Phylogeny of capsid proteins of small icosahedral RNA plant viruses

Valerian V. Dolja\(^1\)* and Eugene V. Koonin\(^2\)

\(^1\)A. N. Belozersky Laboratory of Molecular Biology, Moscow State University, Moscow 119899 and \(^2\)Institute of Microbiology, U.S.S.R. Academy of Sciences, 117811 Moscow, U.S.S.R.

Statistically significant alignment was generated between the amino acid sequences of the (putative) shell (S) domains of the capsid proteins of small RNA plant viruses with icosahedral capsids in the tombusvirus, carmovirus, dianthovirus, sobemovirus and luteovirus groups. Inspection of the alignment showed good correspondence between the experimentally defined \(\beta\)-strands and \(\alpha\)-helices of the capsid proteins of tomato bushy stunt, southern bean mosaic and turnip crinkle viruses, allowing prediction of the secondary structure elements in proteins with unresolved tertiary structure. It is concluded that this set of viral capsid proteins forms a tight evolutionary cluster. Comparison of the alignment of the proteins of this family with the sequences of other capsid proteins of icosahedral RNA viruses revealed more distant similarities to the satellites of tobacco necrosis, panicum mosaic, tobacco mosaic and maize white line mosaic viruses, as well as to nepo- and comoviruses. The tentative phylogenetic tree derived from the capsid protein alignment separated into three main lineages: (I) carmo-, tombus- and dianthoviruses, (II) southern bean mosaic, tobacco necrosis and maize chlorotic mottle viruses, and (III) luteoviruses. Comparison of this tree topology with the tentative evolutionary schemes for the respective virus RNA-dependent RNA polymerases suggested that gene shuffling is the universal trend in the evolution of small RNA plant virus genomes.

Introduction

Capsid proteins (CPs) of icosahedral positive-strand RNA viruses are among the classical objects for fine structural studies by X-ray crystallography (for a recent review see Rossmann & Johnson, 1989). Typically, these proteins consist of four distinct domains: the N-terminal disordered positively charged R domain involved in the interaction with RNA, the connecting arm (a), the central domain forming the virion shell (S), and the C-terminal projecting (P) domain. Some of the viruses lack either R or P domains (Liljas, 1986). Fine structure studies of these proteins culminated in the elucidation of the eight \(\beta\)-strand anti-parallel twisted sheet conformation, the so-called jellyroll (Richardson, 1981). This topology is conserved in S domains of all solved icosahedral virus CPs, except the recently reported structure of MS2 bacteriophage protein (Valegard et al., 1990). In addition, the jellyroll conformation was found in certain cellular ligand-binding proteins, e.g. concanavalin A, and it was speculated that all such domains might have a common evolutionary ancestry (Argos et al., 1980). In contrast to this remarkable spatial structure conservation, the conservation of the CP amino acid sequences is only limited, and it is generally believed that comparative sequence analysis is not of much use for understanding the structural and evolutionary relationships among these proteins (e.g. Rossmann & Johnson, 1989). On the other hand, it is impossible to dismiss convergence as the basis underlying the observed tertiary structure similarity, as very similar protein topologies can be secured by quite different amino acid sequences (see Lau & Dill, 1990). The neutralistic theory of molecular evolution (Kimura, 1983) suggests that, for inferring reliable phylogenies, it is essential to produce a statistically significant sequence alignment of the proteins under comparison.

Carrington et al. (1987) reported reasonable sequence alignments between capsid proteins of turnip crinkle virus (TCV), carnation mottle virus (CarMV), tomato bushy stunt virus (TBSV) and southern bean mosaic virus (SBMV). More recently, relatively close sequence similarity was observed between the CPs of tobacco necrosis virus (TNV) and SBMV (Meulewaeter et al., 1990), and moderate similarity was found to exist between the CPs of red clover necrotic mosaic dianthovirus (RCNMV) and carmoviruses (Xiong & Lommel, 1989).

