Similarities Between Putative Transport Proteins of Plant Viruses

Article *in* Journal of General Virology · May 1990

DOI: 10.1099/0022-1317-71-5-1009

Journal of General Virology (1990), 71, 1009–1018. *Printed in Great Britain*

Similarities between putative transport proteins of plant viruses

Ulrich Melcher

Department of Biochemistry, Oklahoma State University, Stillwater, Oklahoma 74078-0454, U.S.A.

The nucleic acids of many plant viruses encode proteins with one or more of the following properties: an M_r of approximately 30000, localization in the cell wall of the infected plant and a demonstrated role in cell-to-cell transport of infection. A progressive alignment strategy, aligning first those sequences known to be similar, and then aligning the resulting groups of sequences, was used to examine further the relatedness of the amino acid sequences of putative transport proteins of caulimoviruses, of proteins similar to the putative transport protein of alfalfa mosaic virus (AlMV) and of those similar to the tobacco mosaic virus (TMV) 30K protein. The strategy first identified regions in which multiple dipeptides of one group were similar to those of another group. The regions of similarity were brought into alignment by the conservative introduction of gaps. The positions of the introduction of gaps were adjusted to optimize similarity. Statistical significances of the resulting alignments, determined both by comparison with shuffled amino acid sequences and with the sequence alignment off-set by 1 to 15 residues in each direction, suggest that the amino acid sequences of the three groups of viruses are distantly related. Nevertheless, significant relationships between members of the caulimoviral group of sequences and members of each of the AlMV-like and TMV-like groups were found. These relationships and the analysis of the number of insertions/deletions between present sequences and a hypothetical common ancestor suggest that the sequences of the caulimoviral proteins are less diverged from the ancestor than either the AlMV-like or TMV-like proteins. The alignment identified common regions of predicted secondary structure and regions of similar hydropathy, regions possibly crucial for proper functioning of the proteins.

Introduction

Tobacco mosaic virus (TMV) RNA encodes a polypeptide with an M_r of approximately 30000 (30K) that is required for transport of the infection from cell to cell in plants (Meshi *et al.*, 1987; Deom *et al.*, 1987). The polypeptide has been located in the nuclear fraction of infected protoplasts (Watanabe *et al.*, 1986). However, immunogold labelling suggests that it is associated with plasmodesmata of cell walls of infected plants (Tomenius *et al.*, 1987). The protein accumulates in the cell wall fraction of the systemic host *Nicotiana tabacum* cv. Samsun but disappears from this fraction during the hypersensitive response in cv. Samsun NN hosts (Moser *et al.*, 1988).

The assignment of a transport function to proteins encoded by other viruses (reviewed in Atabekov & Dorokhov, 1984, and Hull, 1989) is supported by subliminal infection by subsets of RNAs of multicomponent viruses, complementation of movement of unrelated viruses, cytological location of the putative transport protein in cell walls, and amino acid sequence similarity. Although the TMV 30K (Tomenius *et al.*, 1987), the cauliflower mosaic virus (CaMV) gene I product (Albrecht *et al.*, 1988; Linstead *et al.*, 1988) and the alfalfa mosaic virus (AlMV) 3a (Godefroy-Colburn *et al.*, 1986) proteins are located in cell walls, the cucumber mosaic virus (CMV) protein was identified immunologically in nucleoli and not found in cell walls of tobacco plants (MacKenzie & Tremaine, 1988). The amino acid sequences of polypeptides of approximately 30K encoded by tobacco rattle virus (TRV), cucumber green mottle mosaic virus (CGMMV), and two caulimoviruses, CaMV and carnation etched ring virus (CERV) have been shown to be significantly similar to those of the proteins of TMV isolates (Hull *et al.*, 1986). The AlMV amino acid sequence has significant similarity to sequences of proteins encoded by tobacco streak virus (TSV) (A. Gibbs *et al.*, unpublished, cited in Cornelissen *et al.*, 1984), brome mosaic virus (BMV) and CMV (Savithri & Murthy, 1983). However, no significant sequence similarity was found between sequences of the proteins of the tripartite genome viruses (TSV, AlMV, BMV, CMV) and the sequences related to the TMV transport protein (Cornelissen & Bol, 1984; Boccara *et al.*, 1986). A regional sequence similarity of uncertain significance was noted between the BMV protein and the TMV 30K protein (Hull, 1989). Zimmern (1983) noted

similarity between the amino acid sequences of the 30K proteins of TMV and sunn-hemp mosaic virus (SHMV) with those of two peptides encoded by mitochondrial introns of *Saccharomyces cerevisiae*. Possibly significant similarities were found between the CaMV protein and several cellular proteins (Hull *et al.*, 1986; Martinez-Izquierdo *et al.*, 1987).

Amino acid sequences derived from nucleotide sequences are available from a substantial number of plant viruses. Amino acid sequences corresponding to the viral coat protein can be identified by similarities in amino acid composition of virions and open reading frame (ORF) products or by comparison of the predicted sequence with coat polypeptide sequences, if such are available. Predicted polypeptides responsible for nucleic acid polymerase function have been identified based on the presence of stretches of amino acid sequence conserved in known polymerases (Kunin *et al.*, 1987). The assignment of other functions, including transport of infection, to the remaining polypeptides has been difficult. Identification of genes for transport proteins in other plant viruses is important, not only for an understanding of pathogenesis by those viruses, but also for understanding the evolutionary relationships among plant viruses (Goldbach, 1987; Malyshenko *et al.*, 1989). I have re-examined the similarities between the amino acid sequences of the putative transport polypeptides to find conserved sequence features which could be used to identify transport proteins encoded by other viruses. I report that caulimoviral sequences [CaMV, CERV and figwort mosaic virus (FMV)] were significantly similar to sequences of proteins similar to the AlMV 3a protein (encoded by RNAs of CMV, BMV, TSV and AlMV) and of those similar to the TMV 30K protein (encoded by RNAs of TRV, CGMMV, SHMV and two strains of TMV). The inter-relatedness of the sequences suggests a common evolutionary origin or a common function for the putative transport polypeptides encoded by these plant viruses and should assist in the identification of transport protein genes in other plant viruses.

