JOURNAL OF VIROLOGY, June 1991, p. 2910–2920
0022-538X/91/062910-11$02.00/0

Vol. 65, No. 6

Equine Arteritis Virus Is Not a Togavirus but Belongs to the Coronaviruslike Superfamily

JOHAN A. DEN BOON,[1,2] ERIC J. SNIJDER,[1]* EWAN D. CHIRNSIDE,[2]† ANTOINE A. F. DE VRIES,[2] MARIAN C. HORZINEK,[2] AND WILLY J. M. SPAAN[1]

Department of Virology, Institute of Medical Microbiology, Faculty of Medicine, Leiden University, Postbus 320, 2300 AH Leiden,[1] and Department of Virology, Institute of Infectious Diseases and Immunology, Veterinary Faculty, University of Utrecht, Utrecht,[2] The Netherlands

Received 13 November 1990/Accepted 20 February 1991

The nucleotide sequence of the genome of equine arteritis virus (EAV) was determined from a set of overlapping cDNA clones and was found to contain eight open reading frames (ORFs). ORFs 2 through 7 are expressed from six 3′-coterminal subgenomic mRNAs, which are transcribed from the 3′-terminal quarter of the viral genome. A number of these ORFs are predicted to encode structural EAV proteins. The organization and expression of the 3′ part of the EAV genome are remarkably similar to those of coronaviruses and toroviruses. The 5′-terminal three-quarters of the genome contain the putative EAV polymerase gene, which also shares a number of features with the corresponding gene of corona- and toroviruses. The gene contains two large ORFs, ORF1a and ORF1b, with an overlap region of 19 nucleotides. The presence of a "shifty" heptanucleotide sequence in this region and a downstream RNA pseudoknot structure indicate that ORF1b is probably expressed by ribosomal frameshifting. The frameshift-directing potential of the ORF1a/ORF1b overlap region was demonstrated by using a reporter gene. Moreover, the predicted ORF1b product was found to contain four domains which have been identified in the same relative positions in coronavirus and torovirus ORF1b products. The sequences of the EAV and coronavirus ORF1a proteins were found to be much more diverged. The EAV ORF1a product contains a putative trypsinlike serine protease motif. Our data indicate that EAV, presently considered a togavirus, is evolutionarily related to viruses from the coronaviruslike superfamily.

Equine arteritis virus (EAV) was first isolated from a fetus aborted during an endemic disease outbreak in pregnant mares (9). Serological evidence suggests that the virus is widespread in the horse population and only rarely causes disease. If it does, the clinical features are acute anorexia and fever, usually accompanied by palpebral edema, conjunctivitis, nasal catarrh, and edema of the legs, genitals, and abdomen. In infected pregnant mares, abortion is common (10, 15). A carrier state exists in seropositive stallions, in which EAV is produced in semen (47). These "shedding stallions" may consequently infect brood mares by a venereal route. Field isolates are rare and may be difficult to propagate in cell culture, and consequently the biology of EAV is poorly understood.

EAV has been classified as a member of the togavirus family (34, 53). The spherical enveloped EAV particle has a diameter of 50 to 70 nm (25). It consists of an icosahedral core structure of 35 nm (23) surrounded by an envelope carrying ringlike structures with a diameter of 12 to 15 nm (30). The viral genome is a single-stranded RNA of positive polarity with an estimated M_r of 4,000,000 (51). Four or five virion proteins have been described: a nucleocapsid (N) protein of 12 kDa, an unglycosylated 18-kDa envelope protein, a 21-kDa envelope protein, and one or two larger envelope proteins which, probably due to glycosylation, give rise to heterogeneous material of between 28 and 40 kDa (8a, 25, 54). During EAV replication, a 3′-coterminal nested set of seven virus-specific RNAs are produced, ranging in size from genome length (13 kb) to 0.8 kb (8, 48, 49). Viral subgenomic RNAs (sgRNAs) are composed of leader and body sequences which are not contiguous on the EAV genome. The sgRNAs may be produced by alternative splicing, since their leader sequence is derived from the extreme 5′ end of the EAV genome (8).

In order to study the genome organization, replication strategy, and polypeptide composition of EAV, we have determined its genomic sequence. Our data reveal an ancestral relationship between EAV and the coronaviruses and toroviruses, two virus groups for which we recently proposed a superfamily status (40). Interestingly, coronaviruses and toroviruses display similar genome sizes and virion architecture which are completely different from those of the arteriviruses.

MATERIALS AND METHODS

cDNA clones and nucleotide sequence analysis. Generation of EAV cDNA, mapping of clones, subcloning of restriction fragments in M13 vectors, and DNA sequence analysis were performed as described previously (8).

Computer analysis of sequence data. Nucleotide sequence data were assembled and analyzed with the computer software designed by Staden (43). Amino acid sequence similarity searches were carried out with the FASTA program (31) and the NBRF protein identification resource (release 22.0). Dot matrix comparisons, sequence alignments, and polypeptide structure analysis were carried out by using the COMPARE, GAP, and PEPTIDE STRUCTURE options from the software provided by the Computer Genetics Group/University of Wisconsin (version 5, 1989 [7]). Sequence alignments were optimized by visual inspection.

* Corresponding author.
† Present address: Equine Virology Unit, The Animal Health Trust, Lanwades Park, Kennett, Suffolk, England CB8 7PN.

Construction of pFSEAV and pFSEAVδ. Construct pFSEAV (see Fig. 4A) was similar to vector pBSFS, which had been developed to test ribosomal frameshifting during Berne virus polymerase gene expression (40). An intermediary construct from the pBSFS construction, named pBSMBB (40), was used to create pFSEAV. Vector pBSMBB contained a copy of the mouse hepatitis virus (MHV) A59 membrane (M) protein gene into which *Bgl*II and *Bam*HI linkers had been inserted. Following digestion of pBSMBB with *Bgl*II and *Bam*HI, pFSEAV was generated by cloning a 425-bp *Sau*3A-*Sau*3A fragment (nucleotide [nt] positions 5368 to 5793 in Fig. 2) from EAV clone 579 into pBSMBB. This 425-bp fragment contained the EAV polymerase open reading frame 1a (ORF1a)/ORF1b overlap region (see Fig. 4A).

Construct pFSEAVδ was obtained by deleting the region between the *Sma*I (nt position 5512 in Fig. 2) and *Pvu*II (nt position 5788 in Fig. 2) sites from the EAV ORF1b sequence in pFSEAV. A *Hin*dIII site, which is present downstream of the hybrid gene in the pBS multiple cloning region, was used for this purpose (see Fig. 4A). pFSEAV was digested with *Pvu*II and *Hin*dIII, and the restriction fragment which contained the EAV ORF1b/MHV M (C terminus) junction was purified. This fragment was recloned into *Sma*I- and *Hin*dIII-digested pFSEAV, resulting in a 276-bp deletion (nt positions 5515 and 5790 in Fig. 2), which leaves the EAV ORF1b reading frame intact.

The orientations of inserts and nucleotide sequences at the MHV M/EAV polymerase junctions were examined by sequence analysis. The hybrid genes in pFSEAV and pFSEAVδ were under the control of the pBS T7 promoter.

In vivo expression of pFSEAV and pFSEAVδ. pFSEAV and pFSEAVδ were expressed in vivo by transfection of HeLa cells. The HeLa cells had previously been infected with vaccinia virus recombinant vTF7-3, which contains the T7 polymerase gene under the control of a vaccinia virus promoter (11). Infections, transfections, and metabolic labeling were performed as described previously (40). Immunoprecipitation of expression products with antisera directed against the N and C termini of the MHV M protein were carried out as described before (40). In order to reduce aspecific precipitation, 0.2% sodium dodecyl sulfate (SDS) (instead of 0.1%) was used in immunoprecipitations with the N-terminal antiserum.

Nucleotide sequence accession number. The sequence reported here has been assigned EMBL accession number X53459.

