Gene, 19 (1982) 239–249
Elsevier Biomedical Press

Nucleotide sequence of DNA from an altered-virulence isolate D/H of the cauliflower mosaic virus

(Plant DNA virus; Maxam and Gilbert sequencing; genome organization; open reading frames (ORF); deletions; start points for transcription)

Ervin Balàzs, Hubert Guilley, Gérard Jonard and Kenneth Richards

Laboratoire de Virologie, Institut de Biologie Moléculaire et Cellulaire du CNRS, 15 rue Descartes, 67084 Strasbourg Cedex (France)

(Received June 17th, 1982)
(Accepted July 12th, 1982)

SUMMARY

The double-stranded DNA from the isolate D/H with an altered virulence of the cauliflower mosaic virus (CaMV) contains 8016 bp. The DNA is circular and possesses, like the DNA of most CaMV strains, three sequence interruptions.

The comparison of its sequence with the previously published sequences of two other CaMV strains (Cabb-S and CM1841) leads to the following conclusions: (1) The genetic organization of all three CaMV strains is identical with six potential genes (open reading frames) and two intergenic regions; (2) considered pairwise, the three DNAs differ from one another by only about 5% with base substitutions accounting for most of the changes although several deletions and insertions are also observed. The sequence differences among the three strains are spread in a uniform manner upon the genome except for the two intergenic regions, which are more highly conserved. The stability of the noncoding regions is probably linked to the fact that they carry sequences important for the initiation and termination of transcription. On the other hand, the sequence variation in the open reading frames has relatively little effect on the sequence of the corresponding polypeptides as changes occur preferentially in the third position of the reading frame triplets.

It is anticipated that knowledge of the DNA sequences of several CaMV strains will facilitate construction of inter-strain recombinants which, once available, can be used to correlate gene structure and function.

INTRODUCTION

The type member of the DNA-containing caulimoviruses is CaMV. Recent developments in recombinant DNA technology and associated techniques have heightened interest in the molecular biology of CaMV and its potential as a vector for gene transfer in higher plants. The genome of CaMV is a double-stranded circular DNA molecule of about 8000 bp. Virion DNA of most CaMV strains contains three single-strand discon-

Abbreviations: bp, base pairs; CaMV, cauliflower mosaic virus; ORF, open reading frame.

0378-1119/82/0000–0000/$02.75

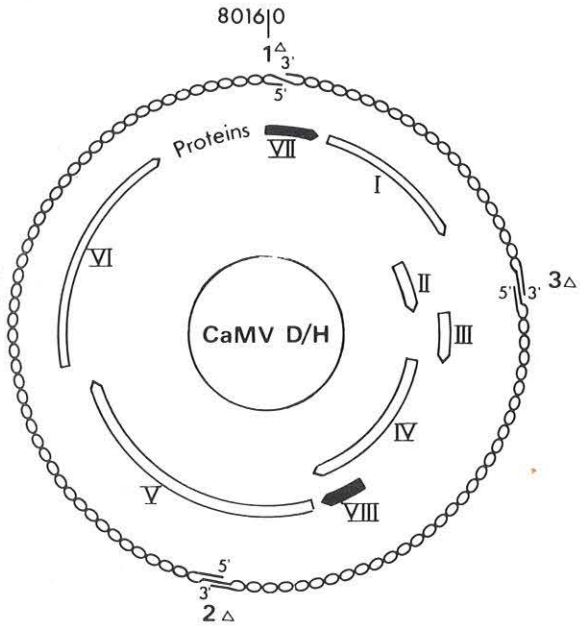


Fig. 1. Physical map of cauliflower mosaic virus DNA. Outermost portion: The double-stranded virion DNA with the three discontinuities (Δ). The 3′ and 5′ free ends at each discontinuity are shown. Inner segments (arrows) show location and direction of the open reading frames in the three phases. Solid black arrows show location and direction of ORFs VII and VIII.

tinuities, two in one strand and one in the other. The single break, 1Δ, in the so-called α-strand of the DNA is taken as the origin of the conventional restriction map. Complete nucleotide sequences are available for the virion DNA of the Cabb-S isolate (Franck et al., 1980) and for an infectious clone of CM1841 (Gardner et al., 1981). Analysis of the distribution of translation initiation and termination codons in these sequences has revealed the existence of six long ORFs on the DNA α-strand, the strand which is transcribed (Fig. 1). It seems reasonable to hypothesize that the six ORFs correspond to virus-coded polypeptides. So far, however, only one such polypeptide has been detected, a protein of M_r 62000 which is the major virus-specific in vitro translation product directed by polyadenylated RNA from infected plants and is coded for by ORF VI (Odell and Howell, 1980; Covey and Hull, 1981). Indirect evidence suggests that the viral coat protein is encoded by ORF IV (Franck et al., 1980; Hahn and Shepherd, 1982).

