Complete nucleotide sequence of pepper huasteco virus: analysis and comparison with bipartite geminiviruses

I. Torres-Pacheco,† J. A. Garzón-Tiznado,† L. Herrera-Estrella and R. F. Rivera-Bustamante*

Departamento de Ingeniería Genética, Centro de Investigación y de Estudios Avanzados del I.P.N., Unidad Irapuato, Apartado Postal no. 629, Irapuato, Gto. México

The complete nucleotide sequence of pepper huasteco geminivirus (PHV) isolated in Northern Mexico was determined. The PHV genome consists of two circular ssDNA molecules of 2631 bases (PHV A) and 2589 bases (PHV B). PHV has a genome organization typical of a bipartite geminivirus with four open reading frames (ORFs) (AR1, AL1, AL2 and AL3) in component A and two (BR1 and BL1) in component B. An unexpected ORF was found in the complementary sense strand of PHV A. This ORF, termed AL5, is found entirely inside, but in the opposite orientation to AR1 (encoding the coat protein). AL5 shows some homology to equivalent but smaller ORFs predicted in other geminiviruses. Phylogeny trees based on pairwise comparisons of AR1, AL2, AL3, BL1 and BR1 predicted proteins placed PHV among the western hemisphere geminiviruses. A phylogeny tree based on AL1 (replicase-encoding ORF), on the other hand, placed PHV with eastern hemisphere geminiviruses, i.e. African cassava mosaic virus and the Sardinia and Israel isolates of tomato yellow leaf curl virus. Possible mechanisms for the ‘hybrid or transition nature’ of PHV are discussed.

Introduction

Geminiviruses are a group of plant viruses that have a small ‘twinned’ isometric particle and a genome containing one or two ssDNA molecules. Most geminiviruses can be classified into four subgroups according to their host-range, insect vector and genome organization.

All bipartite geminiviruses described so far show similar genomic organization with four open reading frames (ORFs) in component A and two ORFs in component B. Component A has one ORF (AR1) in the positive or viral sense and three ORFs (AL1, AL2 and AL3) in the negative or complementary sense. In component B, one ORF is found in each orientation (BR1 and BL1). In both cases, the ORFs extend in opposite directions from a 180 to 200 base common region shared by both components. Some sequences in this common region have been reported to interact with the respective putative replicase (AL1 protein) and are unique for each geminivirus except for a conserved 30 base sequence element with a potential hairpin structure (Lazarowitz, 1992).

Genetic analyses have suggested the possible function for some ORF products. AR1 ORF is considered to be the coat protein gene, although to our knowledge direct evidence of this has been reported for only two cases [e.g. sequencing of the African cassava mosaic virus (ACMV) coat protein and immunological identification of the tomato golden mosaic virus (TGMV) AR1 gene product after expression in Escherichia coli] (Stanley & Gay, 1986; Murayama et al., 1991). The products of the AL1, AL2 and AL3 ORFs have been implicated in several steps of the replication process (Elmer et al., 1988; Sunter et al., 1990). In addition, the AL2 protein has been reported as a possible trans-activator needed for expression of AR1 and BR1 genes (Sunter & Bisaro, 1991). The ORFs found in component B, BL1 and BR1, on the other hand, have been suggested as being responsible for the movement of the virus throughout the plant (Lazarowitz, 1992; von Arnim & Stanley, 1992).

Recently we reported the cloning of both components of pepper huasteco virus (PHV), a new whitefly-transmitted bipartite geminivirus affecting pepper crops in northern Mexico (Garzón-Tiznado et al., 1993). This non-mechanically transmissible geminivirus has been detected in several horticultural areas in the country, usually in association with other non-characterized geminiviruses. These complex viral diseases are currently the most important phytopathological problem in Mexican horticulture. Monomeric clones of PHV DNAs (A + B) are infectious with a 60 to 70% efficiency when
introduced into pepper plants by a biolistic procedure. Neither A nor B DNA produces any symptoms when inoculated alone (Garzón-Tiznado et al., 1993).

Here, we report the complete nucleotide sequence of PHV DNAs. We also compared PHV predicted proteins with those of previously reported geminiviruses in an attempt to address the question of the possible origin of this important geminivirus problem in Mexico.

