The 3’ untranslated region of cucumber mosaic virus (CMV) subgroup II RNA3 arose by interspecific recombination between CMV and tomato aspermy virus

Jeremy R. Thompson and Mark Tepfer

Recombination in single-stranded RNA viruses is one of the principal mechanisms responsible for their evolution. Here we show, using a variety of different methods, that the 3’ untranslated region (3’UTR) of subgroup II strains of cucumber mosaic virus (CMV(II)) is related more closely to that of tomato aspermy virus (TAV) than to those of CMV(I) strains. These results suggest that the CMV(II) 3’UTR arose by interspecific CMV/TAV recombination. The putative crossover is close to the 5’ end of the 3’UTR, at a recombination hot spot previously observed in short time-frame experiments. The CMV(II) strains show divergence from TAV at specific points along the 3’UTR that most probably indicate adaptive changes due to natural selection. Thus, for the large majority of CMV(II) strains examined, the 3’UTR has two discrete regions, W (nt 1902–1971) and Y (nt 2126–2184), that are more similar to the corresponding regions of TAV than to those of CMV(I) strains.

Genetic variation in RNA viruses is assumed to exist principally because of three mechanisms: mutation, recombination and reassortment. Due to a relatively short life cycle and an RNA-dependent RNA polymerase that lacks proofreading, RNA viruses have a mutation rate that has been described as being close to the viable maximum (Drake & Holland, 1999; Roossinck et al., 2001). Nevertheless, studies in natural plant virus populations appear to indicate that RNA viruses might be more genetically stable and less diverse than the above would suggest (reviewed by Garcia-Arenal et al., 2001) with, in the case of strains of cucumber mosaic virus (CMV) (genus *Cucumovirus*, family *Bromoviridae*), different parts of the genome being subjected to different evolutionary constraints (Bonnet et al., 2005; Lin et al., 2004). The accumulation of deleterious mutations in a virus population leading to a loss in overall fitness can be counteracted by both natural selection and/or recombination, where the unfit genotype is either diluted and eventually lost (in the first case) or repaired (in the latter). Recombination therefore can be seen to have a dual role, i.e. eliminating negative combinations of mutants from the population, but also introducing gross alterations in the virus genome that can, in some instances, result in the successful establishment of new taxa (Codoñer & Elena, 2008; Miller & Koev, 1998; Valli et al., 2007).

CMV has the largest host range (>1000 plant species) of any known plant virus, and recombinant CMV has been observed both in the field (Bonnet et al., 2005; Fraile et al., 1997) and in the laboratory (Aaziz & Tepfer, 1999; Chen et al., 2002; de Wispelaere et al., 2005; Turturo et al., 2008). The genome is tripartite, with RNA1 encoding the 1a protein (helicase/methyltransferase), RNA2 encoding the 2a protein (RNA polymerase) and also the silencing suppressor protein, encoded by overlapping open reading frame (ORF) 2b, and the bicistronic RNA3 encoding the 3a movement protein (MP) and the 3b coat protein (CP). Taxonomically, CMV falls into two main subgroups (I and II), subgroup I being further subdivided into IA and IB (reviewed by Palukaitis & Garcia-Arenal, 2003; Palukaitis et al., 1992). Of 159 CMV strains taken in Spain at various locations across a range of 13 years, 5% were reassortants and 17% were recombinants. The majority of the recombinants (15.7% of the strains) were derived from RNA3, of which 96% were of a type where the MP ORF was from subgroup IB [CMV(IB)] and the CP ORF was from subgroup IA [CMV(IA)]. CMV(I)/CMV(II) recombinants were found at a frequency of 0.6% for RNA1 and RNA2, whereas none were found for RNA3 (Bonnet et al., 2005). Therefore, the nature of the recombinant and reassortant populations found supports the idea that there is in general an associated fitness cost, in particular when the recombinant genome encodes a hybrid protein. In a study of the biological properties of CMV(I)/CMV(II) RNA3 recombinants, even those that accumulated to levels...
equivalent to those of the parental strains when tested individually were outcompeted when co-inoculated with the parentals (Pierrugues et al., 2007). Nevertheless, there are laboratory examples where it has been demonstrated that novel cucumoviral recombinants can outcompete the parental virus. A pseudorecombinant virus consisting of RNAs 1 and 2 of strain trk7 [CMV(II)] and a recombinant RNA3 of tomato aspermy virus (TAV), with most of its 3′ untranslated region (3′ UTR) belonging to RNA2 of trk7, showed an increase in fitness relative to either TAV or CMV in tobacco (Fernandez-Cuartero et al., 1994).