CaMV is widely distributed throughout the temperate zones of the world with the DNA of some 30 different isolates having been characterized by restriction analysis (Hull, 1980). One isolate examined by Hull (1980), the D/H (or Hungary) isolate, is of particular interest to us because its symptoms differ significantly at the subcellular level from those provoked by Cabb-S with respect to the structure of the viral inclusion bodies which appear in the cytoplasm of CaMV-infected cells (Xiong et al., 1982). It is hoped that recombination experiments (Lebeurier et al., 1982) between the two strains, leading to formation of infectious hybrid DNA molecules, will enable us to pinpoint the portions of the DNA sequence responsible for these cytological changes. As a preliminary to such work we report the complete nucleotide sequence (8016 bp) of Cabb-D/H DNA. The Cabb-D/H sequence is compared to those of Cabb-S and CM1841 both at the nucleotide level and at the amino acid level in the various open reading frames.

MATERIALS AND METHODS

CaMV-D/H isolate (inoculum, a gift of Dr. J. Horvàth in Budapest) was propagated on turnip plants (*Brassica rapa* L. cv. Just Right) grown under growth chamber conditions (22°C, 2000 lux, 16 h/day). Virus was purified using the Triton X-100-urea method (Hull et al., 1976) 4 weeks after inoculation. CaMV DNA was prepared from purified virus according to Hohn et al. (1980). Digestion of the DNA with restriction enzymes was performed as recommended by the suppliers (Biolabs, BRL, Boehringer-Mannheim). Restriction fragments obtained with the following enzymes were characterized in the course of sequence analysis: *Hin*fI, *Dde*I, *Taq*I, *Bgl*II, *Hin*dIII, *Hae*III, *Eco*RI, *Ava*I, *Ava*II, *Cla*I, *Hga*I, *Hha*I and *Sau*3A. A complete restriction site computer listing is available on request.

Preparation of 5′-^{32}P-labelled restriction fragments and sequence analysis followed the method of Maxam and Gilbert (1977) as outlined in detail by Franck et al. (1980). Numerous restriction fragments were also 3′-^{32}P-labelled by fill-in of the 3′ recessed extremities with reverse transcriptase (obtained from Dr. J. Beard, Life Science, Inc.) and

the appropriate [α-^{32}P]deoxynucleoside triphosphate (Smith and Calvo, 1980). After 3′ end labelling the protocol for strand separation and sequence analysis was as for the 5′ end-labelled material. Approx. 50% of the sequence was determined on both strands and $\simeq 90\%$ of the sequence was read from several independent sequence gels. Computer programmes (for which we thank G. de Marcillac) were used to compile data and to compare the D/H sequence to that of Cabb-S and CM1841. Information concerning the details of sequence determination is available on request.

RESULTS AND DISCUSSION

Prior to sequencing, a map of D/H isolate CaMV DNA for a number of commonly used restriction enzymes was established. Some enzymes gave the same digestion profile as for Cabb-S

