VIROLOGY 206, 1129–1132 (1995)

A Recombinational Event in the History of Luteoviruses Probably Induced by Base-Pairing between the Genomes of Two Distinct Viruses

M. J. GIBBS[1] AND J. I. COOPER

NERC Institute of Virology and Environmental Microbiology, Mansfield Road, Oxford OX1 3SR, United Kingdom

Received August 9, 1994; accepted October 21, 1994

Alignments of luteovirus readthrough protein amino acid sequences show they consist of two distinct regions, here named the N domain and the C domain. N domain sequences were classified, and comparison of this gene phylogeny to phylogenies of other luteovirus genes revealed an anomaly in the relationships between beet western yellows luteovirus, cucurbit aphid-borne yellows luteovirus (CABYV), and pea enation mosaic RNA1 (PEMV1). Together with alignments of virion protein and readthrough protein amino acid sequences, these gene phylogenies indicate the anomaly to be the result of two recombinational events, probably between ancestors of CABYV and PEMV1 and leading to the transfer of RNA coding for the N domain to an ancestor of CABYV. Two likely recombination sites were identified from the alignments, one at the 5′ end of the readthrough protein gene and the other at the 5′ end of the sequence coding for the C domain. Alignments of the nucleotide sequences encompassing the probable recombination sites suggest that base-pairing between the genomes of the two ancestral luteoviruses, resulting from local sequence similarity at the 5′ end of the readthrough protein gene, probably induced one of the interspecies recombinational events.

The isometric virions of luteoviruses are assembled from the virion protein (VP) and a fusion protein consisting of the VP and a second polypeptide known as the readthrough protein (RT protein) (*1–3*). Alignments show luteovirus VPs to have homology with the virion protein shell domain (S domains) of carmoviruses, sobemoviruses, and tombusviruses (*4*) and hence to have the eight-stranded anti-parallel β-barrel structure (*5*). The fusion protein is produced by suppression of the VP gene stop codon and subsequent translation of the in frame RT protein gene (*6, 7*). RT proteins are exposed on the surface of virions (*8*). Alignments of RT protein amino acid sequences show they consist of two distinct regions probably corresponding to separate domains (*2, 9*). The N-terminal sequences (N domains) are relatively conserved whereas the C-terminal sequences (C domains) are more variable.

Classification of luteovirus S domain and C domain amino acid sequences, shown in Fig. 1, place beet western yellows luteovirus (BWYV) (*2*) and cucurbit aphid-borne yellows luteovirus (CABYV) (*9*) as sister taxa. This relationship was also found in classifications of amino acid sequences inferred from the second open reading frame and polymerase gene (M.J.G., unpublished data). Hence the apparent relationship between the N domains of CABYV and pea enation mosaic virus RNA1 (PEMV1) (*13*), also shown in Fig. 1, is anomalous.

The anomaly in tree topology may have been due to sequence convergence (homoplasy), if for example the sequences were poorly conserved. However, the conservation of N domain sequences suggests a different explanation. Alignments, Fig. 2, and diagon plots (*22*) (not shown), confirm that the VP and C domain of CABYV are more closely related to those of BWYV than PEMV1, but also show that the converse is true of the N domains of these viruses. Percentage identity scores, calculated from the alignment shown in Fig. 2, summarize the relationships. The VP, N domain, and C domain sequences of CABYV and BWYV are 67, 27, and 47% identical, respectively, whereas the VP, N domain, and C domain sequences of CABYV and PEMV1 are 30, 49, and 9% identical, respectively.

Together the alignments and classifications indicate the anomaly to be the result of recombination probably between ancestors of CABYV and PEMV1 and leading to the transfer of RNA coding for the N domain to an ancestor of CABYV. Two alternative explanations, that either BWYV or PEMV1 evolved from a recombinant, are difficult to reconcile with the phylogeny of the other characterized luteoviruses. The reestablishment of the close relationship between CABYV and BWYV in the C domain sequence suggests that two crossover events took place and that apart from the sequence encoding the N domain, the ancestor of CABYV remained intact.

Identical and closely similar amino acid residues found on aligning sequences from the three viruses are shown in Fig. 2. The regions in the sequences where relationships change (transitional regions), one at the N terminus of the RT protein and the other at the N terminus of the C domain within the RT protein, as marked in Fig.

[1] To whom reprint requests should be addressed.

