# Nucleotide Sequence of Yellow Fever Virus: Implications for Flavivirus Gene Expression and Evolution

Charles M. Rice, Edith M. Lenches, Sean R. Eddy
Se Jung Shin, Rebecca L. Sheets, James H. Strauss

Abstract. *The sequence of the entire RNA genome of the type flavivirus, yellow fever virus, has been obtained. Inspection of this sequence reveals a single long open reading frame of 10,233 nucleotides, which could encode a polypeptide of 3411 amino acids. The structural proteins are found within the amino-terminal 780 residues of this polyprotein; the remainder of the open reading frame consists of nonstructural viral polypeptides. This genome organization implies that mature viral proteins are produced by posttranslational cleavage of a polyprotein precursor and has implications for flavivirus RNA replication and for the evolutionary relation of this virus family to other RNA viruses.*

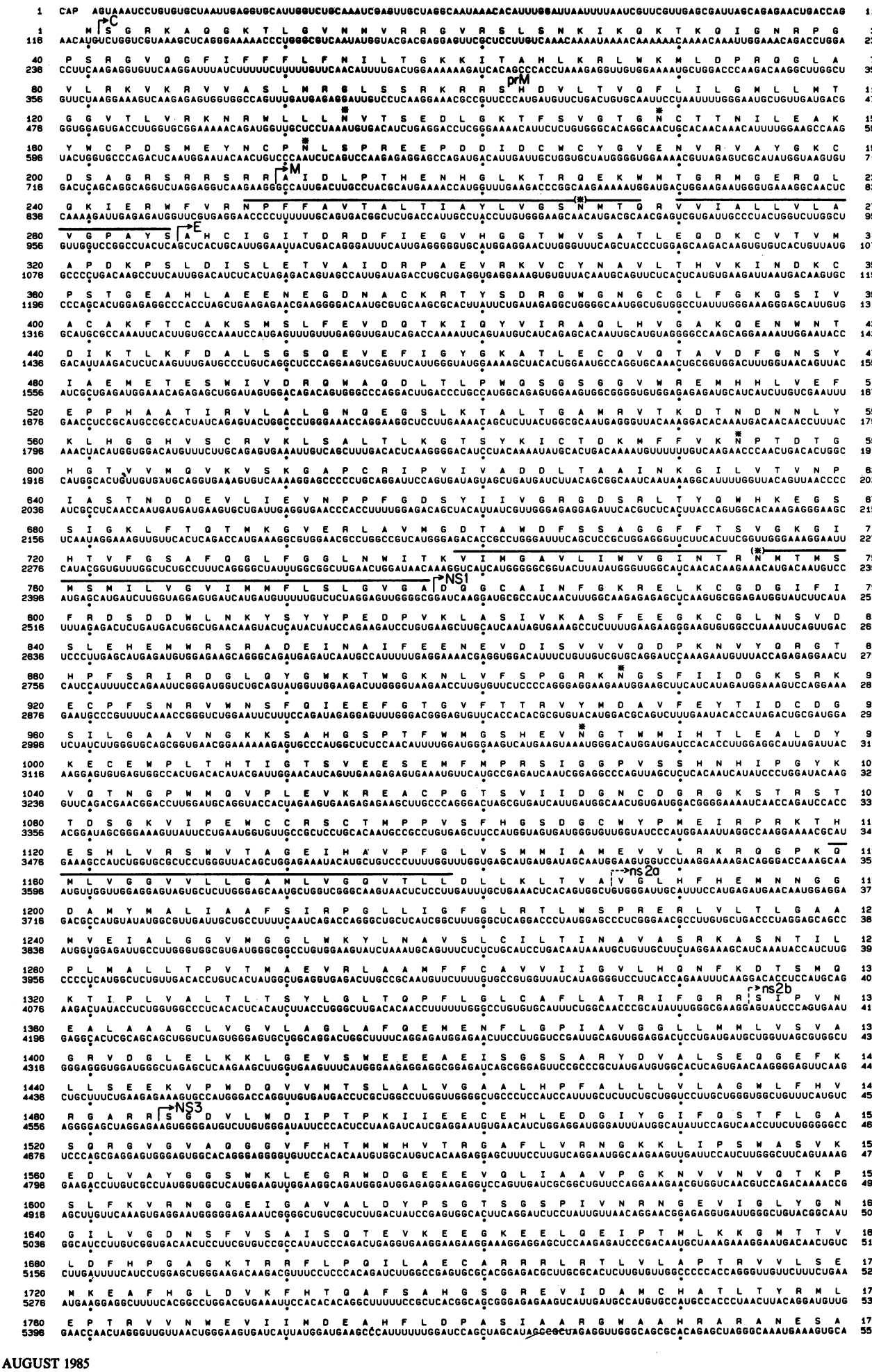
The *Flavivirus* genus, family Flaviviridae, consists of a group of some 70 closely related human or veterinary pathogens causing many serious illnesses, including dengue fever, Japanese encephalitis, St. Louis encephalitis, Murray Valley encephalitis, tick-borne encephalitis, and yellow fever (*1*). Most flaviviruses are transmitted to vertebrate hosts by blood-sucking arthropods, mosquitoes or ticks, although some evidently lack an arthropod vector (*2*). Arthropod-transmitted flaviviruses replicate in the arthropod host as well as the vertebrate host. Human flavivirus diseases have diverse and complex pathologies and different viruses exhibit marked tissue tropisms. Many are neurotropic, causing encephalitic symptoms; others, such as the dengue group, replicate preferentially in host macrophages, whereas yellow fever is usually viscerotropic.

