J. theor. Biol. (2002) **217,** 195–201
doi:10.1006/yjtbi.3023, available online at http://www.idealibrary.com on IDEAL®

Third Position Codon Composition Suggests Two Classes of Genes Within the *Cauliflower Mosaic Virus* Genome

S. M. Leisner*† and D. A. Neher‡

†*Department of Biological Sciences, College of Arts and Sciences, The University of Toledo, Toledo, OH* 43606, *U.S.A. and* ‡*Department of Earth, Ecological and Environmental Sciences, College of Arts and Sciences, The University of Toledo, Toledo, OH* 43606, *U.S.A.*

(*Received on* 15 *May* 2001, *Accepted in revised form on* 22 *February* 2002)

The translation of viral mRNAs by host ribosomes is essential for infection. Hence, codon usage of virus genes may influence efficiency of infection. In addition, composition of nucleotides in the third position within codons of genes can reflect evolutionary relationships. In this study, third position codon composition was examined for the seven genes of eight *Cauliflower mosaic virus* isolates. Genes IV–VII had similar codon composition values and were termed Class 1 genes. Genes I–III possessed corresponding codon composition values and were termed Class 2 genes. The codon composition values of Class 1 and genes differed significantly. Neither Class 1 nor Class 2 genes had codon composition values identical to that of the host plant, *Arabidopsis thaliana*. However, Class 1 genes possessed codon composition values closer to those of the host than Class 2 genes. Examination of the genomes of three *Rous sarcoma virus* isolates indicated that codon composition values were similar for the *gag*, *pol*, and *env* genes but these genes differed significantly from the *src* genes. Since codon composition values for *Rous sarcoma* virus distinguished a "foreign" gene from the rest of the viral genome, it is possible that the *Cauliflower mosaic virus* genome is composed of genes from two different sources. Others have suggested that *Cauliflower mosaic virus* evolved in this manner and our data provide support for this hypothesis.
© 2002 Elsevier Science Ltd. All rights reserved.

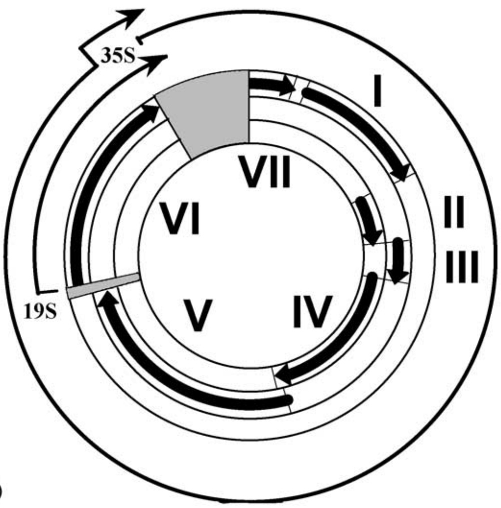
1. Introduction

Viruses are efficient pathogens of most cellular organisms because of their tight integration with host physiology (Knipe, 1990; Matthews, 1991). For viral gene expression to occur, translation of virus mRNAs by host ribosomes is essential. Some viruses encode their own tRNAs to facilitate viral protein synthesis (Strauss *et al.*, 1990). Other viruses, such as *Cauliflower mosaic virus* (CaMV), depend entirely on the translation machinery of the host, suggesting that the codon usage of their genes correlates with their translation (Strauss *et al.*, 1990).

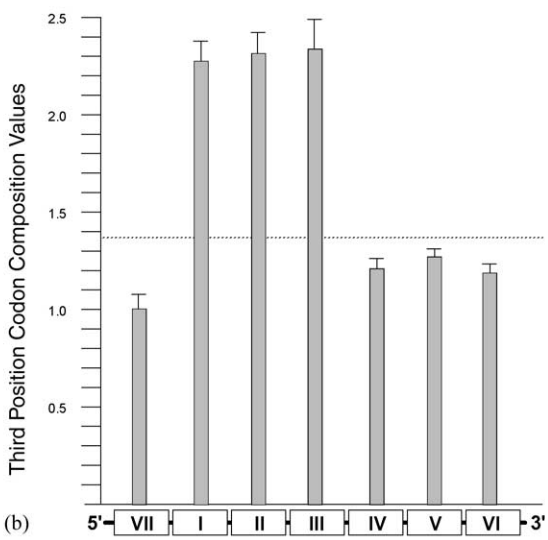
CaMV particles harbor an 8 kbp double-stranded circular DNA genome, shown in Fig. 1(A) (Hull & Covey, 1985; Mason *et al.*, 1987; Matthews, 1991; Rothnie *et al.*, 1994). Following invasion of a host cell, viral DNA is targeted to the nucleus where it serves as a

*Corresponding author. Tel.: +1-419-530-1549; fax: +1-419-530-7737.
E-mail address: sleisne@uoft02.utoledo.edu (S.M. Leisner).