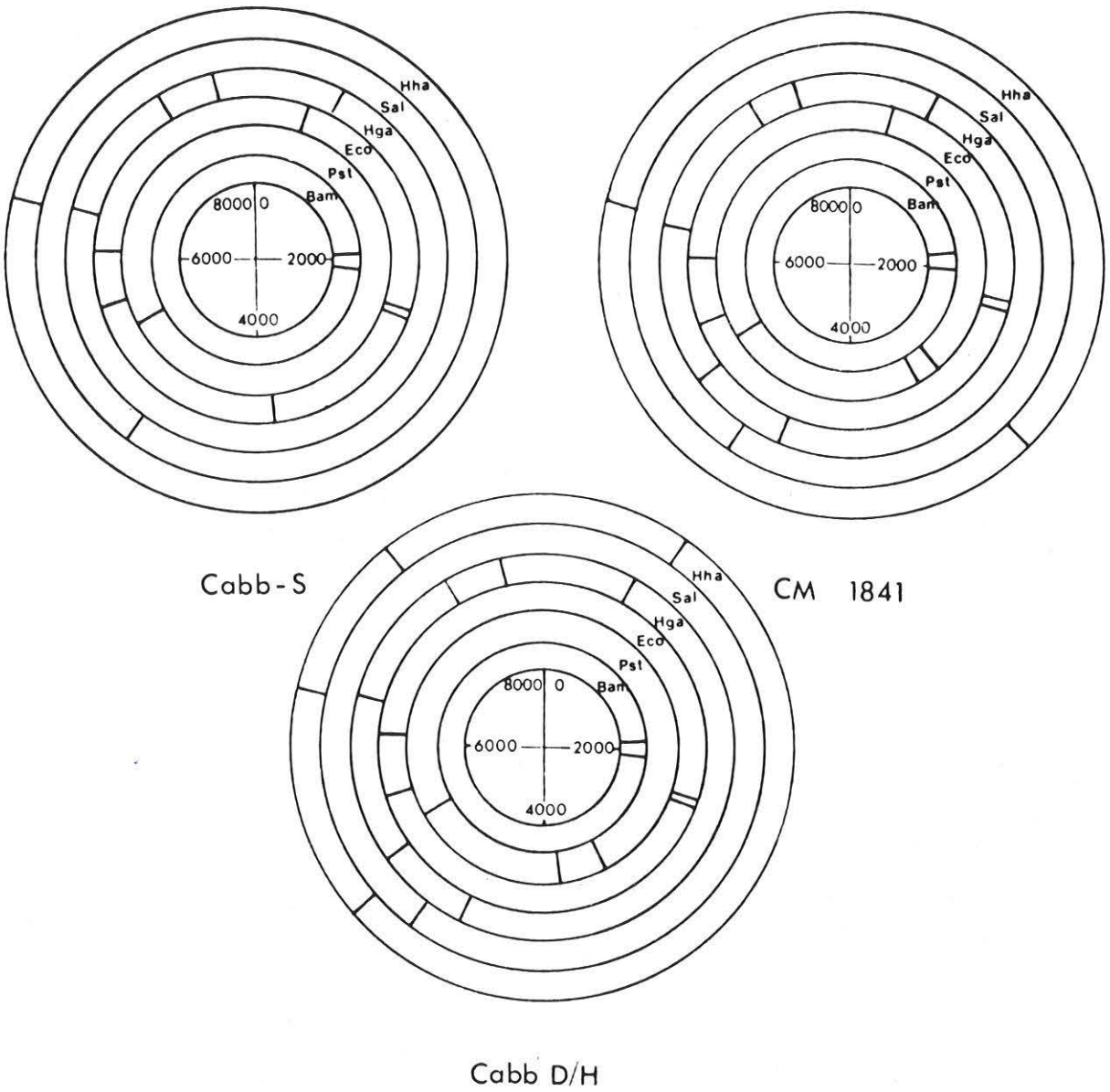


Fig. 2. Restriction maps of the three sequenced strains (Cabb-S: Franck et al., 1980; CM1841: Gardner et al., 1981; and Cabb-D/H).

Fig. 3. Complete nucleotide sequence of D/H DNA. The sequence of the β-strand is presented along with the nucleotide changes compared to the Cabb-S (upper row) and CM1841 (lower row) (deletions are marked as dashes). The sequence is part of the EMBL Sequence Data Library that is available on magnetic tape to the users.