0042-6822/95 $6.00

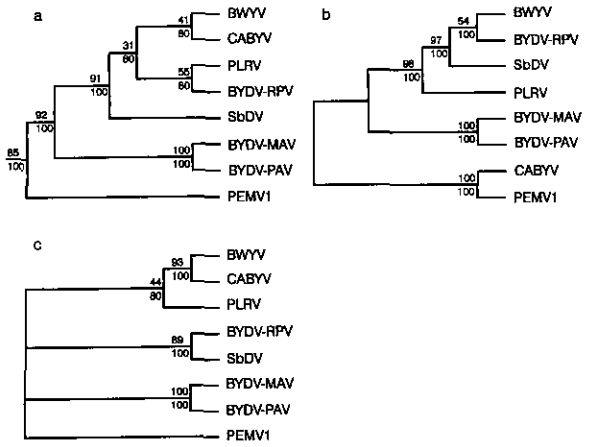

FIG. 1. Majority rule consensus trees inferred from classifications of the aligned amino acid sequences of luteovirus S domains (a), N domains (b), and C domains (c). Genomic sequences for the following viruses were obtained from the EMBL Nucleotide Sequence Data Library: BWYV (*2*), barley yellow dwarf luteovirus type MAV (BYDV-MAV) (*10*), BYDV type PAV (BYDV-PAV) (*11*), BYDV type RPV (BYDV-RPV) (*12*), CABYV (*9*), pea enation mosaic RNA1 (PEMV1) (*13*), potato leaf roll luteovirus (PLRV) (*14, 15*), and soybean dwarf luteovirus (SbDV) (*16*). S domain sequences were delineated from the VP sequences as in Dolja and Koonin (*4*). N domain and C domain sequences were delineated from the RT protein sequences as shown in Fig. 2. Sets of amino acid sequences were progressively aligned using the program package ClustalV (*17*). Classifications were inferred by the neighbor-joining method (*18*), implemented in ClustalV, and by the maximum parsimony method using the heuristic searching technique tree bisection-reconnection (TBR), implemented in the program package PAUP version 3.0 (*19*). Three equally parsimonious trees were found using the S domain and C domain data and one most-parsimonious tree was found using the N domain data. A tree found using maximum parsimony and bootstrapping (*20*) was included. Majority rule consensus trees were inferred from all tree topologies found for a set of sequences using COMPONENT version 2.0 (*21*). To the left of each node and below the branch is the percentage of trees in agreement with the nodes shown. To the left of each node and above the branch are boot-strap values obtained for each node. Tdraw (J. H. W. Fergeson, unpublished program) was used to infer the midpoint between the most distant taxa in the neighbor-joining trees. The root of the S domain tree was inferred from a classification including sequences from viruses from the aphthovirus, carmovirus, dianthovirus, necrovirus, tombusvirus, and tymovirus genera. The PEMV1 C domain sequence is considerably shorter than those of the other viruses (see Fig. 2) and therefore its inclusion in the C domain set may be invalid; however, exclusion of the sequence did not affect the tree topology, and a maximum parsimony tree inferred using only the positions that aligned with the PEMV1 sequence had the same topology.

2, were found by comparing the frequency of identities and similarities at the junctions of the three polypeptides and from diagon plots (not shown). Equivalent transitional regions were found by alignment of the nucleotide sequences (not shown), confirming that the 5′ end of the RT protein gene and the 5′ end of the sequence encoding the C domain within the RT protein gene are the likely sites of recombinational events. However, the location of each recombination site could not be defined more accurately using these alignments.

Template switching, the likely cause of the two events, is thought to be induced in some instances by base-pairing between the template RNAs (*23*). Consequently, alignments were also used to search for any nucleotide sequence similarities within, or nearby, the two probable recombination sites. Little likelihood of complementation was found between nucleotide sequences encoding the N termini of the C domains. However, at the 5′ ends of the RT protein genes, similarities were found between the nucleotide sequences of all three viruses. These similar sequences, mostly consisting of matching cytosine residues coding for proline, span the sequence coding for the transitional region.

