J Mol Evol (1994) 39:496–505

Journal
of Molecular
Evolution

Phylogenetic Relationships Reveal Recombination Among Isolates of Cauliflower Mosaic Virus

Kelly D. Chenault,* Ulrich Melcher

Department of Biochemistry and Molecular Biology, Oklahoma State University, Stillwater, OK 74078, USA

Received: 25 June 1993 / Revised: 24 April 1994 / Accepted: 8 May 1994

Abstract. Isolates of cauliflower mosaic virus (CaMV) differ in host range and symptomatology. Knowledge of their sequence relationships should assist in identifying nucleotide sequences responsible for isolate-specific characters. Complete nucleotide sequences of the DNAs of eight isolates of CaMV were aligned and the aligned sequences were used to analyze phylogenetic relationships by maximum likelihood, bootstrapped parsimony, and distance methods. Isolates found in North America clustered separately from those isolated from other parts of the world. Additional isolates, for which partial sequences were available, were incorporated into phylogenetic analysis of the sequences of genome segments corresponding to individual protein coding regions or the large intergenic region of CaMV DNA. The analysis revealed several instances where the position of an isolate on a tree for one coding region did not agree with the position of the isolate on the tree for the complete genome or with its position on trees for other coding regions. Examination of the distribution of shared residue types of phylogenetically informative positions in anomalous regions suggested that most of the anomalies were due to recombination events during the evolution of the isolates. Application of an algorithm that searches for segments of significant length that are identical between pairs of isolates or contain a significantly high concentration of polymorphisms suggested two additional recombination events between progenitors of the isolates studied and an event between the XinJing isolate and a CaMV not represented in the data set. An earlier phylogenetic origin for CaMV than for carnation etched ring virus, the caulimovirus used as outgroup in these analyses, was deduced from the position of the outgroup with North American isolates in some trees, but with non–North American isolates in other trees.

Key words: Caulimovirus — Pararetroviruses — Recombination — Reverse transcription — Viral isolates

Introduction

Cauliflower mosaic virus (CaMV), the type member of the caulimovirus group of plant viruses (Shepherd 1989), has a double-stranded circular DNA genome of 8 kbp packaged in icosahedral particles of 50-nm diameter. Aphids transmit CaMV to cruciferous and solanaceous plants. In plants, the CaMV genome serves as template for transcription of a 35S RNA. The RNA has a long 5′ untranslated region (intergenic region) and seven open reading frames which, respectively, encode a protein needed for movement of the infection from cell to cell, an aphid acquisition factor, a DNA binding protein of unknown function, a capsid precursor protein, a reverse transcriptase, an inclusion body matrix protein, and a nonessential protein of unknown function. Genome replication occurs by reverse transcription of the 35S RNA (Mason et al. 1987). Such replication is typical of pararetroviruses, which also include the hepadna- and badnaviruses, and is similar to that of retroviruses in which the RNA transcript is packaged into virions.

* *Present address:* Department of Plant Pathology, Oklahoma State University, Stillwater, OK 74078, USA
Correspondence to: U. Melcher

Table 1. Geographic and plant sources of cauliflower mosaic virus isolates

Isolate	Geographic source	Plant source	Reference	Accession number
Bari 1	Bari, Italy	*Diplotaxis tenuifolia*	(Hull 1980)	D00335
BBC[a]	California, USA	*Brassica rapa*	(Chenault and Melcher 1993)	M90542
Cabbage B-JI[a]	Wisconsin, USA	*Brassica* sp.	(Hull 1980)	—
Cabbage S[a]	Bari, Italy	*B. ruvo*	(Franck et al. 1980)	J02048
Campbell	California, USA	*B. oleracea*	(Woolston et al. 1983)	M17415
CM4-184[a]	California, USA	*Brassica* sp.	(Dixon et al. 1986)	M10385
CM1841[a]	California, USA	*B. campestris*	(Sanger et al. 1991)	J02046
CMV-1[a]	California, USA	—	(Stenger et al. 1988)	M90543
D-4	California, USA	*B. campestris*	(Schoelz et al. 1986)	M23620
D/H[a]	Budapest, Hungary	*B. oleracea*	(Sanger et al. 1991)	J02047
NY8153[a]	New York, USA	*Brassica* sp.	(Lung and Pirone 1972)	M90541
PV147	Wisconsin, USA	*B. rapa*	(Shepherd et al. 1970)	X53860
S-Japan	Yokohama, Japan	*Armoracia rusticana*	(Sanger et al. 1991)	X14911
W	California, USA	—	(Choe et al. 1985; Walden and Howell 1983)	M32811
W260	Mendoza prov., Argentina	Unspecified Crucifer	(Gracia and Shepherd 1985)	M94887, L09053
XinJing[a]	XinJiang, China	*B. oleracae*	(Sanger et al. 1991)	(Rongxiang et al. 1985)

[a] Isolates whose complete genomic sequence is known

CaMV isolates have been found in all parts of the world in many cruciferous and solanaceous plants (Gracia and Shepherd 1985; Hull 1980). Isolates vary in infectivity to solanaceous species (Schoelz and Shepherd 1988) and in the symptoms they cause on common hosts such as turnip (*Brassica rapa* L.) and *Arabidopsis thaliana* Heyn (Melcher 1989). Some isolates act synergistically with other viruses in disease development, while others do not (Melcher et al. 1992). The genomes of nine isolates have been completely sequenced (Table 1) and partial sequences of several others are available. Nucleotide sequences of the isolates differ from one another in about 5% of the nucleotide positions (Balàzs et al. 1982; Chenault and Melcher 1994). Experiments with chimeric CaMV DNAs have identified regions associated with some isolate-specific CaMV properties (Daubert et al. 1984; Schoelz et al. 1986; Stratford and Covey 1989). Further identification of nucleotide sequences responsible for isolate-specific phenotypes should be aided by knowledge of the phylogenetic relationships among CaMV isolates.

Phylogenetic analysis of CaMV nucleotide sequences may be complicated by high substitution frequencies and recombination. Viruses that use reverse transcription in their replication cycles have high rates of nucleotide substitution (Steinhauer and Holland 1987). Estimates for HIV-1 are in the range of 10^{-3} changes per nucleotide per cycle (Gojobori et al. 1990). Such frequent mutation could obscure phylogenetic relationships among isolates. If frequent mutations occur, isolate sequences should be distributed in sequence space as a quasi-species population, much as are members of a virion population obtained from a single individual (Holland et al. 1992). For CaMV, substitution frequencies may be sufficiently low so as to not obscure phylogenetic relationships. For one complete growth cycle in a plant, $4–6 \times 10^{-4}$ changes per nucleotide were observed (Pennington and Melcher 1993). Further, sequence analysis of CaMV virion DNA did not identify ambiguities due to population heterogeneity (Balàzs et al. 1982; Franck et al. 1980). The lower substitution frequencies for caulimoviruses than for RNA-containing viruses are supported by phylogenetic comparisons of proteins involved in the movement of viral infections from cell to cell. The caulimoviral proteins have diverged substantially less from a common ancestor than similar proteins of RNA-containing viruses (Melcher 1990).