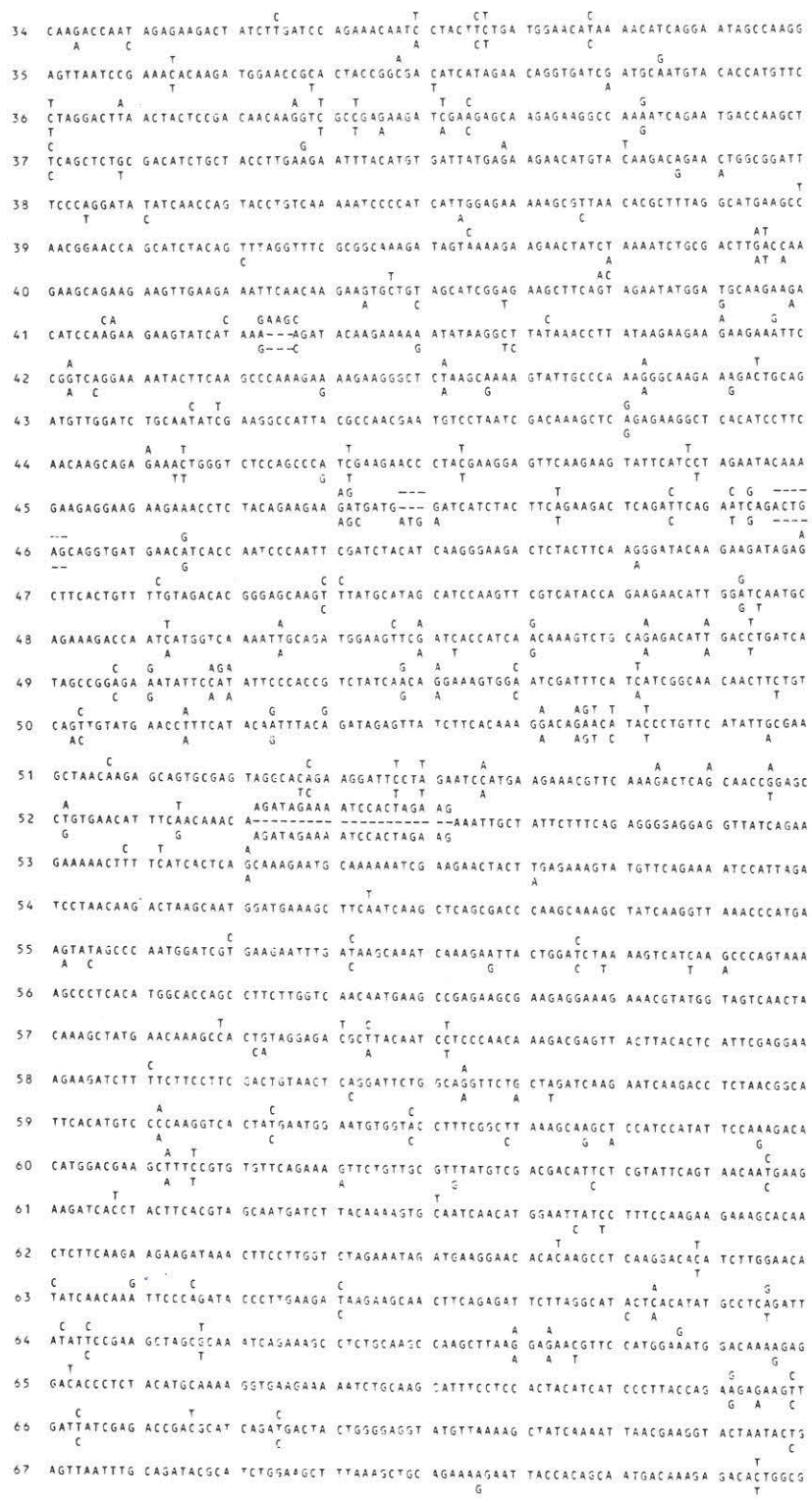


Fig. 3 (continued).

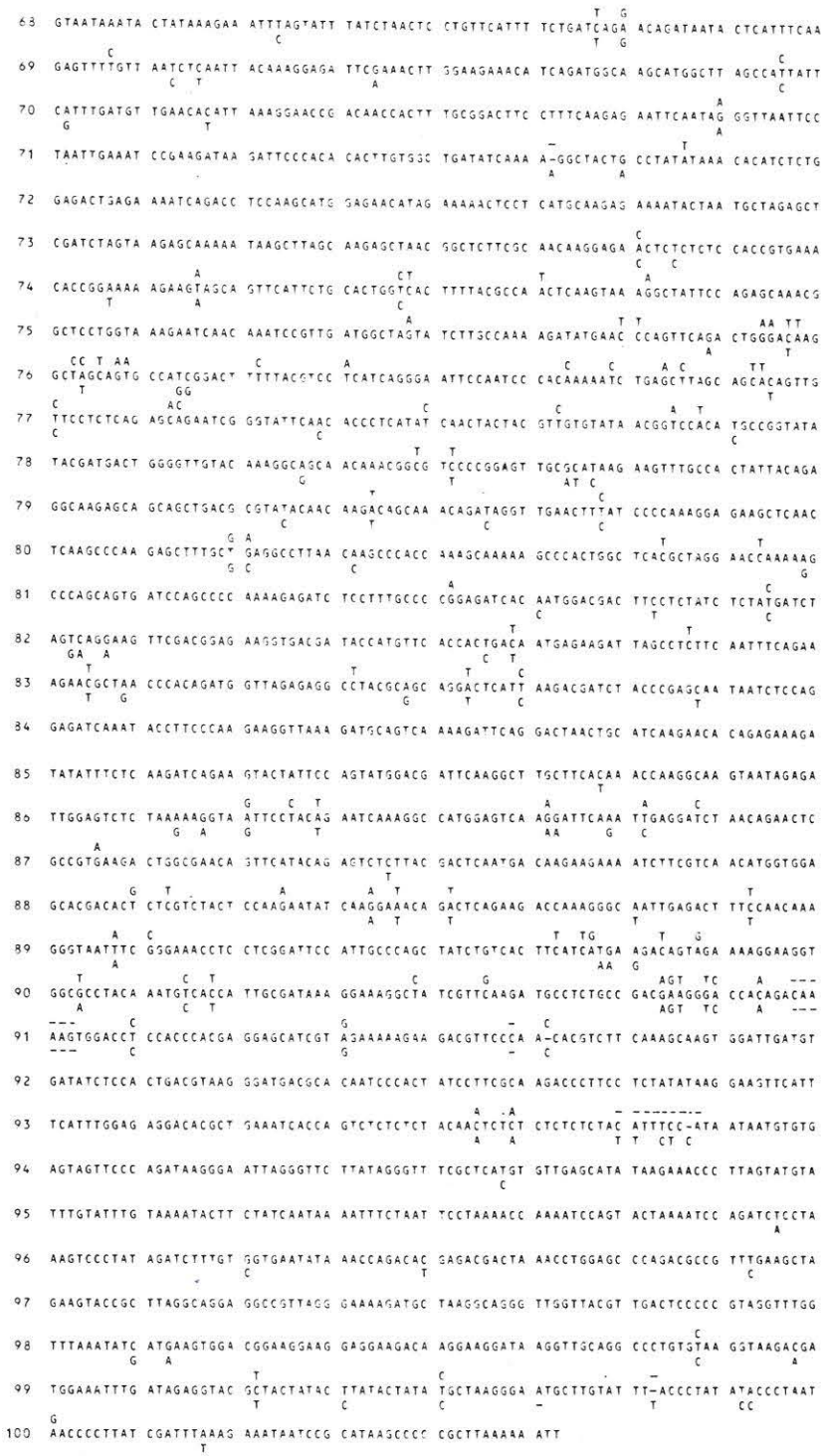


Fig. 3 (continued).

