Proc. Natl. Acad. Sci. USA
Vol. 81, pp. 4358–4362, July 1984
Biochemistry

Striking similarities in amino acid sequence among nonstructural proteins encoded by RNA viruses that have dissimilar genomic organization

(alfalfa mosaic virus/brome mosaic virus/tobacco mosaic virus/Sindbis virus/protein sequence homology)

JAMES HASELOFF*, PHILIP GOELET*†, DAVID ZIMMERN*, PAUL AHLQUIST‡§, RANJIT DASGUPTA‡¶, AND PAUL KAESBERG‡¶

*Medical Research Council Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, United Kingdom; and ‡Biophysics Laboratory and Departments of ¶Biochemistry and §Plant Pathology, University of Wisconsin–Madison, Madison, WI 53706

Communicated by T. O. Diener, April 5, 1984

ABSTRACT **The plant viruses alfalfa mosaic virus (AMV) and brome mosaic virus (BMV) each divide their genetic information among three RNAs while tobacco mosaic virus (TMV) contains a single genomic RNA. Amino acid sequence comparisons suggest that the single proteins encoded by AMV RNA 1 and BMV RNA 1 and by AMV RNA 2 and BMV RNA 2 are related to the NH_2-terminal two-thirds and the COOH-terminal one-third, respectively, of the largest protein encoded by TMV. Separating these two domains in the TMV RNA sequence is an amber termination codon, whose partial suppression allows translation of the downstream domain. Many of the residues that the TMV read-through domain and the segmented plant viruses have in common are also conserved in a read-through domain found in the nonstructural polyprotein of the animal alphaviruses Sindbis and Middelburg. We suggest that, despite substantial differences in gene organization and expression, all of these viruses use related proteins for common functions in RNA replication. Reassortment of functional modules of coding and regulatory sequence from preexisting viral or cellular sources, perhaps via RNA recombination, may be an important mechanism in RNA virus evolution.**

Viruses with single-stranded RNA genomes that infect higher eukaryotic hosts form a diverse group displaying wide variation in genomic organization (reviewed in ref. 1). The genome of tobacco mosaic virus (TMV), for example, is a single RNA molecule of 6.4 kilobases (kb) (ref. 2; reviewed in ref. 3). It encodes at least four proteins in three open reading frames. That nearest the 5′ end contains an in-phase amber termination codon that is partly suppressed during translation *in vitro* or *in vivo* to give two products, the larger (known from its molecular weight as p183) being a read-through extension of the smaller (p126). The template for translation of both of these proteins is the genomic RNA, the two remaining genes being expressed via subgenomic RNAs.

The genomes of alfalfa mosaic virus (AMV) and brome mosaic virus (BMV), in contrast, each consist of three RNA segments, termed RNAs 1, 2, and 3 in order of decreasing size (ref. 4–8; reviewed in ref. 9). The two larger RNAs of each virus are monocistronic. The smallest is dicistronic, with the 3′ proximal gene in both cases encoding the coat protein that is translated from a subgenomic mRNA. Although both viruses require all three RNAs for infection, AMV, unlike BMV, also requires either coat protein or the subgenomic mRNA for coat. Conversely, all three BMV RNAs, unlike the AMV RNAs, are aminoacylatable with tyrosine. In this respect, the BMV RNAs resemble TMV RNA (which accepts either histidine or valine according to the strain). Each virus has a different morphology, TMV being rod-shaped, AMV bacilliform, and BMV icosahedral.

All three viruses are thus clearly distinguished by conventional criteria. Nevertheless, we show in this paper that the amino acid sequences of the proteins encoded by AMV RNA 1 and BMV RNA 1 are strikingly similar both to each other and to that of TMV p126. Furthermore, the proteins encoded by AMV RNA 2 and BMV RNA 2 are also similar to each other and to the COOH-terminal read-through domain in TMV protein p183. We suggest that despite different strategies of viral gene expression, these proteins are related in function, and perhaps origin. We also show that one of these two groups of related proteins has a counterpart in a protein expressed by translational read-through and proteolytic processing that is encoded by the animal alphaviruses Sindbis and Middelburg (ref. 10; reviewed in ref. 11). We discuss these relationships and their possible implications.

