The ‘30K’ superfamily of viral movement proteins

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

Relationships among the amino acid sequences of viral movement proteins related to the 30 kDa (‘30K’) movement protein of tobacco mosaic virus – the 30K superfamily – were explored. Sequences were grouped into 18 families. A comparison of secondary structure predictions for each family revealed a common predicted core structure flanked by variable N- and C-terminal domains. The core consisted of a series of β-elements flanked by an α-helix on each end. Consensus sequences for each of the families were generated and aligned with one another. From this alignment an overall secondary structure prediction was generated and a consensus sequence that can recognize each family in database searches was obtained. The analysis led to criteria that were used to evaluate other virus-encoded proteins for possible membership of the 30K superfamily. A rhabdoviral and a tenuiviral protein were identified as 30K superfamily members, as were plant-encoded phloem proteins. Parsimony analysis grouped tubule-forming movement proteins separate from others. Establishment of the alignment of residues of diverse families facilitates comparison of mutagenesis experiments done on different movement proteins and should serve as a guide for further such experiments.

Introduction

Plasmodesmal transport of macromolecules (Lucas et al., 1993) occurs for selected molecules, such as phloem proteins, synthesized in companion cells and deposited in sieve elements (Ding, 1998), and viral genomes, which spread throughout a plant after establishing infection in a single cell (Carrington et al., 1996; Nelson & van Bel, 1998). Most plant viral genomes contain a movement protein (MP) gene or genes required for spread of the infection (Leisner, 1999). Recently, a phloem protein cross-reacting with an antiserum against a viral MP was identified, shown to move from cell to cell and to carry RNA to the next cell (Xoconostle-Cazares et al., 1999).

Recognized MPs fall into four superfamilies: the products of the triple gene block of potexviruses and related viruses; the tymoviral MPs; a series of small polypeptides, less than 10 kDa, encoded by carmo-like viruses and some geminiviruses; and the ‘30K’ superfamily, related to the 30 kDa tobacco mosaic virus (TMV) MP. A variety of activities has been demonstrated for 30K superfamily MPs (Leisner, 1999), including abilities to bind nucleic acids, to increase the size-exclusion limit of plasmodesmata, to localize to and accumulate in plasmodesmata, to move to neighbouring cells on microinjection, to facilitate movement of RNA to neighbouring cells on microinjection, to form tubular structures and to interact with cytoskeletal elements. Which of these activities is relevant to movement of infection is unclear. At least two mechanisms have been proposed for the 30K superfamily MPs. In one, typified by the TMV MP, the MP expands the plasmodesmal connections between cells, allowing the viral genome, bound to protein, to pass into the neighbouring cell. In the second, typified by comoviruses, MP-containing tubules are established between cells and serve as conduits for particles.

Despite their common function, 30K superfamily members have but few conserved motifs in their amino acid sequences (Melcher, 1990, 1993; Koonin et al., 1991; Mushegian & Koonin, 1993). An LXDXG motif was the only conserved feature noted in a study of a limited number of sequences (Melcher, 1990). A hydrophobic sequence resides just N-terminal of the LXDXG motif (Koonin et al., 1991). The poor sequence conservation suggests that the 30K superfamily members may have a common three-dimensional structure obscured by divergence in sequence. To discover secondary structure elements contributing to three-dimensional structures, 30K MP sequences were aligned and the alignments used to predict secondary structures. Similarities to plasmodesmata proteins and to heat shock proteins of the hsp70 class were evaluated.
Methods

Groupings of similar MP amino acid sequences were designated ‘families’, in a non-taxonomic sense. For each virus MP family, available sequences were identified by BLAST search (Altschul et al., 1990) of a non-redundant compilation of protein sequences, using the MP of one family member as query. When sequences of two or three isolates of the same virus were available, one sequence was arbitrarily chosen as representative. When sequences of more than three isolates were available, a consensus was first generated, using Profilemake (GCC) (Gribskov et al., 1989), and the consensus was used in the alignment. Sequences within each family were preliminarily aligned using Pileup (GCC). The Pileup alignment was inspected visually for unlikely gaps, which were then removed or consolidated. The positioning of gaps was refined by gap translation (Melcher, 1993) using the SCM2 structural correlation matrix (Niefind & Schomburg, 1991).

