Arch Virol (1998) 143: 945–962

Archives of Virology



# Cassava vein mosaic virus (CsVMV), type species for a new genus of plant double stranded DNA viruses?

**A. de Kochko, B. Verdaguer, N. Taylor, R. Carcamo, R. N. Beachy,** and **C. Fauquet**

ILTAB/ORSTOM-TSRI, The Scripps Research Institute, Division of Plant Biology, La Jolla, California, U.S.A.

Accepted December 16, 1997

**Summary.** The complete sequence of 8159 nucleotides of the double stranded DNA genome of cassava vein mosaic virus (CsVMV) was determined (# U59751) and revealed a significant difference in genome organization when compared with a previous report (# U20341). When transferred to cassava plants by microbombardment, the full length CsVMV clone was infectious, confirming the genome organization here described. Sequence comparisons between CsVMV and members of the genera *Caulimovirus* and *Badnavirus* revealed high homologies between consensus sequences of several proteins that are indispensable for virus replication, including a potential transactivator factor not reported previously. The presence of a sequence complementary to a plant Met tRNA confirms that CsVMV is a plant pararetrovirus and is most closely related to members of the genus *Caulimovirus* as previously assessed. However, differences in genome organization, number and size of the ORFs, in addition to sequence comparisons with other plant pararetroviruses, shows that either the genetic variability of caulimoviruses is much greater than previously thought, or that CsVMV is the unique representative of a new genus within the *Caulimoviridae* family. On the basis of this study, it is proposed to upgrade the floating genus *Caulimovirus* to the family level and to divide the *Caulimoviridae* family into at least three genera with CsVMV being the type member of a new genus.

## Introduction

Double stranded (ds) DNA plant viruses are currently classified into one of two floating genera; *Caulimovirus* or *Badnavirus*, of which cauliflower mosaic virus (CaMV) and commelina yellow mottle virus (ComYMV) are the respective type species [24, 25]. All caulimoviruses identified to date are aphid-transmitted, infect only dicotyledonous plants and have isometric particles of about 50 nm in diameter which accumulate in inclusion bodies in the cytoplasm of infected cells.

Their genome is a circular molecule of dsDNA which replicates through a terminally redundant full-length RNA transcript which is used as template by a virus encoded reverse transcriptase [16]. In addition to CaMV, several caulimoviruses have been characterized, including carnation etched ring virus (CERV) [20], figwort mosaic virus (FMV) [32], soybean chlorotic mottle virus (SbCMV) [14], peanut chlorotic streak virus (PCSV) [31] and strawberry vein banding virus (SVBV, X97304).

Members of the genus *Badnavirus* can infect both monocotyledonous and dicotyledonous plant species, and their particles have a bacilliform shape. They are insect transmitted and their dsDNA genome is replicated in a manner similar to that of caulimoviruses [19]. In addition to ComYMV, the badnaviruses sequenced to date include; rice tungro bacilliform virus (RTBV) [15, 22, 30], sugarcane bacilliform virus (ScBV) [3] and cacao swollen shoot virus (CSSV) [13].

Cassava vein mosaic virus (CsVMV) is found primarily in the Amazon basin in Brazil, where it infects cassava plants (*Manihot* spp), it causes only mild symptoms and its incidence in cassava crops is apparently low [23]. Some of its characteristics, such as the shape of the particles, their accumulation in cytoplasmic inclusion bodies of infected cells [23] and the nature and size of its genome, make it a tentative member of the genus *Caulimovirus*. We report here the complete sequence of an infectious clone of CsVMV. Confirmed by sequencing PCR products from naturally infected plants, this sequence differs from that previously published [4], primarily by the addition of a single nucleotide, the result of which modifies the number of predicted open reading frames (ORFs). After microbombardment of cassava plants with cloned DNA, analysis of inoculated plants confirmed the clone to be infectious. Sequence comparisons were used to classify plant viruses and the results obtained agree with the previous classifications made on the basis of serological and biological properties [29]. As for many of the plant dsDNA viruses, little information concerning their biological characteristics is available, therefore it is more convenient to use sequence comparison and molecular properties to establish their phylogenetic relationships. By these means we show that although certain molecular features support the assignment of CsVMV to the genus *Caulimovirus*, major differences in genome organization and coding capacities, suggest either that caulimoviruses comprise several subgenera, CsVMV being the sole member of one subgenus, or that CsVMV is the sole member of a new genus of dsDNA plant viruses in the family *Caulimoviridae* comprising *Caulimovirus*, *Badnavirus* and the new genus.

## Materials and methods

### *Infection*

Micropropagated plantlets of three different cassava cultivars, Bonoua Rouge (Ivory Coast), MMex 55 (Mexico) and MCol 22 (Colombia), were bombarded with gold particles coated with viral DNA linearized by *Bgl*II, using the same parameters as reported earlier for cassava transformation [33]. Three weeks later, cuttings from the plantlets were transferred to magenta boxes containing 30 ml of Murashige and Skoog (Murashige and Skoog, 1962) (MS) basal medium supplemented with 2% (w/v) sucrose and solidified with 7.5 $gl^{-1}$ agar. After a further

7 to 10 days, upper leaves were removed, DNA was extracted and a PCR performed using CV6 and CV7R primers. An amplified fragment was purified and sequenced using CV4 and CV5R primers. Total DNA from PCR positive plants was blotted after digestion by *Bgl* II and hybridized using a probe made of a 978 bp *Xba* I fragment of the CsVMV genome (nts 6385-7238).

### *Sequencing*

A full length clone of the CsVMV genome was kindly provided by Robert J. Shepherd (University of Kentucky). This was cloned into pCKIZ at the *Bgl* II site. Six overlapping fragments were isolated using convenient restriction sites and cloned into pUC119. Automatic sequencing, using the Applied Biosystem model 373A apparatus, was carried out on both strands by means of *Taq* mediated elongation with dye labeled primers. When necessary, PCR reactions were performed and direct sequencing of the PCR product was carried out. The sequence was compiled using the SeqEd software provided by the manufacturer.

In order to verify the sequence of the first isolate, infected tissue was kindly provided by Prof. A. Lima (Universidade Federal do Ceara, Fortaleza-Brazil). Two different sets of primers were designed to amplify the ambiguous region; nucleotides 6 817–6 818 of # U20341. Primer CV4 sequence 5′-GCTCTTCTATTGAAGAC-3′ corresponding to CsVMV nucleotides (nts) 6 449 to 6 465 was used with primer CV5R (5′-CAGCCATTGCACTGTA-3′) which corresponds to nts 6 844 to 6 859 (inverted, complementary), to amplify a fragment of 410 nts. A fragment of 551 nts was amplified between primer CV6 (5′-GTCGAAGAACAATATCAG-3′, nts 6 511–6 528 of CsVMV) and primer CV7R (5′-GTCTTCCATCTAGGTTGGAG-3′, nts 7 062–7 081 reverse complement of CsVMV). These two PCR fragments were sequenced immediately after amplification without cloning in order to avoid possible selection during amplification.

### *Genome analysis and alignments*

Analysis of DNA sequences, including determination of ORFs and derivation of predicted protein sequences, homology searches and multiple alignments, were carried out using programs from the DNASTAR package for Macintosh (DNASTAR Inc., Madison, WI, USA). To predict the functions of the ORFs, the genomic sequence of the Strasbourg strain of CaMV and of one isolate of RTBV were used. For multiple alignments and establishing phylogenetic trees, all available plant pararetrovirus sequences were compared. Sequences were aligned using the clustal method of the Megalign program from the DNASTAR package with a gap penalty of 10 and a gap length penalty of 10. The multiple alignments obtained were used for constructing phylogenetic trees.

