Comparisons of the genomic sequences of erysimum latent virus and other tymoviruses: a search for the molecular basis of their host specificities

Pattana Srifah, Paul Keese, Georg Weiller and Adrian Gibbs.

The nucleotide sequence of the genome of erysimum latent tymovirus (ELV) has been determined. It closely resembles those of the other four sequenced tymoviral genomes in its gene organization and composition, but is the smallest (6034 nucleotides) and most distinct of them. Furthermore the 78 non-coding nucleotides at the 3' terminus of the ELV genome are unable to form a complete tRNA-like structure like that reported for other tymoviruses. Comparisons of the five tymovirus genomes and their encoded proteins indicate that they have probably evolved from the progenitor tymovirus by independent progressive mutational change without genetic recombination. Comparisons of the sequences of the two non-virion proteins of five tymoviruses, and virion proteins of 17 tymoviruses, revealed no specific similarities between those of ELV and turnip yellow mosaic virus that could explain why their host ranges and symptoms are so similar, yet differ, in this respect, from ononis yellow mosaic, kennedya yellow mosaic and eggplant mosaic tymoviruses.

Introduction

Erysimum latent tymovirus (ELV), which was first isolated and described by Shukla & Schmelzer (1972), came from symptomless Erysimum helveticum (Jacq.) DC. growing in East Germany. ELV, like turnip yellow mosaic virus (TYMV), causes diseases of wild and crop brassicas and preferentially infects (Guy et al., 1984) crassinucellate dicotyledonous plants (Young & Watson, 1970). However, despite these clear biological similarities to TYMV, its particles are serologically unrelated to those of TYMV and only distantly related to those of Andean potato latent (APLV) and ononis yellow mosaic (OYMV) tymoviruses (Shukla & Gough, 1980).

ELV has icosahedral particles with a diameter of about 27 nm, and these sediment as two components (Shukla et al., 1980). Those particles which sediment at approximately 113S are infectious, and contain both the genomic RNA and subgenomic RNA (the mRNA of the virion protein). Particles of the other component sediment at approximately 59S, and contain no nucleic acid.

The genome of ELV is a linear, single-stranded, positive-sense RNA molecule with an $M_r$ of about $2 \times 10^6$. It constitutes 32% of the nucleoprotein particle weight (Gough et al., 1982), and like those of other tymoviruses has a small guanine content (15.6%) and an unusually large cytosine content (34.5%) (Shukla et al., 1980).

The particles of ELV contain a single species of virion protein (VP), the sequence of which has been reported (Srifah et al., 1990). It has 202 amino acid residues, and hence is somewhat larger than the VPs of other tymoviruses. Its sequence is the most distinct of the known tymoviral VPs, and this feature correlates with the very distant, or lack of, serological relationships between the particles of ELV and other tymoviruses.

Thus ELV resembles TYMV closely in the species of plants it infects and the symptoms it causes, yet their particles are serologically so distant that they are related only through intermediate tymoviruses. This implies that the molecular basis of the biological similarities of ELV and TYMV may reside in other proteins they encode or in their nucleic acids.

In this paper, we report the complete genomic sequence of ELV, and compare that sequence with those reported for four other tymoviruses, namely eggplant mosaic virus (EMV-Trin; Osorio-Keese et al., 1989), kennedya yellow mosaic virus (KYMV-JB; Ding et al., 1990c), OYMV (OYMV-Tin; Ding et al., 1989) and TYMV (TYMV-CL; Keese et al., 1989). The main aim of these comparisons was to search for specific sequence features that could be responsible for the biological similarities of ELV and TYMV-CL.
Methods

Preparing cloned DNA encoding the ELV genome. The ELV used for cloning and sequencing was kindly supplied by Dr D. D. Shukla. The virus was propagated and maintained in *Brassica campestris* ssp. *pekinesis* L. Viral genomic RNA was extracted from particles purified as described previously (Gibbs et al., 1966; Shukla et al., 1973; Both & Air, 1979). A library of clones containing DNA encoding parts of the ELV genome was generated as described previously (Keese et al., 1989) using *Del*. *HaeIII*, *Hind* I, *MspI*, *NcoI*, *SacI* or *TaqI* restriction endonucleases and bacteriophage M13 or pTZ plasmid vectors.