The multiple sequence alignments (or single sequences in cases of one-member families) were used as input to PHD protein secondary structure prediction (Rost et al., 1994). The PHD subset prediction for three-state secondary structure (expected average accuracy greater than 82%) was used. A consensus was generated via Profilemake and modified at positions occupied by blanks in 50% or more of the sequences by replacing consensus residues with blanks.

A Pileup preliminary alignment of the resulting 18 family sequences was subjected to a phylogenetic analysis by parsimony using PROTPARS in the Phylip package (Felsenstein, 1989). For this analysis, the less reliably alignable N- and C-terminal one-thirds of the sequences were ignored. The pattern of relatedness revealed by the most parsimonious tree was used to refine the alignment using a progressive alignment strategy (Feng & Doolittle, 1987). This consisted of aligning first the most closely related sequences and then adding those pairs to other related pairs or sequences, progressively building pairwise to alignment of the two most distant groups. Alignment used the gap translation strategy and the SCM2 matrix. The cycle of protein parsimony followed by realignment was repeated until no further changes in the tree resulted (Berti & Storer, 1995). The reliability of relationships inferred by parsimony was tested by analysis of 100 bootstrap replicates using SEQBOOT, PROTPARS and CONSENSE (Phylip package). Branches not supported in over 80% of the bootstrap trees were collapsed. PROTDIST (Phylip package) was used to calculate distances.

Before construction of plots of similarity, residue positions occupied in three or fewer sequences were removed. To differentiate peaks in similarity due to alignment from those due to stretches of similar amino acids, similarity plots were also calculated with one set offset by one residue position relative to the other. A 10 position window size at each position and the SCM2 scoring matrix (Niefind & Schomburg, 1991) were used. SCM2 was also used in the determination of significance scores for alignment of pairs of multiple sequence alignments (or single sequences) by 30 randomizations of the order of residues as previously described (Melcher, 1993).

Profilemake and BLAST generated a single 30K superfamly consensus sequence. Initially, each family sequence had equal weight in consensus construction. Weights for families identified with the lowest expect scores in BLAST 2.0 were reduced, while those for families not identified by BLAST were increased. Consensus sequence generation, BLAST search and weighting adjustment were repeated until the BLAST search identified at least one member of each of the input families using default parameters. Relationships among families were also explored using PSI-BLAST (Altschul et al., 1997) seeded with family consensus sequences. All alignments with sources of the sequences used are available at http://opbs.okstate.edu/virevol/index.html.

Results

Group alignment and secondary structure prediction

Previously, 20 groups of MPs were identified as 30K superfamly members (Koonin et al., 1991; Melcher, 1990, 1993; Mushegian & Koonin, 1993; Mushegian, 1994): the MPs of bromo-, cucumo-, alfamo-, ilar-, diantho-, tombus-, tobamo-, tobra-, umbra-, begomo-, como-, nepo-, tospoS-, sequi-, capillo-, tricho-, furo-, idaeo-, caulimo- and badnaviruses. For the present analysis, sequences of the alfamo- and ilarviral MPs were grouped as one family because of the close similarity of those sequences and the few alfamoviral sequences available. For similar reasons, tricho- and capilloviral MPs were grouped as one family. Of the furoviruses, only the soilborne wheat mosaic virus MP was used, because other sequenced furoviral RNAs encode MPs of the triple gene block superfamily. The nepoviral MPs were divided into two families (A and B) because of only distant similarities between members of the two (Mushegian, 1994).

Members of the 30K superfamly contain a ‘common core’ domain consisting of two α-helices separated by a series of β-elements. Within each of the 19 30K families, the sequences were aligned with one another. The family alignments (or single sequences for furo-, idaeo- and sequiviruses) were used for secondary structure prediction. Predictions (Fig. 1) for all but one of the families contained, consistent with earlier predictions (Melcher, 1990, 1993), a series of β-elements. The N-terminal region of the sequiivaral polyprotein, thought to contain movement function (Mushegian & Koonin, 1993), was predicted to be predominantly helical, and may thus not be a superfamly member. The smallest of the 30K MPs, those of the tombusviruses, should have one α-helical element about 15 residues from the C terminus, preceded by a series of more than five β-elements. An equivalent series of β-elements in the 17 other predictions was, in most cases, flanked on the C-terminal side by an α-helical segment. In all but the bromoviral MPs an α-helical segment was also predicted to flank the β-structure rich region on the N-terminal side.

