Evidence for diversifying selection in *Potato virus Y* and in the coat protein of other potyviruses

Benoît Moury,1, 2 Caroline Morel,1 Elisabeth Johansen2 and Mireille Jacquemond1

1 Station de Pathologie Végétale, Institut National de la Recherche Agronomique, F-84143 Montfavet cedex, France
2 Danish Institute of Agricultural Sciences, Biotechnology Group, Thorvaldensvej 40 1, DK-1871 Frederiksberg C, Denmark

The modes of evolution of the proteins of *Potato virus Y* were investigated with a maximum-likelihood method based on estimation of the ratio between non-synonymous and synonymous substitution rates. Evidence for diversifying selection was obtained for the 6K2 protein (one amino acid position) and coat protein (24 amino acid positions). Amino acid sites in the coat proteins of other potyviruses (*Bean yellow mosaic virus*, *Yam mosaic virus*) were also found to be under diversifying selection. Most of the sites belonged to the N-terminal domain, which is exposed to the exterior of the virion particle. Several of these amino acid positions in the coat proteins were shared between some of these three potyviruses. Identification of diversifying selection events in these different proteins will help to unravel their biological functions and is essential to an understanding of the evolutionary constraints exerted on the potyvirus genome. The hypothesis of a link between evolutionary constraints due to host plants and occurrence of diversifying selection is discussed.

Introduction

*Potato virus Y* (PVY) is the type member of the genus *Potyvirus* in the family *Potyviridae* (Shukla et al., 1994). The monopartite genome is composed of a single-stranded, positive-sense RNA molecule of about 9.7 kb. During the infection process, this RNA is translated into a large precursor polyprotein that is cleaved co- and post-translationally into 10 mature proteins (Dougherty & Carrington, 1988; Riechmann et al., 1992). PVY is the cause of major diseases in solanaceous crops including potato, tobacco, pepper and tomato, and also infects many solanaceous and non-solanaceous weeds. The symptoms induced by PVY and its host range are highly variable. These traits are used for the classification of PVY strains into different groups. For example, strains from potato are classified into three main subgroups (O, N and C) according to symptomatology and serology (with monoclonal antibodies).

PVY subgroups obtained by phylogenetic analyses of sequence data correlate only partially with the biological traits of the strains. No molecular determinants of symptomatology or host range have yet been isolated, with the exception of the NLa proteinase of PVY, required for elicitation of the Ry-mediated resistance in potato (Mestre et al., 2000). Knowledge of modes of evolution in the different genome regions and in the different phylogenetic lineages would help to identify symptomatology and host-range molecular determinants as well as external (including plant hosts, vector insects and physical environment) and internal (biological functions of PVY proteins) evolutionary constraints exerted on PVY.

The ratio ($\omega$) of non-synonymous (amino acid-altering) to synonymous (silent) substitution rates provides an estimate of the selective pressure at the protein level (Kimura, 1983). A value of $\omega > 1$ means that non-synonymous mutations offer fitness advantages to the protein and have higher fixation probabilities than synonymous mutations (diversifying or positive selection). On the other hand, $\omega$ values close to 0 mean that the protein is essentially conserved at the amino acid level (purifying or negative selection) and $\omega = 1$ corresponds to neutral evolution. Calculation of $\omega$ as a mean over all codons in the gene and over the entire evolutionary time that separates the sequences provides only limited information and impedes the detection of diversifying selection events in many cases. Indeed, many proteins appear to be under purifying
selection most of the time (Li, 1997) and a large proportion of their amino acids are largely invariable (with $\omega$ close to 0) due to structural constraints. When knowledge of the functional domains of the protein is not available, or when only a few codons or a few lineages undergo diversifying selection, a better approach is to devise statistical models that allow for heterogeneous $\omega$ ratios among codons or among lineages (Yang, 1998; Nielsen & Yang, 1998). This can be achieved by maximum-likelihood methods (Yang, 1997; Yang et al., 2000).

Most of the PVY genomic sequences available are from potato strains. To obtain more information about the diversity and evolution of PVY, we have cloned and sequenced the genomes of two PVY strains, one from tomato and the other from pepper. Using a maximum-likelihood algorithm, we looked for positive selection events in each of the PVY-encoded proteins. Finally, we also examined the presence of substitution saturation and the occurrence of recombination events among the sequences, since these can interfere with the detection of positive selection events. Amino acid sites in two PVY proteins [the 6K2 and coat or capsid proteins (CP)] were found to be under positive selection. We also examined the CPs of other potyviruses to confirm our findings and to evaluate the involvement of the CP in host adaptation.