Sequencing. The viral inserts in 276 randomly selected M13 clones (approximately 40000 nucleotides), mainly from *HaeIII* and *MspI* fragments of cDNA, were sequenced using the dideoxynucleotide chain termination method (Sanger et al., 1980), and the sequence was compiled by computer. Most of the sequence was obtained from independent clones which formed an overlapping set in both orientations and extended over about 95% of the genome. No clones were obtained for a region that was shown subsequently to span nucleotides 1950 to 2034, therefore the primer PS2: 5' CTGGAAAGG-AYCTCTGAAAGG 3' (complementary to nucleotides 2035 to 2056), was used as a first-strand primer to synthesize, by the method of Gubler & Hoffman (1983), a dsDNA *NcoI* fragment of about 100 bp to span this region. This was sequenced in the same way as the other clones. The 5'-proximal sequence was determined as described by Forman et al. (1988) using an oligonucleotide primer (PS3, 5' CCGGAGACTTG-AATGCCC 3') complementary to nucleotides 261 to 276.

Computer comparisons. Nucleic acid and deduced amino acid sequences were analysed using programs of the SEQ library (Research School of Biological Sciences, Australian National University, Canberra, Australia) and the GCG (Version 6) library (University of Wisconsin Genetic Computing Group) in a VAX computer. In addition the sequences were aligned by the progressive alignment method of Feng & Doolittle (1987), and the dissimilarities of the aligned sequences, calculated by the method of Feng & Nei (1987) as implemented by Studier & Keppler (1988). A program, SEQCORR, was used to assess the correlation between amino acid or nucleotide sequence differences at each position in the aligned sequences and known differences in the host preferences of the viruses providing those sequences.

Results and Discussion

Nucleotide sequence of the ELV genome

The sequence of the ELV genome is shown in Fig. 1. The ELV genome is 6034 nucleotides in length. Only at nucleotide position 5004 was sequence heterogeneity, A and G, found. The sequence has a base composition of 16-24% G, 23-72% A, 25-46% U and 34-55% C, which is close to the chemically determined composition (Shukla et al., 1980). Despite its large cytosine content, the genome has only 12 strings of five or six cytosine residues (C<sub>5</sub> or C<sub>6</sub>); this is many fewer than the 22 strings of C<sub>5</sub> five of C<sub>6</sub>, one of C<sub>7</sub> and one of C<sub>11</sub> found in TYMV (Morch et al., 1988; Keese et al., 1989).

There are two guanines at the 5' terminus of the genome. One of these may correspond to the 7'-methyl-

guanosine cap structure that probably occurs in all tymoviruses (Klein et al., 1976; Pleij et al., 1976) as the nucleotide in the cap structure of some viruses has been found to contribute to the determined sequence (Ahquist & Janda, 1984; Gupta & Kingsbury, 1984). Thus it is likely that the sequence at the 5' terminus of ELV is m<sup>7</sup>GpppGAAAUU—, although the possibility that it is m<sup>7</sup>GpppGGAAAAU— has not been excluded.

Five separate clones with a 3'-terminal poly(A) sequence were found. It was concluded that these encoded the 3' terminus of the genome as that had been polyadenylated in order to obtain cDNA complementary to the genome. The sequence of each clone was determined, in both orientations. Their 5' portions were homologous, and they terminated in -C or -C(A) like other tymoviruses (van Belkum et al., 1987). It is unlikely that the ELV genome has a longer poly(A) tail as the primer dT<sub>8</sub>G failed to assist transcription unless the genomic RNA had been polyadenylated.

The ELV genome was found to be very compact and had no intercistronic regions; 1-59% of the genome at its 5' terminus and 1-29% at its 3' terminus seems not to encode amino acids. Computer analysis of the genomic sequence revealed three major open reading frames (ORFs) encoding proteins with predicted *M*<sub>s</sub> of 48597, 193811 and 21483. Their amino acid sequences are shown in Fig. 1. The ORF that starts closest to the 5' terminus at nucleotide 97 and terminates with UGA at nucleotide 1419 encodes a protein of 440 amino acid residues. This ORF is similar in size and genomic position to that encoding the overlapping protein (OP) of previously described tymoviruses. The function of the OP is at present unknown. The second ORF, which encodes a protein of 1747 amino acids, begins at nucleotide 103 seven nucleotides after the OP ORF and hence out-of-phase with it, and terminates with a UGA codon at nucleotide 5350. This protein is the largest ELV protein. It has sequence motifs typical of viral replicase proteins (RPs) and is homologous to the RP of TYMV (Morch et al., 1988; Keese et al., 1989). The third ORF is located near the 3' terminus of the genome and has been shown to encode the VP (Srifah et al., 1990). The termination codon (UGA) of the RP ORF overlaps the AUG start codon of the VP ORF at nucleotide 5347, and thus the VP ORF, which ends at nucleotide 5955, is in the same phase as that of the OP.