In this work, we demonstrate in silico that CMV(II) has obtained almost all of its 3′ UTR from TAV by means of recombination, thereby providing clear evidence of the establishment and maintenance in nature of a recombinant taxon in the genus Cucumovirus. First, phylogenetic analysis of the nucleotide sequence of each coding and non-coding region of RNA3 of members of the genus Cucumovirus was carried out on complete sequences in GenBank for CMV, peanut stunt virus (PSV) and TAV (see Supplementary Table S1, available in JGV Online). Complete sequences were split into their component coding and non-coding regions by using Vector NTI Advance 9.1 software (Invitrogen). The 5′ UTR was excluded from the analyses because of the poor quality of the resulting alignments. All phylogenograms were produced by using PAUP* 4.0 (Swofford, 1998) after selecting the best nucleotide-substitution method with MODELEST (Posada & Crandall, 1998) (Fig. 1). Regions encoding MP and CP revealed similar clustering for each virus species, with CMV(I) and (II) grouping together, as expected (Fig. 1b, d). Subgroup IB strains lay outside the main IA cluster as a dispersed group. Interestingly, subgroup IA strains 207, Fny, I17F, M, Ns, Ri-8, Rs and Sny formed a subclade within CMV(IA) for both the MP and CP. CMV(II) strains formed a much less divergent group, as also attested to by the uniformity of the alignments (not shown). The groupings in the intercistronic region (ICR) (Fig. 1c) also reflect the distributions found for the coding region, except that the subclade within CMV(IA) is not distinguished. For the 3′ UTR (Fig. 1e), the picture is very different, with all CMV(II) strains grouping with PSV and TAV strains on a branch separate from CMV(I). Furthermore, CMV(II) forms a separate branch with TAV. There is still, however, consistency in the arrangement of subgroup IB strains, except that C7-2 (IB) falls in with IA strains and D8 (IA) is found on a branch with SD (IB). These results support and expand the analyses of Roossinck (2002).

The closer phylogenetic relationship between the 3′UTRs of CMV(II) strains and TAV than between those of the two CMV subgroups could in principle be due to two processes: either gradual convergence via mutation and selection, or recombination. To test whether it was the latter, we analysed full-length alignments between I17F-CMV (I), R-CMV (II) and P-TAV, using the seven methods for detection of recombination breakpoints contained in the RDP software (Martin et al., 2005), and found two distinct regions in the 3′ UTR at R-CMV nt 1902–1971 (region W) and 2126–2184 (region Y) that show high identity to P-TAV (Fig. 2). This analysis was then extended to the remaining CMV(II) strains. There was consistency in all strains for a breakpoint at nt 1902, yet its end point ranged from nt 1951 to 2043 (see Supplementary Table S2, available in JGV Online). For the second region, all strains showed a breakpoint at nt 2126, except for trk7 and Xb, where it was located at nt 2141. The end points were predominantly at nt 2184, with some minor variation (see Supplementary Table S3, available in JGV Online). The putative recombinant regions were detected on average by three methods, consistent with the conservative cut-off proposed by Codoñer & Elena (2008). For the whole length of RNA3, only one other putative recombinant site was found: between P-TAV and I17F-CMV in the MP, although this was discarded as not significant, being detected by only one method (not shown). To confirm these findings, full-length alignments of I17F-CMV (I), P-TAV and all CMV(II) strains were subjected to both GARD (multiple breakpoints) and single breakpoint analyses of the GARD program (Kosakovsky Pond et al., 2006). A single, consistent dominant breakpoint was detected for all alignments at nt 1895/6–1897/8 (R-CMV) (see Supplementary Table S4, available in JGV Online), with an evidence ratio ranging from 2.1×10^4 to 4.9×10^6 – significance using GARD is shown by an evidence ratio [exp(ΔAIC/2)] of >100, where ΔAIC is defined as the AIC (Akaike’s information criterion) score improvement (Burnham & Anderson, 2003). GARD analysis identified multiple breakpoints [n=13–19 (not shown)], although the dominant breakpoint of nt 1895/6–1897/8 was always the first breakpoint and the only one to be identified consistently. It is also important to note that the breakpoints downstream of nt 1902 identified by RDP3 were not verified here.