Recombination between CaMV genomes may also complicate phylogenetic analysis. Recombination occurs when plants are inoculated with pairs of CaMV DNAs, each member of the pair having a lethal mutation in a different gene (Choe et al. 1985; Howell et al. 1981). Resulting diseased plants harbor recombined CaMV genomes. Analysis of the recombination junctions suggested a major role for reverse transcription in recombination (Vaden and Melcher 1990). In addition, two natural isolates have been identified whose genomes appear to consist of chimeric DNAs probably formed when the obligatory template switch during DNA (–) strand synthesis was an interisolate switch (Dixon et al. 1986; Vaden and Melcher 1990). On the other hand, two observations suggest that the DNAs of multiple isolates do not often exist together in the same cell. Plants infected as little as 2 days previous are cross-protected against infection with a second isolate (Zhang and Melcher 1989). Also, in mixed infections, lesions that develop on noninoculated leaves contain one or the other of the coinoculated isolates, but not both (Riederer et al. 1992). Thus, though recombination is possible, the opportunity for its occurrence may not be frequent.

We report here the phylogenetic analysis of available CaMV nucleotide sequences. The results, presented with the above reservations in mind, suggest that neither rapid substitution rates nor extensive recombination has obscured CaMV phylogenetic relationships. Even so, recombination has frequently played a part in generating nucleotide sequences of present-day CaMV isolates.

Materials and Methods

Sequences and Alignment. Alignment of the complete nucleotide sequences of eight CaMV isolates (Table 1) was produced with the aid of gap translation (Melcher 1990) to optimally position gaps. Optimal positioning was scored using a matrix in which identities were assigned a value of 1.0, transitions 0.5, and transversions and gaps, 0. For phylogenetic analysis of CaMV isolates, we chose the nucleotide sequence of another caulimovirus as an outgroup. Carnation etched ring virus (CERV) was chosen since comparisons of four caulimoviral sequences (Hasegawa et al. 1989; Hull et al. 1986) indicated that CaMV was more closely related to CERV than to the other two caulimoviruses. To align the CERV sequence with that of CaMV CMV-1, we used the location of coding regions, alignment of the amino acid sequences, and local nucleotide sequence similarities identified by the ''align'' option of MacVector. The positioning of gaps was accomplished by gap translation as described above. The CERV–CMV-1 sequence alignment was then added to that of the other CaMV isolates. In some analyses, the outgroup was omitted. The nucleotide sequence of CM4-184 was not included in species tree phylogenetic analysis (except for the intergenic region) due to its similarity to that of CM1841 (Dixon et al. 1986).

Phylogenetic Analysis. Phylogenetic trees were constructed by three methods—bootstrapped parsimony, maximum likelihood, and distance—as implemented in the PHYLIP phylogenetic inference package (Felsenstein 1989). A bootstrap value for each internal branch in parsimony trees produced by DNAPARS was calculated by DNABOOT using 500 random resamplings. Maximum likelihood trees were constructed with DNAML using the default assumptions. For data sets of more than nine isolates, two calculations were performed, each containing the outgroup CERV and one non–North American isolate (XinJing) and one North American isolate (CM1841). One calculation included, in addition, the other North American isolates, and the second the other non–North American isolates. Resulting trees were joined using the common isolates. Minimum mutation distances between isolate pairs were calculated by DNADIST using the Kimura two-parameter option (Kimura 1980). Distance trees were constructed from the resulting distance matrices using FITCH. Each analysis was executed at least three times and, where possible, with randomization of data input order using the ''jumble'' option and global rearrangement of trees. The trees derived by the three methods were combined into one tree according to the strict consensus principle (Sokal and Rohlf 1980) and were displayed using DRAWTREE.

Sawyer Test for Recombination. The VTDIST program (Sawyer 1989) was used to search pairs of isolates for stretches of identical sequence that were significantly larger than expected based on random distributions of polymorphic sites. The length of such stretches, or fragments, is measured in total residues (uncondensed fragment) or number of polymorphic loci (condensed fragment). We considered a fragment significant if its P value (the fraction of permuted fragment lengths greater than or equal to the observed fragment) was ≤ 0.05. Options were involved to test for outer recombination (between a sequence in the sample and one from outside the sample) and inner recombination (between pairs of sequences within the sample).

Recombination Junctions. Isolate pairs were selected for analysis of recombination junctions either because the relative positions of the isolates in a gene tree were inconsistent with the species tree or because Sawyer analysis suggested that a recombination event had occurred between ancestors of the isolates. For selected isolate pairs, phylogenetically informative positions were scanned visually to identify positions near which a significant change in the density of differences between the isolates may have occurred. Segments extending leftward and rightward from these junctions were defined. In cases where junctions were in close proximity, the length of the segments was defined as the number of positions between the two junctions. Percentages of positions with identical residues (sharing percentages) were calculated. Junctions were judged significant if the sharing percentages of adjacent segments differed by more than one standard deviation. In cases where only one junction was found, length was chosen to assure differences of greater than two standard deviations between sharing percentages of adjacent segments.

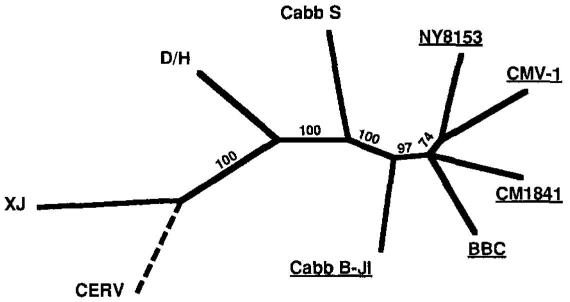

Fig. 1. Unrooted phylogenetic species tree of CaMV isolates. The tree is a consensus of trees obtained by bootstrapped parsimony, distance, and maximum-likelihood methods. Nodes were collapsed until a tree consistent with each of the methods was obtained. Values on branches represent the percentage of 500 bootstrapped parsimony resamplings of the data in which the isolates on one side of the branch grouped separately from those on the other side. Lengths of *solid lines* are proportional to distances determined by maximum likelihood. The length of the *dashed line* connecting the outgroup is reduced 50-fold relative to the other lines. North American isolates are *underlined.* Abbreviations used for isolates Cabbage S, Cabbage B-JI, and XinJing are, respectively, *Cabb S, Cabb B-JI,* and *XJ.*

