Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

VIROLOGY 177, 225–238 (1990)

Sequence of the Genome RNA of Rubella Virus: Evidence for Genetic Rearrangement during Togavirus Evolution

GERALDINA DOMINGUEZ, CHIN-YEN WANG, AND TERYL K. FREY[1]

Department of Biology and Laboratory for Microbial and Biochemical Sciences, Georgia State University, P. O. Box 4010, Atlanta, Georgia 30302-4010

Received January 24, 1990; accepted March 19, 1990

The nucleotide sequence of the rubella virus (RUB) genomic RNA was determined. The RUB genomic RNA is 9757 nucleotides in length [excluding the poly(A) tail] and has a G/C content of 69.5%, the highest of any RNA virus sequenced to date. The RUB genomic RNA contains two long open reading frames (ORFs), a 5′-proximal ORF of 6656 nucleotides and a 3′-proximal ORF of 3189 nucleotides which encodes the structural proteins. Thus, the genomic organization of RUB is similar to that of alphaviruses, the other genus of the Togavirus family, and the 5′-proximal ORF of RUB therefore putatively codes for the nonstructural proteins. Sequences homologous to three regions of nucleotide sequence highly conserved among alphaviruses (a stem-and-loop structure at the 5′ end of the genome, a 51-nucleotide conserved sequence near the 5′ end of the genome, and a 20-nucleotide conserved sequence at the subgenomic RNA start site) were found in the RUB genomic RNA. Amino acid sequence comparisons between the nonstructural ORF of RUB and alphaviruses revealed only one short (122 amino acids) region of significant homology, indicating that these viruses are only distantly related. This region of homology is located at the NH_2 terminus of nsP3 in the alphavirus genome. The RUB nonstructural protein ORF contains two global amino acid motifs conserved in a large number of positive-polarity RNA viruses, a motif indicative of helicase activity and a motif indicative of replicase activity. The order of the helicase motif and the nsP3 homology region in the RUB genome is reversed with respect to the alphavirus genome indicating that a genetic rearrangement has occurred during the evolution of these viruses.

INTRODUCTION

Rubella virus (RUB) is a single-stranded, positive-polarity RNA virus classified in the Togavirus family as the only member of the genus *Rubivirus* (Matthews, 1982). In the rubella virion, the RNA genome is enclosed in an icosahedral nucleocapsid composed of multiple copies of a single protein, the capsid or C protein (M_r = 34 kDa) (Ho-Terry and Cohen, 1982; Waxham and Wolinsky, 1983; Oker-Blom *et al.*, 1983). The nucleocapsid is surrounded by a lipid bilayer envelope in which the two virus-specific glycoproteins, E1 and E2 (M_r's = 58 kDa and 42–47 kDa, respectively) are embedded. In infected cells, in addition to the genomic RNA, a subgenomic RNA is synthesized (Oker-Blom *et al.*, 1984) which consists of the 3′-terminal 3327 nucleotides of the genomic RNA (Frey *et al.*, 1989). The subgenomic RNA contains a single long open reading frame (ORF) which is translated into a 110-kDa polyprotein and is post-translationally processed into the structural proteins. The order of the structural proteins within the 110-kDa precursor is NH_2–C–E2–E1–COOH (Oker-Blom, 1984). Both positive-polarity RNA species are transcribed from a genome-length, negative-polarity RNA template (Hemphill *et al.*, 1988). In these aspects of virion structure and replication strategy, RUB is similar to the alphaviruses which have been extensively characterized (Strauss and Strauss, 1986). The sequence of the 3′-terminal 4500 nucleotides of the RUB genome which contains the structural protein ORF has been reported (Frey *et al.*, 1986; Clarke *et al.*, 1987; Vidgren *et al.*, 1987; Takkinen *et al.*, 1988; Frey and Marr, 1988). No significant homology exists between the structural protein coding regions of RUB and the alphaviruses, indicating that these viruses are only distantly related.

In the alphavirus genome, a single ORF of approximately 7.4 kb spans the 5′ two-thirds of the genomic RNA. The polyprotein translated from this ORF is post-translationally processed into four nonstructural proteins, nsP1, nsP2, nsP3, and nsP4 (Strauss and Strauss, 1986). In Sindbis virus (SIN), Ross River virus, and Middelburg virus, there is an in-frame opal termination codon between the nonstructural proteins nsP3 and nsP4 which is occasionally read through (Strauss *et al.*, 1983). However, Semliki Forest virus and O'Nyong-nyong virus lack this opal codon (Strauss *et al.*, 1988). Some of the functions associated with these proteins have been determined. Temperature-sensitive (ts^-) RNA^- mutants have been mapped to nsP1, nsP2,

Sequence data reported in this article have been submitted to GenBank and assigned the accession number M32735.

[1] To whom requests for reprints should be addressed.

0042-6822/90 $3.00

and nsP4 and from the nature of the phenotypes of these ts$^-$ mutants it seems likely that nsP4 is the RNA polymerase, nsP1 functions in minus-polarity RNA synthesis, and nsP2 functions in subgenomic RNA synthesis (Hahn *et al.*, 1989a,b). Mutations in the methyltransferase activity required for capping of the 5′ terminus of the genomic and subgenomic RNAs have also been mapped to nsP1 (Mi *et al.*, 1989). nsP2 has been shown to contain the autoprotease required for posttranslational processing of the nonstructural polyprotein (Ding and Schlesinger, 1989; Hardy and Strauss, 1989).

In contrast, the nonstructural proteins of RUB have not been characterized. RUB does not shut off host cell macromolecular synthesis, making the small quantities of both structural and nonstructural proteins synthesized very difficult to detect over the host cell background (Hemphill *et al.*, 1988).

In this paper we present the entire sequence of the RUB genome. The RUB genomic RNA is 9757 nucleotides in length and has a high G/C content (69.5%: 38.7%C, 30.8%G). In addition to the structural protein ORF at the 3′ end of the genomic RNA, a second long ORF is present from nucleotides 41 to 6656 which putatively codes for the nonstructural proteins. Comparison of the 5′-proximal ORFs of RUB and alphaviruses revealed only one short region (122 amino acids) of significant homology. Two global motifs indicative of replicase and helicase function were also found in the RUB 5′ proximal ORF. The relative positioning within the 5′-proximal ORFs of RUB and alphaviruses of the short region of homology and the two global amino acid motifs suggests that a rearrangement in this region of the genome has occurred in the evolution of these viruses.

MATERIALS AND METHODS

Cells and virus

Vero cells obtained from the American Type Culture Collection were maintained at 35° under 4% CO_2 in Eagle Minimal Essential Medium containing Earle's salts and supplemented with 10% tryptose phosphate and 5% fetal bovine serum. Plaque-purified RUB stocks (Therien strain) prepared as described previously (Hemphill *et al.*, 1988) were used in this study.

RNA isolation

Virion RNA was isolated by phenol–chloroform extraction of virions purified as described by Waxham and Wolinsky (1983). Intracellular RNA was extracted from infected Vero cells (m.o.i. = 0.1) 72 hr postinfection as previously described (Frey *et al.*, 1986). The extracted RNA was chromatographed over oligo(dT) cellulose.

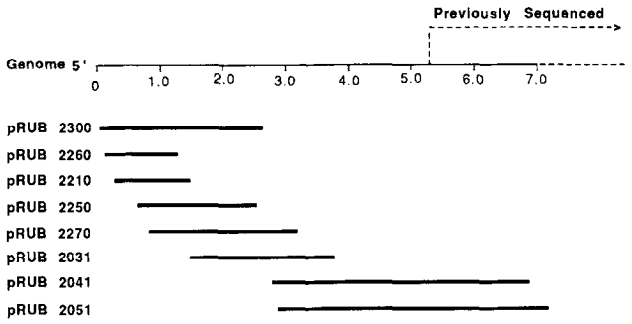

FIG. 1. Map of cDNA clones used in sequence determination. Numbers in the RUB genome scale refer to distances from the 5′ end in kilobases. Determination of the sequence from nucleotide 5249 to the 3′ end of the genome was previously reported (Frey *et al.*, 1986; Frey and Marr, 1988).

After ethanol precipitation, the poly(A)+ fraction was dissolved in 90% DMSO and heated at 55° for 5 min to denature double-stranded RNA replicative forms and intermediates [which bind to oligo(dT) cellulose]. The DMSO-denatured RNA was ethanol precipitated twice and dissolved in 0.01 *M* Tris (pH 8.0), 0.001 *M* EDTA.

Derivation and sequencing of cDNA clones

Virion RNA was used as the template for first-strand cDNA synthesis primed with random deoxyhexamers (Pharmacia) as described by Rice *et al.* (1985). Second-strand DNA synthesis, deoxycytidine (dC) tailing of the double-stranded cDNA with terminal transferase, annealing of dC-tailed cDNA with dG-tailed pUC 9, and transformations were done as previously described (Frey *et al.*, 1986) with the following modifications: double-stranded cDNA was chromatographed on a Sepharose CL-4B column (Pharmacia) to eliminate cDNAs less than 700 nucleotides in length (Eschenfeldt and Berger, 1987) and transformation was done using competent DH5-α cells (Bethesda Research Labs).

Colonies with cDNA clones containing sequences overlapping the previously determined RUB sequence [the 3′-terminal 4508 nucleotides of the genome (Frey and Marr, 1988)] were identified by colony blot hybridization using as probes ^{32}P-labeled restriction fragments and oligonucleotides from the 5′ end of this sequence. This set of cDNA clones was restriction mapped and a restriction fragment from the 5′ end of this set of clones was used as a probe to isolate new overlapping clones. In this manner 18 clones were identified and mapped which covered the region between 4500 nucleotides from the 3′ end of the genome and the 5′ end of the genome.

Sequencing strategy

The cDNA inserts from eight clones (shown in Fig. 1) were subcloned into M13 for sequencing. Subcloning

by use of convenient restriction sites, shotgun cloning of sonicated DNA (Bankier and Barrell, 1983), and exonuclease III digestion to produce directional deletions (Henikoff, 1984) were all employed. Several gaps which remained were sequenced using synthetic oligonucleotide primers on the appropriate templates. Oligonucleotides were synthesized using an Applied Biosystems Model 381A DNA Synthesizer. All sequencing was done by dideoxy sequencing (Sanger *et al.*, 1977) using the procedure recommended by Sequenase Version 2.0 kit (USBiochemicals) with [^{35}S]dATP label and 7-deaza-dGTP in place of dGTP.

