Evolution and Taxonomy of Positive-Strand RNA Viruses: Implications of Comparative Analysis of Amino Acid Sequences

Eugene V. Koonin[1]* and Valerian V. Dolja[2]

[1]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, and [2]Biology Department, Texas A&M University, College Station, TX 77843.

Referee: Dr. T. Jack Morris, School of Biological Sciences, 348 Manter Hall, University of Nebraska, Lincoln.

* Corresponding author.

ABSTRACT: Despite the rapid mutational change that is typical of positive-strand RNA viruses, enzymes mediating the replication and expression of virus genomes contain arrays of conserved sequence motifs. Proteins with such motifs include RNA-dependent RNA polymerase, putative RNA helicase, chymotrypsin-like and papain-like proteases, and methyltransferases. The genes for these proteins form partially conserved modules in large subsets of viruses. A concept of the virus genome as a relatively evolutionarily stable "core" of housekeeping genes accompanied by a much more flexible "shell" consisting mostly of genes coding for virion components and various accessory proteins is discussed. Shuffling of the "shell" genes including genome reorganization and recombination between remote groups of viruses is considered to be one of the major factors of virus evolution.

Multiple alignments for the conserved viral proteins were constructed and used to generate the respective phylogenetic trees. Based primarily on the tentative phylogeny for the RNA-dependent RNA polymerase, which is the only universally conserved protein of positive-strand RNA viruses, three large classes of viruses, each consisting of distinct smaller divisions, were delineated. A strong correlation was observed between this grouping and the tentative phylogenies for the other conserved proteins as well as the arrangement of genes encoding these proteins in the virus genome. A comparable correlation with the polymerase phylogeny was not found for genes encoding virion components or for genome expression strategies. It is surmised that several types of arrangement of the "shell" genes as well as basic mechanisms of expression could have evolved independently in different evolutionary lineages.

The grouping revealed by phylogenetic analysis may provide the basis for revision of virus classification, and phylogenetic taxonomy of positive-strand RNA viruses is outlined. Some of the phylogenetically derived divisions of positive-strand RNA viruses also include double-stranded RNA viruses, indicating that in certain cases the type of genome nucleic acid may not be a reliable taxonomic criterion for viruses.

Hypothetical evolutionary scenarios for positive-strand RNA viruses are proposed. It is hypothesized that all positive-strand RNA viruses and some related double-stranded RNA viruses could have evolved from a common ancestor virus that contained genes for RNA-dependent RNA polymerase, a chymotrypsin-related protease that also functioned as the capsid protein, and possibly an RNA helicase.

KEY WORDS: virus evolution, multiple alignment, phylogenetic taxonomy, RNA dependent RNA polymerase.

I. INTRODUCTION

One of the major conceptual assets of molecular virology during the past decade was the finding that relationships between apparently quite disparate groups of plant, animal, and bacterial viruses could be unraveled by use of comparative analysis of amino acid sequences of viral proteins. This revelation was far from being a truism

as the potential of viruses to evolve rapidly has been recognized early enough (Holland et al., 1982). Positive-strand RNA viruses represent the largest class of viruses (Francki et al., 1991), and the contribution of computer-assisted comparative studies to the understanding of the relationships within this class has been perhaps the greatest. The first studies that have revealed non-trivial links between positive-strand RNA viruses in-

fecting plants and animals (Franssen et al., 1984; Haseloff et al., 1984; Argos et al., 1984; Kamer and Argos, 1984; Blinov et al., 1984; Gorbalenya et al., 1985; Ahlquist et al., 1985) have given the start to the rapid development of the avenue of research that may be called "molecular macroevolution of viruses". We use the term macroevolution to indicate that the emphasis in this type of research is the analysis of relationships between superficially highly diverse groups of viruses. If such different groups of viruses are evolutionarily related, this evolution should have included drastic changes associated with macroevolution in classic evolutionary theory (Simpson, 1944).

Unexpected as it may seem, now, 8 years after these initial daring attempts, the area appears to be becoming self-contained and summation of the results seems timely. Table 1 shows that complete genome sequences are already available for most of the groups of positive-strand RNA viruses at the genus level as defined in the latest virus taxonomy (Francki et al., 1991). Clearly, the situation is quite different at the species and lower levels. With the progress of the sequencing techniques, numerous new isolates and strains of viruses are being sequenced at a high rate but they increasingly tend to fall within already recognized groups. Obviously, the collection of viral sequences now available is biased toward a disproportionally high representation of viruses infecting man and economically important animals, plants, and microorganisms. It seems very unlikely that this trend might change in the foreseen future to accommodate massive investigation of viruses infecting such organisms as, say, Fungi, Protozoa, or Algae, which actually appear to account for the major part of the evolutionary spun of the eukaryotes (Sogin, 1991). We only have to hope that the diversity of the genome structure of positive-strand RNA viruses as a whole does not reach far beyond the limits set by the current collection. The available information on viruses infecting taxonomically distant hosts, however incomplete, is not incompatible with such a hope.

In the past few years several reviews have dealt with different aspects of the evolution of positive-strand RNA viruses (Matthews, 1985; Goldbach, 1986, 1987; Goldbach and Wellink,

1988; Zimmern, 1988; Strauss and Strauss, 1988; Goldbach et al., 1991; Strauss et al., 1991; Dolja and Carrington, 1992). Very recently, the results of computer-assisted analysis of the proteins involved in replication and expression of these viruses have been described in considerable detail (Gorbalenya and Koonin, 1993a). So why another review? An obvious reason is an update taking into account new sequence information and the latest results of analysis. Much more importantly, however, we here try to address the issues of virus evolution and taxonomy in a more direct way than it has been done before. After briefly summarizing the relevant results of sequence comparisons, we will proceed with an explicit discussion of possible scenarios for the evolution of the extant diversity of positive-strand RNA viruses and formulation of proposals for the amendment of the existing virus taxonomy. Although the speculative element is substantial in this type of analysis, we hope that it will be useful in at least delineating the space of logical possibilities, in which hypotheses on virus evolution should be developed.

While we intended to make this review reasonably comprehensive, some important virus proteins and the respective evolutionary problems are only mentioned briefly. These include the proteins mediating the cell-to-cell movement of plant viruses whose classification is discussed elsewhere (Koonin et al., 1991a; Mushegian and Koonin, 1993), viral glycoproteins and membrane proteins, and some other less widespread proteins. Neither did we address the evolutionary significance of overlapping open reading frames that have been implicated recently as possible intermediates in the emergence of new genes (Keese and Gibbs, 1992). Importantly, we did not attempt to include detailed discussion of the possible evolutionary pathways within well-studied virus families, for example, Picornaviridae, which have been analyzed in considerable depth (Palmenberg, 1989; Stanway, 1990). Furthermore, while we tried to cite as many papers devoted directly to virus evolution and/or reporting complete virus genome sequences as possible, work on biochemical characterization of viral proteins is cited very selectively; in part, this is complemented by a recent review (Gorbalenya and Koonin, 1993a).

TABLE 1
Sequencing of Positive-Strand RNA Virus Genomes

Family/Genus/group[a,b]	Species[c]	Abbreviation	Genome size, nt[d]	Ref.
Togaviridae				
Alphavirus	Sindbis	SNBV	11,703	Strauss et al., 1984
	Eastern equine encephalitis	EEEV	11,675	Volchkov et al., 1991.
	O'nyong-nyong	ONNV	11,835	Levinson et al., 1990
	Ross River	RRV	11,657	Faragher et al., 1988
	Semliki Forest	SFV	11,442	Takkinen, 1986
	Venezuelan equine encephalitis	VEEV	11,422	Kinney et al., 1992
Rubivirus	Rubella	RubV	9,755	Dominguez et al., 1990
Arterivirus	Equine arteritis	EAV	12,720	den Boon et al., 1991
	Lactate dehydrogenase	LDV	14,222	Godeny et al., 1993
	Lelystad	LV	15,101	Meulenberg et al., 1993
Flaviviridae				
Flavivirus	Yellow fever	YFV	10,862	Rice et al., 1993
	Cell-fusing agent	CFAV	10,695	Cammisa-Parks et al., 1992
	Dengue (serotype 1,2,3,4)	DEN	10,723 (type 2)	Irie et al., 1989
	Japanese encephalitis	JEV	10,976	Nitayaphan et al., 1990
	Kunjin	KUN	10,664	Coia et al., 1988
	West Nile	WNV	10,960	Castle et al., 1986
	Tick-borne encephalitis	TBEV	10,477	Pletnev et al., 1989
Pestivirus	Bovine viral diarrhea	BVDV	12,573	Collett et al., 1988
	Hog cholera	HoCV	12,284	Meyers et al., 1989
Hepatitis C virus group(?)	Hepatitis C	HCV	9,413	Kato et al., 1990
Coronaviridae				
Coronavirus	Infectious bronchitis	IBV	27,608	Boursnell et al., 1987
	Murine hepatitis	MHV	>31,000	Lee et al., 1991
Torovirus	Berne	BeV	partial, >20,000	Snijder et al., 1990
Caliciviridae				
Calicivirus	Feline calicivirus	FCV	7,690	Carter et al., 1992
	Rabbit hemorrhagic disease	RHDV	7,437	Meyers et al., 1991
	Southampton	SRSV	7,696	Lambsen et al., 1993
Hepatitis E virus group(?)	Hepatitis E	HEV	7,207	Tam et al., 1991
Carmovirus	Carnation mottle	CarMV	4,003	Guilley et al., 1985
	Melon necrotic spot	MNSV	4,266	Riviere and Rochon, 1990
	Turnip crinkle	TCV	4,050	Carrington et al., 1989
Leviviridae				
Levivirus	MS2 phage		3,569	Fiers et al., 1976
	GA phage		3,466	Inokuchi et al., 1986
Allolevivirus	SP phage		4,276	Hirashima et al., 1988
Luteovirus	Barley yellow dwarf	BYDV-PAV	5,677	Miller et al., 1988
		BYDV-NY-RPV	5,600	Vincent et al., 1991
	Beet western yellows	BWYV	5,641	Veidt et al., 1988
	Potato leafroll	PLRV	5,883	Van der Wilk et al., 1989
Maize Chlorotic Dwarf virus Group	Rice tungro spherical	RTSV	12,307	Shen et al., 1993
Marafavirus	Maize rayado fino	MRFV	None, ~8,000	Francki et al., 1991

RIGHTSLINK

TABLE 1 (continued)
Sequencing of Positive-Strand RNA Virus Genomes

Family/Genus/group[a,b]	Species[c]	Abbreviation	Genome size, nt[d]	Ref.
Necrovirus	Tobacco necrosis	TNV	3,759	Coutts et al., 1991
Parsnip yellow Fleck virus Group	Parsnip yellow fleck	PYFV	9,871	Turnbull-Ross et al., 1992
Picornaviridae				
Enterovirus	Polio 1	PV	7,440	Racaniello and Baltimore, 1981
	Coxsackie	CoxV-B4	7,395	Jenkins et al., 1987
	Bovine entero	BoEV	7,414	Earle et al., 1988
Hepatovirus	Hepatitis A	HAV	7,478	Najarian et al., 1985
	Simian Hepatitis A	SHAV	7,400	Tsarev et al., 1991
Cardiovirus	Encephalomyocarditis	EMCV	7,825	Bae et al., 1989
	Theiler murine encephalomyelitis	TMEV	8,098	Pevear et al., 1988
Rhinovirus	Human rhino 1,2,14,89	HRV2	7,102	Skern et al., 1985
		HRV14	7,212	Stanway et al., 1984
Aphthovirus	Foot-and-mouth disease	FMDV	~8,500	Forss et al., 1984
Echovirus 22	Echo 22	ECHO22	7,339	Hyypia et al., 1992
Cricket paralysis virus	Cricket paralysis	CRPV	Partial, 7,500	King et al., 1987; Koonin and Gorbalenya, 1992
Sobemovirus	Southern bean mosaic	SBMV	4,194	Wu et al., 1987
Tetraviridae				
Nudaurelia β virus group	Nudaurelia β virus and numerous other insect viruses		Partial, ~6,000	Agrawal and Johnson, 1992
Tombusvirus	Tomato bushy stunt	TBSV	4,776	Hearne et al., 1990
	Cucumber necrosis	CNV	4,701	Rochon and Tremaine, 1989
	Cymbidium ringspot	CyRV	4,733	Grieco et al., 1989
Tymovirus	Turnip yellow mosaic	TYMV	6,319	Morch et al., 1988
	Eggplant mosaic	EPMV	6,331	Osorio-Keese et al., 1989
	Kennedya mosaic	KYMV	6,362	Ding et al., 1990
	Ononis yellow mosaic	OYMV	6,211	Ding et al., 1989
Capillovirus	Apple stem grooving	ASGV	6,496	Yoshikawa et al., 1992
Carlavirus	Potato M	PMV	8,535	Zavriev et al., 1991
	Shallot virus X(?)	ShVX	8,890	Kanyuka et al., 1992
Closterovirus	Beet yellows	BYV	15,480	Agranovsky et al., 1991b; 1993
	Citrus tristeza	CTV	Partial, 20,000	Sekiya et al., 1991
	Apple chlorotic leafspot	ACLV	7,555	German et al., 1990
Potexvirus	Potato virus X	PVX	6,432	Kraev et al., 1988
	Foxtail mosaic	FMV	6,151	Bancroft et al., 1991
	Narcissus mosaic	NMV	6,955	Linthorst et al., 1989
	Papaya mosaic	PMV	6,656	Sit et al., 1989
	White clover mosaic	WClMV	5,845	Forster et al., 1988
	Strawberry mild yellow edge-associated	SMYEAV	5,966	Jelkmann et al., 1992

TABLE 1 (continued)
Sequencing of Positive-Strand RNA Virus Genomes

Family/Genus/ group[a,b]	Species[c]	Abbreviation	Genome size, nt[d]	Ref.
Potyvirus	Potato virus Y	PVY	9,704	Robaglia et al., 1989
	Pea seed-borne mosaic	PSBMV	9,925	Johansen et al., 1991
	Plum pox	PPV	9,741	Maiss et al., 1989
	Tobacco etch	TEV	9,497	Allison et al., 1986
	Tobacco vein mottling	TVMV	9,472	Domier et al., 1986
	Pepper mottle virus	PeMV	9,640	Vance et al., 1992
Bymovirus	Barley yellow mosaic	BaYMV	11,217(2)	Kashiwazaki et al., 1990, 1991
Tobamovirus	Tobacco mosaic	TMV	6,395	Goelet et al., 1982
Comovirus	Cowpea mosaic	CPMV	9,370(2)	Lomonossoff and Shanks, 1983; van Wezenbeck et al., 1983
	Red clover mottle	RCMV	9,576(2)	Shanks et al., 1986; Shanks and Lomonossoff, 1992
Dianthovirus	Red clover necrotic mosaic	RCNMV	5,388(2)	Lommel et al., 1988; Xiong and Lommel, 1989
Fabavirus	Broad bean wilt	BBWV	None	Francki et al., 1991
Nepovirus	Tomato black ring	TBRV	12,018 (2)	Meyer et al., 1986; Greif et al., 1988
	Hungarian grapevine chrome mosaic	GCMV	11,653 (2)	Le Gall et al., 1989; Brault et al., 1989
	Grapevine fanleaf	GFLV	11,116 (2)	Serghiniet al., 1990; Ritzenhaler et al., 1993;
Nodaviridae Nodavirus	Black beetle	BBV	4,504(2)	Dasgupta et al., 1984; Dasmahapatra et al., 1985
Pea enation Mosaic virus Group	Pea enation mosaic	PEMV	9,959(2)	Demler and de Zoeten, 1991; Demler et al., 1993
Furovirus	Soilborne wheat mosaic	SBWMV	10,690 (2)	Shirako and Wilson, 1993
	Beet necrotic yellow vein	BNYVV	11,358 (2)	Bouzoubaa et al., 1986, 1987
Tobravirus	Tobacco rattle	TRV	8,590(2)	Angenent et al., 1986; Hamilton et al., 1987
	Pea early browning	PEBV	9,429(2)	MacFarlane et al., 1989; Goulden et al., 1990
Idaeovirus(?)	Raspberry bushy dwarf	RBDV	7,680(2)	Natsuaki et al., 1991; Ziegler et al., 1992
Bromovirus	Brome mosaic	BMV	8,210(3)	Ahlquist et al., 1981, 1984
	Broad bean mottle	BBMV	8,250(3)	Dzianott and Bujarski, 1991; Romero et al., 1992
	Cowpea chlorotic mottle	CCMV	8,118(3)	Allison et al., 1989; Dzianott and Bujarski, 1991
Cucumovirus	Cucumber mosaic	CMV	8,641(3)	Gould and Symons, 1982; Rezaian et al., 1984,1985
	Peanut stunt	PSV	8,487(3)	Karasawa et al., 1991,1992
	Tomato aspermy	TAV	8,698(3)	Bernal et al., 1991; Moriones et al., 1991; O'Reilly et al., 1991

TABLE 1 (continued)
Sequencing of Positive-Strand RNA Virus Genomes

Family/Genus/group[a,b]	Species[c]	Abbreviation	Genome size, nt[d]	Ref.
llarvirus	Tobacco streak	TSV	Partial, ~9,000	Cornelissen et al., 1984
Alfalfa mosaic Virus group	Alfalfa mosaic	AlMV	8,274(3)	Cornelissen et al., 1983a,b; Barker et al., 1983
Hordeivirus	Barley stripe mosaic	BSMV	9,848(3)	Gustafson et al., 1986, 1987, 1989

[a] The table includes all groups at the genus level approved by the International Committee for Classification and Taxonomy of Viruses (ICTV) and listed in the same order as in the latest ICTV Report (Francki et al., 1991), with minor amendments (indicated by "?") based on the subsequent publications.

[b] The currently used classifications of plant and animal viruses differ in that the former does not include families but only groups that are considered to be roughly equivalent to genera (Francki et al., 1991). Accordingly, the names of animal virus families and plant virus groups are shown flush left, whereas the names of animal virus genera are shown in lower case and indented.

[c] Generally, the type member is indicated first, with all viruses, for which complete sequences were available in GenBank (Release 74, amended at the National Center for Biotechnology Information, NIH, as of February 4, 1993), listed below. For Picornaviridae, only selected data are included. Where complete sequences were unavailable, partial sequences are cited, with the best available genome size estimate. The word "virus" is omitted in each case for brevity.

[d] The sequences were from GenBank; some of them may lack short terminal regions. The number of genome segments is indicated in parentheses.

II. BRIEF NOTES ON THE METHODOLOGY

It is beyond the scope of this review to describe in any detail the methods for computer-assisted analysis of amino acid sequences. Many of such methods have been compiled in several recent volumes (Doolittle, 1986, 1990; Von Heijne, 1987). However, it is appropriate to discuss here some of the peculiarities in the application of these methods to proteins encoded by positive-strand RNA virus genomes. As already mentioned, a striking feature of RNA viruses is their rapid evolution resulting in enormous sequence divergence even among apparently closely related viruses (Holland et al., 1982; Domingo et al., 1985; Steinhauer and Holland, 1987). The degree of sequence conservation between homologous viral proteins, for example, RNA-dependent RNA polymerases that are the principal enzymes of virus replication, is much lower than between enzymes with analogous functions (DNA-dependent DNA and RNA polymerases) from eubacteria, archaebacteria, and eukaryotes (data not shown).

As a rule, among distant groups of viruses, only short amino acid sequence motifs thought to be directly involved in enzymatic function are conserved (Koonin and Gorbalenya, 1989). This situation directly influences the choice of methods for computer analysis of viral protein sequence, seriously delimiting the use of database searches that generally are central to the area of sequence analysis. Instead, of critical importance are methods for knowledge-based identification of functional motifs. Usually finding such motifs constitutes the first step in comparative studies on viral genomes. An alternative is a detailed search for sequence segments of statistically most significant similarity regardless of their functional assignment. These initial steps of analysis define those portions of viral proteins that are used further for construction of multiple sequence alignments. Aligning viral proteins is a difficult exercise because of the limited conservation observed but it is greatly facilitated by the availability of multiple sequences with widely varying degree of similarity to each other. Generation of multiple

alignments in the stepwise manner starting with the closest sequences and proceeding to more distant ones is common practice (Doolittle, 1986) but this is particularly important for groups of viral proteins, in which the overall conservation may be limited to a few conserved amino acid residues in the principal functional motifs. An important element of the research strategy in comparative studies of small genomes like those of positive-strand RNA viruses is what may be called "gene context analysis". As it will be obvious from the subsequent sections, not only sequence motifs in individual proteins but in part also the gene order is conserved even among distantly related viruses allowing predictions to be made for new species. This provides for even very weak motifs that cannot be considered significant as such to serve as identifiers of function and basis for subsequent analysis if they are found to be encoded in the expected portion of the genome (e.g., Gorbalenya et al., 1989a; Koonin et al., 1992).

These initial steps of comparative sequence analysis account for the peculiarities of comparative studies in virology. The subsequent phylogenetic analysis, and specifically construction and interpretation of phylogenetic trees for virus genes and proteins, share all the generic virtues and pitfalls of these approaches. Interpretation of trees for viral proteins may be particularly difficult because the number of unambiguously aligned residues is very limited. Attempts have been made to produce trees based on short conserved motifs, but the results of such analysis have proven hard to interpret in a definitive way (Candresse et al., 1990). Hence, once again, the prerequisite for any reliable phylogenetic analysis is the generation of the best possible multiple alignments.

The size of positive-strand RNA viral genomes lies within the interval from about 3.5 to about 30 kb (Table 1). This relatively small size makes positive-strand RNA virus genomes particularly attractive for a thorough phylogenetic study as construction of a comprehensive evolutionary scenario appears to be a tractable goal. Hopefully, such a scenario may serve as a model in studies of other, more complex genomes.

III. THE PARADIGM OF POSITIVE-STRAND RNA VIRUS EVOLUTION

We believe that the comparative studies on positive-strand RNA viruses performed by now have been complete enough to allow an explicit formulation of the main principles of their evolution. To make the subsequent discussion straightforward, we will put this formulation here and then will deal with each of these principles at some length.

1. RNA viruses evolve rapidly. Hence, only important functional motifs are conserved in a wide range of virus groups.

2. Positive-strand RNA virus genomes are made of a limited number of building blocks. The universal blocks are the genes for: (i) the RNA-dependent RNA polymerases (RdRp), and (ii) the coat protein. Only RdRp contains universal sequence motifs that are conserved in all positive-strand RNA viruses with known genome sequences. Therefore, phylogenetic analysis of the RdRp sequences inevitably forms the groundwork for our efforts to produce a coherent picture of the evolution of this virus class in general.

3. Evolution of positive-strand RNA viruses is shaped by two opposite trends: (i) conservation of distinct arrays of genes, primarily those encoding proteins mediating virus RNA replication, and (ii) recombinational shuffling of genes and gene blocks.

4. The wide spread of recombination, even among distantly related viruses, makes it impossible in principle to depict the evolutionary history of positive-strand RNA viruses as a single phylogenetic tree. An adequate description can only be a complex scenario providing for both vertical ("tree-like"), and horizontal flow of genetic information.

5. Correlation between virus phylogeny and strategy of genome replication and expression is only limited, suggesting that fundamental expression and replication mechanisms could have evolved more than once.

IV. THE BUILDING BLOCKS: ALIGNMENTS, CONSERVED MOTIFS, AND TENTATIVE PHYLOGENIES

A. Sequence Conservation in Positive-Strand RNA Viruses

When amino acid sequences of positive-strand RNA virus proteins were compared to each other using the BLAST program (Altschul et al., 1990), it was shown that statistically significant similarities were rather sparse and were observed mostly between relatively closely related viruses comprising compact groups (unpublished observations). These similarities were found in distinct conserved domains with already known or suspected function. The RNA-dependent RNA polymerase (RdRp) was by far the best conserved region showing the highest score for most of the compared virus pairs, with the RNA helicase holding the second position. Only in a few cases the segment with the highest similarity resided in the chymotrypsin-related protease or in the methyltransferase domain. Below we discuss the multiple alignments and the conserved sequence motifs for these and some other domains of positive-strand RNA viruses in the order of decreasing sequence conservation.

Generation of tentative phylogenetic trees is a logical sequel to construction of multiple alignments. Clearly, information on "macroevolution" of viruses could be derived only by phylogenetic analysis of the limited set of viral proteins that are conserved in a wide range of virus groups. The tentative phylogenies for these proteins should be analyzed separately because of the wide spread of gene shuffling. Subsequent comparison of different trees could provide information on the congruency of the evolution of viral genes, and on major recombination events.

B. RNA-Dependent RNA Polymerase

1. Amino Acid Sequence Comparisons

All nondefective positive-strand RNA viruses encode the RdRp. Following the seminal work by Kamer and Argos (1984), it has been recognized that all these polymerases share a set of conserved

sequence motifs (Zimmern, 1988; Morozov and Rupasov, 1985; Koonin et al., 1987; Poch et al., 1989; Koonin, 1991a). The latest, most-detailed comparative analysis (Koonin, 1991a) revealed eight such motifs (Figure 1). It has to be realized that the overall sequence similarity among the positive-strand RNA viral polymerases is quite poor. Only three motifs, IV, V, and VI, showed completely unequivocal conservation throughout the whole class, with six invariant amino acid residues (Figure 1). Despite this modest conservation, the signature $Dx_3[FYWLCA]x_{0-1}Dx_n[STM]Gx_3Tx_3[NE]x_n[GS]DD$ (x – any amino acid residue; alternative residues shown in brackets) could serve as an identifier of the RdRps of positive-strand RNA viruses and some related dsRNA viruses (and by inference of the viruses themselves), with only very few false positives retrieved in a search of the available amino acid sequence databases.

The importance of the "core" RdRp motifs IV, V, and VI for the polymerase activity has been proven by site-directed mutagenesis of the EMCV polymerase (Sankar and Porter, 1992). Although their actual function is not known, it is thought that these motifs may be involved in the binding of the NTP substrate (Koonin, 1991a).

Alignment of the remaining motifs is more tentative, based on superposition of alignments of distinct groups of RdRps. There were 11 such groups comprising three larger supergroups (Koonin, 1991a; Figure 1). The sequences within each of the groups, and in some cases the sequences belonging to different groups within a supergroup showed statistically significant similarity in pairwise comparisons (Koonin, 1991a, and unpublished observations). On the other hand, such similarities were generally not observed between the supergroups. The significance of the similarity in this case could be demonstrated only by comparing multiple alignments (Koonin, 1991a).

Positive-strand RNA virus RdRps also showed sequence similarity to other polymerases. The closest relationship was observed with the putative RdRps of double-stranded RNA viruses. Importantly, at least some of the double-stranded RNA virus polymerases showed obvious affinity with specific groups of positive-strand RNA viruses and could be considered to belong to the

```
                    I              II
cons1             ..K.E .    .&....&...& .. R&&....... .&.....
                                    K
      PV    156  YVKDE-L 12  LIEASSLNDSV-AM-RMAFGNLYA--AFHKNPG ┐
      FMDV  161  FLKDE-I 12  IVDVLPVEHIL-YT-RMMIGRFCA--QMHSNNG │
      ECHO22 163 CLKDE-L 12  CIEACEVDYCI-VY-RMIMMEIYD--KIYQTPC │
      HAV   164  CPKDE-L 12  AIDACPLDYSI-LC-RMYWGPAIS--YFHLNPG │
      FCV    ?   GLKDE-L 12  MIWGCDVGVAT-VCAAAFKGVSDAI-TANHQYG │
      RTSV   ?   CLKDE-R 13  TFTILSPEVNI-LF-RQYFGDF-A--AMVMSTR │
      CPMV  199  CPKDE-K 13  CFTILPMEYNL-VV-RRKFLNF-V--RFIMANR │
      GCMV   ?   CPKDE-L 12  LFEIMPLHYNL-LL-RVKTCAF-T--AFLQHNR │
      TEV   169  SLKAE-L 12  TFTAAPIDTLL-AG-KVCVDDFNN--QF-YDLN │
      BaYMV 197  SLKAE-L 12  VFTASPITSLF-AM-KFYVDDFNK--KF-YATN │ 1
      dsHyAV 1822 FAKSQ-A 15 TVVSEDLSAYM-VD-QIFQIEANK--RITWETY │
      dsScV 422  STKYE-W 12  STLITNFAMFR-CE-DVLTHKF----PVGDQAE │
      dsLRV1 392 SAKLE-H 12  SYMWFEYALRP-VE-RIWENSN----VILDPGS │
      BBV   512  FNKNE-- 7   IISGFPDILFI-L--KVSRYTLAYS-DIVLHAE │
      PLRV  553  FVKGE-P 12  LIMSVSLVDQL-VA-RVLFQNQNKR-EISLWRS │
      PEMV1  ?   FVKLE-P 12  LIASVSIVDQL-VA-RMLFRDQNEE-ELLQHMA │
      SBMV  215  FVKQE-P 12  LISSVSIVDQL-VE-RMLFGAQNEL-EIAEWQS │
      IBV   517  NLKYA-I 7   TVAGVSILSTM-TN-RQFHQKI----LKSIVNT │
      BEV   412  ITKFA-L 7   TVSSCSFIAST-IF-RFAHKPVTSK-MVEVAQN │
      EAV    ?   LEKYN-L 15  AYLKEEIGDAP-PL--YLPSTVPSK-NSQAGIN ┘

cons2             &.K.E K    ....R...&.......&&.........&.....
                  N R
      CarMV 358  FIKAE-K 12  VIQPRSPRYNVELG-RYLKKYEH--HAYKALDK ┐
      TBSV  446  FVKAE-K 11  VIQPRNPRYNVELG-RYLRHMES--KLMKAVDG │
      MCMV  584  FVKAE-K 12  VIQPRAPRYNVELG-RYLRPVEH--PIYHAIDK │
      PEMV2 174  FVKAE-K 12  VIQPRDPRSNIVLA-KYIKPLEP--MLYKALGK │
      RCNMV 352  FIKVE-K 11  TIQPRSKRYNLAIG-QILRLNEK--KMLDSIDD │
      BYDV  525  FLKKE-K 10  LICPRSKRYNIILGTRLK-FNEK--KIMHAIDS │
      BVDV   ?   IPKNE-K 12  EKRPRVIQYP-EAKTRL--AITK--VMYNWVKQ │
      HCV    ?   MAKNE-V 8   RKPARLIVFP-DLGVRV--CEKM--ALYDVVTK │ 2
      WNV   456  MGKRE-K 6   AKGSRAIWYMW-LGARFL-EFEA--LGFLNEDH │
      YFV   454  MGKRE-K 6   AKGSRAIWYMW-LGARYL-EFEA--LGFLNEDH │
      TBEV  454  MGKRE-K 6   AKGSRAIWYMW-LGSRFL-EFEA--LGFLNEDH │
      CFAV  438  MGKKE-K 6   AKGSRTIWYMW-LGSRFL-EFEA--LGFLNADH │
      MS2   194  VPKNN-K 0   --IDRAACKEP-DMNMYL-QKGV--GAFIRRR- │
      SP    211  VPKNS-K 0   --TDRCIAIEP-GWNMFF-QLGV--GAVLRDR- │
      W-RNA 291  SVVRE-R 0   GHKVRVVSAME-THELVLGHAARRRLFKGLRRE │
      T-RNA 321  AVVPE-R 0   GFKARIVTTHS-ASRVTFGHQFRRYLLQGIRRH ┘

cons3             &&K...K    ......U.......U...&.P&.  ..&...&.
      BSMV  444  MIKNDVK 9   YAALQTVVYPDK-IVNAFFGPII--KEINERII ┐
      TRV   1371 MIKSDVK 9   YSALQTVVYHEK-LINSLFGPIF--KEINERKL │
      TMV   1300 MYKAQPK 9   YPALQTIVYHSK-KINAIFGPLF--SELTRQLL │
      CMV   434  MIKSDVK 9   RPVPATITFHKK-TITSQFSPLF--ISLFERFQ │
      BMV   385  MLKSDVK 9   RAVAATITFHSK-GVTSNFSPFF--TACFEKLS │
      A1MV  448  MIKSVLK 9   RPMPATITYHDK-DIVMSSSPIF--LAAAARLM │
      SNBV  287  DMKRDVK 9   RPKVQVIQAAEP-LATAYLCGIH--RELVRRLT │ 3
      HEV    ?   FQKDCNK 9   GKVGQGISAWSKTFCALLFGPWF--RAIEKAIL │
      RuBV  162  FLKATLK 18  GKAGLEIRAWAKEWVQVVMSPHF--RAIQKIIM │
      BNYVV  ?   QLKDIEK 9   AKAGQGILAWSKEAHVKKFMVAF--RVLNDLLL │
      PVX   1161 FLKSQWV 9   IKPGQTIAAFYQ-QTVMLFGTM---ARYMRWFR │
      TYMV   ?   FAKAQHK 8   WKACQTLALMHD-YVILVLGPV---KKYQRIFD │
      ACLV   ?   FMKSQLC 8   AKAGQTLACFPH-KILVEFSPW---CRYTEKVL │
      ASGV   ?   FMKSQYC 8   AKAGQTLACFQH-IVLFRFGPM---LRAIESAF ┘
CONS              &.K.e.k    ......&.........&&&........&.....
```