Methods

Nucleotide sequences. PVY nucleotide sequences available in the GenBank database were used (Table 1). Strain LYE84.2 was isolated from tomato and has been repeatedly passed through Lycopersicon spp. genotypes (Legnani, 1995). Strain SON41 was isolated from the weed Solanum nigrum and has been repeatedly passed through Capsicum annum cultivar ‘Florida VR2’ (Gebré-Selassié et al., 1985). Degenerate oligonucleotide primers were defined in regions conserved among PVY strains. Overlapping cDNA fragments covering the whole genome of SON41 were amplified by PCR using Taq DNA polymerase and cloned in pCR2.1-TOPO (Invitrogen). Nucleotide sequence reactions (ABI Prism BigDye Terminator cycle sequencing ready reaction kit, Applied Biosystems) were analysed on an automated ABI 377XL DNA sequencer (Applied Biosystems). Specific PCR primers were designed and overlapping PCR fragments were amplified with Pfu DNA polymerase (Stratagene) and cloned in pCR-blunt-II-TOPO (Invitrogen). The nucleotide sequences of these clones were determined as described previously.

Two strategies were used to obtain the cDNA of the strain LYE84.2. A 3.8 kb clone covering the 3’-terminal region was obtained using the procedure of Gubler & Hoffman (1983). The remaining part of the virus was amplified by PCR and cloned in pZEoO-2 (Invitrogen). Sequencing was performed by Genome Express (Grenoble, France). The nucleotide sequences at the 5’ termini of the SON41 and LYE84.2 RNAs have not been determined. Therefore, a consensus primer corresponding to the first 28 nucleotides was used as the 5’ primer for cDNA amplification of this part of the genome.

Nucleotide sequences of the coat proteins of Bean yellow mosaic virus (BYMV), Lettuce mosaic virus (LMV), Potato virus A (PVA), Potato virus V (PVV), Turnip mosaic virus (TuMV) and Yarm mosaic virus (YMV) were obtained from the GenBank database.

Sequence alignments. Because the data contained variable numbers of sequences along the PVY genome (Table 1), each of the 10 proteins resulting from the cleavage of the PVY polyprotein was analysed separately, unless indicated. The nucleotide sequences were

Table 1. PVY sequences used for phylogenetic analyses

Sequences corresponding to accession numbers M95491, U09509, X12456 and X97895 and to strains SON41 (AJ439544) and LYE84.2 (AJ439545) are full-length and hence are common to all analyses. Accession numbers or strain names of additional sequences are shown in the table.

<table>
<thead>
<tr>
<th>Protein</th>
<th>Sequences analysed (n)</th>
<th>Accession numbers or names of strains*</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>11</td>
<td>M37180, M38377, N-Nysat, NN†, N-Wilga†</td>
</tr>
<tr>
<td>HC-Pro</td>
<td>8/12‡</td>
<td>AF166115, M37180, M38377, Z50041, Z50042, Z50043</td>
</tr>
<tr>
<td>P3</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>6K1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>CI</td>
<td>8</td>
<td>SK1663, SK1664</td>
</tr>
<tr>
<td>6K2</td>
<td>8</td>
<td>D12539, Z29526</td>
</tr>
<tr>
<td>Nla VPG</td>
<td>8</td>
<td>D12539, Z29526</td>
</tr>
<tr>
<td>Nla Pro</td>
<td>7</td>
<td>D12539</td>
</tr>
<tr>
<td>NIb</td>
<td>9/11‡</td>
<td>AF062999, D12539, U06789, U09508, Y16879</td>
</tr>
<tr>
<td>CP</td>
<td>25</td>
<td>AF012026, AF012027, AF012028, AF012029, AJ005639, D12570, U06789, U10378, X14136, X54058, X54636, X68221, X68222, X68223, X68224, X68225, X68226, X79305, P2§</td>
</tr>
</tbody>
</table>

* Most of the sequences are from potato strains. Underlined and bold entries are respectively from tobacco and pepper strains.
† From Marie-Jeanne Tordo et al. (1995) and personal communication.
‡ Some sequences are only partial. Several analyses were performed with different numbers of sequences and different sequence lengths.
§ From Fahalha et al. (1995); sequence entered by hand from the original publication.
aligned using the clustal method of aligning multiple sequences (MegAlign program) available with the DNASTAR package for the Apple Macintosh (version 1.02, DNASTAR Inc.). No gaps were observed, except for P3 and Nb. For P3, sequence X12456 shows a one-nucleotide deletion after nucleotide 509 followed by a one-nucleotide insertion after nucleotide 612. In the Nb coding region of sequence X12456, a similar local frame shift was identified by a nucleotide deletion after nucleotide 266 followed by a one-nucleotide insertion after nucleotide 311. Both of these local frame shifts resulted in a deduced amino acid sequence for X12456 that is highly divergent in these regions from the five other strains. It is highly probable that these one-nucleotide deletions and insertions are sequencing errors (C. Robaglia, personal communication) and they were considered as such in the following analyses. For Nb, sequence X12456 shows a two-amino acid codon insertion after amino acid codon 45 in comparison with all other sequences. These positions were discarded in the following analyses. For other potyviruses, only CP sequences that did not show any gaps were kept for further analysis.