Primer extension and dideoxy sequencing from an RNA template were both performed on poly(A)+ RNA from RUB-infected cells using as a primer a 5′-^{32}P-labeled oligonucleotide with the sequence dTGGTCTCT-TACCCAACT, which is complementary to nucleotides 101 to 117 of the genomic RNA. The primer extension reaction was done as previously described (Frey *et al.*, 1989), while RNA sequencing was done by dideoxy sequencing modified for an RNA template (Rico-Hesse *et al.*, 1987; Zimmern and Kaesberg, 1978).

Computer analysis

The analysis of the sequence was performed on the Centers for Disease Control VAX using the University of Wisconsin GCG package (Devereux *et al.*, 1984).

RESULTS AND DISCUSSION

Sequencing strategy

The complete sequence of the RUB genomic RNA is given in Fig. 2. The regions of the RUB genome covered by each of the eight cDNA clones used in sequence determination of the 5′ half of this sequence are shown in Fig. 1. All sequences were determined in both directions and from at least two cDNA clones. Primer extension using poly(A)+ RNA from RUB-infected cells and an oligonucleotide primer complementary to the sequence near the 5′ end of the already determined cDNA sequence indicated that the cDNA sequence extended to within 20 nucleotides of the 5′ end of the genome. Therefore, the same oligo was used to prime dideoxy sequencing from poly(A)+ RNA from RUB-infected cells. Dark bands were present in all four lanes of the top two steps of this sequencing ladder, a common occurrence encountered in sequencing and primer extension using capped RNAs as templates due to termination of reverse transcription at the penultimate and ultimate nucleotides and possible copying of the cap (Gupta and Kingsbury, 1984; Ahlquist and Janda, 1984). Thus the 5′-terminal nucleotides are reported as N–N in Fig. 2. The cDNA sequence was found to terminate 12 nucleotides from the 5′ end of the RUB genomic RNA. To reconcile the absence of two nucleotides at the exact 3′ terminus of the RUB RNA in cDNA clones used to determine the sequence in the earlier report (Frey *et al.*, 1986) but present in other RUB sequence reports (Clarke *et al.*, 1987; Vidgren *et al.*, 1987), additional 3′-cDNA clones were derived and found to contain these additional two nucleotides. The length of the RUB genomic RNA was 9757 nucleotides [excluding the poly(A) tail] which is in good agreement with size estimates of 9800 to 10,000 nucleotides determined by gel electrophoresis (Oker-Blom *et al.*, 1984; Hemphill *et al.*, 1988).

ORF analysis of the RUB sequence

An analysis of coding potential of the RUB genome, shown in Fig. 3, reveals two long ORFs in the positive orientation. The predicted amino acid sequence encoded by these ORFs is given below the nucleotide sequence in Fig. 2. The 3′-most ORF beginning at nucleotide 6507 and ending at nucleotide 9696 (3189 nucleotides in length encoding a polypeptide of 1063 amino acids) codes for the structural proteins as discussed elsewhere (Frey *et al.*, 1986; Frey and Marr, 1988). The 5′-most ORF begins at nucleotide 41 and terminates at position 6656 with an opal codon (UGA) which is followed 12 nucleotides downstream by a second in-frame opal codon. This ORF is 6615 nucleotides in length and encodes a 2205-amino acid polypeptide. These two ORFs overlap by 149 nucleotides. The 3′-terminal 1407 nucleotides of the 5′ ORF was reported earlier (Frey and Marr, 1988). The organization of the RUB genome is thus similar to that of the alphavirus genome. Therefore, the 5′-proximal ORF of the RUB genome will be referred to as the nonstructural protein ORF. In terms of favorability of context of neighboring nucleotides as compiled by Kozak (1987), the AUG initiating the nonstructural protein ORF (CC*AUG*G) has the preferred nucleotides at positions −2, −1, and +4, but lacks the A at position −3.

Interestingly, there is an AUG beginning at nucleotide 3 of the RUB genomic RNA. Translation initiated at this AUG would terminate at a UAG codon beginning at nucleotide 54, yielding a 16-amino acid product. It is not known if translation of this short ORF occurs. In eukaryotic mRNAs, most AUGs at which translation is initiated are 20 to 100 nucleotides from the 5′ end of the mRNA (Kozak, 1987); however, AUGs as close to the 5′ end as nucleotide 4 at which initiation of translation occurs have been reported (Kelley *et al.*, 1982). Initiation of translation at AUGs less than 15 nucleotides from the 5′ end of eukaryotic mRNAs is less efficient than at AUGs greater than 15 nucleotides from the 5′

a

```
   1  NNAUGGAAGCUAUCGGACCUCGCUUAGGACUCCCAUUCCCAUGGAGAAACUCCUAGAUGAGGUUCUUGCCCCCGGUGGGCCUUAUAACUUAACCGUCGGCAGUUGGGUAAGAGACCACGU  120
                                                 M  E  K  L  L  D  E  V  L  A  P  G  G  P  Y  N  L  T  V  G  S  W  V  R  D  H  V   27

 121  CCGAUCAAUUGUCGAGGGCGCGUGGGAAGUGCGCGAUGUUGUUACCGCUGCCCAAAAGCGGGCCAUCGUAGCCGUGAUACCCAGACCUGUGUUCACGCAGAUGCAGGUCAGUGAUCACCC  240
  28   R  S  I  V  E  G  A  W  E  V  R  D  V  V  T  A  A  Q  K  R  A  I  V  A  V  I  P  R  P  V  F  T  Q  M  Q  V  S  D  H  P   67