The common core was flanked by additional sequences on N-terminal and C-terminal sides. The N-terminal flanking region was of variable length. The region was longer for those MPs known to form virion-carrying tubules and for the idaeoviral MP. These had numerous predicted α-helical segments. Loop structure was also commonly predicted for this region. The predicted structures on the C-terminal sides were predominantly loops except for the probable extensively helical tospoS- and furoviral MP C termini. The C-terminal region also varied in length among the families, with the tombusviral MPs having only 15 residues in this region.

Multiple sequence alignment

To examine the relationship of the similarity in predicted secondary structure to sequence similarity, a multiple sequence
alignment of 30K families was performed. To simplify alignment, consensus sequences (Gribskov et al., 1989) were first generated for each family with multiple sequences. The sequiviral sequence was not included in the overall alignment since its predicted secondary structure excluded it from the 30K superfamily. An alignment of the resulting 18 sequence families was generated by an iterative progressive alignment–parsimony analysis strategy (Berti & Storer, 1995). The alignment and parsimony were repeated until realignment did not alter the most parsimonious tree obtained (22 cycles).

The alignment of family consensus sequences was also used to produce an overall secondary structure prediction (Fig. 1, lowest line). The central region of the overall prediction consisted of seven β-elements and four α-helical segments. The outer α-helical segments, α-A and α-D, corresponded to those noted in the predictions for individual families. α-B corresponded only to a helical segment predicted in the nepoviral A, badnaviral, comoviral and tobaviral families. α-C was not predicted for any of the constituent families individually. The β-elements β-1, β-2, β-5 and β-6 corresponded to β-elements in most of the family predictions. Variable structures were found in the family predictions for the region between β-2 and β-5 and around β-7. The C-terminal region was predicted to be a predominantly random coil. The N-terminal extension specific to tubule-forming 30K superfamily members was predicted to be predominantly helical.

Examination of the aligned sequences for conserved sequence motifs (Fig. 2) failed to reveal any absolutely conserved residues. Five regions of moderate sequence conservation were noted. The first, characterized by a PLX(P/D) motif, preceded α-A. The second spanned β-1 and β-2 and contained the LXD motif at the C terminus of β-2. The β-3–β-4 region contained positions rich in aromatic and proline residues. The fourth region contained the almost completely conserved Gly residue in the N terminus of β-6. The last region occurred unexpectedly near the C termini of 14 of the families and consisted of variations of a SIS tripeptide.

### 30K superfamily consensus and similarity plots

An overall consensus sequence generated from the alignment of the family consensuses (Fig. 3) was used in a BLAST search of protein databases and recognized at least one member of each family that contributed to the consensus. This recognition suggested that a satisfactory alignment of related sequences had been obtained. As a further test of the overall accuracy of the alignment, each family multiple sequence alignment was tested for significance relative to the superfamily alignment. For this assessment, the consensus for the family being evaluated was removed from the superfamily alignment. For most families, the alignment scores were between 7.0 and 13.0 standard deviations (SD) above the mean for scrambled comparisons (Fig. 1). Nepoviral and badnaviral MPs scored lower. A self-comparison of the 18 family sequences (positive control) yielded 18 SD, while the average of 30 offset self-comparisons (negative control) gave 1.0 SD.

The degree of sequence similarity varied along the alignment (Fig. 4). The β-1 and β-2 elements were the most strongly similar. The loop between them (L12) was less conserved. The degree of similarity of L12 was close to that of the loop preceding α-A (LA), L34 and L56. The α-D segment and the region containing the SIS motif also yielded similarity peaks. Similarity in the α-D region was due, in part, to contiguous similar residues, as reflected in a peak in the control similarity calculation using a dataset pair offset by one residue. The proximity of the similarity profiles for aligned and offset pairs at the N and C termini suggests that the alignment in these regions was not reliable.

Iterative searching of databases using PSI-BLAST (Altschul et al., 1997) allows exploration of relationships among 30K
superfamily members alternate to the consensus approach. Searches were seeded with the consensus sequences for each of the acknowledged 30K superfamily families and reiterated until no further MP sequences were retrieved. The search did not find previously unrecognized candidates for 30K superfamily membership. The consensuses for five families did not identify other groups even though reciprocal searches identified members of these five families. Neither of the two nepoviral sequences identified any MPs outside the family nor were nepoviral MPs identified in searches with other sequences. In contrast, the tobraviral consensus sequence identified MPs belonging to 12 families, including bromo- and furoviruses, and umbraviral MPs were identified by searches initiated with eight family consensus sequences including those of cucumo-, furo- and dianthoviruses. These results strongly suggest sequence similarity between MPs of the tripartite viruses and MPs such as those of umbra- and tobraviruses that are related to tobamoviral MPs.