Phylogenetic relationships among PVY strains. When the genetic distance increases between sequences, the number of observed transitions (substitutions between T and C and between A and G) relative to that of transversions (all other changes) gradually decreases due to substitution saturation. This can lead to detection of artefactual diversifying selection events. For instance, a lower transition/transversion ratio may reflect a positive, diversifying selective pressure on the region (and vice versa) since, overall, transitions are more likely to be synonymous changes than are transversions. Simulation studies show that phylogenetic information is essentially lost when the observed saturation is equal to or greater than half of full substitution saturation (Xia, 1999). The presence of substitution saturation was tested in our sequence data sets with DAMBE version 4.0.43 (Xia & Xie, 2001). Comparison of the saturation index expected when assuming half of full saturation with the observed saturation index was performed with a t test for each of the proteins. A plot of the transition and transversion ratios (estimated with DAMBE) versus the divergence of pairwise comparisons between the six full-length ORFs also offered a visual display of substitution saturation.

Phylogenetic reconstruction and evaluation was done using the neighbour-joining (NJ), Fitch and Margoliash and maximum-likelihood methods implemented in the MEGA software package (Felsenstein, 1993). One thousand bootstrap replications were performed to place confidence estimates on groups contained in the most-parsimonious unrooted trees. Nodes with low reliability (bootstrap support below 70%, Hillis & Bull, 1993) were collapsed and the subsequent tree topology was used for maximum-likelihood analyses of codon substitution.

Determination of recombination events. Because recombination events cause different regions of a gene to have different evolutionary histories, it is important to detect them before performing analyses of evolutionary rates and searches for positive selection. The initial search for recombination in PVY was conducted using split decomposition, a method that depicts all the shortest pathways linking sequences, including those that produce an interconnected network, as expected under recombination (Bandelt & Dress, 1992). This analysis was performed with the program SplitsTree2 (Huson, 1998). When putative recombinant sequences were detected, the use of Lard version 2.2 (Holmes et al., 1999) and SScan version 1.01 (Smith et al., 2000) allowed mapping of the recombination events and assessment of the statistical support for them. Further analyses were performed either discarding the putative recombinant strains or with sequences that did not contain any recombination breakpoints.

Tests for positive selection. In order to identify regions submitted to positive selection in the PVY genome, we employed paml version 3.0c, which estimates the occurrence of ω values greater than 1:0 and which has proven useful in the documentation of positive selection in other viruses (Yang et al., 2000). Using the program coxmol and the tree structures obtained for each of the 10 PVY-encoded proteins, we employed two categories of maximum-likelihood models. The first category of models aims to identify codon sites submitted to positive selection. These models assume that the non-synonymous/synonymous substitution rate ratio, ω, is constant across all the lineages but varies across the codons. Different evolutionary models were tested (Yang et al., 2000): a model (M0) assuming a constant ω ratio, models (M1 and M7) assuming that amino acid sites are either neutral or deleterious (ω ≤ 1) and models (M3 and M8) allowing the occurrence of positively selected sites (ω > 1). Several other models available in paml were not exploited because of their lack of power or because they converged particularly slowly, according to the recommendations of Yang et al. (2000). For each codon, the probability of observing the data was computed using the proportion of sites belonging to these different categories. The log likelihood is the sum of these probabilities over all codons in the sequence. When ω > 1 for some codons, the likelihood ratio of the two models to be compared (M3 versus M1 or M8 versus M7) tests whether the positive selection model fits the data significantly better than the null hypothesis: twice the difference in log likelihood between the two models is compared with a χ² distribution with n degrees of freedom, n being the difference between the numbers of parameters of the two models. An empirical Bayesian approach implemented in paml was used to infer to which category (neutral, deleterious or advantageous) each amino acid most likely belongs. Codons with posterior probabilities above 95% were considered significant.

The second category of models aims to identify lineages undergoing positive selection. They assume that ω is constant across all the codons but varies across the N different branches in the tree. Comparison of the scenario where each branch has a particular ω value (‘N-ratio’ model) to the null hypothesis, where ω is constant across all branches and all codons (model M0; Yang et al., 2000), constitutes a likelihood ratio test of the constancy of ω among evolutionary lineages. When branches with ω > 1 were obtained, we tested whether the ω values associated with those particular branches were significantly greater than 1:0 using Fisher’s exact test of homogeneity (Zhang et al., 1997).

Care was taken to run paml with several initial ω values (at least the four initial values 0:2, 1:0, 1:2 and 2:1) for each model to avoid local maxima. In almost all analyses, at least three initial values led to an identical maximum likelihood (that was larger than the one obtained with the fourth initial value, when this was different). When an ambiguity remained, three more initial values were tested and allowed to produce the maximum likelihood with satisfactory confidence.

A maximum-likelihood method, also implemented in paml, allowed the estimation of ω values in pairwise sequence comparisons (Yang & Nielsen, 2000). Significant departures from neutral evolution (ω estimates significantly greater than 1:0) were assessed with Fisher’s exact test of homogeneity (Zhang et al., 1997).