MATERIALS AND METHODS

Initial homology searches were made with the matrix comparison program DIAGON (12) run on a VAX 11/780 computer. Detailed alignments were made by using the interactive facility of DIAGON and with the objective alignment programs BESTFIT and GAPOUT (13).

RESULTS AND DISCUSSION

Homologous Nonstructural Proteins in Three Plant Viruses. The recently determined nucleotide sequences of the AMV (4–6), BMV (7, 8), and TMV (2) genomes, together with earlier *in vitro* and *in vivo* viral translation studies, suggest that each virus encodes four major proteins. For brevity, we will refer to the products of AMV RNAs 1 and 2 and of BMV RNAs 1 and 2 as A1 and A2 and as B1 and B2, respectively, and to the products of each dicistronic RNA 3 as A3 and B3 (products of the 5′ proximal genes) and A4 and B4 (the coat proteins), respectively. Similarly, we will refer to the TMV open reading frames in their 5′ to 3′ order along the RNA as T1 (p126), T2 (the remainder of the first open reading frame downstream of the amber stop codon of p126, which forms the COOH terminus of the read-through product p183), T3, and T4 (the coat protein).

The amino acid sequences of these proteins predicted from the genomic RNA sequences were compared by using the computer program DIAGON. For these comparisons the program recorded as dots on a graph all pairs of 31 residue blocks whose similarity in terms of both identical and related

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviations: AMV, alfalfa mosaic virus; BMV, brome mosaic virus; TMV, tobacco mosaic virus; kb, kilobases.

†Present address: Center for Neurobiology and Behavior, College of Physicians and Surgeons of Columbia University, New York, NY 10032.

amino acids exceeded a double matching probability (12) of $\approx 10^{-5}$, compared to random pairs of 31 residue blocks drawn from pools with the same amino acid composition as the proteins under comparison. Extensive sequence similarities were detected between proteins A1, B1, and T1 and between proteins A2, B2, and T2 (Fig. 1). It can be seen from both the diagonal plots and from more detailed alignments (Figs. 2 and 3) that the proteins A1, B1, and T1 share two main regions of related sequence situated within the NH_2-terminal and COOH-terminal one-thirds of the proteins, with substantially less homology existing in the middle thirds, where protein B1 is about 130 amino acids shorter than the other two. Proteins A2, B2, and T2 share a main region of related sequence that is in the central to COOH-terminal 400 amino acids of A2 and B2. T2, which is some 300 residues shorter than A2 and B2, apparently lacks sequences corresponding to the NH_2-terminal portions of these proteins. Mutants in A2 are known to map in two complementation groups (14), suggesting the existence of two functional domains. It is possible that only one of these is represented in the TMV genome.

The significance of the observed sequence homology is supported by the conserved arrangement and clustered nature of homologous residues in both groups of proteins. For example, A1 residues 836–846, B1 residues 683–693, and 9 of 11 T1 residues 831–841 are identical (positions 932–942 in the alignment shown in Fig. 2). Similarly, 13 of 15 A2 residues 528–542, 13 of 15 B2 residues 463–477, and 12 of 15 T2 residues 1404–1418 are identical (positions 284–298 in Fig. 3). A2 and B2 have identical amino acids in about 30% of positions, rising to about 40% when conservative substitutions are also counted. These figures are clearly above the background of random resemblances in protein sequences (15). A1 and B1 and both TMV proteins are less closely related overall (20%), but random resemblances would not be expected to cluster in the manner we observe in six independent pairwise comparisons. Accordingly, we suggest that all three representatives of both groups of proteins are structurally related and potentially functionally homologous, although there may be some latitude for specialization due to differences in folding outside the most highly conserved domains.

Comparisons of the sequences of the remaining AMV, BMV, and TMV encoded proteins revealed a limited number of matches significant at the 10^{-4} level between A3 and B3 and between A4 and B4 that fell on the diagonal, but none of any significance between the others. It is not clear if any of the proteins are related in three-dimensional structure, but have insufficient conservation of primary sequence for their similarity to be apparent, or whether they are completely unrelated. The sequence relationships between the proteins of AMV, BMV, and TMV are shown schematically in Fig. 4.