Methods

Sources of sequences. Sequences of transport proteins of TMV used were: common strain (TMVC) (Goelet *et al.*, 1982), tomato strain (TMVL) (Takamatsu *et al.*, 1983), and cowpea strain (Meshi *et al.*, 1982), also known as SHMV (Kassanis & Varma, 1975). The sequence of the common strain differs from that of a Japanese isolate (Meshi *et al.*, 1982) in five positions. Sequences of the 30K protein of CGMMV (Meshi *et al.*, 1983; Saito *et al.*, 1988) and the 29K protein of RNA-1 of TRV (Boccara *et al.*, 1986) were also used. The AlMV protein sequence used was that of the 5′ extreme ORF of RNA3 of the M strain (Barker *et al.*, 1983). The amino acid sequences of proteins of the S (Ravelonandro *et al.*, 1984) and L (Langereis *et al.*, 1986) strains were not separately analysed, since they contained only a limited number of amino acid substitutions relative to the M isolate. Other tripartite virus sequences used were: the 31·7K polypeptide of TSV RNA3 (Cornelissen *et al.*, 1984), the 3a polypeptides of BMV RNA3 (Ahlquist *et al.*, 1981), and of CMV RNA3 strain Q (Davies & Symons, 1988). The sequence of proteins encoded by CMV strain O (Hayakawa *et al.*, 1989) and the very similar strain Y (Nitta *et al.*, 1988) were not separately analysed. Three ORF I product caulimoviral sequences were used: CaMV, Cabb S isolate (Franck *et al.*, 1980), CERV (Hull *et al.*, 1986) and FMV (Richins *et al.*, 1987). Sequences of ORF I polypeptides of several other isolates of CaMV (Balazs *et al.*, 1982; Gardner *et al.*, 1981; Dixon *et al.*, 1986; Hirochika *et al.*, 1985) were not separately analysed since they differ only slightly from that of Cabb S. For comparison purposes, significance scores were determined as detailed below for published alignments that differed significantly from the present one. These included the alignment by Savithri & Murthy (1983) of AlMV (Barker *et al.*, 1983), BMV (Ahlquist *et al.*, 1981) and CMV (Gould & Symons, 1982) sequences and that of Saito *et al.* (1988) for TMV-like proteins. Alignments of globin sequences (Dayhoff, 1972) were used as standards for determination of the significance of relationships.

Alignment method. Alignment of multiple sequences was achieved in steps (Feng & Doolittle, 1987). First, pairwise alignments of sequences known to be highly related (BMV and CMV, AlMV and TSV, TMVC and TMVL, and CaMV and CERV) were performed. Then SHMV, CGMMV and TRV sequences were added, in that order, to the TMV-like group. The FMV sequence was added to the two other caulimoviral sequences and the BMV–CMV pair was aligned with the TSV–AlMV pair. Three groups of aligned sequences resulted: the TMV-like group, the caulimoviral group, and the AlMV-like group. Next, the TMV-like group was aligned with the caulimoviral group. Finally, the AlMV-like group was aligned with the TMV–caulimoviral groups.

At each step, one sequence or group of sequences was designated 'master' and the other 'slave'. A dipeptide look-up table (Lipman & Pearson, 1985) was constructed of the positions of each possible dipeptide in every member of the master group. A reduced alphabet (in which Asp is equivalent to Glu, Asn to Gln, Lys to Arg, Ser to Thr, Tyr and Phe to Trp, and Ile and Leu and Val to Met) was used for this construction. For each dipeptide in every member of the slave group, the differences between its position and those of the matching dipeptides in the lookup table, the off-set values, were calculated. The number of matches at each off-set value was counted. Off-set values with the 10 highest number of matches were identified. The master group of sequences was set off relative to the slave group by the number of residues suggested by the identified off-set values, and similarity scores over a stretch of 10 residues were calculated for each position. Similarity scores were the sum of values of the log odds matrix (Dayhoff, 1972, Table 10-1) for all possible comparisons of master with slave sequences at each position. The approach identified stretches of sequences that produced high scores when aligned at the selected off-sets.

High scoring stretches of sequence aligned at different off-sets were brought into alignment by gap translation. A gap of length equal to the difference in off-set values for adjacent high scoring stretches was introduced into the appropriate sequence set at a position several residues N-terminal to the position where the gap was expected to belong. The gap was then moved, one residue position at a time, a specified distance towards the C terminus. At each step the effect on the similarity score of translating the gap was determined. The position at which the effect on the score was maximally positive was accepted as the position at which the gap should be introduced. If no maximum was found the gap introduction was rejected.

After gap introduction, the alignment was inspected visually to identify gaps that could be removed without greatly reducing the similarity score. Gaps of equal length in each of the master and the

slave sets (or a subset of one, but not both) within 20 residues of one another were not allowed unless the sequences between the gaps showed unmistakable similarity ('once a gap always a gap'; Feng & Doolittle, 1987). Gaps within 20 residues of one another in the same sequence set were consolidated and their position re-examined by gap translation, unless there were compelling reasons to allow them to exist as two separate gaps. Gap consolidation was not permitted when the vicinal gaps were present in two different subsets of the sequences being compared. Calculation of the off-set distributions and introduction of gaps by gap translation were repeated until no further significant improvements in similarity scores resulted. The positioning of gaps, both newly introduced and pre-existing, was re-examined by gap translation prior to the next step in the progressive alignment.

Determination of significance. In determining the significance of alignments obtained by completely objective algorithms, randomized sequences of the same amino acid composition were subjected to the alignment algorithm. The scores of the final alignment were compared with the mean of a sufficiently large number of alignments of random sequences. Since the present method is not completely objective, randomized sequences cannot be optimally aligned objectively. Two other approaches were used to assess the significance of each resulting pairwise alignment.

Shuffled sequences (Feng & Doolittle, 1987) were generated by sequential placement, from N to C terminus, of residues randomly chosen from the amino acid composition of the protein. The lengths and positions of gaps were not altered. A similarity score between the second sequence and the shuffled sequence was determined. The shuffling and scoring was repeated 30 times to generate a mean shuffled score and a standard deviation. The significance score was calculated as the difference between the similarity score for the test alignment and the mean of the shuffled scores divided by the standard deviation. Significance was also assessed by off-set comparison. The pair of aligned sequences to be evaluated were first scanned to remove blanks that occurred in common positions. A similarity score for the two sequences was calculated. Sequence A was then moved one residue towards the C terminus and the similarity score calculated again. The process of moving by one residue was repeated for a total of 15 residues towards the C terminus. Sequence B was similarly set off one residue at a time toward the C terminus. The mean and standard deviation of the 30 off-set similarity scores were calculated. The significance score was determined as described above.

Computation. Secondary structure predictions using the Chou & Fasman (1978) algorithm and hydropathy values determined according to Hopp & Woods (1981) and to Kyte & Doolittle (1982) were obtained using the Bionet-Intelligenetics computer resource. Secondary structure predictions using the Garnier *et al.* (1978) algorithm were performed with the Macintosh version of the Molgenjr program (Lowe, 1986). Computer programs implementing the off-set value determinations, gap translation, significance score calculations by shuffling and by off-set were written in Zbasic for the Apple IIe and are available on request from the author.

Results

Alignment

The progressive alignment strategy produced alignments of the sequences of each of the three groups of putative transport proteins: the TMV-like group, the caulimovirus group, and the AlMV-like group. As expected, each intragroup comparison resulted in one or more off-set values at which large numbers of dipeptides in one set of sequences were similar to those in the other set. These off-set values allowed for easy alignment of the sequences. The three possible two-way dipeptide similarity off-set comparisons between the groups were then performed. Consistent with the reported relationship of caulimoviruses to the TMV group (Hull *et al.*, 1986) and the reported lack of relationship between the AlMV-like group and the TMV-like group (Cornelissen & Bol, 1984; Boccara *et al.*, 1986), the highest dipeptide similarity was noted between the caulimovirus group and the TMV-like group. Therefore, the progressive alignment was continued aligning these two groups and then adding the AlMV-like group.