Results

Sequence data

We have cloned and sequenced the entire genomes of the LYE84.2 and SON41 PVY strains. No differences were observed between the clones obtained with Taq or Pfu DNA
Fig. 1. Plot of the transition rate (●) and transversion rate (○) against the Tamura & Nei (1993) (TN93) distance for pairwise comparisons of the six full-length PVY strains. Linearity is an indication of lack of saturation of the phylogenetic signal.

Fig. 2. Unrooted neighbour-joining phylogenetic tree of the CP gene of PVY. Bootstrap analysis was applied using 1000 bootstrap samples. Some bootstrap percentages at internal nodes are reported. The different subgroups of PVY strains from potato plants are indicated. Underlined sequences and sequences in boxes are respectively from tobacco and pepper strains.

Fig. 3. (A) Map of recombination breakpoints identified in the PVY genome. These recombination events were also shown in Revers et al. (1996) and Glais et al. (2002) except for the two recombination breakpoints in the Nb gene. (B) Z scores for total nucleotide identity scores of the Nb gene were calculated with SICan for two sets of comparisons: strains X12456 and M95491 (dashed line) and strains X12456 and D12539 (unbroken line), showing two putative recombination events for strain X12456. Z values above 3.67 are significant (P < 5 %).

AJ439545 and AJ439544, respectively. These are the first full-length sequences of non-potato strains and the first sequences of non-potato strains for several PVY proteins (P3, 6K1, CI, 6K2 and Nb).

Analysis of the phylogenetic signal
Evidence for lack of substitution saturation comes from linear regression of the transition and transversion rates versus the divergence of the sequences (Fig. 1). In the case of saturation, the number of observed transitions relative to that of transversions gradually decreases with increasing divergence (Xia, 1999). The plot of pairwise comparisons between the six full-length ORFs of PVY showed that this is not the case here (Fig. 1). Expected and observed saturation indices were also compared for each data set, corresponding to each protein, with DAMBE. The significance of their differences showed that there was no evidence for substitution saturation (data not shown).

Topology of the PVY trees
As expected, the trees obtained with the different PVY genes can be divided into three main parts (illustrated for the polymerases (SON41) or between the different sequence reactions (LYE84.2). The nucleotide sequences and deduced amino acid sequences of LYE84.2 and SON41 have been submitted to the EMBL database under accession numbers
Table 2. Maximum-likelihood analysis of the evolution of PVY proteins with models allowing $\omega$ to vary across amino acid sites

Results were obtained with the codeml program implemented in PAML (Yang, 1997).

<table>
<thead>
<tr>
<th>Parameter</th>
<th>P1</th>
<th>HC-Pro (12 sequences)</th>
<th>P3</th>
<th>6K1</th>
<th>CI</th>
<th>6K2</th>
<th>Nla VPg</th>
<th>Nla Pro</th>
<th>Nlb (9 sequences)</th>
<th>CP</th>
</tr>
</thead>
<tbody>
<tr>
<td>— Log likelihood for model:*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M0</td>
<td>3421-68</td>
<td>4042-28</td>
<td>2924-02</td>
<td>374-00</td>
<td>4962-57</td>
<td>452-95</td>
<td>1492-61</td>
<td>1929-83</td>
<td>2963-69</td>
<td>3601-08</td>
</tr>
<tr>
<td>M1</td>
<td>3409-16</td>
<td>4070-60</td>
<td>2878-09</td>
<td>385-35</td>
<td>4983-15</td>
<td>452-78</td>
<td>1484-22</td>
<td>1936-87</td>
<td>2938-72</td>
<td>3525-35</td>
</tr>
<tr>
<td>M3</td>
<td>3378-30</td>
<td>4025-08</td>
<td>2876-80</td>
<td>374-00</td>
<td>4944-00</td>
<td>446-87</td>
<td>1478-01</td>
<td>1918-22</td>
<td>2920-16</td>
<td>3488-00</td>
</tr>
<tr>
<td>M7</td>
<td>3378-37</td>
<td>4025-81</td>
<td>2881-56</td>
<td>374-01</td>
<td>4948-28</td>
<td>449-26</td>
<td>1481-01</td>
<td>1920-66</td>
<td>2931-62</td>
<td>3511-03</td>
</tr>
<tr>
<td>M8</td>
<td>3378-32</td>
<td>4025-09</td>
<td>2876-87</td>
<td>374-00</td>
<td>4944-00</td>
<td>446-63</td>
<td>1477-95</td>
<td>1918-21</td>
<td>2920-98</td>
<td>3488-60</td>
</tr>
<tr>
<td>Best fit among models M0, M1 and M3</td>
<td>M3</td>
<td>M3</td>
<td>M1</td>
<td>M1</td>
<td>M3</td>
<td>M3</td>
<td>M3</td>
<td>M3</td>
<td>M3</td>
<td>M3</td>
</tr>
<tr>
<td>Mean $\omega$†</td>
<td>0-226</td>
<td>0-079</td>
<td>0-213</td>
<td>0-065</td>
<td>0-032</td>
<td>0-116</td>
<td>0-069</td>
<td>0-053</td>
<td>0-067</td>
<td>0-162</td>
</tr>
<tr>
<td>Maximum $\omega$ category and corresponding proportion, p, of sites‡</td>
<td>$\omega = 0-562$, $\omega = 0-625$, $\omega = 1-000$, $\omega = 0-065$, $\omega = 0-233$, $\omega = 2-002$, $\omega = 0-402$, $\omega = 0-861$, $\omega = 0-850$, $\omega = 1-086$,</td>
<td>$p = 0-363$, $p = 0-034$, $p = 0-213$, $p = 1-000$, $p = 0-101$, $p = 0-029$, $p = 0-169$, $p = 0-026$, $p = 0-042$, $p = 0-125$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M3 versus M1‡</td>
<td>0-019</td>
<td>4 × 10⁻¹⁵</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M8 versus M7‡</td>
<td>0-072</td>
<td>2 × 10⁻¹⁰</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* Model names after Yang et al. (2000). M0 assumes a constant $\omega$ ratio for amino acid sites. M1 and M7 assume that no amino acid sites are under diversifying selection whereas M3 and M8 allow sites with $\omega > 1$.
† For the best-fit model among M0, M1 and M3.
‡ Probabilities that twice the difference between the log likelihoods is smaller than a $\chi^2$ with 2 (M8 versus M7) or 4 (M3 versus M1) degrees of freedom. Values significant at the 5% threshold are in bold.
CP gene in Fig. 2). PVY strains were mainly divided into the N and O strains and a third group, composed mainly of non-potato strains. For the CP, PVY strains of the C group are joined to the non-potato strains (C1 subgroup) or consist of a separate group (C2 subgroup) between the non-potato strains and the N and O strains. No other sequences are available for strains in the C groups, except one for the P1 protein that clusters with SON41 and LYE84.2. Tobacco strains can be joined to any group. This clustering was supported by high bootstrap values and was consistent for all PVY genes, with the exception of recombination events (see below). The different methods used to investigate the PVY phylogeny gave the same topologies with slight differences in the bootstrap values (data not shown).