All the sequenced luteovirus genomes encode a series of proline residues, usually alternating with other amino acid residues, at the N terminus of the RT protein. Codons for proline have the sequence "CCN," hence the 5′ ends of the RT protein genes are rich in cytosine residues and these genomes are very similar in this region. For example, although the entire genomic sequences of BWYV and PEMV1 are 37% identical at the nucleotide level when aligned using the program GAP (*24*), the nucleotide sequences from these two genomes spanning the 5′ recombination site, 54 and 51 bases long, respectively, are 63% identical when aligned in this way. In comparison, the polymerase genes, the most conserved open reading frames, of these two viruses are only 48% identical when aligned using GAP. Alignment of the nucleotide sequences that encode the transitional region, based on the amino acid alignment, Fig. 3, show CABYV and PEMV1 to have 34 identities, 24 of which are cytosine, and BWYV and PEMV1 to have 33 identities, 20 of which are cytosine. In the genomes of CABYV and PEMV1 repetition of cytosine residues continues for a further 24 nucleotides on the 3′ side of the transitional region.

The location of a recombination site within this region of significant homology is unlikely to be coincidental. Cytosine residues on the positive strand of the genome of one ancestral luteovirus may have base-paired with the complementary guanine residues on the negative strand copy of the genome of the other ancestral luteovirus and so induced a replicase to switch templates. This suggestion, together with findings concerning the evolution of tobacco rattle tobravirus (*25*), supports the notion that the formation of intermolecular duplexes contributes to gene transfer among RNA viruses. The divergence of BWYV, CABYV, and PEMV since the recombinational events, as judged from alignments (Figs. 2 and 3), obscures the precise location of each recombination site and prevents prediction of the likely base-pairing between the ancestors of CABYV and PEMV1. However, the evolutionary distance between BWYV and PEMV1 may be similar to that between the parents of the recombinant and so comparison of these sequences gives some measure of the possibilities (Fig. 3).