0022-5193/02/$35.00/0
© 2002 Elsevier Science Ltd. All rights reserved.



(a)



(b)

FIG. 1. The average third position codon composition values of Cauliflower mosaic virus genes. (A) The organization and structure of the CaMV genome is shown. The heavy arrows indicate the seven CaMV genes (labeled with roman numerals) in the 8 kb pair genome, while the thin arrows show the two viral transcripts, and the shaded portions indicate the intergenic regions. Genes VII, I, and VI are in a different reading frame from II and IV, which differs, from that for III and V. It is important to note that genes III–V overlap slightly. Gene names located within the inner circle are Class 1 genes while the others are Class 2 genes. (B) The average CC values for CaMV genes VII–VI are indicated. Also shown are the standard deviations. The dashed line indicates the average *A. thaliana* CC value. Below the graph is a drawing of the CaMV 35S RNA [in a linearized form from (A)] showing the approximate positions of the seven genes. Note that genes are not drawn to scale.

template for plant RNA polymerase II. Unlike retroviruses, the integration of the pararetroviral DNA into a host genome is not required for viral replication. As a member of the pararetrovirus family, a CaMV DNA genome replicates via an RNA intermediate, the 35S RNA. In addition to its role as a template for reverse transcription, CaMV 35S RNA serves as a polycistronic mRNA for viral protein synthesis. Reverse transcription of CaMV 35S RNA is believed to occur in cytoplasmic inclusion bodies and viral DNA generated is packaged concomitantly into viral particles.

The CaMV genome encodes seven proteins (Hull & Covey, 1985; Matthews, 1991; Rothnie *et al*., 1994). The arrangement of the genes on the 35S RNA is shown in Fig. 1(B) and the functions of their gene products are indicated in Table 1. Genes I–III encode proteins involved mainly in plant-related functions (Bonneville & Hohn, 1993; Bonneville *et al*., 1987; Mason *et al*., 1987). Genes IV–VI products primarily effect the processes of replication and virion assembly (Bonneville & Hohn, 1993; Chenault & Melcher, 1994a, b; Hull & Covey, 1985; Mason *et al*., 1987; Matthews, 1991; Rothnie *et al*., 1994). In many respects, genes IV, V and VI of CaMV resemble the *gag*, *pol* and *env* genes of retro-elements, respectively.

Synthesis of CaMV proteins requires that the virus exploit the translation machinery of the host. Therefore, we expect CaMV codon usage to resemble that of the host. To test this hypothesis, we examined the composition of the nucleotide in the third position of all codons for each viral gene. Third position codon composition (CC) has been used to examine codon preference of genes within an organism, and indicate evolutionary relationships among species (Campbell & Gowri, 1990; Lawrence & Roth, 1996). This study was undertaken to answer two basic questions. First, how do the CC values for CaMV genes compare with that of a plant host? Second, how similar are the CC values of the various CaMV genes when compared with one another? Based on CC values, we report that CaMV genes fall into two classes. Neither of these classes had CC values equivalent to that of the host plant *Arabidopsis thaliana*.

TABLE 1

Functions of the CaMV genes examined in this study

CaMV gene	Function of gene product	Reference
VII	Protein unstable in plants, function unknown	Wurch *et al.* (1990)
I	Facilitates transport of viral nucleic acid from one plant cell to another	Thomas & Maule (1995)
II	Involved in transmission of the virus by plant-feeding insects (aphids)	Blanc *et al.* (1993)
III	Acts as a linker connecting the gene II product to the viral capsid, non-specific double-stranded DNA-binding protein	Leh *et al.* (2001), Mesnard *et al.* (1990)
IV	Viral capsid protein; analogous to retroviral *gag* protein	Bonneville & Hohn (1993), Gardner *et al.* (1981), Hull & Shepherd (1976)
V	Reverse transcriptase replication enzyme, containing proteinase and RNase H domains; analogous to retroviral *pol* protein	Bonneville & Hohn (1993)
VI	Major inclusion body protein, binds to capsid protein and thought to be involved in the virion assembly process, determines host range and symptom severity, translational transactivator; somewhat analogous to retroviral *env* protein	Bonneville *et al.* (1989), Chenault & Melcher (1994*a*), Covey & Hull (1981), Himmelbach *et al.* (1996), Schoelz *et al.* (1986), Stratford & Covey (1989)