**Detection of recombination in the PVY genome**

Several putative recombination events were identified among the sequences examined and were found to be statistically significant (Fig. 3A). The recombinations, occurring in CP and in the 3' untranslated region, have already been documented (Revers et al., 1996). The recombinations affecting strain N-Wilga (in the P1 protein) and strain M95491 (in the P3 and VPg proteins) have also been identified (Glais et al., 2002) and resulted in changes of phylogenetic group (between PVY N and O subgroups) that were supported by high bootstrap values (70%). Finally, two recombination events in the NIb gene were found to be significant for strain X12456, with the central part of NIb linked to the N group and the rest of the genome linked to the O group of PVY (Fig. 3B). Reanalysis of the data with SplitsTree2 without these recombinant strains did not reveal any other putative recombinant strains.

**Non-synonymous/synonymous substitution rate analyses in PVY**

**Heterogeneous evolution across amino acid sites.** Log likelihood values and parameter estimates for the best-fit model among M0, M1 and M3 (see Methods) are listed in Table 2 for each PVY protein. As a whole, the evolution of the PVY proteins appears to be conservative, with mean $\omega$ values ranging from 0.03 (CI) to 0.23 (P1). There was no evidence of selection for codon usage (effective number of codons is 52–58). Occurrence of sites submitted to positive selection was shown for the 6K2 and CP proteins. The $\omega$ values associated with the positively selected sites were weak (respectively 0.2 and 1.1) and comprised one (amino acid position 14 of the 6K2 protein) and 24 (CP; Fig. 4) amino acid positions with a posterior probability greater than 95% that they belong to the positive selection $\omega$ category. Comparison of models M8 and M7 (Yang et al., 2000) confirmed the significance of positive selection for CP but was slightly above the 5% threshold for the 6K2 protein (Table 2). For the 6K2 protein, the significance of positive selection (comparisons of M3 and M1 and of M8 and M7) increased when a larger region was analysed (from the end of CI to the end of VPg) with the six sequences available in the data set (data not shown). It is unlikely that detection of positive selection in CP was due to the large number of sequences used for the analysis. Positive selection was still detected with only the six CP sequences corresponding to the full-length PVY genomes [using PAML allowed detection of a category of codon positions with $\omega = 1.5$ (model M8) and showed that model M8 is significantly better than model M7 ($P = 0.02$)]. A proportion of amino acid sites with relatively high $\omega$ values was also observed for Nla-Pro (2.6% of sites with $\omega = 0.86$), NIb (4.2% of sites with $\omega = 0.85$) and...
Diversifying selection in potyviruses

Occurrence of diversifying selection in the CP of potyviruses with varying host-range properties