Phylogenetic relationships were established by a cladistic parsimony method using the version 3.1.1 of the PAUP software (Phylogenetic Analysis Using Parsimony, D. L. Swofford, The Smithsonian Institution, Washington, DC, USA) by building phylogenetic trees according to the heuristic search, while assessment of the strength of the trees was achieved by performing bootstrap analysis with 100 replicates.

CsVMV sequences were compared with published sequences of the following caulimoviruses and badnaviruses: cauliflower mosaic virus, CaMV Strasbourg strain (J02048); CaMV-cm1841 strain (J02046); and CaMV D/H strain (M10376); carnation etched ring virus, CERV (X04658); figwort mosaic virus, FMV (X06166); strawberry vein banding virus, SVBV (X97304); peanut chlorotic streak virus, PCSV (U13988); soybean chlorotic mottle virus, SbCMV (X15828); three isolates of the rice tungro bacilliform virus, RTBV (M65026, X57924 and D10774); commelina yellow mottle virus, ComYMV (X52938); cacao swollen shoot virus, CSSV (L14546); and sugarcane bacilliform virus, ScBV (M89923).

The ORF1 represents the putative movement protein (MP), of the caulimoviruses and was considered as the standard to which 330 aa from the C terminus of the polyprotein corresponding to ORF1 of CsVMV (aa 900–1230), and 330 aa from the N-terminus of the polyprotein encoded by ORF3 of badnaviruses (aa 1–330) were compared. The core sequence corresponding to the coat protein (CP) was estimated as being 200 aa in length and is upstream to the RNA binding site. A sequence corresponding to the proteinase was estimated to be 120 aa long beginning 23 aa upstream of the consensus DXGX sequence. After a first set of alignments was performed on the estimated full length sequence of the reverse transcriptase (RT), we kept for each virus, a sequence showing a significant level of homology ($\geq 24\%$) of about 260 aa. in length. The inclusion body protein (IBP) of caulimoviruses was compared with the protein encoded by ORF4 of CsVMV, an ORF absent in the former report [4]. In addition, a sequence of 300 nts that ends at the first ATG 3′ to the TATA box was also aligned for phylogenetic analysis. A random sequence of equal length and composition was generated from four different virus sequences using the GeneAlign software [34] and has been included in each alignment.

The PAUP program was unable to perform an accurate bootstrap analysis on multiple alignment established with the entire genomic sequences of all the viruses. After aligning the genomes by pair, we determined that only a segment of about 5500 nts, starting at ORF1 for the caulimoviruses and CsVMV, and ORF3 for the badnaviruses, presented a significant level of homology among all viruses; the remaining part of the genome was not different from random alignment. A bootstrap analysis on such an alignment could then assess the strength of the deduced phylogenetic tree. As CsVMV genome is rearranged in relation to all currently characterized plant pararetroviruses, another alignment was made in which the homologous sequences were arranged in the same order as it exists in the plant pararetroviruses (MP-CP vs. CP-MP).

## Results

### *Infection*

Three out of fourteen plantlets tested of the cultivar MMex 55, one out of six MCol 22 and one out of four Bonoua Rouge plants were PCR positive for CsVMV DNA one month after microbombardment with the linearized clone of the virus. The sequence of the amplified fragment was found to be identical to the corresponding region of the clone used to infect the plants (data not shown). A DNA blot performed on total DNA from a PCR positive plantlet, after digestion by *Bgl* II, a single cutter in the original clone, revealed a band of approximately 8 kb corresponding to the full length genome of CsVMV (data not shown). Virus particles were also identified by electron microscopy in sections of infected tissue (Fig. 1a–b). All PCR positive plants developed symptoms while in tissue culture, including discoloration of the veins, mottling, reduced leaf lamina and necrosis which started from the tip and extended to the whole leaf. Under growth chamber conditions symptom development was very slow, and only leaves more than one month old, produced symptoms similar to those observed in tissue culture. Moreover, necrosis did not extend to the entire leaf (Fig.1c), viral DNA was not detectable in an ethidium bromide stained gel, and virus particles were rare in the infected leaves of these plants.

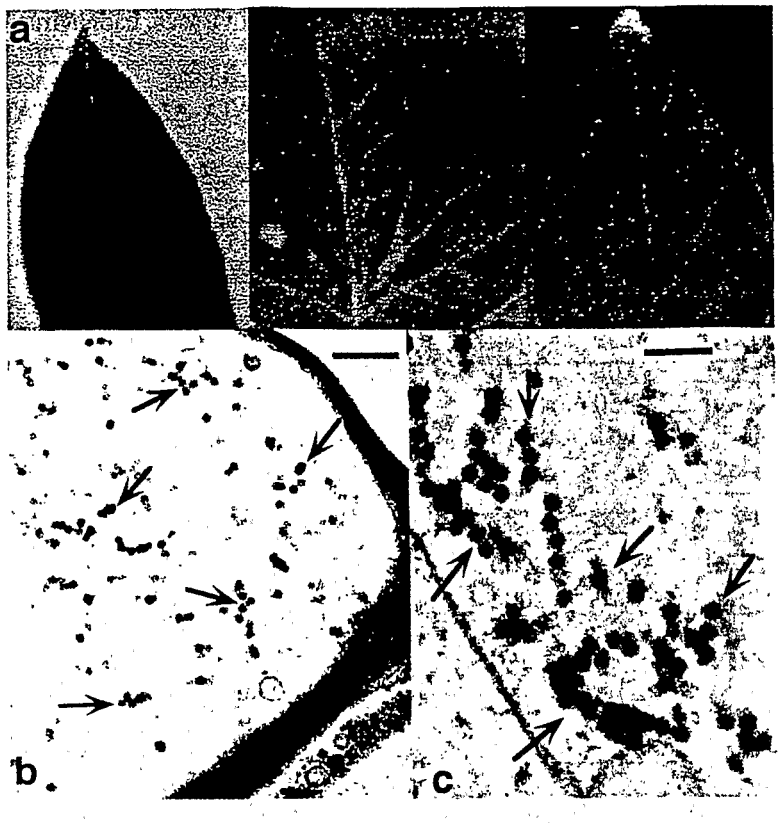


**Fig. 1.** **a** Symptoms of cassava vein mosaic virus developing on an infected cassava plant under growth chamber conditions, from an early stage (left) to a later one (right). Electron micrographs of CsVMV particles in cassava mesophyll cells, **b** in an inclusion body, magnification 40 k, bar = 250 nm and **c** as a cluster in the cytoplasm, magnification 80 k, bar = 125 nm

### *Genome organization*

The CsVMV genome is a circular, double stranded DNA molecule. At 8159 bp long it is slightly longer than that of other caulimoviruses except for those which infect the legumes (PCSV and SbCMV) (Table 1 and Fig. 2) and with the exception of RTBV, is almost 1 kb longer than the genome of badnaviruses. Table 1 shows that the base composition of CsVMV is substantially different from both the caulimoviruses and badnaviruses with a GC content of only 25% compared with 34% to 40% for former and 33% to 44% for latter.

The organization of the CsVMV promoter region is also distinctive in that the distance between the TATA box and the first ATG (which is not necessarily the ATG of the first identified ORF) is slightly greater than for the caulimoviruses (102/66-93). The distance between the AS1 *cis* element and the TATA box is also substantially greater in CsVMV (Table 1), than for many of the caulimoviruses (168/19-142). The significance of these differences has yet to be determined.