TABLE I

Comparison of nucleotide sequence between the CaMV isolates

The differences among the nucleotide sequences of Cabb-S, CM1841 and Cabb-D/H were examined in each open reading frame and intergenic region of the genome. Deletions (−) and insertions (+) in Cabb-S and CM1841 with respect to the D/H sequence are indicated. The putative protein products of each open reading frame were compared pairwise and the number and % of changes are shown in columns *7* and *8*. Columns *9* and *10* give the fraction of nucleotide changes that result in amino acid changes for each region. The putative coding regions start with the first in-phase methionine signal in each ORF.

Region	Map location (first Met-stop)	Base substitutions		Other changes		Net changes [a]		Amino acid change				Amino acid change per base	
		in S	in CM1841	in S	in CM1841	S%	CM1841 %	in S	%	in CM1841	%	in S	in CM1841
		(1)	*(2)*	*(3)*	*(4)*	*(5)*	*(6)*	*(7)*		*(8)*		*(9)*	*(10)*
ORF I	365–1348	43	45	+5	+5	4.5	4.7	7	2.1	8	2.4	0.16	0.17
ORF II	1345–1824	14	13	+5	+5	3.1	2.9	4	2.5	8	5.0	0.28	0.61
ORF III	1826–2215	9	14	+1 −1	+1 −1	2.8	4.1	3	2.3	6	4.6	0.33	0.42
ORF IV	2197–3669	83	96	−6 +3	−3 +3	5.8	6.7	13	2.6	24	4.8	0.15	0.25
ORF V	3623–5650	94	110	−6 +21	+3 −6 +21	4.7	5.6	14	2.1	19	2.8	0.15	0.17
ORF VI	5754–7322	83	71	+1 −1 −6	+1 −1 −6	5.5	4.7	33	6.3	27	5.2	0.39	0.38
small intergenic	5651–5753	1	2	−	−	1	2.0	–	–	–	–	–	–
long intergenic	7323–8016	6	20	−8	+1	1	3.0	–	–	–	–	–	–
	1– 364	12	17	−1	−1	3.5	4.9	–	–	–	–	–	–
Possible coding regions													
VII	13– 303	11	14	−	−	3.8	4.8	3	3.1	11	11.5	0.27	0.78
VIII	3260–3583	23	21	+3	−	7.4	6.5	11	10.3	15	14.0	0.47	0.71

[a] Each deletion or insertion counts as a single change in the %. Calculated as sum of (1)+(3) or (2)+(4) divided by length of the interval in bp (second column).

or CM1841 DNA while others gave a quite different pattern. Fig. 2 presents a comparison of the restriction maps for the three strains.

The complete nucleotide sequence of D/H DNA is presented in Fig. 3. The sequence consists of 8016 bp; numbering begins with the first dG nearest the 5′ boundary of 1Δ. Only the sequence of the β-strand, which is of the same polarity as viral mRNA, is presented along with nucleotide changes compared to the Cabb-S and CM1841 sequences. The sequence around the three discontinuities in the virion DNA is conserved in all three isolates and the position upon the DNA sequence of the 5′ extremity of the interrupted strand at each discontinuity in D/H DNA is identical to that observed for Cabb-S DNA.

Although the three CaMV DNA sequences have not been subjected to detailed statistical analysis, the base changes appear to be distributed in a roughly uniform manner upon the genome with two exceptions: the small intergenic region separating ORFs V and VI (residues 5651–5753) and the intergenic region between the end of ORF VI and 1Δ (Table I). This sequence stability is no doubt due, at least in part, to the localization of transcription initiation and termination signals in these regions (Covey et al., 1981; Dudley et al., 1982; Guilley et al., 1982). It is interesting to note that one of the two changes distinguishing the small intergenic region of the Cabb-S sequence from that of D/H and CM1841 is the substitution of a T for an A at position 5702 of the D/H sequence (Fig. 4A). The change in question occurs in the middle of a TATA-box type sequence (TATATAA for D/H and CM1841; TATTTAA in Cabb-S) which falls 33 bp upstream of one of the two major viral transcription start points (Guilley et al., 1982). An unusual change in the long intergenic region is a 9- or 8-bp insertion found in the CM1841 and D/H sequences, respectively but lacking from Cabb-S (residue 7480). The absence of this insert creates a new TATA-box type sequence, TATAATA, in the long intergenic region of Cabb-S DNA (Fig. 4B) 45 bp downstream from the second major transcription start point (Guilley et al., 1982). There is no evidence, however, that this new TATA box can function as a transcription start signal in vivo or in vitro (Guilley et al., 1982).

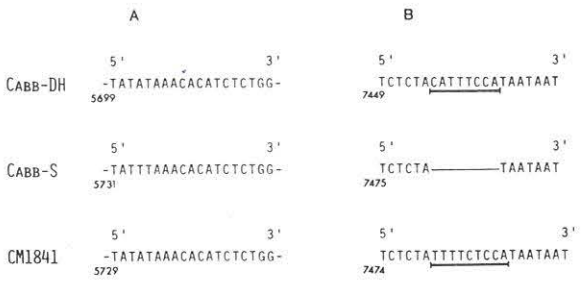


Fig. 4. Sequence changes in the intergenic regions. (A) Change in the middle of a TATA-box type sequence which falls 33 bp upstream of one of the two major transcription start points (Guilley et al., 1982). (B) Insertions found in CM1841 and Cabb-D/H sequences. The absence of this insert creates a new TATA-box type sequence in Cabb-S.