The phenotypic consequences of the recombinational

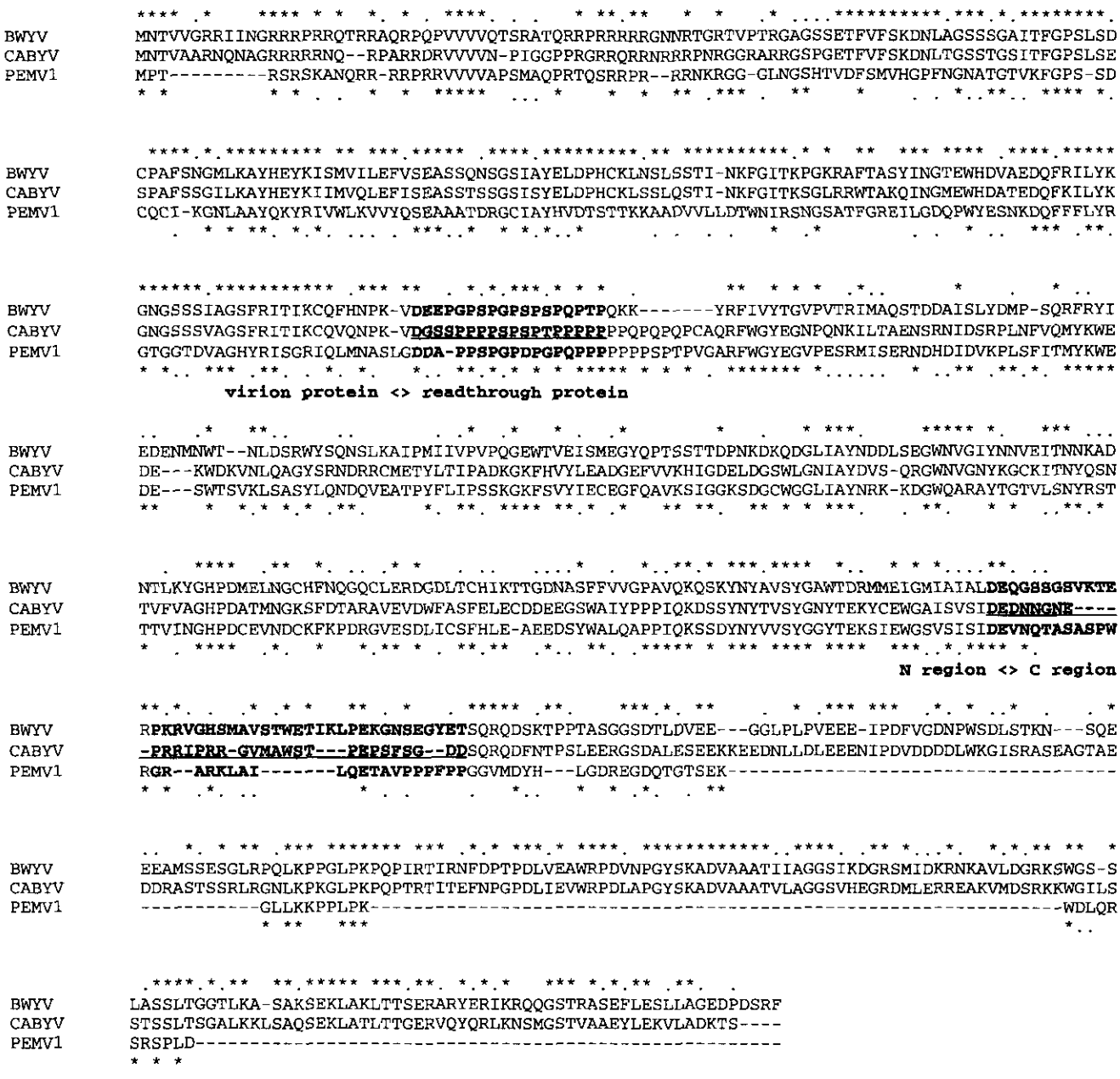

FIG. 2. An alignment of the VP and RT protein N and C domain amino acid sequences of BWYV, CABYV, and PEMV1. Each of the three sets of sequences was aligned independently using ClustalV (*17*). The two regions in the sequences where transitions in phylogenetic relationship occur are marked in bold type and the CABYV sequence is underlined at each of these locations. Identical and closely similar amino acid residues found between the sequences are marked by asterisks and dots. Those found between the sequences of BWYV and CABYV are marked above the alignment, and those found between the sequences of CABYV and PEMV1 are marked below the alignment.

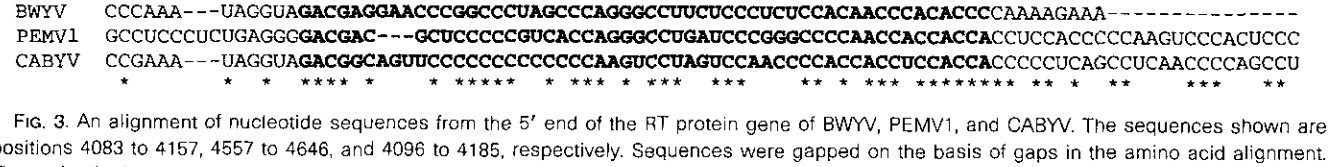

FIG. 3. An alignment of nucleotide sequences from the 5′ end of the RT protein gene of BWYV, PEMV1, and CABYV. The sequences shown are positions 4083 to 4157, 4557 to 4646, and 4096 to 4185, respectively. Sequences were gapped on the basis of gaps in the amino acid alignment. The region in the sequences where the transition in phylogenetic relationship occurs is marked in bold. Identities between the sequences of BWYV and PEMV1 are marked by asterisks above the alignment. Identities between the sequences of CABYV and PEMV1 are marked by asterisks below the alignment.

event are unclear. It has been suggested RT proteins determine the specificity of luteoviruses for their aphid vectors, possibly permitting transport of virions from the haemocoel to the salivary duct within the aphid (*1, 8*). Hence the recombinational event may have altered the vector specificity of the ancestral virus and so offered an advantage. If this were true CABYV and PEMV1 might be expected to have similar vector specificities, but the specificities of CABYV (*26*) appear to be closer to those of BWYV than PEMV1 (reviewed in *27* and *28*). There may be other possible explanations for the success of a recombinant as RT proteins may have additional as yet unidentified functions (*13, 29*).