The origin of DNA replication with relation to the start of the first ORF, differs among caulimoviruses. Only PCSV and SbCMV have two putative ORFs located

**Table 1.** Comparison of selected genomic features of caulimoviruses, badnaviruses and CsVMV

| | Virus name | GC content % | Genome length nts | TATA box position nts | Distance TATA-ATG nts | Distance AS1-TATA nts | Distance TATA-org. of rep.nts |
|---|---|---|---|---|---|---|---|
| | CaMV Str. | 39.97 | 8024 | 7405-11 | 75 | 34 | 613 |
| | CaMV 1841 | 40.19 | 8031 | 7402-08 | 85 | 34 | 623 |
| | CaMV D/H | 40.11 | 8016 | 7388-94 | 83 | 34 | 622 |
| Caulimo | FMV | 35.37 | 7743 | 6896-6902 | 86 | 19 | 841 |
| viruses | CERV | 36.39 | 7932 | 7052-58 | 93 | 33 | 874 |
| | SVBV | 39.08 | 7876 | 7220-26 | 66 | N/A | 650 |
| | PCSV | 34.50 | 8174 | 6049-55 | 89 | 124 | 2119 |
| | SbCMV[a] | 34.01 | 8175 | 6044-50 | 18 | 142 | 2125 |
| | | | | 6147-53 | 220 | 245 | 2022 |
| CsVMV | CsVMV | 24.93 | 8159 | 7571-77 | 102 | 168 | 582 |
| | RTBV Ph1 | 33.67 | 8000 | 7373-79 | 116 | N/A | 621 |
| | RTBV Ph2 | 33.70 | 8002 | 7373-79 | 116 | N/A | 623 |
| Badna | RTBV Ph3 | 33.54 | 8002 | 7373-79 | 116 | N/A | 623 |
| viruses | ComYMV | 43.40 | 7489 | 7322-29 | 121 | N/A | 160 |
| | CSSV | 44.09 | 7161 | 6962-68 | 95 | N/A | 193 |
| | ScBV | 43.30 | 7568 | 7373-79 | 43 | 226 | 189 |

[a]SbCMV has two putative TATA boxes in the large intergenic region

between the TATA box in the intergenic region and the origin of DNA replication (Fig. 2). CsVMV has greater similarity to the other caulimoviruses than to the badnaviruses, for which the intergenic region is closer to the sites of initiation of DNA replication.

After establishing the sequence of the infectious clone, it could be seen that the genome of CsVMV is organized in five ORFs (Fig.2), and not the six as previously reported [4]. Two Adenine (AA) at position 6817-18, instead of only one A at 6817 described by Calvert et al. induces a frameshift such that ORF4 and 5 of the former description become actually only one ORF (ORF4). There would also appear to be a large intergenic region comprising nts 7453 to 7973 (735 nts long) which contains an eucaryotic consensus TATA box and the promoter of the virus [38].

In addition to ORF1, 2 and 3 previously described [4], the CsVMV genome contains:

– ORF4, 1178 nts long which overlaps by 28 nts with ORF3, it encodes a predicted protein of 46.3 kDa and has a low homology with the IBP of the caulimoviruses.

– An additional, short ORF5 (nts 7973–8136), is found downstream of the intergenic region. It is predicted to encode a protein of 6.3 kDa and has no apparent sequence homology with any other caulimovirus ORF, nor any protein in the standard sequence data bases.

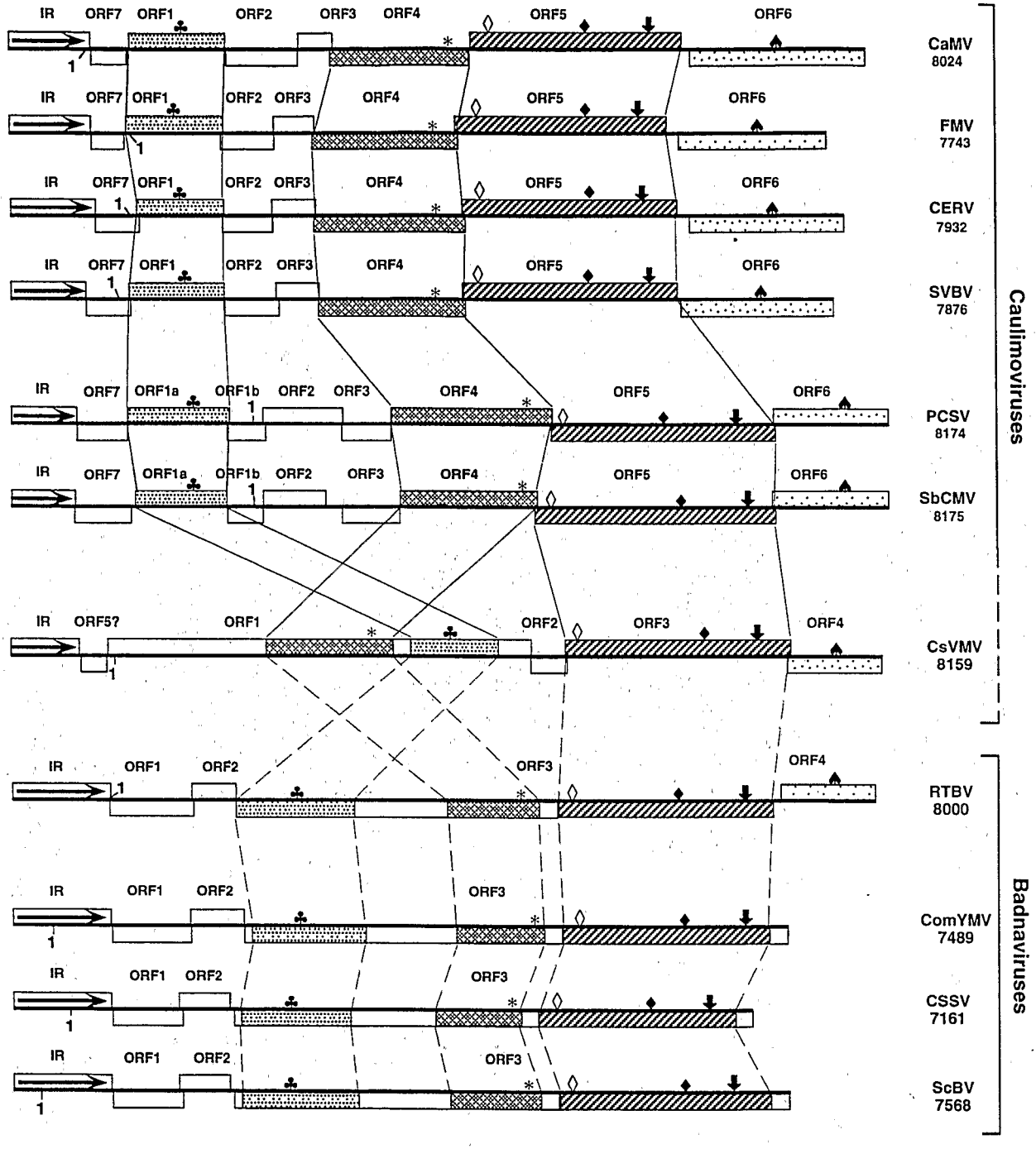


**Fig. 2.** Comparison of the genomic organization of CsVMV with the caulimoviruses and badnaviruses. ORFs or ORF segments encoding similar putative function are linked by vertical lines; the number 1 indicates the origin of DNA replication. ♣ MP active site, * RNA binding site, ◇ PR active site, ◆ RT active site, ♠ TAV active center, ⬇ RNAse H consensus sequence. All maps commence at the beginning of the intergenic region because it is a common region defined just by the sequence and produces a coherent figure. Starting at the site of DNA replication will give a scrambled figure as this site varies considerably among all the viruses. As the initiation of transcription is not known for all of the viruses it cannot be used as starting point