We sought to assess more thoroughly the limits of sequence conservation among the capsid proteins of positive-strand RNA viruses with icosahedral capsids. It was demonstrated that CP sequences from a wide range

0001-0087 © 1991 SGM
of positive-strand RNA plant virus groups with small (4 to 5-8 kb) genomes could be readily aligned with statistically highly significant scores. The tentative evolutionary schemes emerging from this alignment include gene shuffling as the main evolutionary mechanism in this virus class.

**Methods**


**Computer-assisted sequence analysis.** Initial comparisons were by the program DOTHELIX which is a version of a diagonal comparison plot generating a full local similarity map and allowing precise determination of the lengths of similar segments (Brodsky et al., 1991). Stepwise multiple sequence alignment was by the program OPTAL as previously described (Gorbalenya et al., 1989). Briefly, the program implements the algorithm of Sankoff (1972) and includes statistical evaluation of alignments by calculation of the adjusted alignment score, AS = (S' - S)/σ. Here, S' is the observed alignment score, S is the mean score obtained upon 25 random simulations of the alignment procedure, and σ is the standard deviation of the score. Thus AS is measured as the number of S.D. above the random expectation. The log odds matrix of Dayhoff et al. (1983) was used for the calculation of sequence comparison scores. Cluster dendrograms were generated by the program TREE implementing the unweighted pair group maximum averages (UPGMA) clustering algorithm (Sneath & Sokal, 1973) from distance matrices calculated according to Feng et al. (1985). Parsimony trees were generated by the program PROTPARS program of the PHYLIP package, using the aligned sequences as the input data (Felsenstein, 1989). Rate-independent trees were generated also by the program TF implementing the maximal topological similarity principle allowing construction of a tree in which the set of nearest neighbour species quartets is best compatible with that derived from the distance matrix (Chumakov & Yushmanov, 1988). Secondary structure predictions were by the program PROTEIN2 implementing the modified algorithm of Garnier (Biou et al., 1988). Programs DOTHELIX, SITE, TREE, TF and PROTEIN2 are modules of the GENEBEE program package for biopolymer sequence analysis (Brodsky et al., 1991).

**Results and Discussion**

**Generation of the sequence alignments**

For analysis of the viral capsid protein sequences, our strategy was first to delimit regions of significant similarity, and then to perform detailed multiple alignments of selected subsequences. Preliminary comparisons revealed three groups of small icosahedral RNA plant virus capsid proteins. Group I encompassed the CPs of carmo-, tombus- and dianthoviruses. We first aligned the CPs of tombusviruses and MNSV; the CP of MNSV was reported earlier to be more closely related to those of tombusviruses than to those of the other carmoviruses (Riviere et al., 1989). We also aligned those of CarMV and TCV carmoviruses. The two alignments were then combined over their entire length, yielding an alignment with AS above 22 S.D. Inspection of this alignment showed that the sequence conservation was restricted mainly to the S domains delineated experimentally in TBSV and TCV. Only a few conserved positions were found in the R domains, arms and the P domains (not shown). Thus only the (putative) S domain sequences were used for further comparisons. The dianthovirus RCNMV CP sequence was readily joined to the alignment of the tombusvirus and carmovirus CPs, yielding an AS of over 20 S.D. Thus this protein is related not only to the CPs of carmoviruses, as originally noted (Xiong & Lommel, 1989), but also to those of tombusviruses.

Group II brought together the CPs of SBMV, TNV and MCMV. For the first pair of proteins, a close relationship has been well documented (Meulewaeter et al., 1990). On the other hand, it was claimed that the CP of MCMV is unrelated to those of any other viruses (Nutter et al., 1989). We, however, succeeded in aligning this sequence with those of SBMV and TNV, with a significant, though moderate, AS value (about 8 S.D.).

Group III consisted of the CPs of luteoviruses, which are closely related to each other (Mayo et al., 1989; Van der Wilk et al., 1989). Even for the alignment of the most diverged sequence of BYDV CP with those of BWYV and PLRV, a striking AS value of over 25 S.D. was observed.