Results

Species Tree

Phylogenetic trees were constructed for the aligned complete nucleotide sequences of CaMV isolates by distance, bootstrapped parsimony, and maximum-likelihood methods. The phylogenetic tree shown in Fig. 1 is a strict consensus of trees generated by the three methods. Branching orders were identical in the parsimony and maximum-likelihood trees. These trees differed from the distance tree only in that by parsimony and maximum likelihood, BBC diverged from the path leading to CMV-1 and NY8153 before CM1841 did, while the distance tree suggested BBC diverged from CM1841 after their common ancestor diverged from the CMV-1–NY8153 lineage (data not shown). Greater than 95% of the bootstrap resamplings supported the earlier placement of BBC divergence as did a significantly ($P < 0.01$) nonzero length for the branch separating the three termi-

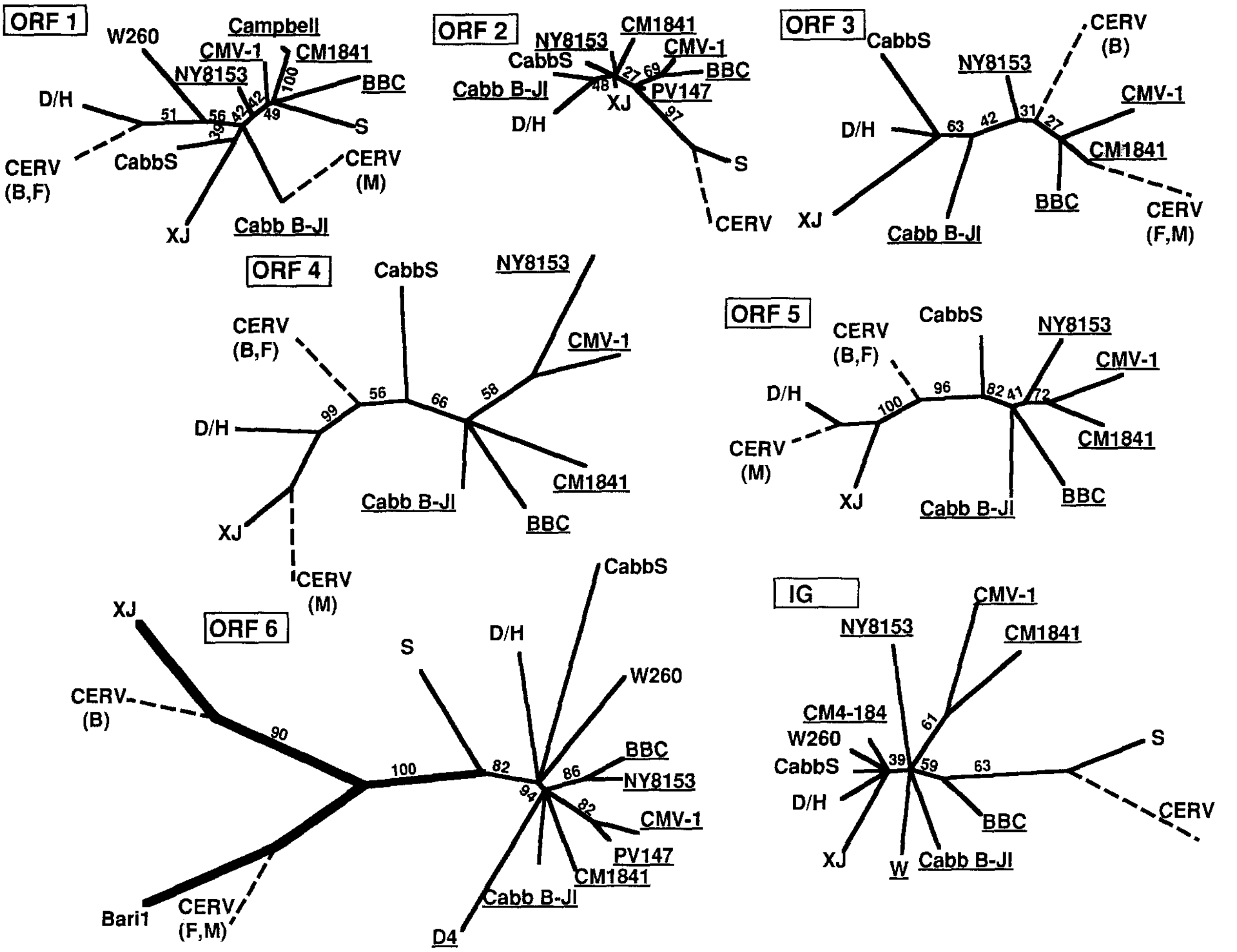

Fig. 2. Unrooted phylogenetic "gene" trees of CaMV isolates for ORFs 1–6 and the intergenic region (*IG*). Trees were drawn as described for Fig. 1 except that outgroup nodes from different methods were not collapsed. In cases of inconsistent outgroup placement, the source of the placement is indicated by letters; *B*, bootstrapped parsimony; *F*, distance; and *M*, maximum likelihood. Lengths of some branches (*thick lines*) in ORF 6 are reduced twofold relative to other lines.

nal isolates (NY8153, CMV-1, and CM1841) from BBC and the other isolates.

In the consensus tree, the lengths of terminal branches for all but the XinJing isolate were within a narrow range (standard deviations of 10 and 11% for maximum-likelihood and distance trees, respectively). The XinJing branches were more than 2.5 times the mean lengths of other terminal branches in both trees. The XinJing terminal branch, in all three trees, contained the node connecting the CaMV tree with the CERV outgroup. For both maximum-likelihood and distance trees, all five internal branch lengths were shorter than all terminal branch lengths. The distribution of isolates on the trees correlated with geographic location of their isolation. One cluster of isolates (Cabbage B-JI, BBC, NY8153, CMV-1, CM1841) consisted of isolates from North America and a second cluster (XinJing, D/H, Cabbage S) consisted of isolated from non–North American locations. For both maximum-likelihood and distance trees, all three internal branches in the North American cluster were shorter than the internal branch separating the clusters and the internal branch of the non–North American cluster.