Having established which parts of the A1/B1/T1 and A2/B2/T2 amino acid sequences were strongly conserved by analyzing the plant viral sequences, we searched for related sequences in other viruses. In particular, we examined the sequence of the nonstructural proteins of two alphaviruses, Sindbis virus and Middelburg virus, where an opal termination codon interrupts a long open reading frame (10). We found that the 616-residue read-through portion of this open reading frame, encoding a protein known as ns72, contained many of the same clusters of conserved residues identified in A2, B2, and T2 and that these were arranged in the same order, giving rise to diagonal lines on a matrix comparison (Fig. 1) and enabling us to align the sequences as shown in Fig. 3. This is consistent with conservation of functionally important residues within homologous, but distantly related, proteins.

It is intriguing that the potentially homologous T2 and ns72 proteins are both expressed by suppression of translational

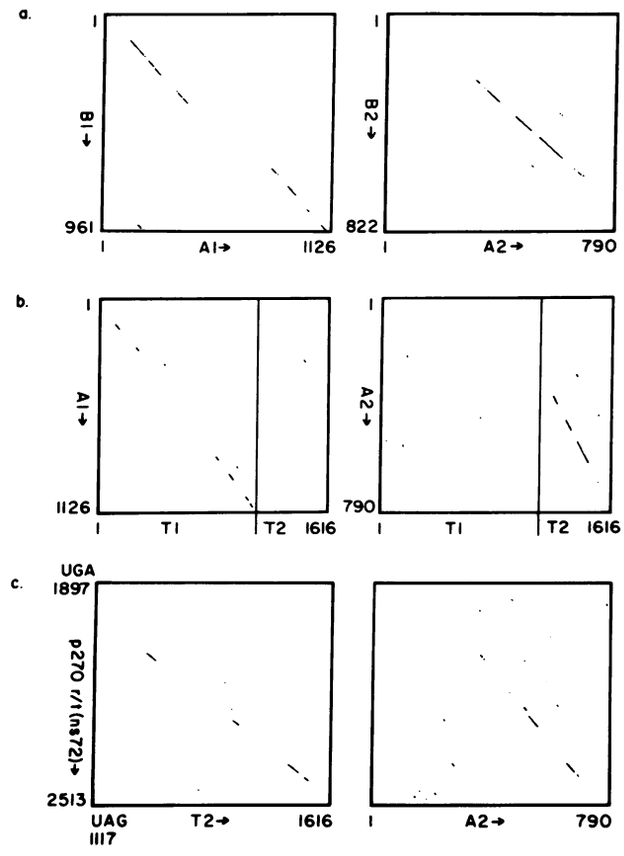

FIG. 1. DIAGON amino acid homology plots comparing B1 vs. A1 and B2 vs. A2 (*a*), A1 vs. T1 + T2 and A2 vs. T1 + T2 (*b*), and the read-through portion of Sindbis virus p270 (ns72) vs. T2 and A2 (*c*).

termination. Exploitation of translational read-through probably does not depend on chance events alone, since there is evidence for specialized natural opal suppressor tRNAs in both cattle and chickens (16, 17), whereas amber suppression of the T1 terminator may utilize a naturally occurring undermodified form of tyrosine tRNA (18).

Functional Implications. Although RNAs 1, 2, and 3 (together with coat protein or its subgenomic RNA for AMV) are required for a productive infection by AMV or BMV, inoculation of cowpea protoplasts with AMV particles containing RNAs 1 and 2 only results in symmetric synthesis of plus and minus viral RNA strands (19). Inoculation with RNA 1 alone, RNA 2 alone, RNAs 1 + 3, or RNAs 2 + 3 does not result in RNA replication. Mutations in AMV RNAs 1 and 2 interfere with RNA synthesis (14, 20). Similarly, inoculation of barley protoplasts with BMV RNAs 1 and 2 alone resulted in synthesis of B1 and B2 translation products consistent with amplification of their templates (21). Mutants in RNA 1 of the closely related virus cowpea chlorotic mottle virus are deficient in RNA replication (22). These studies indicate that the products encoded by AMV and BMV RNAs 1 and 2 are required for viral RNA replication. Replication-deficient TMV mutants are also known (3), although these have not been mapped.