(a): TATA box:

| | | |
|---|---|---|
| CaMV: | TCCTC**TATATAA**GGAAG | nts 7400-7416 |
| FMV: | CCCTC**TATATAA**GAAGG | nts 6891-6907 |
| CERV: | CTGGC**TATATAA**AGGGA | nts 7046-7062 |
| SVBV: | CTCTC**TATATAA**AGAGC | nts 7215-7231 |
| PCSV: | TTGCC**TATAAAT**AAGTT | nts 6044-6060 |
| SbCMV: | TACTT**TATTAAA**AGTGG | nts 6039-6055 |
| SbCMV: | CGCAT**TATAAAT**AAGAG | nts 6142-6158 |
| **CsVMV:** | TTTCC**TATATAA**GGAAC | nts 7566-7582 |
| RTBV: | TCCAG**TATATAA**GGAGC | nts 7368-7384 |
| ComYMV: | TTTCC**TTATTTAA**GCAC | nts 7317-7333 |
| CSSV: | CCATC**TATAAAT**GAGAG | nts 6958-6973 |
| ScBV: | CTGCC**TATATAA**GGCAC | nts 7368-7384 |

(b): Activating Sequence 1 (AS1):

| | | |
|---|---|---|
| OCS(*A.t.*): | AA **ACGTAAGC/GCTTACGT** AC | |
| CaMV: | TG **ACGTAAGG/GATGACGC** AC | nts 7354-7373 |
| FMV: | TG **ACGAACGC/AGTGACGA** CC | nts 6860-6879 |
| CERV: | AG **ACGTCATG/CATGACGT** TT | nts 7001-7020 |
| PCSV: | TG **ACGTAAGG/GCTTACGC** CA | nts 5908-5927 |
| SbCMV: | GA **ACGTCGGC/AATGACGA** AA | nts 5885-5904 |
| **CsVMV:** | AG **ACGTAAGC/ACTGACGA** CA | nts 7386-7405 |
| ScBV: | AG **ACGTAAGC/AATGACGA** TT | nts 130-7149 |

(c): Coat protein (CP) RNA binding site:

| | | |
|---|---|---|
| CaMV: | **CRCWICNIEGHYANECP** | (ORF4 aa 412-428) |
| FMV: | **CRCWICTEEGHYANECP** | (ORF4 aa 409-425) |
| CERV: | **CRCWVCNIEGHYANECP** | (ORF4 aa 418-434) |
| SVBV: | **CRCWICNEIGHFAKDCR** | (ORF4 aa 397-413) |
| PCSV: | **CRCWICQEEGHYANECP** | (ORF3 aa 390-406) |
| SbCMV: | **CQCWLCHEEGHYANECP** | (ORF4 aa 380-396) |
| **CsVMV:** | **CKCYNCGEEGHISPNCK** | (ORF1 aa 739-755) |
| RTBV: | **CRCYICQDENHLANRCP** | (ORF3 aa 772-788) |
| ComYMV: | **CKCYICGQEGHYANQCR** | (ORF3 aa 879-895) |
| CSSV: | **CKCYLCGDEGHFARECP** | (ORF3 aa 786-802) |
| ScBV: | **CRCYVCGSPDHLMKDCK** | (ORF3 aa 736-752) |
| Consensus: | **CXCXXCXXXXHXXXXC** | |

(d): Movement protein (MP):

| | | |
|---|---|---|
| | **iHlgavkilika Fr Gidtp k aliDdRi** | Caulimoviruses consensus |
| CaMV: | VHLGAVKILLKAQFRNGIDTPIKIALIDDRINS | (ORF1 aa 127-159) |
| FMV: | IHLGAVKILLTAQFRQGIDTSVKMALIDDRIVN | (ORF1 aa 124-156) |
| CERV: | IHFGAIKVLIKARFREGINSPIKMALIDDRITD | (ORF1 aa 119-151) |
| SVBV: | IHIGSVKIMIKSTFRTGIDSPISVALLDRRMKN | (ORF1 aa 118-150) |
| PCSV: | IHVNVVQIVIRSTFREGITTPVIIRVEDNRIQD | (ORF1 aa 113-145) |
| SoyMV: | VHISTLQVLIKSTFLKGLDTPLELTLRDNRLLN | (ORF1 aa 102-134) |
| **CsVMV:** | IHLAAVEIVVKAYFREGIDTPFEIILCDDRITY | (ORF1 aa 1001-1033) |
| RTBV: | YHIGMMAIGVKGLHRRKIGTKVMIMFYDDSFGK | (ORF3 aa 113-145) |
| ComYMV: | IHIGVMLVRIQILHRKFAGTMALIVFRDTRWSD | (ORF3 aa 140-172) |
| CSSV: | IHIGILQVRIQILHRQEEGTMALVVFRDNRWSG | (ORF3 aa 140-172) |
| ScBV: | IHPGILAVRIQPLHPDWSGKLVFIVFRDIRDNP | (ORF3 aa 136-160) |
| | **iHiGilavriqiLHr Gtm livFrD rws** | Badnaviruses consensus |

(e): Proteinase (PR):

| | | |
|---|---|---|
| CaMV: | FV **DTGA** SLCIA | (ORF5 aa 43-53) |
| FMV: | YV **DTGA** SLCIA | (ORF5 aa 52-62) |
| CERV: | YV **DTGS** SLCMA | (ORF5 aa 32-42) |
| SVBV: | YV **DTGA** SMCTA | (ORF5 aa 77-87) |
| PCSV: | YI **DTGA** TICLA | (ORF5 aa 21-31) |
| SbCMV: | YI **DTGA** TLCFG | (ORF5 aa 34-44) |
| **CsVMV:** | LF **DTGA** NICIC | (ORF3 aa 24-34) |
| RTBV: | LI **DSGS** THNII | (ORF3 aa 985-995) |
| ComYMV: | IV **DTGA** TACLI | (ORF3 aa 1218-1228) |
| CSSV: | IL **DTGA** TTCCI | (ORF3 aa 1076-1086) |
| ScBV: | LL **DTGA** TRSCI | (ORF3 aa 1081-1091) |

(f): Reverse transcriptase (RT):

| | | |
|---|---|---|
| CaMV: | FCCV **YVDDI** LVFSN | (ORF5 aa 398-411) |
| FMV: | FCMV **YVDDI** IVFSN | (ORF5 aa 390-403) |
| CERV: | YCCV **YVDDI** LVFSN | (ORF5 aa 380-393) |
| SVBV: | FCAV **YVDDI** IVFSK | (ORF5 aa 430-443) |
| PCSV: | ISLA **YIDDI** IVFTK | (ORF5 aa 339-352) |
| SbCMV: | IYLA **YIDDI** LIFTK | (ORF5 aa 354-367) |
| **CsVMV** | FIIV **YIDDI** LVFSK | (ORF3 aa 358-371) |
| RTBV: | FAIL **YIDDI** LIASN | (ORF3 aa 1335-1348) |
| ComYMV: | FIAV **YIDDI** LVFSE | (ORF3 aa 1560-1573) |
| CSSV: | FIAV **YIDDI** LVFSE | (ORF3 aa 1439-1452) |
| ScBV: | FIAV **YIDDI** LVFSE | (ORF3 aa 1412-1425) |

## *Sequence identification*

### Transcriptional promoter

An eukaryotic consensus TATA box is located in the intergenic region at nucleotide position 7571–7577. This region shows a high homology with other pararetroviruses genomes (Fig. 3a) and its function as a strong constitutive promoter has been demonstrated [38]. In addition, a sequence with high homology to the OCS element of *Agrobacterium tumefaciens*, also called activating sequence 1 (AS1), has been identified at position 7388–7403, 168 nts upstream of the TATA box. This sequence is known to activate plant virus promoters [2] and is present in all but one of the caulimoviruses (SVBV) (Fig. 3b) and in only one badnavirus (ScBV). ComYMV has a AS1-like sequence with low homology to the OCS element and a different role to that of caulimoviruses AS1 [26].