A

FIGURE 1. Conserved sequence motifs in the RNA-dependent RNA polymerases of positive-strand RNA viruses and related double-stranded RNA viruses. Only a sampling of the available RdRp sequences is shown. An attempt was made to make it representative by including the diverse sequences within groups and excluding closely related sequences. The representation of some of the virus groups in this and the subsequent figures may differ depending on the total number of the available sequences and the level of conservation. The motifs are designated as in Koonin (1991a). The number of amino acid residues between the motifs and the distances from the protein termini are shown (question marks show that the latter are unknown, that is, in viruses that produce their RdRp by processing of a precursor polyprotein at unidentified cleavage sites). The three superfamilies are designated and the consensus including residues conserved in at least 75% of the presented sequences is shown separately for each superfamily. The general consensus includes the residues present in all three patterns (upper case) or in two patterns (lower case). In the consensus lines U designates a bulky aliphatic residue (I, L, V, M), @ designates an aromatic residue (F, Y, W), & designates a bulky hydrophobic residue (either aliphatic or aromatic), and (·) dot designates any residue). The residues conforming to the consensus are highlighted by bold typing. The sources of the positive-strand RNA virus sequences are listed in Table 1. The double-stranded (ds) RNA virus sequences were from Shapira et al., 1991 (*Cryphonectria parasitica* hypovirulence-associated virus, HyAV), Icho and Wickner, 1989 (*Saccharomyces cerevisiae* virus L-A, ScV), Stuart et al., 1992 (*Leishmania* RNA virus 1, LRV1), Rodriguez-Cousino et al., 1991 (W RNA), Esteban et al., 1992 (T RNA).

```
              III                 IV                     V
cons1    &G.......&..       .D&..0 D....          ...SG...T...NS.&...&.&....
                              C N                   I      T
   PV   5  -VGCDP-DLFWSK   11 FDYTGY-DASLS   40 KGGMPSGCSGTSIFNSMIN-NLIIRTLL ⌉
  FMDV   5  -VGCNP-DVDWQR   13 VDYSAF-DTNHC   43 EGGMPSDCSATGIINTILN-NIYVLYAL |
ECHO22   5  -VGINP-YKDWHF   11 MDYSQY-DGSLS   42 HGGMPSGSPCTTVLNSLCN-LMMCIYTT |
   HAV   5  -IGIDP-DRQWDE   14 LDFSAF-DASLS   43 CGSMPSGSPCTALLNSIIN-NVNLYYVF |
   FCV   6  -MDSPSVEALFQR    9 VDYSKW-DSTQS   43 SSGLPSGMPLTSVINSLNH-CLYVGCAI |
  RTSV   5  -VGINPESMEWSD   14 GDYSKF-DGIGS   46 SQGMPSGFAMTVIFNSFVN-YYFWALAW |
  CPMV   5  -VGINPYSMEWSR   14 CDYSSF-DGLLS   46 ECGIPSGFPMTVIVNSIFN-EILIRYHY |
  TBRV   5  -VGTNPYSREWGH   15 CDYSGF-DGLLT   50 NCGLPSGFALTVVMNSIFNEEILIRYAY |
   TEV   5  -VGMTKFYQGWNE   13 ADGSQF-DSSLT   47 HKGNNSGQPSTVVDNTLMV-IIAMLYT- |
 BaYMV   5  -VGINKFGRGWEK   14 GDGSRF-DSSID   45 NVGNNSGQPSTVVDNTLVL-MTAFLYAY |1
dsHyAV  6  -L-SQSMARIWDE   14 ADATAY-DSNCK  178 NRGGGTGQSATSWDNTATF-KLGVISAW |
 dsScV   4  ----KR-VNMMKD    6 FDYDDF-NSQHS   50 QGTLLSGWRLTTFMNTVLN-WAYMKLAG |
dsLRV1  4  --IATR-INGWRN    9 VDYDDF-NSQHT   44 AGTLMSGHRATSFINSVLN-RAYIICAG |
   BBV   7  YPGRNP-TEIADG   14 TDFSNL-DGRVS   47 GVGVKSGSSTTTPHNTQYNGCVEFTALT |
  PLRV   5  GFGLST-DTQTAE   27 TDCSGF-DWSVA   50 PGVQKSGSYNTSSSNSRIR--VMAAYHC |
 PEMV1   5  GLGFSQ-DHQVLA   26 TDCSGF-DWSVP   50 PGIQKSGSFNTSSTNSRMR--YMLALYA |
  SBMV   5  GMGLSV-IHQADA   16 ADISGF-DWSVQ   49 PGIMKSGSYCTSSTNSRIR--CLMAELI |
   IBV   5  VIGTTKFYGGWDN   16 WDYPKC-DRAMP   49 PGGTSSGDATTAYANSVFNIIQATSANV |
   BEV   5  LIGVSKYGLKFSK   16 SDYTKC-DRTFP   45 PGGTSSGDATTAHSNTFYNYMVHYVVAF |
   EAV  80  YLGKSKFDPIPAP    6 TDLESC-DRSTP   42 RGGLSSGDPITSISNTIYSLVLYTQHML ⌋

cons2    ...&........&       .D...&.D..US          ...R.SG...T...N.&U.&........
                               A     CT                    E      A
 CarMV   7  MKGYTTEEVAQHI   14 FDMSRF-DQHVS   46 EGCRMSGDMNTALGNCLL-ACLITKHLM ⌉
  TBSV   7  IKGYTADEVGAIF   14 LDASRF-DQHCS   46 EGCRMSGDINTSLGNYLL-MCAMVHGYM |
  MCMV   7  MKGYSVEQIGRHI   14 FDASRF-DQHVS   46 DGKRMSGDMNISLGNCIL-ATAITHDFV |
 PEMV2   7  AKGFNAVETGEII   14 LDASRF-DQHVS   45 KGRRMSGDMDTSLGNCVL-NVLLTRNLC |
 RCNMV   7  LSGLDNRAQGRAI   14 LDASRF-DQHCS   47 KGCRMSGDINTGLGNKIL-MCSMVHAFL |
  BYDV   7  LSGYDNFKQGRII   14 VDASRF-DQHVS   45 RGHRMSGDINTSMGNKLI-MCGMMHAYL |
  BVDV   8  EGKTPLFNIFDKV   14 FDTKAW-DTQVT   45 NGQRGSGQPDTSAGNSMLNVLTMMYAFC |
   HCV  10  GFQYSPGQRVEFL   14 YDTRCF-DSTVT   50 RASRASGVLTTSCGNTLICYIKARAA-C |2
   WNV  10  VEGEGLHRLGYIL   13 DDTAGW-DTRIT   52 KEQRGSGQVGTYGLNTFTNMEVQLSRQM |
   YFV  10  VEGIGLQYLGYVI   13 DDTAGW-DTRIT   53 RDQRGSGQVVTYALNTITNLKVQLIRMA |
  TBEV  10  VEGISLNYLGWHL   13 DDTAGW-DTKVT   53 RDQRGSGQVVTYALNTLTNIKVQLIRMM |
  CFAV  10  VGGVGVNYFGYYL   12 DDIAGW-DTKIS   57 RDQRGSGQVVTYALNTITNGKVQVARVL |
   MS2  0  LKSVGIDLNDQSI   16 IDLSSASDS-IS   33 ELFSTMGNGFTFELESMI-FWAIVKAT- |
    SP  0  LRLWKIDLNDQST   16 IDLSAASDS-IS   34 EKISSMGNGYTFELESLI-FAAIARSVC |
 W-RNA  0  RRLRDTLKGDFEA   14 SDMKSASDL-IP   44 RRGILMGLPTTWAILNLM-HLWCWDSAD |
 T-RNA  0  PALVDVIGGDHRR   20 ADLTSASDR-IP   47 RQGILMGLPTTWPLLCLI-HLFWVELSD ⌋

cons3    ........&...&       .D&S.F D.SQ.          .&.R.SGD..T&..NT&..U......&.
                             A       T    T            C K T E      S  A
  BSMV   9  NSRMTADELNETV   12 IDFSKF-DKSKT   44 LYQQKSGNCDTYGSNTWSAALALLDCLP ⌉
   TRV   9  NTRMTSSDLNDRV   13 IDMSKF-DKSAN   44 WYQQKSGDADTYNANSDRTLCALLSELP |
   TMV  10  FTRKTPAQIEDFF   13 IDISKY-DKSQN   44 WYQRKSGDVTTFIGNTVIIAACLASMLP |
   CMV   9  PVGKISSLEMTGF    9 IDLSKF-DKSQG   44 SFQRRTGDAFTYFGNTIVTMAEFAWCYD |
   BMV   9  PIGKISSLELKNV    9 ADLSKF-DKSQG   44 SFQRRTGDAFTYFGNTLVTMAMIAYASD |
  ALMV   9  PSGKFHQLFSIDA   11 IDFSKF-DKSQN   44 DFQRRTGDALTYLGNTIVTLACLCHVYD |
  SNBV   9  LFDMSAEDFDAII   12 TDIASF-DKSQD   44 GAMMKSGMFLTLFVNTVLNVVIASRVLE |3
   HEV   9  GDAFDDTVFSAAV   10 NDFSEF-DSTQN   42 FWKKHSGEPGTLLWNTVWNMAVITHCYD |
  RuBV   9  AAGHTEPEVDAWW   10 VDFTEF-DMNQT   42 GCERTSGEPATLLHNTTVAMCMAMRMVP |
 BNYVV   9  DNTMSETEFVGKI   15 IDAAAC-DSGQG   42 SYVKTSGEPGTLLGNTILMGAMLNAMLR |
   PVX  10  NCETTPEDMSAWA   12 NDYTAF-DQSQD   38 SIMRLTGEGPTFDANTECNIAYTHTKFD |
  TYMV  10  HCGKTPNQLRDWC   12 NDYTAF-DQSQH   38 TCMRLTGEPGTYDDNTDYNLAVIYSQYD |
  ACLV  10  HQRKNFSELEDFA   11 SDYTAF-DVSQD   37 AIMRFTGEFSTFLFNTLANMVFTFCRYE |
  ASGV  10  HSGKNFFCLDSFV   13 SDYTAF-DSSQD   38 AIMRFTGEFCTFLFNTFANMLFTQLKYK ⌋
  CONS     ...........&       .D...&.D....         ...r.SG...T...N&&..&........
                               k T                  t       a
```

FIGURE 1B

positive-strand RNA virus RdRp class (Koonin et al., 1989, 1991b; Bruenn, 1991; Koonin, 1992a; Figure 1).

Counterparts to the "core" RdRp motifs were detected also in RNA-dependent DNA polymerases, DNA-dependent DNA polymerases, and RdRps of negative-strand RNA viruses (Poch et al., 1989; Delarue et al., 1990; Xiong and Eickbusch, 1990). However, in contrast with the double-stranded RNA viruses, these were remote relationships and each of these enzyme classes clearly did not overlap with the positive-strand RNA virus RdRps.

2. Phylogenetic Analysis

RdRp is the only domain of positive-strand RNA viruses allowing an all-inclusive phyloge-

```
                  VI                VII              VIII
cons1         . &.&GDD.&&        ....@U...        ......U...
                  S                 C

   PV     13  MIAYGDDVIA  36   ENVTFLKRF  13   VMPMKEIHES   61
 FMDV     13  MISYGDDIVV  38   TDVTFLKRH  12   VMASKTLEAI   59
 ECHO22   10  PIVYGDDVIL  37   MEVEFLKRK  13   LLDTENMIQH   63
  HAV     16  ILCYGDDVLI  42   SELTFLKRS  10   AISEKTIWSL   69
  FCV     20  MMTYGDDGVY  39   NSVVFLKRT  10   LLDRSSILRQ  102
 RTSV     24  IVAYGDDNVV  43   TKMSFLKRG  13   PLDKTSIEER  126
 CPMV     22  LVTYGDDNLI  42   EECDFLKRT  11   PEDKASLWSQ  245
 GCMV     17  LLVYGDDNLI  42   SELDFLKRK  12   PLDKSAIFSC  361
  TEV     12  YYVNGDDLLI  33   TQLWFMSHR  10   KLEEERIVSI   77
 BaYMV    17  FVCNGDDNKF  34   CENPYMSLT  10   SLPVERIIAI   80  1
 dsHyAV   16  LYNTSDDTVW  33   TEVEYLSKL  14   WR-QGRIENM  949
 dsScV     9  SVHNGDDVMI  30   SISEFLRVE  11   QYLSRSCATL  217
 dsLRV1    5  SMHVGDDILM  30   TSGEFLRVA  11   RVISSAVSGN  256
  BBV     15  -PKCGDDGLS  27   IGLCFLSRV  10   IQDPLRTLRK  145
 PLRV      4  AMAMGDDALE  19   -ELEFCSHI   9   VNTNKMLYKL   51
 PEMV1     4  AVTMGDDALE  19   -EFDFCSHL   9   KNLEKMVYGL   50
 SBMV      4  CIAMGDDSVE  31   YAVEFCSHV   8   TSWPKTLYRF  100
  IBV     50  LMILSDDGVV  43   GPHEFCSQH  15   PDPSR-ILGA   90
  BEV     43  LNFLSDDSFI  37   HIEEFCSAH  12   PSRGR-LLAS   90
  EAV     30  VYIYSDDVVL  36   PSFLGCRFK  15   TRSLLYHIGA    ?

cons2         U...GDDU&U        ....FC...        &..UR.....
                  C                                 CK

 CarMV     5  LINNGDDCVL  34   EKIRFCQMA   7   WLMVRDPLVS  133
 TBSV      9  LANCGDDCVL  34   EEVEFCQAH   7   WKMVRNVRTA  128
 MCMV      8  LINNGDDNVL  34   EQVEFCQMR   7   YTMMRDPRTT  136
 PEMV2     8  LFNNGDDCIV  34   EKIEFCQTQ   7   WRTVRCI-SS  164
 RCNMV     8  LANNGDDCVL  34   EKVAFCRSQ   7   WAMVRQL-GS  129
 BYDV      8  LCNNGDDCVI  33   EQLEFCQSK   7   YRMVRRP-DS  142
 BVDV     15  IHVCGDDGFL  39   EDIEFCSHT  12   HMAGRDTAVI  197
  HCV      9  MLVCGDDLVV  37   ELITSCSSN  12   YYLTRDPTTP  200  2
  WNV     35  MAISGDDCVV  36   QEVPFCAHH  12   VVPCRNQDEL  166
  YFV     36  MAVSGDDCVV  36   ENVPFCSHH  12   VVPCREQDEL  167
 TBEV     33  MLVSGDDCVV  36   EEVPFCSHH  12   IVPCRDQDEL  168
 CFAV     27  MVIAGDDVVV  36   EKVEFCSHH  12   IAPCRHENEV  166
  MS2      9  IGIYGDDIIC  29   -RRESCGAH   8   F-YIKKPVDN  145
   SP      9  VSVYGDDIII  29   -RRESCGKH   8   F-YIRRPIRC  158
 W-RNA    17  CRVCGDDLIG  30   -RGVFLERL  12   VIYRKVGHRR  312
 T-RNA    17  FRICGDDLIA  30   -WGIFTEKV   3   PVKMKVRVRS  396

cons3         ..&.GDD.&U        ..P.&CG.&        ...UK& U..
                  CA                G   S            AR  A

 BSMV      6  CVFGGDDSLI  30   KYPAFCGKF  11   PDAAKF-ITK   87
 TRV       6  VTYGGDDSLI  30   DVPMFCGKF  11   PDPVKV-LTK   98
 TMV       6  GAFCGDDSLL  30   QYGYFCGRY  11   YDPLKL-ISK   76
 CMV       6  LLFSGDDSLA  28   AVPYICSKF  15   REIQRL-GTK  169
 BMV       6  AIFSGDDSLI  26   SVPYVCSKF  16   REIQRL-AKR  200
 ALMV      8  VVASGDDSLI  27   NQPFICSKF  16   PNPLKL-LIR  101
 SNBV      8  AAFIGDDNII  31   RPPYFCGGF  13   ADPLKR-LFK   80  3
 HEV       6  AAFKGDDSIV  29   PIGLYAGVV   8   PDVVRF-AGR   85
 RuBV      6  GIFQGDDMVI  34   PTPSFCGHV   7   HDVMHQ-AIK  178
 BNYVV     6  MAMKGDDGFK  30   VPITFCGYA   7   PSVSRK-LTK  120
 PVX       6  QVYAGDDSAL  33   SWPEFCGWL  13   HVSLKL-AEA   61
 TYMV      5  IMVSGDDSLI  28   SHPLFCGYY  13   FCKLMI-AVD  119
 ACLV      6  ICFAGDDMCA  27   KVPMFCGWR  13   YERLQV-AIE   97
 ASGV      7  ILFAGDDMCS  28   KFPMFCGWY  13   WARIKM-MSE    ?
 CONS         ..&.GDD.&.        ....@C...        ....k.....
                                                     r
```

FIGURE 1C

netic analysis. As the final alignment of the RdRps was rather uncertain beyond the motifs shown in Figure 1, we used these concatenated motifs to derive the overall tree topology. The branching order within distinct groups was determined using the complete unambiguous alignments for the three polymerase supergroups (Koonin, 1991a).

Figure 2 illustrates the results of phylogenetic analysis of the RdRps. In accord with the qualitative observations described above, the separation of the three vast supergroups was obvious (Figure 2A). A degree of uncertainty remains regarding the inclusion of the polymerases of RNA bacteriophages and the related dsRNA elements in supergroup 2 as they were isolated into a separate lineage by some of the tree-generating algorithms.

In turn, each of the RdRp supergroups split into well-defined lineages compatible with the
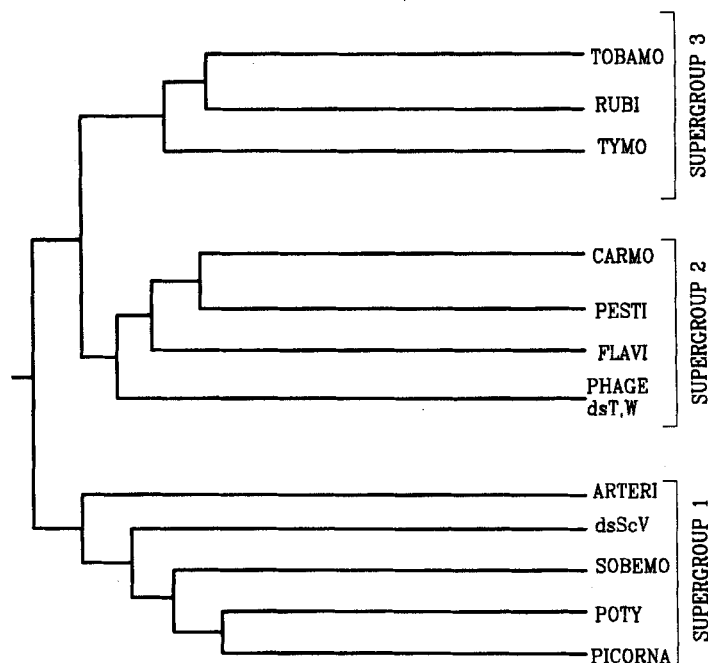
A

FIGURE 2. Tentative phylogeny of the RNA-dependent RNA polymerases of positive-strand RNA viruses and related double-stranded RNA viruses. The presented dendrograms are tentative phylogenetic schemes derived by consensus of the following algorithms for phylogenetic tree generation implemented in the PHYLIP package (Felsenstein, 1989): UPGMA (unweighted pair group minimum average) clustering (Sneath and Sokal, 1973); neighbor-joining method (Saitou and Nei, 1987), two versions of the least square distance matrix method (Fitch and Margoliash, 1967) implemented in programs FITCH and KITSCH, and protein parsimony method (Felsenstein, 1989). The distance matrices used as input by all these methods except protein parsimony were generated using the CLUSTAL-V program (Higgins and Sharp, 1988). The root position was inferred from the results obtained with UPGMA. The branch lengths are approximate. (a) The principal branching order of the RdRp tree. Only the large divisions designated by the names of the prototype virus groups are shown. (b) Supergroup 1; (c) Supergroup 2; (d) Supergroup 3.

groups revealed at the qualitative level during construction of the multiple alignments. We designate these lineages after prototype viruses. Supergroup 1 (Figures 2A,B) consisted of the picorna-like lineage (picornaviruses, comoviruses, nepoviruses, RTSV and PYFV, and caliciviruses), the poty-like lineage (potyviruses, bymoviruses, and the dsRNA hypovirulence-associated virus from *Cryphonectria parasitica*), the lineage in-

cluding viruses with small genomes (nodaviruses, sobemoviruses, luteoviruses), the dsRNA virus lineage (ScV, LRV1), and the arteri-like lineage (coronaviruses, toroviruses, and arteriviruses). In addition, phylogenetic analysis of the putative RdRp of human astrovirus whose complete genome sequence has been determined very recently indicated that it is a peripheral member of supergroup 1 that probably has branched from the tree
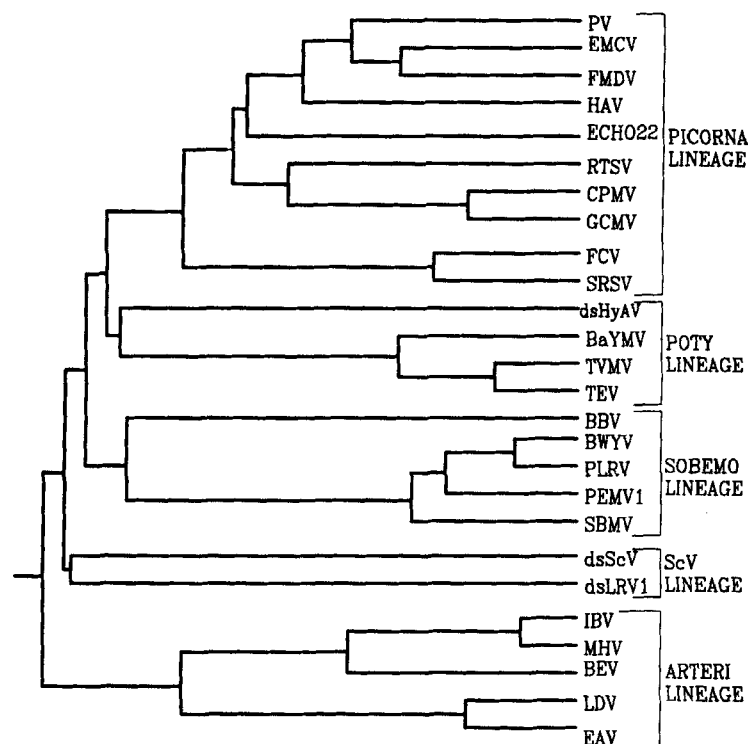
FIGURE 2B
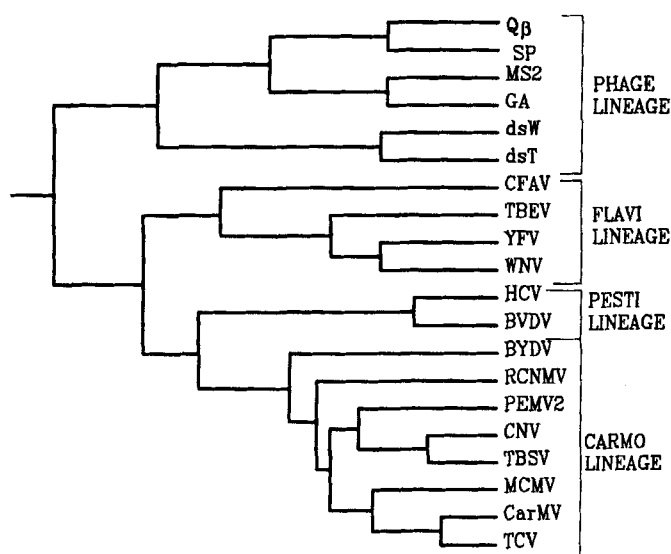
FIGURE 2C

at about the same point as the arteri-like viruses (Jiang et al., 1993). The specific position of the dsRNA viral lineage within the supergroup and the inclusion of the nodavirus polymerase in one group with the sobemoviruses and luteoviruses remained tentative.

Supergroup 2 (Figures 2A,C) included the phage lineage comprised by the RNA coliphages
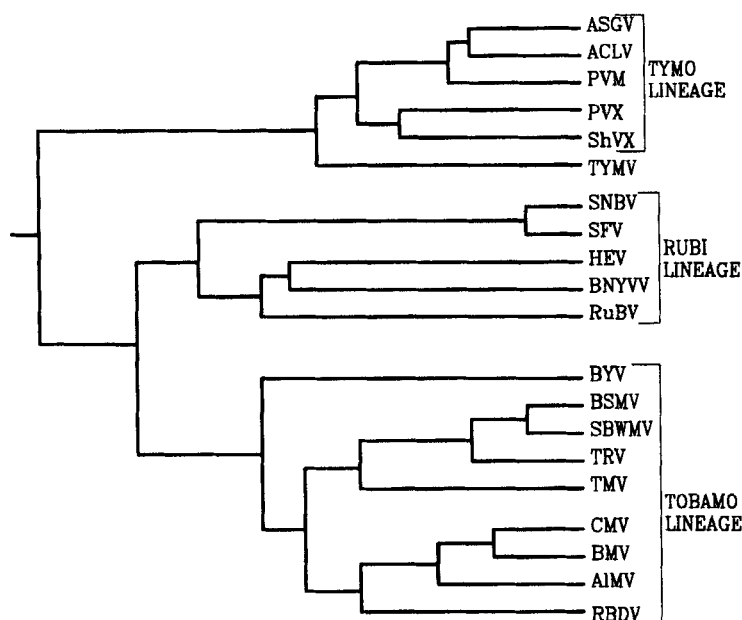
FIGURE 2D

and the related yeast dsRNA elements (Rodriguez-Cousino et al., 1991; Esteban et al., 1992; Koonin, 1992a), the flavivirus lineage, the pestivirus lineage (pestiviruses and HCV), and the lineage compiling plant viruses with small genomes (BYDV-PAM luteovirus, dianthoviruses, carmoviruses, tombusviruses, necroviruses).

Finally, supergroup 3 (Figures 2A,D) was composed of the tymo-like lineage (tymoviruses, carlaviruses, potexviruses, capilloviruses), the rubi-like lineage (rubella virus, HEV, BNYVV, and possibly alphaviruses), and the tobamo-like lineage (tobamoviruses, tricornaviruses, RBDV, hordeiviruses, tobraviruses, closteroviruses). The position of the alphavirus RdRps in the tree was uncertain, with some algorithms including them in the rubi-like lineage, and others in the tobamo-like lineage. As discussed below, the former grouping appeared to be more compatible with the results of comparison of genome organization.

An important corollary of the phylogenetic analysis of the positive-strand RNA virus RdRps is that each of the three supergroups and even the smaller groups comprising them compile viruses with very different genome size, organization,

and expression strategy. Moreover, it seems clear that at least on three occasions specific groups of positive-strand RNA viruses have given rise to dsRNA-containing viruses or virus-like agents (Figure 2; Koonin, 1992a). On the other hand, the opposite trend, that is, association of certain conserved gene arrays and modes of expression with distinct clusters of phylogenetic trees, was no less obvious (see below).

C. RNA Helicases

1. Amino Acid Sequence Comparisons

All positive-strand RNA viruses with genome size over 6 kb encode a (putative) RNA helicase thought to be involved in duplex unwinding during viral RNA replication, and perhaps also translation (Gorbalenya and Koonin, 1989). Very recent work has revealed an interesting exception to this "rule" by demonstrating that human astrovirus whose genome is about 7.2 kb in length does not encode a helicase (Jiang et al., 1993). These results once more emphasize the relative, "opportunistic" nature of even the most prominent regularities in the organization of viral genomes.

Positive-strand RNA virus helicases belong to three distinct superfamilies, each of them also including cellular and DNA viral helicases (Gorbalenya et al., 1988a,b, 1989b, 1990; Hodgman, 1988; Lain et al., 1989; Gorbalenya and Koonin, 1989, 1993b). Like the polymerases, the helicases revealed the "patchy" sequence conservation, with seven conserved motifs in superfamilies 1 and 2 (Figure 3), and three conserved motifs in superfamily III (Figure 4). Superfamilies I and II appeared to be distantly related to each other (Figure 3). Several viruses encode two putative helicases of superfamily I, and these proteins showed the highest degree of divergence in the whole superfamily (Gorbalenya et al., 1988b; Figure 3).

Actual helicase activity has been demonstrated only for CI protein of plum pox potyvirus (Lain et al., 1990, 1991). RNA-dependent ATPase activity has been found in flavivirus NS3 protein (Wengler and Wengler, 1991; Warrener et al., 1993), and in poliovirus 2C protein (Mirzayan and Wimmer, 1992a), and the essential importance of the principal NTP-binding motif (Figure 3) for virus reproduction has been confirmed by site-directed mutagenesis of the latter protein (Mirzayan and Wimmer, 1992b; Teterina et al., 1992). This limited experimental support notwithstanding, the degree of sequence similarity left virtually no doubt that all the proteins belonging to the three superfamilies possess the helicase activity.

It remains an unresolved and tantalizing question whether the positive-strand RNA viruses with smaller genomes do not require helicase action at all, or they recruit cellular helicases. One hypothesis linking the presence of the helicase gene with the genome size is that the helicase might ensure the fidelity of complementary nucleotide incorporation that is necessary to replicate relatively large genomes (Koonin, 1991b).

2. Phylogenetic Analysis

The three superfamilies, to which the (putative) helicases of positive-strand RNA viruses belong, share only two conserved motifs making generation of an overall phylogenetic tree in the

normal sense impractical. Nevertheless, first, all these helicases belong to the single class of NTPases, and, second, superfamilies 1 and 2 are related to each other and distinct from superfamily 3 (Gorbalenya and Koonin, 1989, 1993b; Gorbalenya et al., 1989b, 1990), it is illustrative to present the relationship between them in a tree-like scheme (Figure 5a). Actual trees were generated separately for each of the superfamilies.

