The Use of 3′ Non-coding Nucleotide Sequences in the Taxonomy of Potyviruses: Application to Watermelon Mosaic Virus 2 and Soybean Mosaic Virus-N

By M. J. FRENKEL, C. W. WARD AND D. D. SHUKLA*
CSIRO, Division of Biotechnology, Parkville Laboratory, 343 Royal Parade, Parkville, Victoria 3052, Australia

(Accepted 16 June 1989)

SUMMARY
The sequence of the 3′ 1106 nucleotides of the watermelon mosaic virus 2 (WMV 2) genome has been determined. The sequence contains the complete coding region of the viral coat protein followed by a 3′ untranslated sequence of 251 nucleotides. When these sequences were compared with the equivalent regions of the N strain of soybean mosaic virus (SMV-N), the coat protein coding regions were 82% homologous, whereas the 3′ untranslated sequences were 78% homologous. Optimal alignment of the 3′ untranslated regions of RNA from 13 strains of seven other distinct potyviruses revealed that the degree of homology between strains was in the range 83 to 99%. In contrast, the sequences from distinct viruses had identities in the range 39 to 53%, comparable to the level of identity found between the 3′ non-coding regions of viruses from unrelated plant virus groups. On the basis of these results, WMV 2 and SMV-N could be regarded as strains of one virus. These results also suggest that the sequence of the 3′ untranslated region of the potyvirus genome may be an accurate marker of genetic relatedness and could serve as an aid to identification and classification of potyviruses.

INTRODUCTION
The number of viruses now recognized as potyviruses has expanded rapidly in recent years to a current estimate of at least 175 definitive and possible members (Milne, 1988; Shukla & Ward, 1989b), or 30% of all known plant viruses. Thus the potyvirus group is by far the largest of the 28 plant virus groups and much of its taxonomy is complex, inconsistent and confused. As Francki et al. (1985) pointed out, the existence of so many members and possible members of the potyvirus group, and the frequent description of new candidates, emphasize the need to agree on the characteristics required to assign viruses to the group and those that will distinguish distinct viruses from strains.

In a review of the historical development of general taxonomic principles, Mayr (1982) drew attention to Darwin’s comment that taxonomy reflects propinquity of descent and that all true classifications are genealogical. Mayr (1982) also traced the changes in the criteria used for general taxonomy from the use of descriptive morphological characters to the application of biochemical techniques that characterize the variation and evolution of molecules. Plant virus taxonomy is going through similar phases of development from the initial reliance on morphological, biological and serological properties to assignments based on coat protein and nucleic acid sequences. Thus, of the physical, biological and chemical properties used to classify potyviruses, protein and genome sequence information should represent the ultimate criteria.

We have shown recently that coat protein sequence data can be used to discriminate between distinct viruses and strains (Shukla & Ward, 1988). Distinct potyviruses have coat protein sequence identities of 38 to 71%, compared to 90 to 99% for related strains (see Shukla & Ward, 1989a, for review). These relationships can be readily identified from comparative HPLC
profiles of coat protein peptides (Shukla et al., 1988a) or serology with polyclonal antibodies directed towards virus-specific N termini of the coat proteins (Shukla et al., 1989a, b, c). Despite these advances, there are some problems, such as the occurrence of unexpected and inconsistent paired serological relationships between biologically distinct potyviruses and single sequence changes in key contact residues in the virus-specific epitopes of some strains. These problems still remain if the identification and classification is based solely on coat protein properties (Shukla & Ward, 1989b).

In this report we draw attention to an alternative approach to discriminating between independent potyviruses and strains. It developed from our attempts to establish whether watermelon mosaic virus 2 (WMV 2; Yu et al., 1989) and soybean mosaic virus-N (SMV-N; Eggenberger et al., 1989) were strains of one virus or distinct but closely related viruses. In a previous report, Yu et al. (1989) showed that the homology (83%) between the amino acid sequences of the coat proteins of WMV 2 and SMV-N lay midway between the homology found between distinct potyviruses and that for related strains. In order to investigate the genetic changes associated with these sequence differences and the degree of genetic relatedness between WMV 2 and SMV-N, we have determined the nucleotide sequences of the coat protein-coding region and the 3' untranslated region of WMV 2 RNA and compared them with the sequence data for RNA of SMV-N (Eggenberger et al., 1989) and other potyviruses (Allison et al., 1985a, b, 1986; Dougherty et al., 1985; Domier et al., 1986; Gough et al., 1987; Gunyuzlu et al., 1987; Lain et al., 1988; Maiss et al., 1989; Ravelonandro et al., 1988; Rosner & Raccah, 1988; van der Vlugt et al., 1989). The comparisons show that nucleotide sequences of the 3' untranslated region of the potyviral genome can serve as an aid to identification and classification of potyviruses.