RESULTS AND DISCUSSION

Nucleotide sequence analysis of the EAV genome. The synthesis of a genomic EAV cDNA library and the mapping of clones which cover all but the 18 nucleotides at the 5′ end of the EAV genome have been described recently (8). By similar methods, additional cDNA clones were obtained which allowed sequence analysis of the EAV genome on at least two independent cDNA clones (Fig. 1). However, nt 19 to 355 were present only in cDNA clone 586. This clone contained the EAV genomic leader sequence, which is adjoined to the 5′ end of all EAV RNA species. Partial sequence analysis of clone 586 has been reported previously (8).

The consensus nucleotide sequence of the EAV genome (12.7 kb) and the deduced amino acid sequences of EAV proteins are presented in Fig. 2. The primer extension experiments which were used to determine the length of the

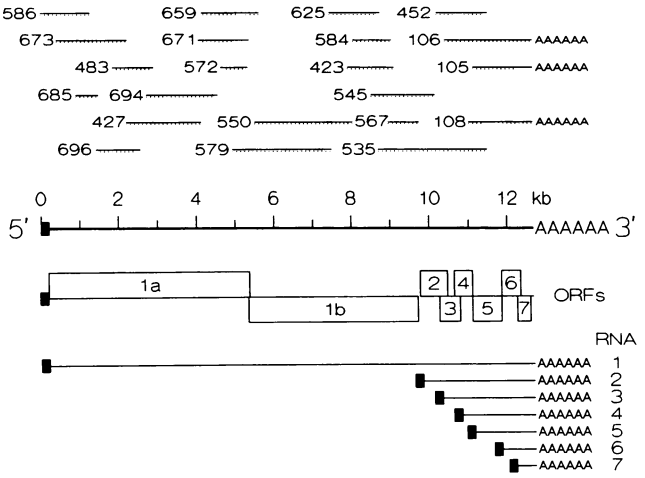


FIG. 1. Organization of the EAV genome. The cDNA clones which were used in the sequence analysis of the EAV genome are indicated in the upper part of the figure. Dotted lines indicate the parts which were sequenced. The lower part of the figure shows the positions of the ORFs (based on the sequence) and the positions of the sgRNAs (based on RNA lengths estimated from the gel; see text). The solid box indicates the EAV leader sequence.

EAV leader sequence (8) were repeated with an oligonucleotide which is complementary to positions 81 to 100 in the sequence presented in Fig. 2 (data not shown). The EAV leader sequence was calculated to be 207 nt long (1 nt shorter than reported previously [8]), indicating that the sequence in Fig. 2 lacks only the 18 most-5′-terminal nucleotides of the EAV genome. Experiments to obtain this sequence are in progress.

EAV genome structure and expression. Eight open reading frames (ORFs) were identified in the EAV genome (Fig. 1 and 2). We assume that ORF1a and ORF1b, which take up about 75% (9.7 kb) of the genome, are translated from the genomic RNA and encode the EAV polymerase (POL) or replicase (see below). The other six ORFs, which are partially overlapping, are thought to be expressed from six subgenomic mRNAs generated in EAV-infected cells. The bodies of these 3′-coterminal mRNAs are homologous to the 3′ end of the genome (8). In addition, they contain a common leader sequence at their 5′ end (8) (Fig. 1). We have reestimated the lengths of the EAV sgRNAs (reported earlier by van Berlo et al. [48]) on the basis of the agarose gels presented by de Vries et al. (8). Sizes of 3.2, 2.7, 2.2, 1.9, 1.2, and 0.8 kb [including a poly(A) tail] were now calculated for EAV mRNAs 2, 3, 4, 5, 6, and 7, respectively. Some characteristics of and our proposed nomenclature for the EAV RNAs and ORFs are listed in Table 1.

The sequence 5′-UCAAC-3′ has been identified as the site where the EAV leader sequence is joined to the body of sgRNAs 6 and 7 (8). As a result, ORFs 6 and 7 are located in the "unique" 5′-terminal regions of these RNAs. The EAV genome was found to contain 18 UCAAC motifs (indicated in Fig. 2), and all ORFs (except ORF1b) are preceded by one or more of these motifs. The sizes of the bodies of EAV sgRNAs 2 through 5 closely correspond to the positions of UCAAC junction motifs and ORFs 2 through 5 on the genome (Fig. 1 and 2).

Many of the EAV ORFs are preceded by multiple UCAAC motifs, but only one of these potential junction sites seems to be used in sgRNA generation. For example, ORF7

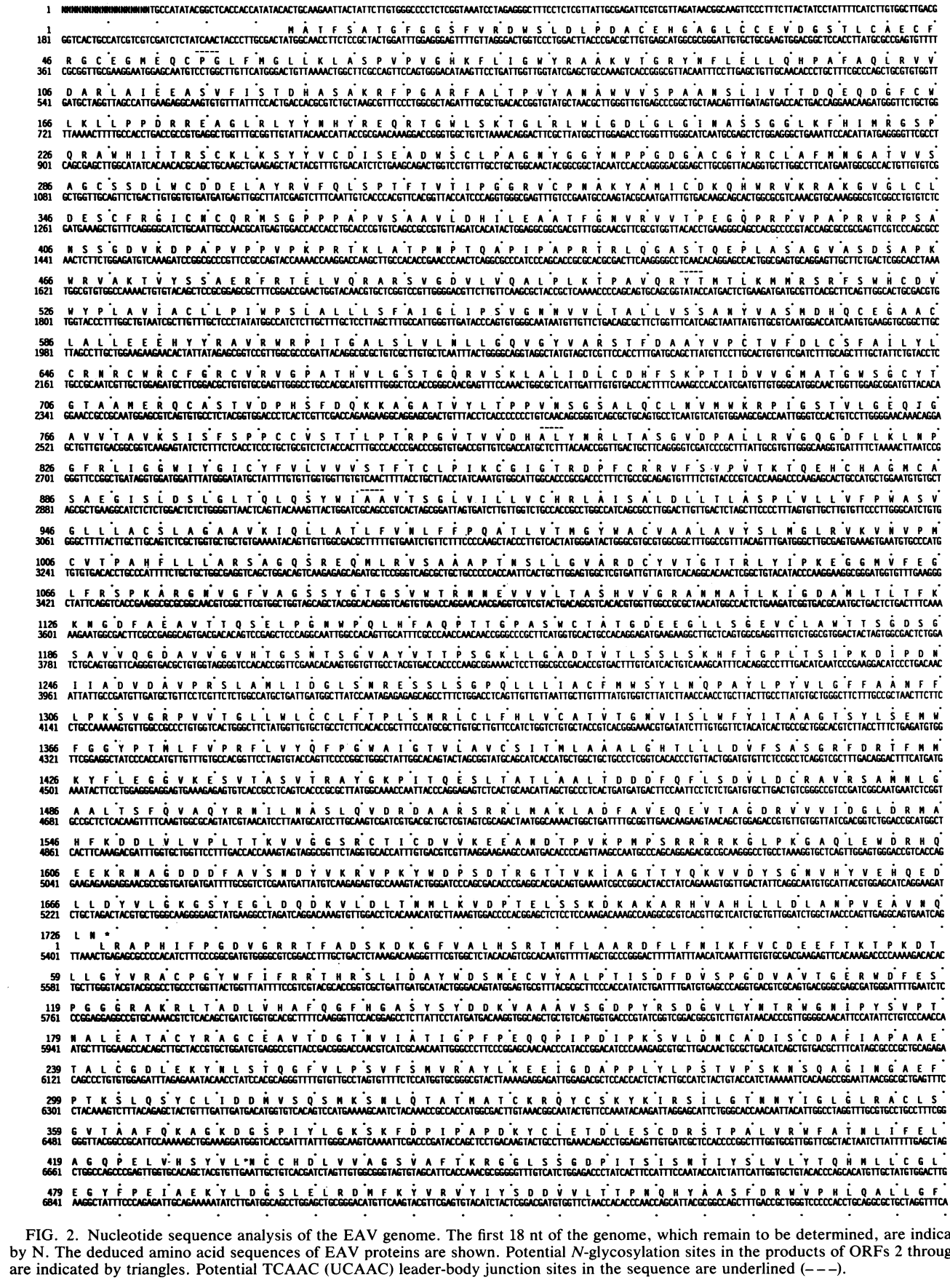


FIG. 2. Nucleotide sequence analysis of the EAV genome. The first 18 nt of the genome, which remain to be determined, are indicated by N. The deduced amino acid sequences of EAV proteins are shown. Potential *N*-glycosylation sites in the products of ORFs 2 through 5 are indicated by triangles. Potential TCAAC (UCAAC) leader-body junction sites in the sequence are underlined (– – –).