Gene Trees

Separate phylogenetic trees, "gene trees" (Nei 1987), were constructed by all three methods for each of the major CaMV ORFs and the large intergenic region. Isolates used for the gene trees included those found in the species tree (Fig. 1) and partially sequenced isolates for which a complete gene nucleotide sequence was available. The ORF 1 trees required only a single node collapse (Sokal and Rohlf 1981) to produce a consensus tree (Fig. 2). A single node collapse was also sufficient to produce consensus gene trees for ORFs 3 and 5, while two collapses were required for ORF 4 (Fig. 2). Trees for ORFs 2 and 6 and the intergenic region required multiple node collapses to achieve consistency (Fig. 2). The node collapses were consistent with low frequencies of separation in bootstrapped parsimony, with short branch lengths in distance trees, and with marginally or nonsignificant branch lengths in maximum-likelihood trees. The frequencies of separation in bootstrapped parsimony were lower than those obtained by bootstrapped distance analysis. For ORF 3, the frequencies ranged from 27.4 to 63.2% for parsimony analysis but from 50 to 99% for distance analysis. Only for the parsimony and maximum likelihood trees of ORF 4 was the branching pattern consistent with that of the consensus species tree (Fig. 1). The ORF 5 and ORF 3 maximum-likelihood trees had identical branching patterns, except for the relative placements of NY8153 and BBC isolates. The ORF 5 parsimony branch order was the same as each of the three intergenic region trees for isolates included in both ORF 5 and intergenic region trees. Branching orders in trees for ORFs 1, 2, 4, and 6 did not closely resemble the branching order in any other gene tree. For all but two gene trees the position of the CERV outgroup in the maximum-likelihood trees differed from its position in the bootstrapped parsimony trees. This discrepancy is due to the high content of adenine in the base composition of the CaMV DNA (+) strands (Chenault and Melcher 1994) and due to the assumption in the maximum-likelihood method of a mutation preference to bases of the input composition. The CERV outgroup coincided with the parsimony position when maximum-likelihood analysis assumed that each base was equally likely to occur.

The size of the gene trees was estimated by calculating the maximum-likelihood distances between the most separated isolates. For ORF 6, Bari 1 and XinJing isolates were not considered for reasons discussed below. The ORF 2 tree was the smallest (0.0245 expected substitutions/site, for D/H vs S); trees for ORFs 1, 3, and 5 were next largest (0.059 for D/H vs S, 0.066 for XinJing vs CMV-1, and 0.061 for XinJing vs CMV-1 expected substitutions/site, respectively); those for ORFs 6 and the intergenic region were still larger (0.082 for S vs Cabbage S, and 0.088 for S vs CMV-1 expected substitutions/site, respectively). The ORF 4 tree was the largest (0.102 expected substitutions/site for XinJing vs NY8153). Branches separating internal nodes involving North American isolates of the species tree were short, none being over 0.01 expected substitutions/site. With three exceptions, terminal branch lengths were consistently between 0.002 and 0.03. Trees for ORFs 1–5 and the intergenic region had terminal branch lengths for XinJing (the isolate with the unusually long terminal branch in the species tree) similar to those for other isolates. XinJing and Bari 1 ORF 6 branch lengths were significantly longer than those of other ORF 6 terminal branches. The S Japan intergenic region branch was longer than other intergenic region terminal branches, but not significantly so.

In the trees for ORFs 1, 3–6, and the large intergenic region, the non–North American isolates of the species tree clustered apart from the North American isolates. Among the isolates not represented in the species tree, the North American isolates PV147, W, and D4 clustered with the North American isolates, while the Italian Bari I and the Argentinian W260 isolates clustered with the non–North American isolates. Four exceptions to the separate clustering of North American and non–North American isolates occurred. In the intergenic region tree, the North American isolate CM4-184 clustered with the European Cabbage S, consistent with previous reports (Dixon et al. 1986). Isolate S-Japan clustered with the North American isolates in the intergenic region and ORF 1 trees but with the non–North American isolates in the ORF 6 trees. In the ORF 2 tree, North American isolate Cabbage B-JI branched with the European isolate D/H. Also in this gene tree, the branching of NY8153 and CM1841 could not be separated from branching of the non–North American isolates Cabbage S and XinJing.

Sawyer Test for Recombination

The method of Sawyer (1989) was used to test for recombination between pairs of sequences within the CaMV alignment (inner-recombination) and between an aligned sequence and one not included in the alignment (outer-recombination). Analysis of all completely sequenced isolates identified inner fragments in common between CM1841 and CM4-184 covering all but ORF 2 and the intergenic region. Fragment ends corresponded to positions polymorphic between CM4-184 and CM1841. The identification is consistent with the common origin of these isolates and with the CM4-184 ORF 2 deletion (Howarth et al. 1981) and the substitution of a Cabbage S–like intergenic region in CM4-184 (Dixon et al. 1986). To identify other possible significant inner fragments, CM4-184 was omitted from subsequent analyses. With CM4-184 omitted, no uncondensed fragments were significantly longer than expected from a random distribution of polymorphic sites. Significant ($P \leq 0.05$) outer- and inner-condensed fragments are listed in Table 2. Five outer condensed fragments identified as significant were in ORF 6 of the XinJing isolate. They were 20–50 nucleotides long and contained seven to nine polymorphic loci unique to XinJing. One of the predicted fragments was within the 3′ hypervariable region of ORF 6. Examination of the ORF 6 nucleotide sequence alignment revealed that these short patches of polymorphisms unique to XinJing were part of a larger region, from nucleotide 6075 (Cabbage S coordinates) of ORF 6 to its end, characterized by 241 noninformative polymorphic positions, of which 51.0% were due to variations in XinJing. An additional 26.1% were due to residues of Bari 1. Because the complete nucleotide sequence of Bari 1 was not available, it was not included in the analysis by the Sawyer algorithm. Inner-condensed fragments varied in length from 115 to 249 nucleotides. Significant inner

Table 2. Significant (P value ≤ 0.05) condensed fragments detected by the Sawyer test for recombination among CaMV isolates[a]

Isolate(s)	Nucleotide position	Fragment length	No. polymorphic sites	P value
CMV-1/BBC	6554	246	63	0.0001
CMV-1/CM1841	6947	224	55	0.0007
D/H/Cabbage S	6678	210	46	0.0066
D/H/Cabbage S	7484	400	38	0.0152*
BBC/CM1841	6815	168	42	0.0220
NY8153/Cabbage S	7224	177	42	0.0220
NY8153/CM1841	7196	172	42	0.0220
BBC/CM1841	5894	249	40	0.0364, 0.0213*
NY8153/CMV-1	6966	153	39	0.0476
NY8153/CM1841	6966	153	39	0.0476
XinJing	6997	43	9	0.0019
XinJing	6638	28	9	0.0019
XinJing	7262	20	9	0.0019
XinJing	6686	50	7	0.0278
XinJing	6939	26	7	0.0278

[a] Fragments identified as significant upon omission of XinJing from the analysis are identified by asterisked P values. Listing of two isolates indicates putative recombination between those two isolates. Listing of one isolate indicates recombination between that isolate and a sequence not present in the data set. Starting position numbering is that of the Cabbage S isolate (Franck et al. 1980). Lengths of uncondensed fragments and the numbers of polymorphic sites in those fragments are shown.

fragments were found only in ORF 6. Since these condensed fragments overlap the large polymorphic XinJing fragment, their lengths were artefactually high. To assess the influence of the XinJing sequence, the analysis was repeated omitting XinJing. The only significant inner fragment detected with and without XinJing present was a 249-nucleotide fragment containing 40 polymorphic sites shared between BBC and CM1841. A 400-nucleotide fragment containing 38 polymorphic sites shared between Cabbage S and D/H in all but the last 150 nucleotides of the large intergenic region was significant when XinJing was omitted.