Close to the 3' terminus of the RP gene is a 16 nucleotide sequence (underlined in Fig. 1; nucleotides 5315 to 5330) conserved in all tymoviruses that have been sequenced, and called the tymobox by Ding et al. (1990a). The tymobox probably acts as the transcriptional promoter for the subgenomic mRNA of the VP gene. Eleven of the 15 known tymobox sequences are identical; that of ELV, like that of KYMV-JB, has U
instead of C at the fifth position and, as a result, codes for the sequence -EFELL- at the C-terminal end of the RP, rather than for -ESELL- as in most other tymoviruses.

One consequence of the large cytosine content of the genome is the large number of cytosine-rich codons, and hence of amino acids such as leucine (CNN), proline (CCN) and serine (NCN) in the encoded polypeptides.

The plan of tymovirus genomes

The complete genomic sequences of six tymoviruses are known. These viruses are ELV, EMV-Trin, KYMV-JB, OYMV-Tin, TYMV-CL and TYMV-type (Morch et al., 1988); the genomic sequences of TYMV-CL and TYMV-type differ by only 54%, so the comparisons reported here were made using the TYMV-CL sequence only. Comparisons of the genomic sequences of the five tymoviruses show that the basic plan of their genomes is closely similar (Fig. 2); ELV has the smallest genome. They all have three ORFs (OP, RP and VP) and two of these overlap. In each, the OP ORF starts seven nucleotides before the RP ORF, but is only about one-third as long. The ORF for the third protein, the VP, is at the 3' end of the genome and, in different tymoviruses, is in any of the three reading frames with respect to the OP/RP ORFs. All components of the genomes vary in length; among the ORFs the OPs are the most variable whereas, of the intergenic regions, that between the RP and VP genes varies from +14 to -20 nucleotides in length. It is 14 and four nucleotides long in TYMV and EMV respectively, whereas in ELV and KYMV-JB there is no intergenic region and in OYMV-Tin the RP gene overlaps that of the VP by 20 nucleotides.

The terminal untranslated regions of the genomes

The 5' non-coding regions of the known tymoviral genomes vary from 78 to 171 nucleotides in length. They have a smaller guanine content (8·3 to 12·9%) than the remainder of the genome (15·1 to 16·8%), and a slightly smaller cytosine content (24·4 to 30·2%) compared with 34·6 to 39·4%). The sequences of the 5'-terminal regions of ELV, EMV-Trin, KYMV-JB, OYMV-Tin and TYMV-CL were examined directly and using the Squiggles program of the GCG package. They formed structures with two, three, two, four and one stem–loops, and these had free energies of ΔG = -3·5, -3·2, -3·1, -3·5 and -5·5 kJ/mol, respectively (Zuker & Steigler, 1981). However it has been reported previously that this region of the OYMV genome might form two stem–loops with such similar sequences that they may have arisen by duplication (Ding et al., 1989). The two stem–loops of ELV may also have a similar history as the base-paired parts of the first loop (nucleotides 17 to 45) and the second loop (nucleotides 52 to 106) have similar sequences.

The 3' non-coding regions have an above average guanine content (18·9 to 25·2%) and their cytosine content is similar to that of the remainder of the genome. They vary from 78 to 147 nucleotides in length; that of ELV (nucleotides 5936 to 6034) is the shortest. The Squiggles program failed to predict a tRNA-like secondary structure for the ELV 3'-terminal sequence like those found at the 3' termini of other tymoviral genomes, therefore it was aligned by eye with those of other tymoviruses and found to have only part of the traditional cloverleaf or L-shaped tRNA-like structure. It probably has four stem–loops with short arms, that have six, four, three and two base-paired nucleotides. In Fig. 3 this structure is compared with the similar regions of four other tymoviral genomes. It is clear that the 3'-terminal tRNA-like structures of most tymoviruses are similar but not identical; the numbers of stem–loops are the same, but their details differ; however, that of ELV is different. The 3'-most terminal stem–loop structure and accompanying pseudo-knot of the tymovirus tRNA-like structure has a homologous counterpart in ELV RNA, however no other structural components, including the valine anticodon loop, appear to be present.