**Methods**

**Sequencing of PHV.** All standard procedures utilized during the cloning and sequencing of viral DNA were according to Sambrook et al. (1989). Viral inserts contained in plasmids pIGV21 and pIGV22 were sequenced using Sequenase 2.0 kits from United States Biochemicals according to the manufacturer’s instructions. In some cases, restriction fragments from pIGV21 and pIGV22 were subcloned into Bluescript SK+ (Stratagene) for sequencing. Synthetic oligonucleotides were used as primers to sequence regions where no convenient restriction sites could be found. Oligonucleotides were produced in a DNA Synthesizer Model 381A from Applied Biosystems.

**Computer analysis.** Sequences were assembled and analysed using GeneWorks 2.2 and PCGene 6.5 (IntelliGenetics) and MacDNAsis 2.0 (Hitachi) software packages. The following geminivirus sequences were obtained directly from the EMBL and GenBank databases and used for comparisons: ACMV (X17095, X17096) (Stanley & Gay, 1986); abutilon mosaic virus, AbMV (X15983, X15984) (Frischmuth et al., 1990); beet curly top virus, BCTV (X04144) (Stanley et al., 1986); bean golden mosaic virus, Puerto Rico isolate, BGMV-PR (D00200, D00201) (Howarth et al., 1985); Brazil isolate, BGMV-BZ (M88686, M88687), and Guatemala isolate, BGMV-GA (M91604, M91605) (Gilbertson et al., 1991); bean dwarf mosaic virus, BDMV (M88179, M88180) (Hidayat et al., 1993); potato yellow mosaic virus, PYMV (D00940, D00941) (Courts et al., 1991); TGMV (K02029, M73794) (Hamilton et al., 1984; von Arnim & Stanley, 1992); tomato yellow leaf curl virus, Sardinia, TYLCV-S (X61153) and Israel isolates, TYLCV-I (X15656) (Kheyr-Pour et al., 1991; Navot et al., 1991); squash leaf curl virus, SqLCV (M38182, M38183) (Lazarowitz & Lazdins, 1991).

Phylogeny trees were produced with the GeneWorks software package which uses the UPGMA (unweighted pair group method with arithmetic mean) procedure (Nei, 1987). The program algorithm does pairwise comparisons and first clusters the two most similar sequences and calculates the average similarity of those sequences. The program then clusters the next two most similar objects.

**Results and Discussion**

**PHV genomic organization**

Sequence analysis revealed PHV as a typical bipartite geminivirus whose genome is divided in two single-stranded molecules termed PHV A and PHV B. Components A and B are 2631 and 2589 bases long, respectively. Fig. 1 presents the nucleotide sequence of both components. Also shown is the 30 base stem–loop region shared by all geminiviruses infecting dicotyledons (GGCCAT<sub>7</sub>AACCCGT<sub>7</sub>/A<sub>7</sub>/TAAATATTACCGG<sub>7</sub>/TGGCC). Nucleotide 1 in both cases corresponds to the first base of the 30 base conserved region. This region contains the typical loop structure TAATATTAC, found in hairpin structures of all monocotyledon- and dicotyledon-infecting geminiviruses, and is believed to be involved in virus replication (Lazarowitz, 1987).

Similarly to all previously reported geminiviruses, the base composition of PHV showed that thymidine is present at the highest percentage, whereas cytosine is the lowest in occurrence. Individually, the base composition of each component was: PHV A, 26.25 % A, 19.45 % C, 23.82 % G and 30.47 % T; PHV B, 28.23 % A, 18.35 % C, 21.71 % G and 31.71 % T.

The genomic organization of PHV is similar to that of all bipartite geminiviruses. Fig. 2 illustrates the coding capacity of PHV. Component A contains five putative ORFs and component B contains two. The ORFs are found in both orientations: viral or positive-sense and complementary or negative-sense. The nomenclature used for PHV ORFs is based on analogies with previously reported geminiviruses.