Preliminary comparisons by the DOTHELIX program showed that the proteins of these three groups are related to each other. The sequences of the first two groups were aligned with the fully convincing AS value of over 14 S.D., despite a long insert in the proteins of SBMV, TNV and MCMV. The luteovirus sequences were added to the alignment with an AS of about 13 S.D.

The resulting alignment of 14 virus CPs is shown in Fig. 1. The alignment includes only three strictly conserved residues (two Gly and one Asp), and in addition nine positions occupied by similar residues in all (or all but one) sequences. Thus the lack of specific residue conservation contrasted with the high statistical significance of the alignment, suggesting that it is the general structural organization that is conserved in the CPs. In agreement with this conjecture, in 16 positions of the alignment the hydrophobic residue character was found to be conserved (Fig. 1). Recently we performed an analogous analysis for the CPs of plant viruses with rod-shaped and filamentous capsids and found a very similar type of sequence conservation (Dolja et al., 1991).
### Small RNA plant virus capsid proteins

<table>
<thead>
<tr>
<th>Virus</th>
<th>AMCV</th>
<th>CyRSV</th>
<th>CNV</th>
<th>MNSV</th>
<th>CarMV</th>
<th>TCV</th>
<th>RCNMV</th>
<th>SBMV</th>
<th>TNV</th>
<th>MCMV</th>
<th>BWYV</th>
<th>PLRV</th>
<th>BYDV</th>
</tr>
</thead>
<tbody>
<tr>
<td>TBSV 101</td>
<td>SVTYT-HREYLSQVNNSVTG----FQVNGVGIVNLQPLNGLTLFWSVPAIASNFDQYTFNSVLHLYVL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AMCV 100</td>
<td>SVTGYHREYITVNNSSSG----FQVNGVGIVNLQPLNGLTLFWSVPAIASNFDQYTFNSVLHLYVL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CyRSV 98</td>
<td>SVRIT-HREYLSQVNNSSTG----FQVNGVGIVNLQPLNGLTLFWSVPAIASNFDQYTFNSVLHLYVL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CNV 92</td>
<td>SRVIT-HREYLSQVNNSSTG----FQVNGVGIVNLQPLNGLTLFWSVPAIASNFDQYTFNSVLHLYVL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MNSV 94</td>
<td>SVKIV-HREYLSQVNNSSTG----FQVNGVGIVNLQPLNGLTLFWSVPAIASNFDQYTFNSVLHLYVL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CarMV 81</td>
<td>SMTMS-KTEELTKTVGTGTVIPSFE------AVVSPNVRVATFQLGSSALTFNKYRITALVYKSPA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TCV 80</td>
<td>GITRSHGKTVNMTKDPMTG------TVLLSPERTNQLIKAQAEYKRTLSLRFVSPM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RCNMV 46</td>
<td>ARGSL-HQVMSVGSVSQVRNN---GKSNNRFLNPNSALPFTLAEANMYMRKLKLTVLYVL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| SBMV 64 | DVTILSHE-LSTELA-VIT-TFV------TSEELVMFVPTL-WLRVQAQSWSKYAWVAVITYLVP | | | | | | | | | | | | |
| TNV 85 | DSTRV يتمELNQL-AI-GFAN------TT--NSALIAAMFWLATTLYKVRSLCE11YIKP | | | | | | | | | | | | |
| MCMV 51 | E1-LVAYTADQTLKEDPTV-CAG1-----DPAGTMYRIGAAGKLAQEGMNTQWNTSVRLWTEIF | | | | | | | | | | | | |
| BWYV 57 | TRGAGSSETTVFSDKNLAG-----SSSG---AITFGPSLSDCPFASNGMLKAVHEYKISMVILEFVSE | | | | | | | | | | | | |
| PLRV 62 | PRGGRSSSETTVFSDKNLAVG--NTGQ-----SFTGFPSSDCDFKDI[KAYHEKITSILQVFVSE | | | | | | | | | | | | |
| BYDV 54 | FRPTGTEVFVSVDNLKQ--NSSG---AIKFGPSLSQCPALSQGLKISYHRKITSIRVEFKSH | | | | | | | | | | | | |