Recombination Junctions

Recombination during the genesis of isolates may result in branching inconsistent between a gene tree and the species tree or between two gene trees (Dykhuizen and Green 1991). Branching patterns of gene (Fig. 2) and species (Fig. 1) trees were compared to detect inconsistencies. Fourteen inconsistencies suggesting recombination and the two possible recombination events suggested by the Sawyer algorithm (Table 2) were further examined. For the relevant pairs of isolates, phylogenetically informative positions were scanned to identify putative recombination junctions as described in Materials and Methods. All 16 areas examined had at least one significant junction (Fig. 3). A second junction could not be located for the S vs D/H comparison because of lack of sequence information in adjacent regions. Suspected junctions in ORF 6 for the BBC vs CM1841 comparison and in ORF 3 for the BBC vs CMV-1 comparison were not significant. For the 13 other comparisons, two significant junctions were found. Two pairs of isolates (BBC and Cabbage B-JI, and CM1841 and CMV-1) each had three segments with high percentages of shared residues. Six junctions were near the start site of 35S transcription. These included those previously identified for CM4-184 and W isolates (Dixon et al. 1986; Vaden and Melcher 1990). Six junctions (two from the previosly reported CM4-184 and W recombination segments) were near the start site of DNA(−) strand synthesis. Four junctions were found in the region encompassing ORF 3 and the N-terminus of ORF 4, a region previously identified as a hot spot for recombination in recombinants generated experimentally (Vaden and Melcher 1990). In addition, a position near the 19S RNA transcription initiation site contained one recombination junction both in the present isolate comparisons (NY8153 vs BBC, Fig. 3) and in the analysis of experimentally selected recombinants. The upstream boundary of the intergenic region segment of BBC which was related to the Cabbage B-JI sequence corresponded to the sequence immediately downstream of the small bulge, sb2, in the proposed secondary structure of the CaMV 35S RNA (Fütterer et al. 1988). This bulge has previously been implicated in intermolecular recombination (Pennington and Melcher 1993) and occurs just beyond the 5′ end of the sequence complementary to the 3′ end of the 35S RNA. Because of the possible involvement of RNA secondary structure in stimulating recombination, regions surrounding putative recombination junctions were examined for possible secondary structure. Numerous potential elements of secondary structure were observed. Their significance is uncertain since they were not found at a consistent distance from the putative recombination junction and were also found in regions where no junctions were detected.

Discussion

The two factors considered in this work that could obscure phylogenetic relationships among CaMV isolates were a high nucleotide substitution rate and extensive recombination. A sufficiently high substitution rate would cause isolates to be related to one another as if they were members of a quasi-species. The topology of quasi-species sequences should be a star (Penny et al. 1991) with all sequences radiating from a central master sequence. Yet, members of a true quasi-species will produce a bifurcating tree when analyzed by methods based on a bifurcating evolutionary model. Still, trees obtained from different segments of a quasi-species genome should be unrelated to one another. Two distinct CaMV lineages, North American and non–North American, were found in the species tree (Fig. 1) and the majority of gene trees (Fig. 2), regardless of the method used or region analyzed. Thus, at least at the level of these lineages, excessive substitution rates did not obscure evolutionary relationships among CaMV isolates. Whether each of the two lineages represents a quasi-species or whether the isolates in each lineage are derived by divergence and radiation from common ancestors is not clear. The short internodes for the North American isolates suggest that those sequences are consistent with the star topology expected from a quasi-species relationship.

Two methods were used to search for possible recombination events between progenitors of the CaMV isolates studied. The Sawyer test (Sawyer 1989) detects stretches of sequence similar between two isolates. Two statistically significant inner fragments were found: one for isolates D/H and Cabbage S in the large intergenic region and the other for BBC and CM1841 at the 5′ end of ORF 6. Five outer-condensed fragments for XinJing were located in ORF 6. They were separated by small stretches of nucleotides containing polymorphic positions shared with other CaMV isolates. The interruption of fragments by shared polymorphic residues is a major limitation of the ability of the Sawyer test to detect recombination events. The Sawyer test did not predict any other significant recombination events between the CaMV isolates considered in this study.

The second method used to identify recombination events was the comparison of gene trees (Dykhuizen and Green 1991) with each other and with the species tree. Variation among gene-tree branching patterns has been used previously to identify recombination events between viral isolates (Li et al. 1988). For CaMV DNA, 14 instances of inconsistent branching were identified (Fig. 3). In most cases, the inconsistent branching patterns probably reflected recombination events in the genesis of the CaMV isolates. Thus, for detecting possible recombination events among CaMV isolates, gene-tree comparative analysis is more sensitive than the Sawyer test.

Confidence that inconsistent branching patterns reflect recombination events derives from several observations. In most of the regions of suspected events, abrupt changes in the percentage of shared residue types at phylogenetically informative positions were found (Fig. 3). The method identified the previously reported (Dixon et al. 1986) apparent recombination event between a CM1841-like isolate and a Cabbage S–like isolate. Many of the putative recombination junctions correspond to junctions observed in experimentally generated recombinants. Many of these have junctions at initiation sites for DNA (–) strand synthesis or for 35S or 19S RNA transcription (Wintermantel and Schoelz 1992; Gal et al. 1992; Grimsley et al. 1986; Vaden and Melcher 1990). These junctions likely result from the initiation of reverse transcription and template switches that occur when the enzyme reaches the 5′ end of its template. Nine of the junctions newly identified (Fig. 3) correspond, within the resolution of the data, to junctions expected from the reverse transcription mode of replication. Junctions not obviously related to reverse transcription template switches have been identified in experimentally generated recombinants. The correspondence between recombination junctions detected by sequence comparison of isolates and junctions identified in recombinants obtained under selective pressure suggests, consistent with previous reports (Dixon et al. 1986; Vaden and Melcher 1990; Melcher et al. 1986), that at least for CaMV, selective pressure is not necessary to detect recombination (Falk and Bruening 1994).