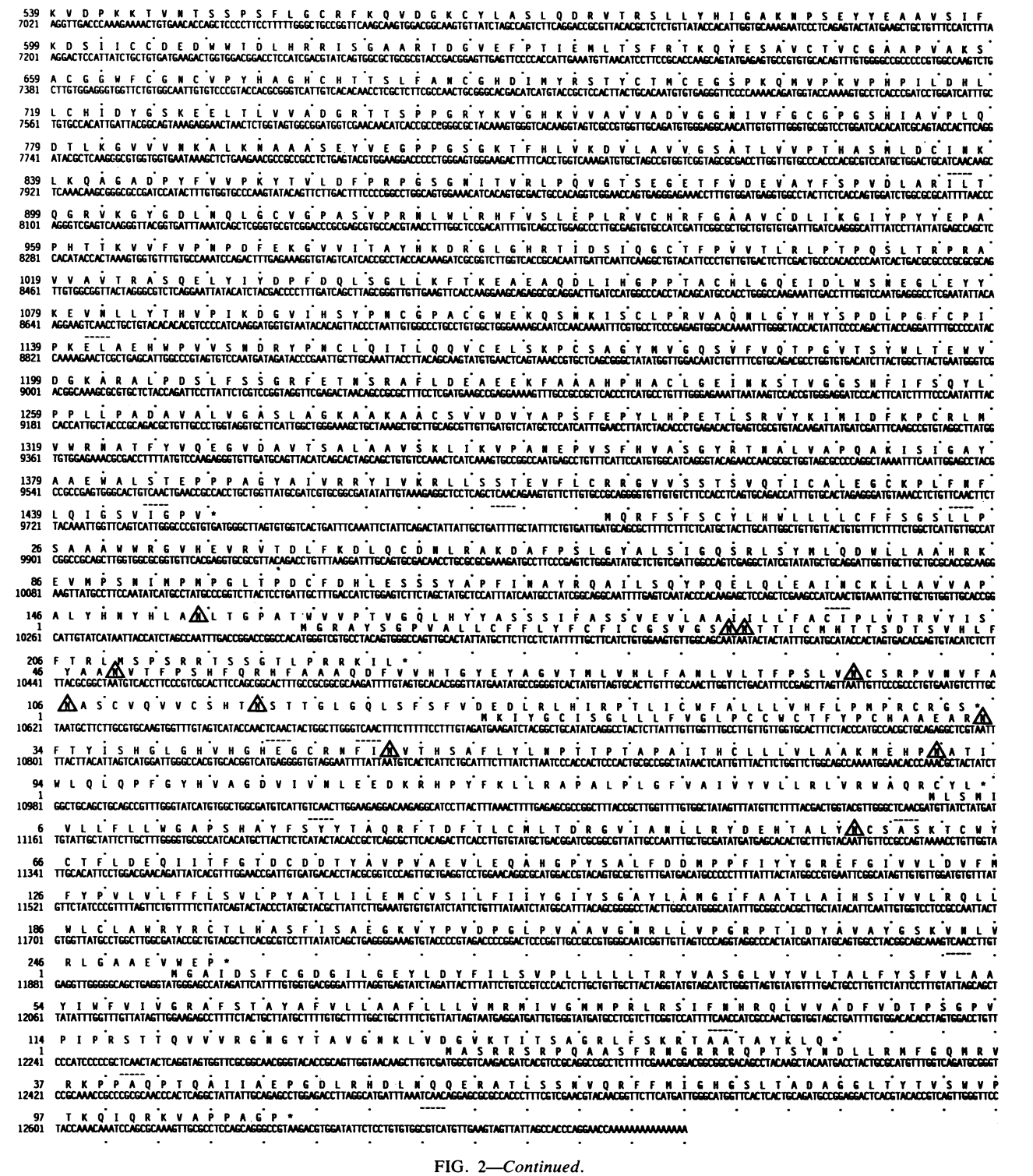


FIG. 2—*Continued.*

is preceded by two UCAAC motifs, but no heterogeneity in the size of RNA 7 has been detected (8). In addition, the presence of UCAAC motifs at positions which do not correspond to an sgRNA confirms the view that it is not only the primary RNA structure at this position which determines the functionality of a UCAAC motif in leader-body joining.

On the basis of a remarkable similarity with sequences involved in the splicing of *Tetrahymena* pre-rRNA (8), alternative splicing has been proposed as a possible mechanism for the generation of EAV sgRNAs. However, in view of the similarities between the EAV and coronavirus polymerases described in this article (see below), a role for the

TABLE 1. Characteristics of EAV RNAs and ORFs

RNA	Estimated size (kb)	ORF no.	Nucleotides (first–last)	No. of amino acids encoded	Size of product (kDa)
1	13	1a	226–5,406	1,727	186.9
		1b	5,406–9,749	1,448	159.0
2	3.2	2	9,825–10,505	227	25.6
3	2.7	3	10,307–10,795	163	18.0
4	2.2	4	10,701–11,156	152	17.2
5	1.9	5	11,147–11,911	255	28.7
6	1.2	6	11,902–12,387	162	17.7
7	0.8	7	12,314–12,643	110	12.3

EAV leader sequence and the conserved pregenic UCAAC motifs in a coronaviruslike transcription mechanism would not be surprising. The presence of negative-stranded sgRNAs in infected cells, which has recently been described for coronaviruses (35, 36), would be an important indication for a coronaviruslike transcription mechanism in EAV. Experiments to study the negative-stranded RNAs in EAV-infected cells are now in progress.

Analysis of ORFs encoded by sgRNAs. EAV ORFs 2, 3, and 4 encode polypeptides with hydrophobic N and C termini and predicted sizes of 25.6, 18.0, and 17.2 kDa, respectively (Table 1). ORF5 encodes a 28.7-kDa product with a hydrophobic N terminus and an internal hydrophobic domain. The predicted products of ORFs 2 through 5 contain 1, 6, 3, and 1 potential *N*-glycosylation sites, respectively (Fig. 2). In view of these characteristics, each of these ORFs may encode an envelope protein in the range between 28 and 40 kDa. ORF6 encodes a polypeptide of 17.7 kDa which is predicted to be an unglycosylated triple-spanning membrane protein. The deduced ORF7 product is 110 amino acids long (12.3 kDa). This is in good agreement with the size of the product obtained after in vitro translation of the corresponding sgRNA and the identification of this protein as the EAV N protein (8a, 24, 50).

Amino acid sequence comparisons between the deduced products of EAV ORFs 2 through 7 and the sequences currently available in the NBRF Protein Identification Resource did not reveal any significant similarities with other viral or cellular proteins. The results of a further study of the EAV structural proteins will be presented elsewhere (8a).

EAV polymerase gene resembles the polymerase gene of coronaviruses and toroviruses. By definition, the genomic RNA of a positive-stranded RNA virus encodes the viral polymerase or replicase. ORFs of 5.2 kb (ORF1a) and 4.2 kb (ORF1b) were identified in the region of the EAV genome which is not transcribed into sgRNAs. These ORFs overlap over a distance of 19 nt, ORF1b being in the −1 reading frame with respect to ORF1a. To exclude possible gel-reading errors, the ORF1a/ORF1b overlap sequence was determined from three independent cDNA clones.

The organization of the EAV POL gene was found to be remarkably similar to that of the POL genes of two other evolutionarily related groups of viruses which produce nested sets of subgenomic mRNAs, the coronaviruses (3, 4) and the toroviruses (40). In these two virus groups, the POL gene also consists of two ORFs, which overlap over a short distance. A subgenomic mRNA from which ORF1b could be expressed is lacking in cells infected with coronaviruses (for a review, see reference 42), toroviruses (38, 39), and arteriviruses (8, 48).