### Origin of replication

As in the genomes of other pararetroviruses, CsVMV genome contains a sequence that is complementary to a plant met tRNA. This corresponds to the initiation site of DNA replication used by the viral reverse transcriptase on the viral RNA when primed by the tRNA [16]. The first nucleotide at this site is generally designated nucleotide 1 of pararetrovirus genomes. The origin of replication is located downstream of the large intergenic region in caulimoviruses and RTBV, but inside the large intergenic region in the four other badnaviruses. Only the caulimoviruses which infect legumes have two ORFs between the end of the large intergenic region and the origin of DNA replication, of which ORF1 encodes the putative movement protein (Fig. 2).

### Structural and catalytic functions; coat protein (CP), movement protein (MP), replicase (RT), and proteinase (PR)

The sequence reported here does not differ from the one reported previously with regards to the ORFs containing the consensus for the RNA binding site of the coat protein (CP), the putative movement protein (MP), the proteinase (PR) or the reverse transcriptase (RT). Figure 3c–f shows all these consensus and their position in the different ORFs. The RNA binding site of the CP [1], with the consensus sequence of a zinc finger CXCXXC4XH4XC found in all plant pararetroviruses coat proteins was identified in CsVMV at amino acids (aa) 739–755 of the polyprotein encoded by ORF1.

**Fig. 3.** **a** Homologies among the TATA box sequences of different plant pararetroviruses. Some are putative. The first six sequences are caulimoviruses, the following four are badnavirus sequences. For SbCMV two putative TATA boxes were proposed. **b** Sequence homologies between the activating sequence 1 (AS1) of different caulimoviruses and the corresponding palyndromic sequence of the octopin synthase gene of *Agrobacterium tumefaciens*. Homologous sequences do not exist in badnaviruses except for ScBV. Consensus sequences for: **c** coat protein RNA binding site, **d** putative movement proteins (MP), **e** proteinase (PR) and **f** replicase (RT) active site of plant pararetroviruses

In addition, a significant aa similarity (up to 61%) with the consensus sequence of the caulimoviruses movement proteins (MP) [27], comprising the RNA binding domain of the CaMV MP [35], was found in the aa 1001–1033 of the polyprotein encoded by ORF1. The most peculiar aspect of the CsVMV genome is the respective order of the CP and MP coding sequences compared with the genomes of both the caulimoviruses and badnaviruses. All viruses from both genera encode the MP upstream of the CP; while CsVMV apparently encodes first the CP, then the MP (Fig. 2). This unusual order has also been reported by Calvert et al. [4], who confirmed this genomic organization by sequencing PCR products obtained from CsVMV infected tissues.

ORF3 of CsVMV corresponds to ORF5 of caulimoviruses, and encodes a polyprotein in which the consensus sequence of the aspartic proteinase DTGA is located (aa 26–29) (see Fig. 3e) [28, 36]. This ORF also includes the consensus sequence YI/VDDI represented in the reverse transcriptase of plant pararetroviruses at aa 364-366 (Fig. 3f) [9].

It is interesting to note that for CsVMV, in contrast to the general situation found in retroelements, the equivalent of the *gag* (CP) and *pol* (RT) ORFs are not found on the same ORF nor in the two overlapping ORFs [8, 17].

## Inclusion body protein (IBP) or transactivator (TAV)

The inclusion body protein (IBP) encoded by ORF6 of CaMV is reported to be involved in several functions during the virus cycle, the most important being the main component of the viral inclusion body matrix [6] and the regulating translation of polycistronic mRNA. In the latter function, the product of ORF6 is also termed transactivator (TAV) [7]. The mRNA encoding ORF6 is transcribed by the 19S promoter in CaMV [11]. An alignment performed with the encoded products of caulimoviruses ORF6 and CsVMV ORF4, showed a low level of similarity (from 11.5% to 13%) which increases from 13% to 20% (11% to 13% for the random sequence) when a core sequence of 91aa, including the putative active center of the TAV defined by De Tapia et al. [7], is considered (Fig. 4a). We assume that CsVMV ORF4 encodes a protein with a role equivalent to the IBP (or TAV), despite the molecular weight differences with the caulimoviruses IBP (46 kDa compared to 56–61 kDa). This function was not identified in the previous report of CsVMV sequence [4].

A region of 17 aa located in the ORF2 of CsVMV which encodes a 71aa protein, was found to be homologous (up to 58% similarity) to a sequence in CaMV, FMV and CERV ORF6 (Fig. 4b). However, this region of homology does not cover the putative active center of the TAV [7], indicating that this similarity may not have any biological significance. Interestingly, sequences homologous to the IBP were also found in ORF4 of RTBV (Fig. 5). This was the only badnavirus found to possess a fourth ORF at the same genomic location as ORF6 of caulimoviruses. The product of RTBV ORF4 has no known function but it is assumed that it does not act as a translational transactivator, equivalent to the CaMV TAV [10].

```
a CsVMV:   YSLTDYNKLVADIYTDRNLV   (ORF4 aa 123-142)
  CaMV:    NYYVVYNGPHAGIYDDWGCT   (ORF6 aa 139-158)
  FMV:     SWFAVYKGPNKEFFTEWEIV   (ORF6 aa 155-174)
  CERV:    DFYVVYNGPYAGIYDHWGTA   (ORF6 aa 133-152)
  SVBV:    KTYVIYDGPNQGIYDSWALV   (ORF6 aa 145-164)
  PCSV:    RYYVIYNGPGKGIYDEWGKA   (ORF5 aa 79-98)
  SbCMV:   KAYVIFDGPWKGIYQDWHIV   (ORF6 -1 aa ?)

           yv ynGP  giyd w        Caulimoviruses consensus

b CsVMV:     IKRFNKHLIKTTYKRIF   (ORF2 aa 44-60)
  CaMV:      VKRFRTNCIKNTEKDIF   (ORF6 aa 324-340)
  FMV:       IKNFRKKVLNAKDAAIF   (ORF6 aa 316-332)
  CERV:      VQRFRKKCIKDSEKEIF   (ORF6 aa 315-331)

             krFrk  ik   k IF    consensus
```

**Fig. 4.** **a** Sequence homologies found between caulimovirus ORF6, encoding for the inclusion body protein (IBP), and the protein encoded by ORF4 of CsVMV. The presumed active site is indicated in bold. For SbCMV the consensus is found in a frameshift (−1) ORF as mentioned by De Tapia et al. [7]. **b** Short homologies found between the IBP of caulimoviruses and the 8.8 kDa protein encoded by ORF2 of CsVMV

```
consensus          P       aGL   IYPs NLqE          i s IP  W

CaMV: ORF6 aa 289 PQMVREAYAAGLIKTIYPSNNLQE 312-30-343 IRSTIPV-W 350
FMV:  ORF6 aa 281 PYLVHTAFRAGLAKVIYPSPNLQE 304-30-335 IFSSIPD-W 342
CERV: ORF6 aa 280 PEGILESFKAGLVRFIYPSTNLQE 303-30-334 IKSTIPC-W 341
SVBV: ORF6 aa 298 PEYVREAFHCGLISLIYPGENLKE 321-42-364 DIL-IPYHW 370

RTBV: ORF4 aa 232 PRFIGDLYAHGFIKQINFTTKVPE 255-32-278 IEREIPR-W 285
```

**Fig. 5.** Sequence homologies between the encoded product of caulimoviruses ORF6 and RTBV ORF4

## Other functions

Computer assisted searches for other functions that are characteristic of the caulimoviruses, including the insect transmission factor (ITF) encoded by their ORF2, and the DNA binding protein encoded by ORF3, did not lead to the identification of such sequences in the CsVMV genome.