 241  AGCACUCCACGCAAUUUCGCGGUAUACCCGCCGCCAUUGGAUCGAGUGGGGCCCUAAAGAAGCCCUACACGUCCUCAUCGACCCAAGCCCGGGCCUGCUCCGCGAGGUCGCUCGCGUUGA  360
  68   A  L  H  A  I  S  R  Y  T  R  R  H  W  I  E  W  G  P  K  E  A  L  H  V  L  I  D  P  S  P  G  L  L  R  E  V  A  R  V  E  107

 361  GCGCCGCUGGGUCGCACUGUGCCUCCACAGGACGGCACGCAAACUCGCCACCGCCCUGGCCGAGACGGCCAGCGAGGCGUGGCACGCUGACUACGUGUGCGCGCUGCGUGGCGCACCGAG  480
 108   R  R  W  V  A  L  C  L  H  R  T  A  R  K  L  A  T  A  L  A  E  T  A  S  E  A  W  H  A  D  Y  V  C  A  L  R  G  A  P  S  147

 481  CGGCCCCUUCUACGUCCACCCUGAGGACGUCCCGCACGGCGGUCGCGCCGUGGCGGACAGAUGCUUGCUCUACUACACCCAUGCAGAUGUGCGAGCUGAUGCGUACCAUUGACGCCAC  600
 148   G  P  F  Y  V  H  P  E  D  V  P  H  G  G  R  A  V  A  D  R  C  L  L  Y  Y  T  P  M  Q  M  C  E  L  M  R  T  I  D  A  T  187

 601  CCUGCUCGUGGCGGUUGACUUGUGGCCGGUCGCCCUUGCGGCCCACGUCGGCGACGACUGGGACGACCUGGGCAUUGCCUGGCAUCUCGACCAUGACGGCGGUUGCCCCGCCGAUUGCCG  720
 188   L  L  V  A  V  D  L  W  P  V  A  L  A  A  H  V  G  D  D  W  D  D  L  G  I  A  W  H  L  D  H  D  G  G  C  P  A  D  C  R  227

 721  CGGAGCCGGCGCUGGGCCCACGCCCGGCUACACCCGCCCCUGCACCACACGCAUCUACCAAGUCCUGCCGGACACCGCCCACCCCGGGCGCCUCUACCGGUGCGGGCCCCGCCUGUGGAC  840
 228   G  A  G  A  G  P  T  P  G  Y  T  R  P  C  T  T  R  I  Y  Q  V  L  P  D  T  A  H  P  G  R  L  Y  R  C  G  P  R  L  W  T  267

 841  GCGCGAUUGCGCCGUGGCCGAACUCUCAUGGGAGGUUGCCCAACACUGCGGGCACCAGGCGCGCGUGCGCGCCGUGCGAUGCACCCUCCCUAUCCGCCACGUGCGCAGCCUCCAACCCAG  960
 268   R  D  C  A  V  A  E  L  S  W  E  V  A  Q  H  C  G  H  Q  A  R  V  R  A  V  R  C  T  L  P  I  R  H  V  R  S  L  Q  P  S  307

 961  CGCGCGGGUCCGACUCCCGGACCUCGUCCAUCUCGCCGAGGUGGGCCGGUGGCGGUGGUUCAGCCUCCCCCGCCCCGUGUUCCAGCGCAUGCUGUCCUACUGCAAGACCCUGAGCCCCGA 1080
 308   A  R  V  R  L  P  D  L  V  H  L  A  E  V  G  R  W  R  W  F  S  L  P  R  P  V  F  Q  R  M  L  S  Y  C  K  T  L  S  P  D  347

1081  CGCGUACUACAGCGAGCGCGUGUUCAAGUUCAAGAACGCCCUGUGCCACAGCAUCACGCUCGCGGGCAAUGUGCUGCAAGAGGGGUGGAAGGGCACGUGCGCCGAGGAAGACGCGCUGUG 1200
 348   A  Y  Y  S  E  R  V  F  K  F  K  N  A  L  C  H  S  I  T  L  A  G  N  V  L  Q  E  G  W  K  G  T  C  A  E  E  D  A  L  C  387

1201  CGCAUACGUAGCCUUCCGCGCGUGGCAGUCUAACGCCAGGUUGGCGGGGAUUAUGAAAGGCGCGAAGUGCGCCGCCGACUCUUUGAGCGUGGCCGGCUGGCUGGACACCAUUUGGGACGC 1320
 388   A  Y  V  A  F  R  A  W  Q  S  N  A  R  L  A  G  I  M  K  G  A  K  C  A  A  D  S  L  S  V  A  G  W  L  D  T  I  W  D  A  427

1321  CAUUAAGCGGUUCCUCGGUAGCGUGCCCCUCGCCGAGCGCAUGGAGGAGUGGGAACAGGACGCCGCGGUCGCCGCCUUCGACCGCGGCCCCCUCGAGGACGGCGGGCGCCACUUGGACAC 1440
 428   I  K  R  F  L  G  S  V  P  L  A  E  R  M  E  E  W  E  Q  D  A  A  V  A  A  F  D  R  G  P  L  E  D  G  G  R  H  L  D  T  467

1441  CGUGCAACCCCCAAAAUCGCCGCCCCGCCCUGAGAUCGCCGCGACCUGGAUCGUCCACGCAGCCAGCGAAGACCGCCAUUGCGCGUGCGCUCCCCGCUGCGACGUCCCGCGCGAACGUCC 1560
 468   V  Q  P  P  K  S  P  P  R  P  E  I  A  A  T  W  I  V  H  A  A  S  E  D  R  H  C  A  C  A  P  R  C  D  V  P  R  E  R  P  507

1561  UUCCGCGCCCGCCGGCCAGCCGGAUGACGAGGCGCUCAUCCCGCCGUGGCUGUUCGCCGAGCGCCGUGCCCUCCGCUGCCGCGAGUGGGAUUUCGAGGCUCUCCGCGCGCGCGCCGAUAC 1680
 508   S  A  P  A  G  Q  P  D  D  E  A  L  I  P  P  W  L  F  A  E  R  R  A  L  R  C  R  E  W  D  F  E  A  L  R  A  R  A  D  T  547

1681  GGCGGCCGCGCCCGCCCCGCCGGCUCCACGCCCCGCGCGGUACCCCACCGUGCUCUACCGCCACCCCGCCCACCACGGCCCGUGGCUCACCCUUGACGAGCCGGGCGAGGCUGACGCGGC 1800
 548   A  A  A  P  A  P  P  A  P  R  P  A  R  Y  P  T  V  L  Y  R  H  P  A  H  H  G  P  W  L  T  L  D  E  P  G  E  A  D  A  A  587

1801  CCUGGUCUUAUGCGACCCACUUGGCCAGCCGCUCCGGGGCCCUGAACGCCACUUCGCCGCCGGCGCGCAUAUGUGCGCGCAGGCGCGGGGGCUCCAGGCUUUUGUCCGUGUCGUGCCUCC 1920
 588   L  V  L  C  D  P  L  G  Q  P  L  R  G  P  E  R  H  F  A  A  G  A  H  M  C  A  Q  A  R  G  L  Q  A  F  V  R  V  V  P  P  627

1921  ACCCGAGCGCCCCUGGGCCGACGGGGGCGCCAGAGCGUGGGCGAAGUUCUUCCGCGGCUGCGCCUGGGCGCAGCGCUUGCUCGGCGAGCCAGCAGUUAUGCACCUCCCAUACACCGAUGG 2040
 628   P  E  R  P  W  A  D  G  G  A  R  A  W  A  K  F  F  R  G  C  A  W  A  Q  R  L  L  G  E  P  A  V  M  H  L  P  Y  T  D  G  667

2041  CGACGUGCCACAGCUGAUCGCACUGGCUUUGCGCACGCUGGCCCAACAGGGGGCCGCCUUGGCACUCUCGGUGCGUGACCUGCCCGGGGGUGCAGCGUUCGACGCAAACGCGGUCACCGC 2160
 668   D  V  P  Q  L  I  A  L  A  L  R  T  L  A  Q  Q  G  A  A  L  A  L  S  V  R  D  L  P  G  G  A  A  F  D  A  N  A  V  T  A  707

2161  CGCCGUGCGCGCUGGCCCCCGCCAGUCCGCGGCCGCGUCACCGCCACCCGGCGACCCCCCGCCGCCGCGCCGCGCACGGCGAUCGCAACGGCACUCGGACGCUCGCGGCACUCCGCCCCC 2280
 708   A  V  R  A  G  P  R  Q  S  A  A  A  S  P  P  P  G  D  P  P  P  P  R  R  A  R  R  S  Q  R  H  S  D  A  R  G  T  P  P  P  747

2281  CGCGCCUGCGCGCGACCCGCCGCCGCCCGCCCCCAGCCCGCCCGCGCCACCCCGCCGCUGGUGACCCGGUCCCUCCCAUUCCCGCGGGGCCGGCGGAUCGCGCGCGUGACGCCGAGCUGGA 2400
 748   A  P  A  R  D  P  P  P  P  A  P  S  P  P  A  P  P  R  A  G  D  P  V  P  P  I  P  A  G  P  A  D  R  A  R  D  A  E  L  E  787

2401  GGUCGCCUGCGAGCCGAGCGGCCCCCCCACGUCAACCAGGGCAGACCCAGACAGCGACAUCGUUGAAAGUUACGCCCGCGCCGCCGGACCCGUGCACCUCCGAGUCCGCGACAUCAUGGA 2520
 788   V  A  C  E  P  S  G  P  P  T  S  T  R  A  D  P  D  S  D  I  V  E  S  Y  A  R  A  A  G  P  V  H  L  R  V  R  D  I  M  D  827
```