Ribosomal frameshifting, resembling that in retroviruses

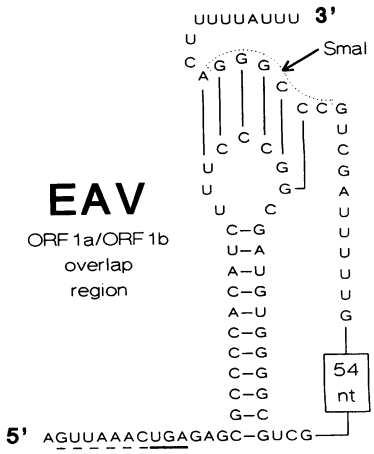


FIG. 3. Predicted secondary and tertiary RNA (pseudoknot) structure of the EAV ORF1a/ORF1b overlap region. The slippery sequence G UUA AAC is indicated by a dashed line. The ORF1a termination codon is underlined. The *Sma*I site in the downstream pseudoknot domain, which was digested in the construction of pFSEAVδ, is indicated by a dotted line.

(26), has been found to be the ORF1b expression mechanism for the coronavirus infectious bronchitis virus (IBV [5]). A "slippery" heptanucleotide sequence (5′-UUUAAAC-3′) and a downstream RNA pseudoknot (32) were shown to be essential for efficient frameshifting in the ORF1a/ORF1b overlap region (6). Sequence analysis of the POL genes of the mouse hepatitis coronavirus (MHV-A59 [4]) and the Berne torovirus (BEV [40]) revealed that this expression mechanism and the essential sequence elements in the ORF1a/ORF1b overlap region have been conserved during evolution of these virus groups.

The frameshift-directing RNA pseudoknot, consisting of a stem-loop structure base-paired to a downstream sequence element, could also be folded at the 3′ side of the EAV ORF1a/ORF1b overlap region (Fig. 3). A potential slippery sequence (5′-GUUAAAC-3′) was identified just upstream of the stem of the EAV hairpin and was positioned at the 5′ side of the termination codon for translation of ORF1a. These findings implied that ribosomal frameshifting was also the probable mechanism used to express EAV ORF1b. In addition, they were the first indications of an ancestral relationship between coronavirus, torovirus, and arterivirus polymerases.

EAV ORF1a/ORF1b overlap region directs ribosomal frameshifting. To study the frameshift-directing potential of the EAV ORF1a/ORF1b overlap region, we used the expression of a reporter gene similar to that used in the study of translational frameshifting in MHV-A59 (4) and BEV (40). The EAV ORF1a/ORF1b overlap region from clone 579 was cloned into a copy of the MHV-A59 26-kDa membrane protein (M) gene positioned downstream of a T7 promoter. The resulting construct, pFSEAV, consisted of the 5′ portion of the MHV M ORF fused in-frame to the EAV ORF1a/ORF1b overlap sequence, which in turn was fused in frame to the 3′ portion of the MHV M ORF (Fig. 4A). Termination of translation at the ORF1a termination codon should produce a 24-kDa fusion protein (Fig. 4A). Ribosomal frameshifting would lead to the synthesis of an additional fusion protein of 41 kDa. The two predicted products could be identified by immunoprecipitation with antibodies di-

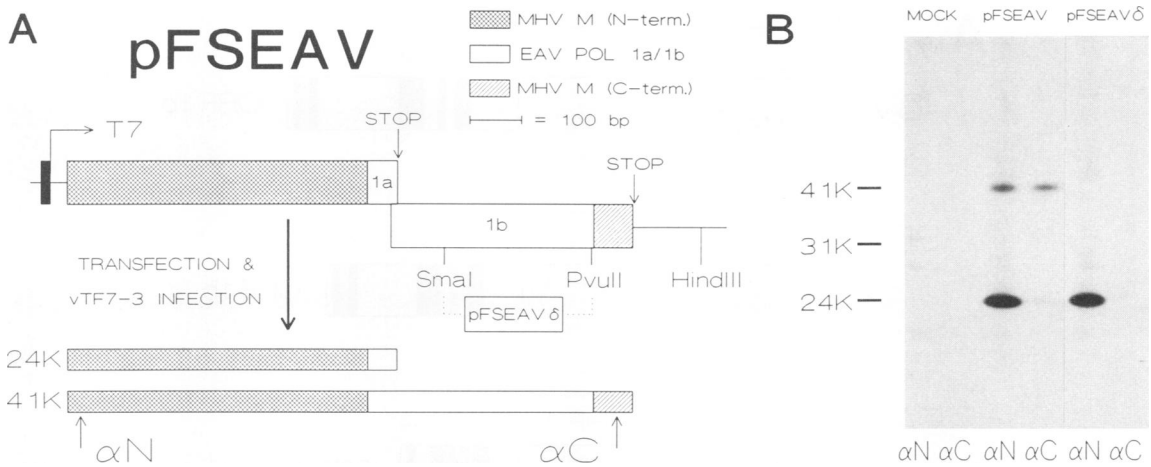


FIG. 4. Analysis of EAV ribosomal frameshifting in vivo with constructs pFSEAV and pFSEAVδ. (A) Schematic representation of the pFSEAV construct. The solid box represents the T7 RNA polymerase promoter in pBS. Open boxes indicate EAV sequences; N- and C-terminal (term.) MHV M sequences are represented by differently hatched boxes. The composition of the two predicted translation products is shown. The *Sma*I, *Pvu*II, and *Hin*dIII restriction sites which were used in the generation of pFSEAVδ are indicated. (B) SDS electrophoresis of pFSEAV (center) and pFSEAVδ (right) translation products immunoprecipitated by the N-terminal (αN) and C-terminal (αC) MHV M antisera. The two lanes on the left show the result of the same immunoprecipitations with an untransfected control lysate from vaccinia virus vTF7-3-infected cells. Sizes are shown in kilodaltons.

rected against the N- and C-terminal parts of the MHV M protein (Fig. 4A).

Ribosomal frameshifting was studied in vivo by transfection of pFSEAV into HeLa cells which had previously been infected with vaccinia virus recombinant vTF7-3. Expression of pFSEAV led to the synthesis of proteins of 24 and 41 kDa. Both products were precipitated by the antiserum directed against the N terminus of the MHV M protein (Fig. 4B). Only the 41-kDa product was precipitated when the C-terminal antiserum was used (Fig. 4B). The amount of incorporated label in both protein bands was determined by scintillation counting, resulting in an estimated frameshifting efficiency of between 15 and 20% (the 24- and 41-kDa products contained 8 and 10 methionine residues, respectively).

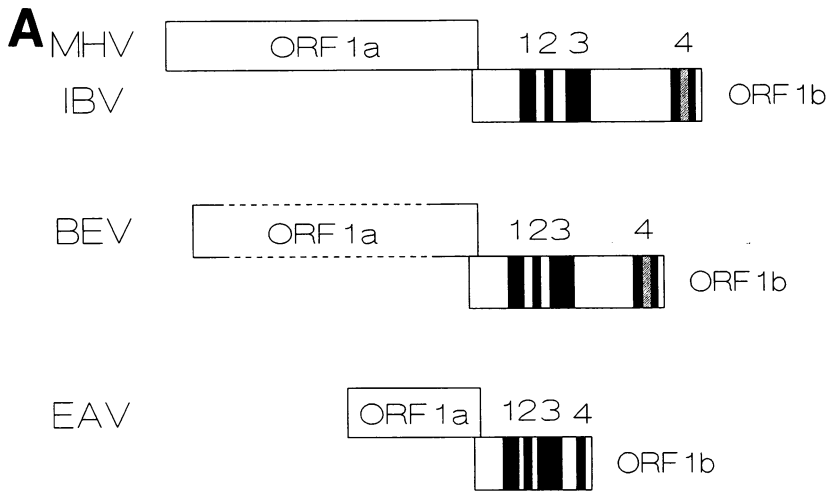
The EAV slippery sequence contains a nucleotide substitution compared with the corresponding sequence of coronaviruses and toroviruses (GUUAAAC instead of UUUA AAC). EAV therefore breaks the rule that slippery sequences consist of runs of three A, U, or G residues followed by the tetranucleotide UUUA, UUUU, or AAAC (26, 46). Although this may seem somewhat conflicting with the simultaneous slippage model proposed by Jacks et al. (26), there are indications that the mRNA-tRNA base-pairing requirements at the peptidyl (P) site of a frameshifting ribosome are more relaxed than those at the aminoacyl (A) site (6, 26). In fact, this nucleotide substitution may explain the lower frameshifting efficiency (in a reporter gene) of EAV than of coronaviruses and toroviruses.