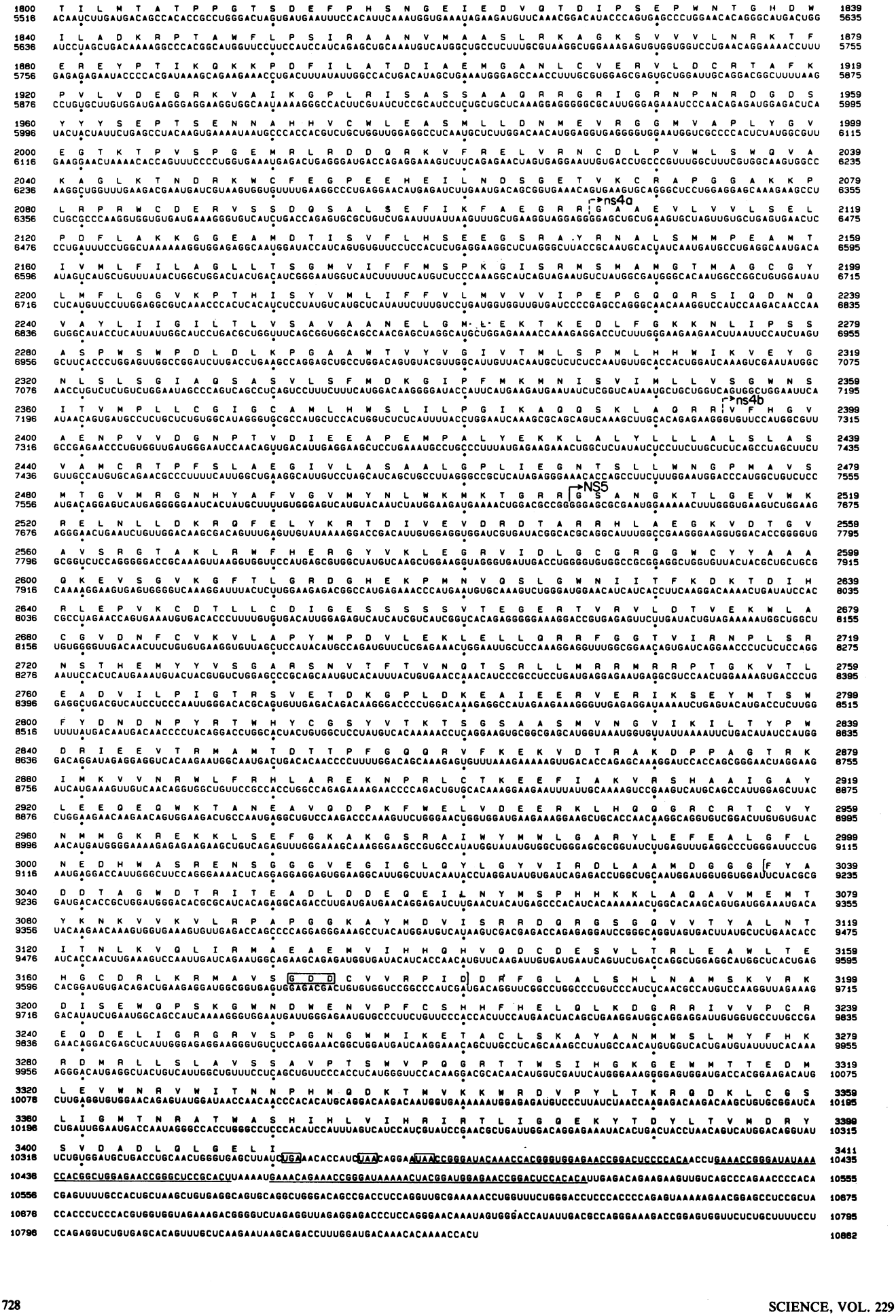
The disease known as yellow fever has been recognized for several hundred years (*3, 4*). Until the early 1900's recurrent epidemics occurred in the Caribbean area which caused great human suffering and had a profound influence on human activities in the area. From its focus in the Caribbean, epidemic yellow fever was spread by ship to ports as far north as Boston and as far east as England, where mortality rates in an epidemic could exceed 20 percent of those contracting the disease. Walter Reed and colleagues in pioneering studies in Cuba in 1900 demonstrated that yellow fever is transmitted by mosquitoes, and 2 years later showed that the disease agent is filterable (*5*). With the recognition that the mosquito *Aedes aegypti* is the vector for urban yellow fever, mosquito control measures rapidly led to the elimination of urban yellow fever. Subsequently, a safe and effective attenuated vaccine strain (17D) was developed by in vitro passage of the virulent Asibi strain in chicken embryo tissue (*6*). However, the virus persists in a sylvan cycle in the forests of South America and Africa, transmitted by numerous mosquito species including those of the genus *Haemagogus* in South America and of the genus *Aedes* in Africa. The vertebrate hosts in this cycle appear to be almost exclusively primates, demonstrating the limited natural host range of yellow fever. From the sylvan cycle periodic outbreaks in neighboring human populations have arisen on both continents. Furthermore, since *Aedes aegypti* is widespread in the world, a situation exacerbated by relaxation of mosquito abatement procedures in the Caribbean and elsewhere, the potential exists for future epidemics of urban yellow fever.