Fig. 1. Amino acid sequence alignment of the (putative) S domains of small spherical positive-strand RNA plant viruses. The sequences constituting the three groups described in the text are separated by horizontal lines. The lengths of N- and C-terminal extensions (the latter forming the P domains in proteins of group I) are indicated. Double asterisks: residues strictly conserved in 14 virus CP sequences; single asterisks: identical or similar residues in 13 or 14 sequences (the residue grouping was G, A; D, E, N, Q, S, T; K, R; I, L, V, M; F, Y, W); plus signs: hydrophobic residues (I, L, V, M, F, Y, W, C, A) in 14 sequences. The β-strands (bB to I) and α-helices (aA to E) determined from the crystal structure of TBSV are designated according to Rossmann et al. (1983). The alignment generated by the program OPTAL was slightly modified by hand to obtain a better fit to the experimentally determined structures.
Correspondence between the alignments and the crystal structures of viral capsids

The boundaries of the conserved regions coincided almost precisely with the experimentally determined boundaries of the S domains of the CPs of TBSV, TCV and SBMV. Thus it seemed most likely that in all other virus proteins the homologous sequences also comprised the S domains. These domains are rather uniform in length (about 160 residues each), except for the approximately 30 residue inserts in SBMV, TNV and MCMV. The β-strands and α-helices defined by X-ray diffraction analysis of the CPs of TCV, TBSV and SBMV corresponded to patches of high conservation in the alignment (Fig. 1), suggesting that the jellyroll structure is well conserved throughout the entire family. The best conserved secondary structure elements were βD, βE and βI. On the other hand, strands C, F and G showed only poor conservation among the three groups of proteins delineated above. The two invariant Gly residues are located near the N termini of βE and βI, respectively, probably serving to maintain structurally important β-turns. The clusters of partially conserved negatively charged residues in the CPs of TBSV, TCV and SBMV have been shown experimentally to constitute Ca²⁺-binding sites (see Carrington et al., 1987). The present alignment showed that conservation of an Asp residue in one such site noted by Carrington et al. (1987) is maintained in the whole family. Moreover, all CPs, except that of MCMV, retain at least two Asp (Glu) residues, in accord with the Ca²⁺-binding hypothesis.

Search for sequence similarity in other capsid proteins

Searches for possible counterparts to the motifs conserved in our alignment identified some similar segments also in the capsid proteins of four virus satellites, those of TNV, panicum mosaic virus, tobacco mosaic virus and maize white line mosaic virus, and of nepo- and comoviruses (not shown). However, detailed comparisons failed to produce alignments of sufficient significance to justify inclusion of these proteins in the family described above. On the other hand, we were definitely unable to detect significant similarity between the group of proteins described above and the other jellyroll capsid proteins of RNA viruses, namely those of noda- and picornaviruses. The CPs of icosahedral single-strand RNA bacteriophages, for which X-ray data have recently been reported (Valeegaard et al., 1990), lack the jellyroll structure. The structures of the icosahedral capsids of tymoviruses and tricornaviruses have not been resolved, but there is evidence that their physicochemical properties are substantially different from those of the viruses with purported jellyroll domains (Hirth & Givord, 1988; Francki, 1985). Not surprisingly, the respective capsid proteins could not be included in the alignment shown in Fig. 1, although those of tymo- and tricornaviruses showed their own, specific patterns of amino acid residue conservation (unpublished observations).