About 85% of the CaMV DNA α-strand sequence can encode protein. The overall organization of this coding capacity into six major open reading frames is closely similar in all three strains. One interesting difference in the D/H sequence occurs at the junction between ORFs I and II. In the Cabb-S and CM1841 sequences the TAA termination codon ending ORF I is separated by a single nucleotide from the first in-phase ATG codon of ORF II while in the D/H sequence the two ORFs overlap by four nucleotides (Fig. 5A). This change is the result of a 5 bp deletion at the ORF I-ORF II junction in the D/H sequence which eliminates the original stop and initiation signals but generates new ones (Fig. 5A). A second distinction occurs at the end of ORF IV where a 6-bp insertion has occurred in the D/H sequence which, when coupled with a mutation eliminating the original TGA termination codon, elongates ORF IV by 6 bp (Fig. 5B).

Compared to the Cabb-S and CM1841 sequences about 75% of the base changes in ORFs I

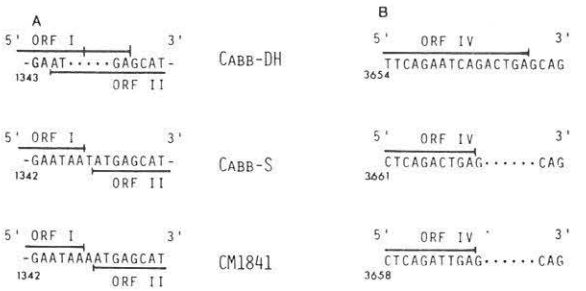


Fig. 5. Changes in sequence at the junctions of open reading frames. (A) Deletion in the D/H sequence at the junction of ORF I and II producing an overlap. (B) Insertion at the end of ORF IV in the D/H sequence elongates ORF IV by 6 bp.

to VI of the D/H strain are silent, that is, without effect on the amino acid assignment at that position. The amino acid changes that do occur are shown in Fig. 6. Note that the highest overall rate of change occurs in ORF VI with an extent of amino acid substitution of 6.3% and 5.2% compared to the Cabb-S and CM1841 sequences, respectively. This variability may be related to the distinct structure of the inclusion bodies provoked by infection with D/H strain as the region VI polypeptide is believed to be a major component of these structures (Xiong et al., 1982). Likewise of interest is the extreme sequence conservation of the carboxyl-terminal half of the ORF V polypeptide in which only two changes, both involving lysine-arginine substitutions, are found (Fig. 6A,B,C).