C. M. Rice, E. M. Lenches, and J. H. Strauss are members of the Division of Biology, California Institute of Technology, Pasadena 91125. S. R. Eddy and S. J. Shin are students at the California Institute of Technology and R. L. Sheets is doing graduate work in the Department of Cellular, Viral and Molecular Biology, University of Utah, University Medical Center, Salt Lake City 84132.

Previous studies have shown that flaviviruses contain single-stranded infectious RNA (thus defining them as plus-stranded RNA viruses in which the virion RNA serves as a messenger) encapsidated in a nucleocapsid possessing icosahedral symmetry and containing a single species of capsid protein [C, apparent mass of about 14 kilodaltons (kD)]. This in turn is surrounded by a lipid bilayer containing an envelope protein (E; about 50 to 60 kD) that is usually but not invariably glycosylated (*7*) and a second, nonglycosylated protein (M; about 8 kD) (*8, 9*). How the envelope is obtained is unclear, as budding flaviviruses are seldom identified in electron microscopic studies, although maturation does appear to occur in association with intracellular membranes (*9, 10*). Replication of flaviviruses in tissue culture is slow, with a long latent period, and only moderate titers of virus are produced. Host cell protein and RNA synthesis are shut off only poorly (vertebrate cells) or not at all (mosquito cells), making study of flavivirus replication and structure somewhat more difficult. Virus-specific protein synthesis appears to be associated with the rough endoplasmic reticulum, and RNA replication is localized in the perinuclear region (*11*). No subgenomic RNA has been detected in cells infected with flaviviruses, and it is believed that the genomic length RNA which is capped but not polyadenylated (*12, 13*) is the only messenger RNA (mRNA) species (*9, 12, 14*). This mRNA is translated into the three structural proteins and several nonstructural proteins. Translation of the flavivirus genome in vitro produces polypeptides related to the structural proteins (*15*) which, in the presence of appropriate membrane fractions, can be processed efficiently to yield C and E (*16*). Peptide mapping of in vitro translation products as well as selective incorporation of *N*-formylmethionine suggest that initiation in vitro occurs only with the capsid protein. Alternatively, studies on the in vivo translation of flavivirus Kunjin have been based on the use of pactamycin or high salt inhibition of translation initiation (*17*) or ultraviolet inactivation of translation (*18*) in an attempt to map the genome order of flavivirus proteins on the assumption that there is just a single site for initiation of translation. These experiments have led Westaway and collaborators to suggest that multiple independent translation initiation sites are used within flavivirus RNA, a situation not typically found with other eukaryotic mRNA's (*19*).

We now present the complete nucleo-


C
prM
M
E
NS1
ns2a
ns2b
NS3


ns4a
ns4b
NS5

tide sequence of the yellow fever genome determined from complementary DNA (cDNA) clones of the 17D vaccine strain. Together with recent $NH_2$-terminal sequence analysis of both structural (*20*) and some nonstructural yellow fever proteins, the amino acid sequences of the encoded proteins have been deduced and a preliminary picture of flavivirus gene organization and expression has begun to emerge.

*Sequence of yellow fever RNA.* The complete sequence of yellow fever RNA is shown in Fig. 1. The 5′- and 3′-terminal sequences presented were derived from several independent clones, are homologous to the 5′ and 3′ termini of West Nile flavivirus genomic RNA (*21*) (see below), and thus probably reflect the extreme ends of the yellow fever genome. Given these assumptions, the RNA genome is 10,862 nucleotides in length and has a mass of $3.75 \times 10^6$ daltons (expressed as the sodium form). Previous reports have shown that flavivirus genomic RNA contains a type 1 cap at the 5′ terminus but lacks a polyadenylate tract at the 3′ terminus (*12, 13*). The base composition of the RNA is 27.3 percent A, 23.0 percent U, 28.4 percent G, and 21.3 percent C.

It is striking that the RNA contains an extremely long open reading frame, which spans virtually the entire length of the genome. This open reading frame, beginning from the first AUG triplet, is 10,233 nucleotides in length, terminating with a single opal codon (UGA), and could encode a polypeptide of 380,763 daltons, leaving 5′- and 3′-noncoding regions of 118 and 511 nucleotides, respectively. Examination of the remaining five possible reading frames (two in the virion RNA and three in the complementary RNA) reveals multiple stop codons in every case, with the longest possible other open reading frame being 804 nucleotides (in the complementary strand). Thus there is no reason to expect that any protein is translated from yellow fever RNA other than the polyprotein encoded by the long open reading frame shown in Fig. 1.