Alphaviruses encode four early proteins that are translated from a 42S (*ca.* 12 kb) mRNA apparently identical to that packaged into virus particles and distinct from the 26S subgenomic mRNA for the structural proteins (11). Translation of 42S RNA results in a major polyprotein that is proteolytically cleaved to give three mature products and a minor read-through polyprotein that is cleaved into four products (10), the additional read-through protein being the one ho-

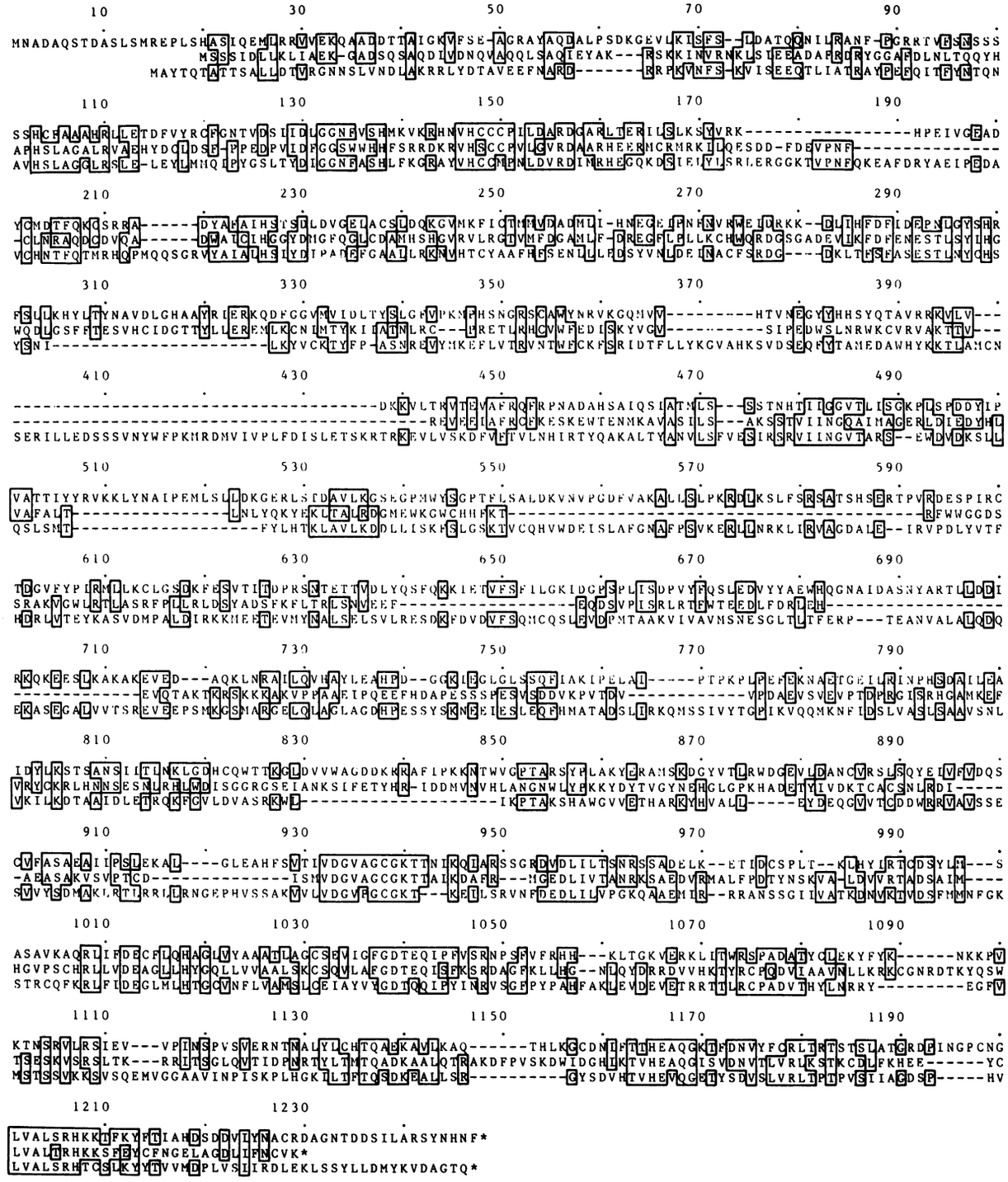

FIG. 2. Alignment of the entire sequences of proteins A1 (AMV RNA 1 product 1a; top row), B1 (BMV RNA 1 product 1a; middle row), and T1 (TMV protein p126; bottom row). Percentage identities in the pairwise alignments (including gaps) are A1 vs. B1, 20.8%; A1 vs. T1, 18.5%; and B1 vs. T1, 17.7%.

mologous to A2, B2, and T2. These four early proteins may correspond to the four complementation groups that have been assigned to replication-defective mutants of Sindbis virus and that are required for elongation (group F), minus strand synthesis (group B), and subgenomic RNA synthesis (groups A and G) (11). Thus, protein ns72 of alphaviruses is also implicated in RNA replication.

Although the available genetic evidence is compatible with the idea that either or both groups of proteins may be components of the viral replicases, they might alternatively, or in addition, be involved in more specialized roles in RNA replication such as those revealed by genetic analysis in Sindbis virus. For example, all four groups of viruses considered here have capped RNAs (5′-terminal m^7GpppG caps for the plant viruses, m^7GpppA for alphaviruses). The alphavirus cap structure is unusual among animal cell or viral RNAs in lacking a ribose methylation on the 5′-terminal nucleotide of the chain proper, and there is evidence that capping is performed by an early viral function (23). An enzyme involved in capping (perhaps coupled to the initiation of plus strand synthesis) might thus be encoded by all four groups of viruses. All four groups also use subgenomic RNA synthesis to