<table>
<thead>
<tr>
<th>Parameter</th>
<th>PVV</th>
<th>PVA</th>
<th>PPV</th>
<th>YMV</th>
<th>LMV</th>
<th>PYY</th>
<th>TuMV</th>
<th>BYMV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source plant genera</td>
<td>1</td>
<td>1</td>
<td>111</td>
<td>247</td>
<td>8</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Nucleotide diversity*</td>
<td>0.028 (0.028)</td>
<td>0.014 (0.054)</td>
<td>0.035 (0.07)</td>
<td>0.026 (0.032)</td>
<td>0.079 (0.029)</td>
<td>0.079 (0.029)</td>
<td>0.079 (0.029)</td>
<td>0.079 (0.029)</td>
</tr>
<tr>
<td>ω category and corresponding ωfl</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>ωfl</td>
<td>0.013 (0.007)</td>
<td>0.016 (0.014)</td>
<td>0.015 (0.010)</td>
<td>0.015 (0.010)</td>
<td>0.015 (0.010)</td>
<td>0.015 (0.010)</td>
<td>0.015 (0.010)</td>
<td>0.015 (0.010)</td>
</tr>
<tr>
<td>ω fl</td>
<td>0.013 (0.007)</td>
<td>0.016 (0.014)</td>
<td>0.015 (0.010)</td>
<td>0.015 (0.010)</td>
<td>0.015 (0.010)</td>
<td>0.015 (0.010)</td>
<td>0.015 (0.010)</td>
<td>0.015 (0.010)</td>
</tr>
<tr>
<td>M3 versus M1</td>
<td>4 × 10^{-10}</td>
<td>4 × 10^{-10}</td>
<td>2 × 10^{-12}</td>
<td>2 × 10^{-12}</td>
<td>2 × 10^{-12}</td>
<td>2 × 10^{-12}</td>
<td>2 × 10^{-12}</td>
<td>2 × 10^{-12}</td>
</tr>
<tr>
<td>M8 versus M7</td>
<td>3 × 10^{-8}</td>
<td>3 × 10^{-8}</td>
<td>3 × 10^{-8}</td>
<td>3 × 10^{-8}</td>
<td>3 × 10^{-8}</td>
<td>3 × 10^{-8}</td>
<td>3 × 10^{-8}</td>
<td>3 × 10^{-8}</td>
</tr>
</tbody>
</table>

Heterogeneous evolution across lineages. For all proteins in the PVY genome except the P1, 6K1 and 6K2 proteins, the ‘N-ratio’ model (see above) fitted the data significantly better than M0, indicating heterogeneity in the mode of evolution along lineages (data not shown). In all 10 proteins, branches were detected with ω estimates greater than 1 (data not shown), but these particular branches involved only a very small number of substitutions (a maximum of eight substitutions was observed for a branch in the P3 tree with ω = 1.9). The ω estimates of these branches were never significantly greater than 1, essentially because of the small number of substitutions along these branches in comparison with the background branches (data not shown).

Evolution modes of the CP of other potyviruses

Few sequence data are available for the 6K2 proteins of other potyviruses. However, the numerous available CP sequences can be used to see whether the CP of other potyviruses share the same evolution pattern as that of PVY. Hence, the same analyses as before were performed with the CP genes of seven potyviruses, with either wide or narrow host ranges (Table 3). For Plum pox virus (PPV), TuMV and YMV, several sequences were discarded because of important gaps in the alignments. Most of the potyviruses showed a varying proportion of fast-evolving amino acid sites, except PVV and TuMV (Table 3). Diversifying selection was statistically significant for YMV and BYMV when models M3 and M1 or M8 and M7 in paml (Yang, 1997) is smaller than a χ² with 4 or 2 degrees of freedom, respectively. Values significant at the 5% threshold are in bold.

For all proteins in the PVY genome except the P1, 6K1 and 6K2 proteins, the ‘N-ratio’ model (see above) fitted the data significantly better than M0, indicating heterogeneity in the mode of evolution along lineages (data not shown). In all 10 proteins, branches were detected with ω estimates greater than 1 (data not shown), but these particular branches involved only a very small number of substitutions (a maximum of eight substitutions was observed for a branch in the P3 tree with ω = 1.9). The ω estimates of these branches were never significantly greater than 1, essentially because of the small number of substitutions along these branches in comparison with the background branches (data not shown).

Evolution modes of the CP of other potyviruses

Few sequence data are available for the 6K2 proteins of other potyviruses. However, the numerous available CP sequences can be used to see whether the CP of other potyviruses share the same evolution pattern as that of PVY. Hence, the same analyses as before were performed with the CP genes of seven potyviruses, with either wide or narrow host ranges (Table 3). For Plum pox virus (PPV), TuMV and YMV, several sequences were discarded because of important gaps in the alignments. Most of the potyviruses showed a varying proportion of fast-evolving amino acid sites, except PVV and TuMV (Table 3). Diversifying selection was statistically significant for YMV and BYMV when models M3 and M1 or M8 and M7 in paml (Yang, 1997) is smaller than a χ² with 4 or 2 degrees of freedom, respectively. Values significant at the 5% threshold are in bold.