Helicases of superfamily 1 are found in viruses with polymerases of supergroup 3, with the only exception of the arteri-like viruses that have a helicase of superfamily 1 combined with an RdRp of supergroup 1. This is the largest group of viral helicases and it caused most of a problem in generating an unambiguous tree topology. The complexity is amended by the existence of two helicases in several groups of viruses. The tree shown in Figure 5b consisted of five distinct lineages: (1) arteri-like virus lineage; (2) lineage compiling the helicases of tobamo-like viruses and alphaviruses; (3) the rubi-like lineage; (4) the tymo-like lineage; and (5) lineage of "second", or "accessory" helicases that are typical of a subset of the tymo-like viruses, hordeiviruses, and BNYVV and are encoded outside of the block of genes mediating RNA replication (Gorbalenya et al., 1988b). However, the relationships between these lineages and the routing of the tree presented a problem. The clustering algorithm of tree generation has isolated the group of the "accessory" helicases as the outgroup, that is, the branch that separated from all the other branches at the root of the tree (not shown). We believe that this may be because of the rapid and anomalous evolution of this group related to a change in their function. This view is supported by the replacement of several otherwise invariant amino acid residues in the conserved helicase motifs in these proteins (Gorbalenya et al., 1988b; Figure 3). It has been shown that the "accessory" helicase-like protein of hordeiviruses, potexviruses, and BNYVV is involved in virus cell-to-cell movement in the infected plant (Petty et al., 1990; Beck et al., 1991; Gilmer et al., 1992). The arteri-like lineage branched next and we believe that the most likely position of the root is between this lin-

		I		IA		II		
CMV	708	ISQVDGVAGCGKTTAIKS	4	STDIIVIANKKSAQDV	31	RVLVDEVV	18	
BMV	680	ISMVDGVAGCGKTTAIKD	4	GEDLVITANRKSAEDV	32	RLLVDEAG	18	
A1MV	820	VTIVDGVAGCGKTTNIKQ	7	DVDLILTSNRSSADEL	30	RLIFDECF	18	
RBDV	853	IKLVDGVTGCGKTTEIVR	3	PGILILSVCKANVDEI	29	ELFIDEYG	18	
BSMV1	832	FELIDGVPGCGKSTMILN	4	RREVVVGEGRNATDDL	36	RFHFDEAL	18	
TMV	828	VVLVDGVPGCGKTKEILS	4	DEDLILVPGKQAAEMI	35	RLFIDEGL	18	
TMGMV	826	MVLVDGVPGCGKYKGDFE	4	DEDLILVPGKQAAAMI	32	RLFIDEGL	18	
TRV	899	FELVDGVPGCGKSTMIVN	4	CVDVVLSTGRAATDDL	34	VLHFDEAL	18	
SNBV	181	TIGVIGTPGSGKSAIIKS	3	ARDLVTSGKKENCREI	30	VLYVDEAF	19	
SFV	181	VVGVFGVPGSGKSAIIKS	3	KHDLVTSGKKENCQEI	30	ILYVDEAF	19	
RUBV	?	IRVWNMAAGAGKTTRILA	3	REDLYVCPTNALLHEI	29	RIYIDEAF	17	
HeEV	?	YQFTAGVPGSGKSRSITQ	1	DVDVVVVPTRELRNAW	20	RVVIDEAP	16	
BNYVV1	?	LEYVKGGPGTGKSFLIRS	4	IRDLVVAPSIKLRSDY	27	IIFVDEFT	18	
TYMV	?	VVHFAGFAGCGKTYPIQQ	7	KDFRVSCPTTELRTEW	26	ILVIDEIY	19	
KYMV	?	MFHLAGFAGCGKTKPLQS	7	HSFRVSTPTTELRNEW	26	ILVIDEIY	19	1
ACLV	?	IYGIFGFAGSGKSHAIQN	9	QGIMVICPRRFLAKDW	24	LFILDEIS	24	
ASGV	?	VGLRLGFAGSGKTHKVLQ	7	VKRMFISPRRMLADEV	26	EVFVDEIG	40	
PVM1	?	LHAIVGTFGSGKSTLFKN	7	KSLDFVSPRRALAEDF	39	VVILDEHQ	19	
WC1MV1	565	MSVIHGAGGSGKSHAIQT	9	RHVTIILPTTDLRNDW	25	IIVFDDYS	19	
PVX1	699	ACVIHGAGGSGKSHAIQK	8	SDITVVLPTNELRLDW	25	IVIFDDYS	19	
BSMV2	264	TGIISGVPGSGKSTIVRT	5	FPAVCALANPALMNDY	22	LLIIDEYT	18	
BNYVV2	128	VGIVLGAPGVGKSTSIKN	8	HKMVLCLPFSQLLEGV	24	TMLVDEVT	18	
PVM2	24	PIVVHCVPGAGKSSLIRE	8	CAYTAGVEDQPRLSGN	13	FVVLDEYT	8	
WC1MV2	23	PIVVHAIAGSGKSTVIRK	9	KAYTLGKPDPYSLSNP	12	LDILDEYG	11	
PVX2	24	PLVVHAVAGAGKSTALRK	8	TNHTLGVPDKVSIRTR	12	FAILDEYT	9	
IBV	?	RTTVQGPPGSGKSHFAIG	7	ARVVFTACSHAAVDAL	52	ILLVDEVS	18	
MHV	?	YCTVQGPPGTGKSHLAIG	7	ARVVYTAASHAAVDAL	52	IIVVDEVS	18	
BEV	?	VTFVMGPPGTGKTTFVYD	9	NRFVYCAPTHRLVGDW	45	VLIADEVS	17	
EAV	?	SEYVEGPPGSGKTFHLVK	6	GSATLVVPTHASMLDC	43	ETFVDEVA	15	

```
             A A    S                                  DA
cons1    ...&.G.PG.GKT..&..      ....&..........&      .&&&DE&.
           :# : :#:### ::        : ::: ::: : :#       :#####:
cons2    ..&&....GSGKT...&P      .R.&UU.PTR.U..E&      &&&&DE.H
             A S                  K A   SK A  N
```

YFV	190	TTVLDFHPGAGKTRRFLP	10	LRTLVLAPTRVVLSEM	48	VIIMDEAH	23	
WNV	186	ITVLDLHPGAGKTRKILP	10	LRTAVLAPTRVVAAEM	48	LFIMDEAH	23	
DEN2	185	LTIMDLHPGAGKTKRYLP	10	LRTLILAPTRVVAAEM	48	LIIMDEAH	23	
JEV	186	MTVLDLHPCSGKTRKILP	10	LRTAVLAPTRVVAAEM	48	LFVMDEAH	23	
TBEV	191	ITVLDMHPSGKTHRVLP	10	LRTLVLAPTRVVLKEM	48	VAIMDEAH	23	
CFAV	183	RRFVTWHPCKGKTRKVIV	10	QRTVILTPTRVVMAEV	46	MIIMDECH	22	
BVDV	?	FKQITLATGAGKTTE-LP	10	KRVLVLIPLRAAAESV	48	YIFLDEYH	23	2
HCV	?	VAHLHAPTGSGKSTK-VP	7	YKVLVLNPSVAATLGF	47	IIICDECH	24	
TEV	76	DFLVRGAVGSGKSTG-LP	6	GRVLMLEPTRPLTDWM	51	FVIIDECH	21	
TVMV	77	DIILMGAVGSGKSTG-LP	6	GGVLLLEPTRPLAENV	51	FIIFDEFH	21	
PSbMV	77	EYLIRGAVGSGKSTG-LP	6	GRVLLLEPTRPLTENV	51	FVIFDECH	21	
PPV	77	DILIRGAVGSGKSTG-LP	6	GHVLLIEPTRPLAENV	51	CIIFDECH	21	
BaYMV	85	WSMVVGHTGSGKSTY-LP	14	QQILICEPTQAATENV	50	AIFLDEAH	20	
dsHyAV	2657	HVTVAAKTASGKSTF-FP	12	KKLWIVMPRKILRDNW	43	LVFFDEFH	16	

A

FIGURE 3. Conserved sequence motifs in the distantly related positive-strand RNA virus helicases of superfamilies 1 and 2. The alignments are updated versions of the previously published ones (Gorbalenya et al., 1988b, 1989b). The tentative superposition of the seven conserved motifs is essentially as suggested in Gorbalenya et al., 1989b. The consensus patterns are shown separately for the two superfamilies. Asterisks indicate positions where the consensus residues are identical or similar, and colons indicate positions where the consensus residue(s) of one of the families is not present in the pattern for the second family but still is found in a significant proportion of its members. Motif I (A) and II (B) together comprise the purine NTP-binding pattern (Walker et al., 1982; Gorbalenya and Koonin, 1989). For viruses encoding two putative helicases, the one comprising a domain of a large protein and involved in genome replication is designated "1", and the "accessory" helicase encoded by a stand alone gene is designated "2". For other details and designations see caption to Figure 1.

	III		IV		V		VI	
	III		IV		V		VI	
CMV	ALCFGDSEQ	22	SDADTTFRS	81	DRIKTVHESQGISEDHVTLVR	10	KFSYCLVAVTRHKV	1?
BMV	VLAFGDTEQ	22	DVVHKTYRC	80	GHIKTVHEAQGISVDNVTLVR	10	HEEYCLVALTRHKK	1?
A1MV	VIGFGDTEQ	21	ERKLITWRS	68	DNIFTTHEAQGKTFDNVYFCR	16	GPCNGLVALSRHKK	2?
RBDV	VTLFGDSEQ	26	EIRSTTYRC	71	SEVRTVHAAQGLSYKNVVYFR	12	KLPYHLVAISRHTD	587
BSMV1	ILAQGDRAQ	21	NPKLASYRI	80	ESISTIHEAQGGTYENVILVR	14	SAPYIVVGTSRHTK	34
TMV	AYVYGDTQQ	24	ETRRTTLRC	64	SDVHTVHEVQGETYSDVSLVR	11	DSPHVLVALSRHTC	3C
TMGMV	AYIYGDTQQ	24	EMRRTTLRC	64	KNVNTVHEIQGETFEDVSLVR	11	SSPHVLVALTRHTK	3C
TRV	CICQGDQNQ	24	TEKRETYRS	66	AKVSTVHESQGETFKDVVLVR	10	GREYLIVALSRHTQ	35
SNBV	VVLCGDPMQ	24	FYKYISRRC	60	HEVMTAAASQGLTRKGVYAVR	11	TSEHVNVLLTRTED	378
SFV	VVLCGDPKQ	21	CHKSISRRC	60	HEVMTAAASQGLTRKGVYAVR	11	ASEHVNVLLTRTED	369
RUBV	VICVGDRDQ	19	ERSRHTWRF	52	IRAYTVREAQGMSVGTACIHV	11	TRDLAIVSLTRASD	?
HeEV	VHLLGDPNQ	22	SWWHVTHRW	46	PGSVTVHEAQGATYTETTIIA	10	SRAHAIVALTRHTE	?
BNYVV1	IYLVGDEQQ	24	HVPIMNFRN	63	VSKTTVRANQGSTYDNVVLPV	9	SAELNLVALSRHRN	?
TYMV	VIILGDPLQ	27	MYCWWSYRI	47	YRSCTISSSQGLTFCDPAIIV	8	SSANGLVALTRSRS	?
KYMV	VIILGDPLQ	26	HYCWWTYRV	51	FPATTISASQGVTHHNRVTIL	8	SPSNTLVALTRSTV	?
ACLV	IVCIGDPLQ	26	NYKWYSYRI	60	GNVMTFGESQGLTFNCGVIVL	7	SDAHIMVAITRFRR	?
ASGV	IRCFGDPLQ	27	KYLFQGYRF	62	VPVATVSESQGMTISKRVLIC	8	GANAAIVAITRSKV	?
PVM1	LFLVGDPAQ	28	NYKVRSHRF	61	AKVLTEGESTGLTFMHGTIYI	7	NERRWITALRRFRF	?
WC1MV1	AILTGDSKQ	27	YYLNITHRN	47	QKSMTYAGCQGLTTKAVQILL	7	SSNVIYTALSRAVD	499
PVX1	VILTGDSRQ	27	YYLNATHRN	48	NDTFTYAGCQGLTKPKVQIVL	7	SANVMYTALSRATD	502
BSMV2	VLLVGDVAQ	17	YRSETTYRL	54	YDCALAIDVQGKEFDSVTLFL	9	DKHLRLVALSRHKS	38
BNYVV2	VICFGDPAQ	17	AECYASRRF	58	IESILYSDAHGQTYDVVTIIL	10	DPNVRAVLLTRARK	30
PVM2	FALFGDPIQ	10	FVCSVSRRF	51	VEALSLQEITGQTFEVVTVRD	6	NRAAAYQCMTRHRR	6
WC1MV2	EFIFTDPYQ	10	TTLETTYRF	54	ASFFKVSDVIGYQWPTVTLYL	10	ERHLLFIGLTRHTE	14
PVX2	QALFADPYQ	9	FYLETSFRV	54	VEFVKPCQVTGLEFKVVTVVS	7	QSTAFYNAITRSK-	9
IBV	VVYVGDPAQ	30	IFLAKCYRC	86	LNVQTVDSSQGSEYDYVIFCV	8	NINRFNVALTRAKR	28
MHV	YVYIGDPAQ	30	IFLGTCYRC	83	LQTQTADSAQGSAYDFVIYSQ	8	NVNRFNVAITRAKK	31
BEV	VVLLGDPFQ	26	RYLTACYRC	75	GDVTTIDSSQGTTAANHLLVL	8	TVNRVIVGCSRSTT	33
EAV	VKGYGDLNQ	21	EPLRVCHRF	46	LGHRTIDSIQGCTFPVVTLRL	6	TRPRAVVAVTRASQ	28
	AACC		SA		S	AC	G S	
cons1	&&&&GD..Q		.......R.		....T&...QG.T&...V.&&.		.....&VAUTR...	
	:&&&		:: ::		: : ● ●: ● : ●:		: : : ●	
cons2	.&.UTATPP		U&&UPS...		&U&.TD&.E.GU.& ....UU		T.....QR.GRUGR	
	S		A AT		N A		S K	
YFV	TILMTATPP	39	AWFLPSIRA	36	FILATDIAEMGANL-CVERVL	27	SASSAAQRRGRIGR	156
WNV	AIFMTATPP	39	VWFVPSVKM	36	FVYTTDISEMGANF-KASRVI	28	TAASAAQRRGRIGR	156
DEN2	GIFMTATPP	39	VWFVPSIKT	36	FVVTTDISEMGANF-KAERVI	28	THSSAAQRRGRIGR	156
JEV	AIFMTATPP	39	VWFVASVKM	36	FVITTDISEMGANF-GASRVI	28	TSASAAQRRGRVGR	156
TBEV	LVLMTATPP	39	AWFVPSSAK	36	FVVNTDISEMGANL-DVSRVI	25	TTASAAQRRGRVGR	155
CFAV	LIYLSATPP	39	ILFVPSHNQ	32	LVISTDISEMGANL-GVDLVI	23	TTSSMIQRRGRTGR	147
BVDV	VVAMTATPA	49	LVFVPTRNM	36	VIVATNAIESGVTLPDLDTVI	30	TVGEQAQRRGRVGR	?
HCV	VVLATATPP	40	LIFCHSKKK	33	VVVATDALMTGYTG-DFDSVI	27	DAVSRTQRRGRTGR	?
TVMV	IIKVSATPP	44	LVYVASYNE	40	FIVATNIIENGVTL-DVDVVV	26	SLGERIQRFGRVGR	277
TEV	VLKVSATPP	44	LVYVASYND	40	FIVATNIIENGVTI-DIDVVV	26	SYGERIQKLGRVGR	276
PSbMV	ILKVSATPP	44	LVYVSSYNE	40	FIVATNIIENGVTL-DIDVVV	24	SYGERIQRLGRVGR	258
PPV	ILKVSATPP	44	LVYVASYNE	40	FVVATNIIENGVTL-DIDVVV	24	SYGERIQRLGRVGR	277
BaYMV	KFYVSATPR	45	LVFLAGRPE	42	IIFTTNIIETGVTL-SVDCVV	24	TRNERQQRIGRAGR	284
dsHyAV	TIFMSATPV	52	MIIVPTYNE	31	GLVCTPYVQTGIDIpAPSILI	20	DEKTNEQRVNRVGR	136

(brackets at right: group 1, group 2)

FIGURE 3B

eage and the remaining viral helicases as shown in Figure 5B. With the arteri-like viruses as the outgroup, the helicase superfamily 1 tree shown in Figure 5B mimicked the RdRp tree topology from Figure 2D, with two apparent exceptions. First, the helicases of alphaviruses belonged to the tobamovirus lineage, not to the rubivirus lineage as their RdRps (however, it has to be noted that while the topology of the helicase tree was stable, the position of the alphavirus RdRps in the tree strongly depended on the method used). Second, in the helicase tree the carlaviruses grouped with the tymoviruses, not with the potexviruses as in the RdRp tree. The topology of the helicase

tree with coronaviruses as the outgroup was compatible with the attractive hypothesis of the origin of the "accessory" helicase group by a gene duplication in the putative common ancestor of the tymo-like viruses.

All viruses with relatively large genomes that have RdRp of supergroup 2 encode a helicase of superfamily 2. In addition, superfamily 2 helicases have been found in the poty-like viruses where it is combined with supergroup 1 polymerase. The topology of the putative phylogenetic tree for superfamily 2 helicases was simple enough, with the most likely root position defined using a related cellular helicase as the outgroup (Figure 5C).

		A		B		C	
PV	120	IEPVCLLVHGSPGTGKSVAT	28	QQGVVIMDD	30	KGILFTSNYVLASTNS	105
BoEV	120	SEPVCALIHGSPGTGKSLAT	28	QQAVVVMDD	30	KGKIFTSKFVLASTNA	105
HRV2	115	CEPVAIVIHGPPGAGKSITT	26	QQSVVIMDD	30	KGKAFDSRFVLCSTNH	105
HRV14	120	IEPVCVLIHGTPGSGKSLTT	28	QQEVVIMDD	30	KGMLFTSNFVLASTNS	105
EMCV	110	CEPVVIVLRGDAGQGKSLSS	30	NQFAAIMDD	30	KGTPFTSQLVVATTNL	109
FMDV	101	PEPVVVCLRGKSGQGKSFLA	31	QQTVVVMDD	30	KGKPFNSKVIIATTNL	137
HAV	145	CEPVVCYLYGKRGGGKSLTS	32	GQLVCIIDD	30	KGRHFSSPFIIATSNW	93
ECHO22	130	IEPIGIWIQGEPGQGKSFLT	32	NQDIHLIDD	31	KGKFYTSKLVVATTNK	93
RTSV	?	IDPLHVCMLGAPGVGKSTIA	33	QEPVILYDD	36	KGHCTSKYVFSCTNV	?
PYFV	?	VDPFHVSLYGSPGVGKSFVM	33	GQTAVKCDD	34	KGRTFTSKYIFSTTNV	?
CPMV	159	KMPFTIFFQGKSRTGKSLLM	31	PQPFVLMDD	31	KGICFDSQFVFVSTNF	326
TBRV	206	CEPVWIYLFGQRHCGKSNFM	30	GQTFFHVDD	31	KPIYFRSPFIISSSNF	?
GCMV	?	KEPVWIYLWGPSHCGKSNFM	31	GQTIMEIDD	32	KPIYFKSQFVISSSNQ	?
RHDV	?	PQPVAVIFKGAPGIGKTYLV	26	GEEVAIADE	32	KNKVFNSKYLLCTTNS	?
FCV	?	QVPVCYILTGPPGCGKTTAA	26	GNEVCIIDE	31	KGKLFTSKYIIMTSNS	?
		D C T		E AC E		TS	
CONS		.EP&.&&&.G..G.GKS...		.Q.&&U&DD		KG..@.S.&U&.STN.	

FIGURE 4. Conserved sequence motifs in the positive-strand RNA virus helicases of superfamily 3. The alignment is updated from the one published previously (Gorbalenya et al., 1990). The designations of the conserved motifs are after Gorbalenya et al., 1990. For the other details and designations see caption to Figure 1.

A
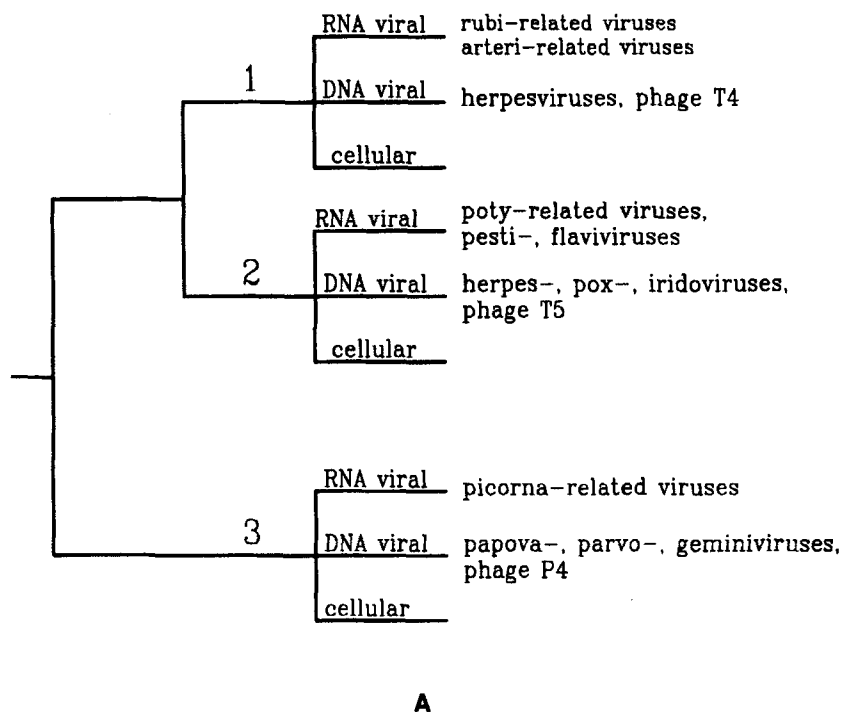
FIGURE 5. Phylogenetic analysis of positive-strand RNA virus RNA helicases. (a) The three helicase superfamilies. An arbitrary scheme is shown based on the qualitative separation of the superfamilies, each of which has a distinct set of conserved motifs (see text and Figures 2 and 3). (b) Superfamily 1. The methods used for tree generation are described in the caption to Figure 8. The root position was inferred from the cluster dendrogram constructed using UPGMA. (c) Superfamily 2. The root position was determined by using the sequence of putative yeast helicase PRP16 as an outgroup. (d) Superfamily 3. The root position was determined by using the sequence of the helicase domain of adeno-associated virus NS1 protein as an outgroup.
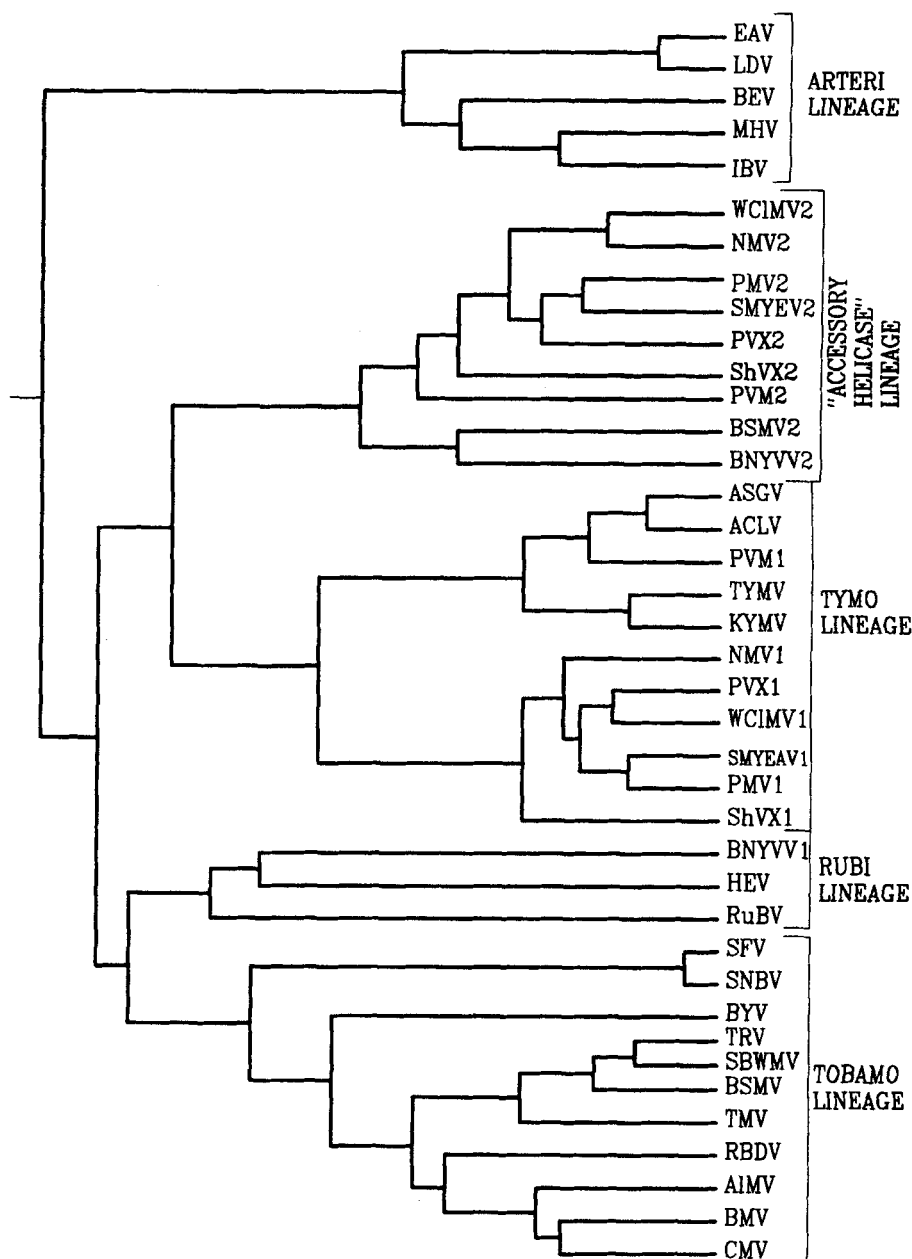
FIGURE 5B

The separation between the poty/bymovirus lineage, and the flavi/pesti/HCV lineage was not unexpected given the topology of the RdRp tree (Figure 2). The only issue of significant uncertainty included the putative helicase of the dsRNA fungal virus HyAV, which was placed with the poty- and bymoviruses by some but not all of the methods of tree generation (Koonin et al., 1991b; Figure 4C).

The helicases of superfamily 3 have been found exclusively in association with supergroup 1 polymerases. As with superfamily 2, phylogenetic analysis of the superfamily 3 helicases was straightforward, with the root position inferred by using a related DNA viral helicase as the outgroup (Figure 5D). The apparent grouping of these helicases conformed to the topology of the respective domain of

JEV
WNV
DEN4
YFV
TBEV
CFAV
HCV
BVDV

FLAVI
LINEAGE

dsHyAV
BaYMV
TVMV
PSBMV
TEV
PPV

POTY
LINEAGE

FIGURE 5C

PV
HRV14
BoEV
HRV2
EMCV
FMDV
HAV
ECHO22

PICORNA
LINEAGE

RTSV
CPMV
GCMV
TBRV

COMO
LINEAGE

SRSV
FCV
RHDV

CALICI
LINEAGE
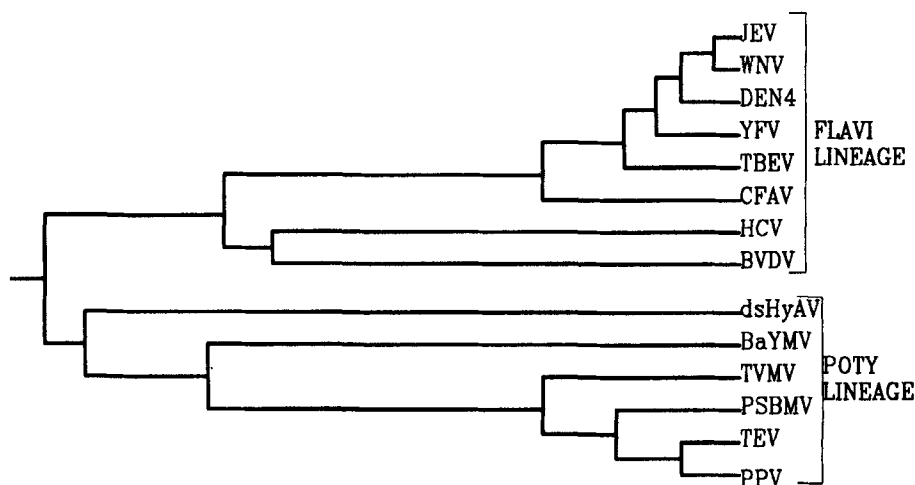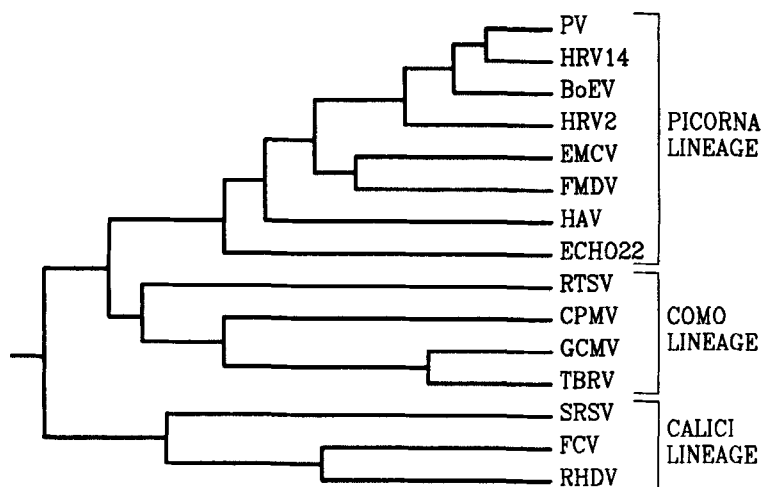
FIGURE 5D

the polymerase tree (compare Figures 5D and 2D).

D. Proteases

Numerous positive-strand RNA viruses express all or part of their proteins via processing of polyprotein precursors (Strauss, 1990). Generally, processing of virion envelope proteins is mediated by host proteases, while nonstructural proteins, and in many cases capsid proteins, are pro-

cessed by virus-encoded proteases. There are two main classes of positive-strand RNA virus proteases.

1. Chymotrypsin-Related Cysteine and Serine Proteases

The initial discovery of a virus-encoded chymotrypsin-related serine protease was really surprising as this activity has been identified with the capsid protein of alphaviruses that functions as an

394

autoprotease cleaving itself from the polyprotein precursor (Hahn et al., 1985). Subsequently, related serine proteases have been identified in flaviviruses, pestiviruses, and hepatitis C virus, first by computer (Gorbalenya et al., 1989c; Bazan and Fletterick, 1989a; Chambers et al., 1990a) and then experimentally (Preugschat et al., 1990; Chambers et al., 1990b; Wengler et al., 1991; Wiskerchen and Collett, 1991). Additionally, the N-terminal protein of potyviruses that has been shown to possess autoprotease activity (Verchot et al., 1991, 1992) also is specifically similar to the alphavirus capsid protease (Figure 6, and E. V. Koonin, unpublished observations).

Unexpectedly, in a parallel series of studies it has been shown that cysteine proteases of picornaviruses and several groups of related viruses also belong to the chymotrypsin-like class, the replacement of the principal catalytic residue in the majority of the proteases of this family notwithstanding (Blinov et al., 1984; Gorbalenya et al., 1986, 1989d; Bazan and Fletterick, 1988, 1989b, 1990). By now, the chymotrypsin-like enzymes comprise the largest class of viral proteases (Figure 6). Some of these proteases, for example, cysteine proteases of picornaviruses and potyviruses, occupy the central position in virus expression and mediate the processing of the majority of viral proteins. Other proteases, for example, the capsid protein of alphaviruses and the N-terminal protein of potyviruses, perform only one cleavage liberating their own C-terminus. The sequence conservation in the viral chymotrypsin-like proteases is largely confined to four conserved motifs, three of which center at the catalytic residues, whereas the fourth distal motif is implicated in substrate binding (Gorbalenya et al., 1989d; Figure 6).

The recent resolution of the three-dimensional structure of the alphavirus capsid protein clearly showed that its conformation is very similar to that of chymotrypsin-like proteases and distinct from that of capsid proteins of other icosahedral viruses (Choi et al., 1991).

It is important to emphasize that in the case of the capsid autoprotease of alphaviruses and the nonstructural proteases of flavi- and pestiviruses,

chymotrypsin-related enzymes with drastically different functions in virus reproduction showed a greater sequence similarity to each other than to any cellular proteases (Gorbalenya et al., 1989c). The same was true of the second large group of viral chymotrypsin-like proteases related to the picornavirus 3C protease, with some reserve regarding the serine proteases of arteriviruses (Gorbalenya et al., 1989d; Gorbalenya and Koonin, 1993a). However, it was very difficult to either prove or disprove the hypothesis that the two large subdivisions of viral chymotrypsin-related protease comprised a single monophyletic superfamily.

Only for the chymotrypsin-like proteases of alphaviruses, flaviviruses, and pestiviruses long enough alignments could be generated to make phylogenetic analysis over long evolutionary distances meaningful. The resulting tree, which has been rooted using a related cellular protease as an outgroup, revealed the expected separation of the alphavirus capsid autoproteases from the nonstructural proteases of flaviviruses and pestiviruses but failed to show grouping between the pestiviruses and HCV suggested by the polymerase and the helicase trees (Figure 7).