**Analysis of PHV ORFs**

Computer analysis suggests a subdivision for bipartite geminiviruses into western (WH) and eastern (EH) hemisphere geminivirus subgroups. WH geminiviruses included BGMV, TGMV and AbMV whereas EH geminiviruses were represented by ACMV and TYLCV (Howarth & Vandemark, 1989). The analysis was based on pairwise homologies of coat proteins (AR1-encoded) and the product of the AL1 ORF (putative replicase). To determine the situation of PHV among these subgroups, the predicted protein from PHV AR1 was compared pairwise with the equivalent proteins from several other geminiviruses (Table 1). PHV AR1 showed a high degree of similarity (82 to 90%) with the WH geminivirus subgroup, closest to the Guatemalan isolate of BGMV (BGMV-GA) (Gilbertson et al., 1991). The degree of similarity decreased (68 %) when the PHV AR1 protein was compared with the equivalent proteins from members of the EH subgroup: ACMV and the Sardinia and Israel isolates of TYLCV (TYLCV-S and TYLCV-I, respectively) (Kheyr-Pour et al., 1991; Navot et al., 1991). A phylogeny tree based on pairwise homology data showed an evident grouping of the viruses placing PHV, as expected, with the WH cluster (Fig. 3a). Alignments of coat proteins from all viruses included in Fig. 3 revealed several highly conserved domains (data not shown). In contrast, no significant homology was detected when the PHV AR1 protein was compared with coat proteins from geminiviruses transmitted by leafhoppers, e.g. BCTV (29 %), maize steak virus and wheat dwarf virus (data not shown).

A similar analysis was applied to the putative replicase, encoded by the AL1 ORF, and several interesting
Fig. 1. Nucleotide sequence of the viral sense strand of both components of PHV (A and B). The underlined regions (30 bases) correspond to the consensus found in all bipartite geminiviruses. In both cases, nucleotide 1 corresponds to the first base of the 30-mer consensus.
Fig. 2. Proposed genomic organization for PHV. Arrows represent ORFs in both orientations. CR represents the 180 base common region for both components which contains the 30 base stem-loop structure found in all bipartite geminiviruses. ORF coordinates are: AR1, 230 to 982; AL1, 1253 to 1467; AL2, 1543 to 1130; AL3, 1380 to 985; AL5, 853 to 341; BR1, 447 to 1214; BL1, 2158 to 1280.

Table 1. Amino acid sequence similarities (%) between putative translation products of PHV and selected geminiviruses

<table>
<thead>
<tr>
<th></th>
<th>AR1</th>
<th>AL1</th>
<th>AL2</th>
<th>AL3</th>
<th>BR1</th>
<th>BL1</th>
</tr>
</thead>
<tbody>
<tr>
<td>AbMV</td>
<td>82</td>
<td>65</td>
<td>54</td>
<td>67</td>
<td>61</td>
<td>63</td>
</tr>
<tr>
<td>ACMV</td>
<td>68</td>
<td>65</td>
<td>50</td>
<td>48</td>
<td>37</td>
<td>46</td>
</tr>
<tr>
<td>BCTV</td>
<td>29</td>
<td>61</td>
<td>30</td>
<td>43</td>
<td>NA†</td>
<td>NA</td>
</tr>
<tr>
<td>BDMV</td>
<td>87</td>
<td>67</td>
<td>59</td>
<td>66</td>
<td>67</td>
<td>82</td>
</tr>
<tr>
<td>BGMV-PR</td>
<td>84</td>
<td>64</td>
<td>48</td>
<td>67</td>
<td>68</td>
<td>82</td>
</tr>
<tr>
<td>BGMV-BZ</td>
<td>86</td>
<td>61</td>
<td>50</td>
<td>67</td>
<td>67</td>
<td>76</td>
</tr>
<tr>
<td>BGMV-GA</td>
<td>90</td>
<td>66</td>
<td>49</td>
<td>68</td>
<td>67</td>
<td>84</td>
</tr>
<tr>
<td>PYMV</td>
<td>87</td>
<td>63</td>
<td>52</td>
<td>66</td>
<td>65</td>
<td>81</td>
</tr>
<tr>
<td>SqLCV</td>
<td>88</td>
<td>48</td>
<td>51</td>
<td>66</td>
<td>53</td>
<td>77</td>
</tr>
<tr>
<td>TGMV</td>
<td>86</td>
<td>66</td>
<td>54</td>
<td>67</td>
<td>68</td>
<td>78‡</td>
</tr>
<tr>
<td>TYLCV-I</td>
<td>68</td>
<td>68</td>
<td>50</td>
<td>50</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>TYLCV-S</td>
<td>68</td>
<td>69</td>
<td>49</td>
<td>48</td>
<td>NA</td>
<td>NA</td>
</tr>
</tbody>
</table>