It has been shown that several tymoviral genomic RNAs can be aminoacylated with valine (Rietveld et al., 1982), however the fact that ELV could not be acylated led van Belkum et al. (1987) to suggest that ELV might lack a tRNA-like structure, and this agrees with our results. It is uncertain how essential are the 3'-terminal tRNA-like structures of tymoviruses. It is clear that in some viruses these structures are involved in viral replication (Ahlquist et al., 1984; Miller et al., 1985; Bujarski et al., 1986; Dreher & Hall, 1988a, b), but the details are uncertain. Variants of tobacco mosaic virus (TMV) that lack the 3'-terminal five to 10 nucleotides are not infectious (Salomon et al., 1976), and certain point mutations within the tRNA-like structure of the brome mosaic virus genome destroy its infectivity (Dreher & Hall, 1988b), are not aminoacylated (Hall, 1979) and do not function as a template for synthesis of the genomic minus strand. By contrast, the fact that the 3'-terminal region of the ELV genome probably forms only part of the tRNA-like structure found in other tymoviruses indicates that the recognition by valine-acyl tRNA synthetase is not essential for tymovirus survival, and that other parts of the ELV genomes can replace its function.

The ORFs and encoded proteins of tymoviruses

Tymovirus genomes encode three proteins. The OPs, and the ORFs that encode them, vary much more in size than...
Fig. 1. The nucleotide sequence of the genome of ELV together with the amino acid sequences of polypeptides encoded by the three main ORFs referred to in the text as RP, OP and VP. Termination codons are indicated by asterisks. The tymobox sequence is underlined.
The OP and RP ORFs all start seven nucleotides to the 5' side of the start of the ORFs, thus their size differences reflect the variability in the position of their termination codons. A major consequence of the relative positions of the start codons of the OP and RP ORFs is that the many cytosine residues found in the third positions of the ORFs (Keese et al., 1989) are in the first codon positions, which is found in the RPs of most virus genomes (Koonin, 1991). Region 1000 to 1250 contains the sequence motif -GDD-, which is found in the RPs of most virus genomes (Gorbalenya et al., 1989a, b). Region 1 to 450 of the RPs is the least conserved, and shows distant sequence similarities to a region associated with methyltransferase activity in alphaviruses (Mi et al., 1989; Candresse et al., 1990).

FJD distances between the aligned RPs were used to calculate a dendrogram (Fig. 4) by the neighbour-joining method. It can be seen that this has the same topology as that for the OPs and VPs (Srifah et al., 1990); that of the OPs had the longest branch lengths and that of the RPs the shortest. In another analysis, each of the RPs was separated into the three parts, aligned residues 1 to 650, 651 to 1350 and 1351 to 1928, and these were classified separately. The three dendrograms (not shown) had the same topology as the dendrogram of the entire RPs, although their mean branch lengths differed; the N-terminal portions had a mean identity of 48.6%, the central regions 44.7% and the C-terminal regions 60.0%.

Fig. 4 also shows a classification of the five VPs using their FJD distances. It can been seen that the five viruses fall into three subgroups; EMV–OYMV, TYMV–KYMV and ELV which has its own long separate branch. The average branch lengths for each protein are, however, different, reflecting differences in the rates of evolutionary change of the three proteins; the maximum FJD distance of RPs, VPs and OPs are about 34%, 54% and 90% respectively showing that the greatest rate of change is in the OPs. These classifications show that the relationships of the five fully sequenced tymoviruses are

(CAY), and also of proline (CCN) and leucine (CUN) in the OPs.

The OP sequences were aligned by the progressive alignment method of Feng & Doolittle (1987). They were found to have pairwise identities varying from 25-7% to 33-5% with most of the conserved residues in their N-terminal 350 residues. The progressive alignment program also calculates pairwise similarity estimates (FJD distances; Feng et al., 1985) between the aligned sequences. These were used to calculate a dendrogram (Fig. 4) by the neighbour-joining method (Saito & Nei, 1987; Studier & Keppler, 1988).