The resulting alignment is shown in Fig. 1. Only one position, glycine 227, is invariant. Proline is found in 11 sequences at position 193. Ten sequences have leucine at position 160. Aspartic acid or asparagine occupy position 162 in all 12 sequences. Numerous positions are occupied exclusively by hydrophobic or hydrophilic amino acids. The larger size of the caulimoviral proteins appears to be due to an N-terminal extension which is not present in the other proteins. Of the 14 different amino acid replacements in sequences of ORF I proteins of three CaMV isolates relative to CabbS and 17 different replacements in sequences of two AlMV isolates relative to the M strain, 13 are conservative replacements. The positions at which the remaining 18 occur are each occupied by a variety of residues in the 12 aligned sequences. Of the 46 differences between the Q and O strains of CMV (Hayakawa *et al.*, 1989), 19 were in the variable C-terminal region and a cluster of six overlapped the five-residue insert specific to sequences of BMV and CMV proteins. Of the remaining differences only two (Arg to Cys and Leu to Thr at positions 127 and 208, respectively; Fig. 1) were not conservative.

The Chou & Fasman (1978) algorithm was employed to predict propensities of regions of the sequences to form α-helices, β-structure and β-turns. As a consequence of alignment, regions likely to assume a β-structure conformation tended to occur at the same positions in each of the 12 sequences (Fig. 2*b*). Similar coincidences of regions in the aligned sequences likely to be α-helical (Fig. 2*a*) and of regions with β-turn propensity (Fig. 2*c*) were also noted. The predicted structures of the C-terminal domains beyond position 320 were rich in turns and random coil conformations. In many specific regions of the remainder more than half of the sequences were predicted to have the same secondary structure: six regions of β-structure, five α-helical regions, and seven turn regions. Glycine 227 and proline 193 occurred in probable turn regions. Positions 160 and 162 were in a region of apparent transition between β-structure and α-helix. Prediction of secondary structure with the Garnier *et al.* (1978) algorithm increased the proportion of α-helix

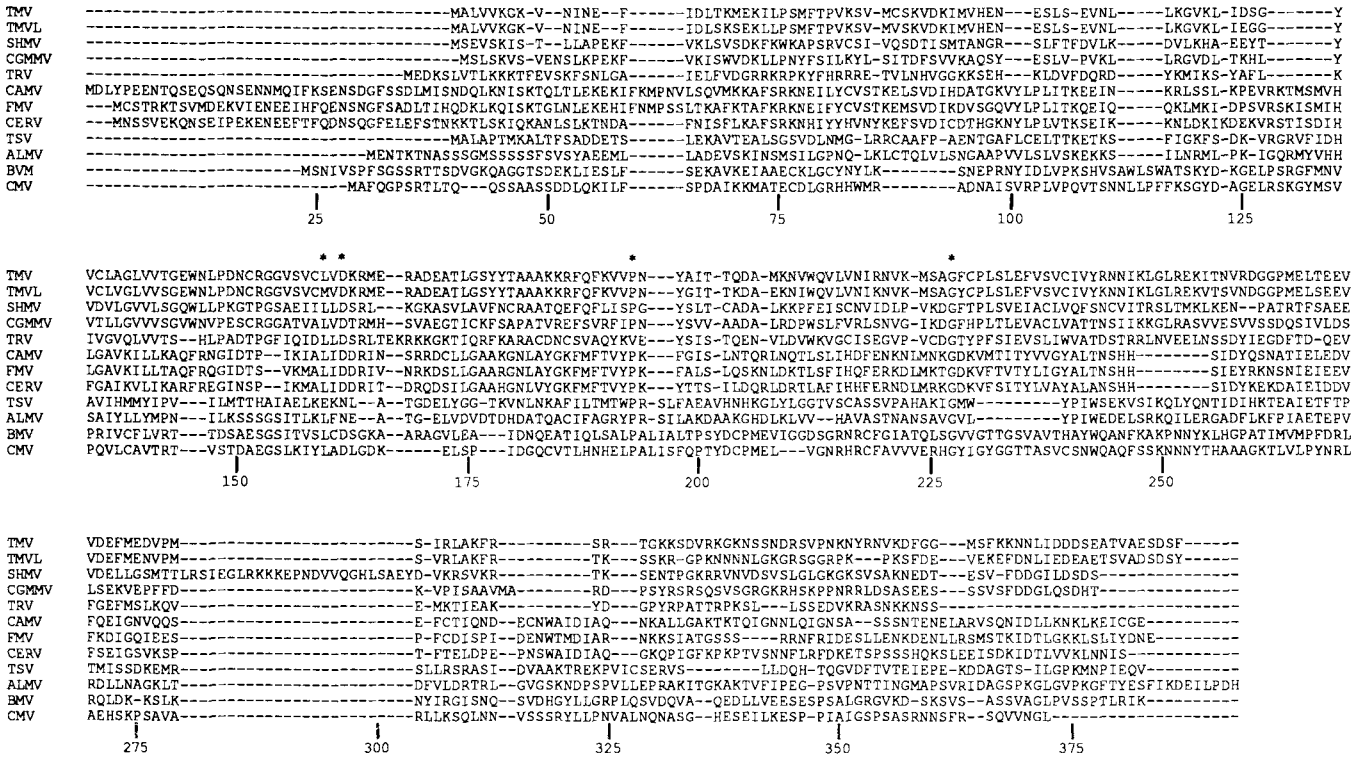

Fig. 1. Amino acid sequence alignment of putative cell-to-cell transport proteins of 12 viruses.

to β-structure for all sequences to an extent that prevented analysis of the randomness of their distribution.

Hydropathy distributions (Hopp & Woods, 1981) for sequences of a representative protein of the caulimoviruses and the TMV-like group and for two of the AlMV-like group show common profiles (Fig. 3). Particularly notable are profiles between positions 130 and 180 and between 220 and 255. The former region includes the almost invariant positions 160 and 162 while the invariant glycine is contained in the latter. Analysis of hydropathy distributions of other proteins in the alignment (not shown) and using the Kyte & Doolittle (1982) values gave results similar to those shown (Fig. 3).

To determine whether the alignment (Fig. 1) and the strategy used to produce it can be used to determine whether a sequence might be related to this family, the ungapped TSV sequence, one of the sequences least related to most of the others (see below), was compared to the other 11 sequences aligned. The distribution of dipeptide similarities as a function of off-set position had significant peaks that led to the alignment of TSV illustrated in Fig. 1.

Significance of the alignment

Whether the alignments of Fig. 1 reflect significant relationships between the sequences was assessed by calculating the number of standard deviations by which the similarity score for each alignment exceeded the mean of scores of a set of off-set alignments (upper right half of Table 1). For comparison, values obtained for aligned sequences of haemoglobin polypeptides (Dayhoff, 1972) determined by the same method ranged from 3·9 for the midge haemoglobin–leghaemoglobin comparison to 12·8 for the comparison of human α and β chains. All comparisons between members of the TMV-like and the caulimoviral groups gave values above 3·9, whereas only some comparisons between AlMV-like proteins and members of the other two groups exceeded this value.