As for the PVY proteins, models allowing the ω ratio to vary across lineages showed significant heterogeneity (‘N-ratio’ model versus M0) in most cases, but did not reveal any branch with ω significantly greater than 1-0 (data not shown).

Estimates of ω in sequence pairwise comparisons (Yang & Nielsen, 2000) for all sequence data sets were rarely greater than 1-0 or involved only a small number of substitutions. An exception was found for PPV, where pairwise comparisons between the whole CP sequences of two strains (corresponding to accession numbers X81078 (unknown host origin) and X81083 (sour cherry origin)) and several other strains revealed ω estimates greater than 1-0 with relatively large numbers of substitutions. The estimate between CP sequences X81078 and X81083 is ω = 3-09, with 163 non-synonymous versus 20 synonymous substitution estimates, very near the 5% significance threshold (P = 0-06).
Discussion

Occurrence of positive selection in potyviruses

Sequence analysis has allowed the detection of genome regions that are undergoing positive selection in several animal viruses (reviewed in Yang & Bielawski, 2000), mainly in domains that interact with the host immune system. It is difficult, however, to derive evidence for positive selection from sequence data when ω ratios are estimated for entire genes. Indeed, ω ratios for plant virus genes are usually low, except for some overlapping ORFs (reviewed in Garcia-Arenal et al., 2001). Positive selection is, however, suggested to occur in natural populations of Tobacco mild green mosaic virus (Frale et al., 1996).

Our study showed that amino acids in two PVY proteins were submitted to positive selection: a single amino acid position in the 6K2 protein and several amino acid positions in the CP (especially in the N terminus). Evidence for diversifying selection was also obtained for the CP of BYMV and YMV. Several of these positions were shared (or nearby) in several potyviruses, strengthening their significance. Heterogeneous modes of evolution were detected not only across amino acid sites but also across lineages. For most of the PVY proteins and for the CP of other potyviruses, the ω ratio is significantly different across the different lineages. However, no specific branches with ω ratios significantly greater than 1.0 were detected, because of the relatively small numbers of substitutions. Diversifying selection events were also suspected for the CP of PVA and of two PPV isolates in comparison with several other isolates. It should be noted that the test of molecular adaptation used in PAML is highly conservative: it will fail if positive selection affects only a few amino acid sites along a few lineages on the phylogeny. Models that allow the selective pressure to vary both among lineages and among codons would be more realistic for detecting adaptive molecular evolution, but should have increased power.

Evolutionary constraints on the potyvirus genome

We have obtained evidence for positive selection events occurring in two proteins of the PVY genome and in the CP of other potyviruses. In the absence of a three-dimensional structural model or with only very partial delineation of the functional domains of these proteins, it is difficult to unravel the evolutionary constraints that are responsible for these events. It is plausible that the regions identified are ligand-binding domains and that positive selection reflects gains or losses of affinity for these ligands. In this respect, interactions with plant or vector-insect factors as well as interactions with other PVY proteins can be the driving forces of this evolution.

The N-terminal part of CP is non-essential for replication and cell-to-cell movement of the potyviruses Tobacco etch virus (TEV) (Dolja et al., 1993, 1994) and Zucchini yellow mosaic virus (Arazi et al., 2001). Mutations in this part of the genome can, however, affect some virus functions quantitatively. Deletion of 25 amino acids in the CP of TEV (8 of these amino acids correspond to amino acid positions submitted to positive selection in the PVY genome; Fig. 4) reduced the speed of cell-to-cell movement in tobacco and completely abolished systemic movement (Dolja et al., 1994). A substitution of the amino acid immediately following the DAG motif in the CP of tobacco vein mottling potyvirus (the corresponding amino acid in PVY is potentially submitted to positive selection; Fig. 4) reduced the efficiency of aphid transmission of the virus (Atreya et al., 1995). This illustrates that amino acid substitutions in this region can affect the fitness of various potyviruses quantitatively and can consequently be the target of selective processes. The N-terminal part of CP is a striking example of the multifunctionality of viral proteins. It is exposed on the virion surface (thus potentially involved in binding ligands) and involved in aphid transmission (Atreya et al., 1990, 1991, 1995) and cell-to-cell and long-distance movement (Dolja et al., 1994, 1995; Rojas et al., 1997). Therefore, it is difficult to know which virus function is involved in these positive selection events.

A three-dimensional structure model was recently proposed for the CP of PVA (Baratova et al., 2001). We compared the amino acid positions previously identified with this model as a reference, using the CP sequence alignment in Fig. 4 and keeping in mind that only weak evidence of diversifying selection was obtained for PVA (Table 3). In all four potyviruses, a large proportion of sites that belong to the positive selection category are in the N-terminal part of the CP, comprising the regions most accessible at the particle surface (Baratova et al., 2001). Several of the positively selected amino acid sites are also encountered in the central and C-terminal (internal) parts of the capsid, without coinciding with a particular domain of the putative protein structure.