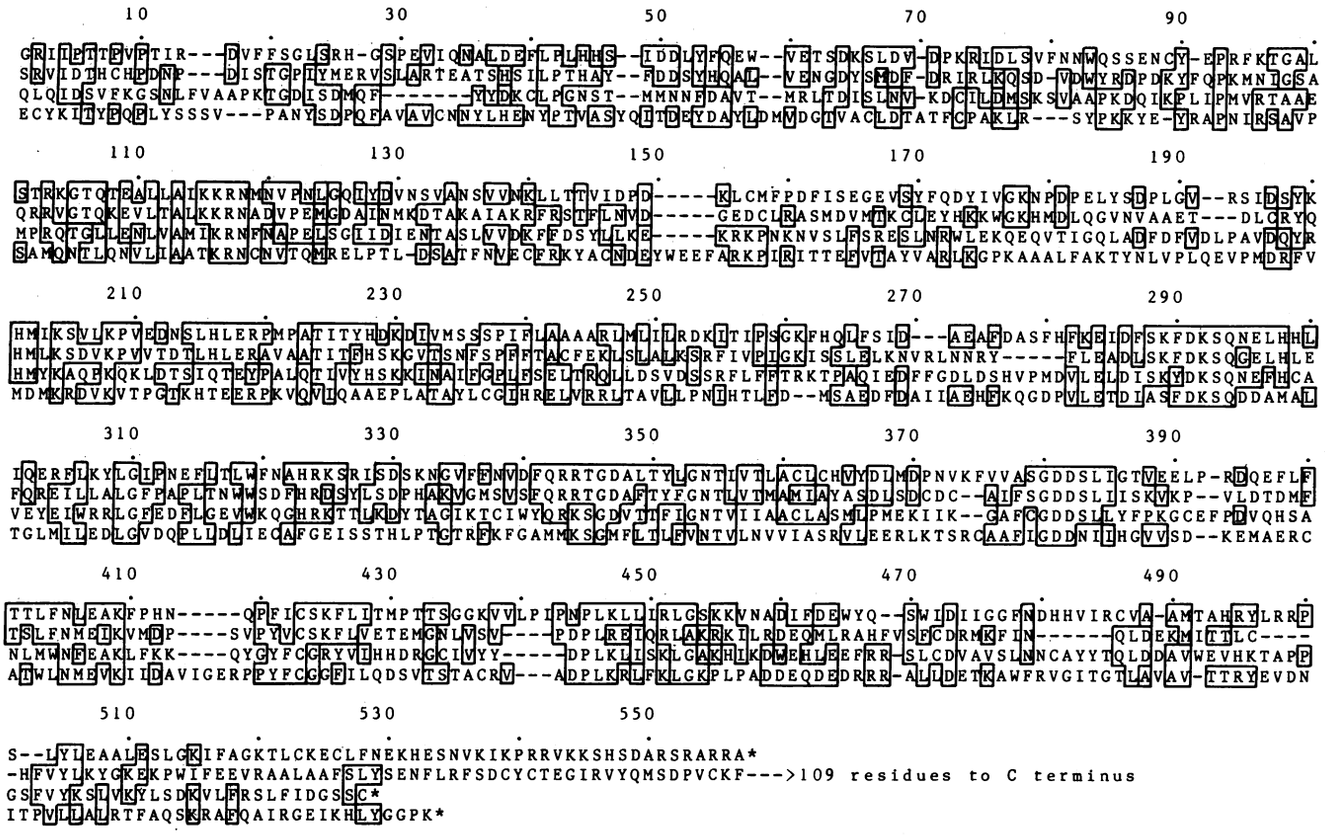

FIG. 3. Alignment of the homologous portions of the amino acid sequences of AMV RNA 2 product 2a (A2; top line), BMV RNA 2 product 2a (B2; second line), the read-through domain of TMV p183 (T2; third line), and ns72, the read-through portion of Sindbis virus p270 (bottom line). The figure is numbered for reference only; sequences begin at residue 265 of A2, 202 of B2, and 1 of T2 (immediately after the p126 termination codon) and, for ns72, 101 residues after the opal codon in the Sindbis nonstructural cistron. Percentage identities in the pairwise alignments (including gaps) are A2 vs. B2, 30.8%; A2 vs. T2, 21.6%; A2 vs. ns72, 18.5%; B2 vs. T2, 21.4%; B2 vs. ns72, 18.3%; T2 vs. ns72, 20.0%.

control expression of their structural proteins temporally and quantitatively. This is therefore another candidate for a potential common function for a homologous protein.

Evolutionary Implications. The structural similarities between these proteins may reflect either convergent evolution due to common functions or common origins in preexisting viral or host genes or some combination of these possibilities, which we consider in turn.

General necessities of RNA replication, such as the ability to bind RNA, might account for a degree of resemblance among the enzymes responsible. In particular, one might question whether the homology observed among the three plant viruses and the alphaviruses is due to convergence. Sequence comparisons alone cannot provide a definite answer. However, the demonstration of amino acid sequence homology between a protein encoded by cauliflower mosaic virus and the reverse transcriptases of both retroviruses and hepadnaviruses (24), and between the replicases of cowpea mosaic virus and picornaviruses (25), suggests that several groups of plant and animal viruses may use similar proteins in their replication. Since each group uses a distinctive set of proteins to achieve the common end of replicating an RNA template, we suggest that each group of proteins owes its common features to descent from a common ancestor.

The genomes of AMV and BMV, which are similar apart from their 3′ termini, clearly evoke a common viral ancestry, but the TMV genome also has many structural motifs in common with the tripartite viruses. Each virus encodes four well-characterized translation products (Fig. 4), of which the two largest show clearly identifiable amino acid sequence homology. The RNA termini show similarities as already outlined. Possibly, a TMV-like virus could be generated by the fusion of all three RNA segments of a tripartite virus to form a single RNA, providing that control sequences appropriate to the expression of T2 and T3 were generated and that the particle became adapted to carry a larger RNA. Conversely, a segmented virus could be derived from a TMV-like progenitor by fission, provided that the fragments became able to replicate and be encapsidated.

The three plant viruses and the alphaviruses might also have descended from a common viral ancestor whose existence predated the divergence of the plant and animal kingdoms. This would necessitate extraordinary selection pressures at the protein level given the high mutation rate of RNA virus genomes (26, 27). Alternatively, a more recent ancestral virus might have existed that could replicate in both plant and animal cells, like the reo- and rhabdoviruses of plants that also multiply in their insect vectors (28), but this does not readily account for the major differences in genomic organization contrasted with the conservation of protein sequence. It seems more attractive to explain this in a different way.