2. Papain-Like Cysteine Proteases

Unlike the unique chymotrypsin-like cysteine proteases, the nature and orientation of the catalytic Cys and His residues in these viral proteases resemble classic cellular papain-like thiol proteases although the actual sequence similarity is quite low (Gorbalenya et al., 1991). In fact, conservation among viral papain-like proteases is observed almost exclusively around the catalytic Cys residue; even identification of the catalytic His in some of these proteases has been a difficult exercise, and no other conserved motifs could be detected (Gorbalenya et al., 1991; Figure 8). Recent studies have shown, perhaps somewhat unexpectedly, that papain-like proteases are almost as widespread among positive-strand RNA viruses as chymotrypsin-like proteases (Oh and Carrington, 1989; Hardy and Strauss, 1989; Gorbalenya et al., 1989a, 1991; Lee et al., 1991; Koonin et al., 1991b, 1992). The leader function

```
                    •         #           !!!  #               ! !  !
PV      3C   32 NVAILPTHA 28 LEITIITLKR 61 PTRAGQCGG-VITCT-G---KVIGMHVGG  19
CVB4    3C   32 RWAVLPRHA 28 LELTLLKLNR 61 PTRAGQCGG-VLMST-G---KVLGIHVGG  19
BoEV    3C   32 TVVVLPRHA 28 LELTIVKLKM 61 PTKAGQCGG-VVISM-G---KIVGVHVGG  19
HRV89   3C   32 QVMVLPTHS 28 LEITVVKLIR 61 PTKAGYCGG-VLYKV-G---SILGIHVGG  19
HRV14   3C   32 RVCVIPTHA 28 LELTVLTLDR 60 ATKTGQCGG-VLCAT-G---KIFGIHVGG  19
EMCV    3C   40 RTLVVNRHM 31 TDVSFIRLSS 64 NTRKGWCGSALLADL-GGSKKILGIHSAG  25
FMDV    3C   40 TAYLVPRHL 35 SDAALMVLHR 64 ATRAGYCGGAVLAKD-GADTFIVGTHSAG  29
HAV     3C   40 DWLLVPSHA 37 QDVVLMKVPT 69 AWRPGMCGGALVSSNQSIQNAILGIHVAG  23
ECHO22  3C   39 DEIILHGHS 35 MDLAILKCKL 62 KSCKGMCGGLLISKVEG-NFKILGMHIAG  20


PV      2A   12 GYKICNYHL 15 RDLLVTESRA 46 FASPGDCGGILRC-HHG----VIGIITAG  23
CVB4    2A   13 NYKVVNRHL 15 RDLLVSTTTA 46 FSEPGDCGGILRC-EHG----VIGLVTMG  23
BoEV    2A   13 SYKILNRHL 15 RDLLVTRVDA 46 FAEPGDCGGLLRC-EHG----VMGILTVG  23
HRV89   2A   10 NLIYRNLHL 14 SDLIIYRTNT 46 PCEPGDCGGKLLC-KHG----VIGMITAG  19
HRV14   2A   12 NVKIMNYHL 15 RDLAIVSTGG 46 PAEPGDCGGILRC-IHG----PIGLLTAG  20


FCV     3Cl   ? HGVASVAHV 18 GEFCCFRSTK 47 ETHPGDCGLPYI-DDNG---RVTGLHTGS   ?
SRSV    3Cl   ? TVFITTTHV 21 GEFTQFRFSK 70 GTIPGDCGAPYV-HKRGNDWVVCGVHAAA   ?
RHDV    3Cl   ? NGLISNTHT 14 TDLCLVKGES 45 QTTHGDCGLPLY-DSSG---KIVAIHTGK   ?
BaYMV   NIa   ? DWILVPGHL 34 -DVIAIRRPA 51 STVLGMCGCQFWTLER----QIDGIHVAT   ?
TEV     NIa 226 PFIITNKHL 34 -DMIIIRMPK 53 QTKDGQCGSPLVSTRDG---FIVGIHSAS  72
TVMV    NIa 221 PYIIANQHL 34 -DIIVIKMAK 53 TTKDGQCGSPLVSIIDG---NILGIHSLT  71
RTSV    3Cl   ? TYVCMPAHY 34 QETVVWDLGP 74 HCMPGFCGRAIMRADATCFRKIIGMHVSG   ?
PYFV    3Cl   ? GFLLAPLHT 32 YDACIIRTDA 66 KTEDGQCGSCLVSTSDKLDGKVFCSLVAG   ?
RCMV    p24  32 RRFIGYSH- 33 SELCVYHSSC 70 PTVMSDCGSMIITNVGG-KTKIVGIHVAG  21
CPMV    p24  32 RRFLACKH- 33 SELVLYSHPS 70 PTIPEDCGSLVIAHIGG-KHKIVGVHVAG  21
TBRV    p24  30 KSVRMTRHQ 35 SEIVTWLAPS 72 ESRNDDCGMIILCQIKG-KMRVVGMLVAG  19
GCMV    p24   ? KCVMMTRHQ 35 SEVVCWLAPS 77 ESRNNDCGMLLTCQLSG-KMKVVGMLVAG   ?
IBV     3Cl  33 DTIYCPRHV 18 HEFEVTTQHG 74 SFLAGACGSVGFNIEKG-VVNFFYMHHLE 143
MHV     3Cl  33 DKVYCPRHV 21 SDFCVMSDRM 65 SFLCGSCGSVGYVLTGD-SVRFVYMHQLE 137


SBMV    3Cl  41 DVLMVPHHV 31 IDFVLVKVPT 53 PTAKGWSGTPLY-TRDG----IVGMHTGY   ?
PLRV    3Cl  39 NALVTAEH- 29 NDISILVGPP 53 NTGPGYSGTGFW-SSKN----LLGVLKGF   ?
BWYV    3Cl  30 NALMTATHV 33 GDVTLLRGPP 55 HTEGGHSGSPYF-NGKT----ILGVHSCA   ?
PEMV    3Cl  52 TGIVLPIHV 27 HDSLIMTSAM 54 DTRPGDSGLPLF-DMKM----NVVAHRGT   ?
EAV     3Cl  31 VVVLTASHV 23 GDFAEAVTTQ 40 WTTSGDSGSAVV-QGDA----VVGVHTGS  67
LV      3Cl  31 RTVVTAAHL 22 GDYAWSHADD 38 FTNCGDSGSPVI-SESG---DLIGIHTGS  68


HCV     NS31  ? GGISSVDHV 25 TDESEYGVKT 51 KNLKGWSGLPIFEASSG---RVVGRVKVG   ?
BVDV    NS31  ? GGISSVDHV 25 TDETEYGVKT 53 KNLKGWSGLPIFEASSG---RVVGRVKVG   ?
HeCV    NS31  ? GVCWTVYHG 21 QDLVGWPAPQ 46 SYLKGSSGGPLLC-PAG---HAVGIFRAA   ?
TBEV    NS3  46 GVLHTMWHV 21 EDVVCYGGAW 48 DLAKGTSGSPILN-SQG---VVVGLYGNG 464
YFV     NS3  45 GVFHTMWHV 21 EDLVAYGGSW 49 DYPSGTSGSPIVN-RNG---EVIGLYGNG 466
WNV     NS3  43 GVFHTLWHT 21 EDRLCYGGPW 48 DYPTGTSGSPIVD-KNG---DVIGLYGNG 465
JEV     NS3  43 GTFHTMWHV 21 KDLISYGGGW 48 DFSPGTSGSPIVD-KKG---KVVGLYGNG 464
CFAV    NS3  47 GVFHTLMHV 21 RDVVSYGGPW 44 DFGKGSSGSPFFINGEPVGFYGCGFYVNG 417


VEEV    CP  144 GKLFRPMHV 20 YD-LEYADVP 39 VGAKGDSGRPILD-NQG---RVVAIVLGG  31
SNBV    CP  133 GKVMKPLHV 20 YD-MEFAQLP 39 VGGRGDSGRPIMD-NSG---RVVAIVLGG  31
SFV     CP  137 DKVMKPAHV 20 YD-LECAQIP 39 AGKPGDSGRPIFD-NKG---RVVAIVLGG  30
WEEV    CP  126 GRLMKPLHV 20 YD-LEYGDVP 39 VGGKGDSGRPILD-NRG---RVVAIVLGG  31


TEV     NtP 206 QLFASVRHM   6 VD-LRIDNWQ 18 KLTFGSSGLVLRQ-GSY---GPAHWYRHG  30
PPV     NtP 208 YFRTHVRHL   6 YD-LVLDEAT 20 EVTPGMSGFVVNP-INL---SDPMQVYDT  31
TVMV    NtP 157 ALFIDVAHA   6 ID-CRMHRRE 18 HLRKGDSGIVLLT-QKI---KGHLSGVRD  31
PVY     NtP 184 SAAVRTAHM   6 VD-FRCDMWT 20 NIRRGDSGVILNT-KSL---KGHFGRSSG  31
PSBMV   NtP 299 LLQVETKHH   6 KD-ASLNNLT 18 SITHGHSGVVFLR-ANI---SGSKSYSID  31
                A           E  C              ....G.SG..&&. ...  A H
consensus       ..&....H.   .D...&....                             .&&G&&...
```

FIGURE 6. Conserved sequence motifs in positive-strand RNA virus chymotrypsin-related cysteine and serine proteases. The alignment was constructed by combining the conserved motifs from updated versions of the previously published alignments for distinct families of viral proteases (Gorbalenya et al., 1989b,c; Verchot et al., 1991). The (putative) catalytic residues are shown by asterisks, and residues thought to participate in substrate binding are shown by exclamation marks. 3Cl indicates a "3C-like protease", NS31 indicates an "NS3-like protein", and NtP indicates an "N-terminal protein". The other designations are as in Figure 1.

with the single C-terminal cleavage is very typical of this class of proteases (e.g., aphthoviruses, potyviruses, coronaviruses, arteriviruses). In alphaviruses (Hardy and Strauss, 1989), and probably also in RubV and HEV (Gorbalenya et al., 1991; Koonin et al., 1992), a papain-like protease

```
                                                    ┌─ WNV    FLAVI-
                                              ┌─────┤           PESTI-
                                        ┌─────┤     └─ YFV     VIRUS
                                   ┌────┤     └──────── TBEV   NON-
                              ┌────┤    └─────────────── CFAV  STRUCT.
                         ┌────┤    └────────────────────── HCV PROTE-
                         │    └───────────────────────────── BVDV  ASES

                         │                           ┌─ EEEV  ALPHA-
                         │                     ┌─────┤          VIRUS
                         └─────┐         ┌────┤     └─ WEEV   CAPSID
                               └────────┤    └────────── VEEV  PROTE-
                                        └───┐      ┌──── SFV   ASES
                                            └──────┤
                                                   └──── SNBV
```
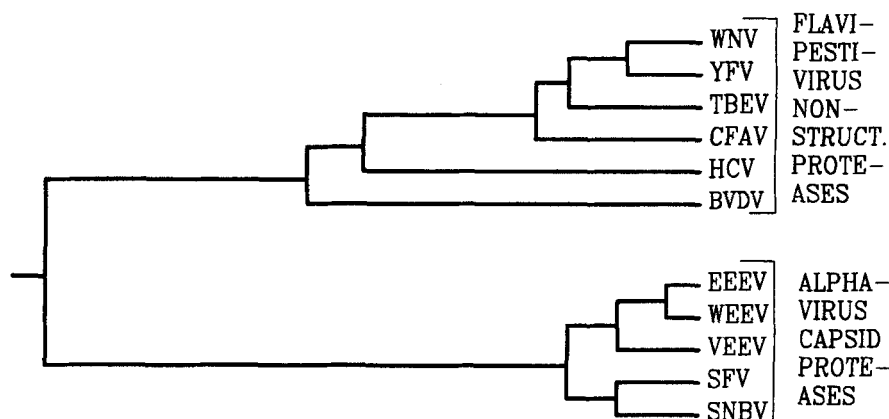
FIGURE 7. Phylogenetic analysis of the family of positive-strand RNA virus chymotrypsin-related proteases including the capsid autoprotease of alphaviruses and the nonstructural protease of flaviviruses and pestiviruses. The root position was inferred by using the sequence of *Staphylococcus aureus* protease V8 as an outgroup.

```
                              *                           *
BNYVV            ?  CADFYNLVSRPNNCLVVAISECLGVT  77 SDGHFIAAPLSS    ?
RUV             ?  RASTRGGELDPNTCWLRAAANVAQAA 105 PTGHFVCAVGGG    ?
HEV             ?  FCCFMKWLGQECTCFLQPAEGAVGDQ  91 PERHNLSFDASQ    ?

IBV             ?  RDNFLILEWRDGNCWISSAIVLLQAA 147 NSGHCYTQAAGQ    ?
MHV1    leader  ?  CGNYFAFKQSNNNCYINVACLMLQHL 141 SVAH-YTHVKCK    ?
MHV2            ?  CG-FYSPAIERTNCWLRSTLIVMQSL 135 NDCHSMAVVDGK    ?
EAV     leader 150 SLIVTTDQEQDGFCWLKLLPPDRREA  50 RAWHITTRSCKL   22
LV1     leader  62 IGIPQVECTPSGCCWLSAVFPLARMT  65 GATHVLTNSPLP    ?
LV2             ?  V---QNPDVFDGKCWLSCFLGQSVEV  53 WIRHLTLDDDVT   32

FMDV    leader  38 KT-FYSRPNNHDNCWLNTILQLFRYV  42 NIKHLLQTGIGT?  84
                                                71 ADFHAGIFMKGQ?  56
                                                81 GQEHAVFACVTS?  46
SNBV nsP2      467 TPRANPFSCKTNVCWAKALEPILATA  63 PVAHWDNSPGTR  240
SFV  nsP2      465 AAPVDAFQNKANVCWAKSLVPVLDTA  56 --NHWDNRPGGR  242
RRV  nsP2      465 STAVDPFQNKAKVCWAKCLVQVLETA  55 --NHWDNRPGGR  243
VEEV nsP2      464 PDPTDVFQNKANVCWAKALVPVLKTA  56 --NHWDNSPSPN  240

TEV  HC        331 LNEEKMYIANEGYCYMNIFFALLVNV  57 KTMHVLDSYGSR   32
PPV  HC        330 AKGGAMFIAKAGYCYINIFLAMLINI  57 KIFHVVDEFGSL   32
TVMV HC        330 EISNLMYIAKEGYCYINIFLAMLVNV  57 KTIHVVDSYGSL   33
PVY  HC        329 GDSEMLYIAKQGYCYINVFLAMLINI  57 QTCHVVDSFGSQ   33
PSBMV HC       331 TETGRMWIAKEGYCYINIXFAMLVNV  57 QIFHVIDSYGSM   33
BaYMV HC1      129 VQTFIAFDFAHGYCYLSLFIPLSFRI  56 LQFHVSDARG-L   33

HyAV p29 leader 148 SRNGSLAQFGQGYCYLS---AIVDSA  43 -VYHVV-----V   30
HyAV p48 leader 327 EIDTLRVPVEEGRCF-----ELLFNN  36 QCVHIVA--GET   24
                                 ▲                       ▲
consensus          ............C●U.....U&...      ...H&.......
```

FIGURE 8. Conserved sequence motifs in positive-strand RNA virus papain-like cysteine proteases. The alignment was an updated superposition of the conserved motifs from previously published alignments of this type of proteases (Gorbalenya et al., 1991; Koonin et al., 1991, 1992). The tentative identification of a papain-like protease of BNYVV has not been described previously. The leader proteases, that is, proteases that comprise the N-terminal domain of a polyprotein and cleave a single site at their own C-terminus, are marked. The catalytic histidine in the leader protease of FMDV could not be identified unambiguously and three candidate segments are shown (Gorbalenya et al., 1991). MHV and LV each encode two putative papain-like proteases, while the related viruses IBV and EAV possess only one domain of this type (Lee et al., 1991; Godeny et al., 1993). The other designations are as in Figure 1.

appears to be responsible for the entire processing of the nonstructural precursor polyprotein. Moreover, it is known that tymoviruses express their nonstructural proteins via polyprotein processing (Morch et al., 1990) and the proteolytic activity has been mapped to a distinct domain (Bransom et al., 1991). Recent experiments have shown that this protease is probably of the papain type (Rozanov et al., 1992a; Bransom and Dreher, 1993; T. W. Dreher, personal communication). Amino acid sequence comparisons indicated that related proteases may be encoded by carlaviruses, capilloviruses, and BNYVV (Figure 8; E. V. Koonin, unpublished observations), suggesting that all these viruses have an expression strategy similar to that of tymoviruses.

E. Methyltransferases

Reproduction of positive-strand RNA viruses that have capped plus-strand RNA re-

quires either viral or cellular capping enzymes. Association of the N-terminal domain of the nonstructural polyprotein with methyltransferase activity has been clearly demonstrated for alphaviruses by genetic and biochemical experiments (Mi et al., 1989; Mi and Stollar, 1991). The same type of methyltransferase domain has been tentatively identified in a number of related virus groups by amino acid sequence comparison; these domains had unique conserved motifs unrelated to those found in cellular methyltransferases (Rozanov et al., 1989b; Figure 9).

This type of methyltransferase domain is strictly associated with supergroup 3 polymerases and superfamily 1 helicases. The tentative phylogenetic tree for the methyltransferases separated into two clear-cut lineages, the tymo-like viruses and the tobamolike viruses together with alphaviruses (Figure 10), in accord with the results of alignment analysis (Rozanov et al., 1989b). The methyltransferase tree failed to reveal the rubi-like

		I		II		III	
SNBV	33	NDHANARAFSHLASK	35	RSPEDPDR	149	GSTL-YPEHRASLQSWH	270
SFV	37	NDHANARAFSHLATK	35	RSAEDPER	151	GSTL-YTESRKLLRSWH	277
TRV	92	MVHGFAAAERKLQAL	36	LDIRDDQR	132	DPSFSYIHDWEEYKKYL	947
BSMV	87	GTHSMAACFRFLETE	38	LSMRDSER	140	DPNAGYSHDLKDYLKYV	834
TMV	78	AVHSLAGGLRSLELE	36	LDVRDIMR	127	ESTLNYCHSYSNILKYV	835
AMV	96	SSHCFAAAHRLLETD	37	LDARDGAR	105	EPNLGYSHRFSLLKHYL	848
BMV	77	APHSLAGALRVAEHY	36	LGVRDAAR	107	ESTLSYLHGWQDLGSFF	701
CMV	78	APHGLAGALRLCETL	39	LGIRDKMR	133	ESTMSYVHDWDNIKSFM	688
RBDV	200	GPHNMAAAHRLLETH	36	LDLRDNER	168	DSTMGYRHDWEVYSKYL	950
HEV	62	WNHPIQRVIHNELEL	32	PVGRDVQR	106	DTSAGYNHDVSNLRSWI	?
BNYVV	215	RNHPVLAALREVMRQ	40	KDSKDLVR	126	CEDVGYNFSVD---AWL	?
RUBV	63	SDHPALHAISRYTRR		170		HPGRLYRCGPR---LWT	?
		G		K		A	
cons1		..H..A.A&..&...		...RD..R		...&.Y...&.....@&	
		* *		:* *		* * * **	
cons2		H.H...K.UE..UL.		U...D..R		.....Y.Q.&.....@U	
						H	
KYMV	64	HPHPAHKTLETHLLF	38	LTPTDSVR	68	HSAGSYNQPITAL-SWL	?
TYMV	64	HPHAAHKTIETFLLC	38	LHPNDSTR	68	HEAGSYNQPSDAH-SWL	?
PVM	68	HSHPVCKTLENYILY	39	VTSADRMR	97	VQSEGYQQPLKG--GYL	?
ACLV	69	HSHPGCKTLENHLLF	43	VTAKDKAR	72	VRSESYTQPLEN--GFL	?
ASGV	75	HSHPISKMIENHLLY	47	IDGKDKYR	95	VASECYEQNLANS-KWP	?
WC1MV	65	HSHGAVKSIENTLLE	38	IEPRDLQR	72	HGGGSYSHEFKQL-EWL	1070
NMV	65	HTHAACKVIENRMLE	38	VEPKDLFR	72	HGGGCYFHPYTTL-EWL	1419
PVX	65	HTHAAAKTIENKLLE	38	IEPRDVAR	72	HGGGAYHHEFSHL-QWL	1232

FIGURE 9. Conserved sequence motifs in type I putative methyltransferase domain of positive-strand RNA viruses. The motifs were taken from an updated version of the previously published alignment (Rozanov et al., 1992). The consensus was derived separately for the rubi-like group and the tymo-like group, and the correspondence between the two patterns is shown as detailed in the caption to Figure 3. In the RubV sequence motif II could not be identified (Rozanov et al., 1992). The other designations are as in Figure 1.

PVM
ACLV
ASGV
TYMV
KYMV
NMV
PMV
PVX
FMV
ShVX

TYMO
LINEAGE

SNBV
SFV
BNYVV
HEV
BYV
TRV
SBWMV
BSMV
TMV
RBDV
AlMV
CMV
BMV

TOBAMO
LINEAGE
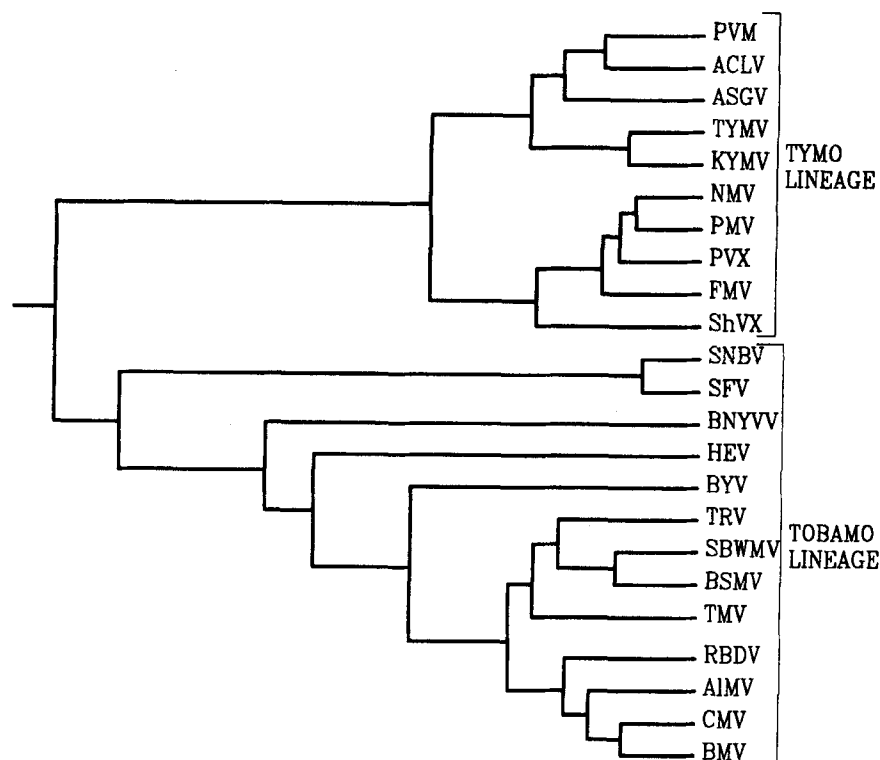
FIGURE 10. Phylogenetic analysis of the type I putative viral methyltransferases. The root position was inferred from the cluster dendrogram.

division, which was apparent in both the RdRp and the superfamily 1 helicase trees. As discussed previously (Rozanov et al., 1989b), the sequence of the putative methyltransferase domain in RuBV was anomalous, precluding unambiguous alignment with other virus methyltransferases. Thus the RuBV sequence even could not be included in the phylogenetic analysis. The methyltransferases of HEV and BNYVV did not group together either, in contrast with the grouping of the RdRps and the helicases. Where discrepancies were observed between the RdRp and the helicase trees, that is, in the positions of the alphavirus and the tymovirus, the methyltransferase tree agreed with the helicase and not with the polymerase tree.

The second type of methyltransferase domain was found exclusively in flaviviruses and showed significant similarity to cellular S-adenosylmethionine-utilizing methyltransferases (Koonin, 1993).

F. Conserved Nonstructural Domains of Unknown Function

In addition to the conserved proteins with known or at least surmised functions, several domains have been delineated in nonstructural proteins of positive-strand RNA viruses whose functions remain to be determined. The most remarkable of these was the so-called X domain, which is associated with the (putative) papain-like proteases in alphaviruses, rubella virus, hepatitis E virus, and coronaviruses (Gorbalenya et al., 1991; Koonin et al., 1992). Other conserved domains have been identified in the arteriviruses and related virus groups (den Boon et al., 1991), and in rubella virus, HEV, and BNYVV (Koonin et al., 1992). Given the high variability of the positive-strand RNA virus genomes, the importance of these domains for virus reproduction is beyond reasonable doubt but search for functional motifs that might have given clues as to their actual roles so far has been unsuccessful.

G. Capsid Proteins

Capsids of positive-strand RNA viruses have either spherical (icosahedral) or helical symmetry. Not unexpectedly, capsid proteins reveal much less sequence similarity than proteins involved in replication and expression. For icosahedral capsid proteins, significant sequence conservation has been observed only among picornaviruses, caliciviruses, and RTSV (Palmenberg, 1989; Gorbalenya et al., 1992; Shen et al., 1993), and among several groups of small plant viruses (Dolja and Koonin, 1991). However, a wide variety of viruses have conserved, so-called jelly roll conformation of the capsid protein (Rossmann and Johnson, 1989; Chelvanayagam et al., 1992). It is suspected that this type of capsid protein has evolved only once. However, capsid proteins of RNA bacteriophages (Valegard et al., 1990) and possibly those of tymoviruses have a different conformation. Taken together with the data on the tertiary structure of the alphavirus capsid protease (Choi et al., 1991), these observations indicate that generally the icosahedral capsid of positive-strand RNA viruses is polyphyletic.

It has been shown that two large families of helical capsid proteins possess several conserved motifs; moreover, two motifs appeared to be common to both families (Dolja et al., 1991; Figure 11). Thus it is likely that all capsid proteins of plant viruses with helical capsids have a common origin. Unexpectedly, a domain with obvious similarity to the rod-shaped capsid proteins was identified in the nonstructural polyprotein encoded by RNA 2 of BaYMV, which is a filamentous virus (E. V. Koonin, unpublished observations). The role for this domain in virus reproduction is unclear, but it may be important as the main conserved motifs of the capsid proteins were intact (Figure 11).

Tentative phylogenetic trees had to be generated separately for the rod-shaped family (Figure 12A) and the filamentous family of capsid proteins (Figure 12B). When these trees were superimposed on the RdRp or the helicase trees, it was clear that viruses with capsid proteins of both families are scattered among viruses with other capsid types (compare Figures 12, 2, and 5). This makes one of the strongest cases for the "shuffling" theory of virus evolution (see below). Par-

ticularly striking is the situation with potyviruses and bymoviruses whose capsid proteins were related to those of potexviruses and carlaviruses, while the nonstructural proteins of these viruses were phylogenetically remote. An interesting cluster in the tree of rod-shaped capsid proteins was formed by the "pseudocapsid" domain of BaYMV, and the capsid protein of SBWMV furovirus (Figure 12A), suggesting the possibility of horizontal transfer of the capsid protein gene from a furovirus to a bymovirus associated with a change in function.

The nucleocapsid protein of coronaviruses appears to represent an independent type of capsid structure (Lai, 1990).

F. A Note on the Reliability of the Results of Phylogenetic Analysis

Ideally, a single optimal method of tree generation should be used for phylogenetic analysis throughout. Unfortunately, however, a method like that does not exist (so far) and we were forced to use the consensus of several methods based on different assumptions. Generally, distinct groups of viral proteins (e.g., the picorna-like division, the poty-like division, the tymo-like division, etc.) were very stable independent of the method of tree generation use. On the other hand, more variation was observed both at the higher level, that is, in the branching order of these groups within supergroups, and at the lower level, that is, in the branching of smaller groups and species within the stable groups. Although this situation called for much caution in the interpretation of the results of phylogenetic analysis, we believe that there is a strong reason to consider the stable divisions to be real evolutionarily compact entities.

VI. CONSERVATION AND VARIABILITY IN THE ARRANGEMENT OF THE BUILDING BLOCKS

When comparing the tentative phylogenetic trees for individual viral proteins, we have already mentioned the prominent correlations between them. Below we will see that despite the known plasticity of the positive-strand RNA virus

400

```
              +                          -
TRV       61 DENTRFPSG-KVY  93 AVVQR---TFEKEY  27┐
PEBV      61 TRFKRFSDG-EEY  95 ALDQE---DFEEKF  33 │
BSMV      73 DVDRRFAGA-R--  87 VYTRK---TFEREL  16 │ rod-shaped
TMV       56 QVTVRFPDS--DF  62 SYNRS---SFESSS  18 │ capsid proteins
RMV       56 APNQRFPDT--GF  62 YMNRA---EFEAI-  17 │
CGMMV     56 DINSRFPDA--GF  62 VYDRA---SFEAAF  20 │
BNYVV     68 SPMTRFPQTLTMY  80 MWTRD---KFEDRF  15 │
NVMV      59 ---TRFPDTMVRT  77 LYTRT---TIENKL  15 │
BaYMVx    65 RREKRFPA---YF  57 CYTRN---IFENDN 521?┘
cons1        ....RFP.....@    .&.R.    &E...
                *               *      **

cons2        ...&R.&C..&S.    ...K&    AFDF&.
                 U     A          R       G
PeMV     144 KPTLRQIMAHFSD  28 SLARY---AFDFYE  71┐
SCMV     195 KPTLRQCMMHFSD  28 SLARY---AFDFYE  56 │
TEV      155 QPTLRQIMTHFSD  28 SLSRY---AFDFYE  56 │
TVMV     156 NPSLRQIMKHFSN  28 NLAPF---AFDFFE  57 │
BaYMV    202 NGGLRRIMRNYSD  22 ANAKY---AFDFFV  49 │
PVX      112 CT-LRQFCMKYAP  21 PEHKF--AAFDFFN  78 │
PMV       95 GTSLRKFCRYFAP  20 PSAKF--AAFDFFD  71 │
WC1MV     95 HCTIRQFCMYFAN  21 EESKF--AGFDFFD  48 │
NMV      108 NITPRQFCMYFAK  21 DDCKF--AQFDFFE  84 │ filamentous
LVX       92 L-PLRQFCRYYAK  21 AEARF--AAFDFFD  61 │ capsid proteins
PVS      180 A-GLRKVCRLYAP  21 WNARF--AAFDTFD  66 │
PVM      189 AETLRRVCRLYAP  21 YEDRF--AAFDCFD  67 │
LSV      178 A-GLRKVCRLYAP  21 YNTRF--AAFDTFD  66 │
ACLV      95 --TFRQVCEAFAP  25 --SKYPELMFDFNK  52 │
ASGV       ? --TFRKLCEPFAD  24 AFEKSPWVAFDFAT  44 │
BYV      104 PNKLRCFCRTFQK  24 AEDHY--LAADFIS  49 │
BYV p24  123 PNPVRTFCATFED  24 SGYEF--LGADFL-  45 │
CTV      132 TNALRVWGRTNDA  24 AGYHY--LCADFL-  43┘
```

FIGURE 11. Conserved sequence motifs in the capsid proteins of positive-strand RNA viruses with elongated particles. Motifs extracted from previously published alignments (Dolja et al., 1991; Boyko et al., 1992) were combined and updated. The conserved positively charged residue (+) in motif I is thought to form a functionally important salt bridge with the conserved negatively charged residue (–) in motif II. BaYMVx is a newly identified region of the RNA2 polyprotein of BaYMV, which is related to the capsid proteins of rod-shaped viruses (see text). p24 of BYV is the diverged copy of the BYV capsid protein that is apparently not included in the virions (Boyko et al., 1992). The consensus was derived separately for the rod-shaped family and the filamentous family, and the correspondence between the two patterns is shown as in Figures 3 and 9. Viruses with partially sequenced genomes not mentioned in Table 1: NVMV, *Nicotiana vellutina* mosaic virus (a putative furovirus); RMV, ribgrass mosaic virus; SHMV, sunnhemp mosaic virus; CGMMV, cucumber green mottle mosaic virus (tobamoviruses); SCMV, sugarcane mosaic virus (a potyvirus); PVS, potato virus S; LSV, lilia symptomless virus (carlaviruses); LVS, lilia virus S (a potexvirus); CTV, citrus tristeza virus (a closterovirus). Sources for these sequences are cited in Dolja et al., 1991; Boyko et al., 1992.

genome, this congruency is complemented by existence of stable conserved gene arrays. We will be using the tentative phylogenetic tree for the RdRps (Figure 2) as the general framework for this discussion.

A. Conserved Gene Arrays

The presence of conserved arrays of genes coding for proteins involved in replication and expression in positive-strand RNA virus genomes is obvious (Figure 13). The most persistent gene combination is the helicase-polymerase, with the helicase gene being typically located upstream of the polymerase gene. This "rule" holds for all three polymerase supergroups (i.e., for viruses with relatively large genomes encoding helicases), and accordingly for all the three helicase superfamilies that are specifically associated with the former (Figure 13). It is violated only in the arteri-like viruses, and in the double-stranded genome

TRV
PEBV
BSMV
TMV
RMV
CGMMV
BNYVV
NVMV
SBWMV
BaYMVx
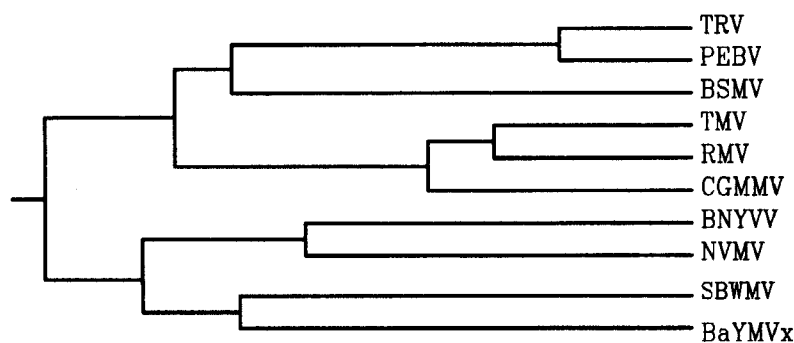
A

FIGURE 12. Phylogenetic analysis of the capsid proteins of positive-strand RNA viruses with elongated particles. (a) rod-shaped viruses; for additional details see caption to Figure 11. (b) filamentous viruses; p26 of CTV is the apparent homologue of the BYV p24. The root position was inferred from the cluster dendrogram.