The presence of a unique *Sma*I restriction site within the putative downstream sequence element of the EAV RNA pseudoknot allowed the deletion of this domain from pFSEAV (Fig. 4). Construct pFSEAVδ contained a 276-bp in-frame deletion (between the *Sma*I and *Pvu*II sites depicted in Fig. 4A) in the EAV ORF1b sequence and should produce a frameshift product of about 31 kDa which should also contain the C-terminal MHV M protein sequence. However, expression of pFSEAVδ yielded only the 24-kDa product which resulted from "normal" termination of ORF1a translation (Fig. 4B). Even after prolonged exposure of the autoradiographs, a 31-kDa polypeptide could not be detected. This result proved that a domain more than 71 nt downstream of the stem-loop structure is essential for efficient frameshifting. In view of the data on retrovirus, coronavirus, and torovirus frameshifting, this domain is likely to be the sequence 5′-CCGGGA-3′ at 70 to 75 nt downstream of the hairpin; the last four nucleotides of this sequence were deleted by the digestion of the *Sma*I site during the construction of pFSEAVδ (Fig. 3).

The second loop (L2, according to the terminology of Pleij et al. [32]) of the pseudoknot, which connects the hairpin with the downstream sequence element, appears to be considerably larger in EAV than in other frameshift-employing viruses. On the basis of the results obtained with pFSEAVδ, the length of L2 in the EAV pseudoknot is assumed to be 69 nt, whereas loops 4 to 32 nt in length have been predicted for other viral frameshifting systems (6, 46). It is unclear whether the length of L2 may influence the stability of the RNA pseudoknot or the frameshifting efficiency.

The ORF1a/ORF1b ribosomal frameshifting mechanism and the RNA structures involved are remarkably conserved in coronaviruses (4–6), toroviruses (40), and arteriviruses (Fig. 3). Its occurrence in EAV identifies translational frameshifting as an ancient and probably essential regulating step in POL gene expression.

EAV ORF1b product has sequence similarity with the polymerase protein of coronaviruses and toroviruses. The conclusion that coronaviruses and toroviruses are evolutionarily related was based on the similar organization and expression of their POL genes and on the presence of four conserved domains in their predicted ORF1b products (40). Although the EAV ORF1b (4.2 kb) is considerably smaller than the corresponding ORFs of coronaviruses (8.0 kb) and toroviruses (6.9 kb), we have identified four domains in the sequence of its product which we postulate are homologous with the conserved motifs from coronavirus and torovirus ORF1b products (Fig. 5A).



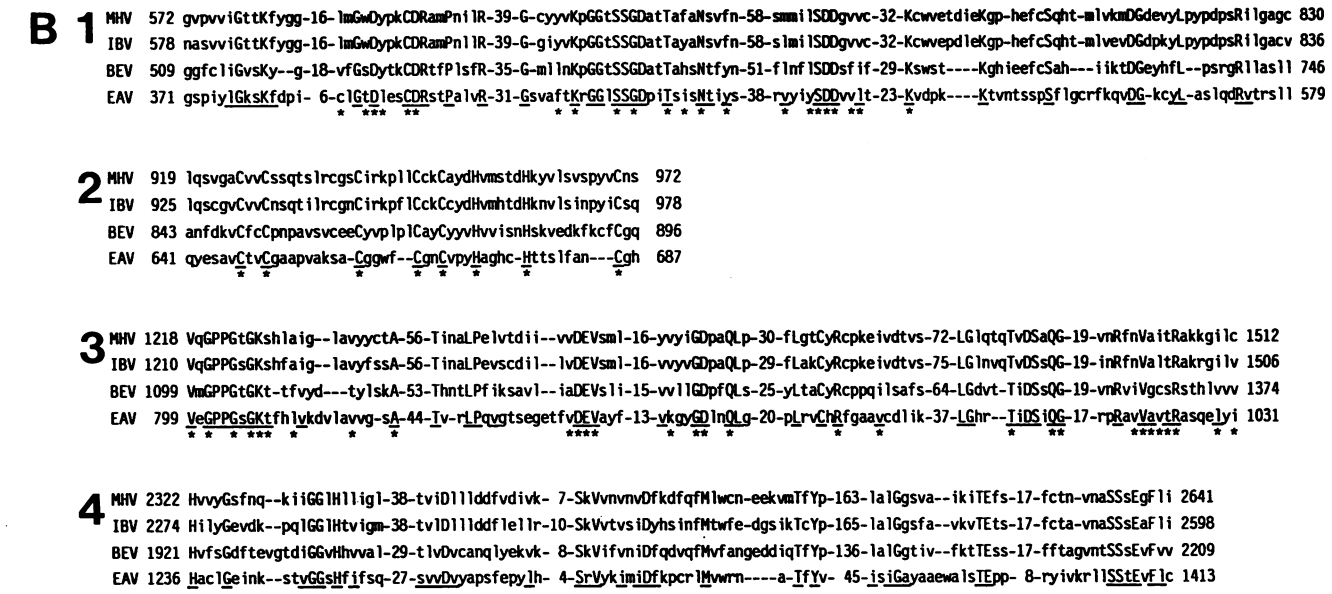


FIG. 5. Positions (A) and aligned sequences (B) of the conserved domains in the ORF1b product of the POL genes of the coronaviruses MHV-A59 (4) and IBV-M42 (3), the torovirus BEV (40), and the arterivirus EAV. Amino acids which are identical in all four sequences are shown in capital letters. Amino acids which are identical or conserved in EAV are underlined. Domain 1, amino acids described as conserved in the majority of positive-stranded RNA viruses (33) are indicated by asterisks. Domain 2, possible conserved cysteine and histidine residues are marked by asterisks. Domain 3, amino acids conserved in "Sindbis virus-like" RNA plant viruses (group A2 [20]) are indicated by asterisks.

Domain 1 represents the well-known polymerase or GDD motif which has been identified in the polymerases of all positive-stranded RNA viruses (for a review, see reference 33). The best match for EAV is in alignment with coronaviruses and toroviruses. EAV also contains a serine instead of a glycine residue in the GDD core of the motif (SDD).

Motif 2 is rich in cysteine and histidine residues and was originally identified by Gorbalenya et al. (18) as a possible zinc-binding finger in the IBV polymerase. Although the spacing between the cysteine and histidine residues is not absolutely conserved, the EAV ORF1b product may also contain a part of this motif.

Domain 3 has been postulated to represent a nucleoside triphosphate (NTP)-binding helicase which is commonly found in proteins of both viral and cellular origin (19, 22). A role for NTP-binding proteins in RNA duplex unwinding has been suggested. Recently, the plum pox potyvirus CI protein, which contains the putative helicase motif, was shown to exhibit an NTP-dependent RNA duplex-unwinding activity in vitro (28). The EAV helicase motif, like that of coronaviruses and toroviruses (40), is more closely related to the alphaviruslike helicase motifs than to the corresponding domains of picornaviruslike origin (20). To date, coronaviruses, toroviruses, and arteriviruses are the only positive-stranded RNA viruses with an unsegmented genome in which the order of the polymerase and helicase domains is reversed (i.e., the helicase domain is positioned downstream of the polymerase motif).