* All values were calculated using the MacDNAsis program. Other programs (GeneWorks and UWGCG) gave similar results.
† NA, Not available; TYLCV and BCTV do not have component B.
‡ For TGMV BL1 the sequence used was as reported by von Arnim & Stanley (1992) because the sequence originally reported by Hamilton et al. (1984) contained a mistake that produced a truncated BL1 protein.

observations were obtained (Table 1 and Fig. 3 b). First, PHV presented the highest similarity with both isolates of TYLCV, Sardinia, 69%, and Israel, 68%. The similarity to the WH geminivirus cluster was slightly lower (61% to 67%). When an AL1 protein-based phylogeny tree was constructed, PHV clearly clustered with the EH geminiviruses (ACMV and TYLCV) (Fig. 3 b). Alignment of AL1 predicted proteins supported the classification displayed in the phylogeny tree. Fig. 4 shows a domain of 40 amino acids for several replicases that contains some residues found in EH geminiviruses (ACMV, TYLCV-I and TYLCV-S) and present also in PHV, but not found in any other WH geminivirus. In addition to this domain, other residues conserved between PHV and EH geminiviruses are V(218), C(241), T(299), F(301) and P(307). A more detailed analysis, however, showed that the AL1 protein also of PHV has some residues found only in WH geminiviruses.
Complete nucleotide sequence of PHV 2229

[PEPW(177 to 180), N(191), C(275)]. Although the AL1 protein comparison suggests a close relationship between PHV and EH geminiviruses, the analyses of the other ORFs as well as the fact that PHV AL1 also possesses some WH characteristics do not allow a definitive overall clustering of PHV. These results propose PHV as a hybrid or transition geminivirus between the WH and EH types.

Trees obtained using other software packages (e.g. PCGene CLUSTAL program) showed minor variations for the position of some geminiviruses inside the main clusters. However in all cases the clustering of PHV remained identical. In any case, the main purpose of the tree is to illustrate similarities between sequences encoded by a specific ORF. Implications for an evolutionary path will require more rigorous analysis.

In contrast to the results found with the AR1 protein where the similarity between PHV and SqLCV was one of the highest (88%), the similarity between products of the PHV AL1 and SqLCV AL1 ORFs was the lowest (48%), even lower than the value obtained with equivalent proteins from BCTV (61%) (Stanley et al., 1986). When SqLCV was included in the AL1 phylogeny tree it was found to stand by itself forming a separate branch, whereas in the AR1 phylogeny tree it was clearly clustered with the WH geminiviruses. BCTV, which is a subgroup II geminivirus transmitted by leafhoppers, was included in the comparisons used for the AL1 tree as an internal control of the method, and clustered with the WH bipartite subgroup as reported previously (Howarth & Vandemark, 1989). Phylogeny trees based on similarities between products of ORFs AL2 and AL3 displayed similar results to the AR1 phylogeny tree. The clustering of PHV was similar in all three cases and the only difference was the degree of similarity that PHV showed with the rest (Table 1).

Computer analysis revealed two possible initiation codons, in the same reading phase, for the PHV AL2 ORF. The first one generates a protein of 208 amino acids. The second initiation codon is found 210 bases downstream producing a protein with only 138 amino acids. Sequence and size comparisons with AL2 proteins from other geminiviruses suggested that the smaller protein is the actual product. Point mutation experiments are being carried out to dismiss the possibility of any initiation of translation at the first ATG codon.

An unexpected fifth ORF (ORF AL5) was detected in component A. The nomenclature of AL5 (instead of AL4) was suggested to avoid confusion with the previously described AL4 ORF for other geminiviruses. This ORF is located entirely inside ORF AR1 (coat protein) but in the opposite orientation. At present it is not known whether this ORF is transcribed and translated. To verify whether this situation is unique for PHV or whether similar proteins could be found in other geminiviruses, the sequences of all previously reported geminiviruses were analysed. The survey included ORFs producing proteins smaller than 10000 kDa, which has been a common size limit for geminiviral proteins. Several viruses (BGMV, BDMV and PYMV) had similar but smaller ORFs. Although the overall pairwise similarity was relatively low in all cases (50%), computer alignments of the predicted AL5 proteins displayed some conserved regions. A genetic study is also underway to determine the importance of AL5.