The two virus lineages detected in the majority of trees constructed may not represent all lineages. The detection of outer-recombination fragments between XinJing and another isolate (Table 2) suggests that at least one other lineage of CaMV, or a closely related virus, exists or existed at one time. The long branch length of Bari 1 in the ORF 6 tree indicates that it may represent still another lineage. Recombination may explain the three exceptions to non-overlapping North American and non–North American lineages. The previously reported region of similarity in the intergenic region of the North American CM4-184 isolate and non–North American isolates is clearly delimited by reverse-transcription junctions (Dixon et al. 1986). A recombination event between progenitors to Cabbage B-JI and D/H isolates may have occurred in the ORF 2-3 region, resulting in the branching of the North American Cabbage B-JI with the non–North American isolates. The placement of the S-Japan isolate with North American isolates for the intergenic region and ORFs 1 and 2 but with the non–North American ones for ORF 6 is less clear. Complete nucleotide sequencing of the S-Japan genome may suggest an explanation for the branching pattern of this isolate. The suggestion of recombination events between isolates of the two lineages is problematical. Recombination between two CaMV isolates would require the presence of both genomes in the same cell and thus in the

```
    ——— ORF 6————————————/——@— Intergenic Region———*/——— ORF 7——————/— ORF 1

     7154---------------89%-------------/---------30%----------7951
  S    CATGACAAGATGTGAGAAGACCCCAAAGACTCACAGGAATTAGCTCTCTAATCCTCCAA
  D/H  ............C.C.G.....A....AG...CTTCACC.-T..CT.CGGGA...TTG.

   7155--------------61%------------------/-------87%-------------------/---------48%---------------603
CM4-184  AAAGCAAAATGCACGGTTCCAA------CTCATTCAAAT-TGCTCTCGGGTCTTTCGA--TAGACTAAAACCGCGGACTGATTGC-ACGATTGGCTGT
Cabb S   GTGATG.T....GG...................AGG..-....CT.....A...............G.CA....ACAAAC.C...TT..C..A..AC

   7155--------------55%------------------/-------87%-------------------/---------45%---------------603
  CM1841 AAAGCAAAATGCGCGGTTCCAA------CTCATTCAAAT-ACTCTCGGGGAATCTCCT--CCGACTAAAACCGCGGACTGATTGC-ACGATTGGCTGT
  CMV-1  G..AGT.TGCT...A.............T...CAGG....T.........T....A....T......G.CAT.T..CAA..CC...TTAT..TA..A.

                 7192---------48%-----------------/---------93%------------/---------52%---------341
           Cabb S TGCGGGGTTCCAA------CTCATAGGAA--TGCCTTCGGGACTTTCGA--TAGACTAGACACGCGACAAACTCGC-TTG
           D/H    ....C..AGA.C.CAAAAG...C.TCACCA.............C..T........ACT.....A.C.A.TG...ATA..A

                 7192----------100%---------------/---------73%------------/--------100%---------341
         W        TTCACGGTTCCAA------CTCAGCAGGAAA-TCCCTGGGGAATTTCGAT-CCGAACTGACACACGGACTAATCGC-ACA
         CabbB-JI ..............................TT...........C......A..A-.........................

         -----------------97%--------------------/-------------44%----------------109
 NY8153  AAGCAATAGTCGCGGTTCCAA------CTCATAGGAAT-TCCCATGGGGAATCTAGA--TAGACCTAAGACCGTGG
 BBC     ..................T......................G.TT.CTAAT.CT....TACCA.........TGCAG

                                            7470----33%----/----100%---/----8%--109
                                      CabbB-JI TTTCCCTGGGGAACTTCGATACCA-ACTGACACACG
                                      BBC      .-.G.T.CTAAT...........ACTAAGACTG.A

                                                                                 44--------------47%---------->
                                                                         Cabb S  GCGACAAACTCGC-TTGACCTGACTACGC
                                                                         XJ      .TCG..TGAC.ATA..A.T..A..A.T..

——————— ORF 1 ——————————————/— ORF 2 ———/— ORF 3 ——/————— ORF 4 ——————

cont'd---/----------81%------------------/----------56%---------------2111
CabbS AACTCAGCCAGCTTCTACGTTTCACTAAGCGTCAAAGGTAATAATACGCCGAAGAAGATTAGTCATAC
XJ    G.T.........A....TA..C..T.G........GA...--..CA...TA...T.A...GA.A.CT.

     630------61%-----------------/----------86%----------1827
BBC    ACTCGAACAGCTTCTATGCCCCAGTGGTTACCATAGGTAATACTACATTAAAAAG
CMV-1  ...TAGC.........CA...TGC..A.......G..........TG.C......

                                 1152---71%--------/---14%-------/---71%---2227
                           NY8153    GCTCTGAGTAATGACATGCCGTAGAGTGCTGACTGCACTCAAC
                           CabbB-JI ...AA.T.....A..GCATT.AG.TAGAT..G.....TA..G.

                        852-------52%-----------/---------81%--------/--------29%-------2452
                 CabbB-JI CTCGCAGGTCGCTAAGTGTAATAACGCATTGAGGTAGATTGGCTGCATACAGCCCCTATCCA---C
                 D/H      TCT....AG..TC...AA..--.....GC......G....A..CA.TC..CATTA.AT..TAGAAA

                                   1158--------41%-----------/--------91%----------/---------50%----------2689
                              CM1841 CCCTGAACGATGACACGCTATAGAGTGCCGGCTGCATCTAGCCCCCATCCC---CTGAAGCCGCAGTTATTT
                              CMV-1  ..A..GGTA..ACT.T.TC.A.A........A..T..........TACT..GAA..---A.AAT.....C..

——————— ORF 4 ———————————————————————————————————//////////——— ORF 5 ——————————

      2776--------58%------------/--------88%--------------/-----------58%------------3824
CabbB-JI CTCTTCTGTTAGTCGCTTGCATTTTCGAACCCTCACAGT---CTCTCAACGTCACTCTAGTGAGTCC------GAGCCTGT
BBC      .C..CA...AGAA...C...G..C..A............AAT.....G..ACTGT...---AG..............T...C

                2905--------   --------------53%----------------------/----------------94%---->
            CM1841 GAACGCCTGCGCACCCAAACCTTCACAGT---CTCTCGACATCGTTCTAGCGAATCT------GAATCAGTAATAG
            CMV-1  .......CA.....T.G.GT.ACTGTCCCAAG....AA.T...AC...---.....C.......C........C..
```

Fig. 3. Putative recombination junctions detected by pairwise comparisons of phylogenetically informative positions of CaMV isolate nucleotide sequences. The first line of each pair shows the nucleotides occupying consecutive informative positions between the coordinates of the Cabbage S isolate indicated above the nucleotides. Positions informative due to insertion or deletion of multiple residues (collectively counted as single positions) are *underlined with a dotted line. Dashes* represent missing residues. Informative positions identical in the second isolate of each pair are represented by *dots. Underlining* identifies segments of informative positions that have significantly different percentages of common nucleotides (percentages given in sequence coordinate row) than adjacent segments. Segments whose percentages differ from those of their nonunderlined neighbors by one or two standard deviations are *singly underscored,* while those differing by two or more standard deviations are *doubly underscored. Solid lines* show the positions of ORFs, of the 35S RNA start site (@), the primer binding site for reverse transcription (*), and the ORF 4-5 overlap (//////).

same geographic location. That the recombination events occurred during laboratory propagation cannot be ruled out.