METHODS

Isolation of WMV 2. The WMV 2 isolate was originally obtained from Mr R. S. Greber (Greber, 1978) and was maintained and propagated in marrow and purified according to methods described previously (Shukla et al., 1989a).

Isolation of RNA and synthesis of a ds cDNA library. RNA was extracted from WMV 2 using a modification of the method described by MacDonald et al. (1987). Briefly, virus particles (in 1 ml) were extracted in 10 ml of 6 M-guanidine hydrochloride, 0.2 M-sodium acetate, pH 5.2 and 10 mM-2-mercaptoethanol. Following the addition of 10 ml of ethanol, the mixture was placed at -20°C for 2 h. Precipitated material was sedimented, resuspended in 5 ml of the extraction buffer containing EDTA (10 mM) in the place of 2-mercaptoethanol and precipitated by the addition of 5 ml of ethanol. The sedimented material was resuspended in 2.5 ml of a solution containing 7 M-urea, 100 mM-Tris-HCl pH 7.5, 0.1 mM-EDTA and 0.1% SDS, the proteins were removed by extraction with buffer phenol and the RNA was precipitated by the addition of ethanol.

Double-stranded cDNA was synthesized using avian myeloblastosis virus reverse transcriptase for the first strand and RNase H/DNA polymerase I (Amersham) for the second strand essentially as described by Gubler & Hoffman (1983). EcoRI linkers were added to the ds cDNA and DNA larger than 500 bp was ligated to λgt11 arms, packaged, and used to produce a library of recombinant bacteriophage in Escherichia coli Y1090 cells (Amersham cDNA-λgt11 cloning system).

Screening of the library with a synthetic oligonucleotide. An aliquot of the cDNA library was screened with a 5' radiolabelled, mixed, synthetic oligonucleotide. The deoxyinosine-containing (Takahashi et al., 1985) oligonucleotide used, 5' [T(I)G(T)TTGG(T)T](T)TA(T)TC 3', was the reverse complement of the derived sequence coding for the peptide EY(3)PKQ(3)NO5 (Yu et al., 1989) and was synthesized using an Applied Biosystems Model 381 DNA synthesizer. The methods used to radiolabel the 5' end of the oligonucleotide with [32P]ATP using polynucleotide kinase, and to screen the library, were those described by Maniatis et al. (1982) and Miyada & Wallace (1987), respectively. A temperature of 30°C was used for both the hybridization and washing of the nitrocellulose filters. Bacteriophage DNA was purified from liquid lysates of two positively reacting clones using the method of Kao et al. (1982).

Subcloning and DNA sequence determination. EcoRI was used to digest the bacteriophage DNA and the sizes of the excised inserts were estimated on agarose gels. Insert DNA was isolated and ligated to EcoRI-digested M13mp19. Following transformation, recombinants were isolated and ssDNA was prepared for sequence determination, essentially as described by Sanger et al. (1980). DNA sequencing reactions were carried out using the chain termination method of Sanger et al. (1980) with either a 17-mer universal sequencing primer (Amersham) or with synthetic oligonucleotides based on previously determined sequences.

Pairs of sequences were aligned using the ALIGN program supplied by the National Biomedical Research
Fig. 1. Comparison of the complete nucleotide sequences of WMV 2 and SMV-N (Eggenberger et al., 1989) coat protein genes, deduced amino acid sequences and their 3' untranslated regions. The SMV-N protein sequence is shown only where it differs from that of WMV 2. The sequences were aligned for maximum identity using the ALIGN program of NBRF with a matrix bias of 0 and a break penalty of 5. Differences in the nucleotide sequences are shown by asterisks and deletions are indicated by dots. The WMV 2 amino acid residue numbers are shown above the protein sequence and the nucleotide numbers (including deletions) are shown at the end of each line.

Foundation (NBRF), with a matrix bias of zero and a break penalty of 5. The percentage identity is based on the number of possible matches and disregards deletions introduced to maximize the homology.
RESULTS

Characterization of a cDNA clone encoding the WMV 2 coat protein sequence

Using a 5' radiolabelled synthetic oligonucleotide as the probe, a screen of 6000 recombinants from the WMV 2 cDNA library yielded 17 positive clones. Two clones were selected that contained inserts of approximately 1600 bp and 800 bp. The inserts were subcloned into M13mp19 in both orientations and the complete sequences of both DNA strands for each clone were determined and found to be identical for the overlapping region. The sequence of the coding strand from the region immediately 5' to the coat protein gene to the poly(A) tail is shown in Fig. 1. The location of the coat protein amino terminus is based on the N-terminal amino acid sequence described by Yu et al. (1989).