The fourth conserved domain is located near the C terminus of the ORF1b product (Fig. 5) and has been described

```
A
TRP     sqwvvsaaHcyks - 37 - nnDimliklks - 77 - kdscqGdSGgpvvcsg-----klqGivswgsgc
CHT     enwvvtaaHcgvt - 38 - nnDitllklst - 77 - vsscmGdSGgplvckkng-awtlvGivswgsst
ELA     qnwvmtaaHcvdr - 41 - gyDiallrlaq - 78 - rsgcqGdSGgplhclvng-qyavhGvtsfvsrl
                 *              *                      *
SIN     egkvmkplHvkgt - 15 - ayDmefaqlpv - 37 - gvggrGdSGrpimdnsgrvvaivlGgadegtrt
YFV     ggvfhtmwHvtrg - 17 - keDlvayggsw - 48 - ldypsGtSGspivnrnge----viGlygngilv
BVDV    qggissvdHvtag - 21 - ltDeteygvkt - 52 - lknlkGwSGlpifeassgr---vvGrvkvgkne
                 *              *                      *
EAV     evvvltasHvvgr - 19 - ngDfaeavttq - 40 - awttsGdSGsavvqgda-----vvGvht-gsnt

B
PV1     dnvailptHaspg - 38 - frDirqhiptq - 46 - fptraGqCGg-vitctg------------kviGmH-vGgng
HRV14   drvcviptHaqpg - 38 - frDirgfised - 45 - yatktGqCGg-vlcatg------------kifGiH-vGgng
FMDV    gtaylvprHlfae - 45 - vrDitkhfrdt - 49 - aatkaGyCGgavlakdgadt---------fivGtHsaGgng
EMCV    grtlvvnrHmaes - 41 - frDntskfvka - 49 - antrkGwCGsalladlggsk---------kilGiHsaGsmg
HAV     ddwllvpsHaykf - 47 - frDitqhfikk - 58 - geglpGmCGgalvssnqsiqn--------ailGiHvaGgns
CPMV    lackhfftHiktk - 43 - cwDlfcwdpdk - 56 - aptipedCGs-lviahiggkh--------kivGvHvaGiqg
                 *              *                      *
SBMV    mdvlmvphHvwyn - 28 - riDfvlvkvpt - 52 - aptakGwSGtplytrd------------givGmHt-Gyvd
                 *              *                      *
EAV     evvvltasHvvgr - 19 - ngDfaeavttq - 40 - awttsGdSGsavvqgd------------avvGvHt-Gsntsgvay
                 *              *                      *
MHV     ddkvycprHvics -          97          - gsflcGsCGsvgyvltgdsvrfvymhqlelstGcHt-Gtdfsgnfy
IBV     gdtiycprHvlgk -         101          - asflaGaCGsvgfniekgvvnffymhhlelpnalHt-Gtdlmgefy

C
PAP     pvknQgscGsCWafsa - 127 - kvdHavaavgyn
MCP     pvknQgqcGsCWafsa - 128 - nldHgvllvgyg
CATB    qirdQgscGsCWafga - 164 - mggHairilgwg
CATH    pvknQgacGsCWtfst - 130 - kvnHavlavgyg
CATL    pvkdQgacGsCWafnt - 128 - dldHgvlvvgyg
CDP     dic-QgalGdCWllaa - 146 - vkgHaysvtapk
                 *                  *
EAV     vttdQeqdGfCWlkll - 201 - vldHileaatfg
                 *                  *
MHV1    fyspaiertnCWlrst - 142 - ndcHsmavvdkg
MHV2    yfafkqsnnnCyinva - 148 - svgH-ythvkck
IBV     flilewrdGnCWissa - 154 - nsgHcytqaagq
```

FIG. 6. Comparative analysis of possible EAV protease domains. (A) Alignment of the putative EAV trypsinlike serine protease motif with cellular and alphavirus, flavivirus, and pestivirus serine proteases. Identical residues are shown in capital letters; putative catalytic residues are indicated with asterisks. The alignment was based on sequence data and comparisons presented in references 2 and 17. (B) Alignment of the putative EAV serine protease and coronavirus 3C-like cysteine protease motifs with the 3C protease sequences of selected picornaviruses and picornaviruslike viruses. The figure was based on sequence data and comparisons presented in references 1, 16, and 29. The alignment of the coronavirus 3C-like sequences with 3C proteases (Fig. 10 in reference 29) was modified to optimize the alignment with the EAV sequence. (C) Alignment of the putative EAV papainlike cysteine protease with selected cellular cysteine proteases and postulated coronavirus papainlike proteases. The figure was based on the alignment presented in Fig. 12 of reference 29. Abbreviations: BVDV, bovine viral diarrhea virus; CATB, rat cathepsin B; CATH, rat cathepsin H; CATL, rat cathepsin L; CDP, chicken calcium-dependent protease; CHT, chymotrypsin; CPMV, cowpea mosaic virus; ELA, elastase; EMCV, encephalomyocarditis virus; FMDV, foot-and-mouth disease virus; HAV, hepatitis A virus; HRV14, human rhinovirus type 14; IBV, infectious bronchitis virus strain M42; MCP, mouse cysteine protease; MHV, mouse hepatitis virus strain JHM; PAP, papain; PV1, poliovirus type 1; SBMV, southern bean mosaic virus; SIN, Sindbis virus; TRP, trypsin; YFV, yellow fever virus.

only for coronaviruses and toroviruses (40). The central part of this domain is missing in EAV. This could imply that motif 4 consists of multiple functional domains, not all of which have been conserved in the EAV polymerase.

Analysis of the EAV ORF1a sequence. Like the ORF1b product, the EAV ORF1a protein (1,727 amino acids) is considerably smaller than the corresponding coronavirus polypeptide (3,951 amino acids for IBV [3], 4,488 amino acids for MHV [29]). Comparison of the EAV and coronavirus ORF1a proteins did not reveal any obvious homologous domains. This is not surprising, since the ORF1a product has been reported to be much more diverged than the ORF1b protein: when the ORF1a-encoded amino acid sequences of the coronaviruses IBV and MHV were compared, only moderate sequence similarities were observed (29).

The EAV ORF1a product contains a number of cysteine-rich motifs (amino acids 25 to 54, 243 to 356, and 632 to 657 in Fig. 2) and some very hydrophobic regions (e.g., amino acids 525 to 575 and 890 to 970 in Fig. 2). An interesting feature is the presence of a serine protease consensus sequence between amino acids 1080 and 1220 (Fig. 6A). The alignment suggests that the histidine, aspartic acid, and serine residues at amino acid positions 1103, 1129, and 1184, respectively, form the catalytic triad of a trypsinlike serine protease. In view of the large size of the predicted EAV RNA1 translation products (187 and 346 kDa), the presence of a viral protease which can process these primary translation products into smaller active units would not be unexpected.

The relative position of the putative EAV trypsinlike protease domain in ORF1a corresponds to a putative prote-

ase motif in the ORF1a products of IBV and MHV (18, 29). However, this coronavirus domain has been proposed to be related to the picornavirus 3C-like cysteine proteases (18, 29). On the basis of sequence comparison and secondary structure predictions, the trypsinlike serine proteases and the 3C-like cysteine proteases are assumed to belong to the same protease superfamily (1, 16). Therefore, the putative EAV and coronavirus proteases can probably be considered related by common ancestry. Whether a more direct relationship exists, i.e., a relatively recent substitution of an active-site cysteine for serine in EAV or serine for cysteine in coronaviruses, is unclear from the alignment in Fig. 6B. Such a substitution has been postulated for the putative serine protease of southern bean mosaic virus, which appears to be more closely related to 3C-like cysteine proteases than to conventional serine proteases (16) (Fig. 6B). In addition to sequence similarities around the active-site residues, there is some similarity between EAV and coronaviruses in the C-terminal region of the domain, which is assumed to be involved in substrate binding (1, 16) (Fig. 6B). On the other hand, the spacing between the motifs which form the putative EAV protease domain is similar to the spacing in alphavirus, flavivirus, and pestivirus serine proteases (Fig. 6A), and EAV appears to contain the catalytic aspartic acid residue which is lacking in coronaviruses (29).

In addition to the 3C-like protease domain discussed above, possible papainlike cysteine protease motifs have been identified in the N-terminal one-third of the MHV and IBV ORF1a products (29). The sequence of the EAV ORF1a protein between amino acids 158 and 178 displays some sequence similarity with the region around the active-site cysteine-tryptophan dipeptide of cellular and alphavirus papainlike proteases (21, 27). The EAV sequence contains several histidine residues which could fulfill the role of active-site histidine in a papainlike protease. The most convincing alignment is obtained when the histidine at position 374 is used (Fig. 6B), although a 201-residue distance between the cysteine- and histidine-containing regions appears to be rather long for a papainlike protease domain. Experimental data will have to confirm the functionality and nature of both the arterivirus and the coronavirus putative proteases.