Phylogeny of capsid proteins

The proteins aligned here constitute a monophyletic family, as suggested by the high statistical significance of the alignments. The alignment of 14 virus CP sequences shown in Fig. 1 was used to generate a tentative phylogenetic tree. The tree shown in Fig. 2 was derived by the clustering procedure, but application of the protein parsimony and maximum topological similarity methods yielded identical tree topologies. The tree consisted of three distinct subdivisions corresponding to the three CP groups delineated above. Also, the association between the proteins of groups I and II was confirmed. The tree provided some intriguing insights into the evolution of virus CPs. CPs of groups II and III lack the P domains (Fig. 1). Comparison of the β-sheet topologies in the S and P domains of the TBSV CP suggested that the latter domain was probably derived from an independent cellular gene (Gibson & Argos, 1990). As CPs with the P domain cluster together in the tree in Fig. 2, it seems likely that this might be a relatively recent evolutionary event, which occurred after the divergence of groups I and II.

Several aspects of the CP tree seem unusual when compared with the evolutionary patterns revealed upon comparative analysis of the large non-structural viral
proteins, including the tentative RNA polymerases, and of virus genome organizations. All viruses with CPs belonging to the family described here lack the (putative) helicase domain found in a number of positive-strand RNA viruses with larger genomes (Gorbalenya & Koonin, 1989; Carrington et al., 1989). The previous claims of identification of a counterpart to the helicase domain in the proteins of sobemoviruses (Wu et al., 1987) and luteoviruses (Habili & Symons, 1989) failed to survive a close scrutiny (Gorbalenya et al., 1989; and unpublished observations).

It has been shown that the polymerases of sobemoviruses and a subset of luteoviruses (BWYV and PLRV), on the one hand, and those of carmo-, tombus-, dianthoviruses, MCMV and the remaining luteovirus (BYDV), on the other hand, belong to different 'supergroups' (Koonin, 1990). Moreover, the viruses with the two polymerase types exploit basically different strategies for their genome expression. SBMV, BWYV and PLRV appear to encode serine proteases (Gorbalenya et al., 1988; Bazan & Fletterick, 1989) and generate at least some of their proteins by polyprotein processing. Carmoviruses and related viruses have no proteases. On the other hand, viruses of both types produce their CPs via subgenomic mRNAs. Along with the fact that two subdivisions of the CP tree, II and III, combine viruses with different polymerase types, these observations suggest recombinational transfer of the CP cistron, together with the subgenomic mRNA promoter, from carmovirus-type viruses to sobemo- and luteo-type viruses. This direction of transfer seems as attractive as subgenomic mRNA formation observed in sobemo- and luteoviruses is unusual for viruses with genome-linked 5' proteins (VPg; Vartapetian & Bogdanov, 1987) and might be an evolutionarily 'young' feature. Recombination events could occur independently in groups II and III, in luteoviruses involving not only the CP cistron, but also the downstream read-through cistron (Miller et al., 1988).

Group I in the tree in Fig. 2, though including viruses with the single polymerase type, also shows an unusual feature. This is the grouping of RCNMV with tombusviruses. Dianthoviruses differ from carmo- and tombus-viruses in that they possess two genomic RNA segments, the second segment coding for the putative movement protein related to the movement proteins of tricornaviruses (Lommel et al., 1988). Thus inspection of each group of the CPs of small spherical RNA plant viruses reveals obvious signs of gene exchange between various virus genomes.

Generally, the present results are fully compatible with the module concepts of positive-strand virus genome evolution (Gibbs, 1987; Zimmern, 1988; Morozov et al., 1989).

**Addendum.** During the processing of this manuscript, a tentative phylogenetic tree generated by a cluster algorithm was published, and included a subset of the viral CPs considered here (Martin et al., 1990). The tree topology was similar to that derived in this paper, with the exception that the grouping of MCMV with SBMV was not observed by these workers.

V. V. D. is grateful to Professor J. G. Atabekov for constant interest and encouragement. The authors thank Dr A. V. Karasev for helpful discussions and critical reading of the manuscript.

**References**


(Received 14 December 1990; Accepted 28 March 1991)