REFERENCES

1. ADAM, G., SANDER, E., and SHEPHERD, R. J., *Virology* **92**, 1–14 (1979).
2. VEIDT, I., LOT, H., LEISER, M., SCHEIDECKER, D., GUILLEY, H., RICHARDS, K., and JONARD, G., *Nucleic Acids Res.*, **16**, 9917–9932 (1988).
3. WATERHOUSE, P. M., MARTIN, R. R., and GERLACH, W. L., *Phytopathology* **79**, 1215 (1989). [Abstract]
4. DOLJA, V. V., and KOONIN, E. V., *J. Gen. Virol.* **72**, 1481–1486 (1991).
5. ROSSMANN, M. G., ABAD-ZAPATERO, C., and MURTHY, M. R. N., *J. Mol. Biol.* **165**, 711–736 (1983).
6. BAHNER, I., LAMB, J., MAYO, M. A., and HAY, R. T., *J. Gen. Virol.* **71**, 2251–2256 (1990).
7. TACKE, E., PRUFFER, D., SALAMINI, F., and ROHDE, W., *J. Gen. Virol.* **71**, 2265–2272 (1990).
8. MARTIN, R. R., KEESE, P. K., YOUNG, M. J., WATERHOUSE, P. M., and GERLACH, W. L., *Annu. Rev. Phytopathol.* **28**, 341–363 (1990).
9. GUILLEY, H., WIPF-SCHEIBEL C., RICHARDS, K., and LECOQ, H., and JONARD, G., *Virology* **202**, 1012–1017 (1994).
10. UENG, P. P., VINCENT, J. R., KAWATA, E. E., LEI, C.-H., LISTER, R. M., and LARKINS, B. A., *J. Gen. Virol.* **73**, 487–492 (1992).
11. MILLER, W. A., WATERHOUSE, P. M., and GERLACH, W. L., *Nucleic Acids Res.* **16**, 6097–6111 (1988).
12. VINCENT, J. R., LISTER, R. M., and LARKINS, B. A., *J. Gen. Virol.* **72**, 2347–2355 (1991).
13. DEMLER, S. A., and DE ZOETEN G. A., *J. Gen. Virol.* **72**, 1819–1834 (1991).
14. MAYO, M. A., ROBINSON, D. J., JOLLY, C. A., and HYMAN, L., *J. Gen. Virol.* **70**, 1037–1051 (1989).
15. VAN DER WILK, F., HUISMAN, M. J., CORNELISSEN, B. J. C., HUTTINGA, H., and GOLDBACH, R. W., *FEBS Lett.* **245**, 51–56 (1989).
16. RATHJEN, J. P., KARAGEORGOS, L. E., HABILI, N., WATERHOUSE, P. M., and SYMONS, R. H., *Virology* **198**, 671–679 (1994).
17. HIGGINS, D. G., and SHARP, P. M. *CABIOS* **5**, 151–153 (1989).
18. SAITOU, N., and NEI, M., *Mol. Biol. Evol.* **4**, 406–425 (1987).
19. SWOFFORD, D. L., PAUP: *Phylogenetic Analysis Using Parsimony*, Version 3.0. Computer program distributed by the Illinois Natural History Survey, Champaign, IL (1990).
20. EFRON, B., CBMS-NSF Reg. Conf. Ser. Appl. Math. No. 38, Philadelphia, *Soc. Ind. Appl. Math* (1982).
21. PAGE, R. D. M., COMPONENT Version 2.0 Tree comparison software. Copyrighted by the Trustees of The Natural History Museum, London (1993).
22. GIBBS, A. J., and MCINTYRE, G. A., *Eur. J. Biochem.* **16**, 1–11 (1970).
23. NAGY, P. D., and BUJARSKI, J. J., *Proc. Natl. Acad. Sci. USA* **90**, 6390–6394 (1993).
24. DEVEREUX, J., HAEBERLI, P., and SMITHIES, O., *Nucleic Acids Res.* **12**, 387–395 (1984).
25. GOULDEN, M. G., LOMONOSSOFF, G. P., WOOD, K. R., and DAVIES, J. W., *J. Gen. Virol.* **72**, 1751–1754 (1991).
26. LECOQ, H., BOURDIN, D., WIPF-SCHEIBEL, C., BON, M., LOT, H., LEMAIRE, O., and HERRBACH, E., *Plant Pathol.* **41**, 749–761 (1992).
27. DUFFUS, J. E., *In* "Aphids as Virus Vectors" (K. F. Harris and K. Maramarosch, Eds.), pp. 361–383. Academic Press, New York, 1977.
28. COCKBAIN, A. J., *In* "Viruses of Plants in Australia" (C. Buchen-Osmond, K. Crabtree, A. Gibbs, and G. McLean, Eds.), pp. 268–269. The Australian National University. Research School of Biological Sciences, Canberra, 1988.
29. YOUNG, M. J., KELLEY, L., LARKIN, P. J., WATERHOUSE, P. M., and GERLACH, W. L., *Virology* **180**, 372–379 (1990).