FIG. 2. Nucleotide sequence of the RUB genome. The identity of the two 5′ nucleotides has not been determined. The deduced amino acid sequence of the nonstructural protein ORF (5′ proximal) and structural protein ORF (3′ proximal) is given. The 5′ end of the subgenomic RNA (Frey *et al.*, 1989) and the amino termini of the structural proteins (Kalkkinen *et al.*, 1984) are indicated. Asterisks denote termination codons.

b

```
2521 CCCACCGCCCGGCUGCAAGGUCGUGGUCAACGCCGCCAACGAGGGGCUACUGGCCGGCUCUGGCGUGUGCGGUGCCAUCUUUGCCAACGCCACGGCGGCCCUCGCUGCAAACUGCCGGCG 2640
 828  P  P  P  G  C  K  V  V  V  N  A  A  N  E  G  L  L  A  G  S  G  V  C  G  A  I  F  A  N  A  T  A  A  L  A  A  N  C  R  R  867

2641 CCUCGCCCCAUGCCCCACCGGCGAGGCAGUGGCGACACCCGGCCACGGCUGCGGGUACACCCACAUCAUCCACGCCGUCGCGCCGCGGCGUCCUCGGGACCCCGCCGCCCUCCAGGAGGG 2760
 868  L  A  P  C  P  T  G  E  A  V  A  T  P  G  H  G  C  G  Y  T  H  I  I  H  A  V  A  P  R  R  P  R  D  P  A  A  L  E  E  G  907

2761 CGAAGCGCUGCUCGAGCGCGCCUACCGCAGCAUCGUCGCGCUAGCCGCCGCGCGCUCGGUGGGCGUGUGUCGCGUGCCCCCUCCUCGGCGCUGGCGUCUACGGCUGGUCUGCUGCGGAGUC 2880
 908  E  A  L  L  E  R  A  Y  R  S  I  V  A  L  A  A  A  R  R  W  A  C  V  A  C  P  L  L  G  A  G  V  Y  G  W  S  A  A  E  S  947

2881 CCUCCGAGCCGCGCUCGCGGCUACGCGCACCGAGCCCGUCGAGCGCGUGAGCCUGCACAUCUGCCACCCCGACCGCGCCACGCUGACGCACGCCUCCGUGCUCGUCGGCGCGGGGCUCGC 3000
 948  L  R  A  A  L  A  A  T  R  T  E  P  V  E  R  V  S  L  H  I  C  H  P  D  R  A  T  L  T  H  A  S  V  L  V  G  A  G  L  A  987

3001 UGCCAGGCGCGUCAGUCCUCCUCCGACCGAGCCCCUCGCAUCUUGCCCCGCCGGUGACCCGGGCCGACCGGCUCAGCGCAGCGCGUCGCCCCCAGCGACCCCCCUUGGGGAUGCCACCGC 3120
 988  A  R  R  V  S  P  P  P  T  E  P  L  A  S  C  P  A  G  D  P  G  R  P  A  Q  R  S  A  S  P  P  A  T  P  L  G  D  A  T  A  1027

3121 GCCCGAGCCCCGCGGAUGCCAGGGGUGCGAACUCUGCCGGUACACGCGCGUCACCAAUGACCGCGCCUAUGUCAACCUGUGGCUCGAGCGCGACCGCGGCGCCACCAGCUGGGCCAUGCG 3240
1028  P  E  P  R  G  C  Q  G  C  E  L  C  R  Y  T  R  V  T  N  D  R  A  Y  V  N  L  W  L  E  R  D  R  G  A  T  S  W  A  M  R  1067

3241 CAUUCCCGAGGUGGUUGUCUACGGGCCGGAGCACCUCGCCACGCAUUUUCCAUUAAACCACUACAGUGUGCUCAAGCCCCGCGGAGGUCAGGCCCCCGCGAGGCAUGUGCGGGAGUGACAU 3360
1068  I  P  E  V  V  V  Y  G  P  E  H  L  A  T  H  F  P  L  N  H  Y  S  V  L  K  P  A  E  V  R  P  P  R  G  M  C  G  S  D  M  1107

3361 GUGGCGCUGCCGCGGCUGGCAUGGCAUGCCGCAGGUGCGGUGCACCCCCUCCAACGCUCACGCCGCCCCUGUGCCGCACAGGCGUGCCCCCUCGGGCGAGCACGCGAGGCGGCGAGCUAGA 3480
1108  W  R  C  R  G  W  H  G  M  P  Q  V  R  C  T  P  S  N  A  H  A  A  L  C  R  T  G  V  P  P  R  A  S  T  R  G  G  E  L  D  1147

3481 CCCAAACACCUGCUGGCUCCGCGCCGCCGCCAACGUUGCGCAGGCUGCGCGCGCCUGCGGCGCCUACACGAGUGCCGGGUGCCCCAAGUGCGCCUACGGCCGCGCCCUGAGCGAAGCCCG 3600
1148  P  N  T  C  W  L  R  A  A  A  N  V  A  Q  A  A  R  A  C  G  A  Y  T  S  A  G  C  P  K  C  A  Y  G  R  A  L  S  E  A  R  1187

3601 CACUCAUGAGGACUUCGCCGCGCUGAGCCAGCGGUGGAGCGCGAGCCACGCCGAUGCCUCCCCUGACGGCACCGGAGAUCCCCUCGACCCCCUGAUGGAGACCGUGGGAUGCGCCUGUUC 3720
1188  T  H  E  D  F  A  A  L  S  Q  R  W  S  A  S  H  A  D  A  S  P  D  G  T  G  D  P  L  D  P  L  M  E  T  V  G  C  A  C  S  1227

3721 GCGCGUGUGGGUCGGCUCCGAGCAUGAGGCCCCGCCCGACCACCUCCUGGUGUCCCUUCACCGUGCCCCAAAUGGUCCGUGGGGCGUAGUGCUCGAGGUGCGUGCGCGCCCCGAGGGGGG 3840
1228  R  V  W  V  G  S  E  H  E  A  P  P  D  H  L  L  V  S  L  H  R  A  P  N  G  P  W  G  V  V  L  E  V  R  A  R  P  E  G  G  1267

3841 CAACCCCACCGGCCACUUCGUCUGCGCGGUCGGCGGCGGCCCACGCCGCGUCUCGGACCGCCCCCACCUCUGGCUUGCGGUCCCCCUGUCUCGGGGCGGUGGCACCUGUGCCGCGACCGA 3960
1268  N  P  T  G  H  F  V  C  A  V  G  G  G  P  R  R  V  S  D  R  P  H  L  W  L  A  V  P  L  S  R  G  G  G  T  C  A  A  T  D  1307

3961 CGAGGGGCUGGCCCAGGCGUACUACGACGACCUCGAGGUGCGCCGCCUCGGGGAUGACGCCAUGGCCCGGGCGGCCCUCGCAUCAGUCCAACGCCCUCGCAAAGGCCCUUACAAUAUCAG 4080
1308  E  G  L  A  Q  A  Y  Y  D  D  L  E  V  R  R  L  G  D  D  A  M  A  R  A  A  L  A  S  V  Q  R  P  R  K  G  P  Y  N  I  R  1347

4081 GGUAUGGAACAUGGCCGCAGGCGCUGGCAAGACUACCCGCAUCCUCGCUGCCUUCACGCGCGAAGACCUUUACGUCUGCCCCACCAAUGCGCUCCUGCACGAGAUCCAGGCCAAACUCCG 4200
1348  V  W  N  M  A  A  G  A  G  K  T  T  R  I  L  A  A  F  T  R  E  D  L  Y  V  C  P  T  N  A  L  L  H  E  I  Q  A  K  L  R  1387

4201 CGCGCGCGAUAUCGACAUCAAGAACGCCGCCACCUACGAGCGCCGGCUGACGAAACCGCUCGCCGCCUACCGCCGCAUCUACAUCGAUGAGGCGUUCACUCUCGGCGGCGAGUACUGCGC 4320
1388  A  R  D  I  D  I  K  N  A  A  T  Y  E  R  R  L  T  K  P  L  A  A  Y  R  R  I  Y  I  D  E  A  F  T  L  G  G  E  Y  C  A  1427

4321 GUUCGUUGCCAGCCAAACCACCGCGGAGGUGAUCUGCGUCGGUGAUCGGGACCAGUGCGGCCCACACUACGCCAAUAACUGCCGCACCCCCGUCCCUGACCGCUGGCCUACCGAGCGCUC 4440
1428  F  V  A  S  Q  T  T  A  E  V  I  C  V  G  D  R  D  Q  C  G  P  H  Y  A  N  N  C  R  T  P  V  P  D  R  W  P  T  E  R  S  1467

4441 GCGCCACACUUGGCGCUUCCCCGACUGCUGGGCGGCCCGCCUGCGCGCGGGGCUCGAUUAUGACAUCGAGGGCGAGCGCACCGGCACCUUCGCCUGCAACCUUUGGGACGGCCGCCAGGU 4560
1468  R  H  T  W  R  F  P  D  C  W  A  A  R  L  R  A  G  L  D  Y  D  I  E  G  E  R  T  G  T  F  A  C  N  L  W  D  G  R  Q  V  1507

4561 CGACCUUCACCUCGCCUUCUCGCGCGAAACCGUGCGCCGCCUUCACGAGGCUGGCAUACGCGCAUACACCGUGCGCGAGGCCCAGGGUAUGAGCGUCGGCACCGCCUGCAUCCAUGUAGG 4680
1508  D  L  H  L  A  F  S  R  E  T  V  R  R  L  H  E  A  G  I  R  A  Y  T  V  R  E  A  Q  G  M  S  V  G  T  A  C  I  H  V  G  1547

4681 CAGAGACGGCACGGACGUUGCCCUGGCGCUGACACGCGACCUCGCCAUCGUCAGCCUGACCCGGGCCUCCGACGCACUCUACCUCCACGAGCUCGAGGACGGCUCACUGCGCGCGUGCGGG 4800
1548  R  D  G  T  D  V  A  L  A  L  T  R  D  L  A  I  V  S  L  T  R  A  S  D  A  L  Y  L  H  E  L  E  D  G  S  L  R  A  A  G  1587

4801 GCUCAGCGCGUUCCUCGACGCCGGGGCACUGGCGGAGCUCAAGGAGGUUCCCGCUGGCAUUGACCGCGUUGUCGCCGUCGAGCAGGCACCACCACCGUUGCCGCCCGCCGACGGCAUCCC 4920
1588  L  S  A  F  L  D  A  G  A  L  A  E  L  K  E  V  P  A  G  I  D  R  V  V  A  V  E  Q  A  P  P  P  L  P  P  A  D  G  I  P  1627

4921 CGAGGCCCAAGACGUGCCGCCCUUCUGCCCCCGCACUCUGGAGGAGCUCGUCUUCGGCCGUGCCGGCCACCCCCAUUACGCGGACCUCAACCGCGUGACUGAGGGCGAACGAGAAGUGCG 5040
1628  E  A  Q  D  V  P  P  F  C  P  R  T  L  E  E  L  V  F  G  R  A  G  H  P  H  Y  A  D  L  N  R  V  T  E  G  E  R  E  V  R  1667

5041 GUACAUGCGCAUCUCGCGUCACCUGCUCAACAAGAAUCACACCGAGAUGCCCGGAACGGAACGCGUUCUCAGUGCCGUUUGCGCCGUGCGGCGCUACCGCGCGGGCGAGGAUGGGUCGAC 5160
1668  Y  M  R  I  S  R  H  L  L  N  K  N  H  T  E  M  P  G  T  E  R  V  L  S  A  V  C  A  V  R  R  Y  R  A  G  E  D  G  S  T  1707
```