DNA binding protein of caulimoviruses (ORF3) and badnaviruses (ORF2) share little similarity. The binding domain of these proteins is characterized by a lysine rich region containing proline residues, generally specific to AT rich DNA [5, 21], and can be identified at nts 221–229 of CsVMV ORF1. Furthermore, two hydrophilic regions, rich in leucine and isoleucine residues located upstream of the previous motif, may constitute a double leucine zipper (nts 136–157 and 198–219), the presence of which suggests a binding activity of the N terminus of the polyprotein encoded by ORF1 (Emmanuel Jacquot, pers. comm.). The binding activity of this ORF1 domain has still to be confirmed by biological assay.

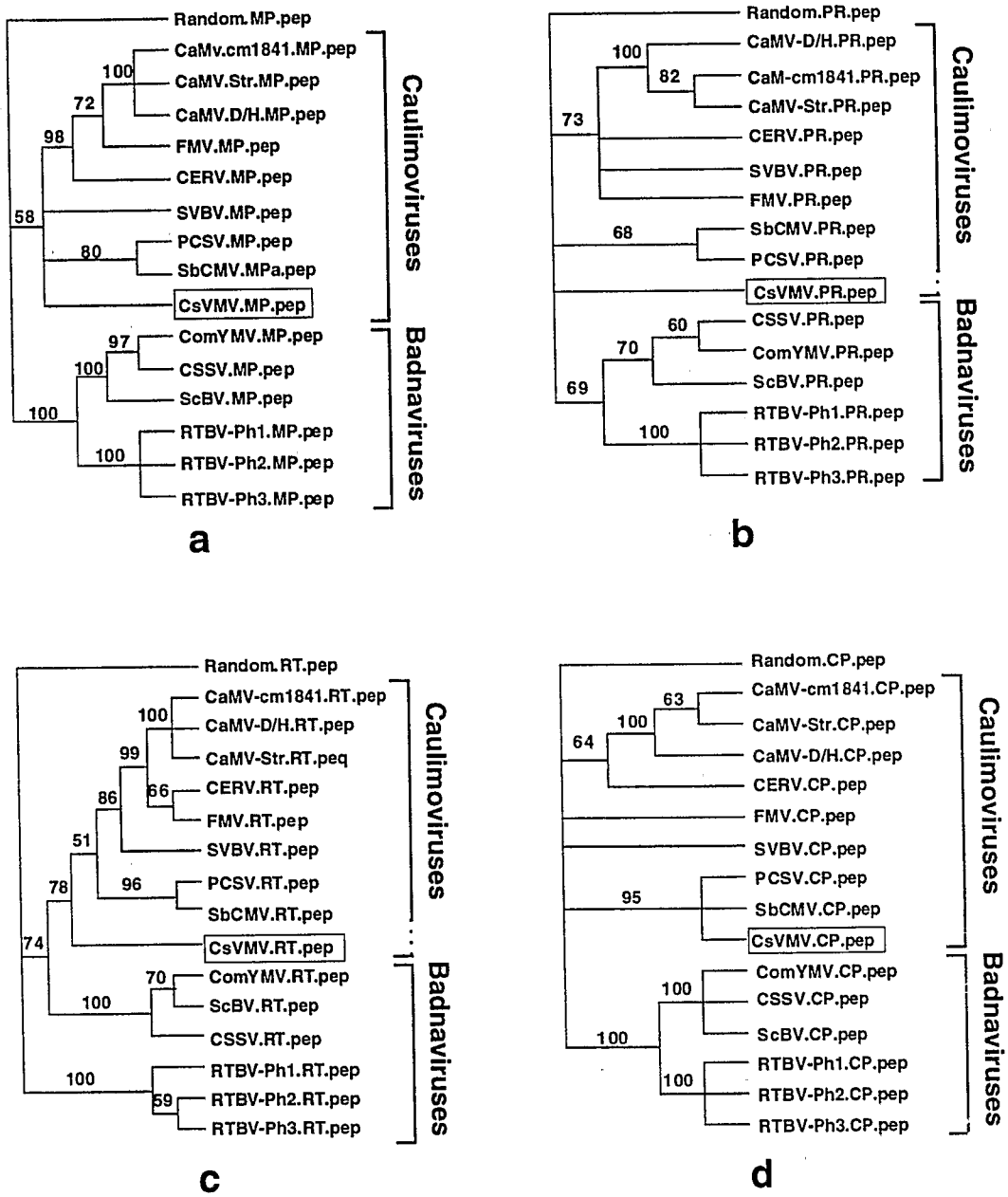


**Fig. 6.** Phylogenetic trees of caulimoviruses, badnaviruses and CsVMV based upon comparison of aa sequences of movement proteins (**a**), proteinases (**b**), reverse transcriptases (**c**), and coat proteins (**d**). Numbers on the branches are the bootstrap values after 100 replicates

No homologies were found in sequence data bases for the remaining part of ORF1, for ORF2 or the putative ORF5. It is interesting to note that a significant level of homology (60% similarity) was found at the ORF1 N-terminus of CsVMV (aa 20 to 68) with an absissic acid inducible protein from barley (aa 17 to 64). However, no role has yet been assigned to this protein [12].

### Multiple alignments and phylogenetic relationships

The phylogenetic tree obtained after alignment of the MP sequences (Fig. 6a) shows a distinct separation between the *Caulimovirus* and *Badnavirus* genera. CsVMV clusters with the caulimoviruses but on a branch independent from other members of the genus. Using the alignment corresponding to the PR sequences (Fig. 6b), a tree is obtained which classifies CsVMV in neither of the recognized genera. However, this is also the case for the legume infecting caulimoviruses. The RT sequences are the most conserved regions among the two genera and their alignment produces a phylogenetic tree in which the main distinction concerns RTBV, which branches apart from the cluster containing all other plant pararetroviruses, and in which CsVMV is the last to be aggregated to the caulimovirus ensemble (Fig. 6c). The tree obtained from the CP sequences alignment clusters CsVMV with the legume caulimoviruses (Fig. 6d). However, this observation is somewhat muted by the fact that the genus *Caulimovirus* is not uniform as far as CP sequences are concerned. The trees given by the IBPs and the promoters (data not shown) show an heterogeneous distribution, allowing no clear relationship to be structured, even among members of the accepted genera.

The tree generated from the comparison of a 5500 nts fragment of the genomes of these viruses clearly reflects the distinction suggested by the protein sequence alignments (Fig. 7). This distinction remains even when the CsVMV sequence is manually adjusted to have the same function order as in the other pararetrovirus genomes (MP–CP). The two genera, *Caulimovirus* and *Badnavirus*, each divided in two subclusters, are strongly differentiated, while among the caulimoviruses a distinction can be made between those infecting legumes and the others. Similarly, RTBV is clustered apart in the genus *Badnavirus*. CsVMV stands on a separate branch and does not cluster with viruses from either of the genera.

## Discussion

The infectivity of the CsVMV clone used in this study indicates that all functions necessary for the virus cycle in the plant, with the possible exception of vector transmission, are present and functional, and that the sequence and genome organization presented here corresponds to a biologically active virus. The slow development of symptoms on infected plants, the low quantity of virus particles, and the low amount of viral DNA extracted from infected leaves, all indicate a low level of virus replication under the growth chamber conditions employed in this study. It is possible however that this observation is the result of an inefficient replication of the clone used, or that, in a natural environment, in different light and/or temperature conditions, the virus replication would be more efficient. Nevertheless, as this virus is not a major threat to cassava cultivation, it is possible that even in its natural environment, replication of CsVMV occurs at low level.