The coat protein gene comprises an open reading frame of 843 nucleotides which codes for a protein of 281 residues, and is followed by a 3' untranslated region of 251 nucleotides, excluding the poly(A) tail (Fig. 1). The sequence derived from the cDNA clone is identical to that determined by direct protein sequencing (Yu et al., 1989).

Comparison of the 3' ends of the WMV 2 and SMV-N genomes

Fig. 1 shows the alignment for optimal homology of the DNA sequences for the coat protein genes and the 3' untranslated regions of WMV 2 and SMV-N. If the 48 nucleotide deletion in the coding region of SMV-N is disregarded, and some gaps in WMV 2 are introduced to maximize the homology in the 3' non-coding region, the coat protein-coding sequences are 82% homologous, whereas the 3' untranslated sequences are 78% homologous. The alignment shown in Fig. 1 indicates that the differences between the sequences are evenly distributed except for the concentration of mismatches at the 5' end of the coat protein-coding region where there are 24 differences in the first 69 nucleotides.

Comparisons among the 3' untranslated sequences of different potyviruses and potyvirus strains

Fig. 2 shows the optimal alignments of the 3' untranslated sequences of known and predicted strains of different potyviruses. The degree of identity between strains (Table 1) was 99% for the two strains of tobacco etch virus (TEV) and 98 or 99% for the different strains of plum pox virus (PPV). The N and I strains of potato virus Y (PVY) were 83% identical, whereas these two PVY strains had identities of 83% and 92%, respectively, with pepper mottle virus (PeMV). The sequence homology of 78% between WMV 2 and SMV-N is very similar to that observed between the PVY strains. These and other comparisons are shown in Table 1.

Although alignments of the 3' untranslated sequences of 13 strains of seven distinct potyviruses revealed degrees of homology of 83 to 99%, similarities in the range 39 to 53% were found in comparisons between unrelated viruses, with most showing about 45% homology (Table 1). The percentage identity between unrelated virus sequences is very similar to that found for comparisons between viruses from different taxonomic groups (data not shown) and presumably this level of similarity arises by chance. An alignment of the coat protein gene sequences of the potyviruses showed that the degree of homology in comparisons of unrelated viruses was generally in the range 56 to 64%, whereas the corresponding sequences of related strains were 88 to 99% homologous (Table 1).

DISCUSSION

The amino acid sequence homology (83%) between the coat proteins of WMV 2 and SMV-N (Yu et al., 1989) lies midway between the ranges of similarity for independent potyviruses (38 to 71%) and related strains (90 to 99%). Most of the differences (including a 16 residue deletion in SMV-N) occurred in the 45 N-terminal residues, with the remainder of the two coat protein molecules being very similar (92% homology). This observation, that the 235 C-terminal amino acid residues are far more conserved than the surface-exposed, immunodominant N-terminal region (Shukla et al., 1988b), suggested that the 3' end of the potyvirus genome including the untranslated region may be an accurate marker of genetic relatedness.

Complete nucleotide sequences are now available for the genomes of three potyviruses, namely the aphid non-transmissible isolate of PPV (PPV-NAT; Maiss et al., 1989), the highly
Fig. 2. Comparisons of 3' untranslated regions from strains of distinct potyviruses. The sequences were aligned for maximum identity using the ALIGN program of NBRF with a matrix bias of 0 and a break penalty of 5. Sequences are shown only where they differ from the full sequence, and deletions are indicated by dots. The nucleotide numbers (including deletions) are shown at the end of each line. (a) NAT and HAT strains of TEV; (b) D, R, NAT and aphid-transmissible (AT) strains of PPV; (c) I and N strains of PVY and PeMV; (d) WMV 2 and SMV-N. The sources of sequence data are given in Table 1.