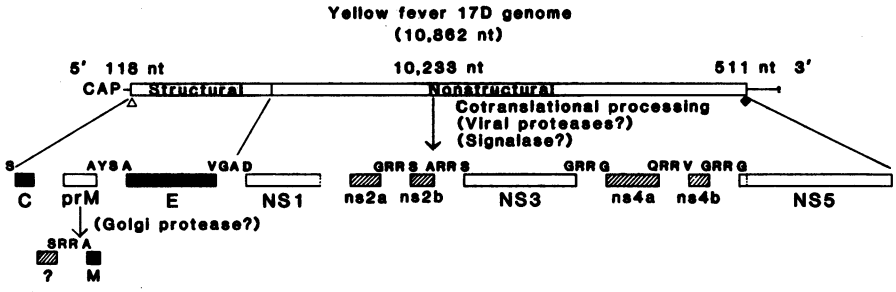



Fig. 2. Organization and processing of proteins encoded by the yellow fever genome. Untranslated regions are shown as single lines and the translated region as an open box. The open triangle is the initiation codon (AUG); the solid diamond the termination codon (UGA). The protein nomenclature is described in Table 1 and (*35*). The single letter amino acid code is used for sequences flanking assigned cleavage sites (solid lines). Two other potential cleavage sites are shown as dotted lines. Structural proteins, identified nonstructural proteins, and hypothesized nonstructural proteins (see text) are indicated by solid, open, and hatched boxes, respectively. Other potential cleavage sites have been found and are described in Table 1, footnote asterisk.

*The structural proteins of yellow fever virus.* The start points of the three yellow fever virus structural proteins (C, M, and E) have been positioned within the translated RNA sequence from $NH_2$-terminal amino acid sequences obtained for the structural proteins isolated from yellow fever virions (*20*) (Fig. 1). The capsid protein is the first protein found in the long open reading frame and begins one residue past the first methionine. Thus, in agreement with in vitro translation data from the flavivirus genomic RNA's of tick-borne encephalitis virus, West Nile virus, and Kunjin virus (*15, 16*), the translation of the yellow fever genome initiates with the capsid protein, and the $NH_2$-terminal methionine is removed during maturation of the protein (*20*). The capsid protein may be released from the precursor polyprotein by cleavage at or just past a series of basic amino acids (Figs. 1 and 2). From this deduced amino acid sequence, the capsid protein is quite basic containing about 25 percent lysine and arginine distributed throughout the protein. The capsid protein of tick-borne encephalitis virus contains a similar proportion of basic amino acids (*22*). Since the capsid protein forms complexes with the RNA, its highly basic character probably acts to neutralize some of the RNA charges in such a compact structure.