The largest ORFs of the tymoviral genomes encode the RPs, which are of Mr 194K to 210K and have closely similar sequences with identities of 50 to 56%. Their sequences were also aligned by the progressive alignment method. The aligned sequences had three regions of close similarity around residues 1 to 450, 1000 to 1250 and 1450 to 1850, which correspond to the regions of sequence conservation noted by Rozanov et al. (1990). Region 1450 to 1850 contains the sequence motif -GDD-, which is found in the RPs of most virus genomes (Koonin, 1991). Region 1000 to 1250 contains the -GXXGXGCGK(T/S)- motif that is characteristic of many NTP-utilizing enzymes, and is found in the majority of viral RPs (Gorbalenya et al., 1989a, b). Region 1 to 450 of the RPs is the least conserved, and shows distant sequence similarities to a region associated with methyltransferase activity in alphaviruses (Mi et al., 1989; Candresse et al., 1990).

FJD distances between the aligned RPs were used to calculate a dendrogram (Fig. 4) by the neighbour-joining method. It can be seen that this has the same topology as that for the OPs and VPs (Srifah et al., 1990); that of the OPs had the longest branch lengths and that of the RPs the shortest. In another analysis, each of the RPs was separated into the three parts, aligned residues 1 to 650, 651 to 1350 and 1351 to 1928, and these were classified separately. The three dendrograms (not shown) had the same topology as the dendrogram of the entire RPs, although their mean branch lengths differed; the N-terminal portions had a mean identity of 48.6%, the central regions 44.7% and the C-terminal regions 60.0%.

Fig. 4 also shows a classification of the five VPs using their FJD distances. It can been seen that the five viruses fall into three subgroups; EMV–OYMV, TYMV–KYMV and ELV which has its own long separate branch. The average branch lengths for each protein are, however, different, reflecting differences in the rates of evolutionary change of the three proteins; the maximum FJD distance of RPs, VPs and OPs are about 34%, 54% and 90% respectively showing that the greatest rate of change is in the OPs. These classifications show that the relationships of the five fully sequenced tymoviruses are
Fig. 3. Sequences of the 3' termini of five tymoviral genomes and their potential tRNA-like secondary structures together with that of yeast tRNA^{Val} (Bonnet et al., 1974).

Fig. 4. Dendrograms showing the relationships of the three proteins (OP, RP and VP) of five tymoviruses. The dendrograms were calculated by the neighbour-joining program of Saito & Nei (1987) from the FJD distances (Feng et al., 1985) of the sequences that had been aligned by the progressive alignment program of Feng & Doolittle (1987).
Table 1. Positions in the aligned proteins of tymoviruses where differences in the amino acids at that position correlated with the taxonomy of the natural hosts of the virus

(a) Overlapping proteins

<table>
<thead>
<tr>
<th>Virus</th>
<th>C/T†</th>
<th>8‡</th>
<th>41‡</th>
<th>81‡</th>
<th>83‡</th>
<th>119‡</th>
<th>204‡</th>
<th>233‡</th>
<th>307‡</th>
<th>341‡</th>
<th>354‡</th>
<th>376‡</th>
<th>378‡</th>
<th>388‡</th>
<th>444‡</th>
</tr>
</thead>
<tbody>
<tr>
<td>ELV</td>
<td>C</td>
<td>S</td>
<td>P</td>
<td>L</td>
<td>P</td>
<td>P</td>
<td>R</td>
<td>P</td>
<td>P</td>
<td>S</td>
<td>R</td>
<td>N</td>
<td>R</td>
<td>R</td>
<td>B</td>
</tr>
<tr>
<td>EMV-Trin</td>
<td>T</td>
<td>C</td>
<td>S</td>
<td>V</td>
<td>S</td>
<td>G</td>
<td>F</td>
<td>S</td>
<td>R</td>
<td>C</td>
<td>D</td>
<td>V</td>
<td>S</td>
<td>S</td>
<td>R</td>
</tr>
<tr>
<td>KYMV-JB</td>
<td>C</td>
<td>S</td>
<td>P</td>
<td>L</td>
<td>P</td>
<td>P</td>
<td>R</td>
<td>P</td>
<td>P</td>
<td>S</td>
<td>R</td>
<td>N</td>
<td>R</td>
<td>R</td>
<td>L</td>
</tr>
<tr>
<td>OYMV-Tin</td>
<td>C</td>
<td>S</td>
<td>P</td>
<td>L</td>
<td>P</td>
<td>P</td>
<td>R</td>
<td>P</td>
<td>P</td>
<td>S</td>
<td>R</td>
<td>N</td>
<td>R</td>
<td>R</td>
<td>L</td>
</tr>
<tr>
<td>TYMV-CL</td>
<td>C</td>
<td>S</td>
<td>P</td>
<td>L</td>
<td>P</td>
<td>P</td>
<td>R</td>
<td>P</td>
<td>P</td>
<td>S</td>
<td>R</td>
<td>N</td>
<td>R</td>
<td>R</td>
<td>L</td>
</tr>
</tbody>
</table>