It is clearly necessary to postulate some form of recombination to account for interconversion of the tripartite and TMV genomes during evolution, even assuming a common viral ancestry. Given recombination, it seems equally possible that similar genes may be incorporated independently into different viral genomes from a separate common source, presumably cellular genes. There are at least two possibilities for such a recombination mechanism. Although reverse transcription is not thought to be involved in normal replication by these viruses, rare reverse transcription such as may

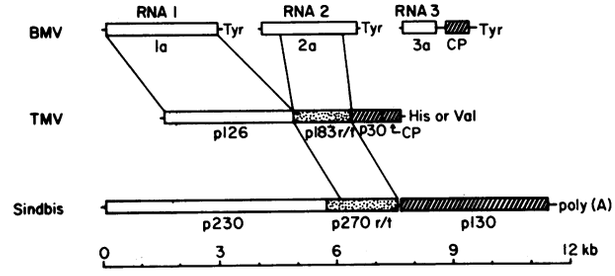

FIG. 4. Schematic diagram of the genomes of BMV (AMV is similar except that the AMV RNAs cannot be aminoacylated), TMV, and Sindbis virus. Protein homologies are marked by connecting lines. Genes expressed by suppression of translational termination are shown stippled and those expressed via subgenomic RNAs are crosshatched. All RNAs shown bear 5′ caps. 3′ poly(A) or amino acid accepting structures are marked as appropriate.

occur during pseudogene formation by cellular mRNAs (29) could be followed by recombination at the DNA level. Another distinct possibility is that recombination may occur by some as yet unspecified mechanism at the RNA level. There is genetic and biochemical evidence for RNA recombination in picornaviruses (30–32), and it is also implicated in the generation of defective interfering RNAs (DI RNAs) in many viruses, including alphaviruses (33, 34). Either or both of these mechanisms might modify viral genomes by recombination with cellular genes or be responsible for assembling genes progressively to form the viruses in the first place. The recent discovery of a Sindbis virus DI RNA with a covalently attached cellular tRNA at its 5′ end (34) is direct evidence that such genetic exchanges are possible, whatever their mechanism. On a formal basis, the differences in genomic organization of these four viruses can be regarded as permutations of modules of related genetic information and of controlling elements appropriate to their distribution along one or more viral RNAs. Ultimately, all of the modules of information whose reassortment we observe as viruses with different structures may be cellular in origin. In that case, the sequence conservation displayed by the proteins considered here may reflect strong cellular evolutionary conservation, maintained after transduction by residual functional constraints and by the need to interact with the products of other genes that remain in the more slowly evolving host cell.

Note Added in Proof. Examination of the recently published complete Sindbis virus RNA sequence (35) shows that the Sindbis nsP1 and nsP2 proteins are related to the A1/B1/T1 group of proteins shown aligned in Fig. 2 above (unpublished results).

P.K. acknowledges the support of National Institutes of Health Public Health Service Grants AI-1466 and AI-15342 and Career Award AI-21942 during this work. J.H. is a Commonwealth Scientific and Industrial Research Organisation Postdoctoral Fellow and P.G. was a Thomas C. Usher Fellow.