TABLE 2

Codon composition values for the eight CaMV isolates used in this study

CaMV isolate	Third position codon composition values							Accession number	References
	VII	I	II	III	IV	V	VI		
B29	0.98	2.19	2.19	2.25	1.17	1.19	1.25	X79465	Pique *et al.* (1995)
BBC	1.06	2.28	2.56	2.61	1.18	1.27	1.19	M90542	Chenault & Melcher (1993b)
Cabb S	1.02	2.28	2.27	2.25	1.28	1.29	1.21	J02048	Franck *et al.* (1980)
CM1841	1.06	2.11	2.33	2.51	1.16	1.31	1.12	V00140	Gardner *et al.* (1981)
CMV-1	0.94	2.34	2.27	2.33	1.16	1.27	1.16	M90543	Chenault & Melcher (993a)
D/H	1.16	2.31	2.33	2.17	1.21	1.31	1.14	J02047	Balazs *et al.* (1982)
NY8153	1.06	2.45	2.27	2.25	1.21	1.27	1.17	M90541	Chenault *et al.* (1992)
Xinjiang	0.94	2.28	2.33	2.33	1.28	1.27	1.24	AF140604	Fang *et al.* (1985)

Note: The third position codon composition value is based on the last nucleotide in the codon being either an A/T or a G/C (XXA/T or XXG/C) and dividing the XXA/T value by the XXG/C value.

Materials and Methods

THIRD POSITION CODON COMPOSITION ANALYSIS OF SEQUENCES

Codon usage for each of the seven genes of eight CaMV isolates (Table 2) was determined with Macintosh DNA Strider 1.2 software. The third position of each codon for every CaMV gene was ranked as an A/T or a G/C nucleotide, and termination codons were included in this total. Numbers of codons ending in A or T were summed separately from those ending in G or C. For each gene, the total number of codons ending in A or T was then divided by those with G or C in the third position to generate a value representing the third position codon composition (CC) of the gene. For comparison, three *Rous sarcoma virus* (RSV) isolates were also examined in this study (Table 3). The CC values of the RSV *gag*, *pol*, *env*, and *src* genes for each isolate were determined as described above for CaMV. The CC value for the bulk genome of the plant host, *A. thaliana,* was determined in the same manner from the data obtained from the Kazusa website (http://www.kazusa.or.jp/codon/).

TABLE 3
Codon composition values for the three RSV isolates used in this study

RSV isolate	Third position codon composition values				Accession number	References
	gag	*pol*	*env*	*src*		
Schmidt–Ruppin B	0.776	0.912	0.992	0.235	AF052428	J. Bouck *et al*., unpublished
Schmidt–Ruppin D	0.767	0.896	1.00	0.242	D10652	Kihara, unpublished
Prague	0.786	0.862	1.07	0.242	J02342	Katz *et al.* (1982), Schwartz *et al.* (1983)

Note: the third position codon composition value is based on the last nucleotide in the codon being either an A/T or a G/C (XXA/T or XXG/C) and dividing the XXA/T value by the XXG/C value.

STATISTICAL ANALYSIS

Two-way analysis of variance was performed with CC values as a dependent variable and gene and isolate as independent variables, using SAS software (SAS/STAT Users Guide, 1989). A single-degree-of-freedom contrast comparing CC values among genes was also performed.

Results

When the CC of the seven genes from eight different CaMV isolates was determined, genes IV–VII (which we termed Class 1 genes) were found to have values near 1.2–1.3 (Table 2). Interestingly, CC values of these genes were consistent among virus isolates ($p = 0.2961$). The CC values of genes I–III (Class 2 genes) were similar with values about 2.3–2.4 for all CaMV isolates ($p = 0.2961$).