ACLV ⎤ CAPILLO
ASGV ⎦ LINEAGE
PeMV ⎤
TEV ⎥
SCMV ⎥ POTY/BYMO
TVMV ⎥ LINEAGE
BaYMV ⎦
PVX ⎤
PMV ⎥
LVX ⎥
WClMV ⎥
NMV ⎥ POTEX/CARLA
PVS ⎥ LINEAGE
PVM ⎥
LSV ⎦
BYV ⎤
CTV ⎥ CLOSTERO
BYVp24 ⎥ LINEAGE
CTVp26 ⎦
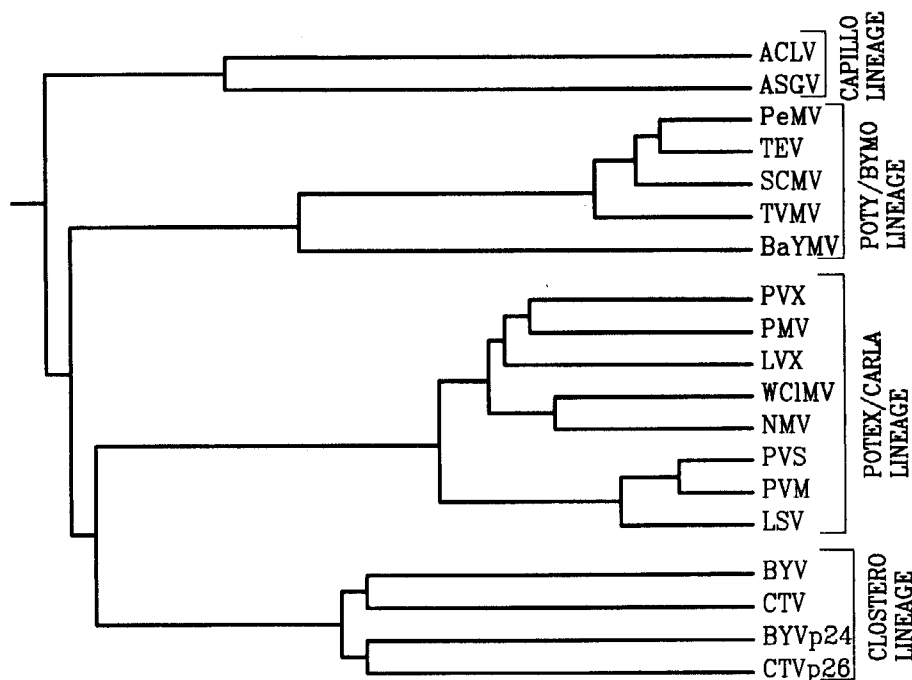
FIGURE 12B

of HyAV. In addition, tricornaviruses and hordeiviruses "ridicule" the rule by assigning the polymerase and the helicase genes to separate genome segments (Figure 13).

A further variation of the RdRp-helicase theme is the presence of two genes for putative helicases flanking the RdRp gene from both sides in potexviruses and carlaviruses, and isolated on a
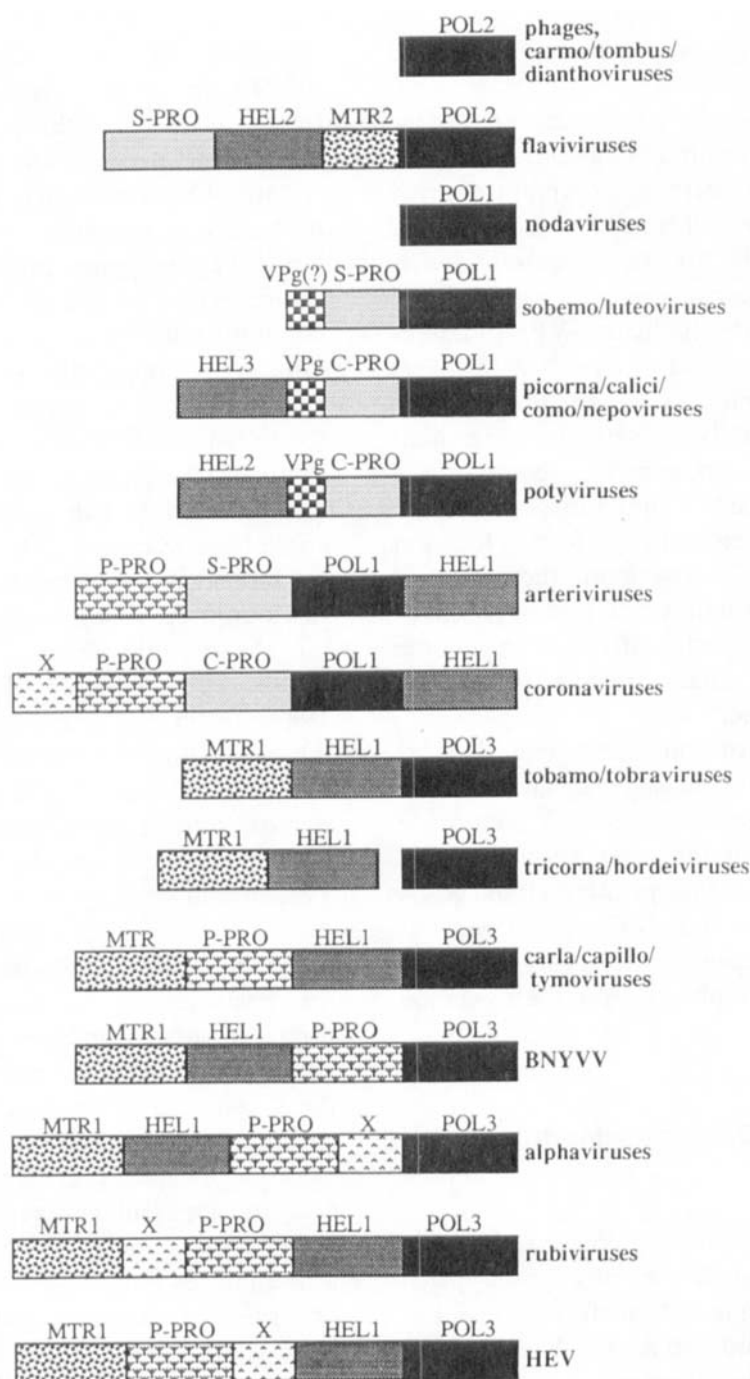
FIGURE 13. Conservation and variability in the arrangement of the "core" genes of positive-strand RNA virus genes encoding proteins involved in genome replication and expression. This and the subsequent schemes are not to scale. Only the principal proteins (domains) involved in replication and expression are shown. Omission of less well-conserved domains in this and the subsequent schemes is not specified in each case. The genes coding for the conserved proteins are highlighted identically in all schemes. POL1,2,3 — RNA-dependent RNA polymerase of the respective supergroup; HEL1,2,3 — (putative) RNA helicase of the respective superfamily; S-PRO — serine chymotrypsin-like protease; C-PRO — cysteine chymotrypsin-like protease; P-PRO — papain-like cysteine protease; MTR1,2 — methyltransferases of type 1 and 2, respectively. The "shell" genes coding for virion and "accessory" proteins are not shown.

separate RNA segment in hordeiviruses and BNYVV.

Other gene clusters are restricted to one of the RdRp-based divisions. These include the methyltransferase-helicase-polymerase array found in all viruses with supergroup 3 RdRps (Rozanov et al., 1992b), the protease-helicase-polymerase array conserved in flaviviruses and pestiviruses, and the (helicase)-VPg-protease-polymerase array typical of viruses with supergroup 1 RdPps (Figure 13). Importantly, the latter array is partially conserved also in small viruses of this division, namely, sobemoviruses and luteoviruses that lack the helicase gene (with some uncertainty remaining as for the location of the VPg gene). This supports the notion of evolutionary compactness of this division of viruses despite the drastic difference in genome size and the concomitant presence or absence of the helicase gene.

The existence of conserved gene clusters, and particularly the quasi-invariant helicase-polymerase arrangement, raise the possibility that not only a common virus ancestor could give rise to all the extant positive-strand RNA viruses and at least some dsRNA viruses but also that certain features of this hypothetical ancestor may be deciphered from their genome organization.

B. Evidence of Gene and Module Shuffling

As described in the previous section, the genes for RdRp, the helicase, the chymotrypsin-like protease, the methyltransferase, and the VPg tend to form ordered arrays. Nonetheless, even the relative arrangement of these genes appears to reflect major recombination events. Two apparent events of this type involve viruses with supergroup 1 polymerases. First, the poty-like viruses encode a superfamily 2 helicase related to the helicases of flaviviruses and pestiviruses instead of superfamily III helicase that is typically associated with supergroup 1 RdRp. Second, arteriviruses and their relatives encode a helicase belonging to superfamily 1 instead of a superfamily 3 helicase. In

the arteri-like viruses and also in the HyAV dsRNA virus, these changes are accompanied by the reversal of the positions of the polymerase and the helicase genes in the genome (Figure 13). Two other, not described previously, acts of recombination within the array of housekeeping genes might have occurred in viruses with supergroup 3 RdRp. First of these, combining the gene array typical of rubi-like viruses (RdRp, papain-like protease, X domain) with the methyltransferase and the helicase from the tobamo-like lineage, may account for the origin of the alphaviruses. The second event, also between the RdRp and the helicase genes, could have occurred in the tymovirus lineage.

Other genes of positive-strand RNA viruses enjoy a much greater freedom, supporting the module concept of virus evolution (Gibbs, 1987; Zimmern, 1988; Morozov et al., 1989; Dolja and Carrington, 1992). Among the genes encoding proteins involved in expression, the gene for the papain-like protease is most subject to "wandering" along the genome (Gorbalenya et al., 1991; Koonin et al., 1992; Figure 13).

As for the genes encoding virion components, variation of both the type of capsid and the capsid gene location appears to be a general principle rather than exception. The typical position of these genes is either at the 5' end or at the 3' end of the genome. Comparison of picornaviruses and caliciviruses provides a clear-cut example of capsid protein gene relocation without any apparent concomitant change in the nonstructural gene complex. Isolation of capsid protein genes on a separate genome segment is observed in different groups, for example, comoviruses and nepoviruses when compared with picornaviruses, RTSV and PYFV; and tricornaviruses, RBDV, and tobraviruses when compared with the tobamoviruses. Change of the capsid type is obvious, for example, after comparison of icosahedral tymoviruses with related viruses having elongated capsids, for example, potexviruses (Rozanov et al., 1990), or of the icosahedral arteriviruses with helical corona- and toroviruses (Godeny et al., 1993). Related viruses may also differ by the presence or absence of genes for envelope proteins as

404

exemplified by the comparison of RuBV and HEV (Koonin et al., 1992).

The position of at least some of the genes in the genome of positive-strand RNA viruses appears to be under selective pressure to ensure the optimal level of expression. Thus the RdRp is usually located downstream of the other proteins involved in replication so that its expression could be regulated, for example, by an upstream leaky termination codon as in tobamoviruses, tobraviruses, and some of the alphaviruses (Dolja and Carrington, 1992). Conversely, the coat protein coding region usually is the 5'-terminal one in an mRNA (which obviously does not necessarily correspond to the 5'-terminal position in the genome) so as to ensure a high level of expression. A notable exception is the poty-like viruses whose capsid protein occupies the 3'-terminal position in the monocistronic genome RNA and appears not to be subject to expression regulation (Dougherty and Carrington, 1988).

The extent of the apparent gene exchange between distantly related groups of positive-strand RNA appears to have been so high that genomes of certain viruses look really patchy, with different gene products having homologs in different other viruses, and sometimes also among cellular genes. This point is illustrated in Figure 14 by using the potyvirus genome as an example.

C. Direct Evidence of Gene Exchange between Positive-Strand RNA Viruses

The above discussion centered at apparent macroevolutionary events when gene exchange between remote viruses seemed to have led to a modification in the genome layout of large virus groups. Regretfully but typically of major evolutionary breakthroughs, such events are hard to register directly. But at a smaller evolutionary scale direct evidence for recombination between different positive-strand RNA viruses has been obtained (Lai, 1992). Examples are, in the order of decreasing evolutionary distance between the recombining partners, BYDV (strain PAV) whose RdRp originates from a carmo-related virus, while the "shell" proteins are clearly related to those of luteoviruses (Martin et al., 1990); sunnhemp mosaic virus, an apparent hybrid between TMV and TYMV (Meshi et al., 1981); and Western equine encephalitis virus, which appears to be an evolutionarily recent hybrid between EEEV and another alphavirus closely related to Sindbis virus (Hahn et al., 1988). In addition, recombinant viruses have been obtained in experimental systems. Intertypic recombinants of poliovirus have been isolated under controlled conditions in tissue culture (King et al., 1982; Tolskaya et al., 1983). Parallel studies in a plant system have
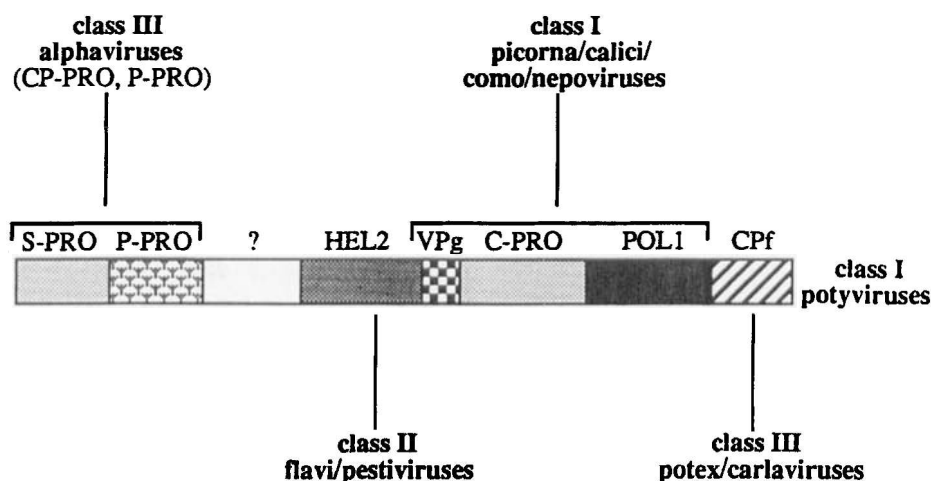

FIGURE 14. Chimeric organization of a positive-strand RNA virus genome. A rough scheme of the potyvirus polyprotein is shown. For each protein (domain) groups of viruses encoding related proteins are indicated. CP-f — filamentous capsid protein; the other abbreviations are as in Figure 13.

demonstrated recombination between different components of the BMV genome (Bujarski and Kaesberg, 1986; Rao et al., 1990). Recently, a chimeric plant virus has been designed from BMV by replacing its movement protein gene by the unrelated gene for the movement protein of TMV (De Jong and Ahlquist, 1992).

A very interesting example of a different type of gene exchange possibly relevant for the origin of divided genomes is provided by PEMV. The genome of this virus consists of two components, one of which encodes a luteovirus-like and the other a carmovirus-like replicative gene complex. On the other hand, the capsid protein and the movement protein are provided in-trans by RNA 1 and RNA 2, respectively (Demler and de Zoeten, 1991; Demler et al., 1993; Mushegian and Koonin, 1993).

D. Gene Duplications

Gene duplication followed by mutational diversification of the two gene copies is the major general evolutionary mechanism (Ohno, 1970; Kimura, 1983). Several examples of gene duplication (or multiplication) are found in positive-strand RNA viruses. These include the apparent duplication of the helicase gene in several groups of plant viruses (Gorbalenya et al., 1988b), chymotrypsin-related protease gene duplication in entero- and rhinoviruses (Blinov et al., 1985; Gorbalenya et al., 1986; Bazan and Fletterick, 1988, 1990), the duplication of the capsid protein gene in closteroviruses (Boyko et al., 1992), and triplication of the VPg gene in FMDV picornavirus (Forss et al., 1984). An exact tandem duplication of the N-terminal portion of the BSMV RdRp has been described (Afanasiev et al., 1986; Gustafson et al., 1987). It is also widely believed that a triplication has taken place in the evolution of picornavirus and comovirus capsid protein genes, although in this case it was evident only at the level of protein tertiary structure (Palmenberg, 1989; Rossmann and Johnson, 1989). In all these examples it could not be ruled out that the duplication might be "fictitious", the presence of two or more gene copies being actually due to recombination between related viruses.

E. Relationships between Positive-Strand RNA Viral and Cellular Proteins.

Many positive-strand RNA viral proteins show varying degree of sequence similarity to cellular proteins (Gorbalenya, 1992). These relationships are of two very different types. Proteins involved in fundamental steps of virus replication and expression and conserved in a wide range of virus groups, like RNA helicases and proteases, are more or less distantly related to cellular proteins with the analogous activity. In these cases the recruitment of cellular genes by the virus genome, if it ever occurred (see discussion below), could only be a very ancient evolutionary event. On the other hand, there are clear-cut examples of recent capture of cellular genes by positive-strand RNA virus genomes including various cellular insertions in pestiviruses (Meyers et al., 1991), and a gene coding for a heat shock protein homolog in closteroviruses (Agranovsky et al., 1991a). As with the examples of intervirus recombination, the disparity is obvious between putative macroevolutionary events that are of very general significance but are difficult to demonstrate directly, and self-evident microevolutionary events of limited impact.

VIII. AN OUTLINE OF PHYLOGENETIC TAXONOMY OF POSITIVE-STRAND RNA VIRUSES.

We feel that the results of phylogenetic analysis and the genome organization comparison are consistent enough to allow the delineation of distinct large taxa of positive-strand RNA viruses. Consider first the viruses with RdRps of supergroups 1 and 3. Clearly, the existence of invariants in the arrangement of genes encoding the main enzymes of replication and expression supports the notion of viruses with these types of RdRp as evolutionarily compact major divisions. The situation with supergroup 2 is less obvious as it represents the ultimate case of divergence, with the RdRp being the only common denominator of the constituent viruses.

The reality of smaller but still very diverse, from the point of view of standard virus taxonomy, groups within these vast divisions is even more obvious. They are supported both by phylogenetic trees for different proteins and by genome context analysis. With the wide spread of recombination, classification of viruses necessarily is to some extent a matter of convention. Nevertheless, the view of a virus as a relatively stable, slowly evolving "core" of the replicative genes accompanied by a much more flexible "shell" of the genes coding for virion components and "accessory" proteins appears to be a strongly preferable, and in a sense the "correct" one. This concept implies that the derived phylogeny of the "core" gene complex should constitute the basis of phylogenetic taxonomy of positive-strand RNA viruses, at least when higher taxa are considered.

Based on these notions and on the examination of individual phylogenetic trees, we outline here a proposal for a new, phylogenetic taxonomy of positive-strand RNA viruses (Table 2). Its purpose is to tentatively define higher-than-family rank taxa, in which viruses are brought together by evolutionary compactness (monophyly) of their replicative gene "core". All of the proposed orders and families are supported by congruent phylogenetic trees for at least two, and more commonly three or four conserved proteins. Generally, an order includes viruses with a common theme in the arrangement of the "core" genes but significant variability of the "shell" organization. A family usually consists of viruses with a common plan of genome organization, including both the "core" and the "shell". In some cases, however, we felt that the apparent evolutionary compactness justified placing viruses with considerable differences in the gene organization in one taxon. Thus, a drastic deviation from the general principle of formation of orders is observed in the proposed order POTYVIRALES, with the double-stranded HyAV grossly differing from potyviruses and bymoviruses in the composition and order of the "core" genes. Less dramatically, in the order TYMOVIRALES, despite the general conservation of the order of "core" genes, some of the viruses have apparently lost the papain-like protease, together with the strategy of expression by

polyprotein processing. At the lower family level examples of diversity are provided by the families Comoviridae, Tricornaviridae, and Tobamoviridae, in all of which viruses belonging to different genera have different number of genome segments (see also the next section).

Inevitably, at some points during the development of this tentative virus taxonomy difficult and ambiguous decisions had to be made. An example of a complicated situation is the family Alphaviridae whose helicase and methyltransferase grouped with those of the *TOBAMO-VIRALES*, while the genome organization, and to some extent the polymerase phylogeny, rather supported association with *RUBIVIRALES* (see above). We placed Alphaviridae in the latter order (Table 2), with the understanding that the upstream portion of the replication gene complex could be derived via recombination with a tobamo-like virus.

Mostly, the proposed taxonomic innovations are amendments to the existing system, sorting out already established families and groups. However, a few instances of regrouping are obvious including the relocation of arteriviruses from Togaviridae to the proposed order *ARTERIVIRALES* (den Boon et al., 1991), transfer of BNYVV from the Furoviruses to the order *RUBIVIRALES*, and relocation of ACLV from the Closteroviruses to the tentative order *TYMOVIRALES* (perhaps its inclusion in Capilloviruses).

A major challenge to the taxonomists is the proposed inclusion of dsRNA viruses or virus-like elements in some of the subdivisions of positive-strand RNA viruses. We believe that as long as this seems to reflect the actual course of evolution, this classification is to be absorbed, however counterintuitive.

Clearly, it cannot be expected that the entire proposed system or even its main features will be officially accepted in the near future. The standard latinized nomenclature is used here solely for convenience and is not intended to create any impression of officialdom. Our hope is, however, that these proposals will be useful in at least bringing the concept of phylogenetic taxonomy of viruses to practical consideration.

TABLE 2
Draft of Phylogenetic Taxonomy of Positive-Strand RNA Viruses

Class[a]	Order	Family	Genera/Groups	Host
I. Picornavirata	Picornavirales	Picornaviridae	Enterovirus Rhinovirus Cardiovirus Aphthovirus Hepatovirus Echovirus (ECHO22)	Vertebrates
		Comoviridae	Comovirus Nepovirus PYFV group	Plants
		Caliciviridae	Calicivirus	Vertebrates
	Potyvirales	Potyviridae	Potyvirus Bymovirus	Plants
		Hypoviridae	HyAV group	Fungi
	Sobemovirales	Sobemoviridae	Sobemovirus	Plants
		Luteoviridae	Luteovirus	Plants
		Nodaviridae(?)	Nodavirus	Insects
	Spheridiplorna-virales	Spheridiplorna-viridae	ScV group LRV group	Fungi Protozoa
	Arterivirales	Arteriviridae	Arterivirus	Vertebrates
		Coronaviridae	Coronavirus Torovirus	Vertebrates
II. Flavivirata	Flavivirales	Flaviviridae	Flavivirus	Vertebrates Arthropods
	Pestivirales	Pestiviridae	Pestivirus	Vertebrates
	Carmovirales	Alloluteo-viridae	MCV group Alloluteovirus (BYDV-PAV)	Plants
		Dianthoviridae	Dianthovirus	Plants
		Carmoviridae	Carmovirus Tombusvirus Necrovirus	Plants
	Levivirales(?)	Leviviridae	Levivirus Allolevivirus	Bacteria
III. Rubivirata	Rubivirales	Rubiviridae	Rubivirus	Vertebrates
		Hepeviridae	HEV group	Vertebrates
		Beneviridae	BNYVV group	Plants
		Alphaviridae	Alphavirus	Vertebrates Insects
	Tobamovirales	Tobamoviridae	Tobamovirus Tobravirus Hordeivirus Furovirus (SBWMV)	Plants
		Tricornaviridae	Cucumovirus Bromovirus AlMV group Idaeovirus (RBDV)	Plants
		Closteroviridae	Closterovirus (BYV)	Plants
	Tymovirales	Tymoviridae	Tymovirus	Plants
		Carlaviridae	Carlavirus Capillovirus (ASGV, ACLV)	
		Potexviridae	Potexvirus	Plants

[a] The standard Latin forms for the (proposed) names of classes, orders, and families were used. Two of the proposed new family names were constructed using the standard method of adopting the prototype virus name: Hepeviridae after HEPatitis E virus, Beneviridae after BEet NEcrotic yellow vein virus, Hypoviridae after HYPOvirulence-associated virus. Other family names had to be proposed ad hoc: Spheridiplornaviridae (dsRNA-containing viruses with spherical virions) and Alloluteoviridae (for BaYMV-PAV, which previously has been included in luteoviruses but has the RdRp related to that of carmoviruses). The proposed genus name *Idaeovirus* is from Ziegler et al., 1992.

IX. TOWARD RECONSTRUCTION OF VIRUS EVOLUTION

The results of phylogenetic analysis of individual viral proteins combined with the observations on conservation and variability in genome organization allow an attempt of reconstruction of the evolutionary history of positive-strand RNA viruses. The phylogenetic trees provide the "arrow of time" necessary to introduce the direction of evolution. There is no way to produce a single unequivocal scenario as it is typical of any evolutionary reconstruction. What can be constructed is the apparently most parsimonious and hence the most likely evolutionary scenario. Qualitatively, this approach is analogous to the maximum parsimony approach in phylogenetic analysis.

As discussed above, the delineation of the main virus divisions is generally supported by phylogenetic trees for three different conserved proteins and by the conservation of the arrangement of the respective genes in the genome. Some departures notwithstanding, the array of these "core" genes in each division appears to comprise a single evolving unit. Considering the tree-like evolution of this unit to be the "backbone" of the evolutionary scenario, the putative events leading to changes in genome organization, expression, and replication can be described in concise terms.

These events include: (1) gene duplication; (2) capture of new (cellular) genes; (3) recombination with distantly related viruses; (4) deletion of "core" (e.g., helicase or papain-like protease) or "shell" genes; (5) gene rearrangement within the replicative gene complex (e.g., "wanderings" of the papain-like protease domain and the associated "X" domain); (6) formation of segmented genomes by either split of the viral genome or capture of a heterologous RNA segment; (7) relocation of the genes encoding virion components and other "shell" genes; (8) insertion/deletion of "accessory" domains without apparent change of the nonstructural gene arrangement; (9) emergence of mechanisms for subgenomic mRNA synthesis.

We believe that all positive-strand RNA viruses and the closely related dsRNA viruses should have evolved from a common ancestral virus.

This conviction is based primarily on the universal conservation of the RdRp sequence. By looking into features of genome organization common to different virus groups, it is possible to envisage the genome arrangement and expression strategy of putative ancestors of large divisions of viruses, and ultimately the original ancestor. Consider first the easier task of deriving the features of the ancestors for the three virus classes. Naturally, all of these hypothetical ancestor viruses should have encoded an RdRp. Beyond that point, we discuss them separately.

A. Class I

All viruses of this class, with the exception of the proposed *POTYVIRALES* and the Coronaviridae family, have icosahedral capsids composed of the jelly roll type proteins. Clearly, this should also be a feature of the ancestor virus. Further, all viruses of this class, with only the exception of nodaviruses, express the majority of their proteins via polyprotein processing mediated by a chymotrypsin-like protease. We surmise that the ancestor virus also encoded such an enzyme. A more difficult question is whether the ancestor virus encoded the RNA helicase. Although the viruses of the proposed order *SOBEMOVIRALES* have small genomes and lack the helicase gene, the majority of the viruses of this class encode a helicase and we consider the presence of this gene in their common ancestor the more likely possibility (see also discussion below). It was of interest to investigate whether the N-terminal region of the replicase proteins of the viruses lacking the helicase gene contain sequence motifs indicative that they might comprise an unusual helicase or isolated C-terminal domain of one of the groups of known helicases. The latter possibility could be especially plausible as an apparent stand-alone C-terminal helicase domain has been described recently in a human transcription factor (Koonin, 1992b). However, detailed searches failed to reveal any "helicase-like" features in the N-terminal domains of the replicative proteins of *SOBEMOVIRALES, NODAVIRALES,* and the astrovirus (E. V. Koonin, unpublished observations).

A
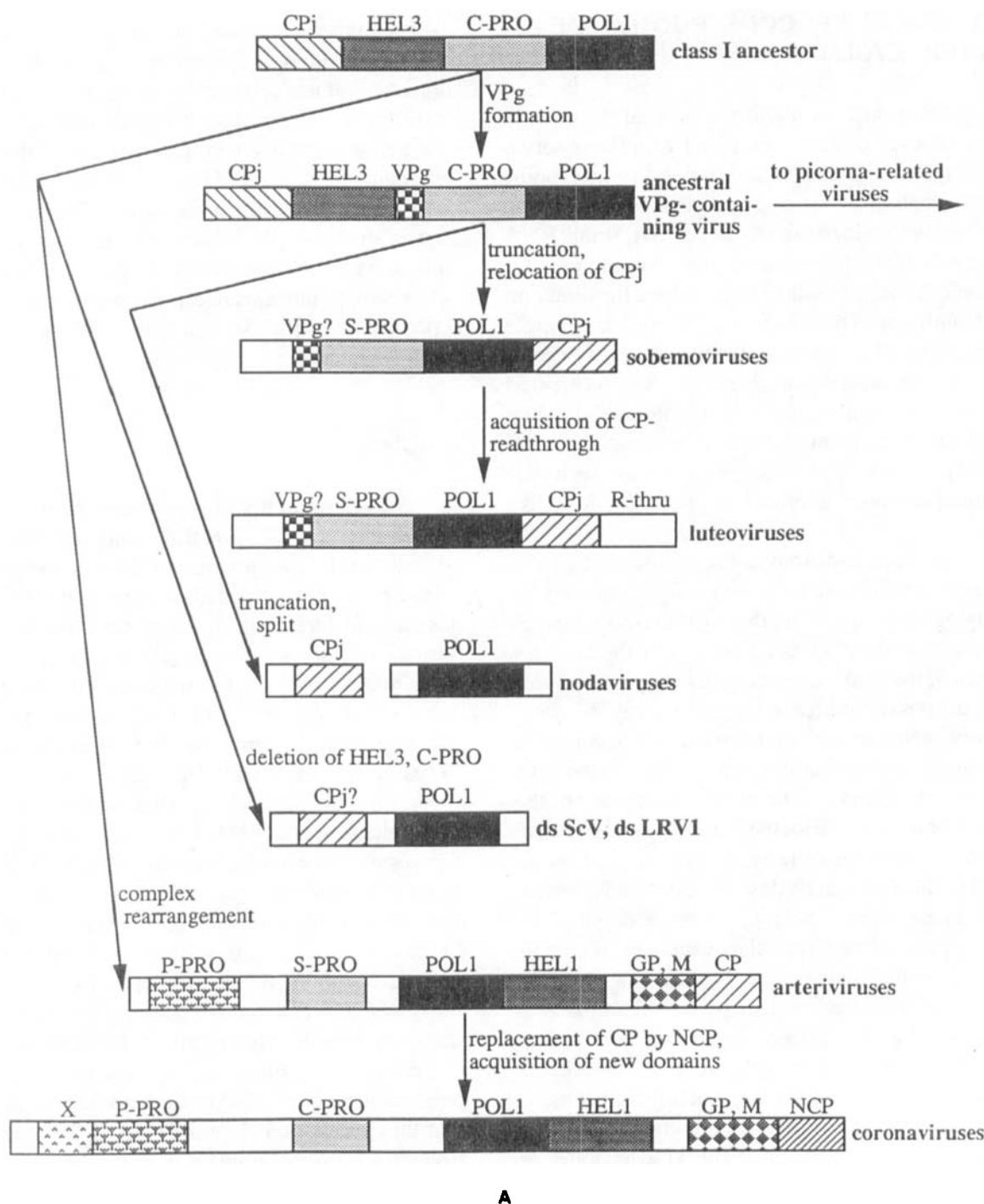
FIGURE 15. Tentative evolutionary scenario for class I of positive-strand RNA viruses. (a) The proposed orders Sobemovirales and Arterivirales; (b) the proposed orders Picornavirales and Potyvirales; This and subsequent schemes are deliberately simplified in that only the main virus genes and the principal postulated evolutionary events are shown. The arrows connecting virus groups are to be interpreted in the sense that, for example, RTSV might have evolved from an ancestral form with genome organization resembling that of picornaviruses, not from extant picornaviruses. The possible events leading to the origin of HyAV and arteri-like viruses are not shown in detail. A scenario for the evolution of HyAV from a poty-like virus has been published previously (Koonin et al., 1991b). The possible pathways of evolution for Arterivirales will be presented elsewhere. CPj — icosahedral capsid protein with jelly roll conformation; CPf — filamentous capsid protein; CPr — rod-shaped capsid protein; R-thru — domain expressed by translation readthrough; X — conserved domain of unknown function accompanying the papain-like proteases of animal viruses (Gorbalenya et al., 1991; Koonin et al., 1992); NCP — nucleocapsid protein; GP — envelope glycoprotein(s); M — membrane protein; MP — plant virus cell-to-cell movement protein. The other designations are as in Figure 13. For further details see text and caption to Figure 13.