(b) Replicase proteins

| Virus   | C/T† | 51  | 116‡ | 36‡ | 41‡ | 122‡ | 133‡ | 119‡ | 187‡ | 203‡ | 206‡ | 375‡ | 376‡ | 408‡ | 41‡ | 585‡ | 677‡ | 835‡ | 845‡ | 1076‡ | 1134‡ | 1166‡ | 1181‡ | 1208‡ | 1228‡ | 129‡ | 1280‡ | 1465‡ | 1525‡ | 1534‡ | 1593‡ | 1687‡ | 1688‡ | 1732‡ | 1770‡ | 1817‡ | 1834‡ | 1836‡ |
|---------|------|-----|------|-----|-----|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|
| ELV     | C    | L   | S   | E   | V   | S   | L   | F   | V   | L   | G   | A   | L   | L   | L   | L   | P   | P   | A   | L   | L   | Y   | F   | S   | Q   | S   | S   | A   | N   | S   | G   | D   | T   | N   | V   | G   | L   | L   | L   |
| EMV-Trin| T    | S   | N   | S   | L   | F   | T   | F   | C   | I   | N   | S   | S   | S   | S   | V   | L   | S   | C   | P   | P   | H   | N   | A   | R   | R   | N   | P   | Q   | Q   | S   | G   | C   | S   | I   | S   | A   | I   | V   |
| OYMV-Tin| C    | L   | A   | Q   | I   | Y   | S   | L   | V   | V   | L   | G   | A   | L   | L   | L   | L   | P   | A   | L   | L   | Y   | F   | S   | Q   | S   | A   | S   | G   | D   | T   | N   | V   | G   | L   | L   | L   |
| TYMV-CL | C    | L   | A   | Q   | I   | Y   | S   | L   | I   | V   | G   | A   | L   | L   | L   | L   | L   | P   | A   | L   | L   | Y   | F   | S   | Q   | S   | A   | N   | S   | G   | D   | T   | N   | V   | G   | L   | L   |