The viral CC values were compared with those of a host plant with large amounts of available codon data, *A. thaliana*. The CC value for the total *A. thaliana* genome in the Kazusa website database was 1.38. Neither Class 1 nor Class 2 genes had CC values that exactly matched that of the CaMV host plant. However, CC values of Class 1 genes were more similar to *A. thaliana* than Class 2 genes (Fig. 1(B)).

The CC values for Class 1 genes differed significantly from those of the Class 2 genes ($p < 0.0001$). Hence, the CaMV genome contains two classes of genes each with a different codon composition. Interestingly, Class 2 genes are contiguous within the genome and are located between Class 1 genes VII and IV (Fig. 1(B)). Perhaps the two classes of CaMV genes originated from different sources. Other virus genomes have obtained genes in this way. For example, the *src* gene of RSV (Coffin, 1990) was likely obtained from a vertebrate host (Rohrschneider *et al*., 1979; Takeya & Hanafusa, 1982). Analysis of three RSV isolates showed that the CC values for the *gag*, *pol* and *env* genes were similar, about 0.9 (Table 3). However, CC values for the *src* genes (close to 0.24), differed significantly from *gag/pol/env* genes ($p < 0.0001$) which was consistent among isolates ($p = 0.7723$).

Discussion

We determined the third position codon composition (CC) values of the seven genes among eight CaMV isolates and compared those values to that of a common host plant, *A. thaliana*. We examined, CC rather than codon usage because it permitted each codon for every gene to be quantified. In addition, the identity of the nucleotide in the third position (XXG/C vs. XXA/T) has been used to examine codon preferences for higher plants, green algae, cyanobacteria and certain bacterial operons (reviewed in Campbell & Gowri, 1990; Lawrence & Roth, 1996). These studies recommended CC as a tool to examine evolutionary relationships. Our results suggest that CaMV genes fall into two classes: Class 1 (genes IV–VII) and Class 2 (genes I–III).

The dramatic difference in codon composition may have at least six explanations. First, it is possible that the different CC values reflect the amount of protein required by the virus. We believe this to be unlikely because the viral coat protein (gene IV product) is more abundant than reverse transcriptase (gene V product) (Kobayashi *et al*., 1998) but both genes possess similar

CC values (see Table 2). Secondly, the amino acid composition of gene products may bias CC values for the different classes of CaMV genes. For example, some of the viral proteins may contain many charged, hydrophobic or hydrophilic amino acids. If so, all four Class 1 genes would show a similar amino acid composition, and one that differed from all three Class 2 genes. None of the CaMV genes appear to encode proteins particularly rich in specific amino acids, making this explanation improbable. Interestingly, Class 1 genes IV–VI all encode RNA-binding proteins (Bonneville & Hohn, 1993; De Tapia *et al*., 1993). Therefore, a third possible explanation is that CC values reflect the ability of a gene product to associate with ribonucleic acid. This is unlikely because Class 2 genes I and III encode RNA-binding proteins (Jacquot *et al*., 1998; Thomas & Maule, 1995). Fourth, CC differences may be essential for folding or packaging the viral DNA into virions. This is improbable because gene II can be replaced with genes having contrasting CC values, yet the virus is still viable (Brisson *et al*., 1984; Lefebvre *et al*., 1987). Fifth, the two classes of genes may be under different selection pressures that maintain divergent CC values. Patterns of nucleotide sequence change for Class 1 genes IV–VI appear different from those for Class 2 genes I–III (Chenault & Melcher, 1994a). Hence, the patterns of nucleotide sequence change correlate, to a large extent, with CC values. Interestingly, 69–79% of the mutations in genes I–IV appear to be silent mutations, compared to 90% for gene V and 54% for gene VI. However, gene IV contains more insertion/deletion mutations and variability than genes I–III; and gene V has the lowest density of coding changes, both suggesting that Class 1 and 2 genes are under different selection pressures. Perhaps, this difference in selection pressure is manifested as contrasting CC values for these genes. It is intriguing that the CC value for gene IV differs from genes I–III even though all four genes possess similar silent mutation percentages. Remarkably, genes V and VI have dissimilar silent mutation percentages and, yet, possess similar CC values. Taken together, these data suggest that mutation selection does not completely explain the two CC classes.