FIG. 2—*Continued*

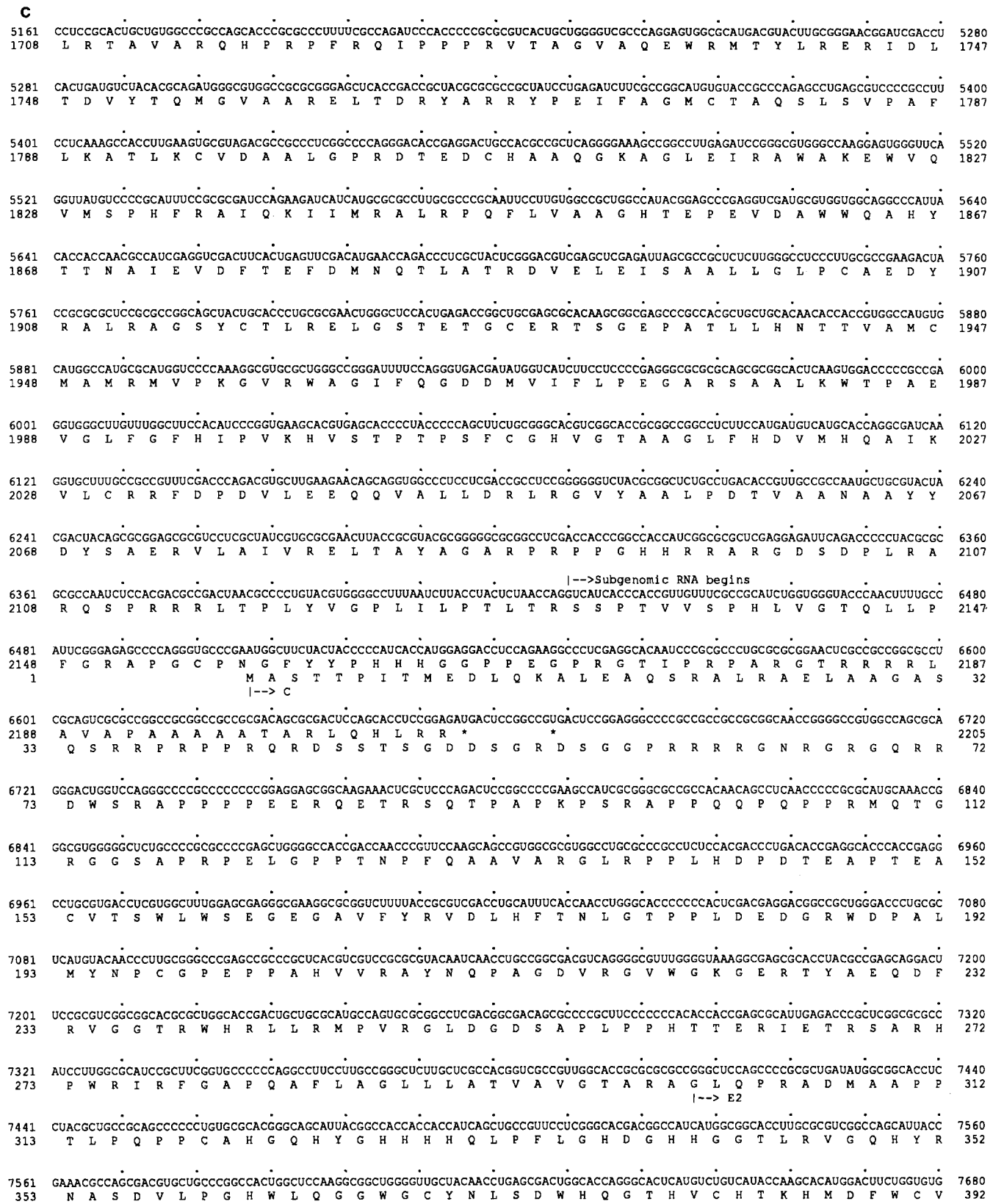

FIG. 2—*Continued*

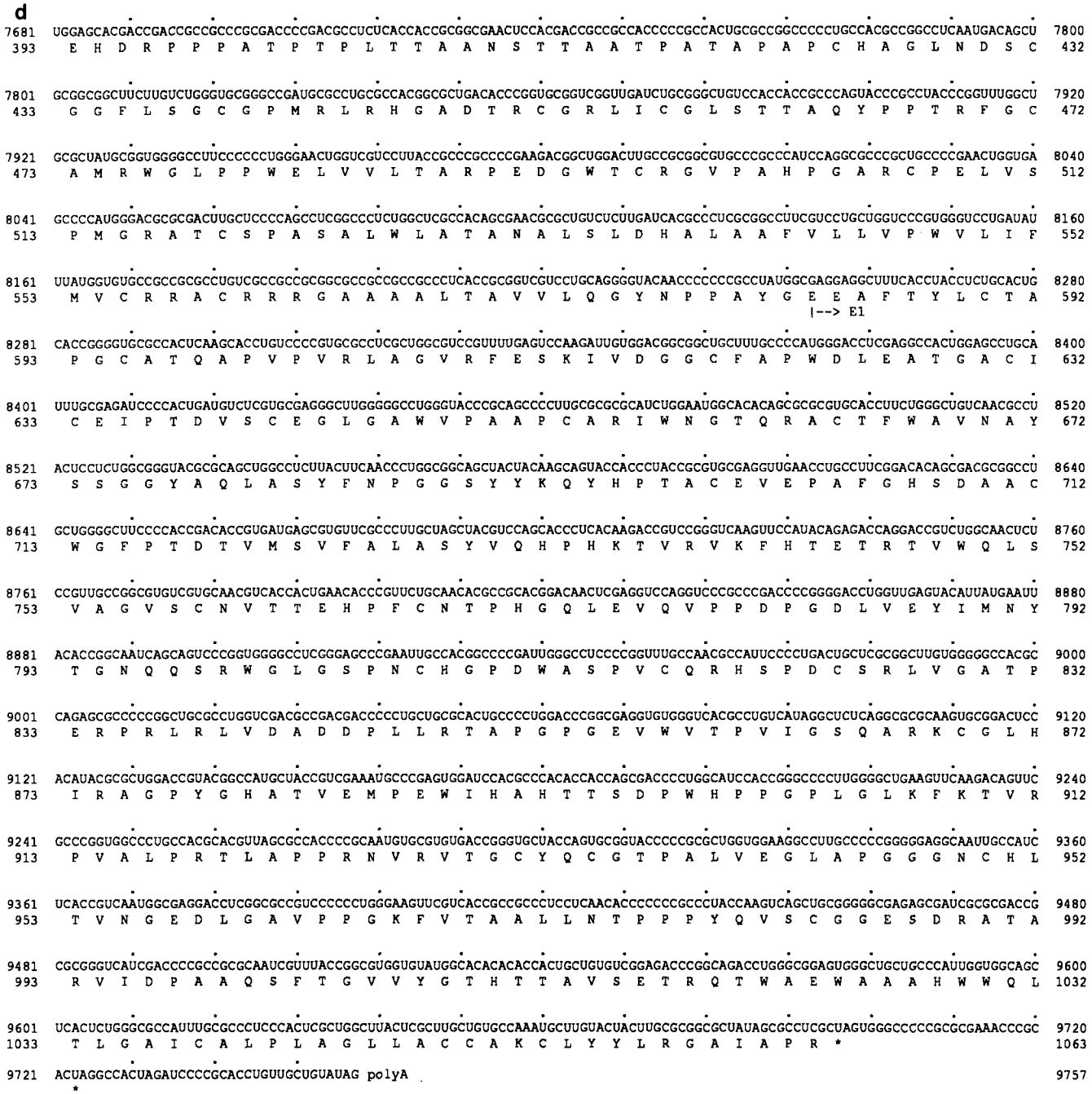

FIG. 2—*Continued*

end (reviewed by Sedman *et al.*, 1990) and thus it is a common finding that mRNAs which contain AUGs close to the 5′ end also contain downstream AUGs at which translation is also initiated (Kozak, 1987), as would be the case in the RUB RNA.

The ORF analysis in Fig. 3 revealed several long ORFs in the negative polarity. The longest of these overlaps the structural protein ORF and would yield a product of 924 amino acids if translated from the first AUG. It is unknown if any of these ORFs are translated. ORFs of similar length are not present in the negative polarity of alphavirus genomic RNAs (Strauss and Strauss, 1986).

Base composition and codon usage in the RUB genomic RNA

The base composition of the RUB genomic RNA is 14.9% A, 15.4% U, 30.8% G and 38.7% C and thus it

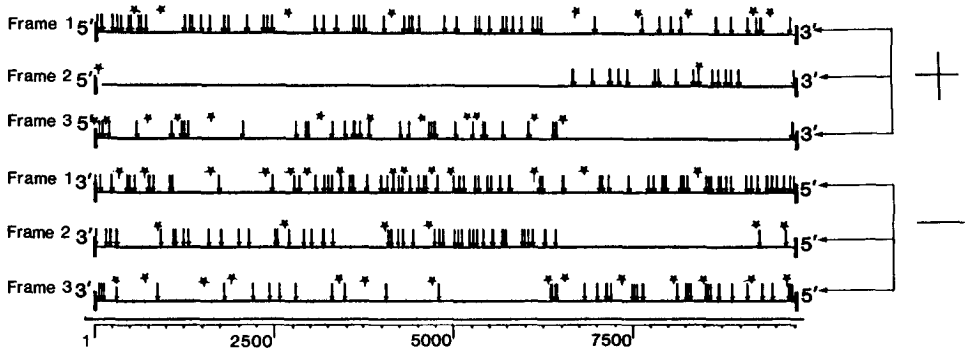

FIG. 3. Open reading frame analysis of the RUB RNA. The positions of stop codons are denoted by arrows and the first AUG in each open reading frame is denoted by a star in all six reading frames.

has the highest G/C content (69.5%) of any RNA virus sequenced. A search of other RNA virus sequences in the GenBank library revealed that G/C content among these sequences ranges between 35 and 53%, the Semliki Forest virus genome having the highest G/C content of the virus sequences analyzed. RUB is not the only virus with a G/C content greater than 60%; for example herpes simplex type 1 has a G/C content of 67% and herpes simplex type 2 has a G/C content of 69% (Roizman and Batterson, 1986). The high G/C content is evenly maintained throughout the RUB genome. The only regions where it differs markedly from the 69.5% average are the first 65 nucleotides of the genome, which have a G/C content of 49%, and 78 nucleotides around the subgenomic RNA start site, which have a G/C content of 47%. The lower G/C content in these regions would result in corresponding regions of the double-stranded replicative form having a lower T_m, possibly facilitating recognition and local denaturation by factors involved in initiation and synthesis of the genomic and subgenomic RNA.

Codon usage in the RUB nonstructural protein ORF was similar to codon usage in the structural protein ORF which was reported earlier (Frey and Marr, 1988). As expected from the high G/C content of the RUB RNA, codon usage in the RUB ORFs was skewed in comparison with eukaryotic mRNAs of lower G/C content. Two selections were particularly apparent. First, there was an extreme selection for G and C residues at the third position, resulting in a third codon position G/C content of 80.9% in the RUB ORFs in comparison to a third codon position G/C content of 54.8% in the SIN ORFs (Strauss and Strauss, 1986) and 60.8% in a compilation of human genes (Maruyama *et al.*, 1986). Second, among isofunctional amino acids encoded in the RUB ORFs, those with G/C-rich codons were favored. For example in the RUB ORFs 87% of the basic residues are arginine (codons are CGN, AGA, and AGG) and only 13% are lysine (codons are AAA and AAG), while in both the SIN ORFs and human genes 45% of the basic residues are arginine and 55% are lysine. A similar selection for valine (GUN) over isoleucine (AUC, AUA and AUU) was also present in the RUB ORFs. In contrast, for isofunctional amino acids for which no difference in G/C content is present, no selection occurs in the RUB ORFs; of the acidic residues, 50% are glutamic acid residues (GAA and GAG) and 50% are aspartic acid residues (GAC and GAU).

Comparison with alphaviruses: Amino acid homologies and global motifs

Homology comparisons generated between the predicted amino acid sequence of the RUB nonstructural protein ORF and the nonstructural protein ORFs of the six alphaviruses thus far sequenced revealed only one short region of homology. An alignment of the region of homology between RUB and a consensus sequence derived from the six alphavirus sequences (J. Strauss, personal communication) is shown in Fig. 4A. This region of homology is at the NH_2 terminus of alphavirus nsP3, a location of interest since alphavirus nsP3 is the only nonstructural protein for which homology has not been found with plant viruses (Ahlquist *et al.*, 1985) and to which no putative function has been assigned.

Two amino acid motifs have been described which are found in the proteins of diverse positive-polarity RNA viruses. The first, centered on a GDD tripeptide, has been found in all positive-polarity RNA viruses analyzed and has been associated with replicase activity due to the fact that several of the proteins in which it occurs have been demonstrated to be replicases (Kamer and Argos, 1984). More recently, a second motif has been found in proteins of alphavirus, flavivirus, coronavirus, herpesvirus, and several plant viruses that has been associated with helicase activity due to its presence in four bacterial helicases (Gorbalenya *et al.*, 1988). Both motifs were found in the RUB predicted