Hohn et al. (1981) have shown that the Cabb-S

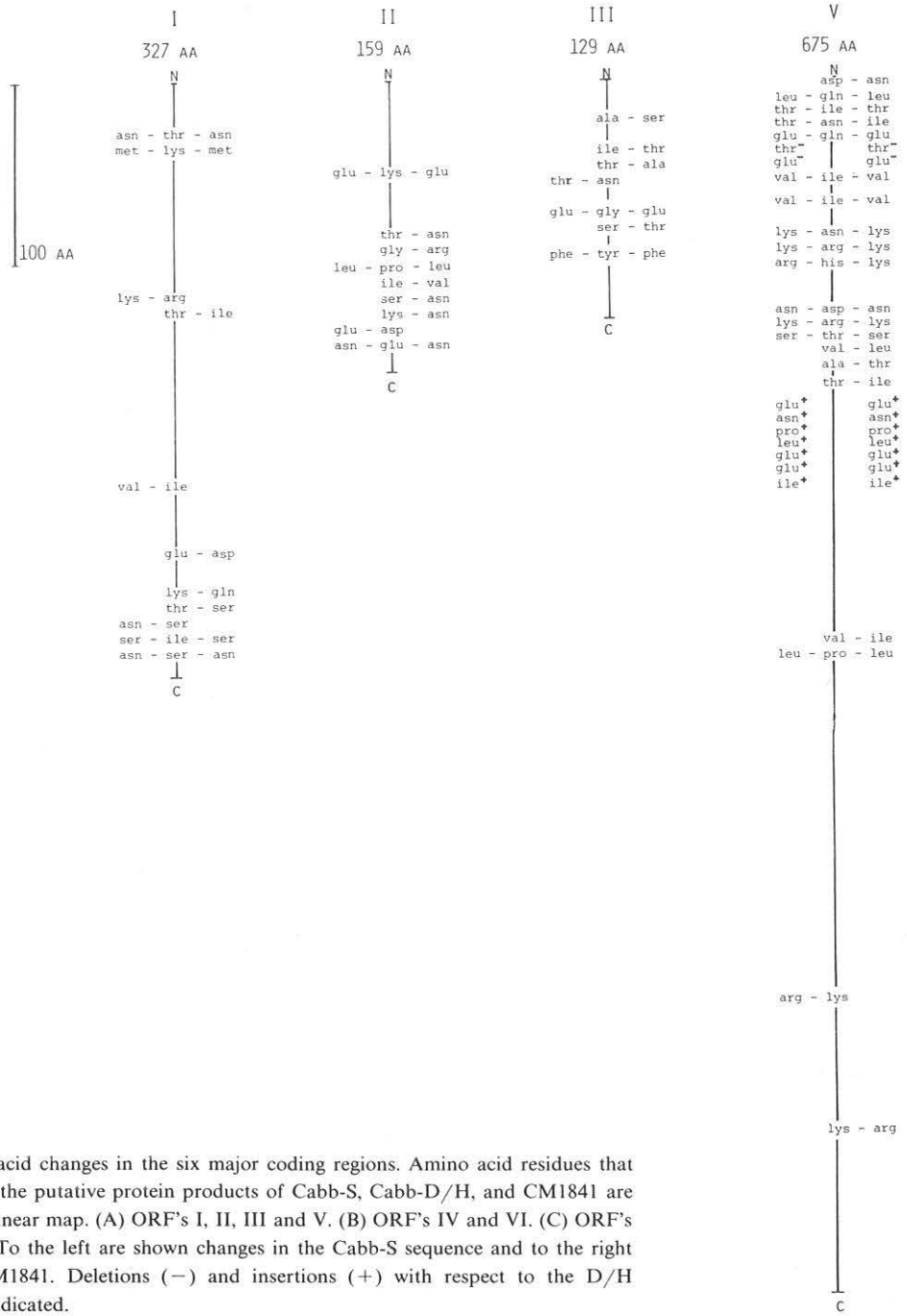
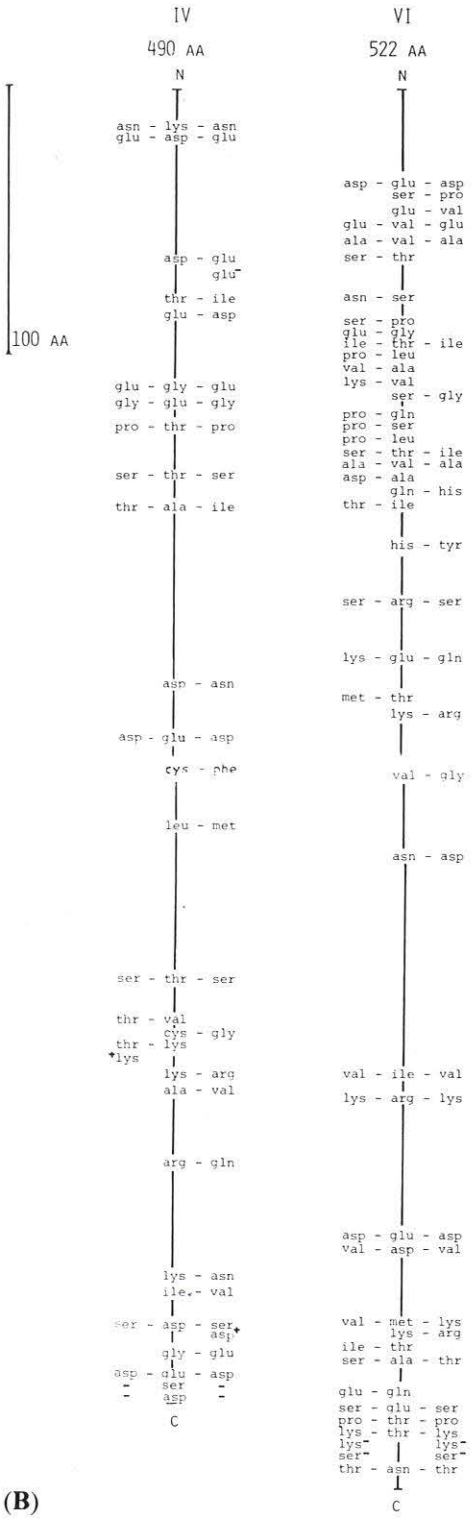


Fig. 6. Amino acid changes in the six major coding regions. Amino acid residues that differ between the putative protein products of Cabb-S, Cabb-D/H, and CM1841 are arranged as a linear map. (A) ORF's I, II, III and V. (B) ORF's IV and VI. (C) ORF's VII and VIII. To the left are shown changes in the Cabb-S sequence and to the right changes in CM1841. Deletions (−) and insertions (+) with respect to the D/H sequence are indicated.