Fig. 1 (preceding page and opposite page). Entire sequence of the genome of yellow fever virus. Yellow fever virus, 17D vaccine strain, was obtained from the American Type Culture Collection. This sample represents in vitro passage 234 of the line originated by Theiler and colleagues who started with the virulent Asibi strain (*6*). After plaque purification in Vero cells and amplification in BHK cells, the virus was grown in SW13 monolayers (*50*) and purified by polyethylene glycol precipitation, in glycerol-tartrate gradients. The purified virus was diluted with aqueous buffer and sedimented in the ultracentrifuge; the RNA was isolated by phenol extraction (*51*). Briefly, single-stranded cDNA was synthesized with avian myeloblastosis virus reverse transcriptase using degraded calf thymus DNA for priming (*47*). Second strand synthesis was carried out essentially as previously described (*52*). After methylation of the Eco RI sites with Eco RI methylase, phosphorylated Eco RI linkers were added with T4 DNA ligase. Following complete digestion with Eco RI, the double-stranded cDNA was sized on an agarose gel and selected size fractions were inserted into the Eco RI site of a plasmid vector derived from pBR322. Colonies containing yellow fever–specific inserts were selected by colony hybridization and were characterized by restriction mapping to obtain clones which represented most of the yellow fever genome. Clones containing the 3′ end of the genome were constructed by poly(A)-tailing (polyadenylation) the genomic RNA with *Escherichia coli* poly(A) polymerase followed by synthesis of double-stranded cDNA with an oligo(dT) primer. Addition of the poly(A) tract was relatively inefficient but after digestion of the double-stranded cDNA with Bgl I, 3′-terminal Bgl I fragments were selectively cloned with a plasmid vector derived from cloned yellow fever DNA (*51*). Clones containing the 5′ end of the genome were constructed by primer extension followed by oligo(dC) tailing with terminal deoxynucleotidyl transferase and oligo(dG) primed second strand synthesis. The entire sequence was obtained by chemical sequencing of both strands of the DNA (*53*). In addition, sequence was obtained throughout from at least two clones. Wherever the sequence differed between two clones (due presumably to heterogeneity in the RNA population or errors introduced during cloning), a third and occasionally a fourth clone was sequenced in this area, and the preferred nucleotide is reported here. Nucleotides are numbered from the 5′ terminus. Amino acids are numbered from the first methionine in the polyprotein sequence. The beginning of each protein is labeled (see Table 1 and text for nomenclature); tentative assignments are indicated by dashed arrows. Putative hydrophobic membrane-associated segments in the structural region are overlined. Potential *N*-linked glycosylation sites are denoted by an asterisk. The region of NS5 homologous to other RNA viruses (see text) is enclosed by brackets and the conserved Gly-Asp-Asp sequence is boxed. Repeated nucleotide sequences are underlined. Closely spaced in phase stop codons that terminate the long open reading frame are boxed. The single letter abbreviations for the amino acid residues are: A, alanine; C, cysteine; D, aspartic acid; E, glutamic acid; F, phenylalanine; G, glycine; H, histidine; I, isoleucine; K, lysine; L, leucine; M, methionine; N, asparagine, P, proline; Q, glutamine; R, arginine; S, serine; T, threonine; V, valine; W, tryptophan; Y, tyrosine.

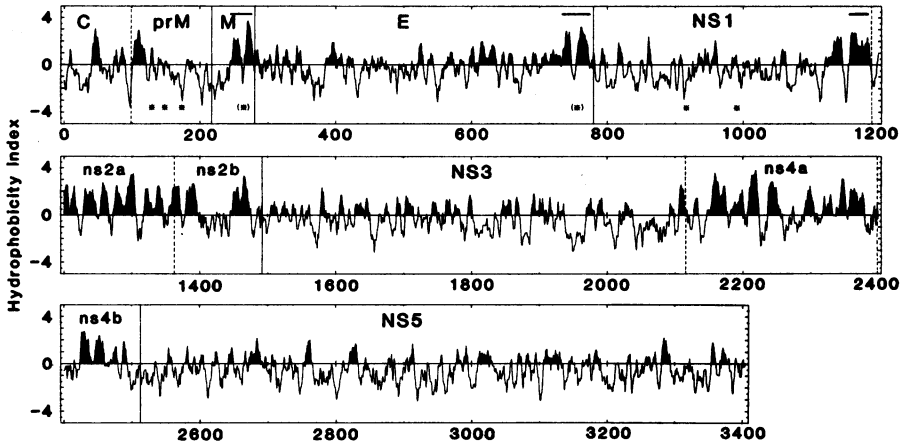



Fig. 3. Hydrophobicity plot of the yellow fever polyprotein sequence. The program of Kyte and Doolittle (*54*) with a search length of seven amino acids was used. Cleavage sites localized by $NH_2$-terminal protein sequence are indicated by solid vertical lines; putative cleavage sites are indicated by dotted vertical lines. The protein nomenclature is described in Table 1 and (*35*). The degree of hydrophobicity increases with distance above the horizontal line; hydrophilicity increases with distance below the horizontal line. Potential *N*-linked glycosylation sites are denoted by asterisks and putative membrane-associated anchors are indicated by solid bars.