A final explanation for the two classes of genes is that they represent the evolutionary history of CaMV. Evidence is two-fold. First, Class 1 genes IV–VI, respectively, resemble the *gag*, *pol* and *env*, genes found in a retroelement (Bonneville & Hohn, 1993; Chenault & Melcher, 1994a,b; Hull & Covey, 1985; Mason *et al*., 1987, Matthews, 1991; Rothnie *et al*., 1994). Second, Class 2 genes form a single continuous block that is located between Class 1 genes VII and IV (see Fig. 1(B)). Together, these suggest a possible scenario for the origin of CaMV. The proto-CaMV may have been a retroelement to which the Class 2 genes were added, possibly via a recombination event. The addition of Class 2 genes may have permitted the new retroelement to adapt efficiently to its hosts by allowing it to spread from cell to cell and plant to plant (Bonneville *et al*., 1987; Mason *et al*., 1987; Rothnie *et al*., 1994). Other workers suggest such an origin for CaMV and that our Class 2 genes may have been obtained from an RNA virus (Bonneville & Hohn, 1993; Bonneville *et al*., 1987; Mason *et al*., 1987; Rothnie *et al*., 1994). We term this the "Dual Origin Hypothesis of CaMV Evolution." Phylogenetic analysis of retroelements based on the sequences of their reverse transcriptases indicates that CaMV is most closely related to members of the Spumavirus genus of the retrovirus family (Li *et al*., 1995). Perhaps proto-CaMV was related closely to Spumaviruses.

In support of the dual origin hypothesis, some CaMV isolates have been generated via recombination (Chenault & Melcher, 1994b). In addition, the region between CaMV genes III and IV, which is a boundary between Class 1 and 2 genes, contains a recombination hotspot (Vaden & Melcher, 1990). Finally, genes with similar CC values appear to have related functions. Class 2 genes appear to be involved primarily in plant-associated functions (Bonneville *et al*., 1987; Hull & Covey, 1985; Mason *et al*., 1987; Matthews, 1991; Rothnie *et al*., 1994), while Class 1 genes play a role in virus genome replication and virion assembly (Bonneville & Hohn, 1993; Chenault & Melcher, 1994a; Hull & Covey, 1985; Mason *et al*., 1987; Matthews, 1991; Rothnie *et al*., 1994). Since Class 2 genes appear to be plant-specific, this suggests that they were added later in the evolution of CaMV than Class 1 genes. This may also explain why Class 1 genes have CC values that more closely resemble that of the *A. thaliana* host plant than Class 2 genes.

To provide further support the hypothesis that the two classes of CaMV genes are from different origins, a viral genome, RSV, known to have obtained a gene via such a mechanism was examined (Coffin, 1990). In addition to the standard retroviral genes, the oncoretrovirus RSV has also obtained the *src* gene, presumably from a host cell (Rohrschneider *et al*., 1979; Takeya & Hanafusa, 1982). The RSV CC values are quite different from those of CaMV, probably reflecting selection pressure imposed by a different host. However, the CC values of the RSV genes show an obvious pattern. The CC values of the three different RSVs for the *gag*, *pol*, and *env* genes are similar, whereas that of the *src* gene is dissimilar. These data suggest that large differences in CC can reflect different sources for viral genes.

The authors thank Drs Stephen Goldman, Lirim Shemshedini (both at The University of Toledo, Toledo, OH), and The University of Toledo Plant Science Research Center, for their assistance. The authors also thank Dr Ulrich Melcher (Oklahoma State University, Stillwater, OK) for his helpful suggestions on the manuscript. This work was supported in part by USDA grant number 96-3503-3284 and the University of Toledo deARCE Memorial Endowment Fund in Support of Medical Research and Development.

REFERENCES

Balazs, 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.

Blanc, S., Cerutti, M., Chaabihi, L. C., Devauchelle, G. & Hull, R. (1993). Gene II product of an aphid-nontransmissible isolate of *Cauliflower mosaic virus* expressed in a baculovirus system possesses aphid transmission factor activity. *Virology* **192,** 651–654.

Bonneville, J.-M. & Hohn, T. (1993). A reverse transcriptase for *Cauliflower mosaic virus* state of the art, 1992. In: *Reverse Transcriptase* (Skalka, A. M. & Goff, S. P., eds), pp. 357-390. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.