<p>| Virus   | C/T† | 51  | 116‡ | 36‡ | 41‡ | 122‡ | 133‡ | 119‡ | 187‡ | 203‡ | 206‡ | 375‡ | 376‡ | 408‡ | 41‡ | 585‡ | 677‡ | 835‡ | 845‡ | 1076‡ | 1134‡ | 1166‡ | 1181‡ | 1208‡ | 1228‡ | 129‡ | 1280‡ | 1465‡ | 1525‡ | 1534‡ | 1593‡ | 1687‡ | 1688‡ | 1732‡ | 1770‡ | 1817‡ | 1834‡ | 1836‡ |
|---------|------|-----|------|-----|-----|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|
| OYMV-Tin| L    | F   | T   | I   | L   | S   | I   | I   | C   | P   | S   | L   | T   | T   | A   | D   | D   | Y   | A   | T   | A   | D   | I   | V   | S   | A   | I   | H   | P   | F   | F   | S   | A   | N   | F   | S   | I   | V   | P   |
| TYMV-CL | B    | N   | L   | H   | L   | Q   | R   | A   | A   | I   | S   | F   | V   | P   | T   | E   | S   | T   | E   | V   | N   | S   | C   | R   | H   | A   | T   | Y   | L   | S   | T   | A   | A   | N   | Q   | S   | Y   | R   | V   | S   | I   |</p>
<table>
<thead>
<tr>
<th>Virus</th>
<th>C/T</th>
<th>30</th>
<th>60</th>
<th>77</th>
<th>99</th>
<th>129</th>
<th>158</th>
<th>179</th>
<th>Host family†</th>
<th>26</th>
<th>68</th>
<th>148</th>
<th>149</th>
<th>181</th>
</tr>
</thead>
<tbody>
<tr>
<td>APLV</td>
<td>T</td>
<td>Y</td>
<td>A</td>
<td>N</td>
<td>N</td>
<td>T</td>
<td>L</td>
<td>K</td>
<td>S</td>
<td>N</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>V</td>
</tr>
<tr>
<td>BMV</td>
<td>T</td>
<td>F</td>
<td>L</td>
<td>G</td>
<td>C</td>
<td>L</td>
<td>I</td>
<td>K</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>V</td>
</tr>
<tr>
<td>CoYMV</td>
<td>C</td>
<td>S</td>
<td>A</td>
<td>L</td>
<td>F</td>
<td>C</td>
<td>T</td>
<td>T</td>
<td>St</td>
<td>N</td>
<td>Q</td>
<td>N</td>
<td>I</td>
<td>P</td>
</tr>
<tr>
<td>CYVV</td>
<td>C</td>
<td>S</td>
<td>S</td>
<td>L</td>
<td>F</td>
<td>C</td>
<td>T</td>
<td>T</td>
<td>L</td>
<td>N</td>
<td>Q</td>
<td>N</td>
<td>I</td>
<td>P</td>
</tr>
<tr>
<td>DMV</td>
<td>T</td>
<td>F</td>
<td>L</td>
<td>G</td>
<td>N</td>
<td>L</td>
<td>I</td>
<td>K</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>V</td>
</tr>
<tr>
<td>ELV</td>
<td>C</td>
<td>S</td>
<td>S</td>
<td>R</td>
<td>H</td>
<td>S</td>
<td>S</td>
<td>T</td>
<td>B</td>
<td>V</td>
<td>F</td>
<td>M</td>
<td>M</td>
<td>S</td>
</tr>
<tr>
<td>EMV</td>
<td>T</td>
<td>F</td>
<td>A</td>
<td>L</td>
<td>N</td>
<td>T</td>
<td>L</td>
<td>K</td>
<td>S</td>
<td>N</td>
<td>A</td>
<td>S</td>
<td>V</td>
<td>P</td>
</tr>
<tr>
<td>KYMV-BP</td>
<td>C</td>
<td>S</td>
<td>S</td>
<td>L</td>
<td>Y</td>
<td>C</td>
<td>T</td>
<td>T</td>
<td>L</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>V</td>
<td>A</td>
</tr>
<tr>
<td>KYMV-JB</td>
<td>C</td>
<td>S</td>
<td>S</td>
<td>L</td>
<td>F</td>
<td>C</td>
<td>T</td>
<td>T</td>
<td>L</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>V</td>
<td>A</td>
</tr>
<tr>
<td>KYMV-PD</td>
<td>C</td>
<td>S</td>
<td>S</td>
<td>L</td>
<td>Y</td>
<td>C</td>
<td>T</td>
<td>T</td>
<td>L</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>V</td>
<td>A</td>
</tr>
<tr>
<td>OYMV-Tin</td>
<td>C</td>
<td>T</td>
<td>S</td>
<td>L</td>
<td>I</td>
<td>N</td>
<td>T</td>
<td>L</td>
<td>S</td>
<td>S</td>
<td>D</td>
<td>S</td>
<td>V</td>
<td>P</td>
</tr>
<tr>
<td>TYMV-BL</td>
<td>C</td>
<td>T</td>
<td>E</td>
<td>F</td>
<td>F</td>
<td>C</td>
<td>Q</td>
<td>P</td>
<td>B</td>
<td>V</td>
<td>I</td>
<td>M</td>
<td>M</td>
<td>S</td>
</tr>
<tr>
<td>TYMV-Caul</td>
<td>C</td>
<td>S</td>
<td>T</td>
<td>L</td>
<td>F</td>
<td>C</td>
<td>Q</td>
<td>T</td>
<td>B</td>
<td>I</td>
<td>L</td>
<td>M</td>
<td>M</td>
<td>S</td>
</tr>
<tr>
<td>TYMV-CL</td>
<td>C</td>
<td>T</td>
<td>E</td>
<td>F</td>
<td>F</td>
<td>C</td>
<td>Q</td>
<td>P</td>
<td>B</td>
<td>V</td>
<td>I</td>
<td>M</td>
<td>M</td>
<td>S</td>
</tr>
<tr>
<td>TYMV-Roth</td>
<td>C</td>
<td>S</td>
<td>T</td>
<td>L</td>
<td>F</td>
<td>C</td>
<td>Q</td>
<td>T</td>
<td>B</td>
<td>I</td>
<td>L</td>
<td>M</td>
<td>M</td>
<td>S</td>
</tr>
<tr>
<td>TYMV-type</td>
<td>C</td>
<td>T</td>
<td>E</td>
<td>F</td>
<td>F</td>
<td>C</td>
<td>Q</td>
<td>P</td>
<td>B</td>
<td>V</td>
<td>I</td>
<td>M</td>
<td>M</td>
<td>S</td>
</tr>
<tr>
<td>WCuMV</td>
<td>C</td>
<td>H</td>
<td>S</td>
<td>L</td>
<td>L</td>
<td>T</td>
<td>S</td>
<td>S</td>
<td>Cu</td>
<td>N</td>
<td>C</td>
<td>S</td>
<td>V</td>
<td>T</td>
</tr>
</tbody>
</table>