The central hydrophobic domain of the 6K2 protein is responsible for binding to the ER (Schaad et al., 1997) and it has consequently been proposed that the 6K2 protein is required for genome amplification and that it anchors the replication apparatus to ER membranes. Chu et al. (1997) showed that substitutions in a region spanning the end of CI, 6K2 and the beginning of VPg of TEV were sufficient (but not absolutely necessary) to induce a wilting response in Tabasco pepper. Since most of the TEV field isolates belong to the ‘wilting type’, the determinants of the wilting response in Tabasco could confer a selective advantage to TEV. However, the two amino acid changes in the 6K2 peptide that distinguish a non-wilting from a wilting isolate of TEV do not coincide with amino acid position 14 of PVY, which we suspect to be under positive selection.

Host adaptation and evolution of the potyvirus genome

Many environmental factors can exert evolutionary constraints on PVY, but one of the most important is certainly the host plant. Occurrence of diversifying selection in the CP of
potyviruses was not found to correlate with the width of host range (Table 3). The three potyviruses that showed evidence of positive selection (PVY, BYMV and YMV) correspond to CP sequences with the largest diversity measured at the nucleotide level (Table 3). Strong conservation of the nucleotide sequences for the other potyviruses can be interpreted as an effect of purifying selection and can also be responsible for the lack of significance of positive selection when \( \omega \) estimates greater than 1:0 were obtained (PVA, PPV and LMV). There was no structuring of the phylogenetic trees as a function of the host species of the isolates (data not shown) except, partially, for PVY, which would be an indication of the lack of involvement of the CP as a host-range determinant. However, it should be emphasized that very limited host-range data are available for the majority of the strains used in these sequence analyses.

Biological arguments have already been reported for host adaptation of PPV in cherry trees versus other Prunus species (Dosba et al., 1987; Nemchinov et al., 1996) and especially for PVY. In PVY, isolates from pepper usually do not infect potato plants systemically (McDonald & Kristjansson, 1993; d’Aquino et al., 1995; Gebre-Selassie et al., 1985) when inoculated manually. Reciprocally, manual inoculations did not allow systemic infection of pepper plants with PVYN strains (Gebre-Selassie et al., 1985; McDonald & Kristjansson, 1993; Valkonen et al., 1996), whereas some of the strains in the O and C groups could infect some pepper cultivars (McDonald & Kristjansson, 1993; Valkonen et al., 1996; Blanco-Urgoiti et al., 1998). Tomato (Stobbs et al., 1994; Legnani, 1995) and tobacco (Blancard, 1998) plants lacking specific resistance genes can be infected by most, if not all, PVY isolates, including those that originate from potato and pepper. Taken together, these data suggest an influence of some host plants (potato and pepper) but not others (tomato and tobacco) in the selection and evolution of PVY isolates. There are also partial arguments for host-range variations according to the source plants for strains of BYMV (Wada et al., 2000) and TuMV (e.g. Stavolone et al., 1998). PVV, PVA, YMV and LMV, which all share very narrow host ranges, do not seem to undergo adaptation or diversification according to host plants (except for overcoming of specific cultivar resistance genes, or propagation of particular virus strains that could be due to vegetative and clonal multiplication of cultivars). However, both kinds of bias should be emphasized. (i) Once again, very limited host-range data are available for the majority of the strains used in these sequence analyses, and (ii) PAML analyses are based on substitution models; very divergent CP sequences (e.g. some cherry isolates of PPV and some YMV and TuMV isolates) could not be included in the analyses because of numerous gaps in the alignments. This will reduce the possibility of detecting divergent evolution events and may introduce a bias for the comparison of CP evolution of the different potyviruses.

In conclusion, evidence of diversifying selection in several potyviruses cannot yet be interpreted in terms of evolutionary constraints. Further biological characterization should be performed and the role of amino acid substitutions in these different proteins investigated through molecular analyses. Confrontation of biological data with the modes of evolution of the CP makes this protein an interesting candidate for host adaptation of some potyviruses, including PVY, BYMV, YMV and, maybe, PPV.

B.M. and C.M. contributed equally to this work and should be considered as joint first authors. B.M. was the recipient of a post-doctoral NATO fellowship and C.M. of a PhD fellowship from the French Research Ministry. We thank F. Garcia-Arenal for improving the manuscript, N. Galtier for helpful comments, V. Marie-Jeanne Tordo for unpublished sequence data and L. Guilbaud and B. Olsen for efficient technical help.

References


and movement of tobacco etch potyvirus in plants. EMBO Journal 13, 1482–1491.


Felsenstein, J. (1993). PHYLIP: phylogenetic inference package, version 3.5c. Distributed by the author, Department of Genetics, University of Washington, Seattle, USA.


Received 27 March 2002; Accepted 17 June 2002