Bonneville, J. M., Fuetterer, J., Gordon, K., Hohn, T., Martinez-Izquierdo, J., Pfeiffer, P. & Pietrzak, M. (1987). The replication cycle of *Cauliflower mosaic virus* in relation to other retroid elements, current perspectives. In: *Molecular Strategies for Crop Protection* (Arntzen, C. J. & Ryan, C., eds), Vol. 48, pp. 267–293. New York: A.R. Liss.

Bonneville, J. M., Sanfacon, H., Futterer, J. & Hohn, T. (1989). Posttranscriptional trans-activation in *Cauliflower mosaic virus*. *Cell* **59,** 1135–1143.

Brisson, N., Paszkowski, J., Penswick, J. R., Gronenborn, B., Potrykus, I. & Hohn, T. (1984). Expression of a bacterial gene in plants by using a viral vector. *Nature* (*London*) **310,** 511–514.

Campbell, W. H. & Gowri, G. (1990). Codon usage in higher plant, green algae and cyanobacteria. *Plant Physiol.* **92,** 1–11.

Chenault, K. D. & Melcher, U. (1993a). *Cauliflower mosaic virus* isolate CMV-1. *Plant Physiol.* **101,** 1395–1396.

Chenault, K. D. & Melcher, U. (1993b). The complete nucleotide sequence of *Cauliflower mosaic virus* isolate BBC. *Gene* **123,** 255–257.

Chenault, K. D. & Melcher, U. (1994a). Patterns of nucleotide sequence variation among *Cauliflower mosaic virus* isolates. *Biochemie* **76,** 3–8.

Chenault, K. D. & Melcher, U. (1994b). Phylogenetic relationships reveal recombination among isolates of *Cauliflower mosaic virus*. *J. Mol. Evol.* **39,** 496–505.

Chenault, K. D., Steffens, D. L. & Melcher, U. (1992). Nucleotide sequence of *Cauliflower mosaic virus* isolate NY8153. *Plant Physiol.* **100,** 542-545.

Coffin, J. M. (1990). Retroviridae and their replication. In: *Fields Virology* (Fields, B. N. & Knipe, D. M., eds), vol. 2, pp. 1437–1500. New York: Raven Press.

Covey, S. & 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.

De Tapia, M., Himmelbach, A. & Hohn, T. (1993): Molecular dissection of the *Cauliflower mosaic virus* translation transactivator. *EMBO J.* **12,** 3305–3314.

Fang, R., Wu, X., Bu, M., Tian, Y., Cai, F. & Mang, K. (1985). Complete nucleotide sequence of *Cauliflower mosaic virus* (Xinjiang isolate) genomic DNA. *Bing Du Xue Bao* **1,** 247–256.

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

Gardner, R. C., Howarth, A. J., Hahn, P., Brown-Luedi, M., Shepherd, R. J. & Messing, J. (1981). The complete nucleotide sequence of an infectious clone of *Cauliflower mosaic virus* by M13mp7 shotgun sequencing. *Nucl. Acids Res.* **9,** 2871–2888.

Himmelbach, A., Chapdelaine, Y. & Hohn, T. (1996). Interaction between *Cauliflower mosaic virus* inclusion body protein and capsid protein: implications for viral assembly. *Virology* **217,** 147–157.

Hull, R. & Covey, S. N. (1985). *Cauliflower mosaic virus*: pathways of infection. *BioEssays* **3,** 160–163.

Hull, R. & Shepherd, R. J. (1976). The coat proteins of *Cauliflower mosaic virus*. *Virology* **70,** 217–220.

Jacquot, E., Geldreich, A., Keller, M. & Yot, P. (1998). Mapping regions of the *Cauliflower mosaic virus* ORF III product required for infectivity. *Virology* **242,** 395–402.

Katz, R. A., Omer, C. A., Weis, J. H., Mitsialis, S. A., Faras, A. J. & Guntaka, R. V. (1982). Restriction endonuclease and nucleotide sequence analyses of molecularly cloned unintegrated avian tumor virus DNA: structure of large terminal repeats in circle junctions. *J. Virol.* **42,** 346–351.

KNIPE, D. (1990). Virus–host interactions., In: *Fields Virology* (Fields, B. N. & Knipe, D. M., eds), Vol. 1, pp. 293–316. New York: Raven Press.

KOBAYASHI, K., NAKAYASHIKI, H., TSUGE, S., MISE, K. & FURUSAWA, I. (1998). Accumulation kinetics of viral gene products in *Cauliflower mosaic virus*-infected turnip protoplasts. *Microbiol. Immunol.* **42,** 65–69.