Sequence shuffling (Feng & Doolittle, 1987), in which one sequence is scrambled, preserving the position and length of gaps, was also used to examine the significance of the alignments. When applied to representative globin sequences, significance scores ranged from 5·3 (lamprey globin versus leghaemoglobin) to 17·9 (α versus β chains of human haemoglobin). The scores for globin pairs were on average higher than off-set significance scores for the same pairs, mainly due to the larger standard deviations in the off-set comparisons. Both the off-set and the shuffle methods yielded significance scores for the least similar globin chains that were substantially higher than those calculated using the Needleman & Wunsch (1970) algorithm (lamprey globin versus leghaemoglobin 2·8;

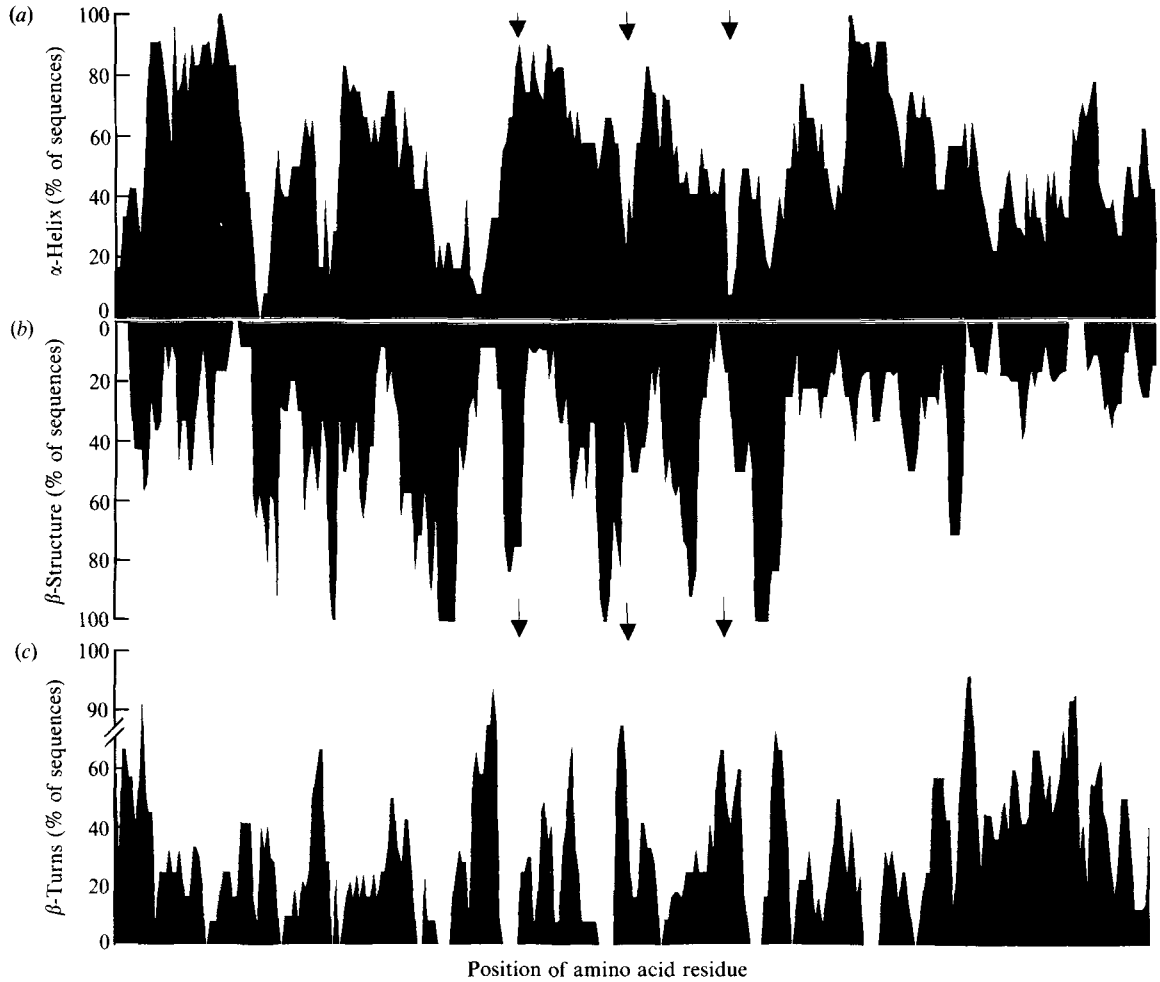

Fig. 2. Secondary structure-forming propensities of the cell-to-cell transport proteins of plant viruses. For each position in the alignment of Fig. 1 (except positions where more than six sequences have gaps) the percentage of sequences in which the residue at that position is likely to form α-helix (*a*), β-sheet (*b*) or β-turn (*c*) structures is plotted. Arrows indicate locations of relatively conserved residues at, from left to right, positions 160, 193 and 227.

Table 1. *Significance scores* for transport protein amino acid sequence alignment*

	TMVC	TMVL	SHMV	CGMMV	TRV	CaMV	FMV	CERV	TSV	AlMV	BMV	CMV
TMVC	–	**38·5**	**11·1**	**12·3**	**7·8**	**5·2**	4·7†	**6·3**	2·9	3·1	6·1†	4·9†
TMVL	35·1	–	**11·4**	**13·0**	**8·2**	**5·4**	4·2†	**4·9**	2·9	2·3	4·8†	5·3†
SHMV	12·2	10·3	–	**12·2**	**7·4**	**8·2**	**7·4**	**8·8**	2·3	2·7	3·5	5·2†
CGMMV	13·1	20·7	13·2	–	**7·0**	**5·0**	4·4†	**4·6**	3·2	2·5	3·5	5·4†
TRV	7·6	6·4	7·4	7·2	–	6·7†	**5·4**	**6·3**	1·8	3·4	2·0	2·7
CaMV	5·4	6·1	9·3	5·9	5·0	–	**22·5**	**20·4**	4·3†	**4·0**	**4·7**	3·6
FMV	4·9	4·2	7·4	4·5	5·9	29·6	–	**19·9**	**4·5**	**5·1**	3·8	3·4
CERV	7·5	5·8	6·2	5·7	5·7	29·0	25·5	–	2·9	3·0	4·5†	3·6
TSV	4·0	3·2	2·7	3·7	1·8	4·1	6·1	5·1	–	**10·6**	4·8†	4·6†
AlMV	3·2	2·6	2·9	2·2	3·0	5·4	6·2	3·6	12·5	–	**10·4**	**6·0**
BMV	3·6	4·4	5·0	4·2	1·6	5·8	3·5	4·3	4·5	8·8	–	**22·6**
CMV	4·4	3·4	4·4	3·6	2·3	2·4	3·4	3·1	5·2	6·3	19·5	–

* Values above the diagonal were determined by the off-set method, those below the diagonal by the sequence shuffling method. Values in bold face represent comparisons for which significance scores were greater than 3·9 by the off-set method and greater than 5·3 by the shuffling method.

† Comparisons for which significance scores were greater than 3·9 by the off-set method but less than 5·3 by the shuffling method.