For the BL1 ORF, PHV showed the largest degree of similarity (84%) to BGMV-GA and the lowest to ACMV (42%). A similar situation, although with a lower degree of similarity, was found for BR1 where BGMV-PR, TGMV (68%) and BDMV, BGMV-BZ, BGMV-GA (67%) showed the higher degree of similarity to PHV and the lowest to ACMV (37%). Phylogeny trees placed all WH geminiviruses in one homogeneous cluster leaving only ACMV on an independent branch (data not shown).

The AR1, AL1, BR1 and BL1 ORFs have typical eukaryotic promoters with CAAT and TATA boxes and a cap signal. The distances between TATA boxes and the cap signal are similar although not identical. The position of the CAAT box showed more divergence. Perhaps the nearby stem–loop structure (e.g. for AL1) could have some effect on the structure of the promoter. The position of the poly(A) signal was also different since the one for AR1 is virtually on the stop codon whereas BR1 has its signal a few bases downstream. This might be a reflection of the fact that the AR1 stop codon overlaps with that of AL3 resulting in the absence of an intergenic region (see Fig. 2). Component B, on the other hand, presents an intergenic region of 59 bases. The poly(A) signal for AL1 was not clearly identified. In addition, no typical regulatory sequences could be found for ORFs AL2, AL3 and AL5. The expression of these overlapping ORFs has yet to be determined.

The overall analysis of the PHV genome showed it to be a WH geminivirus. However, the fact that the PHV AL1 protein is more closely related to the equivalent proteins from ACMV and TYLCV suggests the possibility of some type of recombination between geminiviruses. There are several non-characterized geminiviruses in the same geographical area infecting pepper and tomato crops (Brown et al., 1986; Brown & Nelson, 1988, 1989; Brown & Poulos, 1990). It is possible that a virus similar to TYLCV, at the molecular level, might be present in the area. Mixed infections are common and several diseases have been reported to be caused by a mixture of geminiviruses (Brown et al., 1989; Brown & Nelson, 1989). The presence of several geminiviruses in the same plant provides the opportunity for recombination. This
opportunity is increased in this area since possible hosts for the viruses are found in all seasons. The relatively low degree of similarity (68%) between PHV and TYLCV replicases suggests that this possible recombination event is not a recent one, and that the viruses have had enough time to diverge to the present level. Several experiments in which components of bipartite geminiviruses were exchanged (pseudorecombination) have been recently reported (Gilbertson et al., 1993; von Arnim & Stanley, 1992). This pseudorecombination mechanism can now be added to the intermolecular recombination reported for ACMV (Etessami et al., 1989; Klinkenberg et al., 1989; Stanley & Townsend, 1986). Both mechanisms, independently or acting in a concerted manner, might be important in nature for the generation of new viruses. The unexpected clustering of SqLCV also supports a recombination hypothesis. SqLCV AR1 is highly similar to those of WH geminiviruses whereas SqLCV AL1 is clustered completely on its own.

An alternative explanation is simply that PHV, ACMV and TYLCV share a closer common ancestor and have diverged in adapting themselves to the surrounding environment (e.g. vectors and host plants in the respective geographical area). The characterization of more geminiviruses from the same area where PHV was isolated could provide new information about their phylogenetic relationships.

That Mexico is a centre of diversity for both pepper and tomatoes might be of importance for the number of geminiviruses being found in this area. Nevertheless, the sudden outbreak of diseases caused by geminiviruses in the late 1980s still remains unexplained.

We thank B. Jimenez for her generous help during the sequencing of PHV and A. Alvarez for critical reading of the manuscript. We would also like to thank the anonymous referees for their critical comments towards improvement of the manuscript. This research was supported by a grant from CONACyT-México (1150-N9202) to R.F.R.-B. I.T.-P. also acknowledges fellowship support from CONACyT.

References


(Received 15 March 1993; Accepted 8 June 1993)