**Candidates for 30K superfamily membership**

MPs of rhabdo-, waika-, tenui- and closteroviruses have not been identified or have not been assigned to a MP superfamily. MPs of non-acknowledged 30K superfamily members were not identified in the database search with the overall consensus sequence or the PSI-BLAST search. Nevertheless, secondary structures predicted for candidate MPs were examined to identify those with β-elements flanked by α-helices. For Sonchus yellow net rhabdovirus, protein sc4 had a series of β-elements flanked by α-helical segments at the separation expected for 30K superfamily members. The sc4 protein sequence was thus aligned with previously recognized 30K
superfamily members (Fig. 3). The alignment had a significance score of 80 SD. A similarity plot (Fig. 5A) had peaks above control corresponding to LA, β-1, β-2, L34 and L56, features also found in the plot of the previously recognized 30K superfamily members (Fig. 4). A small peak was in the SIS region. An additional region of similarity was between β-2 and L34. Thus, sc4 probably functions as a rhabdoviral MP, a conclusion supported by identification of the sc4 gene as the only gene not present in animal-infecting rhabdoviruses (Scholthof et al., 1994).

Tenuiviral genomes are multipartite with ambisense RNAs for which no MP genes are known (Toriyama et al., 1998). Of the four ORFs (2V, 3V, 4V and 4C) with no probable function assigned, 4C had a predicted secondary structure that corre-
**Table 1.** PAM 250 distances of closest interfamily neighbours of 30K superfamily movement proteins

Distances were determined using PROTDIST with the PAM250 scoring matrix.

<table>
<thead>
<tr>
<th>MP family</th>
<th>Closest distance</th>
<th>Neighbour family</th>
</tr>
</thead>
<tbody>
<tr>
<td>Furo-</td>
<td>2.89</td>
<td>Diantho-</td>
</tr>
<tr>
<td>Diantho-</td>
<td>2.89</td>
<td>Furo-</td>
</tr>
<tr>
<td>Alfamo-</td>
<td>3.48</td>
<td>Diantho-</td>
</tr>
<tr>
<td>Bromo-</td>
<td>1.34</td>
<td>Cucumo-</td>
</tr>
<tr>
<td>Cucumo-</td>
<td>1.34</td>
<td>Bromo-</td>
</tr>
<tr>
<td>Tenui-</td>
<td>5.61</td>
<td>Umbraco</td>
</tr>
<tr>
<td>Waika-</td>
<td>9.38</td>
<td>Capillo-/Tricho-</td>
</tr>
<tr>
<td>Clostero-</td>
<td>7.12</td>
<td>Gemini-</td>
</tr>
<tr>
<td>Rhabdo-</td>
<td>5.03</td>
<td>Idae-</td>
</tr>
<tr>
<td>Tobra-</td>
<td>2.66</td>
<td>Tobamo-</td>
</tr>
<tr>
<td>Tobamo-</td>
<td>2.66</td>
<td>Tobaro-</td>
</tr>
<tr>
<td>Gemini-</td>
<td>4.05</td>
<td>Tombus-</td>
</tr>
<tr>
<td>Tombus-</td>
<td>3.19</td>
<td>Tobra-</td>
</tr>
<tr>
<td>Umbraco-</td>
<td>4.16</td>
<td>Furo-</td>
</tr>
<tr>
<td>Nepo- (B)</td>
<td>8.25</td>
<td>Gemini-</td>
</tr>
<tr>
<td>Tospo-</td>
<td>4.55</td>
<td>Umbraco</td>
</tr>
<tr>
<td>Idae-</td>
<td>4.55</td>
<td>Diantho-</td>
</tr>
<tr>
<td>Caulimo-</td>
<td>4.06</td>
<td>Nepo- (A)</td>
</tr>
<tr>
<td>Badna-</td>
<td>4.08</td>
<td>Caulimo-</td>
</tr>
<tr>
<td>Nepo- (A)</td>
<td>3.91</td>
<td>Como-</td>
</tr>
<tr>
<td>Como-</td>
<td>3.91</td>
<td>Nepo- (A)</td>
</tr>
<tr>
<td>Capillo-/Tricho-</td>
<td>4.11</td>
<td>Caulimo-</td>
</tr>
<tr>
<td>Phloem*</td>
<td>4.77</td>
<td>Bromo-</td>
</tr>
</tbody>
</table>