Coronaviruslike superfamily. On the basis of computer-assisted analysis of protein sequences and in vitro transcription and translation experiments (8a), a genome organization with the general order 5′-polymerase gene-envelope protein genes-N protein gene-3′ emerges for EAV. This organization is atypical of alphaviruses and rubiviruses, in which only two ORFs are present in the genome and in which the N protein gene is always present upstream of the envelope protein-encoding region (44). However, the organization of the EAV genome, its expression through the production of a 3′-coterminal nested set of mRNAs, and the translation of the second POL ORF by ribosomal frameshifting are remarkably similar to the characteristics of coronavirus (6, 42) and torovirus (39, 40) genome organization and expression.

During the last decade, various replicase modules have been recognized among positive-stranded RNA viruses. They form the basis for superfamilies of plant and animal viruses, each displaying characteristic features. The two largest superfamilies are those of the picornaviruslike and the alphaviruslike virus groups (14, 44). More recently, we proposed a coronaviruslike superfamily (40), which would comprise the coronaviruses and toroviruses and, on the basis of the data presented here, the arteriviruses. The importance of the four conserved domains in coronavirus and torovirus polymerases (40) is underscored by their presence in the putative polymerase of the only distantly related EAV. This is not surprising for the polymerase and helicase activities (domains 1 and 3, respectively), as both are common to the replicases of positive-stranded RNA viruses. However, the conservation, both in sequence and in relative position, of domains 2 and 4 strongly suggests that they also play an important role in the replication of coronaviruslike viruses. This hypothesis is supported by the fact that domain 4 has recently been identified in the putative polymerase of lactate dehydrogenase-elevating virus (LDV) (11a). Although LDV has been referred to as a possible togavirus (53), its genome consists of multiple ORFs (12) and the LDV N protein gene is located at the 3′ end of the genome (13). It is tempting to speculate that this domain performs a function which is specific for coronaviruslike viruses, e.g., the synthesis of multiple subgenomic mRNAs. Since the EAV genome is only 12.7 kb in size, an infectious EAV cDNA clone can be constructed more easily than a similar clone of a coronavirus or torovirus. This will enable us to test the biological functions of domain 4 (and also of domain 2) in the near future.

Our observations may have consequences for virus classification in general. On the basis of the similarities in polymerase expression and amino acid sequence, we postulate that the polymerase genes of arteriviruses, coronaviruses, and toroviruses have descended from a common ancestor. Logically, this evolutionary relationship should be acknowledged by taxonomic position. However, nucleocapsid architecture—a classic trait for viral taxonomy, with the same ranking as the type of nucleic acid or the presence of an envelope—is icosahedral in EAV (23), helical in coronaviruses (for a review, see reference 37), and tubular in toroviruses (52). An additional difference at the structural level is the fact that the EAV envelope does not bear the elongated spikes which are present in both coronaviruses (for a review, see reference 45) and toroviruses (24, 41). The coupling of different arrays of structural genes to the same replicase has been explained by recombination of complete genes or sets of genes (modules). Together with divergence from a common ancestor, modular evolution can account for the diverse composition of viral genomes (14, 44, 55). The joining of a coronaviruslike replicase module to a set of structural genes which confer togavirus morphology to EAV might be another example of modular evolution. Viral taxonomy is faced with the dilemma of designing a system which is both intellectually satisfying and practical.

ACKNOWLEDGMENTS

We thank Peter Bredenbeek, Raoul de Groot, and René Rijnbrand for helpful discussions and assistance.

E.C. was the recipient of a postdoctoral EMBO fellowship (ALTF 131-1988). Part of this work was supported by a research grant from Duphar BV, Weesp, The Netherlands.

REFERENCES

1. **Bazan, J. F., and R. J. Fletterick.** 1988. Viral cysteine proteases are homologous to the trypsin-like family of serine proteases: structural and functional implications. Proc. Natl. Acad. Sci. USA **85**:7872–7876.
2. **Bazan, J. F., and R. J. Fletterick.** 1989. Detection of a trypsin-like serine protease domain in flaviviruses and pestiviruses. Virology **171**:637–639.
3. **Boursnell, M. E. G., T. D. K. Brown, I. J. Foulds, P. F. Green, F. M. Tomley, and M. M. Binns.** 1987. Completion of the sequence of the genome of the coronavirus avian infectious bronchitis virus. J. Gen. Virol. **68**:57–77.

4. **Bredenbeek, P. J., C. J. Pachuk, J. F. H. Noten, J. Charité, W. Luytjes, S. R. Weiss, and W. J. M. Spaan.** 1990. The primary structure and expression of the second open reading frame of the polymerase gene of the coronavirus MHV-A59. Nucleic Acids Res. **18:**1825–1832.
5. **Brierley, I., M. E. G. Boursnell, M. M. Binns, B. Bilimoria, V. C. Blok, T. D. K. Brown, and S. C. Inglis.** 1987. An efficient ribosomal frameshifting signal in the polymerase-encoding region of the coronavirus IBV. EMBO J. **6:**3779–3785.
6. **Brierley, I., P. Diggard, and S. C. Inglis.** 1989. Characterization of an efficient coronavirus ribosomal frameshifting signal: requirement for an RNA pseudoknot. Cell **57:**537–547.
7. **Devereux, J., P. Haeberli, and O. Smithies.** 1984. A comprehensive set of sequence analysis programs for the VAX. Nucleic Acids Res. **12:**387–395.
8. **de Vries, A. A. F., E. D. Chirnside, P. J. Bredenbeek, L. A. Gravestein, M. C. Horzinek, and W. J. M. Spaan.** 1990. All subgenomic mRNAs of equine arteritis virus contain a common leader sequence. Nucleic Acids Res. **18:**3241–3247.

8a. **de Vries, A. A. F., et al.** Unpublished data.

9. **Doll, E. R., J. T. Bryans, W. H. M. McCollum, and M. E. Wallace.** 1957. Isolation of a filterable agent causing arteritis of horses and abortion of mares. Its differentiation from the equine (abortion) influenza virus. Cornell Vet. **47:**3–41.
10. **Doll, E. R., R. E. Knappenberger, and J. T. Bryans.** 1957. An outbreak of abortion caused by the equine arteritis virus. Cornell Vet. **47:**69–75.
11. **Fuerst, T. R., E. G. Niles, F. W. Studier, and B. Moss.** 1986. Eukaryotic transient-expression system based on recombinant vaccinia virus that synthesizes bacteriophage T7 RNA polymerase. Proc. Natl. Acad. Sci. USA **83:**8122–8126.

11a. **Godeny, E. K., and M. A. Brinton.** Personal communication.