410
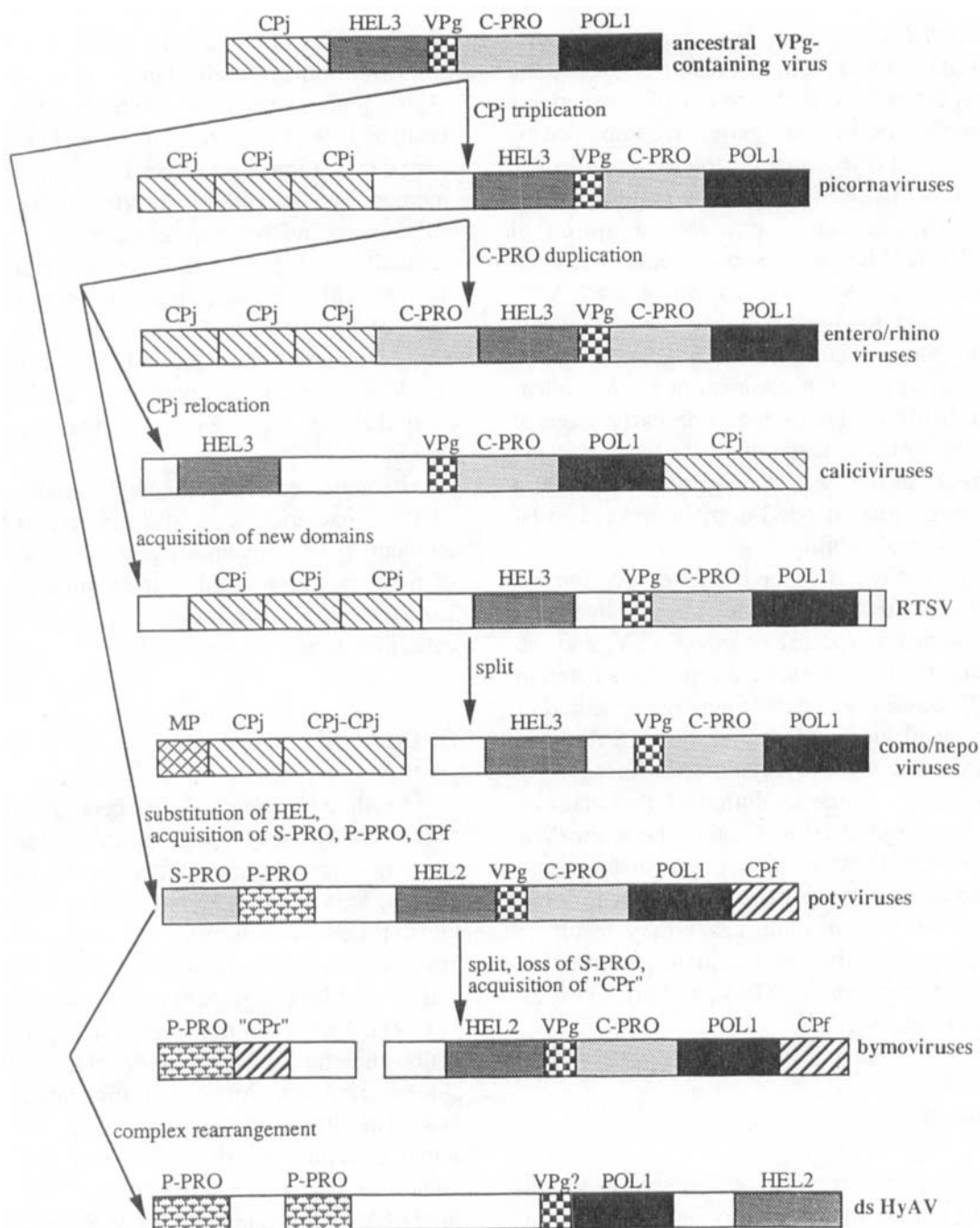
FIGURE 15B

Using the same logic to derive the probable ancestral features, we gather that the ancestral virus should have had a nondivided genome and used polyprotein processing by a virus-encoded protease as the sole method of generation of its mature proteins. With this hypothesis on the genome organization of the probable ancestor, re-construction of the simplest (most parsimonious) evolutionary history of this virus class is straight-forward enough (Figure 15). The major evolutionary events, correlated with the emergence of the virus divisions within this class, include sub-stitution of the helicase gene in the lineages lead-ing to the ancestors of the *POTYVIRALES* and

ARTERIVIRALES orders, apparently via recombinational transfer from viruses of classes II and III, respectively; and truncation of the genome (deletion of the helicase gene), accompanied by relocation of the capsid protein gene in the *SOBEMOVIRALES* lineage. The majority of the viruses of this class, with the exception of *ARTERIVIRALES* and Nodaviridae, have a small protein (VPg) covalently linked to the RNA 5'-terminus and thought to prime replication and transcription (Vartapetian and Bogdanov, 1987). The protein-primed mechanism of RNA replication probably evolved once at an early stage of evolution, concomitantly with the origin of the VPg itself. In the case of VPg, recruitment of a preexisting virus-encoded domain appears to be an attractive possibility.

The additional mechanism of expression via subgenomic mRNA might have been "invented" independently in the calicivirus, RTSV, and the sobemovirus lineages (see the references listed in Table 1). Similarly, bipartite genome appeared to have evolved independently in the RTSV/como/nepovirus and the poty/bymovirus lineages. A typical feature of the evolution of this class of viruses are large insertions within the replicative gene complex. These apparent evolutionary events cause no change of the main features of genome organization, but in some cases may result in more than a twofold difference in the genome size (e.g., picornaviruses vs. RTSV, and arteriviruses vs. coronaviruses).

B. Class II

Like class 1, this class combines viruses with relatively large genomes having genes for a helicase and a chymotrypsin-like protease with small viruses lacking those genes. The RdRp phylogenetic tree (Figure 2C) reveals an association between one of the groups of small viruses (carmovirus-related plant viruses), and pestiviruses that have much larger genomes comprising genes for a helicase and a protease arranged essentially as in flaviviruses (Figure 16). This suggests that at least the common ancestor of the eukaryotic viruses of this class also had these genes and might have been essentially very similar to the class I ancestor (compare Figures 16 and 15). A

similar series of events may have accounted for the evolution of small plant viruses within this class: genome truncation accompanied, in this case, by loss of both the helicase and the protease genes; relocation of the capsid protein gene supplemented by "invention" of the mechanism for subgenomic mRNA formation; and in one case, acquisition of a second RNA segment (Figure 16). As with the small viruses of class I, no vestiges of the helicase and/or the protease domain could be identified in the N-terminal regions of the RdRp-containing proteins of carmoviruses-related viruses (E. V. Koonin, unpublished observations).

As indicated above, the phylogenetic position of RNA bacteriophages and the related dsRNA elements is very uncertain, and so is the pathway of their evolution. With this in mind, evolution from the same common ancestor by genome truncation could not be ruled out.

C. Class III

As already discussed, the genome organization of the viruses belonging to this class is in a sense more homogeneous than in the other two classes. This allows a more confident reconstruction of the common ancestor that apparently should have encoded RNA helicase, methyltransferase, and most likely a papain-like protease (Figure 17). The nature of the ancestral capsid is the major uncertainty for this virus class, which includes numerous viruses with elongated capsids as well as viruses with icosahedral capsids whose actual structure has not been resolved. At this point, we will make a speculation that is not suggested by but is compatible with the observed variation of the capsid structure. We propose that the capsid autoprotease of alphaviruses may be the ancestral capsid protein at least for this class of viruses, and as discussed in the next section, perhaps for positive-strand RNA viruses in general. Interestingly, all known viruses of this class exploit subgenomic mRNAs in their expression, suggesting that this mechanism has evolved in the ancestor virus. The evolution of this virus class appeared to proceed via reorganizations mostly involving the papain-like protease domain, including its independent deletion in the
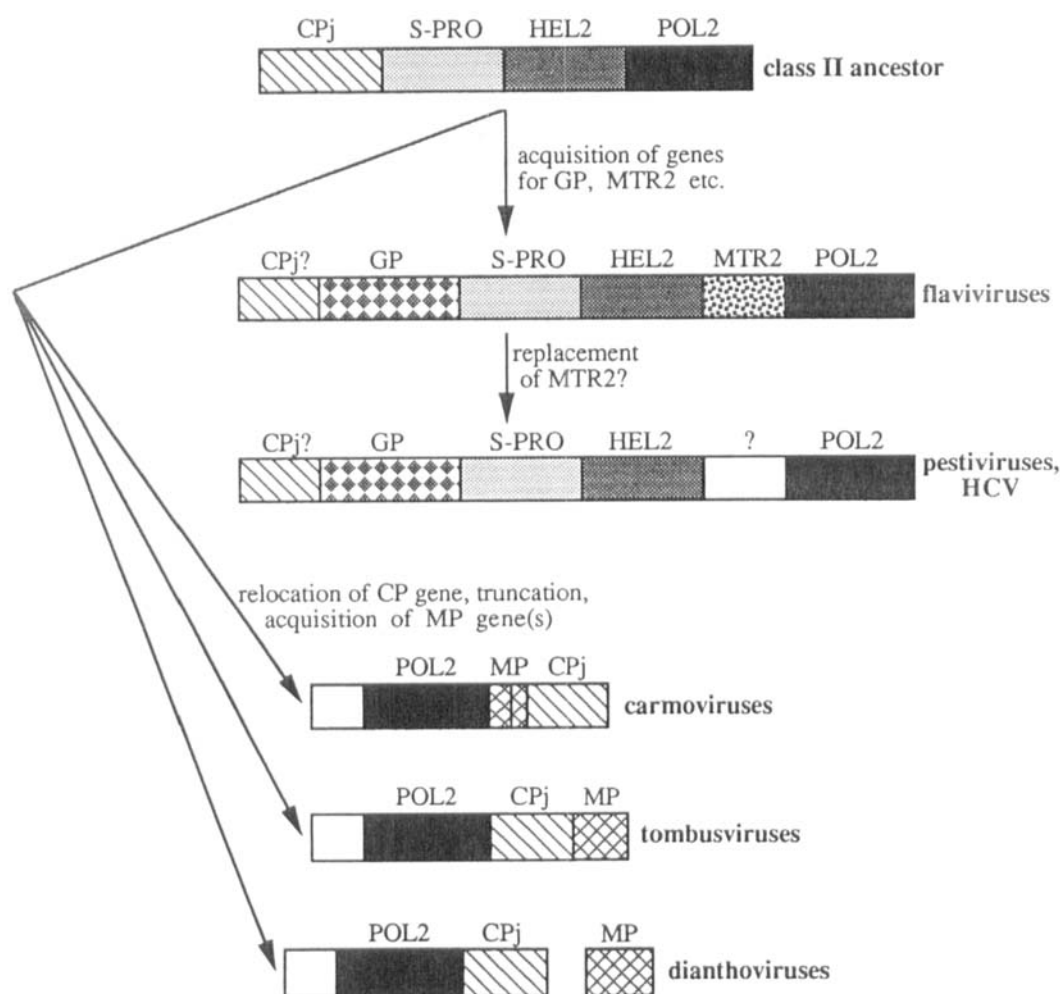
FIGURE 16. Tentative evolutionary scenario for class II of positive-strand RNA viruses. See the captions to Figures 13 and 15 for the designations.

TOBAMOVIRALES and *TYMOVIRALES* lineages, emergence of multipartite genomes, and the postulated recombination downstream from the helicase gene leading to the origin of the alphaviruses and the tymoviruses (Figure 17).

An interesting illustration of the shuffling theory of virus evolution is provided by the "accessory" viral helicase, which together with the accompanying genes for two small proteins comprises the so-called triple gene block (Morozov et al., 1989). The triple block appears to be involved in virus cell-to-cell movement (Petty et al., 1990; Beck et al., 1991; Gilmer et al., 1992), demonstrating a striking example of change in the function of the putative helicase after the gene duplication. This block is found, in different genomic context, in all three proposed orders of the class

III (Figure 17). We suggest that the triple block has evolved in the common ancestor of the *TYMOVIRALES* after the duplication of the helicase gene, and has been captured subsequently by BSMV, BNYVV, and NVMV in the form of a separate RNA segment.

D. The Ultimate Common Ancestor

Clearly, this ultimate "first positive-strand RNA virus" encoded an RdRp that subsequently has given rise to the three main lineages of evolution of this gene. It is also logical to propose that this virus had an icosahedral capsid and expressed its genetic information via polyprotein processing. Then, the simplest way to achieve
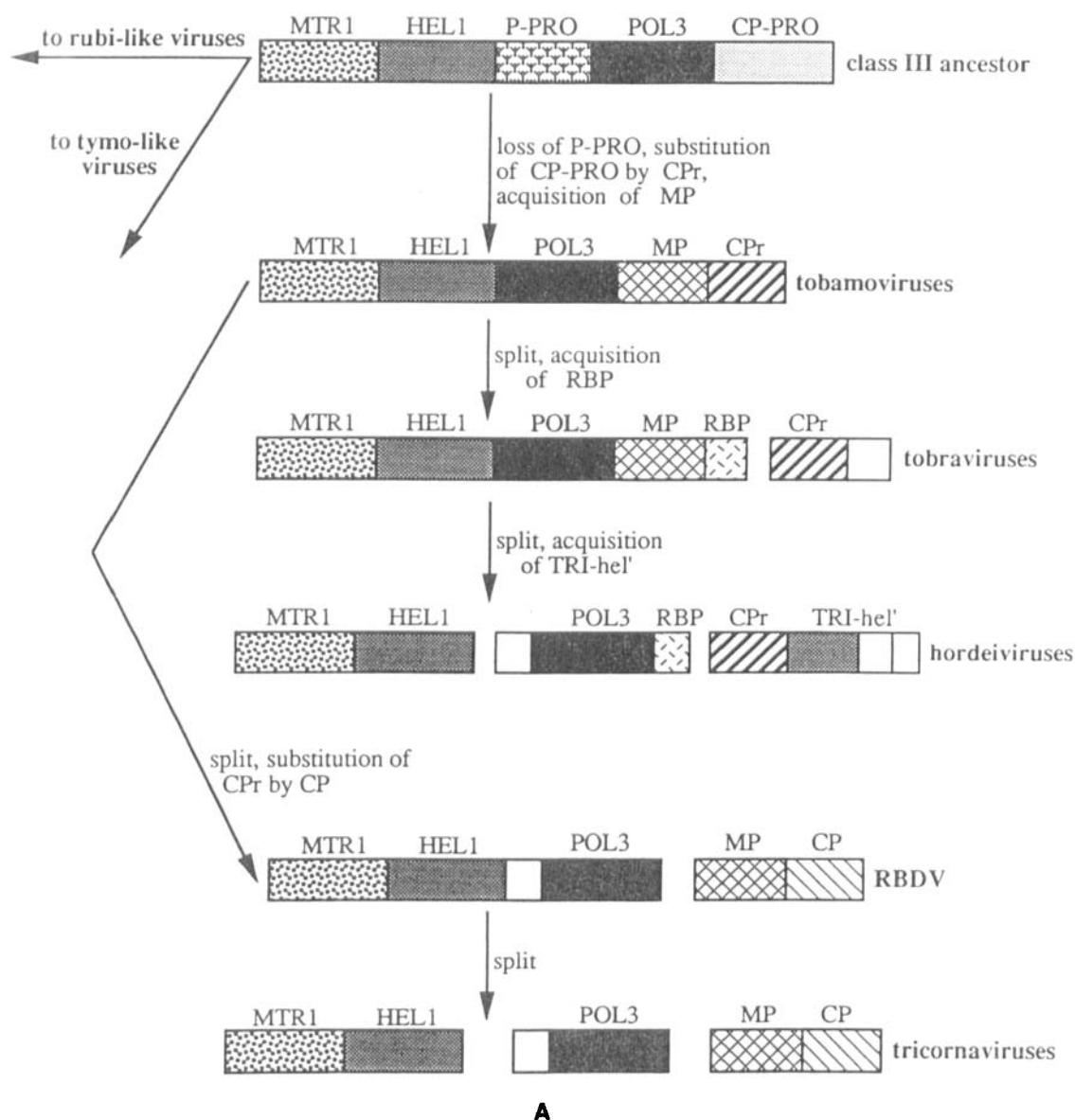
FIGURE 17. Tentative evolutionary scenario for class III of positive-strand RNA viruses. (a) The proposed order Tobamovirales; (b) the proposed order Rubivirales; (c) the proposed order Tymovirales. CP-PRO — capsid autoprotease; TRI-hel — triple gene block including the "accessory" helicase; RBP — (putative) RNA-binding protein. See also captions to Figures 13 and 15.

this is to combine the protease activity and the capsid protein function as it happens in the extant alphaviruses. It seems likely that if the ancestor virus had two or three genes (see below), the polyprotein cleavage by this primitive protease could occur exclusively *in cis,* which is compatible with the properties of the alphavirus capsid protein. Moreover, it has been shown that the cleavage at the C-terminus of the alphavirus capsid

protein is very tolerant to experimentally introduced amino acid substitutions, even affecting the catalytic residues (Hahn and Strauss, 1990). Such plasticity appears to be appropriate for an ancestral protease. We suspect, however, that the primitive protease might contain a catalytic cysteine in place of the catalytic serine found in the extant alphavirus capsid proteases. It has been argued that cysteine catalysis might have pre-
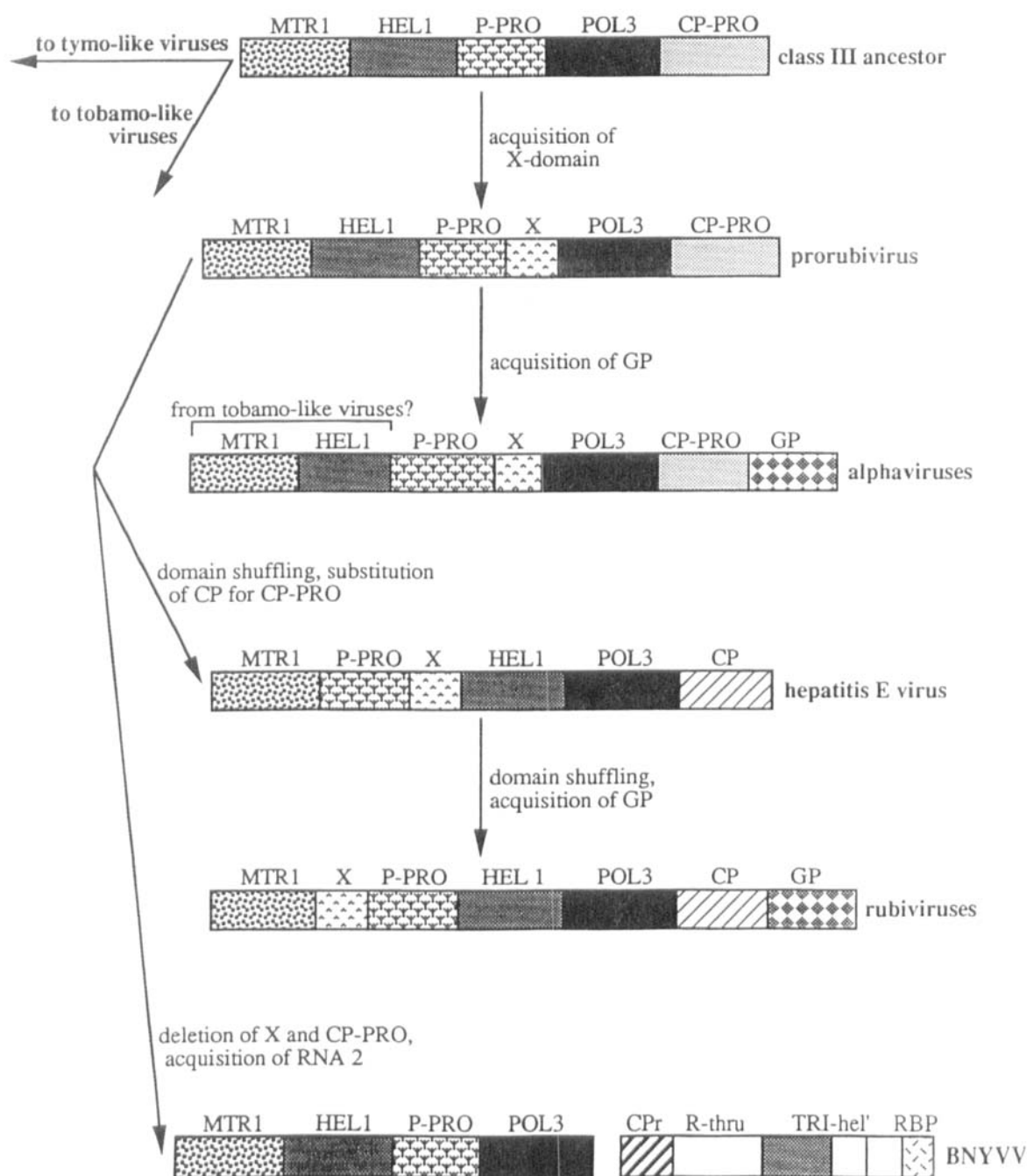
FIGURE 17B

ceded serine catalysis in enzyme evolution (Gorbalenya et al., 1986; Brenner, 1988). Cysteine for serine substitutions in the catalytic site of the Sindbis virus capsid protease resulted in a protein with considerable proteolytic activity (Hahn and Strauss, 1990).

The fact that, of all the extant viruses, the capsid autoprotease is limited to the alphaviruses may be considered an argument against the hypothesis that it is a relic of an ancient virus expression strategy. However, the specific similarity between the capsid protease and the nonstructural proteases of flaviviruses and pestiviruses (see above) is compatible with wide dissemination of an ancient domain. On the other hand, it is conceivable that the majority of the
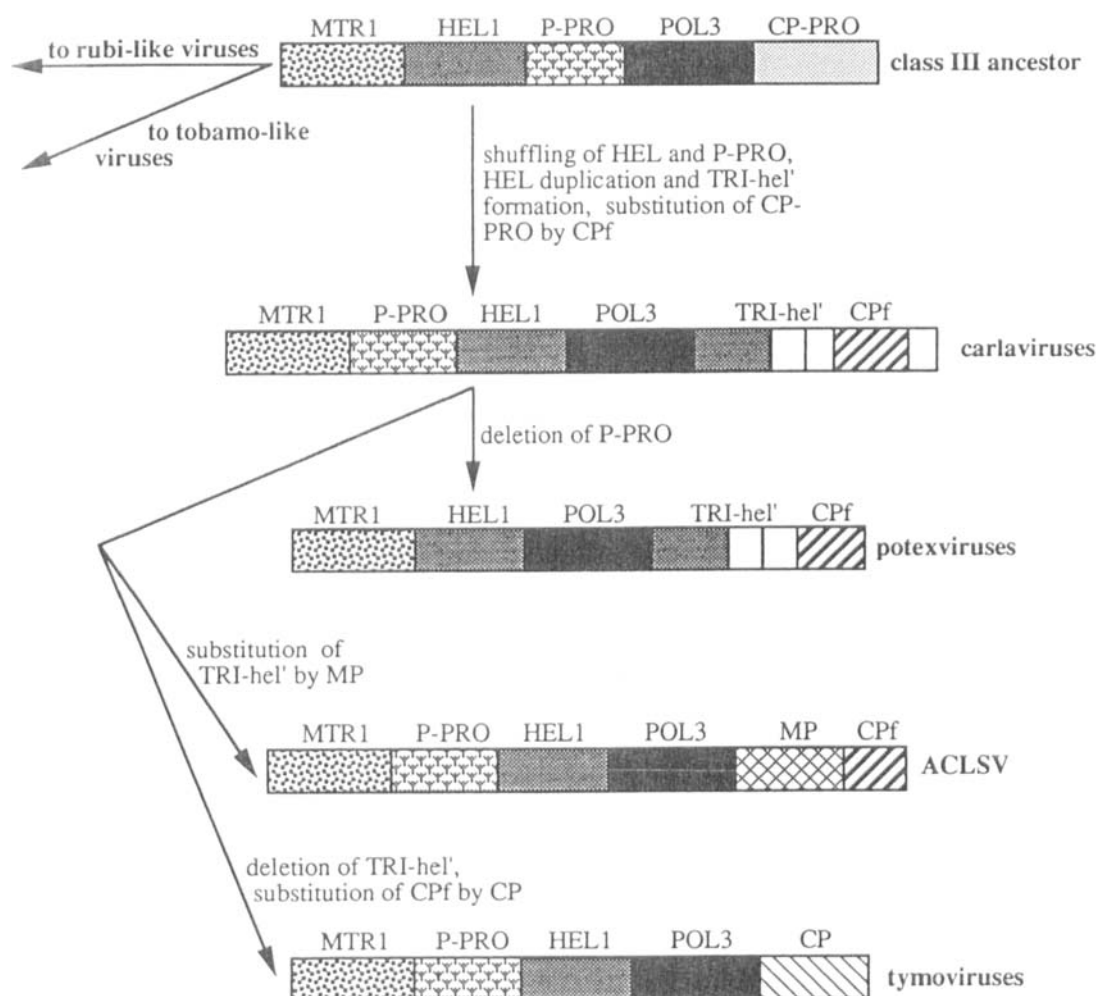
FIGURE 17C

viruses have abandoned the combination of the protease and the capsid functions in one protein for the virtue of more flexible expression strategies.

Perhaps the most dazzling question about the putative ancestor positive-strand RNA virus is whether it encoded an RNA helicase. As discussed above, the three classes of positive-strand RNA viruses encode helicases belonging to three different superfamilies. At least for the helicase superfamilies 1 and 2, multiple members have been identified in both eubacteria and eukaryotes, indicating that their common ancestor should have existed at a very early stage of evolution. As for the origin of RNA viral helicases, probably the most obvious idea is that the hypothetical ancestor of all positive-strand RNA viruses lacked a helicase gene and these

genes have been acquired independently at three different occasions on the evolutionary pathways leading to the three virus classes (Figure 18A). Although not contradicted directly by any available evidence, this scenario leaves us with difficult questions. Why have been the helicase genes captured by the ancestors of all three classes? The question is not trivial as despite the general correlation between the genome size and the presence of the helicase gene, luteoviruses lacking the helicase gene on the one hand and potexviruses and tobamoviruses encoding the helicase on the other hand all have genomes of very similar size (approximately 6 kb; Table 1). Furthermore, why is it that in all three classes the helicase gene is predominantly located upstream of the polymerase gene? The opposite examples, the arteri-related viruses and HyAV,
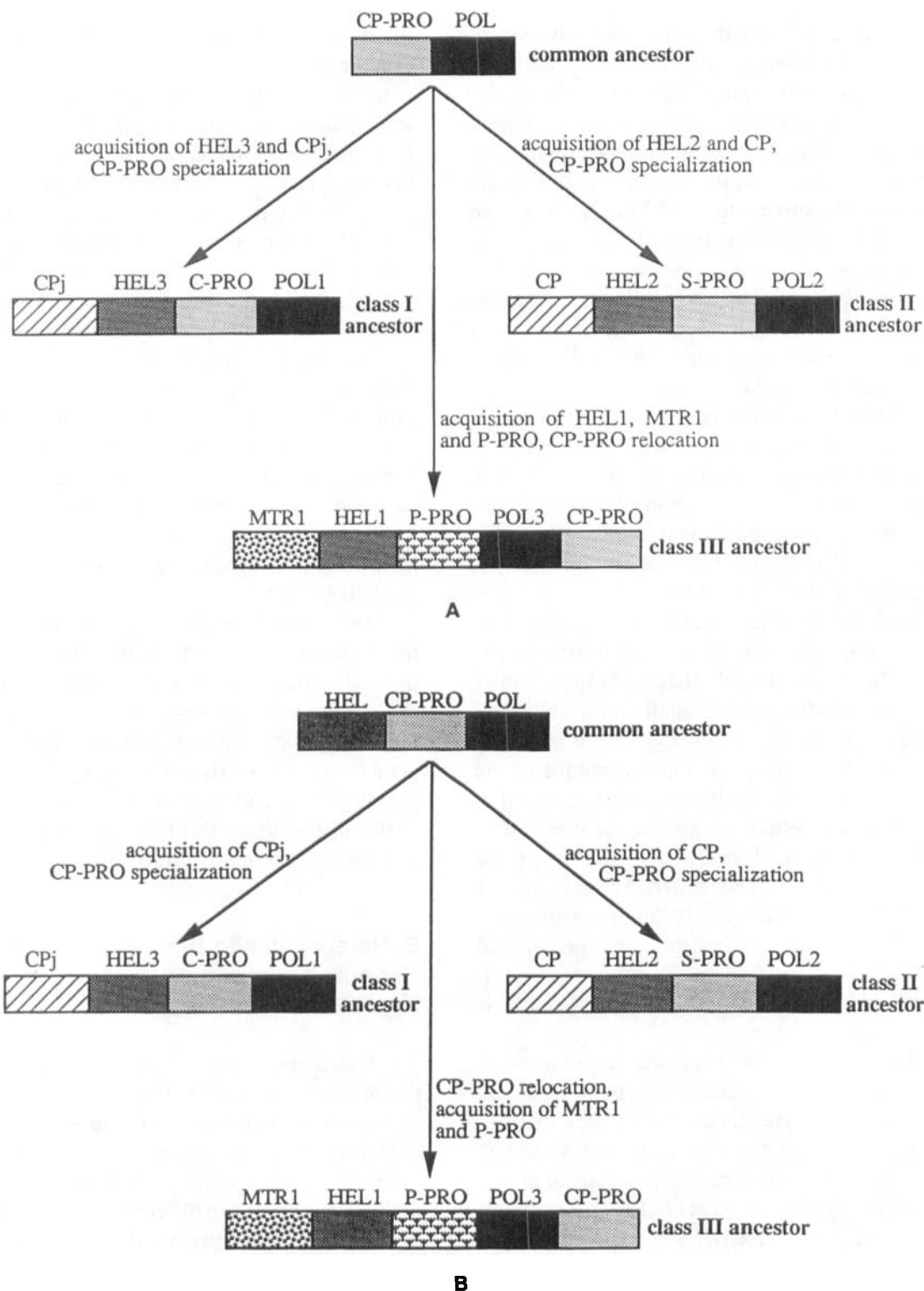
FIGURE 18. Two alternative scenarios for the evolution of the ancestors of the three virus classes from the hypothetical common ancestor virus. (a) "Gene capture" scenario. It is postulated that the ultimate ancestor contained two genes and the capsid autoprotease performed only one cleavage at its own C-terminus, like the contemporary alphavirus capsid protease. (b) "Primordial" scenario. The ultimate ancestor is postulated to contain three genes, with the primitive capsid autoprotease performing two cleavages at its N- and C-termini. For further details see text.

show that this is not the only configuration of these genes allowing virus RNA replication.

The alternative hypothesis is that the ancestral virus already had a helicase gene (Figure 18B). If so, what kind of helicase it might have encoded? One possibility is that it might belong to any of the three superfamilies and has been replaced by a different type of helicase during evolution of two of the three virus classes. We have already considered the possibility of similar recombinational replacements at later stages of evolution accounting for the evolution of the poty-related and arteri-related viruses.

Another possibility is that the divergence of the three helicase superfamilies has occurred in the course of evolution of positive-strand RNA viruses themselves. This hypothesis has profound implications. The separation of helicases into the three superfamilies is observed among cellular, DNA viral, and RNA viral enzymes. Clearly, it is a virtual impossibility that these superfamilies evolved convergently in different types of genomes. Neither is it likely that the helicase genes have been transferred horizontally from RNA viral to cellular genomes. We are forced to conclude that under this hypothesis the divergence of the helicase superfamilies should have occurred *before* viral and cellular genomes became distinguishable, that is, during the evolution of the primeval "RNA world" (Gilbert, 1986; Benner et al., 1987). If so, our hypothetical ancestor virus might have been closely related to the original RNA-based genetic systems, and the extant positive-strand RNA viruses may be considered their direct descent.

The difference between the hypotheses of multiple and single origins of the positive-strand RNA virus helicase genes is in effect that between positive selection and "frozen accident" explanations of evolution. The first type of explanation suggests that the observed genome organization, for example, the relative orientation of the RdRp and helicase genes, is intrinsically strongly advantageous to the virus and hence the chance for its independent emergence in different evolutionary lineages is high enough. The second interpretation concedes that this organization initially appeared purely accidentally in the ancestral genome, and later mechanisms have evolved to take

the full advantage of it, making drastic modifications relatively unlikely. Although most obvious when the evolution of the helicase gene is considered, this dilemma is relevant also for the evolution of other virus genes, most importantly that coding for the chymotrypsin-like protease. We are inclined to prefer the "frozen accident" explanation for the most widespread genes comprising the conserved "core" of the virus genome (helicase, chymotrypsin-like protease). The selectionist scheme allowing repeated independent acquisition of similar genes appears to be better applicable to such genes as that for the papain-like protease. Discrimination between the two types of evolutionary schemes is very difficult. Nevertheless, we believe that the possible relationship between positive-strand RNA viruses and primitive genetic systems implies that exploration of the former may shed some light on the features of the latter.

Aside from the problems with the origin of the helicase gene, the transition from the hypothetical common ancestor to the distinct ancestors of the three virus classes is easily imaginable, the main step being acquisition of the capsid protein gene (classes I and II) or the papain-like protease gene (class III), and the accompanying relegation of one of the functions of the hypothetical primitive capsid protease (Figure 18).

E. Parallelisms in the Evolution of Genome Organization, Expression, and Replication

A striking feature of the evolutionary histories of the three positive-strand RNA virus classes as viewed in our reconstruction is the parallel evolution of crucial features such as the mechanism for subgenomic RNA synthesis and multipartite genome in different lineages. An exciting question is just how far does this parallelism extend and whether we can predict new virus types (Table 3). It will be very interesting to find out whether, for example, viruses exist that belong to class III by phylogenetic criteria but have small genomes and lack the helicase gene as the result of the proposed truncation event. Conversely, it would be an important finding if vi-

TABLE 3

Features of genome organization and expression strategy in different divisions of positive-strand RNA viruses

	HEL	Chymotrypsin-like protease	Papain-like Protease	Icosahedral capsid	Elongated capsid	sg-mRNA[b]	Multipartite genome
Picornavirales	All	All	Aphthovirus	All	?[a]	Calici-viridae RTSV	Comoviridae
Potyvirales	All	Potyviridae	All	?	Potyviridae	?	Bymovirus
Sobemovirales	?	Sobemoviridae Luteoviridae	?	All	?	All	Nodaviridae
Spheridiplorna-virales	?	?	?	All	?	?	?
Arterivirales	All	All	All	Arteriviridae	Coronaviridae	All	?
Flavivirales	All	All	?	All	?	?	?
Pestivirales	All	All	?	All	?	?	?
Carmovirales	?	?	?	All	?	All	Dianthovirus
Rubivirales	All	Alphaviridae	All	Rubiviridae Hepeviridae Alphaviridae	Beneviridae	All	Beneviridae
Tobamovirales	All	?	?	Tricorna-iridae	Tobamoviridae Closteroviridae	All	Tricornaviridae Tobravirus Furovirus Hordeivirus
Tymovirales	All	?	Tymoviridae Carlaviridae	Tymoviridae	Potexviridae Carlaviridae	All	?

a   The question marks indicate that a feature has not been found in the known members of the given division but theoretically may be expected to be identified in new members.
b   sg-mRNA — subgenomic mRNA

ruses were discovered that belong to class II but have an elongated capsid related to those of class III viruses. If viruses of these types could not be found, it would be exciting to try and understand the constraints precluding their existence.