Mean ± SD... 3.67 ± 1.49

* Distances not included in calculation of the mean.

sponded well throughout its length with that of the 30K superfamily. The amino acid sequence was added to the alignment of previously recognized superfamily members (Fig. 3), with a significance score of 7.9. A similarity plot (Fig. 5 C) had a major peak, relative to control, corresponding to L12. Additional areas of above background similarity were near the N terminus to α-A, around β-5 and loops LB3 and L7D. Thus, the tenuiviral 4C protein probably belongs to the 30K MP superfamily.

An MP function was recently identified in a closterovirus-encoded protein (Agranovsky et al., 1998) previously known as a homologue of hsp70 proteins. These proteins include the *Escherichia coli* DnaK protein for which X-ray crystallographic information is available (Zhu et al., 1996). Secondary structure prediction for the closteroviral hsp70 homologues revealed that the C-terminal half of the proteins had an arrangement of β-elements and α-helical segments similar to those predicted for 30K superfamily members. However, the length of the similarly arranged section was longer for the closteroviral hsp70 homologues. To achieve alignment with the 30K superfamily members (Fig. 3), one pair of β-elements had to be deleted from the hsp70 homologue. The closteroviral hsp70 homologues aligned with a score of 4.6 SD. The DnaK–hsp70 homologue alignment (Agranovsky et al., 1998) was used to align DnaK to the 30K superfamily set, yielding a score of only 2.4 SD. The N terminus of the DnaK protein-binding domain aligned to the MP β-2 element. Prominent peaks in the similarity plots of the hsp70 homologues against the 30K superfamily (Fig. 5 D) were at L12, β-1 and L56. Similar peaks, but less pronounced, occurred in the similarity plot of DnaK compared to the 30K superfamily (data not shown).

Antigenic cross-reaction suggests that some phloem proteins may be 30K superfamily members (Xoconostle-Cazares et al., 1999). The secondary structure predicted from a multiple sequence alignment of the cross-reactive CmPP16 and related sequences (Xoconostle-Cazares et al., 1999) was predominantly β-structure with one x-helical segment near the C terminus. An alignment of the phloem sequences with the
30K superfamily (Fig. 3) resulted in a score of 6.5 SD. The alignment began at the β-2 segment, the segment that marked the beginning of the DnaK peptide-binding domain (Zhu et al., 1996) alignment. The similarity plot (Fig. 5B) had a prominent peak in the β-2 element and an additional peak at β-5. Knotted1 and homologues (P24345, JQ2379, Q43484, AAB81079.1, CAA9511.1, Q41330, P46639, BA76904.1, Q41853, AAC32818.1, AAD00692, PP1 (U66277) and PP2 (IQ17311) phloem fibril proteins also traffic between cells. Their predicted structures easily distinguished them from 30K MPs.

Distance measures also tested candidacy for 30K superfamily membership. Family consensuses were generated for the waika-, tenui- and closteroviral proteins and the phloem proteins. Pairwise distances were calculated between all pairs of the 23 family sequences using the evolutionarily derived PAM250 scoring matrix. For each of the 18 families of acknowledged superfamily members, the smallest distance to another family member was identified and the values averaged (Table 1). The mean smallest distance was 3.7 ± 1.5. The distance between the nepoviral B family and its closest 30K superfamily member was more than 2 SDs above this value, thus casting doubt on its membership in the 30K superfamily. The phloem family was within 2 SDs of the mean smallest distance for six families (bromo-, cucumo-, toboamo-, gemini-, umbr- and tospoviruses) with bromoviral MPs being the closest at 4.8. The rhabdoviral sc4 was within 2 SDs of the smallest mean distance for four families (tombus-, tosp-, idaeo- and badnaviruses) with the idaeoviral MP being the closest at 5.0. Only one acknowledged consensus MP, that of umbraviruses, was within the 2 SD range for tenuiviral 4C. The waika- and closteroviral families did not have a closely related family in the 30K superfamily.