Sanger et al. (1991) attempted to infer evolutionary relationships among CaMV isolates based on comparisons of ORF 6 predicted amino acid sequences. Evolutionary relationships were suggested for the following groups of isolates: Bari 1/XinJing, CM1841/D/H, and D-4/CM1841/S-Japan. Our results for the ORF 6 nucleotide sequence support the relationships suggested by Sanger et al. (1991) for Bari 1/XinJing and for D-4/CM1841, but not for CM1841/D/H or for isolates D-4/CM1841/S-Japan.

The CERV DNA sequence (Hull et al. 1986), used as an outgroup, branches from the non–North American isolates in the species tree. This branching is due to two-

```
         ——— ORF 5 ——————————————————————————————————————————————————————————————————

cont'd   ---------------------------/-------------------------68%----------------------------------5114
CM1841   AATCGAAGACATCGAGAAAGTCCACTGGTGGGATAGAAAATCCACTAGAAGATAGTTTAACAATCTTCCAATAACCCGAGTGCCCCCCTTACTC
CMV-1    ...........................AGAT...............T.....C....C...G...C.G....G........T..TTT......T

     3899-----------67%-----------------/------------------90%----------------------------/------------43%----------5537
CabbB-JI AATCGAAGATACTGAAGAAGTCTACTAAAATGATAGAAAATCCACTAGAAGACAGTTTGGCAAGTCCGAAGAACCCGTATTCCTCCTCCGAGCCGTGAGCCATCCCGTCTT
BBC      .........C..CA.G.....TC...G.TGG................A.....................G.....T.A...T.C...TTACC.TA.AGA..G..TT..T.A

                             4097----------------------------------59%-----------------------------------/------96%---->
                         CM1841  CTGGTGGGATAGAAAATCCACTAGAAGATAGTTTAACAATCTTCCAATAACCGAGTGCCCCCCTTACTCATGGACCGTCTTGTCTAA
                         BBC     ...A...................A....C.....GG...GTCCGAGGA.C.T..A.TT....TT...C...A...........T...

                                                                    4613--------------------------58%------------------>
                                                                NY8153  TCTTGAAGTGACCGAGTTCTCCTTTTGCTCATAAGTCGTCTCGTTCTA
                                                                BBC     GTCC..G.AAC.T..A..TC..C...A.C....GAC.....T...TA.


   ORF 5 /———————————— ORF 6 ——————————————————

cont'd   -----------------------------------6142
CM1841   GTACCCTACAGCGGCCATATGGTATGCGTGGTTTGTGCA
BBC      .....T.................................

cont'd   -----------/-----------------93%------------------------------6354
NY8153   ATACCCAACAACGGCCATATGGTATGCGTGGTTTGTGCACGCCTCATTTACTCTAGACTACCGGCC
BBC      G....TT...G.............................GT..AC....................
```

Fig. 3. Continued.

thirds of the CaMV genome contained in ORFs 4–6. The position of CERV in the North American cluster for the intergenic region and ORFs 2 and 3 is inconsistent with divergence of CaMV and CERV from an ancestral caulimoviral lineage followed by CaMV isolate radiation, first of non–North American isolates and then of North American isolates from one of the non–North American isolates. One possible explanation of this apparent contradiction is that CaMV is ancestral to CERV, CERV diverging from one ancient CaMV lineage. After this divergence, there would have been a major recombination event between that lineage and another lineage more distant from CERV. Modern representatives of the complementary recombination products are found in the North American and non–North American isolate clusters. Since the gene-tree comparisons provide reasonable evidence of recombination during the evolution of CaMV isolates, such a recombination event is not unlikely. Other, less favored, explanations require convergent or parallel evolution.

Plant virus evolution may be influenced by various different factors, including both virus–vector (Howarth and Vandemark 1989; Matthews 1991) and virus–host interactions (Dawson 1992; Howarth and Vandemark 1989; Matthews 1991). No CaMV isolates clustered according to whether they are aphid transmissible or nontransmissible. The majority of CaMV isolates used in this study were isolated from *Brassica* species; no branching pattern specific to host source was found for CaMV isolates differing in host genus. Rather, the results suggest that the major factor contributing to divergent lineages of CaMV is CaMV-host geographic distribution. An evolutionary grouping by host geographic distribution has been suggested for other plant viruses (Blok et al. 1987; Howarth and Vandemark 1989; Matthews 1991). Based upon hybridization tests, Blok et al. (1987) suggested that turnip yellow mosaic virus (TYMV) isolates separate into two distinct lineages, one of Australian origin and the other of European origin. Howarth and Vandemark (1989) noted that geminivirus isolates also clustered in phylogenetic trees according to their geographic origin. The effect of host geographic distribution on viral divergence has also been well documented for animal viruses (Donis et al. 1989; Li et al. 1988).

Acknowledgments. We thank Dr. Stanley Sawyer for several helpful discussions about the VTDIST algorithm and for performing the VTDIST calculations reported. The work reported was supported by a Robert Glenn Rapp Foundation Fellowship to K.D.C., the Oklahoma Health Research Program (contract No. HR8-048), and the Oklahoma Agricultural Experiment Station (project 1789), whose director has approved the work for publication. Use of the Oklahoma University Computer Group resource was supported by the University of Oklahoma Computer Center and the Oklahoma Center for Advancement of Science and Technology through a Center of Excellence Contract.

References

Balàzs E, Guilley H, Jonard G, Richards K (1982) Nucleotide sequence of DNA from an altered-virulence isolate D/H of the cauliflower mosaic virus. Gene 19:239–249

Blok J, Mackenzie A, Guy P, Gibbs A (1987) Nucleotide sequence comparisons of turnip yellow mosaic virus isolates from Australia and Europe. Arch Virol 97:283–295

Chenault KD, Melcher U (1993) The complete nucleotide sequence of cauliflower mosaic virus isolate BBC. Gene 123:255–257

Chenault KD, Melcher U (1994) Patterns of nucleotide sequence variation among cauliflower mosaic virus isolates. Biochimie 76:3–8

Choe IS, Melcher U, Richards K, Lebeurier G, Essenberg RC (1985) Recombination between mutant cauliflower mosaic virus DNAs. Plant Mol Biol 5:281–289

Daubert SD, Schoelz J, Debao L, Shepherd RJ (1984) Expression of disease symptoms in cauliflower mosaic virus genomic hybrids. J Mol Appl Gen 2:537–547

Dawson WO (1992) Tobamovirus-plant interactions. Virology 186: 359–367

Dixon L, Nyffenegger T, Delley G, Martinez-Izquierdo J, Hohn T (1986) Evidence for replicative recombination in cauliflower mosaic virus. Virology 150:463–468

Donis R, Bean WJ, Kawaoka Y, Webster RG (1989) Distinct lineages of influenza virus H4 hemagglutinin genes in different regions of the world. Virology 169:408–417

Dykhuizen DE, Green L (1991) Recombination in *Escherichia coli* and the definition of biological species. J Bacteriol 173:7257–7268