Examination of these sequences reveals that the 3' non-coding regions of distinct viruses differ in length (189 to 475 nucleotides) and display no significant sequence homology (Table 1). The degree of homology, ranging from 39 to 53%, is comparable to that obtained when the 3' untranslated regions of unrelated viruses from other plant virus groups are compared with aphid-transmissible isolate of TEV (TEV-HAT; Allison et al., 1986) and tobacco vein mottling virus (TVMV; Domier et al., 1986) with sequences for 3' coding and non-coding regions available for the D, AT and R strains of PPV (Ravelonandro et al., 1988; E. Maiss, personal communication; Lain et al., 1988), the NAT strain of TEV (Allison et al., 1985b), Johnson grass mosaic virus (JGMV) (Gough et al., 1987), PeMV (Dougherty et al., 1985) and the I and N strains of PVY (Rosner & Raccah, 1988; van der Vlugt et al., 1989).
Table 1. Percentage nucleotide sequence homology between coat protein-coding and 3' untranslated regions of potyvirus genomes*

<table>
<thead>
<tr>
<th></th>
<th>PPV-D</th>
<th>PPV-AT</th>
<th>PPV-NAT</th>
<th>PPV-R</th>
<th>JGMV-JG</th>
<th>PVY-I</th>
<th>PVY-PeMV</th>
<th>WMV 2</th>
<th>SMV-N</th>
<th>SMV-V</th>
<th>TVMV</th>
<th>TEV-NAT</th>
<th>TEV-HAT</th>
</tr>
</thead>
<tbody>
<tr>
<td>PPV-D</td>
<td>-</td>
<td>97</td>
<td>97</td>
<td>98</td>
<td>58</td>
<td>63</td>
<td>63</td>
<td>62</td>
<td>61</td>
<td>61</td>
<td>60</td>
<td>60</td>
<td>62</td>
</tr>
<tr>
<td>PPV-AT</td>
<td>99</td>
<td>-</td>
<td>99</td>
<td>99</td>
<td>59</td>
<td>64</td>
<td>64</td>
<td>64</td>
<td>62</td>
<td>61</td>
<td>60</td>
<td>60</td>
<td>64</td>
</tr>
<tr>
<td>PPV-NAT</td>
<td>98</td>
<td>98</td>
<td>-</td>
<td>99</td>
<td>58</td>
<td>63</td>
<td>64</td>
<td>64</td>
<td>62</td>
<td>61</td>
<td>60</td>
<td>60</td>
<td>64</td>
</tr>
<tr>
<td>PPV-R</td>
<td>99</td>
<td>99</td>
<td>99</td>
<td>-</td>
<td>58</td>
<td>63</td>
<td>64</td>
<td>64</td>
<td>62</td>
<td>62</td>
<td>60</td>
<td>60</td>
<td>64</td>
</tr>
<tr>
<td>JGMV-JG</td>
<td>46</td>
<td>46</td>
<td>46</td>
<td>45</td>
<td>-</td>
<td>60</td>
<td>59</td>
<td>60</td>
<td>62</td>
<td>64</td>
<td>59</td>
<td>59</td>
<td>60</td>
</tr>
<tr>
<td>PVY-I</td>
<td>51</td>
<td>49</td>
<td>46</td>
<td>46</td>
<td>46</td>
<td>-</td>
<td>88</td>
<td>92</td>
<td>62</td>
<td>60</td>
<td>62</td>
<td>57</td>
<td>63</td>
</tr>
<tr>
<td>PVY-N</td>
<td>43</td>
<td>44</td>
<td>44</td>
<td>43</td>
<td>44</td>
<td>83</td>
<td>-</td>
<td>83</td>
<td>64</td>
<td>61</td>
<td>62</td>
<td>63</td>
<td>64</td>
</tr>
<tr>
<td>PVY-PeMV</td>
<td>46</td>
<td>44</td>
<td>48</td>
<td>47</td>
<td>43</td>
<td>92</td>
<td>83</td>
<td>-</td>
<td>63</td>
<td>60</td>
<td>64</td>
<td>58</td>
<td>61</td>
</tr>
<tr>
<td>WMV 2</td>
<td>45</td>
<td>44</td>
<td>40</td>
<td>44</td>
<td>44</td>
<td>42</td>
<td>42</td>
<td>-</td>
<td>82</td>
<td>58</td>
<td>59</td>
<td>62</td>
<td>62</td>
</tr>
<tr>
<td>SMV-N</td>
<td>47</td>
<td>45</td>
<td>42</td>
<td>43</td>
<td>43</td>
<td>43</td>
<td>43</td>
<td>50</td>
<td>78</td>
<td>-</td>
<td>59</td>
<td>59</td>
<td>62</td>
</tr>
<tr>
<td>SMV-V</td>
<td>42</td>
<td>41</td>
<td>44</td>
<td>43</td>
<td>48</td>
<td>47</td>
<td>45</td>
<td>44</td>
<td>44</td>
<td>45</td>
<td>-</td>
<td>63</td>
<td>63</td>
</tr>
<tr>
<td>TVMV</td>
<td>41</td>
<td>41</td>
<td>39</td>
<td>42</td>
<td>47</td>
<td>47</td>
<td>41</td>
<td>45</td>
<td>47</td>
<td>39</td>
<td>45</td>
<td>-</td>
<td>60</td>
</tr>
<tr>
<td>TEV-NAT</td>
<td>42</td>
<td>42</td>
<td>42</td>
<td>43</td>
<td>50</td>
<td>48</td>
<td>50</td>
<td>49</td>
<td>42</td>
<td>47</td>
<td>48</td>
<td>52</td>
<td>-</td>
</tr>
<tr>
<td>TEV-HAT</td>
<td>41</td>
<td>42</td>
<td>42</td>
<td>43</td>
<td>49</td>
<td>48</td>
<td>50</td>
<td>49</td>
<td>42</td>
<td>47</td>
<td>48</td>
<td>53</td>
<td>99</td>
</tr>
</tbody>
</table>