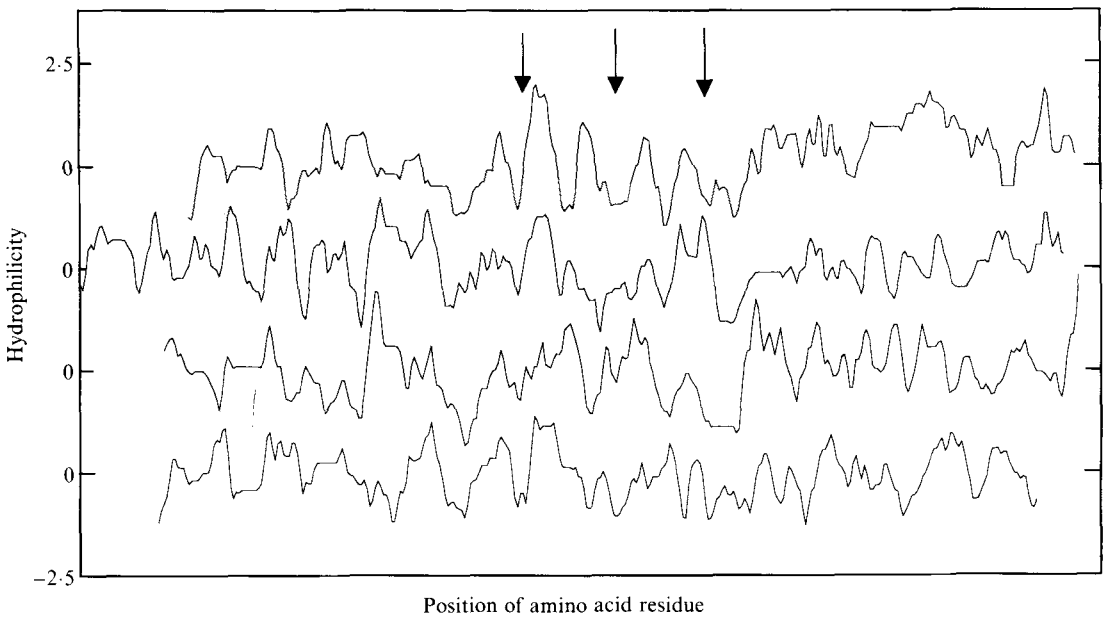

Fig. 3. Hydrophilicity profiles of putative cell-to-cell transport proteins of, from top to bottom, TMV, CaMV, AlMV and CMV in the alignment of Fig. 1. At positions where sequences have gaps, the average of the values for the residues immediately before and immediately after the gap were used. Positions where all four sequences have gaps are not included. Arrows indicate locations of relatively conserved residues at, from left to right, positions 160, 193 and 227.

midge globin versus leghaemoglobin 3·0; Dayhoff, 1972). Significance scores for the viral transport protein alignment determined by shuffling (lower left half of Table 1) were in general agreement with those determined by the off-set method. However, some comparisons produced scores as much as 76% different in the two methods. The off-set method, on average, produced significance scores which were higher, but not significantly so, than those produced by shuffling. Pairs of sequences that generated significance scores greater than those for distantly related globin chains in both methods of determination (Table 1) included all intragroup comparisons in the TMV group and the caulimoviral group and all but the TSV-BMV and TSV-CMV pairs in the AlMV-like group. Also included were all TMV-like caulimoviral comparisons, except FMV paired with TMVC, TMVL and CGMMV, and CaMV paired with TRV. The latter comparison had scores above those of distantly related globins for only the off-set method. Significant relationships involving the AlMV-like proteins were AlMV versus FMV and CaMV, BMV versus CaMV, and TSV versus FMV. Some significance scores for comparisons of BMV, CMV and TSV with the proteins of the TMV-like group and the caulimoviral group exceeded the scores for distantly related globins in only the off-set comparison.

Indel tree

Insertions and deletions (indels) are less frequent evolutionary events than substitutions. Fig. 4 depicts the

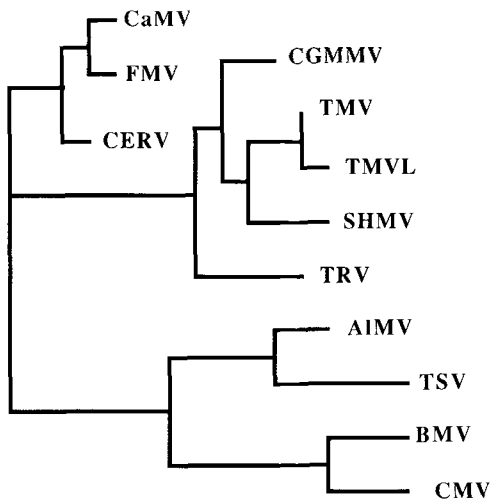

Fig. 4. Diagram of the number of insertions/deletions separating amino acid sequences of putative cell-to-cell transport proteins of the aligned sequences of Fig. 1. Horizontal distance is proportional to the number of indels.

number of indels inserted in the steps of the progressive alignment. Four assumptions resolved ambiguities in constructing the tree. The apparent deletion at position 106 was assumed to have arisen twice, once in TMVC and TMVL and once in CGMMV sequences. The CGMMV residues at 314 to 315 were assumed to have arisen by insertion after deletion in the TMV-group ancestor of residues equivalent to 316 to 323. The insertion at positions 195 to 197 of three residues in BMV

and CMV proteins and only two residues in TSV and AlMV proteins was assumed due to insertion of three residues in their common ancestor followed by deletion of one residue in the TSV–AlMV specific lineage. Similarly, the deletion at 347 in TSV, AlMV and CMV proteins, but not in that of BMV, was assumed to be due to deletion in their common ancestor and subsequent insertion to produce the BMV protein. The former two assumptions were necessitated by the assumed branching order. The latter two only affect the positions of the nodes and were adopted to equalize the length of branches. The average number of indels separating the TMV-like group of sequences or the AlMV-like group of sequences from the common ancestor node with the caulimoviruses was significantly higher than the number of indels separating the caulimoviruses from that node. The distances from the common ancestor node to the members of the TMV-like group and of the AlMV-like group were not significantly different.