F. Positive-Strand RNA Viruses and their Hosts: Coevolution and/or Horizontal Transfer

Positive-strand RNA viruses infect eubacteria, plants, and animals; candidate viruses have also been isolated from *Fungi, Protozoa,* and *Algae,* and related fungal and protozoan dsRNA viruses have been studied in considerable detail (Francki et al., 1991; Koonin, 1992). Eubacterial viruses are represented by a single family of bacteriophages. It is surprising that RNA phages have not developed diversity of genome organization comparable to that observed in eukaryotic viruses. Recombinants between RNA phages could not be obtained and it cannot be ruled out that by some

not yet understood reasons certain types of recombination at the RNA level may be strongly disfavored in bacteria (Horiuchi, 1970). On the other hand, based on the similarity between the phage RdRps and the polymerases of eukaryotic viruses of superfamily 2, it has been speculated that RNA phages might have evolved from eukaryotic viruses by horizontal transfer (Koonin, 1991a). The closer relationship between the phage polymerases and those of yeast dsRNA genetic elements (Rodriguez-Cousino et al., 1991; Esteban et al., 1992; Koonin, 1992a; Figure 2) appears to be compatible with this hypothesis.

The separation between plant and animal viruses is observed in all three virus classes but only in two or three of the 12 proposed orders (Table 2). We do not know virus families including members infecting both plants and animals. Clearly, there are two major, not mutually exclusive explanations for the separation of plant and animal viruses, namely, horizontal transfer or divergence concomitant with that of the hosts. The former hypothesis appears to be the more popular

one, with insects invoked as possible vectors (Haseloff et al., 1984; Goldbach et al., 1991). The recent demonstration of experimental infection of a plant host by an insect virus is compatible with such a scheme (Selling et al., 1990).

On the other hand, it is clear that change of the host type does not occur easily. The points of divergence of plant and animal viruses correspond to deep branchings in the phylogenetic trees, and, although we cannot link them to any specific time scale, virus-host co-evolution seems to be a possibility. An attractive intermediate scenario would include two distinct acts of interkingdom virus transfer in the proposed orders *PICORNAVIRALES* and *RUBIVIRALES*, while the deepest branchings separating plant and animal viruses in each of the three classes would coincide with the plant/animal divergence. The latter two schemes, that is, complete or partial coevolution, imply that a considerable diversity of positive-strand RNA viruses and related dsRNA viruses should have been around already at the time of the plant/animal divergence about one billion years ago.

Positive-strand virus genome organization, primarily that of genes coding for virion components, and expression strategy show limited but obvious correlations with the type of the host. These include: (1) abundance of naked rod-shaped or filamentous virions among plant viruses, a virion type not found in animal viruses; (2) presence of genes for envelope glycoproteins in a subset of animal but not plant viruses; (3) wide spread of multipartite genomes among plant viruses as opposed to their rarity among animal viruses. These features might have evolved differently. It appears likely that the hypothetical gene for the elongated capsid protein has been acquired by plant viruses only once and then has been disseminated by recombination. As the envelope proteins of viruses of different divisions show no appreciable sequence similarity to each other, it seems to be impossible to distinguish between the possibilities of their recombinational transfer and independent acquisition. As discussed above, virus genome split yielding multipartite genomes obviously has occurred more than once in evolution.

IX. FUTURE DIRECTIONS: POSSIBLE TESTS FOR EVOLUTIONARY SCHEMES

It is difficult to propose direct tests for evolutionary scenarios. However, finding new types of virus genome organization compatible with these scenarios and perhaps resembling the postulated ancestor forms may provide indirect evidence. Some of the recent discoveries may be considered such evidence. For example, RTSV and PYFV as well as RBDV bear an obvious resemblance to the proposed ancestors of plant viruses with divided genomes (como/nepoviruses and tricornaviruses, respectively). Directed search for new viruses, particularly those infecting poorly studied organisms, using degenerate PCR primers derived from conserved amino acid sequence motifs, may be a powerful methodology for such analysis.

Very recently experiments have been reported that have direct bearing on the evolution of virus genome organization and expression strategy. In one study a two-component derivative of SNBV has been constructed encoding the nonstructural and the structural polyproteins on separate RNA segments and its efficient reproduction has been demonstrated (Geigenmuller-Gnirke et al., 1991). Another set of experiments included construction of a dicistronic poliovirus expressing the nonstructural proteins by internal translation initiation (Molla et al., 1992). These studies, however limited in scope, illustrate the capability of modern experimental techniques to model putative intermediates in virus evolution. There seems to be no fundamental obstacles in applying this approach to systematically testing the evolutionary scenarios discussed here.

ACKNOWLEDGMENTS

Gorbalenya. We wish to ask all the colleagues whose valuable work is not cited here because of space and/or scope limitations to accept our sincere apologies.

REFERENCES

Afanasiev, B. N., Rupasov, V. V., Fedchenko, V. I., Dolja, V. V., Atabekov, J. G., Chernov, B. K., Kozlov, Yu. V., and Bayev, A. A. 1986. Complete nucleotide sequence of barley stripe mosaic virus RNAs 3, 2b, and 4, and possible mechanisms of their interconversions, *Dokl. Akad. Nauk SSSR,* 290: 724–727.

Agranovsky, A. A., Boyko, V. P., Karasev, A. V., Koonin, E. V., and Dolja, V. V. 1991a. The putative 65K protein of beet yellows closterovirus is a homologue of HSP70 heat shock proteins, *J. Mol. Biol.,* 217: 603–610.

Agranovsky, A. A., Boyko, V. P., Karasev, A. V., Lunina, N. A., Koonin, E. V., and Dolja, V. V. 1991b. Nucleotide sequence of the 3'-terminal half of beet yellows closterovirus RNA genome: unique arrangement of eight virus genes, *J. Gen. Virol.,* 72: 15–23.

Agranovsky, A. A., Koonin, E. V., Boyko, V. P. et al. 1993. Beet yellows closterovirus: complete genome structure and identification of a leader papain-like thiol protease. *Virology,* in press.

Agrawal, D. K. and Johnson, J. E. 1992. Sequence and analysis of the capsid protein of Nudaurelia capensis w virus, an insect virus with T = 4 icosahedral symmetry, *Virology,* 190: 806–814.

Ahlquist, P., Dasgupta, R., and Kaesberg, P. 1984. Nucleotide sequence of the brome mosaic virus genome and its implications for viral replication, *J. Mol. Biol.,* 172: 369–383.

Ahlquist, P., Luckow, V., and Kaesberg, P. 1981. Complete nucleotide sequence of brome mosaic virus RNA 3, *J. Mol. Biol.,* 153: 23–38.

Ahlquist, P., Strauss, E. G., Rice, C. M., Strauss, J. H., Haseloff, J., and Zimmern, D. 1985. Sindbis virus proteins nsP1 and nsP2 contain homology to non-structural proteins from several RNA plant viruses, *J. Virol.,* 53: 536–542.

Allison, R. F., Janda, M., and Ahlquist, P. 1989. Sequence of cowpea chlorotic mottle virus RNAs 2 and 3 and evidence of a recombination event during bromovirus evolution, *Virology,* 172: 321–330.

Allison, R., Johnson, R. E., and Dougherty, W. G. 1986. The nucleotide sequence of the coding region of Tobacco etch virus genomic RNA: evidence for the synthesis of a single polyprotein, *Virology,* 154: 9–20.

Altschul, S. F., Gish, W., Miller, W, Myers, E. W, and Lipman, D. J. 1990. Basic local alignment search tool, *J. Mol. Biol.,* 215: 403–410.

Angenent, G. C., Linthorst, H. J., van Belkum, A. F., Cornelissen, B. J., and Bol, J. F. 1986. RNA 2 of tobacco rattle virus strain TCM encodes an unexpected gene, *Nucleic Acids Res.,* 14: 4673–4682.

Argos, P., Kamer, G., Nicklin, M. J., and Wimmer, E. 1984. Similarity in gene organization and homology between proteins of animal picornaviruses and a plant comovirus suggest common ancestry of these virus families, *Nucleic Acids Res.,* 12: 7251–7267.

Bae, Y.-S., Eun, H.-M., and Yoon, J.-W. 1989. Genomic differences between diabetogenic and nondiabetogenic Encephalomyocarditis virus, *Virology,* 170: 282–287.

Bancroft, J. B., Rouleau, M., Johnston, R., Prins, L., and Mackie, G. A., The complete nucleotide sequence of Foxtail mosaic virus RNA. 1991. *J. Gen. Virol.,* 72: 2173–2181.

Barker, R. F., Jarvis, N. P., Thompson, D. V., Loesch-Fries, L. S., and Hall, T. C. 1983. Complete nucleotide sequence of alfalfa mosaic virus RNA3, *Nucleic Acids Res.,* 11: 2881–2891.

Bazan, J. F. and Fletterick, R. J. 1988. Viral cysteine proteases are homologous to the trypsin-like family of serine protease: structural and functional implications, *Proc. Natl. Acad. Sci. U.S.A.,* 85: 7872–7876.

Bazan, J. F. and Fletterick, R. J. 1989. Detection of a trypsin-like serine protease domain in flaviviruses and pestiviruses, *Virology,* 171: 637–639.

Bazan, J. F. and Fletterick, R. J. 1989. Comparative analysis of viral cysteine protease structural models, *FEBS Lett.,* 249: 5–7.

Bazan, J. F. and Fletterick, R. J. 1990. Structural and catalytic models of trypsin-like viral proteases, *Semin. Virol.,* 1: 311–322.

Beck, D. L., Guilford, P. J., Voot, D. M., Andersen, M. T., and Forster, R. L. 1991. Triple gene block proteins of white clover mosaic potexvirus are required for transport, *Virology,* 183: 695–702.

Benner, S. A., Allemann, R. K., Ellington, A. D., Ge, L., Glasfeld, A., Leanz, G. F., Kruch, T., MacPherson, L. J., Moroney, S., Piccirilli, J. A., and Weinhold, E. 1987. Natural selection, protein engineering and the last riboorganism: rational model building in biochemistry, *Cold Spring Harb. Symp. Quant. Biol.,* 52: 53–63.

Bernal, J. J., Moriones, E., and Garcia-Arenal, F. 1991. Evolutionary relationships in the cucumoviruses: nucleotide sequence of tomato aspermy virus RNA 1, *J. Gen. Virol.,* 72: 2191–2195.

Blinov, V. M., Donchenko, A. P., and Gorbalenya, A. E. 1985. Internal homology of the poliovirus polyprotein primary structure: possible existence of two viral proteinases, *Dokl. Akad. Nauk SSSR,* 281: 984–987 (in Russian).

Blinov, V. M., Gorbalenya, A. E., and Donchenko, A. P. 1984. Sequence similarity between poliovirus cysteine protease P3-7c and cellular serine protease trypsin, *Dokl. Akad. Nauk SSSR,* 279: 502–505 (in Russian).

Boursnell, M. E., Brown, T. D., Foulds, I. J., Green, P. F., Tomley, F. M., and Binns, M. M. 1987. Completion of the sequence of the genome of the coronavirus avian infectious bronchitis virus, *J. Gen. Virol.,* 68: 57–77.

RIGHTSLINK

Bouzoubaa, S., Quillet, L., Guilley, H., Jonard, G., and Richards, K. 1987. Nucleotide sequence of Beet Necrotic Yellow Vein Virus RNA-1, *J. Gen. Virol.*, 68: 615–626.

Bouzoubaa, S., Ziegler, V., Beck, D., Guilley, H., Richards, K., and Jonard, G. 1986. Nucleotide sequence of Beet Necrotic Yellow Vein Virus RNA-2, *J. Gen. Virol.*, 67: 1689–1700.

Boyko, V. P., Karasev, A. V., Agranovsky, A. A., Koonin, E. V., and Dolja, V. V. 1992. Coat protein gene duplication in a filamentous RNA virus of plants, *Proc. Natl. Acad. Sci. U.S.A.*, 89: 9156–9160.

Bransom, K. L., Weiland, J. J., and Dreher, T. W. 1991. Proteolytic maturation of the 206-kDa nonstructural protein encoded by turnip yellow mosaic virus RNA, *Virology*, 184: 351–358.

Brault, V., Hibrand, L., Candresse, T., Le Gall, O., and Dunez, J. 1989. Nucleotide sequence and genetic organisation of Hungarian grapevine chrome mosaic nepovirus RNA2, *Nucleic Acids Res.*, 17: 7809–7819.

Brenner, S. 1988. The molecular evolution of genes and proteins: a tale of two serines, *Nature*, 334: 528–530.

Bruenn, J. 1991. Relationships among the positive strand and double-strand RNA viruses as viewed through their RNA-dependent RNA polymerases, *Nucleic Acids Res.*, 19: 217–226.

Bujarski, J. J. and Kaesberg, P. 1986. Genetic recombination between RNA components of a multipartite plant virus, *Nature*, 231: 528–531.

Cammisa-Parks, H., Cisar, L. A., Kane, A., and Stollar, V. 1992. The complete nucleotide sequence of cell fusing agent (CFA): homology between the nonstructural proteins encoded by CFA and the nonstructural proteins encoded by arthropod-borne flaviviruses, *Virology*, 189: 511–524.

Candresse, T., Morch, M. D., and Dunez, J. 1990. Multiple alignment and hierarchical clustering of conserved amino acid sequences in the replication-associated proteins of plant RNA viruses, *Res. Virol.*, 141: 315–329.

Carrington, J. C., Hillman, B. I., and Morris, T. J. 1989. The genome structure of turnip crinkle virus, *Virology*, 170: 219–226.

Carter, M. J., Milton, I. D., Meanger, J., Bennett, M. F., Gaskell, R. M., and Turner, P. C. 1992. The complete genome sequence of a feline calicivirus, *Virology*, 190: 443–448.

Castle, E., Leidner, U., Nowak, T., Wengler, G., and Wengler, G. 1986. Primary structure of the West Nile flavivirus genome region coding for all nonstructural proteins, *Virology*, 149: 10–26.

Chambers, T. J., Hahn, C. S., Galler, R., and Rice, C. M. 1990a. Flavivirus genome organization, expression, and replication, *Annu. Rev. Microbiol.*, 44: 649–688.

Chambers, T. J., Weir, R. C., Grakouri, A., McCourt, D. W., Bazan, J. F., Fletterick, R. J., and Rice, C. M. 1990b. Evidence that the N-terminal domain of nonstructural protein NS3 from yellow fever virus is a serine proteinase responsible for site-specific cleavages

in the viral polyprotein, *Proc. Natl. Acad. Sci. U.S.A.*, 87: 8898–8902.

Chelvanayagam, G., Heringa, J, and Argos, P. 1992. Anatomy and evolution of proteins displaying the viral capsid jellyroll topology, *J. Mol. Biol.*, 228: 220–241.

Choi, H.-K., Tong, L., Minor, W., Dumas, P., Boege, U., Rossmann, M. G., and Wengler, G. 1991. Structure of Sindbis virus core protein reveals a chymotrypsin-like serine proteinase and the organization of the virion, *Nature*, 354: 37–43.

Coia, G., Parker, M. D., Speight, G., Byrne, M. E., and Westaway, E. G. 1988. Nucleotide and complete amino acid sequences of Kunjin virus: definitive gene order and characteristics of the virus-specific proteins, *J. Gen. Virol.*, 69: 1–21.

Collett, M. S., Larson, R., Gold, C., Strick, D., Anderson, D. K., and Purchio, A. F. 1988. Molecular cloning and nucleotide sequence of the pestivirus bovine viral diarrhea virus, *Virology*, 165: 191–199.

Cornelissen, B. J., Brederode, F. T., Moormann, R. J., and Bol, J. F. 1983a. Complete nucleotide sequence of alfalfa mosaic virus RNA 1, *Nucleic Acids Res.*, 11: 1253–1265.

Cornelissen, B. J., Brederode, F. T., Veeneman, G. H., van Boom, J. H., and Bol, J. F. 1983b. Complete nucleotide sequence of alfalfa mosaic virus RNA 2, *Nucleic Acids Res.*, 11: 3019–3025.

Cornelissen, B. J., Janssen, H., and Bol, J. F. 1984. Complete nucleotide sequence of tobacco streak virus RNA 3, *Nucleic Acids Res.*, 12: 2427–2437.

Coutts, R. H., Rigden, J. E., Slabas, A. R., Lomonossoff, G. P., and Wise, P. J. 1991. The complete nucleotide sequence of tobacco necrosis virus strain D, *J. Gen. Virol.*, 72: 1521–1529.

Dasgupta, R., Ghosh, A., Dasmahapatra, B., Guarino, L. A., and Kaesberg, P. 1984. Primary and secondary structure of black beetle virus RNA2, the genomic messenger for BBV coat protein precursor, *Nucleic Acids Res.*, 12: 7215–7223.

Dasmahapatra, B., Dasgupta, R., Ghosh, A., and Kaesberg, P. 1985. Structure of the black beetle virus genome and its functional implications, *J. Mol. Biol.*, 182: 183–189.

De Jong, W. and Ahlquist, P. 1992. A hybrid plant RNA virus made by transferring the noncapsid movement protein from a rod-shaped to a icosahedral virus is competent for systemic infection, *Proc. Nat. Acad. Sci. U.S.A.*, 89: 6808–6812.

Delarue, M., Poch, O., Tordo, N., Moras, D., and Argos, P. 1990. An attempt to unify the structure of polymerases, *Protein Eng.*, 3: 461–467.

Demler, S. A. and de Zoeten, G. A. 1991. The nucleotide sequence and luteovirus-like nature of RNA 1 of an aphid non-transmissible strain of pea enation mosaic virus, *J. Gen. Virol.*, 72: 1819–1834.

Demler, S. A., Rucker, D. G., and de Zoeten, G. A. 1993. The chimeric nature of the genome of pea enation mosaic virus: the independent replication of RNA 2, *J. Gen. Virol.*, 74: 1–16.

den Boon, J. A., Snijder, E. J., Chirnside, E. D., de Vries, A. A. F., Horzinek, M. C., and Spaan, W. J. M. 1991. Equine arteritis virus is not a togavirus but belongs to the coronavirus-like superfamily, *J. Virol.*, 65: 2910–2920.

Ding, S.-W., Keese, P., and Gibbs, A. 1989. Nucleotide sequence of the Ononis yellow mosaic tymovirus genome, *Virology*, 172: 555–563.

Ding, S., Keese, P., and Gibbs, A. 1990. The nucleotide sequence of the genomic RNA of kennedya yellow mosaic tymovirus-Jervis Bay isolate: relationships with potex- and carlaviruses, *J. Gen. Virol.*, 71: 925–931.

Dolja, V. V., Boyko, V. P., Agranovsky, A. A., and Koonin, E. V. 1991. Phylogeny of capsid proteins of rod-shaped and filamentous plant RNA viruses: two families with distinct patterns of sequence and probably structure conservation, *Virology*, 184: 79–86.

Dolja, V. V. and Carrington, J. C. 1992. Evolution of positive-strand RNA viruses, *Semin. Virol.*, 3: 315–326.

Dolja, V. V. and Koonin, E. V. 1991. Phylogeny of capsid proteins of small icosahedral RNA plant viruses, *J. Gen. Virol.*, 72: 1481–1486.

Domier, L. L., Franklin, K. M., Shahabuddin, M., Hellmann, G. M., Overmeyer, J. H., Hiremath, S. T., Siaw, M. F., Lomonossoff, G. P., Shaw, J. G., and Rhoads, R. E. 1986. The nucleotide sequence of tobacco vein mottling virus RNA, *Nucleic Acids Res.*, 14: 5417–5430.

Dominguez, G., Wang, C.-Y., and Frey, T. K. 1990. Sequence of the rubella virus genome: evidence for genetic rearrangement during togavirus evolution, *Virology*, 177: 225–238.

Domingo, E., Martinez-Salas, E., Sobrino, F., de la Torre, J. C., Portela, A., Ortin, J., Lopez-Galindez, C., Perez-Brena, P., Villanueva, N., Najera, R., VandePol, S., Steinhauer, D., DePolo, N., and Holland, J. J. 1985. The quasispecies (extremely heterogeneous) nature of viral RNA genome populations: biological relevance — a review, *Gene*, 40: 1–8.

Doolittle, R. F., Of URFs and ORFs. A primer on how to analyze derived amino acid sequences, University Science Books, Mill Valley, CA, 1986.

Doolittle, R. F., Ed., Molecular evolution, *Meth. Enzymol.*, 183, 1990.

Dougherty, W. G. and Carrington, J. C. 1988. Expression and function of potyviral gene products, *Annu. Rev. Phytopathol.*, 26: 123–143.

Dzianott, A. M. and Bujarski, J. J. 1991. The nucleotide sequence and genome organization of the RNA-1 segment in two bromoviruses: broad bean bottle virus and cowpea chlorotic mottle virus, *Virology*, 185: 553–562.

Earle, J. A., Skuce, R. A., Fleming, C. S., Hoey, E. M., and Martin, S. J. 1988. The complete nucleotide sequence of a bovine enterovirus, *J. Gen. Virol.*, 69: 253–263.

Esteban, L., Rodriguez-Cousino, N., and Esteban, R. 1992. T double-stranded RNA (dsRNA) sequence re-veals that T and W dsRNAs form a new RNA family in *Saccharomyces cerevisiae*, *J. Biol. Chem.*, 267: 10874–10881.

Faragher, S. G., Meek, A. D., Rice, C. M., and Dalgarno, L. 1988. Genome sequences of a mouse-avirulent and a mouse-virulent strain of Ross River virus, *Virology*, 163: 509–526.

Felsenstein, J. 1989. PHYLIP — Phylogeny inference package (Version 3.2), *Cladistics*, 5: 164–166.

Fiers, W., Contreras, R., Duerinck, F., Haegeman, G., Iserentant, D., Merregaert, J., Jou, W. M., Molemans, F., Raeymaekers, A., Van den Berghe, A., Volckaert, G., and Ysebaert, M. 1976. Complete nucleotide sequence of bacteriophage ms2 ma: primary and secondary structure of the replicase gene, *Nature*, 260: 500–507.

Fitch, W. M. and Margoliash, E. 1967. Construction of phylogenetic trees, *Science*, 155: 279–284.

Forss, S., Strebel, K., Beck, E., and Schaller, H. 1984. Nucleotide sequence and genome organization of foot-and-mouth disease virus, *Nucleic Acids Res.*, 12: 6587–6601.

Forster, R. L., Bevan, M. W., Harbison, S.-A., and Gardner, R. C. 1988. The complete nucleotide sequence of the potexvirus white clover mosaic virus, *Nucleic Acids Res.*, 16: 291–303.

Francki, R. I. B., Fauquet, C. M., Knudson, D. L., and Brown, F., Eds., Classification and nomenclature of viruses. Fifth report of the International Committee on Taxonomy of Viruses, *Archives of Virologyy* (Suppl. 2), 1, 1991.

Franssen, H., Leunissen, J., Goldbach, R., Lomonosoff, G. P., and Zimmern, D. 1984. Homologous sequences in nonstructural proteins from cowpea mosaic virus and picornaviruses, *EMBO J.*, 3: 855–861.

Geigenmuller-Gnirke, U., Weiss, B., Wright, R., and Schlesinger, S. 1991. Complementation between Sindbis viral RNAs produces infectious particles with a bipartite genome, *Proc. Natl. Acad. Sci. U.S.A.*, 88: 3253–3257.

German, S., Candresse, T., Lanneau, M., Huet, J. C., Pernolet, J. C., and Dunez, J. 1990. Nucleotide sequence and genomic organization of apple chlorotic leaf spot virus, *Virology*, 178: 104–112.

Gibbs, A. 1987. A molecular evolution of viruses: "trees," "clocks," and "modules," *J. Cell. Sci.*, (Suppl.) 7: 319–337.

Gilbert, W. 1986. The RNA world, *Nature*, 319: 618.

Gilmer, D., Bouzoubaa, S., Guilley, H., Richards, K., and Jonard, G. 1992. Efficient cell-to-cell movement of beet necrotic yellow vein virus requires 3' proximal genes located on RNA 2, *Virology*, 189: 40–47.

Godeny, E. K., Chen, L., Kumar, S., Methven, S. L., Koonin, E. V., and Brinton, M. A. 1993. Complete genome sequence and phylogenetic analysis of the lactate dehydrogenase-elevating virus (LDV), *Virology*, 194: 585–596.

Goelet, P., Lomonossoff, G. P., Butler, P. J., Akam, M. E., Gait, M. J., and Karn, J. 1982. Nucleotide sequence

of tobacco mosaic virus RNA, *Proc. Natl. Acad. Sci. U.S.A.*, 79: 5818–5822.

Goldbach, R. 1986. Molecular evolution of plant RNA viruses, *Annu. Rev. Phytopathol.*, 24: 289–310.

Goldbach, R. 1987. Genome similarities between plant and animal RNA viruses, *Microbiol. Sci.*, 4: 197–202.

Goldbach, R., Le Gall, O., and Wellink, J. 1991. Alpha-like viruses in plants, *Semin. Virol.*, 2: 19–25.

Goldbach, R. and Wellink, J. 1988. Evolution of plus-strand RNA viruses, *Intervirology*, 29: 260–268.

Gorbalenya, A. E. 1992. Host-related sequences in RNA viral genomes, *Semin. Virol.*, 3: 359–371.

Gorbalenya, A. E., Blinov, V. M., and Koonin, E. V. 1985. Prediction of nucleotide-binding properties of virus proteins from their amino acid sequences, *Molek, Genetika*, No. 11: 30–36 (in Russian).

Gorbalenya, A. E., Blinov, V. M., and Donchenko, A. P. 1986. Poliovirus-encoded proteinase 3C: a possible evolutionary link between cellular serine and cysteine proteinase families, *FEBS Lett.*, 194: 253–257.

Gorbalenya, A. E. and Koonin, E. V. 1989. Virus proteins containing the purine nucleotide-binding proteins, *Nucleic Acids Res.*, 17: 8413–8440.

Gorbalenya, A. E. and Koonin, E. V. 1993a. Comparative analysis of the amino acid sequences of the key enzymes of the replication and expression of positive-strand RNA viruses. Validity of the approach and functional and evolutionary implications, *Sov. Sci. Rev. D. Physicochem. Biol.*, 11: 1–89.

Gorbalenya, A. E. and Koonin, E. V. 1993b. Helicases. Amino acid sequence comparisons and beyond. *Curr. Opin. Struct. Biol.*, in press.

Gorbalenya, A. E., Koonin, E. V., Donchenko, A. P., and Blinov, V. M. 1988a. A conserved NTP-motif in putative helicases, *Nature*, 333: 22.

Gorbalenya, A. E., Koonin, E. V., Donchenko, A. P., and Blinov, V. M. 1988b. A novel superfamily of nucleoside triphosphate-binding motif-containing proteins which are probably involved in duplex unwinding in DNA and RNA replication and recombination, *FEBS Lett.*, 239: 16–24.

Gorbalenya, A. E., Koonin, E. V., Donchenko, A. P., and Blinov, V. M. 1989a. Coronavirus genome: tentative functional mapping of the non-structural polyprotein by theoretical analysis of the amino acid sequence, *Nucleic Acids Res.*, 17: 4456–4469.

Gorbalenya, A. E., Koonin, E. V., Donchenko, A. P., and Blinov, V. M. 1989b. Two related superfamilies of putative helicases involved in replication, recombination, repair and expression of DNA and RNA genomes, *Nucleic Acids Res.*, 17: 4713–4730.

Gorbalenya, A. E., Donchenko, A. P., Koonin, E. V., and Blinov, V. M. 1989c. N-terminal domains of putative helicase of flavi- and pestiviruses may be serine proteases, *Nucleic Acids Res.*, 17: 3889–3897.

Gorbalenya, A. E., Donchenko, A. P., Blinov, V. M., and Koonin, E. V. 1989d. Cysteine proteases of positive strand RNA viruses and chymotrypsin-like serine pro-

teases. A distinct protein superfamily with a common structural fold, *FEBS Lett.*, 243: 103–113.

Gorbalenya, A. E., Koonin, E. V., and Wolf, Yu. I. 1990. A new superfamily of putative NTP-binding domains encoded by genomes of small DNA and RNA viruses, *FEBS Lett.*, 252: 145–148.

Gorbalenya, A. E., Koonin, E. V., and Lai, M. M.-C. 1991. Putative papain-related proteases of positive-strand RNA viruses, *FEBS Lett.*, 288: 201–205.

Gould, A. R. and Symons, R. H. 1991. Cucumber mosaic virus RNA 3: determination of the nucleotide sequence provides the amino acid sequences of protein 3a and viral coat protein, *Eur. J. Biochem.*, 126: 217–226.

Goulden, M. G., Lomonossoff, G. P., Davies, J. W., and Wood, K. R. 1990. The complete nucleotide sequence of PEBV RNA2 reveals the presence of a novel open reading frame and provides insights into the structure of tobraviral subgenomic promoters, *Nucleic Acids Res.*, 18: 4507–4512.

Greif, C., Hemmer, O., and Fritsch, C. 1988. Nucleotide sequence of tomato black ring virus RNA-1, *J. Gen. Virol.*, 69: 1517–1529.

Grieco, F., Burgyan, J., and Russo, M. 1989. The nucleotide sequence of Cymbidium ringspot virus RNA, *Nucleic Acids Res.*, 17: 6383.

Guilley, H., Carrington, J. C., Balazs, E., Jonard, G., Richards, K., and Morris, T. J. 1985. Nucleotide sequence and genome organization of carnation mottle virus RNA, *Nucleic Acids Res.*, 13: 6663–6677.

Gustafson, G. and Armour, S. L. 1986. The complete nucleotide sequence of RNA beta from the type strain of barley stripe mosaic virus, *Nucleic Acids Res.*, 14: 3895–3909.

Gustafson, G. D., Armour, S. L., Gamboa, G. C., Burgett, S. G., and Shepherd, J. W. 1989. Nucleotide sequence of barley stripe mosaic virus RNA-alpha: RNA-alpha encodes a single polypeptide with homology to proteins from other viruses, *Virology*, 170: 370–377.

Gustafson, G., Hunter, B., Hanau, R., Armour, S. L., and Jackson, A. O. 1987. Nucleotide sequence and genetic organization of barley stripe mosaic virus RNA-gamma, *Virology*, 158: 394–406.

Hahn, C. S., Lustig, S., Strauss, E. G., and Strauss, J. H. 1988. Western equine encephalitis virus is a recombinant virus, *Proc. Natl. Acad. Sci. U.S.A.*, 85: 5997–6001.

Hahn, C. S. and Strauss, J. H. 1990. Site-directed mutagenesis of the proposed catalytic amino acids of the Sindbis virus capsid protein autoprotease, *J. Virol.*, 64: 3069–3073.

Hahn, C. S., Strauss, E. G., and Strauss, J. H. 1985. Sequence analysis of three Sindbis virus mutants temperature-sensitive in the capsid protein autoprotease, *Proc. Natl. Acad. Sci. U.S.A.*, 82: 4648–4652.

Hamilton, W. D., Boccara, M., Robinson, D. J., and Baulcombe, D. C. 1987. The complete nucleotide sequence of Tobacco Rattle Virus RNA-1, *J. Gen. Virol.*, 68: 2563–2575.

Hardy, W. R. and Strauss, J. H. 1989. Processing of the nonstructural polyproteins of Sindbis virus: nonstructural proteinase is in the C-terminal half of nsP2 and functions both in *cis* and in *trans, J. Virol.,* 63: 4653–4664.

Haseloff, J., Goelet, P., Zimmern, D., Ahlquist, P., Dasgupta, R., and Kaesberg, P. 1984. Striking similarities in amino acid sequence among nonstructural proteins encoded by RNA viruses that have dissimilar genomic organization, *Proc. Natl. Acad. Sci. U.S.A.,* 81: 4358–4362.

Hearne, P. Q., Knorr, D. A., Hillman, B. I., and Morris, T. J. 1990. The complete genome structure and synthesis of infectious RNA from clones of tomato bushy stunt virus, *Virology,* 177: 141–151.

Higgins, D. G., Bleasby, A. J., and Fuchs, R. (1992) CLUSTAL V: improved software for multiple sequence alignment, *CABIOS,* 8: 189–191.

Hirashima, A., Hirose, T., Inayama, S., Inokuchi, Y., and Jacobson, A. B. 1988. Analysis of the complete nucleotide sequence of the group IV RNA coliphage SP, *Nucleic Acids Res.,* 16: 6205–6221.

Hodgman, T. C. 1988. A new superfamily of replicative proteins, *Nature,* 333: 22–23; 579 (Erratum).

Holland, J. J., Spindler, K., Horodyski, F., Grabau, E., Nichol, S., and Van de Pol, S. 1982. Rapid evolution of RNA genomes, *Science,* 215: 1577–1585.