LAWRENCE, J. G. & ROTH, J. R. (1996). Selfish operons: horizontal transfer may drive the evolution of gene clusters. *Genetics* **143,** 1843–1860.

LEFEBVRE, D. D., MIKI, B. L. & LALIBERTE, J.-F. (1987). Mammalian metallothionein functions in plants. *Biotechnology* **5,** 1053–1056.

LEH, V., JACQUOT, E., GELDREICH, A., HAAS, M., BLANC, S., KELLER, M. & YOT, P. (2001). Interaction between the open reading frame III product and the coat protein is required for transmission of *Cauliflower mosaic virus* by aphids. *J. Virol.* **75,** 100–106.

LI, M. D., BRONSON, D. L., LEMKE, T. D. & FARAS, A. J. (1995). Phylogenetic analyses of 55 retroelements on the basis of the nucleotide and product amino acid sequence of the *pol* gene. *Mol. Biol. Evol.* **12,** 657–670.

MASON, W. S., TAYLOR, J. M. & HULL, R. (1987). Retroid virus genome replication. *Adv. Virus Res.* **32,** 35–96.

MATTHEWS, R. E. F. (1991). *Plant Virology*, 3rd Edn. San Diego; CA: Academic Press.

MESNARD, J.-M., KIRCHHERR, D., WURCH, T. & LEBEURIER, G. (1990). The C*auliflower mosaic virus* gene III product is a non-sequence-specific DNA binding protein. *Virology* **174,** 622–624.

PIQUE, M., MOUGEOT, J. L., GELDRICH, A., GUIDASCI, T., MESNARD, J. M., LEBEURIER, G. & YOT, P. (1995). Sequence of a *Cauliflower mosaic virus* strain infecting solanaceous plants. *Gene* **155,** 305–306.

ROHRSCHNEIDER, L. R., EISENMAN, R. N. & LEITCH, C. R. (1979). Identification of a *Rous sarcoma virus* transformation-related protein in normal avian and mammalian cells. *Proc. Natl Acad. Sci. USA.* **76,** 4479–4483.

ROTHNIE, H. R., CHAPDELAINE, Y. & HOHN, T. (1994). Pararetroviruses and retroviruses: a comparative review of viral structure and gene expression strategies. *Adv. Virus Res.* **44,** 1–67.

SAS/STAT Users Guide (1989). Version 6, 4th Edn. Cary, NC: SAS Institute.

SCHOELZ, J., SHEPHERD, R. J. & DAUBERT, S. (1986). Region VI of *Cauliflower mosaic virus* encodes a host range determinant. *Mol. Cell. Biol.* **6,** 2632–2637.

SCHWARTZ, D. E., TIZARD, R. & GILBERT, W. (1983). Nucleotide sequence of *Rous sarcoma virus*. Cell **32,** 853–869.

STRATFORD, R. & COVEY, S. (1989). Segregation of *Cauliflower mosaic virus* symptom genetic determinants. *Virology* **172,** 451–459.

STRAUSS, E. G., STRAUSS, J. H. & LEVINE, A. J. (1990). Virus evolution. In: *Fields Virology* (Fields, B. N. & Knipe, D. M., eds), Vol. 1, pp. 167–190. New York: Raven Press.

TAKEYA, T. & HANAFUSA, H. (1982). DNA sequence of the viral and cellular *src* gene of chickens. II. Comparison of the *src* genes of two strains of avian sarcoma virus and of the cellular homolog. *J. Virol.* **44,** 12–18.

THOMAS, T. J. & MAULE, A. J. (1995). Identification of structural domains within the *Cauliflower mosaic virus* movement protein by scanning deletion mutagenesis and epitope tagging. *Plant Cell* **7,** 561–572.

VADEN, V. R. & MELCHER, U. (1990). Recombination sites in *Cauliflower mosaic virus* DNAs: implications for mechanisms of recombination. *Virology* **177,** 717–726.

WURCH, T., KIRCHHERR, D., MESNARD, J.-M. & LEBEURIER, G. (1990). The *Cauliflower mosaic virus* open reading frame VII product can be expressed in *Saccharomyces cerevisiae* but is not detected in infected plants. *J. Virol.* **64,** 2594–2598.