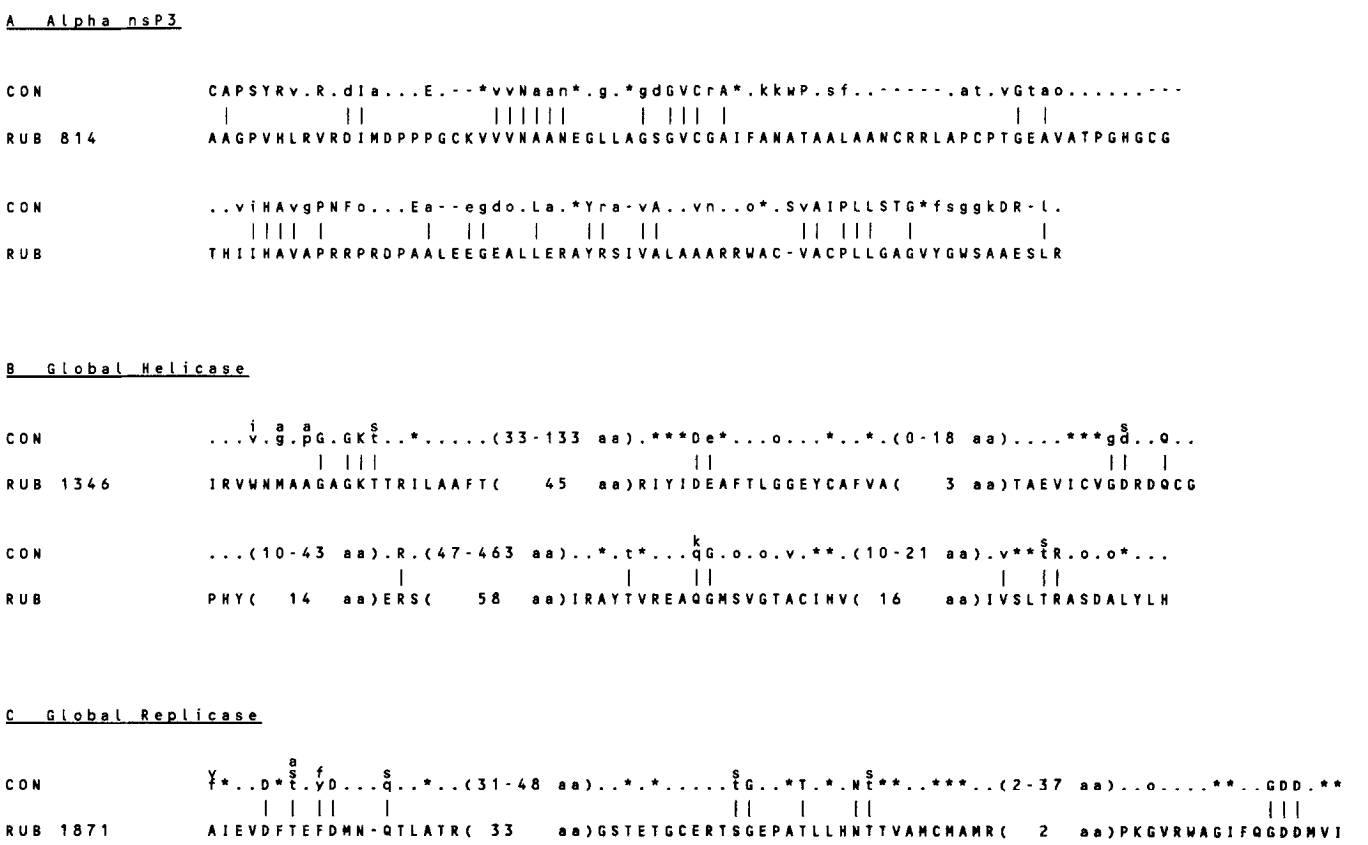

FIG. 4. Alignment of RUB amino acid sequence with homologous alphavirus amino acid sequence and global consensus motifs. (A) Alignment of the only region of significant homology in the nonstructural protein ORFs of RUB and alphaviruses detected by COMPARE amino acid homology program using a window of 19 amino acids with a stringency of 11 (Maizel and Lenk, 1981). Use of lower stringencies failed to detect further regions of homology. The CON sequence is a consensus amino acid sequence beginning at the amino terminus of nsP3 derived from the sequence of six alphaviruses (SIN, Semliki Forest, Middelburg, O'Nyong-nyong, Venezuelan equine encephalitis, and Ross River viruses) (J. Strauss, personal communication). Invariant amino acids are capitalized and those found in at least four of the six alphaviruses are denoted in lowercase. (B) The CON sequence is a motif found in four *Escherichia coli* helicases and proteins in herpesviruses and nine families of plant and animal positive-polarity RNA virus families. Presentation is as in Gorbalenya *et al.* (1988). (C) The CON sequence is a motif found in putative replicases of 11 plant and animal positive-polarity RNA viruses (Kamer and Argos, 1984). Presentation adapted from a compilation by Rice *et al.* (1986). In panels B and C, invariant amino acids are capitalized, majority amino acids are given in lowercase, and lengths of gaps in amino acids between regions of conservation are given in parentheses. In all three panels, · = nonconserved amino acid, - = gap introduced to maximize alignment, * and o = hydrophobic and hydrophilic amino acids, respectively, at a position. The number of the amino acid in the RUB nonstructural ORF at which each alignment begins is given.

nonstructural protein sequence. The alignments are both shown in Figs. 4B and C. The alignment with the replicase motif was previously reported (Frey and Marr, 1988).

Comparison with alphavirus genome: conserved nucleotide stretches

Four regions in the genomic RNAs of alphaviruses have been found to be highly conserved among alphaviruses and thus are thought to be regulatory signals for viral replication (Strauss and Strauss, 1986). The first of these is a stem-and-loop structure found at the 5′ end which is conserved despite divergence of the nucleotides making up the structures. The complementary minus-strand equivalent of this structure is hypothesized to be recognized by the replicase as a binding site for initiation of transcription of the plus-strand genome-length RNA. The RUB genomic RNA has a similar structure (RV-2) at its 5′ end which has a calculated stability similar to that of the alphavirus structures (Fig. 5). Although this structure was originally found by eye, it was also found by the RNA secondary structure analysis program FOLD (Zuker and Stiegler, 1981). Interestingly, a second double-stem-loop structure (RV-1) can be formed by the same sequences. This structure has a lower ΔG, but would be formed first as the newly synthesized genomic RNA is released from the negative-polarity template. Whether this stability is great enough to prevent the more stable stem-and-loop structure from forming under physiological condi-

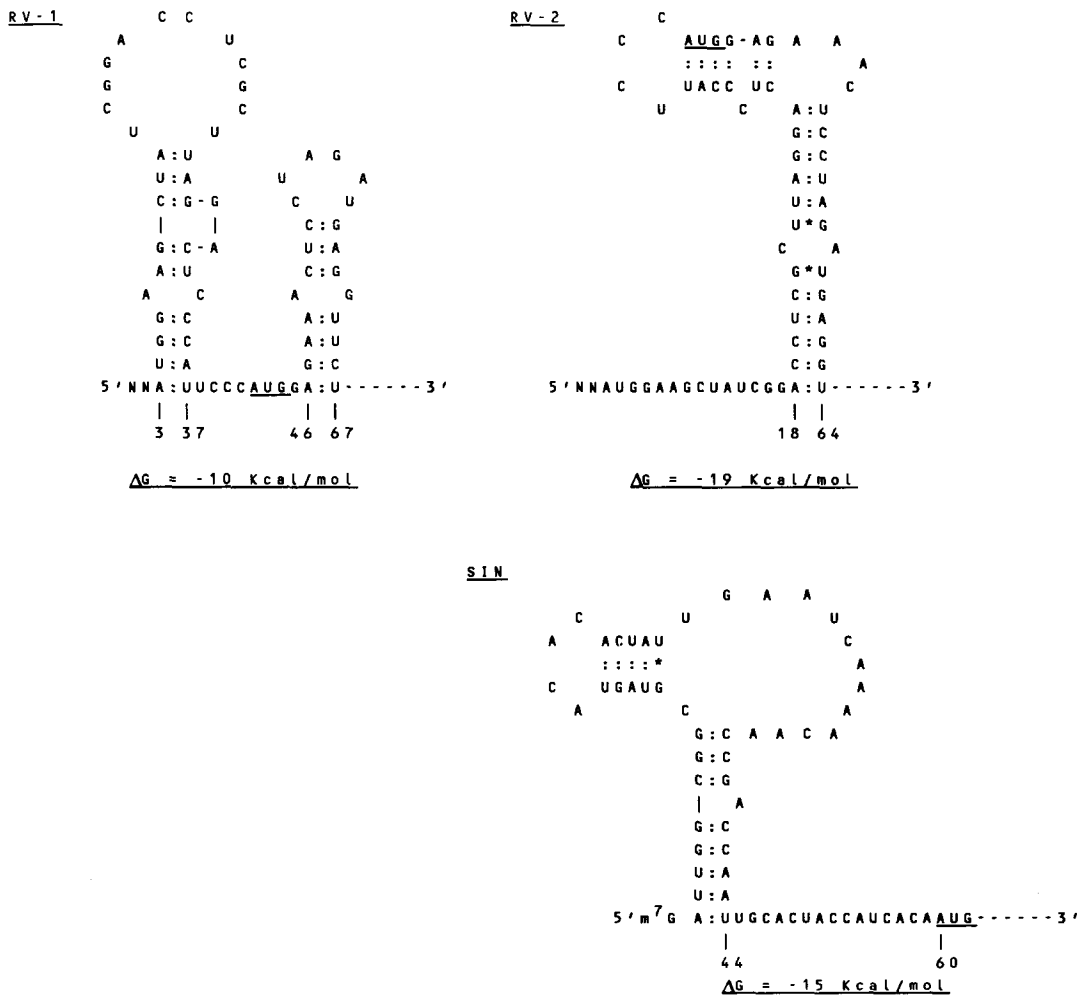

FIG. 5. Potential stem-and-loop structures formed by nucleotides at the 5′ terminus of the RUB RNA. The conserved stem-and-loop structures at the 5′ termini of alphavirus RNAs are represented by the structure at the 5′ terminus of the SIN RNA (Strauss and Strauss, 1986). Free energies were calculated by the method of Tinoco *et al.* (1973). The AUG beginning the nonstructural protein ORF is underlined.

tions is unknown. Primer extension experiments indicated that the more stable stem-and-loop structure (RV-2) could be formed when denatured RUB genome RNA was renatured *in vitro.* Poly(A)+ RNA from RUB-infected cells was mixed with a 5′-^{32}P-labeled oligonucleotide primer with sequence complementary to nucleotides 101 to 117 of the RUB genome, heated to 85°, cooled slowly, and added to a reverse transcription reaction mixture. Electrophoresis of the primer extension products revealed strong stop bands at both the beginning of the RV-2 stem-and-loop structure and the 5′ end of the genome (data not shown).