12. **Godeny, E. K., D. W. Speicher, and M. A. Brinton.** 1990. Sequence analysis of the genome RNA of lactate dehydrogenase-elevating virus, p. 36–40. *In* M. A. Brinton and F. X. Heinz (ed.), New aspects of positive-strand RNA viruses. American Society for Microbiology, Washington, D.C.
13. **Godeny, E. K., D. W. Speicher, and M. A. Brinton.** 1990. Map location of lactate dehydrogenase-elevating virus (LDV) capsid protein (Vp1). Virology **177:**768–771.
14. **Goldbach, R., and J. Wellink.** 1988. Evolution of plus-strand RNA viruses. Intervirology **29:**260–267.
15. **Golnik, W., A. Moraillon, and J. Golnik.** 1986. Identification and antigenic comparison of equine arteritis virus isolated from an outbreak of epidemic abortion of mares. J. Vet. Med. B **33:**413–417.
16. **Gorbalenya, A. E., A. P. Donchenko, V. M. Blinov, and E. V. Koonin.** 1989. Cysteine proteases of positive strand RNA viruses and chymotrypsin-like serine proteases: a distinct protein superfamily with a common structural fold. FEBS Lett. **243:**103–114.
17. **Gorbalenya, A. E., A. P. Donchenko, E. V. Koonin, and V. M. Blinov.** 1989. N-terminal domains of putative helicases of flavi- and pestiviruses may be serine proteases. Nucleic Acids Res. **17:**3889–3897.
18. **Gorbalenya, A. E., E. V. Koonin, A. P. Donchenko, and V. M. Blinov.** 1989. Coronavirus genome: prediction of putative functional domains in the non-structural polyprotein by comparative amino acid sequence analysis. Nucleic Acids Res. **17:**4847–4861.
19. **Gorbalenya, A. E., and E. V. Koonin.** 1989. Viral proteins containing the purine NTP-binding sequence pattern. Nucleic Acids Res. **17:**8413–8440.
20. **Habili, N., and H. Symons.** 1989. Evolutionary relationship between luteoviruses and other RNA plant viruses based on sequence motifs in their putative RNA polymerases and nucleic acid helicase. Nucleic Acids Res. **17:**9543–9555.
21. **Hardy, W. R., and J. H. Strauss.** 1989. Processing the nonstructural proteins 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.
22. **Hodgman, T. C.** 1988. A new superfamily of replicative proteins. Nature (London) **333:**22–23.
23. **Horzinek, M. C., J. Maess, and R. Laufs.** 1971. Studies on the substructure of togaviruses. II. Analysis of equine arteritis, rubella, bovine viral diarrhea and hog cholera viruses. Arch. Gesamte Virusforsch. **33:**306–318.
24. **Horzinek, M. C., J. Ederveen, B. Kaeffer, D. de Boer, and M. Weiss.** 1986. The peplomers of Berne virus. J. Gen. Virol. **67:**2475–2483.
25. **Hyllseth, B.** 1973. Structural proteins of equine arteritis virus. Arch. Gesamte Virusforsch. **40:**177–188.
26. **Jacks, T., D. H. Madhani, F. R. Masiarz, and H. E. Varmus.** 1988. Signals for ribosomal frameshifting in the Rous sarcoma virus *gag-pol* region. Cell **55:**449–458.
27. **Kamphuis, I. G., J. Drenth, and E. N. Baker.** 1985. Thiol proteases: comparative studies based on the high-resolution structures of papain and actinidin, and on amino acid sequence information for cathepsins B and H, and stem bromelain. J. Mol. Biol. **182:**317–329.
28. **Lain, S., J. L. Riechmann, and J. A. Garcia.** 1990. RNA helicase: a novel activity associated with a protein encoded by a positive strand RNA virus. Nucleic Acids Res. **18:**7003–7006.
29. **Lee, H. J., C. K. Shieh, A. E. Gorbalenya, E. V. Koonin, N. la Monica, J. Tuler, A. Bagdzhadzhyan, and M. M. C. Lai.** 1991. The complete sequence (22 kilobases) of murine coronavirus gene 1 encoding the putative protease and RNA polymerase. Virology **180:**567–582.
30. **Murphy, F. A.** 1980. Togavirus morphology and morphogenesis, p. 241–316. *In* R. W. Schlesinger (ed.), The togaviruses: biology, structure and replication. Academic Press, Inc., New York.
31. **Pearson, W. R., and D. J. Lipman.** 1988. Improved tools for biological sequence comparison. Proc. Natl. Acad. Sci. USA **85:**2444–2448.
32. **Pleij, C. W. A., K. Rietveld, and L. Bosch.** 1985. A new principle of RNA folding based on pseudoknotting. Nucleic Acids Res. **13:**1717–1731.
33. **Poch, O., I. Sauvaget, M. Delarue, and N. Tordo.** 1989. Identification of four conserved motifs among the RNA dependent polymerase encoding elements. EMBO J. **8:**3867–3874.
34. **Porterfield, J. S., J. Casals, M. P. Chumakov, S. Y. Gaidamovich, C. Hannoun, I. H. Holmes, M. C. Horzinek, M. Mussgay, N. Oker-Blom, P. K. Russel, and D. W. Trent.** 1978. *Togaviridae*. Intervirology **9:**129–148.
35. **Sawicki, S. G., and D. L. Sawicki.** 1990. Coronavirus transcription: subgenomic mouse hepatitis virus replicative intermediates function in RNA synthesis. J. Virol. **64:**1050–1056.
36. **Sethna, P. B., S. L. Hung, and D. A. Brian.** 1989. Coronavirus subgenomic minus-strand RNAs and the potential for mRNA replicons. Proc. Natl. Acad. Sci. USA **86:**5626–5630.
37. **Siddell, S. G., H. Wege, and V. ter Meulen.** 1983. The structure and replication of coronaviruses. Curr. Top. Microbiol. Immunol. **99:**131–163.
38. **Snijder, E. J., J. Ederveen, W. J. M. Spaan, M. Weiss, and M. C. Horzinek.** 1988. Characterization of Berne virus genomic and messenger RNAs. J. Gen. Virol. **69:**2135–2144.
39. **Snijder, E. J., M. C. Horzinek, and W. J. M. Spaan.** 1990. A 3′-coterminal nested set of independently transcribed mRNAs is generated during Berne virus replication. J. Virol. **64:**331–338.
40. **Snijder, E. J., J. A. den Boon, P. J. Bredenbeek, M. C. Horzinek, R. Rijnbrand, and W. J. M. Spaan.** 1990. The carboxyl-terminal part of the putative Berne virus polymerase is expressed by ribosomal frameshifting and contains sequence motifs which indicate that toro- and coronaviruses are evolutionary related. Nucleic Acids Res. **18:**4535–4542.
41. **Snijder, E. J., J. A. den Boon, W. J. M. Spaan, M. Weiss, and M. C. Horzinek.** 1990. Primary structure and posttranslational processing of the Berne virus peplomer protein. Virology **178:**355–363.
42. **Spaan, W. J. M., D. Cavanagh, and M. C. Horzinek.** 1988. Coronaviruses: structure and genome expression. J. Gen. Virol. **69:**2939–2952.
43. **Staden, R.** 1986. The current status and portability of our sequence handling software. Nucleic Acids Res. **14:**217–233.
44. **Strauss, J. H., and E. G. Strauss.** 1988. Evolution of RNA

viruses. Annu. Rev. Microbiol. **42**:657–683.

45. **Sturman, L. S., and K. V. Holmes.** 1985. The novel glycoproteins of coronaviruses. Trends Biochem. Sci. **10**:17–20.
46. **ten Dam, E. B., C. W. A. Pleij, and L. Bosch.** 1990. RNA pseudoknots; translational frameshifting and readthrough on viral RNAs. Virus Genes **4**:121–136.
47. **Timoney, P. J., W. H. McCollum, A. W. Roberts, and T. W. Murphy.** 1986. Demonstration of the carrier state in naturally acquired equine arteritis virus infection in the stallion. Res. Vet. Sci. **41**:279–280.
48. **van Berlo, M. F., M. C. Horzinek, and B. A. M. van der Zeijst.** 1982. Equine arteritis virus-infected cells contain six polyadenylated virus-specific RNAs. Virology **118**:345–352.
49. **van Berlo, M. F., P. J. M. Rottier, M. C. Horzinek, and B. A. M. van der Zeijst.** 1986. Intracellular equine arteritis virus (EAV)-specific RNAs contain common sequences. Virology **152**:492–496.
50. **van Berlo, M. F., P. J. M. Rottier, W. J. M. Spaan, and M. C. Horzinek.** 1986. Equine arteritis virus (EAV)-induced polypeptide synthesis. J. Gen. Virol. **67**:1543–1549.
51. **van der Zeijst, B. A. M., M. C. Horzinek, and V. Moennig.** 1975. The genome of equine arteritis virus. Virology **68**:418–425.
52. **Weiss, M., F. Steck, and M. C. Horzinek.** 1983. Purification and partial characterization of a new enveloped RNA virus (Berne virus). J. Gen. Virol. **64**:1849–1858.
53. **Westaway, E. G., M. A. Brinton, S. Y. Gaidamovich, M. C. Horzinek, A. Igarashi, L. Kaariainen, D. K. Lvov, J. S. Porterfield, P. K. Russel, and D. W. Trent.** 1985. *Togaviridae*. Intervirology **24**:125–139.
54. **Zeegers, J. J. W., B. A. M. van der Zeijst, and M. C. Horzinek.** 1976. The structural proteins of equine arteritis virus. Virology **73**:200–205.
55. **Zimmern, D.** 1987. Evolution of RNA viruses, p. 211–240. *In* J. J. Holland, E. Domingo, and P. Ahlquist (ed.), RNA genetics, vol. 2. CRC Press, Boca Raton, Fla.