* The positions were numbered from the N terminus of the sequences after they had been aligned by the progressive alignment method of Feng & Doolittle (1987) using its standard parameters.
† The major groupings of dicotyledonous plants, Crassinucelli (C) and Tenuinucelli (T), are as defined by Young & Watson (1970). Thus the Brassicaceae (B), Cucurbitaceae (Cu), Leguminosae (L) and Sterculiaceae (St) are crassinucellate, and the Solanaceae (S) are tenuinucellate. Pairs of viruses with the same natural host family (or major host group) were given a host similarity of one, those which differed were given a similarity of zero, and the similarities of the amino acids were taken from the MDM78 matrix (Dayhoff et al., 1978).
‡ Perfect correlations. For the comparisons between the OPs and RPs and the major host groupings, only positions giving perfect correlations are listed, for the other comparisons all positions giving correlation coefficients above 0.6 are listed.
§ The virion protein sequences of the five fully sequenced tymoviruses discussed in this paper were analysed together with those of APLV (M. Osorio-Keese, unpublished data), belladonna mottle virus (Ding et al., 1990a, b), cacao yellow mosaic virus (Ding et al., 1990d), clitoria yellow vein virus (A. Mackenzie, unpublished data), desmodium yellow mottle virus (J. Gibbs, unpublished data), KYMV-BP (A. Mackenzie, unpublished data), KYMV-PD (P. Keese, unpublished data), OYMV-Tin (Ding et al., 1989), TYMV-BL (D. Meek, unpublished data), TYMV-Cauliflower (A. Mackenzie, unpublished data), TYMV-Roth (M. Torronen, unpublished data), TYMV-type (Morch et al., 1988) and wild cucumber mosaic virus (A. Mackenzie, unpublished data).
closely related other genes. Therefore we examined the aligned sequences for more subtle similarities.

**Viral signals for host specificity**

Guy et al. (1984) showed that tymoviruses preferentially infect species of the same family as their known natural host or hosts (e.g. Brassicaceae, Cucurbitaceae, Leguminosae, Sterculiaceae or Solanaceae), or of the same higher order grouping of dicotyledonous plants, that is the ‘crassinucelli’ or ‘tenuinucelli’ as defined by Young & Watson (1970).

The aligned RPs and OPs of five tymoviruses, and a larger set of VPs, were searched for sequence similarities or motifs that might determine their host range preferences. A computer program, SEQCORR, was written and used to examine each position in the aligned sequences in turn, and to calculate a correlation coefficient between all pairwise comparisons of the amino acids at that position and the known host range preferences of the viruses. Pairs of viruses with the same natural host family were given a host similarity of one, those which differed were given a similarity of zero, and the similarities of the amino acids were taken from the MDM78 matrix (Dayhoff et al., 1978), in order to permit evolutionary redundancy in the comparisons. In another analysis the major host groupings were tested similarly. All positions that gave a correlation coefficient of 0-6 or more were examined, and most of these are listed in Table 1.

The best correlations were between the OPs and RPs and the major host groupings (i.e. the division into crassinucelli and tenuinucelli); Table 1 records only the perfect correlations found in these comparisons. In the comparisons of OPs and RPs with host families, only one OP position and three RP positions correlated perfectly; Table 1 records these and also all positions giving a correlation coefficient greater than 0-6. No perfect correlations were found between host groupings and the VP alignments, but again Table 1 lists all positions with a correlation coefficient greater than 0-6.

The positions that correlated showed some clustering, as has been found in the functional motifs of several proteins, but the number of positions in each protein that correlated was broadly related to the size and conservation of that protein (i.e. was greatest in the RPs), and this may indicate that the observed correlations are fortuitous. These results give no definite clues on which amino acids, if any, encode host preferences, but do indicate which residues might be targeted in an analysis of such preferences by site-directed mutagenesis of cloned viral genomes. Host range preferences are, of course, probably controlled by features of the sequences, but those features may be more subtle than those which the analytical method we used was able to detect.

**References**


(Received 5 November 1991; Accepted 19 February 1992)