### Parsimony-derived relationships among 30K-superfamily members

Fig. 6 shows the relationships deduced by parsimony among the acknowledged 30K superfamily members, the putative tenuiviral and rhabdoviral MPs and the phloem proteins. Relationships among the acknowledged members alone were analysed similarly (data not shown) and differed from those shown only in the clustering of tombus- and begomoviral MPs. Fig. 6 suggests division of the 30K superfamily into three clusters, two of which radiate from common points without much subdefinition of relationships. The third, containing capillo-/tricho-, nepo-, caulimo-, tosp- and idaeo- MPs, had significant substructure. This cluster contained all the MPs thought to induce movement by synthesis of trans-wall tubules. Since it has been proposed that viruses be classified based on the sequences of their RNA or DNA polymerases (Koonin, 1991), the diagram also indicates the type of polymerase encoded by the viruses. No correlation between diagram-position and polymerase type was noted, except that all supergroup I viruses were on the tubule branch. Distance analysis (Felsenstein, 1989) (Table 1 and data not shown) suggests relationships among the proteins.
that differ from those derived by parsimony. For example, the closest MP to the begomoviral family is the idaeoviral protein, which itself is closest to the dianthoviral family. Yet, these families are scattered on the parsimony tree (Fig. 6).

Discussion

Amino acid sequences that are 20–25% identical are generally believed to lie in the twilight zone of relationships, where conclusions about their common ancestry are questionable. The relationships explored among 30K MPs, though generally of sequences with sequence identities in or below the twilight zone, nevertheless produced several important insights.

Previous reports identifying the 30K superfamily (Koonin et al., 1991; Melcher, 1990, 1993; Mushegian, 1994), have not encountered universal acceptance (Giesman-Cookmeyer et al., 1995; Mise et al., 1993). The work reported here supports the concept of a 30K superfamily as a meaningful grouping of proteins related in sequence and structure. Significance scores for membership of sequences in the family were generally well above the range of scores considered questionable. The member families had a similar arrangement of predicted secondary structure elements in the common core. A consensus generated from their alignment identified members of all constituent families in database searches. Though conserved motifs were few, a few residue positions were occupied by the same amino acid in most MPs. These include the LXDXG motif previously recognized (Melcher, 1993). A new relatively conserved motif resembling SIS was found near the C termini of most family members.

The criteria for membership of candidate sequences in the 30K superfamily of movement proteins thus include: a common core of predicted secondary structure elements consisting of a series of beta-elements between two alpha-helices; additional N-terminal and C-terminal sequences (not obligatory, such as in the tombusviral MPs); a distance from at least one previously recognized superfamiliy member that is within the range of distances between close relatives within the superfamily; and a significance score in alignment with acknowledged family members of 5.5 or greater. The presence of LXDXG and SIS motifs give additional support for membership, as does recognition in BLAST searches using the derived consensus sequence. However, failure to be recognized by the consensus sequence is not sufficient for exclusion from the family, since tenuiviral and rhabdoviral MPs and the phloem proteins were not recognized by such a search. These criteria allowed the exclusion of one sequence previously thought to belong, that of the sequiviral N-terminal domain. They also suggested that some nepoviruses might have MPs not related to the 30K superfamily and that walkaviral and closteroviral MPs are not 30K superfamily members.

It is intriguing that the currently known small set of phloem proteins related to CmPP16 had sequences and a predicted structure similar to those of the 30K superfamily proteins. Thus, the knowledge gained through the study of MPs can be used in assessing a possible role in trafficking by plant-encoded proteins. In addition, 30K superfamily membership for these phloem proteins means that tools developed for study of MP function can be applied to intercellular traffic of the plant proteins. The similarities suggest evolutionary homology, a possibility that may lead to hypotheses on the evolutionary origin of viral MPs and thus of plant viruses.

Discrepancies were encountered between positions of 30K superfamliy members on a tree derived by parsimony (Fig. 6) and distance (Table 1) analyses. The parsimony tree was derived from an alignment obtained by repeated optimization of relationships by parsimony. A different set of relationships could have been revealed had alignment been optimized using distances. Distances were not used because the distances of most family members from each other were large. The values suggested too many positions at which multiple substitutions occurred to accurately estimate evolutionary relatedness. When distances in bifurcating evolutionary trees are long, long branch attraction occurs (Lyons-Weiler & Hoelzer, 1997). Long branch attraction describes branch positions on inferred phylogenetic trees that do not reflect their ontogeny. The discrepancies encountered between the parsimony tree (Fig. 6) and distance calculations (Table 1) probably reflect long branch attraction. Perhaps in the course of evolution, MP genes of plant viruses explored most combinations of residues that produce optimally active MPs. These combinations represent fitness peaks in an MP sequence landscape.