Based on the above findings, of the two plausible models for the origin of the sequence of the CMV(II) 3′ UTR, i.e. (i) that recombination with TAV has occurred at various sites (a maximum of four) or (ii) that a recombination event occurred once, followed by gradual changes due to natural selection or genetic drift, the latter is more probable in light of the results obtained by GARD. A model is therefore proposed to explain the present topology based on possible tertiary interactions that are consistent with preserved compensatory nucleotide changes found in related viruses (Fig. 3). The Kinefold program (Xayaphoummine et al., 2005), using renaturation fold parameters while forcing consecutive base pairs, was applied to the R-CMV 3′ UTR structural map (Thompson et al., 2008). Three pseudoknots were found specifically for R-CMV: between loops L and K, between loop G and bulge C (see Supplementary Fig. S1, available in JGV Online) and between loops M and B (Supplementary Fig. S2). It may be significant that five of the six sites involved in potential tertiary interactions are in conserved blocks W and Y. The theoretical feasibility of identified
Fig. 1. Phylogenetic analyses of the coding and non-coding regions of RNA3 of members of the genus Cucumovirus. In (a), the 5' untranslated region (5' UTR), movement protein (MP), intercistronic region (ICR), coat protein (CP) and 3' untranslated region (3' UTR) are shown. Boxes represent open reading frames, lines correspond to untranslated regions and the oval and three circles depict the cap structure and tRNA-like structure, respectively. Numbers in parentheses indicate the range of sequence lengths for each region analysed. (b–e) Phylograms generated by PAUP* 4.0 (Swofford, 1998) of the different RNA3 regions of the strains listed in Supplementary Table S1: (b) MP; (c) ICR; (d) CP; (e) 3' UTR. Bootstrap values from 100 replicates are indicated close to branches. Branches were collapsed when bootstrap values were <70. Sequences containing greater than full-length sequences due to sequence duplications in the 3' UTR had these duplications removed. Thus, the positions of deletions were, for 1-TAV, nt 2064–2231 (168 nt), and for Dhanbad-TAV, Kolkata-TAV, Lucknow-TAV and V-TAV, nt 2065–2229 (165 nt). Bars, 0.1 genetic distance.
tertiary interactions in three dimensions was confirmed with the RNA2D3D application (Martinez et al., 2008) (not shown), but will need to be confirmed in vitro and in vivo.

The main recombination hot spot at or around nt 1900 is clearly biologically important: in short-term experiments, recombination at nt 1902 has been shown to predominate in RNA3 populations originating from two distinct species (e.g. CMV and TAV) (de Wispelaere et al., 2005) and was also observed in singly infected plants transgenic for the CP and 3'UTR (Morroni et al., 2009). It is also present in various TAV strains that have a repeat in RNA3 which links nt 2065 to nt 1902 (equivalent R-CMV positions) (Shi et al., 1997). This therefore demonstrates clearly that recombinants identified in the laboratory can to some extent represent those found in nature, which instead of forming minor components of the population as discussed by Bonnet et al. (2005), can lead to the establishment of a new virus subtype as shown in strains isolated from Alstroemeria (Chen et al., 2002).

The other important feature of nt 1900 is that it is close to the 5' end of subgenomic RNA5, which extends to the 3' end of RNAs 1, 2 and 3 (Blanchard et al., 1997; de Wispelaere & Rao, 2008). RNA5 levels in the plant can reach 50% of total RNA3-derived molecules, thus providing an abundant pool of truncated viral RNA (Thompson et al., 2008). The stable presence of truncated RNAs of related sequences has been shown to stimulate recombination between related molecules (Cheng et al., 2007). Concerning how most of the CMV(II) 3'UTR was received from TAV, the most probable model would be a recombination mechanism similar to that already proposed for precise homologous recombination involving RNA5 (de Wispelaere & Rao, 2008), whereby the polymerase would remain attached to the 3' end of the nascent negative strand of RNA5 of TAV, before switching templates to the positive strand of RNA3 of CMV.

The most probable hypothesis for the origin of the recombinant 3'UTR fragment in CMV(II) strains is that it was the result of a single crossover between CMV and TAV before diversification of CMV(II) around nt 1900, contributing to, along with other unknown events, diversification of the entire genome as a result of gene co-adaptation. If this were not the case and, alternatively, CMV(II) had already diversified or there were repeated recombination events in different CMV strains, one would expect to find some CMV(II) strains without the TAV-like 3'UTR, and none have been described so far. This, however, does not rule out completely the possibility that recombination between TAV and CMV(II) reoccurred subsequently, and might explain some of the variation detected in the location of putative crossover points downstream of nt 1900. This latter point could also be explained by subsequent single nucleotide changes that subtly alter the location of the crossover as determined by the program. In comparison, PSV shows notable variation in a conserved motif (box 1) around nt 1900 (Supplementary Fig. S2). In strain W this motif is missing, whilst in the remaining strains, there are single nucleotide differences compared with the sequence found in CMV(II) and TAV, suggesting a more distant relationship. Looking at the whole genome, the process that has happened in CMV does not appear to have happened in PSV, as the phylogenetic relationship between CMV subgroups remains relatively constant, whereas in contrast, there is strong evidence of interstrain recombination in PSV (Kiss et al., 2008).

Overall, this work has shown that CMV(II) has obtained most or all of its non-coding 3'UTR, a region essential for replication and potentially involved in other essential functions such as translation and encapsidation, from...
another virus species via recombination. After that recombination event, the central part (X) of the recombinant region has changed significantly compared with its flanking regions (W and Y), indicating varying evolutionary pressures. This study also highlights the inherent difficulties of such analyses and the need for cautious interpretation of the results – the application of eight different recombination-detection programs has provided a robust approach.

Acknowledgements

Many thanks to Sergei L. Kosakovsky Pond for his helpful comments. This work was supported in part by the Fondazione Cassamarca.

References