Falk BW, Bruening G (1994) Will transgenic crops generate new viruses and new diseases? Science 263:1395–1396

Felsenstein J (1989) PHYLIP—phylogeny inference package. Cladistics 5:164–166

Franck A, Guilley H, Jonard G, Richards K, Hirth L (1980) Nucleotide sequence of cauliflower mosaic virus DNA. Cell 21:285–294

Fütterer J, Gordon K, Bonneville JM, Sanfaçon H, Pisan B, Penswick J, Hohn T (1988) The leading sequence of caulimovirus large RNA can be folded into a large stem-loop structure. Nucleic Acids Res 16:8377–8390

Gal S, Pisan B, Hohn T, Grimsley N, Hohn B (1992) Agroinfection of transgenic plants leads to viable cauliflower mosaic virus by intermolecular recombination. Virology 187:525–533

Gojobori T, Moriyama EN, Kimura M (1990) Molecular clock of viral evolution, and the neutral theory. Proc Natl Acad Sci U S A 87: 10015–10018

Gracia O, Shepherd RJ (1985) Cauliflower mosaic virus in the nucleus of *Nicotiana.* Virology 146:141–145

Grimsley N, Hohn T, Hohn B (1986) Recombination in a plant virus: template-switching in cauliflower mosaic virus. EMBO J 5:641–646

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 17:9993–10013

Holland JJ, De La Torre JC, Steinhauer DA (1992) RNA virus populations as quasispecies. Curr Top Microbiol Immunol 176:1–20

Howarth AJ, Gardner RC, Messing J, Shepherd RJ (1981) Nucleotide sequence of naturally occurring deletion mutants of cauliflower mosaic virus. Virology 112:678–685

Howarth AJ, Vandemark GJ (1989) Phylogeny of geminiviruses. J Gen Virol 70:2717–2727

Howell SH, Walker LL, Walden RM (1981) Rescue of *in vitro* generated mutants of cloned cauliflower mosaic virus genome in infected plants. Nature 293:483–486

Hull R (1980) Structure of the cauliflower mosaic virus genome III. Restriction endonuclease mapping of thirty-three isolates. Virology 100:76–90

Hull R, Sadler J, Longstaff M (1986) The sequence of carnation etched ring virus DNA: comparison with cauliflower mosaic virus and retroviruses. EMBO J 5:3083–3090

Kimura M (1980) A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. J Mol Evol 16:111–120

Li W-H, Tanimura M, Sharp PM (1988) Rates and dates of divergence between AIDS virus nucleotide sequences. Mol Biol Evol 5:313–330

Lung MCY, Pirone TP (1972) *Datura stramonium,* a local lesion host for certain isolates of cauliflower mosaic virus. Phytopathology 62:1473–1474

Mason WS, Taylor JM, Hull R (1987) Retroid virus genome replication. Adv Virus Res 32:35–96

Matthews R (1991) Plant virology, 3rd ed. Academic Press, New York

Melcher U, Choe IS, Lebeurier G, Richards K, Essenberg RC (1986) Selective allele loss and interference between cauliflower mosaic virus DNAs. Mol Gen Genet 203:230–236

Melcher U (1989) Symptoms of cauliflower mosaic virus infection in *Arabidopsis thaliana* and turnip. Bot Gazette 150:139–147

Melcher U (1990) Similarities between putative transport proteins of plant viruses. J Gen Virol 71:1009–1018

Melcher U, Lartey RT, Pennington RE (1992) Isolate-specific synergy in symptom production between a caulimo- and a tobamovirus. Phytopathology 82:1173–1174

Nei M (1987) Molecular evolutionary genetics. Columbia University Press, New York

Pennington RE, Melcher U (1993) *In planta* deletion of DNA inserts from the large intergenic region of cauliflower mosaic virus DNA. Virology 192:188–196

Penny D, Hendy MD, Steel MA (1991) Testing the theory of descent. In: Miyamoto MM, Cracraft J (eds) Phylogenetic analysis of DNA sequences. Oxford University Press, New York, pp 155–183

Riederer MA, Grimsley NH, Hohn B, Jiricny J (1992) The mode of cauliflower mosaic virus propagation in the plant allows rapid amplification of viable mutant strains. J Gen Virol 73:1449–1456

Rongxiang F, Xiaojun W, Ming B, Yingchuan T, Faxing C, Kequiang M (1985) Complete nucleotide sequence of cauliflower mosaic virus (Xinjing isolate) genomic DNA. Chin J Virol 1:247–256

Sanger M, Daubert S, Goodman RM (1991) The regions of sequence variation in caulimovirus gene VI. Virology 182:830–834

Sawyer S (1989) Statistical tests for detecting gene conversion. Mol Biol Evol 6:526–538

Schoelz J, Shepherd RJ, Daubert S (1986) Region VI of cauliflower mosaic virus encodes a host range determinant. Mol Cell Biol 6: 2632–2637

Schoelz JE, Shepherd RJ (1988) Host range control of cauliflower mosaic virus. Virology 162:30–37

Shepherd RJ (1989) Biochemistry of DNA plant viruses. In: Marcus A (ed) The biochemistry of plants. Academic Press, New York, pp 563–616

Shepherd RJ, Bruening GE, Wakeman RJ (1970) Double-stranded DNA from cauliflower mosaic virus. Virology 41:339–347

Sokal RR, Rohlf FJ (1981) Taxonomic congruence in the Leptodomorpha reexamined. Syst Zool 30:309–325

Steinhauer DA, Holland JJ (1987) Rapid evolution of RNA viruses. Annu Rev Microbiol 41:409–433

Stenger DC, Mullin RH, Morris TJ (1988) Isolation, molecular cloning, and detection of strawberry vein banding virus DNA. Phytopathology 78:154–159

Stratford R, Covey SN (1989) Segregation of cauliflower mosaic virus symptom genetic determinants. Virology 172:451–459

Vaden VR, Melcher U (1990) Recombination sites in cauliflower mosaic virus DNAs: implications for mechanisms of recombination. Virology 177:717–726

Walden RM, Howell SH (1983) Uncut recombinant plasmids bearing nested cauliflower mosaic virus genomes infect plants by intragenomic recombination. Plant Mol Biol 2:27–31

Wintermantel WM, Schoelz JE (1992) Cauliflower mosaic virus is capable of recombination with transgenic *Nicotiana bigelovii* that contain CaMV coding sequences. Phytopathology 82:1110

Woolston CJ, Covey SN, Penswick JR, Davies JW (1983) Aphid transmission and a polypeptide are specified by a defined region of the cauliflower mosaic virus genome. Gene 23:15–23

Zhang XS, Melcher U (1989) Competition between isolates and variants of cauliflower mosaic virus in infected turnip plants. J Gen Virol 70:3427–3437