Relationship to other alignments

The present alignment is in reasonable agreement with some complete or partial published alignments of some members of the group of sequences examined here, but not with others. In aligning BMV and CMV proteins there is one fewer indel in the present alignment than in that of Murthy (1983) and the indels that are common are in approximately, though not precisely, the same places. The alignment of proteins of AlMV, BMV and CMV (Fig. 1) differs substantially from that suggested by Savithri & Murthy (1983). The present alignment resulted in higher significance scores for all three comparisons than those obtained from the Savithri & Murthy alignment (6·7, 4·8, 18·0 for BMV versus AlMV, CMV versus AlMV and CMV versus BMV, respectively, in the shuffle method). On the basis of similar periodicities of β-turn, β-sheet and hydrophilicity profiles, Davies & Symons (1988) implied an alignment of AlMV and CMV sequences that differs from the present one. The comparison of the TSV protein with those of BMV and CMV yielded significance scores higher than those for distantly related globins for only one comparison method, consistent with the lack of similarity noted by A. Gibbs *et al.* (unpublished, cited in Cornelissen *et al.*, 1984). The alignment of portions of TMVC and TRV sequences (N-terminal 100 residues) shown by Domier *et al.* (1987) differs substantially from the present one except for the last 19 amino acids. The partial alignment of proteins of TRV, SHMV, TMVC and CGMMV of Boccara *et al.* (1986) and that of Meyer *et al.* (1986) of TMVC, TMVL, SHMV and CGMMV proteins are similar but not identical to that in Fig. 1. Significance scores for the Saito *et al.* (1988) alignment determined either by the off-set or the sequence shuffling method are essentially the same as those for the alignment presented here. The complete alignment of this group of sequences presented by Saito *et al.* (1988) differs primarily from the present one in that the former is more liberal in the introduction of gaps (59 as opposed to 14). The liberal introduction of gaps obscures a large insert in the SHMV sequence apparent in the alignment presented here (Fig. 1). The conserved 90 amino acid central domain found by Hull *et al.* (1986) between tobamoviruses and caulimoviruses matches the present one (except for their mistaken TRV sequence). The second domain aligned by Hull *et al.* (1986) differs significantly from that in Fig. 1. Some of the regional similarities noted (Meyer *et al.*, 1986) between 30K sequences of tobamoviruses and a portion of the polyprotein encoded by tomato black ring virus (TBRV) RNA2 occur in the more conserved regions of Fig. 1, including positions 160 to 162. Other regional similarities to tobamoviral 30K sequences for TBRV and for a portion of the polyprotein encoded by the RNA of the cowpea mosaic virus (CPMV) middle component did not correspond to more conserved regions of Fig. 1. The possible relationship of the TBRV, CPMV, potyviral (Domier *et al.*, 1987) and other (Hull, 1989) proteins to those in Fig. 1 requires further investigation.

Yeast mitochondrial intron sequences were aligned with putative transport protein sequences essentially as suggested by Zimmern (1983), making a few improvements to adjust for the more conservative gap policy used in the present alignment. The significance score for the alignment of the two intron-encoded sequences was 5·5 by shuffling. The shuffling method confirmed a significant relationship of the polypeptide of the fourth intron of the apocytochrome *b* gene and the TMVC and TMVL proteins (significance scores of 4·9 and 8·3, respectively). Significance scores for the apocytochrome *b* peptide compared against the remaining viral proteins and for the peptide from an intron of the yeast cytochrome oxdase subunit 1 against all 12 viral protein sequences ranged from $-1{\cdot}8$ to $+2{\cdot}3$ and were thus judged not to be significant. Alignment of these intron-encoded regions with the transport protein sequences by the present strategy failed to improve the scores significantly.

Discussion

In addition to the known significant relationships between the caulimoviral group and the TMV-like group of amino acid sequences of putative transport proteins (Hull *et al.*, 1986), such relationships also exist between sequences of the AlMV-like and the caulimoviral proteins (FMV versus TSV and AlMV, CaMV versus

AlMV and BMV, Table 1). The transport proteins of the AlMV-like viruses are thus indirectly related to those of the TMV-like group. Indeed, six of 10 comparisons of BMV or CMV sequences with those of the TMV-like group had significance scores calculated by the off-set method that were higher than the score for distantly related globin chains (Table 1). Confidence that the significance scores measure similarity adequately is enhanced by the observation that, with only one exception (TMVC versus apocytochrome *b*), the alignment of transport protein sequences with those of peptides specified by mitochondrial introns (Zimmern, 1983) did not yield significance scores higher than those for distantly related globins. Confidence is enhanced also by the failure of the strategy to align haemoglobin and cytochrome *c* sequences to the putative transport protein sequences (data not shown). Clearly defined regions of common predicted secondary structures (Fig. 2) and the observation of regions with common hydropathic profiles (Fig. 3) also support the hypothesis that the alignment is meaningful. That non-conserved sequence differences between isolates of the same virus are found at positions whose amino acid residues are also not conserved among different viruses also supports the accuracy of the alignment. The present alignment should prove useful in testing whether unknown reading frames in sequenced plant viruses belong to the transport protein gene class.

Intergroup similarities of sequences may be due in part to the proposed similar function of these proteins. The similarities are not uniformly distributed along the length of the sequence. As noted for the TMV group (Saito *et al.*, 1988) N-terminal and C-terminal regions are less similar in sequence than internal regions. Dissimilar regions could be responsible for functions specific to individual proteins, such as nucleolar or cell wall localization (Tomenius *et al.*, 1987; MacKenzie & Tremaine, 1988). The regions of highest similarity in the alignment correspond well to those noted by Saito *et al.* (1988) for proteins of the TMV group, except that the regions rich in basic and in acidic amino acids are not as prominent in non-TMV sequences. Temperature-sensitive transport mutations in TMV alter a residue seven positions N-terminal (Zimmern & Hunter, 1983) and another three positions C-terminal (Ohno *et al.*, 1983) of glycine 227, the only invariant residue. The identification of residues, such as glycine 227, proline 193, leucine 160 and aspartic acid 162, and regions that are relatively conserved should facilitate examination of the function of the transport protein by site-directed mutagenesis. *In vivo* complementation studies (Taliansky *et al.*, 1982) suggested that transport proteins are host-specific rather than virus-specific. Host-specific sequences were not apparent in the alignment and some significance scores for proteins of viruses with different hosts (BMV with TMVC and TMVL) were higher than those for proteins of viruses with the same hosts (TSV with TMVC and TMVL). However, the opposite expectation, that transport protein sequences should be highly conserved due to their interaction with a conserved host cell component, is also not corroborated. Studies of the biological functions of these polypeptides should clarify the roles of similar and dissimilar domains.

The most striking aspect of the alignment is the relative absence of residues conserved in proteins of most or all viruses. A considerable number (19 out of 66) of the aligned sequence pairs did not produce significance scores higher than those of distantly related globins by either the off-set or the shuffle methods. Thus, significant similarity may not be necessary for proteins to have the transport function. Capsid proteins of icosahedral viruses have very similar three-dimensional structures, yet may be not related perceptibly in amino acid sequence (Rossmann & Johnson, 1989). The lack of significant similarity between some transport proteins suggests that transport proteins may not be best suited to reveal distant relationships among viruses. However, the sequences of other proteins common to the expression repertoire of all plant viruses may reflect differences related to function more strongly than differences related to evolution. Thus, the sequences of virus-encoded polymerases depend on the replication strategy of the viral nucleic acid (Kunin *et al.*, 1987) and the sequences of capsid proteins can be expected to depend on virion morphology as much as on viral evolution.