Horiuchi, K. 1975. Genetic studies of RNA phages, in *RNA Phages,* Zinder, N. D., Ed., Cold Spring Harbor Laboratory, pp. 29–50.

Hyypia, T., Horsnell, C., Maarone, M., Khan, M., Kalkkinen, N., Auvinen, P., Kinnunen, L., and Stanway, G. 1992. A distinct picornavirus group identified by sequence analysis, *Proc. Natl. Acad. Sci. U.S.A.,* 89: 8847–8851.

Icho, T. and Wickner, R. B. 1989. The double-stranded RNA genome of yeast virus L-A encodes its own putative RNA polymerase by fusing two open reading frames, *J. Biol. Chem.,* 264: 6716–6723.

Inokuchi, Y., Takahashi, R., Hirose, T., Inayama, S., Jacobson, A. B., and Hirashima, A. 1986. The complete nucleotide sequence of the group II RNA coliphage GA, *J. Biochem.,* 99: 1169–1180.

Irie, K., Mohan, P. M., Sasaguri, Y., Putnak, R., and Padmanabhan, R. 1989. Sequence analysis of cloned dengue virus type 2 genome (New Guinea-C strain), *Gene,* 75: 197–211.

Jelkmann, W., Maiss, E., and Martin, R. R. 1992. The nucleotide sequence and genome organization of strawberry mild yellow edge-associated potexvirus, *J. Gen. Virol.,* 73: 475–479.

Jenkins, O., Booth, J. D., Minor, P. D., and Almond, J. W. 1987. The complete nucleotide sequence of coxsackievirus B4 and its comparison to other members of the Picornaviridae, *J. Gen. Virol.,* 68: 1835–1848.

Jiang, B., Monroe, S. S., Koonin, E. V. et al. 1993. RNA sequence of astrovirus: unique genome organization and a putative retrovirus-like ribosomal frameshifting

signal directing synthesis of the viral repliease. *Proc. Natl. Acad. Sci. U.S.A.,* 90, in press.

Johansen, E., Rasmussen, O. F., Heide, M., and Borkhardt, B. 1991. The complete nucleotide sequence of pea seed-borne mosaic virus RNA, *J. Gen. Virol.,* 72: 2625–2632.

Kamer, G. and Argos, P. 1984. Primary structural comparison of RNA-dependent polymerases from plant, animal and bacterial viruses, *Nucleic Acids Res.,* 12: 7269–7282.

Kanyuka, K., Vishnichenko, V., Levay, K., Kondrikov, D., Ryabov, E., and Zavriev, S. K. 1992. Nucleotide sequence of shallot virus X RNA reveals a 5'-proximal cistrons, *J. Gen. Virol.,* 73: 2553–2560.

Karasawa, A., Nakaho, K., Kakutani, T., Minobe, Y., and Ehara, Y. 1992. Nucleotide sequence analyses of peanut stunt cucumovirus RNAs 1 and 2, *J. Gen. Virol.,* 73: 701–707.

Karasawa, A., Nakaho, K., Kakutani, T., Minobe, Y., and Ehara, Y. 1991. Nucleotide sequence of RNA 3 of peanut stunt cucumovirus, *Virology,* 185: 464–467.

Kashiwazaki, S., Minobe, Y., Omura, T., and Hibino, H. 1990. Nucleotide sequence of barley yellow mosaic virus RNA 1: a close evolutionary relationship with potyviruses, *J. Gen. Virol.,* 71: 2781–2790.

Kashiwazaki, S., Minobe, Y., and Hibino, H. 1991. Nucleotide sequence of barley yellow mosaic virus RNA 2, *J. Gen. Virol.,* 72: 995–999.

Kato, N., Hijikata, M., Ootsuyama, Y., Nakagawa, M., Ohkoshi, S., Sugimura, T., and Shimotohno, K. 1990. Molecular cloning of the human hepatitis C virus genome from Japanese patients with non-A, non-B hepatitis, *Proc. Natl. Acad. Sci. U.S.A.,* 87: 9524–9528.

Keese, P. K. and Gibbs, A. 1992. Origins of genes: "big bang" or continuous creation?, *Proc. Natl. Acad. Sci. U.S.A.,* 89: 9489–9493.

Kimura, M., *The Neutral Theory of Molecular Evolution,* Cambridge University Press, 1983.

King, A. M., McCahon, D., Slade, W. R., Newman, J. W. 1982. Recombination in RNA, *Cell,* 29: 921–928.

King, L. A., Pullin, J. S., Stanway, G., Almond, J. W., and Moore, N. F. 1987. Cloning of the genome of cricket paralysis virus: Sequence of the 3' end, *Virus Res.,* 6: 331–344.

Kinney, R. M., Tsuchiya, K. R., Sneider, J. M., and Trent, D. W. 1992. Genetic evidence that epizootic Venezuelan equine encephalitis (VEE) viruses may have evolved from enzootic VEE subtype I-D virus, *Virology,* 191: 569–580.

Koonin, E. V. 1991a. The phylogeny of RNA-dependent RNA polymerases of positive-strand RNA viruses, *J. Gen. Virol.,* 72: 2197–2206.

Koonin, E. V. 1991b. Similarities in RNA helicases, *Nature,* 352: 290.

Koonin, E. V. 1992a. Evolution of double-stranded RNA viruses: a case for polyphyletic origin from different

groups of positive-stranded RNA viruses, *Semin. Virol.*, 3: 327–339.

Koonin, E. V. 1992b. A new group of putative helicases, *Trends Biochem. Sci.*, 17, 496–497.

Koonin, E. V. 1993. Computer-assisted identification of a putative methyltransferase domain in NS5 protein of flaviviruses and lambda2 protein of reovirus, *J. Gen. Virol.*, 73, 733–740.

Koonin, E. V., Mushegian, A. R., Ryabov, E. V., and Dolja, V. V. 1991a. Diverse groups of plant RNA and DNA viruses share related movement proteins that may possess chaperone-like activity, *J. Gen. Virol.*, 72: 2895–2903.

Koonin, E. V., Choi, G., Nuss, D. L., Shapira, R., and Carrington, J. C. 1991b. Common ancestry for a chestnut blight hypovirulence-associated dsRNA virus and a group of positive-strand RNA plant viruses: evidence for evolution by genome rearrangement, *Proc. Natl. Acad. Sci. U.S.A.*, 88: 10647–10651.

Koonin, E. V., Chumakov, K. M., and Gorbalenya, A. E. 1989. Tentative identification of the RNA-dependent RNA polymerases of dsRNA viruses, *FEBS Lett.*, 252: 42–46.

Koonin, E. V. and Gorbalenya, A. E. 1989. Evolution of RNA genomes: Does the high mutation rate necessitate high rate of evolution of viral proteins?, *J. Molec. Evol.*, 28: 524–527.

Koonin, E. V. and Gorbalenya, A. E. 1992. An insect picornavirus may have genome organization similar to that of caliciviruses, *FEBS Lett.*, 297: 81–86.

Koonin, E. V., Gorbalenya, A. E., Chumakov, K. M., Donchenko, A. P., and Blinov, V. M. 1987. Evolution of RNA-dependent RNA polymerases of positive-strand RNA viruses, *Molek. Genetika*, No. 7: 27–39. (in Russian).

Koonin, E. V., Gorbalenya, A. E., Purdy, M. A., Rozanov, M. N., Reyes, G. R., and Bradley, D. W. 1992. Computer-assisted assignment of functional domains in the non-structural polyprotein of hepatitis E virus: delineation of an additional group of positive-strand RNA plant and animal viruses, *Proc. Natl. Acad. Sci. U.S.A.*, 89: 8259–8263.

Kraev, A. S., Morozov, S. Y., Lukasheva, L. I., Rozanov, M. N., Chernov, B. K., Simonova, M. L., Golova, Y. B., Belzhelarskaya, S. N., Pozmogova, G. E., Skryabin, K. S., and Atabekov, I. G. 1988. Primary structure and organization of the genome of potato X-virus, *Dokl. Akad. Nauk. SSSR*, 300: 711–716.

Lai, M. M. C. 1990. Coronavirus: organization, replication, and expression of genome, *Annu. Rev. Microbiol.*, 44: 303–333.

Lai, M. M. C. 1992. RNA recombination in animal and plant viruses, *Microbiol. Rev.*, 56: 61–79.

Lain, S., Riechmann, J. L., Martin, M. T., and Garcia, J. A. 1989. Homologous potyvirus and flavivirus proteins belonging to a superfamily of helicase-like proteins, *Gene*, 82: 357–362.

Lain, S., Riechmann, J. L., and Garcia, J. A. 1990. RNA Helicase: a novel activity associated with a protein encoded by a positive strand RNA virus, *Nucleic Acids Res.*, 18: 7003–7006.

Lain, S., Martin, M. T., Riechmann, J. L., and Garcia, J. A. 1991. Novel catalytic activity associated with positive-strand RNA virus infection: nucleic acid-stimulated ATPase activity of the Plum Pox potyvirus helicase-like protein, *J. Virol.*, 65: 1–6.

Lambden, P. R., Caul, O., Ashley, C., and Clarke, I. N. 1993. Sequence and genome organization of a human Small Round Structured (Norwalk-like) virus, *Science*, 259: 516–519.

Lee, C.-J., Shieh, C.-K., Gorbalenya, A. E., Koonin, E. V., La Monica, N., Tuler, J., Bagdzhadzhyan, A., and Lai, M. M. C. 1991. The complete sequence (22 kb) of murine coronavirus gene 1 encoding the putative proteases and RNA polymerase, *Virology*, 180: 567–582.

Le Gall, O., Candresse, T., Brault, V., and Dunez, J. 1989. Nucleotide sequence of Hungarian grapevine chrome mosaic nepovirus RNA1, *Nucleic Acids Res.*, 17: 7795–7807.

Levinson, R. S., Strauss, J. H., and Strauss, E. G. 1990. Complete sequence of the genomic RNA of O'Nyong-nyong virus and its use in the construction of alphavirus phylogenetic trees, *Virology*, 175: 110–123.

Linthorst, H. J., Huisman, M. J., Asjes, C. J., and Bol, J. F. 1989. Nucleotide sequence of narcissus mosaic virus RNA, *J. Gen. Virol.*, 70: 267–276.

Lommel, S. A., Weston-Fina, M, Xiong, Z., and Lomonossoff, G. P. 1988. The nucleotide sequence and gene organization of red clover necrotic mosaic virus RNA-2, *Nucleic Acids Res.*, 16: 8587–8602.

Lomonossoff, G. P. and Shanks, M. 1983. The nucleotide sequence of cowpea mosaic virus B RNA, *EMBO J.*, 2: 2253–2258.

MacFarlane, S. A., Taylor, S. C., King, D. I., Hughes, G., and Davies, J. W. 1989. Pea early browning virus RNA1 encodes four polypeptides including a putative zinc finger protein, *Nucleic Acids Res.*, 17: 2245–2252.

Maiss, E., Timpe, U., Brisske, A., Jelkmann, W., Casper, R., Himmler, G., Mattanovich, D., and Katinger, H. W. 1989. The complete nucleotide sequence of plum pox virus RNA, *J. Gen. Virol.*, 70: 513–524.

Martin, R. R., Keese, P. K., Young, M. G., Waterhouse, P. M., and Gerlach, W. L. 1990. Evolution and molecular biology of luteoviruses, *Annu. Rev. Phytopathol.*, 28: 341–363.

Matthews, R. E. F. 1985. Virus taxonomy for the non-virologist, *Annu. Rev. Microbiol.*, 39: 451–474.

Meshi, T., Ohno, T., Iba, H., and Okada, Y. 1981. Nucleotide sequence of a cloned cDNA copy of TMV (cowpea strain) RNA, including the assembly origin, the coat protein cistron, and the 3' non-coding region, *Mol. Gen. Genet.*, 184: 20–25.

Meulenberg, J. J., Hulst, M. M., de Meijer, E. J., Moonen, P. L., den Besten, A., de Kluyver, E. P., Wensvoort, G., and Moorman, R. J. 1993. Lelystad Virus, the causative agent of porcine epidemic abortion and respi-

ratory syndrome (PEARS), is related to LDV and EAV, *Virology*, 192: 62–72.

Meyer, M., Hemmer, O., Mayo, M. A., and Fritsch, C. 1986. The nucleotide sequence of tomato black ring virus RNA-2, *J. Gen. Virol.*, 67: 1257–1271.

Meyers, G., Tautz, N., Dubovi, E. J., and Thiel, H.-J. 1991. Viral pathogenicity correlated with integration of ubiquitin-coding sequences, *Virology*, 180: 602–616.

Meyers, G., Ruemenapf, T., and Thiel, H.-J. 1989. Molecular cloning and nucleotide sequence of the genome of hog cholera virus, *Virology*, 171: 555–567.

Meyers, G., Wirblich, C., and Thiel, H.-J. 1991. Rabbit hemorrhagic disease virus — molecular cloning and nucleotide sequencing of a Calicivirus genome, *Virology*, 184: 664–676.

Mi, S., Durbin, R., Huang, H. V., Rice, C. M., and Stollar, V. 1989. Association of the Sindbis virus RNA methyltransferase activity with the nonstructural protein nsP1, *Virology*, 170: 385–391.

Mi, S. and Stollar, V. 1991. Expression of Sindbis virus nsP1 and methyltransferase activity in *Escherichia coli*, *Virology*, 178: 429–434.

Miller, W. A., Waterhouse, P. M., and Gerlach, W. L. 1988. Sequence and organization of barley yellow dwarf virus genomic RNA, *Nucleic Acids Res.*, 16: 6097–6111.

Mirzayan, C. and Wimmer, E. 1992a. Genetic and biochemical characterization of poliovirus polypeptide 2C, *Abstracts of the 3rd International Symposium on Positive Strand RNA Viruses*, Clearwater, FL, September 19–24, Abstract P4-28.

Mirzayan, C. and Wimmer, E. 1992b. Genetic analysis of an NTP-binding motif in poliovirus polypeptide 2C, *Virology*, 189: 547–555.

Molla, A., Jang, S. K., Paul, A. V., Reuer, Q., and Wimmer, E. 1992. Cardioviral internal ribosomal entry site is functional in a genetically engineered dicistronic poliovirus, *Nature*, 356: 255–257.

Morch, M. D., Boyer, J. C., and Haenni, A. L. 1988. Overlapping open reading frames revealed by complete nucleotide sequencing of turnip yellow mosaic virus genomic RNA, *Nucleic Acids Res.*, 16: 6157–6173.

Morch, M. D., Drugeon, G., Szafranski, P., and Haenni, A. L. 1989. Proteolytic origin of the 150-kilodalton protein encoded by turnip yellow mosaic virus genomic RNA, *J. Virol.*, 63: 5153–5158.

Moriones, E., Roossinck, M. J., and Garcia-Arenal, F. 1991. Nucleotide sequence of tomato aspermy virus RNA 2, *J. Gen. Virol.*, 72: 779–783.

Morozov, S. Yu., Dolja, V. V., and Atabekov, J. G. 1989. Probable reassortment of genomic elements among elongated RNA-containing plant viruses, *J. Mol. Evol.*, 29: 52–62.

Morozov, S. Yu. and Rupasov, V. V. 1985. On the possibility of a common origin of the genes encoding the RNA polymerases of bacterial, plant, and animal positive-strand RNA viruses, *Biol. Nauki*, No. 10: 19–24 (in Russian).

Mushegian, A. R. and Koonin, E. V., Movement proteins of plant viruses: sequence motifs, protein families, and evolutionary pathways, *Arch. Virol.*, in press.

Najarian, R., Caput, D., Gee, W., Potter, S. J., Renard, A., Merryweather, J., Van Nest, G., and Dina, D. 1985. Primary structure and gene organization of human hepatitis A virus, *Proc. Natl. Acad. Sci. U.S.A.*, 82: 2627–2631.

Natsuaki, T., Mayo, M. A., Jolly, C. A., and Murant, A. F. 1991. Nucleotide sequence of raspberry bushy dwarf virus RNA-2: a bicistronic component of a bipartite genome, *J. Gen. Virol.*, 72: 2183–2189.

Nitayaphan, S., Grant, J. A., Chang, G.-J. J., and Trent, D. W. 1990. Nucleotide sequence of the virulent SA-14 strain of Japanese encephalitis virus and its attenuated vaccine derivative, *Virology*, 177: 541–552.

Oh, C.-S. and Carrington, J. C. 1989. Identification of essential residues in potyvirus proteinase HC-Pro by site-directed mutagenesis, *Virology*, 173: 692–699.

Ohno, S., *Evolution by Gene Duplication*, Springer, 1970.

O'Reilly, D. R., Thomas, C. J., and Coutts, R. H. 1991. Tomato aspermy virus has an evolutionary relationship with other tripartite RNA plant viruses, *J. Gen. Virol.*, 72: 1–7.

Osorio-Keese, M. E., Keese, P., and Gibbs, A. 1989. Nucleotide sequence of the genome of eggplant mosaic tymovirus, *Virology*, 172: 547–554.

Palmenberg, A. C. 1989. Sequence alignments of picornaviral capsid proteins, in *Molecular Aspects of Picornavirus Infection and Detection*, Semler, B. L. and Ehrenfeld, E., Eds., p. 211–230.

Petty, I. T., French, R., Jones, R. W., and Jackson, A. O. 1990. Identification of barley stripe mosaic virus genes involved in viral RNA replication and systemic movement, *EMBO J.*, 9: 3453–3457.

Pevear, D. C., Borkowski, J., Calenoff, M., Oh, C. K., Ostrawski, B., and Lipton, H. L. 1988. Insights into Theiler's virus neurovirulence based on a genomic comparison of the neurovirulent GDVII and less virulent BeAn strains, *Virology*, 165: 1–12.

Pletnev, A. G., Yamschikov, V. F., and Blinov, V. M. 1989. Nucleotide sequence of the genome and complete amino acid sequence of the polyprotein of tick-borne encephalitis virus, *Virology*, 174: 250–263.

Poch, O., Sauvageut, I., Delarue, M., and Tordo, N. 1989. Identification of four conserved motifs among the RNA-dependent polymerase encoding elements, *EMBO J.*, 8: 3867–3874.

Preugshat, F., Yao, C.-W., and Strauss, J. H. 1990. In vitro processing of Dengue virus type 2 nonstructural proteins NS2A, NS2B, and NS3, *J. Virol.*, 64: 4364–4374.

Racaniello, V. R. and Baltimore, D. 1981. Molecular cloning of poliovirus cDNA and determination of the complete nucleotide sequence of the viral genome, *Proc. Natl. Acad. Sci. U.S.A.*, 78: 4887–4891.

Rao, A. L. N., Sullivan, B. R., and Hall, T. C. 1990. Use of *Chenopodium hybridum* facilitates isolation of

BMV RNA recombinants, *J. Gen. Virol.*, 71: 1403–1407.

Rezaian, M. A., Williams, R. H., Gordon, K. H., Gould, A. R., and Symons, R. H. 1984. Nucleotide sequence of cucumber-mosaic-virus RNA 2 reveals a translation product significantly homologous to corresponding proteins of other viruses, *Eur. J. Biochem.*, 143: 277–284.

Rezaian, M. A., Williams, R. H., and Symons, R. H. 1985. Nucleotide sequence of cucumber mosaic virus RNA 1: presence of a sequence complementary to part of the viral satellite RNA and homologies with other viral RNAs, *Eur. J. Biochem.*, 150: 331–339.

Rice, C. M., Lenches, E. M., Eddy, S. R., Shin, S. J., Sheets, R. L., and Strauss, J. H. 1985. Nucleotide sequence of yellow fever virus: implications for flavivirus gene expression and evolution, *Science*, 229: 726–733.

Ritzenthaler, C., Viry, M., Pinck, M., Fuchs, M., and Pinck, L. 1993. Complete nucleotide sequence and genetic organization of grapevine fanleaf nepovirus RNA1, *J. Gen. Virol.*, in press.

Riviere, C. J. and Rochon, D. M. 1990. Nucleotide sequence and genomic organization of melon necrotic spot virus, *J. Gen. Virol.*, 71: 1887–1896.

Robaglia, C., Durand-Tardif, M., Tronchet, M., Boudazin, G., Astier-Manifacier, S., and Casse-Delbart, F. 1989. Nucleotide sequence of potato virus Y (N strain) genomic RNA, *J. Gen. Virol.*, 70: 935–947.

Rochon, D. M. and Tremaine, J. H. 1989. Complete nucleotide sequence of the cucumber necrosis virus genome, *Virology*, 169: 251–259.

Rodriguez-Cousino, N., Esteban, L. M., and Esteban, R. 1991. Molecular cloning and characterization of W double-stranded RNA, a linear molecule present in *Saccharomyces cerevisiae*. Identification of its single-stranded RNA form as 20 S RNA, *J. Biol. Chem.*, 266: 12772–12778.

Romero, J., Dzianott, A. M., and Bujarski, J. J. 1992. The nucleotide sequence and genome organization of the RNA2 and RNA3 segments in broad bean mottle virus, *Virology*, 187: 671–681.

Rossmann, M. G. and Johnson, J. E. 1989. Icosahedral RNA virus structure, *Annu. Rev. Biochem.*, 58: 533–573.

Rozanov, M. N., Drugeon, G., Seron, K., and Haenni, A.-L. 1992a. Papain-related proteinase of turnip yellow mosaic virus, Abstract Book, *NATO Advanced Study Institute and EEC Course on the Regulation of Gene Expression by Animal Viruses*, Mallorca, Spain, May 30–June 8, Abstract S16.

Rozanov, M. N., Koonin, E. V., and Gorbalenya, A. E. 1992. Conservation of the putative methyltransferase domain: a hallmark of the "Sindbis-like" supergroup of positive-strand RNA viruses, *J. Gen. Virol.*, 73: 2129–2134.

Rozanov, M. N., Morozov, S. Yu., and Skryabin, K. G. 1990. Unexpected close relationship between the large non-virion proteins of filamentous potexviruses and spherical tymoviruses, *Virus Genes*, 3: 373–379.

Saitou, N. and Nei, M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees, *Mol. Biol. Evol.*, 4: 406–425.

Sankar, S. and Porter, A. G. 1992. Point mutations which drastically affect the polymerization activity of encephalomyocarditis virus RNA-dependent RNA polymerase correspond to the active site of *Escherichia coli* DNA polymerase. I, *J. Biol. Chem.*, 267: 10168–10176.

Sekiya, M. E., Lawrence, S. D., McCaffery, M., and Cline, K. 1991. Molecular cloning and nucleotide sequencing of the coat protein of citrus tristeza virus, *J. Gen. Virol.*, 72: 1013–1020.

Selling, B. H., Allison, R. F., and Kaesberg, P. 1990. Genomic RNA of an insect virus directs synthesis of infectious virions in plants, *Proc. Natl. Acad. Sci. U.S.A.*, 87: 434–438.

Serghini, M. A., Fuchs, M., Pinck, M., Reinbolt, J., Walter, B., and Pinck, L. 1990. RNA2 of Grapevine Fanleaf Virus: sequence analysis and coat protein cistron location, *J. Gen. Virol.*, 71: 1433–1441.

Shanks, M., Stanley, J., and Lomonossoff, G. P. 1986. The primary structure of Red clover mottle virus middle component RNA, *Virology*, 155: 697–706.

Shanks, M. and Lomonossoff, G. P. 1992. The nucleotide sequence of red clover mottle virus bottom component RNA, *J. Gen. Virol.*, 73: 2473–2477.

Shapira, R., Choi, G. H., and Nuss, D. L. 1991. Virus-like genetic organization and expression strategy for a double-stranded RNA genetic element associated with biological control of chestnut blight, *EMBO J.*, 10: 731–739.

Shen, P., Kaniewska, M., Smith, C., and Beachy, R. N. 1993. Nucleotide sequence and genomic organization of rice tungro spherical virus, *Virology*, 193: 621–630.

Shirako, Y. and Wilson, T. M. A. 1993. Complete nucleotide sequence and organization of the bipartite RNA genome of soil-borne wheat mosaic virus, *Virology*, 195: 16–32.

Simpson, G. G. 1944. *Tempo and Mode in Evolution*, Columbia University Press, New York.

Sit, T. L., Abouhaidar, M. G., and Holy, S. 1989. Nucleotide sequence of papaya mosaic virus RNA, *J. Gen. Virol.*, 70: 2325–2331.

Skern, T., Sommergruber, W., Blaas, D., Gruendler, P., Fraundorfer, F., Pieler, C., Fogy, I., and Kuechler, E. 1985. Human rhinovirus 2: complete sequence and proteolytic processing signals in the capsid protein region, *Nucleic Acids Res.*, 13: 2111–2126.

Sneath, P. and Sokal, R., *Principles of Numerical Taxonomy*, Freeman, San Francisco, 1973.

Snijder, E. J., den Boon, J. A., Bredenbeek, P. J., Horzinek, M. C., Rijnbrand, R., and Spaan, W. J. M. 1990. The carboxyl-terminal part of the

428

putative Berne virus polymerase is expressed by ribosomal frameshifting and contains sequence motifs which indicate that toro- and coronaviruses are evolutionarily related, *Nucleic Acids Res.*, 18: 4535–4542.

Sogin, M. L. 1991. Early evolution and the origin of eukaryotes, *Curr. Opin. Genet. Dev.*, 1: 457–463.

Stanway, G. 1990. Structure, function and evolution of picornaviruses, *J. Gen. Virol.*, 71: 2483–2501.

Stanway, G., Hughes, P. J., Mountford, R. C., Minor, P. D., and Almond, J. W. 1984. The complete nucleotide sequence of a common cold virus: human rhinovirus 14, *Nucleic Acids Res.*, 12: 7859–7875.

Steinhauer, D. and Holland, J. J. 1987. Rapid evolution of RNA viruses, *Annu. Rev. Biochem.*, 41: 409–433.

Strauss, J. H., Ed., Viral proteases, *Sem. Virol.*, 1, No. 1, 1990.

Strauss, E. G., Rice, C. M., and Strauss, J. H. 1984. Complete nucleotide sequence of the genomic RNA of Sindbis virus, *Virology*, 133: 92–110.

Strauss, J. H. and Strauss, E. G. 1988. Evolution of RNA viruses, *Annu. Rev. Microbiol.*, 42: 657–683.

Strauss, E. G., Strauss, J. H., and Levine, A. J. 1991. Virus evolution, in *Fundamental Virology*, Fields, B. N. and Knipe, D. M., Eds., Raven Press, New York, pp. 167–190.

Stuart, K. D., Weeks, R., Guilbride, L., and Myler, P. J. 1992. Molecular organization of Leishmania RNA virus 1, *Proc. Natl. Acad. Sci. U.S.A.*, 89: 8596–8590.

Takkinen, K. 1986. Complete nucleotide sequence of the nonstructural protein genes of Semliki Forest virus, *Nucleic Acids Res.*, 14: 5667–5682.

Tam, A. W., Smith, M. M., Guerra, M. E., Huang, C.-C., Bradley, D. W., Fry, K. E., and Reyes, G. R. 1991. Hepatitis E virus (HEV): molecular cloning and sequencing of the full-length viral genome, *Virology*, 185: 120–131.

Teterina, N. L., Kean, K. M., Gorbalenya, A. E., Agol, V. I., and Girard, M. 1992. Analysis of the functional significance of amino acid residues in the putative NTP-binding pattern of the poliovirus 2C protein, *J. Gen. Virol.*, 73: 1977–1986.

Tolskaya, E. A., Romanova, L. A., Kolesnikova, M. S., and Agol, V. I. 1983. Intertypic recombination in poliovirus: genetic and biochemical studies, *Virology*, 124: 121–132.

Tsarev, S. A., Emerson, S. U., Balayan, M. S., Ticehurst, J. R., and Purcell, R. H. 1991. Simian hepatitis a virus (HAV) strain AGM-27: comparison of genome structure and growth in cell culture with other HAV strains, *J. Gen. Virol.*, 72: 1677–1683.

Turnbull-Ross, A. D., Reavy, B., Mayo, M. A., and Murant, A. F. 1992. The nucleotide sequence of parsnip yellow fleck virus: a plant picorna-like virus, *J. Gen. Virol.*, 73: 3203–3211.

Valegard, K., Liljas, L., Fridborg, K., and Unge, T. 1990. The three-dimensional structure of the bacterial virus MS2, *Nature*, 345: 36–41.

Vance, V. B., Moore, D., Turpen, T. H., Bracker, A., and Hollowell, V. C. 1992. The complete nucleotide sequence of pepper mottle virus genomic RNA: comparison of the encoded polyprotein with those of other sequenced potyviruses, *Virology*, 191: 19–30.

Van der Wilk, F., Huisman, M. J., Cornelissen, B. J., Huttinga, H., and Goldbach, R. W. 1989. Nucleotide sequence and organization of potato leafroll virus genomic RNA, *FEBS Lett.*, 245: 51–56.

van Wezenbeck, P., Verver, J., Harmsen, J., Vos, P., and van Kammen, A. 1983. Primary structure and gene organization of the middle-component RNA of cowpea mosaic virus, *EMBO J.*, 2: 941–946.

Vartapetian, A. B. and Bogdanov, A. A. 1987. Proteins covalently linked to the viral genome, *Prog. Nucl. Acids Res. Molec. Biol.*, 34: 209–251.

Veidt, I., Lot, H., Leiser, R. M., Scheidecker, D., Guilley, H., Richards, K., and Jonard, G. 1988. Nucleotide sequence of beet western yellows virus RNA, *Nucleic Acids Res.*, 16: 9917–9932.

Verchot, J., Koonin, E. V., and Carrington, J. C. 1991. The 35-kDa protein from the N-terminus of the potyviral polyprotein functions as a third virus-encoded proteinase, *Virology*, 185: 527–535.

Verchot, J., Herndon, K. L., and Carrington, J. C. 1992. Mutational analysis of the tobacco etch potyviral 35-kDa proteinase: identification of essential residues and requirements for autoproteolysis, *Virology*, 190: 298–306.

Vincent, J. R., Lister, R. M., and Larkins, B. A. 1991. Nucleotide sequence analysis and genomic organization of the NY-RPV isolate of barley yellow dwarf virus, *J. Gen. Virol.*, 72: 2347–2355.

Volchkov, V. E., Volchkova, V. A., and Netesov, S. V. 1991. Complete nucleotide sequence of the genomic RNA of eastern equine encephalomyelitis virus, *Molek. Genetika*, No. 5: 8–15.

Von Heijne, G., *Sequence Analysis in Molecular Biology. Treasure Trove or Trivial Pursuit*, Academic Press, 1987.

Walker, J. E., Saraste, M., Runswick, M. J., and Gay, N. J. 1982. Distantly Related Sequences in the a- and b-Subunits of ATP synthase, myosin, kinases and other ATP-requiring enzymes and a common nucleotide binding fold, *EMBO J.*, 1: 945–951.

Warrener, P., Tamura, J., and Colett, M. S. 1993. RNA-stimulated NTPase activity associated with yellow fever virus NS3 protein expressed in bacteria, *J. Virol.*, 67: 989–996.

Wengler, G., Czaya, G., Farber, P. M., and Hegemann, J. H. 1991. In vitro synthesis of West Nile virus proteins indicates that the aminoterminal segment of the NS3 protein contains the active centre of the protease which cleaves the viral polyprotein after multiple basic amino acids, *J. Gen. Virol.*, 72: 851–858.

Wengler, G. and Wengler, G. 1991. The carboxy-terminal part of the NS3 protein of the West Nile flavivirus can

RIGHTSLINK

be isolated as a soluble protein after proteolytic cleavage and represents an RNA-stimulated NTPase, *Virology*, 184: 707–715.

Wiskerchen, M. and Collett, M. 1991. Pestivirus gene expression: protein p80 of bovine viral diarrhea virus is a proteinase involved in polyprotein processing, *Virology*, 184: 341–350.

Wu, S., Rinehart, C. A., and Kaesberg, P. 1987. Sequence and organization of Southern Bean mosaic virus genomic RNA, *Virology*, 161: 73–80.

Xiong, Y. and Eickbusch, T. H. 1990. Origin and evolution of retroelements based upon their reverse transcriptase sequences, *EMBO J.*, 9: 3353–3362.

Xiong, Z. and Lommel, S. A. 1989. The complete nucleotide sequence and genome organization of red clover necrotic mosaic virus RNA 1, *Virology*, 171: 543–554.

Yoshikawa, N., Sasaki, E., Kato, M., and Takahashi, T. 1992. The nucleotide sequence of apple stem grooving capillovirus genome, *Virology*, 191: 98–105.

Zavriev, S. K., Kanyuka, K. V., and Levay, K. E. 1991. The genome organization of potato virus M RNA, *J. Gen. Virol.*, 72: 9–14.

Ziegler, A., Natsuaki, T., Mayo, M. A., Jolly, C. A., and Murant, A. F. 1992. The nucleotide sequence of RNA-1 of raspberry bushy dwarf virus, *J. Gen. Virol.*, 73: 3213–3218.

Zimmern, D. 1988. Evolution of RNA viruses, in *RNA Genetics*, Holland, J. J., Domingo, E., and Ahlquist, P., Eds., CRC Press, Boca Raton, FL, pp. 211–240.