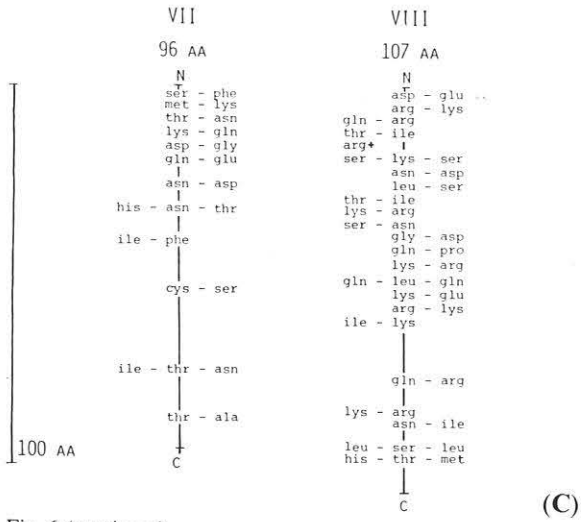
(A)



(B)

Fig. 6 (continued).

sequence contains two additional open reading frames, termed VII (96 triplets) and VIII (107 triplets), with ORF VII falling between Δ1 and



(C)

Fig. 6 (continued).

ORF I and ORF VIII overlapping the end of ORF IV but in a different reading frame (see Fig. 1). In view of their small size, it is impossible to judge from the Cabb-S sequence data alone whether these ORFs have occurred by chance or whether they correspond to authentic viral polypeptides. Nevertheless, it is of interest that both ORF VII and VIII are also present in the D/H and CM1841 sequences; the failure of mutational events to introduce an in-phase termination codon into either the D/H or CM1841 DNA in these regions can be taken as weak support for the hypothesis that ORFs VII and VIII are functional coding regions. If, on the other hand, we compare the putative polypeptides encoded by ORFs VII and VIII in the three strains it is clear that the extent of amino acid substitution in each is notably higher than for coding regions I–VI (Table I; Fig. 6A,B,C) considerably weakening the protein coding hypothesis. It is evident that characterization of the viral-coded polypeptides present in CaMV-infected plant tissue will be required to resolve this issue.

ACKNOWLEDGEMENTS

We are highly indebted to Prof. L. Hirth for his constant interest in the work and many suggestions and discussions. This work was supported by ATP No. 2080 of the Centre National de la Recherche Scientifique.

REFERENCES

Covey, S.N. and Hull, R.: Transcription of cauliflower mosaic virus DNA. Detection of transcripts, properties and location of the gene encoding the virus inclusion body protein. Virology 111 (1981) 463–474.

Covey, S.N., Lomonossoff, G.P. and Hull, R.: Characterization of cauliflower mosaic virus DNA sequences which encode major polyadenylated transcripts. Nucleic Acids Res. 9 (1981) 6735–6747.

Dudley, R.K., Odell, J.T. and Howell, S.H.: Structure and 5′-termini of the large and 19 S RNA transcripts encoded by the cauliflower mosaic virus genome. Virology 117 (1982) 19–28.

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

Gardner, R.C., Howarth, A.J., Hahn, P., Brown-Luedi, M., Shepherd, R.J. and Messing, J.: The complete nucleotide sequence of an infectious clone of cauliflower mosaic virus by M13mp7 shot gun sequencing. Nucleic Acids Res. 9 (1981) 2871–2887.

Guilley, H., Dudley, R.K., Jonard, G., Balàzs, E. and Richards, K.E.: Transcription of cauliflower mosaic virus DNA: Detection of promotor sequences by in vitro transcription and characterization of the polyadenylated transcripts found in virus infected plants. Submitted.

Hahn, P. and Shepherd, R.J.: Evidence for a 58-kilodalton polypeptide as a precursor of the coat protein of cauliflower mosaic virus. Virology 116 (1982) 480–488.

Hohn, T., Hohn, B., Lesot, A. and Lebeurier, G.: Restriction map of original and cloned cauliflower mosaic virus DNA. Gene 11 (1980) 21–31.

Hohn, T., Richards, K.E. and Lebeurier, G.: Cauliflower mosaic virus on its way to becoming a useful plant vector, in Hofschneider, P.H. and Goebel, W. (Eds.), Current Topics in Microbiology and Immunology 96. Springer-Verlag, Berlin, Heidelberg and New York, 1981, pp. 193–236.

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

Hull, R., Shepherd, R.J. and Harvey, J.D.: Cauliflower mosaic virus: An improved purification procedure and some properties of the virus particles. J. Gen. Virol. 31 (1976) 93–100.

Lebeurier, G., Hirth, L., Hohn, B. and Hohn, T.: In vivo recombination of cauliflower mosaic virus DNA. Proc. Natl. Acad. Sci. USA 79 (1982) 2932–2936.

Maxam, A.M. and Gilbert, W.: A new method for sequencing DNA. Proc. Natl. Acad. Sci. USA 74 (1977) 560–564.

Odell, J.T. and Howell, S.H.: The identification, mapping and characterization of mRNA for P66, a cauliflower mosaic virus-coded protein. Virology 102 (1980) 349–359.

Smith, D.R. and Calvo, J.M.: Nucleotide sequence of the E. coli gene coding for dihydrofolate reductase. Nucl. Acids Res. 8 (1980) 2255–2274.

Xiong, C., Balàzs, E., Lebeurier, G., Hindenlang, C., Stoeckel, M.E. and Porte, A.: Comparative cytology of two isolates of cauliflower mosaic virus. J. Gen. Virol. 61 (1982) 75–81.

Communicated by W. Fiers.