Nonetheless, significant inter-group similarity exists and may thus reflect a common evolutionary origin (homology). The findings that fewer indels separate the caulimoviruses from the TMV-like group and from the AlMV-like group than separate the TMV group from the AlMV-like group and that significant similarities are more prominent for comparisons with caulimoviral proteins than for comparisons between proteins of the AlMV-like group and proteins of the TMV-like group are consistent with a common evolutionary origin of these sequences. The unique position of the caulimoviral sequences as links in similarity between proteins of the TMV-like group and the AlMV-like group suggests that fewer indels and substitutions have occurred during evolution of the caulimoviral sequences from a hypothetical common ancestor than in the evolution of the other two groups. The apparently slow rate of caulimoviral transport protein evolution could be explained if the modern caulimoviruses arose more recently than the RNA viruses and the transport protein gene was a captured host gene, changing only slowly while in the host genome. Alternatively, the DNA-containing caulimoviruses could be evolutionary fossils, evolving more

slowly than their RNA-containing cousins. Rates of nucleotide substitution have not been measured for caulimoviruses. Rates of nucleotide substitution in vertebrate retroviruses, viruses that also use reverse transcription in replication, are comparable to those of other vertebrate RNA viruses (Steinhauer & Holland, 1987). The apparent rapid fixation of substitutions in FMV DNA under selective pressure (Gowda *et al.*, 1987) is consistent with rapid substitutions, although a role for selection of a pre-existing variant in the adaptation has not been ruled out. On the other hand, several observations suggest that mechanisms exist for the *in planta* correction and homogenization of CaMV DNA sequences (Choe *et al.*, 1985; Melcher *et al.*, 1986). Such mechanisms could limit the rate of change in the genomes of caulimoviruses.

This research was supported in part by the Oklahoma Health Research Program (HR8-048), by the NSF (grant DMB-8515397 to J. Sherwood and U. M.), by an NIH grant to Intelligenetics, and by the Oklahoma Agricultural Experiment Station of which this is Journal article no. J-5669.

References

AHLQUIST, P., LUCKOW, V. & KAESBERG, P. (1981). Complete nucleotide sequence of brome mosaic virus RNA3. *Journal of Molecular Biology* **153**, 23–38.

ALBRECHT, H., GELDREICH, A., MENISSIER DE MURCIA, J., KIRCHNER, D., MESNARD, J. & LEBEURIER, G. (1988). Cauliflower mosaic virus gene I product detected in a cell-wall-enriched fraction. *Virology* **163**, 503–508.

ATABEKOV, J. G. & DOROKHOV, Y. L. (1984). Plant virus-specific transport function and resistance of plants to viruses. *Advances in Virus Research* **29**, 313–364.

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.

BARKER, R. F., JARVIS, N. P., THOMPSON, D. V., LOESCH-FRIES, L. S. & HALL, T. C. (1983). Complete nucleotide sequence of alfalfa mosaic virus RNA3. *Nucleic Acids Research* **11**, 2881–2891.

BOCCARA, M., HAMILTON, W. D. O. & BAULCOMBE, D. C. (1986). The organisation and interviral homologies of genes at the 3′ end of tobacco rattle virus RNA1. *EMBO Journal* **5**, 223–229.

CHOE, I. S., MELCHER, U., RICHARDS, K., LEBEURIER, G. & ESSENBERG, R. C. (1985). Recombination between mutant cauliflower mosaic virus DNAs. *Plant Molecular Biology* **5**, 281–289.

CHOU, P. Y. & FASMAN, G. D. (1978). Prediction of the secondary structure of proteins from their amino acid sequence. *Advances in Enzymology and Related Areas of Molecular Biology* **47**, 45–148.

CORNELISSEN, B. J. C. & BOL, J. F. (1984). Homology between the proteins encoded by tobacco mosaic virus and two tricornaviruses. *Plant Molecular Biology* **3**, 379–384.

CORNELISSEN, B. J. C., JANSSEN, H., ZUIDEMA, D. & BOL, J. F. (1984). Complete nucleotide sequence of tobacco streak virus RNA3. *Nucleic Acids Research* **12**, 2427–2437.

DAVIES, C. & SYMONS, R. H. (1988). Further implications for the evolutionary relationships between tripartite plant viruses based on cucumber mosaic virus RNA3. *Virology* **165**, 216–224.

DAYHOFF, M. O. (1972). *Atlas of Protein Sequence and Structure*, vol. 5. Washington, D.C.: National Biomedical Research Foundation.

DEOM, C. M., OLIVER, M. J. & BEACHY, R. N. (1987). The 30-kilodalton gene product of tobacco mosaic virus potentiates virus movement. *Science* **237**, 389–394.

DIXON, L., NYFFENEGGER, T., DELLEY, G., MARTINEZ-IZQUIERDO, J. & HOHN, T. (1986). Evidence for replicative recombination in cauliflower mosaic virus. *Virology* **150**, 463–468.

DOMIER, L. L., SHAW, J. G. & RHOADS, R. E. (1987). Potyviral proteins share amino acid sequence homology with picorna-, como- and caulimoviral proteins. *Virology* **158**, 20–27.

FENG, D. F. & DOOLITTLE, R. F. (1987). Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *Journal of Molecular Evolution* **25**, 351–360.

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. *Nucleic Acids Research* **9**, 2871–2888.

GARNIER, J., OSGUTHORPE, D. J. & ROBSON, B. (1978). Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *Journal of Molecular Biology* **120**, 97–120.

GODEFROY-COLBURN, T., GAGEY, M.-J., BERNA, A. & STUSSI-GARAUD, C. (1986). A non-structural protein of alfalfa mosaic virus in the walls of infected tobacco cells. *Journal of General Virology* **67**, 2233–2239.

GOELET, P., LOMONOSSOFF, G. P., BUTLER, P. J. G., AKAM, M. E., GAIT, M. J. & KARN, J. (1982). Nucleotide sequence of tobacco mosaic virus RNA. *Proceedings of the National Academy of Sciences, U.S.A.* **79**, 5818–5822.

GOLDBACH, R. (1987). Genome similarities between plant and animal RNA viruses. *Microbiological Sciences* **4**, 197–202.

GOULD, A. R. & SYMONS, R. H. (1982). Cucumber mosaic virus RNA3. Determination of the nucleotide sequence provides the amino acid sequences of protein 3A and viral coat protein. *European Journal of Biochemistry* **126**, 217–226.

GOWDA, S., RICHINS, R. D. & SHEPHERD, R. J. (1987). Host adaption by figwort mosaic virus. *Phytopathology* **77**, 1704.

HAYAKAWA, T., MIZUKAMI, M., NAKAJIMA, M. & SUZUKI, M. (1989). Complete nucleotide sequence of RNA 3 from cucumber mosaic virus (CMV) strain O: comparative study of nucleotide and amino acid sequences among CMV strains O, Q, D and Y. *Journal of General Virology* **70**, 499–504.

HIROCHIKA, H., TAKATSUJI, H., UBASAWA, A. & IKEDA, J.-E. (1985). Site-specific deletion in cauliflower mosaic virus DNA: possible involvement of RNA splicing and reverse transcription. *EMBO Journal* **4**, 1673–1680.

HOPP, T. P. & WOODS, K. R. (1981). Prediction of protein antigenic determinants from amino acid sequences. *Proceedings of the National Academy of Sciences, U.S.A.* **78**, 3824–3828.

HULL, R. (1989). The movement of viruses in plants. *Annual Review of Phytopathology* **27**, 213–240.

HULL, R., SADLER, J. & LONGSTAFF, M. (1986). The sequence of carnation etched ring virus DNA: comparison with cauliflower mosaic virus and retroviruses. *EMBO Journal* **5**, 3083–3090.

KASSANIS, B. & VARMA, A. (1975). Sunn-hemp mosaic virus. *CMI/AAB Descriptions of Plant Viruses*, no. 153.