1. Strauss, E. G. & Strauss, J. H. (1983) *Curr. Top. Microbiol. Immunol.* **105,** 1–97.
2. Goelet, P., Lomonossoff, G. P., Butler, P. J. G., Akam, M. E., Gait, M. J. & Karn, J. (1982) *Proc. Natl. Acad. Sci. USA* **79,** 5818–5822.
3. Hirth, L. & Richards, K. E. (1981) *Adv. Virus Res.* **26,** 145–199.
4. Cornelissen, B., Brederode, F., Moorman, R. & Bol, J. (1983) *Nucleic Acids Res.* **11,** 1253–1265.
5. Cornelissen, B., Brederode, F., Veeneman, G., van Boom, J. & Bol, J. (1983) *Nucleic Acids Res.* **11,** 3019–3025.
6. Barker, R., Jarvis, N., Thompson, D., Loesch-Fries, L. & Hall, T. (1983) *Nucleic Acids Res.* **11,** 2881–2891.
7. Ahlquist, P., Dasgupta, R. & Kaesberg, P. (1984) *J. Mol. Biol.* **172,** 369–383.
8. Ahlquist, P., Luckow, V. & Kaesberg, P. (1981) *J. Mol. Biol.* **153,** 23–38.
9. van Vloten-Doting, L. & Jaspars, E. M. J. (1977) in *Comprehensive Virology*, eds. Fraenkel-Conrat, H. & Wagner, R. R. (Plenum, New York), Vol. 11, pp. 1–53.
10. Strauss, E. G., Rice, C. M. & Strauss, J. H. (1983) *Proc. Natl. Acad. Sci. USA* **80,** 5271–5275.
11. Schlesinger, R. W., ed. (1980) *The Togaviruses* (Academic, New York).
12. Staden, R. (1982) *Nucleic Acids Res.* **10,** 2951–2961.
13. Devereaux, J., Haeberli, P. & Smithies, O. (1984) *Nucleic Acids Res.* **12,** 387–395.
14. Roosien, J. & van Vloten-Doting, L. (1982) *J. Gen. Virol.* **63,** 189–198.
15. Doolittle, R. F. (1981) *Science* **214,** 149–159.
16. Diamond, A., Dudock, B. & Hatfield, D. (1981) *Cell* **25,** 497–506.
17. Hatfield, D. L., Dudock, B. S. & Eden, F. C. (1983) *Proc. Natl. Acad. Sci. USA* **80,** 4940–4944.
18. Bienz, M. & Kubli, E. (1981) *Nature (London)* **294,** 188–190.
19. Nassuth, A. & Bol, J. (1983) *Virology* **124,** 75–85.
20. Nassuth, A., Alblas, F. & Bol, J. (1981) *J. Gen. Virol.* **53,** 207–214.
21. Kiberstis, P. A., Loesch-Fries, L. S. & Hall, T. C. (1981) *Virology* **112,** 804–808.
22. Dawson, W. O. (1981) *Virology* **115,** 130–136.
23. Cross, R. K. (1983) *Virology* **130,** 452–463.
24. Toh, H., Hayashida, H. & Miyata, T. (1983) *Nature (London)* **305,** 827–829.
25. Franssen, H., Leunissen, J., Goldbach, R., Lomonossoff, G. & Zimmern, D. (1984) *EMBO J.* **3,** 855–861.
26. Domingo, E., Sabo, D., Taniguchi, T. & Weissmann, C. (1978) *Cell* **13,** 735–744.
27. Holland, J., Spindler, K., Horodyski, F., Grabeau, E., Nichol, S. & VandePol, S. (1982) *Science* **215,** 1577–1585.
28. Matthews, R. E. F. (1981) *Plant Virology* (Academic, New York), 2nd Ed., pp. 591–598.
29. Hollis, G. F., Hieter, P. A., McBride, O. W., Swan, D. & Leder, P. (1982) *Nature (London)* **296,** 321–325.
30. Cooper, P. D. (1977) in *Comprehensive Virology*, eds. Fraenkel-Conrat, H. & Wagner, R. R. (Plenum, New York), Vol. 9, pp. 133–207.
31. King, A. M. Q., McCahon, D., Slade, W. R. & Newman, J. W. I. (1982) *Cell* **29,** 921–928.
32. Tolskaya, E. A., Romanova, L. A., Kolesnikova, M. S. & Agol, V. I. (1983) *Virology* **124,** 121–132.
33. Lehtovaara, P., Soderlund, H., Keranen, S., Pettersson, R. F. & Kaariainen, L. (1981) *Proc. Natl. Acad. Sci. USA* **78,** 5353–5355.
34. Monroe, S. S. & Schlesinger, S. (1983) *Proc. Natl. Acad. Sci. USA* **80,** 3279–3283.
35. Strauss, E. G., Rice, M. & Strauss, J. H. (1984) *Virology* **133,** 92–110.