The isometric shape of its particles, their accumulation in cytoplasmic inclusion bodies of infected cells [23] and a genome consisting of a dsDNA molecule, led previous authors to classify CsVMV as a tentative member of the *Caulimovirus* genus [18]. However, very little is known about the biology of this virus, such as

its natural vector, serology or tissue tropism. The CsVMV promoter, in a manner similar to *Caulimovirus* promoters, displays a constitutive pattern of expression in transgenic plants [38] and virus particles have been observed in most types of infected cassava tissue [23].

The low G+C content of the CsVMV genome (25%) was shown to be very different from that of the genomes of caulimoviruses (34–40%) and badnaviruses (34–44%) (Table 1), which could have significance for its classification and evolution, as this criterion is usually uniform within a viral genus. The genome organization of CsVMV shows some similarities both with caulimoviruses and badnaviruses. A long intergenic region in which the replicative promoter is located is common to both genera and to CsVMV, as is the site of initiation of DNA replication with the binding site for Met tRNA which primes the synthesis of the minus strand DNA using the greater-than-full-length RNA as a template. However, location of the DNA replication initiation site differs greatly, not only between genera but also within a genus. In caulimoviruses genomes, it immediately follows the large intergenic region; except for those that infect legumes, in which the DNA replication initiation site is about 2000 nts downstream of the large intergenic region. For badnaviruses, the DNA replication initiation site is located inside the intergenic region, except for RTBV in which it follows the intergenic region, in the manner similar to caulimoviruses (Table 1 and Fig. 2). However, the short intergenic region containing a second promoter responsible for the transcription of ORF6 which is characteristic of the caulimoviruses (19S of CaMV) was not identified on the CsVMV genome. This would make it comparable to the badnaviruses. A second promoter might be present in an ORF, but to identify such a putative second promoter, it would be necessary to carry out in vitro transcription analyses.

Some of the functions encoded by the caulimoviruses genomes were identified in CsVMV by consensus homologies (MP, CP, PR and RT), but other important functions, such as the insect transmission factor (ITF) were not found. However, the product of ORF2 in caulimoviruses (ITF) is not well conserved among members of the genus. Biological studies are needed to clarify the exact role of unassigned ORFs and of some regions of the polyprotein encoded by ORF1 of CsVMV, such as the putative DNA binding domain.

The organization of CP-*pol* functions in CsVMV genome (equivalent to *gag-pol* of retroviruses), makes this virus unusual amongst pararetroviruses. In all others pararetroviruses, retroviruses and retrotransposons [8, 17] the sequences are ordered *gag-pol*, while in CsVMV this order is modified (Fig. 2). Furthermore, the arrangement of MP with respect to CP is unique. The CsVMV genome also contains an additional putative ORF between ORF1 (*gag*) and ORF3 (*pol*) making CsVMV a unique representative among all the described retroelements. It is possible that the low homology shown by the short ORF2 with the IBP of caulimoviruses indicates that this ORF is the result of an intragenomic recombination event which lead to ORF2 and ORF4.

Construction of phylogenetic trees on the basis of homologous sequences yield different results according to the sequences that are compared (Fig. 6a–d).

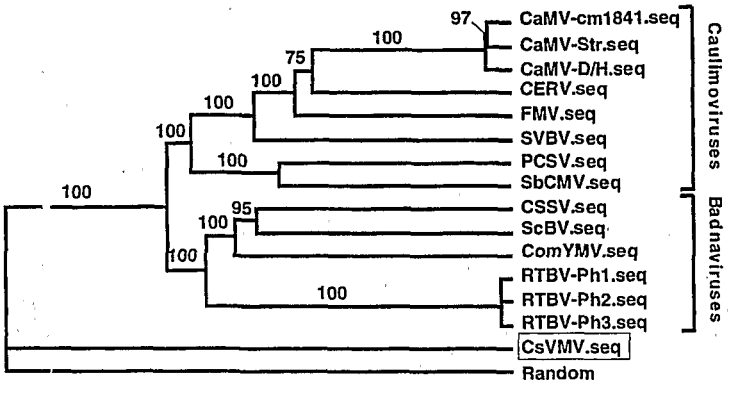


**Fig. 7.** Phylogenetic tree of plant dsDNA viruses obtained upon comparison of 5 500 nts starting at ORF1 for caulimoviruses and CsVMV and ORF3 for badnaviruses

The relationships indicated by the trees obtained from MP and RT sequences link CsVMV with the caulimoviruses. RT sequences, the most conserved sequence among the dsDNA viruses, also place CsVMV within this genus. On the contrary, the tree produced from the PR sequences, which are also well conserved, indicates no relationship with either genera; and this is also the case for the caulimoviruses which infect the legumes. The tree constructed from CP sequences shows a clustering of CsVMV with PCSV and SbCMV. This is the least informative tree in terms of relationships for the caulimoviruses, but it does group all the badnaviruses together. The tree obtained with genomic sequences (5500 nts) (Fig. 7) clearly shows a distinct position for CsVMV that is related neither to the caulimoviruses nor to the badnaviruses.

Even with the few representatives sequenced to date, it can be seen that the genus *Caulimovirus* is not uniform in genome length, the position of the intergenic region nor for the initiation site of DNA replication. Additionally, their genomes also show very low sequence similarities within certain ORFs. These differences are clearly reflected on the current phylogenetic trees (Fig. 6 a–d).

Despite some common biological properties between CsVMV and the members of the *Caulimovirus* genus, on the basis of the results presented in this study, we propose to create a new family named *Caulimoviridae*, and comprising at least three genera; *Caulimovirus*, *Badnavirus* and a new genus with CsVMV as type species and unique member.

All the members of the *Caulimovirus* and *Badnavirus* genera, as well as CsVMV, are poorly related on a molecular basis, with sequence comparisons somewhat at the limit of interpretation for phylogenetic relationships. Parts of their sequence have no detectable relationships, but some domains of their genome, encoding equivalent non-structural (MP, PR, RT) or structural (CP) functions, are conserved enough to show significant degrees of relatedness. Part of the observed genetic variability among plant pararetroviruses, therefore, may be the result of cumulative mutations after an early differentiation between caulimoviruses and badnaviruses. Alternatively, the example of CsVMV's unusual genomic organi-

zation (Fig. 2) clearly shows that genome recombination has been a major factor in the evolution of this virus. Within a virus genus, genomic organization is generally very well conserved and no changes in the order of domains are observed [37]. CsVMV is very unusual in this respect, by the simple fact that the order of the encoded CP and MP is reverse to that in all the other plant dsDNA viruses. In addition the *gag-pol* continuum is disorganized compared to all the retroelements. This atypical situation could have resulted from intra/intergenomic recombination. It would appear therefore that both mutation and recombination mechanisms led to the appearance of CsVMV, a unique virus representing a very unusual type of plant pararetrovirus.

## Acknowledgements

The authors are very grateful to Dr. Robert J. Shepherd for providing the full length clone of CsVMV and Prof. Ambrosio A. Lima for sending CsVMV infected tissues. They are also grateful to Dr. Lisa Bibbs for managing the automatic sequencing and Dr. Malcolm Wood and Mr. Laurent Gillot for their help with the electron microscope. This work was supported by the French Institute for Scientific Research for Development in Cooperation (ORSTOM) and the Rockefeller Foundation.