Despite doubt that the parsimony tree (Fig. 6) accurately represents evolutionary relationships, it accurately reflects a subclassification of 30K superfaimly members based on hypothesized mechanism of movement. Three sub-superfamilies were distinguished, two as clusters and one as an extended branch. Those MPs (badna-, nepo-, tospo-, caulimoviral and comoviral MPs) for which tubule formation is thought to provide a conduit for virus particles in cell-to-cell movement (Cheng et al., 1998; Wieczorek & Sanfacon, 1993; Storms et al., 1995; Kasteel et al., 1996) lie on the extended branch of the tree. Conversely, MPs that are not thought to use such a mechanism were found in the clusters. The tubule-forming branch included idaeo-, capillo- and trichoviral MPs. It would therefore be interesting to determine whether these also form tubules important for movement.

A common sequence feature of the tubule-forming sub-superfamily was the extra length of the N-terminal domain. Thus this domain may be responsible for polymerization into tubules and other functions specifically associated with this mode of movement of viral infections. Indeed, deletions in the N-terminal, but not in the C-terminal, region of the cauliflower mosaic virus MP were recently reported to destroy its ability to form tubules in cultured insect cells (Thomas & Maule, 1999). However, MPs lacking the long N-terminal domain can form tubules in protoplasts (Zheng et al., 1997; Kasteel et al.,
1997), suggesting that the N-terminal domain is not required for polymerization. The C-terminal domain, predicted to not readily assume a definable structure, does not perform an essential function for several MPs. Both C-terminal truncation (Gafny et al., 1992; Berna et al., 1991) and fusion of the C terminus to reporter proteins such as green fluorescent protein (Heinlein et al., 1995; Itaya et al., 1997) fail to interfere noticeably with function. The likelihood that the common core domain contains a nucleic acid binding domain (Citovsky et al., 1992; Fujita et al., 1998; Giesman-Cookmeyer & Lommel, 1993; Schoumacher et al., 1994; Thomas & Maule, 1995) raises the possibility that the C-terminal domain may be a flexible tail which regulates access to the binding cavity of the common core domain. The variations on the SIS motif near the C terminus could represent sites for regulation by phosphorylation. Indeed, phosphorylation of tobamoviral and trichoviral MPs has been demonstrated (Watanabe et al., 1992; Citovsky et al., 1993; Sato et al., 1995).

Increasingly, protein domains are found to belong to a limited number of ‘fold’ groupings unified by similar three-dimensional arrangements of secondary structure elements. Thus, the profile of MP secondary structure elements (Fig. 1) may lead to fold recognition. Previously, it was suggested, based on more limited MP sequences and available folds, that the MPs may fold like aspartic proteases, in particular the lentiviral proteinases (Melcher, 1993). The current work suggested some similarity of the 30K superfamily members to the closteroviral MPs (Agranovsky et al., 1998), known homologues of DnaK, whose three-dimensional structure has been determined (Zhu et al., 1996). Further analysis could not firmly reject the suggestion, leaving open the possibility that 30K MPs have an hsp70 peptide-binding fold. A third possibility is an OB fold (Horvath et al., 1998), a fold characteristic of many proteins that bind single-stranded nucleic acids. It shares with the 30K MP prediction a similar array of β-elements flanked by one or two α-helices, an analogous placement of regions variable in secondary structure, and a high frequency of aromatic residues in selected loop regions (L34 and L56, Fig. 2). That the RNA-binding domains of several MPs map to the central domain is consistent with the region having an OB fold. On the other hand, that the phloem proteins are missing the first β-element argues against an OB fold for 30K MPs.

Regardless of the true fold assumed by the common core of the 30K MPs, the alignments and predicted structures presented here should aid the interpretation of molecular genetic studies in which mutations in the MP genes were created without knowledge of the structures mutated (Giesman-Cookmeyer & Lommel, 1993; Kahn et al., 1998). These studies have been performed with a variety of genes from different families, making comparison of the results from mutagenesis of one family impossible to compare with results from a different family. Further, the predicted structures enable better choices of residues as targets for future mutagenesis.

This work was supported by the Oklahoma Agricultural Experiment Station and is being published with the approval of its Director.

References


Received 14 June 1999; Accepted 24 September 1999