The second conserved region in the genomic RNAs of alphaviruses is a stretch of 51 nucleotides located 156 nucleotides from the 5′ end. A double stem-and-loop secondary structure can be formed by these nucleotides. A stretch of 46 nucleotides located 224 nucleotides from the 5′ end of the RUB genome has 50% overall homology with the alphavirus sequence, including two stretches in which nine of ten nucleotides match (Fig. 6A). No stable secondary structure can be configured from this stretch of RUB nucleotides. Interestingly, the RUB and alphavirus sequences in this region are in-frame translationally yielding a small pocket of amino acid homology. The function of this region of the alphavirus genome is not clear. It is present in naturally occurring DI RNAs (and is sometimes duplicated) but can be deleted from infectious DI clones, although the ability of RNAs transcribed from such clones to be replicated is less than that of RNAs transcribed from clones containing the sequence (Levis *et al.*, 1986; Tsiang *et al.*, 1988). Conservation of this sequence at similar locations in the genomes of viruses as diverse as RUB and alphaviruses argues that it does play an important role in the replication of these viruses. The lack of a predicted secondary structure for the RUB version of this nucleotide sequence indicates that the double hairpin structure is not necessary for function. It seems

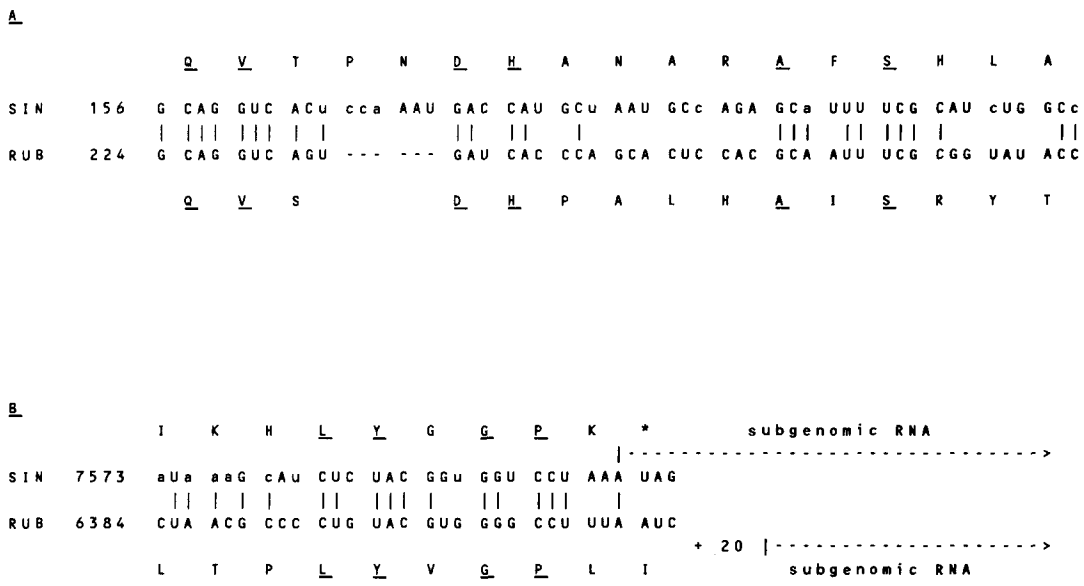

FIG. 6. Alignment of RUB sequences with two sequences conserved among all alphaviruses. (A) The 51-nucleotide conserved sequence, (B) the subgenomic RNA start site (junction region) conserved sequence. For simplicity, the content of these sequences within the prototype SIN RNA is shown. The nucleotides found in all alphavirus RNAs are capitalized. The amino acids encoded by these sequences within the nonstructural ORF are shown; amino acids which are identical in SIN and RUB are underlined.

unlikely that maintenance of the amino acid sequence is the selective pressure maintaining the nucleotide sequence since several of the conserved nucleotides occur at the third codon position.

Shown in Fig. 6B is the third stretch of nucleotides highly conserved among alphaviruses located at the start site of subgenomic RNA synthesis. This region has been aligned with a homologous sequence in the RUB genome (Frey and Marr, 1988). This sequence, which occurs immediately adjacent to the subgenomic start site in alphaviruses, has been shown experimentally to be necessary for alphavirus subgenomic RNA synthesis (Grakoui *et al.,* 1989). The homologous sequence in the RUB genome is 20 nucleotides upstream from the RUB subgenomic RNA start site (Frey *et al.,* 1989). This RUB sequence can form a secondary structure while the comparable alphavirus sequence cannot. Interestingly, these sequences are also in the same translation frame, giving rise to another small pocket of amino acid homology. Whether selective pressure on conservation of this sequence is due to its function in subgenomic RNA synthesis or its coding capacity can be investigated by deletion mutagenesis once an infectious RUB clone is developed. As discussed earlier (Frey *et al.,* 1986), no region in the RUB genome has been found to be homologous with the fourth alphavirus conserved region, the 3′-terminal 19 nucleotides.

Similarities and differences in the genomes of RUB and alphaviruses: evidence for rearrangement during Togavirus evolution

Figure 7 shows a comparative diagram of the genomes of RUB and alphaviruses (represented by the prototype SIN), including the locations of the various regions of nucleotide and amino acid homology discussed previously. While the genetic organizations of these two virus genera are similar, two points of difference are of interest. First, the RUB genome (9757 nucleotides) is 1946 nucleotides shorter than the SIN genome (11,703 nucleotides). The compression of the RUB genome with respect to the SIN genome is exhibited throughout the RUB genome in that both the nonstructural protein ORF (2205 amino acids in RUB; 2513 amino acids in SIN) and the structural protein ORF (1063 amino acids in RUB; 1245 amino acids in SIN), as well as the 3′ nontranslated sequence (62 nucleotides in RUB; 319 nucleotides in SIN), are shorter in the RUB genome than in the SIN genome. In the RUB genome the two ORFs overlap by 149 nucleotides while in the alphavirus genome the two ORFs do not overlap. The difference in the structural protein ORFs is due to the lack of a 6K protein encoded by SIN and differences in the size of the other structural proteins. In this regard, it should be pointed out that processing of the structural proteins of these viruses differs in that the two proteolytic cleavages in the RUB structural protein precursor are putatively mediated by signalase in the lumen of the endoplasmic reticulum (Frey and Marr, 1988), while the SIN capsid protein is autocatalytically removed from the precursor and cleavages on either side of the 6K protein are mediated by signalase (Strauss and Strauss, 1986). These cleavages leave a precursor (PE2) which is processed into E2 and E3 in the Golgi. No such cleavage occurs in the maturation

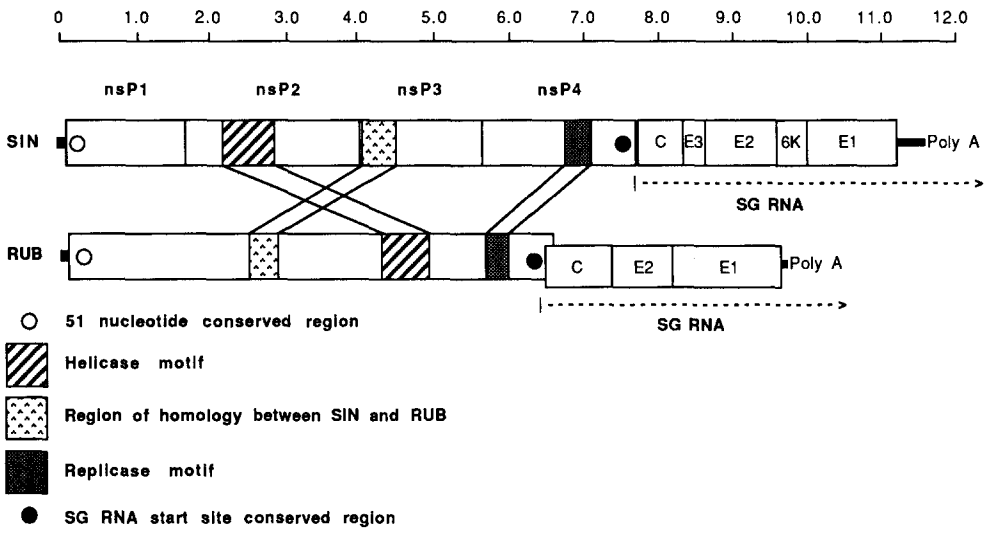

FIG. 7. Diagram of the genomes of RUB and SIN. The locations within the nonstructural protein ORF of regions of nucleotide homology (51-nucleotide conserved region and SG RNA start site conserved region) and regions encoding homologous amino acid sequence (helicase motif, SIN and RUB homology, and replicase motif) are shown. ORFs are denoted by boxes and untranslated regions by lines. The depression of the structural protein ORF relative to the nonstructural protein ORF in the RUB genome reflects the fact that these ORFs overlap and are in different translational frames. The SIN ORFs do not overlap and are in the same translational frame.

of RUB E2. Not enough is known about the RUB nonstructural protein ORF to account for the size difference with the SIN nonstructural protein ORF.

The second point of difference in the two genomes is that while most of the regions of homology are in the same relative position within the two genomes, the order of the helicase domain and nsP3 homology region is reversed. Thus a genetic rearrangement has occurred in the nonstructural protein coding region during Togavirus evolution.

From analysis of amino acid homologies among positive-polarity RNA viruses of animals and plants, it has become apparent that diverse viruses are distantly related and that two superfamilies exist to which many positive-polarity RNA viruses belong: a picorna virus-like superfamily and an alpha virus-like superfamily (Goldbach, 1986, 1987, 1990). RUB belongs to the latter of these. Among members of the alphavirus-like superfamily, some domains such as the helicase and replicase motifs are conserved among all members, while other domains such as the plant virus transport proteins are restricted to a few members. The relative location of the conserved domains thus far distinguished has been found to be similar in the genomes of the viruses in which they occur (with the caveat that some of these domains occur on different segments of di- and tripartite plant virus genera). Thus, the reversed order of the helicase and nsP3 domains in RUB and alphavirus genomes is unique. Two events which could create the reversed order of the helicase and nsP3 domains in RUB can be hypothesized. First, a complicated rearrangement could have occurred within the same genome during either RUB or alphavirus evolution from a common progenitor. Second, the nsP3 domain could have been donated by another virus to a common Togavirus ancestor by interviral recombination occurring twice independently: to the 3′ side of the helicase domain in the alphavirus lineage and to the 5′ side of the helicase domain in the rubivirus lineage. As more functional domains of the nonstructural proteins of these viruses are characterized, more information about the interesting evolution of these genera will be elucidated.

ACKNOWLEDGMENTS

We thank Lee D. Marr for expert technical assistance, Anthony Sanchez and Cynthia A. Derdeyn for help with some of the experiments, Phil Pellett for assistance with GCG programs and critical reading of the manuscript, John Houghton for assistance with graphics, Margo A. Brinton for critical reading of the manuscript, and Sarah McKneally for VAX system management at CDC. This research was supported by Grant A1-21389 from NIH. T.K.F. is the recipient of a Research Career Development Award (AI-00923) from NIH.

REFERENCES

AHLQUIST, P., and JANDA, M. (1984). cDNA cloning and *in vitro* transcription of the complete brome mosaic virus genome. *Mol. Cell. Biol.* **4,** 2876–2882.

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 nonstructural proteins from several RNA plant viruses. *J. Virol.* **53,** 536–542.