* The sources of sequence data were: PPV-D (Ravelonandro et al., 1988); PPV-NAT and PPV-AT (Maiss et al., 1989; E. Maiss, personal communication); PPV-R (Lain et al., 1988); JGMV-JG (Gough et al., 1987); PVY-I (Rosner & Raccah, 1988); PVY-N (van der Vlugt et al., 1989); PeMV (Dougherty et al., 1985); SMV-N (Eggenberger et al., 1989); SMV-V (Gunyuzlu et al., 1987); TVMV (Domier et al., 1986); TEV-NAT and TEV-HAT (Allison et al., 1985a, b, 1986). Above the diagonal, coat protein-coding region sequence identities; below the diagonal, 3' untranslated region sequence identities.
potyviruses and are probably close to the value expected for coincidental matching. In contrast, the 3' untranslated regions of related strains are very similar in length and nucleotide sequence (Table 1, Fig. 2).

We have previously commented (Shukla et al., 1986, 1988c; Shukla & Ward, 1988) on the close structural relationship between the coat proteins of PeMV and strains of PVY, and the possibility that PeMV is a strain of PVY. Comparison of the 3' untranslated region of PeMV (Dougherty et al., 1985) with that published for PVY-I (Rosner & Raccah, 1988) shows that the 3' untranslated sequences are 92% homologous (Table 1), supporting our contention that they are closely related genetically and should be considered strains of one virus. In a more recent paper, van der Vlugt et al. (1989) have determined the nucleic acid sequence of the coat protein-coding and 3' untranslated region of PVY-N and have reached the same conclusion. Moreover, they point out that PeMV was originally classified as PVY-S, the speckling strain of PVY (Zitter, 1972).

As shown in Table 1, the 3' untranslated regions of RNA of WMV 2 and SMV-N are as similar in sequence as those of some strains of PVY, which strongly supports the suggestion that WMV 2 and SMV-N are strains of the one virus. Yu et al. (1989) reached the same conclusion, based on the homology observed for the coat protein sequences, and suggested that SMV-N could be renamed the soybean N strain of WMV (WMV-SN). However, more information is required on the range of viruses causing mosaic disease of soybean before deciding whether SMV-N and WMV 2 should be considered strains of SMV or strains of WMV 2.

There are numerous reports of the use of cDNA hybridization to detect strains of plant viruses (Hull, 1984). However the use of randomly primed cDNAs to the RNAs of some viruses have resulted in strong hybridization with heterologous viruses, including viruses from different groups (Koenig et al., 1988).

The sequence of the 3' untranslated region of potyvirus RNA may have great value as a marker for distinguishing viruses and strains and can be readily obtained from cDNA clones generated by oligo(dT)-primed synthesis on a viral RNA template. In addition, synthetic oligonucleotides, Taq polymerase-generated fragments (Saiki et al., 1988) or cloned cDNA from the 3' untranslated region should find ready application as sensitive probes (Landegren et al., 1988) for potyvirus detection and classification.

We are most grateful to Drs R. N. Beachy (St Louis), E. Maiss (Basel), S. A. Tatin (Blacksburg) and R. Goldbach (Wageningen) for their generous provision of sequence data before publication. We thank Mr N. Bartone for synthetic oligonucleotide probes, Ms S. L. Tracy for excellent technical assistance, Ms L. Monarch for the photographs and Ms B. Wood for typing the manuscript. This work was supported by the Rural Credits Development Fund of the Reserve Bank of Australia.

REFERENCES


3' Non-coding sequences and potyvirus taxonomy


(Received 14 March 1989)