## References

1. Berg JM (1986) Potential metal-binding domains in nucleic acid binding proteins. Science 232: 485–487
2. Bouchez D, Tokuhisa JG, Llewellyn DJ, Dennis ES, Ellis JG (1989) The ocs-element is a viral component of the promoters of several T-DNA and viral plant genes. EMBO J 8: 4 197–4 204
3. Bouhida M, Lockhart BEL, Olszewski NE (1993) An analysis of the complete sequence of a sugarcane bacilliform virus genome infectious to banana and rice. J Gen Virol 74: 15–22
4. Calvert LA, Ospina MD, Shepherd RJ (1995) Characterization of cassava vein mosaic virus: a distinct plant pararetrovirus. J Gen Virol 76: 1 271–1 276
5. Churchill MEA, Travers AA (1991) Protein motifs that recognize structural features of DNA. Trends Biochem Sci 16: 92–97
6. Covey SN, Hull R (1981) Transcription of cauliflower mosaic virus DNA. Detection of transcripts, properties and location of the gene encoding the virus inclusion body protein. Virology 111: 463–474
7. De Tapia M, Himmelbach A, Hohn T (1993) Molecular dissection of the cauliflower mosaic virus translation transactivator. EMBO J 12: 3 305–3 314
8. Doolittle RF, Feng DF, McClure MA, Johnson MS (1990) Retrovirus phylogeny and evolution. Curr Top Microbiol Immunol 157: 1–18
9. Fütterer J, Hohn T (1987) Involvement of nucleocapsids in reverse transcription: a general phenomenon? Trends Biol Sci 12: 92–95
10. Fütterer J, Potrykus I, Valles Brau MP, Dasgupta I, Hull R, Hohn T (1994) Splicing in a plant pararetrovirus. Virology 198: 663–670.
11. Guilley H, Dudley RK, Jonard G, Balazs E, Richards KE (1982) Transcription of cauliflower mosaic virus DNA: detection of promoter sequences, and characterization of transcripts. Cell 30: 763–773
12. Gulli M, Maestri E, Hartings H, Raho G, Perrotta C, Devos KM, Marmiroli N (1995) Isolation and characterization of absisic acid inducible genes in barley seedlings and their responsiveness to environmental stress. Life Sci Adv – Plant Physiol 14: 89–96

13. Hagen LS, Jacquemond M, Lepingle A, Lot H, Tepfer M (1993) Nucleotide sequence and genomic organization of cacao swollen shoot virus. Virology 196: 619–628
14. Hasegawa A, Verver J, Shimada A, Saito M, Goldbach R, Van Kammen A, Miki K, Kameya-Iwaki M, Hibi T (1989) The complete sequence of soybean chlorotic mottle virus DNA and the identification of a novel promoter. Nucleic Acids Res 17: 9 993–10 013
15. Hay JM, Jones MC, Blakebrough ML, Dasgupta I, Davies JW, Hull R (1991) An analysis of the sequence of an infectious clone of rice tungro bacilliform virus, a plant pararetrovirus. Nucleic Acids Res 19: 2 615–2 621
16. Hohn T, Hohn B, Pfeiffer P (1985) Reverse transcription in CaMV. Trends Biol Sci 8: 205–209
17. Hull R (1992) Genome organization of retroviruses and retroelements: evolutionary considerations and implications. Semin Virol 3: 373–382
18. Hull R (1995) *Caulimovirus*. In: Murphy FA, Fauquet CM, Bishop DHL, Ghabrial SA, Jarvis AW, Martelli GP, Mayo MA, Summers MD (eds) Virus Taxonomy. Classification and Nomenclature of Viruses. Sixth Report of the International Committee on Taxonomy of Viruses. Springer, Wien New York, pp 189–192 (Arch Virol [Suppl] 10)
19. Hull R (1996) Molecular biology of rice tungro viruses. Annu Rev Phytopathol 34: 275–297
20. Hull R, Sadler J, Longstaff M (1986) The sequence of carnation etched ring virus DNA: comparison with cauliflower mosaic virus and retroviruses. EMBO J 5: 3 083–3 090
21. Jacquot E, Hagen LS, Jacquemond M, Yot P (1996) The open reading frame 2 product of cacao swollen shoot badnavirus is a nucleic acid-binding protein. Virology 225: 191–195
22. Kano H, Koizumi M, Noda H, Hibino H, Ishikawa K, Omura T, Cabautan PQ, Koganazawa H (1992) Nucleotide sequence of capsid protein of rice tungro bacilliform virus. Arch Virol 124: 157–163
23. Kitajima EW, Costa AS (1966) Particulas esferoidais associadas ao virus do mosaico das nervuras da mandioca. Bragantia 25: 211–221
24. Lockhart BE (1990) Evidence for a double-stranded circular genome in a second group of plant viruses. Phytopathology 80: 127–131
25. Medberry SL, Lockhart BE, Olszewski NO (1990) Properties of Commelina yellow mottle virus's complete DNA sequence, genomic discontinuities and transcript suggest that it is a pararetrovirus. Nucleic Acids Res 18: 5 505–5 513
26. Medberry SL, Olszewski NE (1993) Identification of cis elements involved in Commelina yellow mottle virus promoter activity. Plant J 3: 619–626
27. Mushegian AR, Koonin EV (1993) Cell to cell movement of plant viruses. Insights from amino acid sequence comparison of movement proteins and from analogies with cellular transport systems. Arch Virol 133: 239–257
28. Oroszlan S, Luftig RB (1990) Retroviral proteinase. Curr Top Microbiol Immunol 157: 153–185
29. Padidam M, Beachy RN, Fauquet CM (1995) Classification and identification of geminiviruses using sequence comparisons. J Gen Virol 76: 249–263
30. Qu R, Bhattacharyaa M, Laco G, Kochko de A, Subba Rao BL, Kaniewska M, Elmer JS, Rochester DE, Smith CE, Beachy RN (1991) Characterization of the genome of rice tungro bacilliform virus: comparison with commelina yellow mottle virus and caulimoviruses. Virology 185: 354–364
31. Reddy DVR, Richins RD, Rajeshwari R, Iizuka N, Manohar SK, Shepherd RJ (1993) Peanut cholorotic streak virus, a new caulimovirus infecting peanuts (*Arachis hypogaea*) in India. Phytopathology 83: 129–133

32. Richins RD, Scholthof HB, Shepherd RJ (1987) Sequence of figwort mosaic virus DNA (caulimovirus group). Nucleic Acids Res 15: 8 451–8 466
33. Schöpke C, Taylor N, Carcamo R, Konan NK, Marmey P, Henshaw GG, Beachy R, Fauquet C (1996) Regeneration of transgenic cassava plants (*Manihot esculenta* Crantz) from microbombarded embryogenic suspension cultures. Nature Biotechnology 14: 731–735
34. Sobel E, Martinez HM (1985) A multiple sequence alignment program. Nucleic Acids Res 14: 363–374
35. Thomas CL, Maule AJ (1995) Identification of structural domains within the cauliflower mosaic virus movement protein by scanning deletion mutagenesis and epitope tagging. Plant Cell 7: 561–572
36. Toruella M, Gordon K, Hohn T (1989) Cauliflower mosaic virus produces an aspartic proteinase to cleave its polyproteins. EMBO J 8: 2 819–2 825
37. Van Regenmortel MHV, Bishop DHL, Fauquet C, Mayo MA, Malinoff J, Calisher CH (1997) Guidelines to the demarcation of virus species. Arch Virol 142: 1 505–1 518
38. Verdaguer B, Kochko de A, Beachy RN, Fauquet C (1996) Isolation and expression in transgenic tobacco and rice plants of the cassava vein mosaic virus promoter. Plant Mol Biol 31: 1 129–1 139

Authors' address: Dr. C. M. Fauquet, ILTAB, Division of Plant Biology, BCC206, 10550 North Torrey Pines Road, La Jolla, CA 92037, U.S.A.

Received July 31, 1997