BANKIER, A. T., and BARRELL, B. G. (1983). Shotgun DNA sequencing. Techniques in the Life Sciences, B5. *Nucleic Acid Biochem.* **B508,** 1–34.

CLARKE, D. M., LOO, T. W., HUI, I., CHONG, P., and GILLAM, S. (1987). Nucleotide sequence and *in vitro* expression of rubella virus 24 S subgenomic messenger RNA encoding the structural proteins E1, E2, and C. *Nucl. Acids Res.* **15,** 3041–3057.

DEVEREUX, J., HAEBERLI, P., and SMITHIES, O. (1984). A comprehensive set of sequence analysis programs for the VAX. *Nucl. Acids Res.* **12,** 387–395.

DING, M., and SCHLESINGER, M. J. (1989). Evidence that Sindbis virus nsP2 is an autoprotease which processes the virus nonstructural polyprotein. *Virology* **171,** 280–284.

ESCHENFELDT, W. H., and BERGER, S. L. (1987). Purification of large double-stranded cDNA fragments. *In* "Methods in Enzymology" (S. L. Berger and A. R. Kimmel, Eds.), Vol 152, 335–337. Academic Press, New York.

FREY, T. K., and MARR, L. D. (1988). Sequence of the region coding for virion proteins C and E2 and the carboxy terminus of the nonstructural proteins of rubella virus: Comparison with alphaviruses. *Gene* **62,** 85–99.

FREY, T. K., MARR, L. D., HEMPHILL, M. L., and DOMINGUEZ, G. (1986). Molecular cloning and sequencing of the region of the rubella virus genome coding for glycoprotein E1. *Virology* **154,** 228–232.

FREY, T. K., MARR, L. D., SANCHEZ, A., and SIMMONS, B. (1989). Identification of the 5′ end of the rubella virus subgenomic RNA. *Virology* **168,** 191–194.

GOLDBACH, R. W. (1986). Molecular evolution of plant RNA viruses. *Annu. Rev. Phytopathol.* **24,** 289–310.

GOLDBACH, R. W. (1987). Genomic similarities between plant and animal RNA viruses. *Microbiol. Sci.* **4,** 197–202.

GOLDBACH, R. W. (1990). Genome similarities between positive strand RNA viruses from plants and animals. *In* "New Aspects of Positive Strand RNA Viruses" (M. A. Brinton and F. X. Heinz, Eds.), ASM Press, Washington, DC.

GORBALENYA, A. E., KOONIN, E. V., DONCHENKO, A. P., and BLINOV, V. M. (1988). A novel superfamily of nucleotide triphosphate-binding motif containing proteins which are probably involved in duplex unwinding in DNA and RNA replication and recombination. *FEB Lett.* **235,** 16–24.

GRAKOUI, A., LEVIS, R., RAJU, R., HUANG, H. V., and RICE, C. M. (1989). A *cis*-acting mutation in the Sindbis virus junction region which affects subgenomic RNA synthesis. *J. Virol.* **63,** 5216–5227.

GUPTA, K. C., and KINGSBURY, D. W. (1984). Complete sequences of the intergenic and mRNA start signals in the Sendai virus genome: Homologies with the genome of vesicular stomatitis virus. *Nucl. Acids Res.* **12,** 3829–3841.

HAHN, Y. S., GRAKOUI, A., RICE, C. M., STRAUSS, E. G., and STRAUSS, J. H. (1989a). Mapping of RNA$^-$ temperature-sensitive mutants of Sindbis virus: Complementation group F mutants have lesions in nsP4. *J. Virol.* **63,** 1194–1202.

HAHN, Y. S., STRAUSS, E. G., and STRAUSS, J. H. (1989b). Mapping of RNA$^-$ temperature-sensitive mutants of Sindbis virus: Assignment of complementation groups A, B, and G to nonstructural proteins. *J. Virol.* **63,** 3142–3150.

HARDY, W. R., and STRAUSS, J. H. (1989). Processing 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.

HEMPHILL, M. L., FORNG, R. Y., ABERNATHY, E. S., and FREY, T. K. (1988). Time-course of virus-specific macromolecular synthesis in rubella virus infected Vero cells. *Virology* **162,** 65–75.

HENIKOFF, S. (1984). Unidirectional digestion with exonuclease III creates targeted breakpoints for DNA sequencing. *Gene* **28,** 351–359.

HO-TERRY, L., and COHEN, A. (1982). Rubella virion polypeptides: Characterization by polyacrylamide gel electrophoresis, isoelectric focusing and peptide mapping. *Arch. Virol.* **72,** 47–54.

KALKKINEN, N., OKER-BLOM, C., and PETTERSSON, R. F. (1984). Three genes code for rubella virus structural proteins E1, E2a, E2b and C. *J. Gen. Virol.* **65,** 1549–1557.

KAMER, G., and ARGOS, P. (1984). Primary structural comparison of RNA-dependent polymerases from plant, animal, and bacterial viruses. *Nucl. Acids Res.* **12,** 7269–7282.

KELLEY, D. E., COLECLOUGH, C., and PERRY, R. P. (1982). Functional significance and evolutionary development of the 5′-terminal regions of immunoglobulin variable-region genes. *Cell* **29,** 681–689.

KOZAK, M. (1987). An analysis of 5′-noncoding sequences from 699 vertebrate messenger RNAs. *Nucl. Acids Res.* **15,** 8125–8148.

LEVIS, R., WEISS, B. G., and TSIANG, M. (1986). Deletion mapping of Sindbis virus DI RNAs derived from cDNAs defines the sequences essential for replication and packaging. *Cell* **44,** 137–145.

MAIZEL, J. V., and LENK, R. P. (1981). Enhanced graphic matrix analysis of nucleic acid and protein sequences. *Proc. Natl. Acad. Sci. USA* **78,** 7665–7669.

MARUYAMA, T., GOJOBORI, T., AOTA, S., and IKEMURA, T. (1986). Codon usage tabulated from the GenBank genetic sequence data. *Nucl. Acids Res.* **14,** r151–r197.

MATTHEWS, R. E. F. (1982). Classification and nomenclature of viruses. *Intervirology* **17,** 1–199.

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.

OKER-BLOM, C. (1984). The gene order for rubella virus structural proteins is NH_2–C–E2–E1–COOH. *J. Virol.* **51,** 354–358.

OKER-BLOM, C., KALKKINEN, N., KAARIAINEN, L., and PETTERSSON, R. F. (1983). Rubella virus contains one capsid protein and three envelope glycoproteins E1, E2a, and E2b. *J. Virol.* **46,** 964–973.

OKER-BLOM, C., ULMANEN, I., KAARIAINEN, L., and PETTERSSON, R. F. (1984). Rubella virus 40 S genome RNA specifies a 24 S subgenomic mRNA that codes for a precursor to structural proteins. *J. Virol.* **49,** 403–408.

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.

RICE, C. M., STRAUSS, E. G., and STRAUSS, J. H. (1986). Structure of the flavivirus genome. *In* "The Togaviridae and Flaviviridae" (S. Schlesinger and M. J. Schlesinger, Eds.), pp. 279–326. Plenum, New York.

RICO-HESSE, R., PALLANSCH, M. A., NOTTAY, B. K., and KEW, O. M. (1987). Geographic distribution of wild poliovirus type 1 genotypes. *Virology* **160,** 311–322.

ROIZMAN, B., and BATTERSON, W. (1986). Herpesviruses and their replication. *In* "Fundamental Virology" (B. N. Fields and D. M. Knipe, Eds.), pp. 607–636. Raven Press, New York.

SANGER, F., NICKLEN, S., and COULSON, A. R. (1977). DNA sequencing with chain terminating inhibitors. *Proc. Natl. Acad. Sci. USA* **74,** 5463–5467.

SEDMAN, S. A., GELEMBIUK, G. W., and MERTZ, J. E. (1990). Translation initiation at a downstream AUG occurs with increased efficiency when the upstream AUG is located very close to the 5′ cap. *J. Virol.* **64,** 453–457.

STRAUSS, E. G., LEVINSON, R., RICE, C. M., DALRYMPLE, J., and STRAUSS, J. H. (1988). Nonstructural proteins nsP3 and nsP4 of Ross River and O'Nyong-nyong viruses: Sequence and comparison with those of other alphaviruses. *Virology* **164,** 265–274.

STRAUSS, E. G., RICE, C. M., and STRAUSS, J. H. (1983). Sequence coding for the alphavirus nonstructural proteins is interrupted by an opal termination codon. *Proc. Natl. Acad. Sci. USA* **80,** 5271–5275.

STRAUSS, E. G., and STRAUSS, J. H. (1986). Structure and replication of the

alphavirus genome. *In* "The Togaviridae and Flaviviridae" (S. Schlesinger and M. J. Schlesinger, Eds.), pp. 35–90. Plenum, New York.

TAKKINEN, K., VIDGREN, G., ULF-HELLMAN, J. E., KALKKINEN, N., WERNSTEDT, C., and PETTERSSON, R. F. (1988). Nucleotide sequence of the rubella virus capsid protein gene reveals an unusually high G/C content. *J. Gen. Virol.* **69,** 603–612.

TINOCO, I., BORER, P. N., DENGLER, B., LEVINE, M. D., UHLENBECK, O. C., CROTHERS, D. M., and GRALLA, J. (1973). Improved estimation of secondary structure in ribonucleic acids. *Nature New Biol.* **246,** 40–41.

TSIANG, M., WEISS, B. G., and SCHLESINGER, S. (1988). Effects of 5′-terminal modifications on the biological activity of defective interfering RNAs of Sindbis virus. *J. Virol.* **62,** 47–53.

VIDGREN, G., TAKKINEN, K., KALKKINEN, N., KAARIAINEN, L., and PETTERSSON, R. F. (1987). Nucleotide sequence of the genes coding for the membrane glycoproteins E1 and E2 rubella virus. *J. Gen. Virol.* **68,** 2347–2357.

WAXHAM, M. N., and WOLINSKY, J. S. (1983). Immunochemical identification of rubella virus hemagglutinin. *Virology* **126,** 194–203.

ZIMMERN, D., and KAESBERG, P. (1978). 3′-Terminal nucleotide sequence of encephalomyocarditis virus RNA determined by reverse transcriptase and chain terminating inhibitors. *Proc. Natl. Acad. Sci. USA* **75,** 4257–4261.

ZUKER, M., and STIEGLER, P. (1981). Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucl. Acids Res.* **9,** 133–148.