KUNIN, E. V., GORBALENYA, A. E., CHUMAKOV, K. M., DONCHENKO, A. P. & BLINOV, V. M. (1987). Evolution of RNA-dependent RNA polymerases of positive strand riboviruses. *Molekularnaya Genetika, Mikrobiologia i Virusologia* **7**, 27–39.

KYTE, J. & DOOLITTLE, R. F. (1982). A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology* **157**, 105–132.

LANGEREIS, K., MUGNIER, M.-A., CORNELISSEN, B. J. C., PINCK, L. & BOL, J. F. (1986). Variable repeats and poly(A)-stretches in the leader sequence of alfalfa mosaic virus RNA3. *Virology* **154**, 409–414.

LINSTEAD, P. J., HILLS, G. J., PLASKITT, K. A., WILSON, I. G., HARKER, C. L. & MAULE, A. J. (1988). The subcellular location of the gene 1 product of cauliflower mosaic virus is consistent with a function associated with virus spread. *Journal of General Virology* **69**, 1809–1818.

LIPMAN, D. J. & PEARSON, W. R. (1985). Rapid and sensitive protein similarity searches. *Science* **227**, 1435–1441.

LOWE, J. R. (1986). Transportable microcomputer programs for DNA and protein analysis. *Federation Proceedings* **45**, 1852.

MACKENZIE, D. J. & TREMAINE, J. H. (1988). Ultrastructural location of non-structural protein 3A of cucumber mosaic virus in infected tissue using monoclonal antibodies to a cloned chimeric fusion protein. *Journal of General Virology* **69**, 2387–2395.

MALYSHENKO, S. I., KONDAKOVA, O. A., TALIANSKY, M. E. & ATABEKOV, J. G. (1989). Plant virus transport function: complementation by helper viruses is non-specific. *Journal of General Virology* **70**, 2751–2757.

MARTINEZ-IZQUIERDO, J. A., FUETTERER, J. & HOHN, T. (1987). Protein encoded by ORF I of cauliflower mosaic virus is part of the viral inclusion body. *Virology* **160**, 527–530.

MELCHER, U., CHOE, I. S., LEBEURIER, G., RICHARDS, K. & ESSENBERG, R. C. (1986). Selective allele loss and interference between cauliflower mosaic virus DNAs. *Molecular and General Genetics* **203**, 230–236.

MESHI, T., OHNO, T. & OKADA, Y. (1982). Nucleotide sequence of the 30K protein cistron of cowpea strain of tobacco mosaic virus. *Nucleic Acids Research* **10**, 6111–6117.

MESHI, T., KIYAMA, R., OHNO, T. & OKADA, Y. (1983). Nucleotide sequence of the coat protein cistron and the 3′ noncoding region of cucumber green mottle mosaic virus (watermelon strain) RNA. *Virology* **127**, 54–64.

MESHI, T., WATANABE, Y., SAITO, T., SUGIMOTO, A., MAEDA, T. & OKADA, Y. (1987). Function of the 30kD protein of tobacco mosaic virus: involvement in cell-to-cell movement and dispensability for replication. *EMBO Journal* **6**, 2557–2563.

MEYER, M., HEMMER, O., MAYO, M. A. & FRITSCH, C. (1986). The nucleotide sequence of tomato black ring virus RNA-2. *Journal of General Virology* **67**, 1257–1271.

MOSER, O., GAGEY, M.-J., GODEFROY-COLBURN, T., STUSSI-GARAUD, C., ELLWART-TSCHÜRTZ, M., NITSCHKO, H. & MUNDRY, K.-W. (1988). The fate of the transport protein of tobacco mosaic virus in systemic and hypersensitive tobacco hosts. *Journal of General Virology* **69**, 1367–1373.

MURTHY, M. R. N. (1983). Comparison of the nucleotide sequences of cucumber mosaic virus and brome mosaic virus. *Journal of Molecular Biology* **168**, 469–475.

NEEDLEMAN, S. B. & WUNSCH, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology* **48**, 443–453.

NITTA, N., MASUTA, C., KUWATA, S. & TAKANAMI, Y. (1988). Comparative studies on the nucleotide sequence of cucumber mosaic virus RNA3 between Y strain and Q strain. *Annals of the Phytopathological Society of Japan* **54**, 516–522.

OHNO, T., TAKAMATSU, N., MESHI, T., OKADA, Y., NISHIGUCHI, M. & KIHO, Y. (1983). Single amino acid substitution in 30K protein of tobacco mosaic virus defective in virus transport function. *Virology* **131**, 255–258.

RAVELONANDRO, M., PINCK, M. & PINCK, L. (1984). Complete nucleotide sequence of RNA3 from alfalfa mosaic virus, strain S. *Biochimie* **66**, 395–402.

RICHINS, R. D., SCHOLTHOF, H. B. & SHEPHERD, R. J. (1987). Sequence of figwort mosaic virus DNA (caulimovirus group). *Nucleic Acids Research* **15**, 8451–8466.

ROSSMANN, M. G. & JOHNSON, J. E. (1989). Icosahedral RNA virus structure. *Annual Review of Biochemistry* **58**, 533–573.

SAITO, T., IMAI, Y., MESHI, T. & OKADA, Y. (1988). Interviral homologies of the 30K proteins of tobamoviruses. *Virology* **167**, 653–656.

SAVITHRI, H. S. & MURTHY, M. R. N. (1983). Evolutionary relationship of alfalfa mosaic virus with cucumber mosaic virus and brome mosaic virus. *Journal of Biosciences* **5**, 183–187.

STEINHAUER, D. A. & HOLLAND, J. J. (1987). Rapid evolution of RNA viruses. *Annual Review of Microbiology* **41**, 409–433.

TAKAMATSU, N., OHNO, T., MESHI, T. & OKADA, Y. (1983). Molecular cloning and nucleotide sequence of the 30K and the coat protein cistron of TMV (tomato strain) genome. *Nucleic Acids Research* **11**, 3767–3778.

TALIANSKY, M. E., MALYSHENKO, S. I., PSHENNIKOVA, E. S. & ATABEKOV, J. G. (1982). Plant virus-specific transport function II. A factor controlling virus host range. *Virology* **122**, 327–331.

TOMENIUS, K., CLAPHAM, D. & MESHI, T. (1987). Localization by immunogold cytochemistry of the virus-coded 30K protein in plasmodesmata of leaves infected with tobacco mosaic virus. *Virology* **160**, 363–371.

WATANABE, Y., OOSHIKA, I., MESHI, T. & OKADA, Y. (1986). Subcellular localization of the 30K protein in TMV-inoculated tobacco protoplasts. *Virology* **152**, 414–420.

ZIMMERN, D. (1983). Homologous proteins encoded by yeast mitochondrial introns and by a group of RNA viruses from plants. *Journal of Molecular Biology* **171**, 345–352.

ZIMMERN, D. & HUNTER, T. (1983). Point mutation in the 30K open reading frame of TMV implicated in temperature-sensitive assembly and local lesion spreading of mutant NI2519. *EMBO Journal* **2**, 1893–1900.

(Received 27 September 1989; Accepted 24 January 1990)