There is also a hydrophobic stretch of 16 uncharged amino acids beginning with residue 42 from the $NH_2$ terminus (see Fig. 3), which is conserved among flaviviruses (*23*) and may be involved in protein-protein or specific protein-RNA interactions (or both) which assemble the nucleocapsid and lead to acquisition of the lipoprotein envelope by the capsid.

The start point of the virion M protein is also shown in Fig. 1. This protein contains a charged $NH_2$-terminal domain and two long uncharged stretches at its COOH terminus; these two stretches are separated by a single basic residue (Figs. 1 and 3) and could act as membrane spanning anchors similar to those observed in many virus envelope proteins. Protein M has not been identified in infected cells and is postulated to be derived from a precursor glycoprotein which we call prM (Table 1), which is also called by others GP23, GP19, or NV2 (*8, 24*). The sequence data support this hypothesis. A possible start point of prM, as deduced by limited homology with the $NH_2$-terminal sequence of the flavivirus St. Louis encephalitis NV2 (*20*) and homology in this region with Murray Valley encephalitis virus (*23*), follows the capsid protein; prM may begin with an uncharged stretch of amino acids which could function as an $NH_2$-terminal signal sequence for its cotranslational insertion into the endoplasmic reticulum (Fig. 3). After this hydrophobic domain, which may or may not be removed by signalase, the prM sequence contains three possible glycosylation sites of the type Asn-X-Ser/Thr. The $NH_2$ terminus of M (*20*) follows the sequence Arg-Ser-Arg-Arg in prM, indicating that the cleavage to produce M may be effected by the same enzyme that cleaves a number of viral envelope precursors at the sequence Arg-X-Arg/Lys-Arg and that has been postulated to be a host protease localized in the Golgi apparatus or Golgi-derived vesicles (*25*), perhaps similar to the cathepsins (*26*). As a result of this cleavage, which apparently occurs late during virus maturation and release, an 11.4-kD (not including carbohydrate) glycopeptide would be removed leaving the nonglycosylated M protein embedded in the virion membrane. Trace quantities of small virus-specific glycoproteins have been detected in cytoplasmic extracts (*27, 28, 29*), but whether the glycopeptide fragment remains cell-associated and is rapidly degraded or is released into the extracellular medium is unknown.

The E protein follows M. The $NH_2$ terminus of E is charged, and the more hydrophobic COOH-terminal domain of M (or its precursor, prM) may function as the signal sequence for the translocation of E across the rough endoplasmic reticulum. The protein E contains two sites of the form Asn-X-Ser/Thr which could serve as carbohydrate attachment sites, and both glycosylated and nonglycosylated forms have been detected in infected cells (*7, 27, 30*). The COOH-terminal domain of E contains uncharged stretches that could serve as a transmembrane anchor. Cleavage between M and E occurs after a Ser residue, and could be catalyzed by a host protease such as signalase. Since the COOH terminus of the mature M protein has not been determined, a small peptide, analogous to the 6 kD protein of alphaviruses (*25, 31*) could be produced during maturation of M and E. However, the apparent size of the M protein agrees well with the predicted molecular weight if cleavage occurs after the Ser at position 285.

This model for translation and processing of structural proteins and the features mentioned above predict that most of the E protein and some of the M protein should be exposed on the mature virion surface, and therefore sensitive to digestion by appropriate proteases. Protease digestion of purified tick-borne encephalitis virus (*32*) and also yellow fever virus (*29*) support this hypothesis. Thus, the M protein (or prM) of flaviviruses is an integral membrane protein and may interact specifically with both the E protein as well as the capsid protein-RNA complex during virus assembly.

*The nonstructural proteins.* In addition to prM, at least four and as many as 12 nonstructural proteins have been described in flavivirus-infected cells (*9, 28, 33, 34*). Some or all of these proteins must be active in the replication of the viral RNA. The start points of the three largest nonstructural proteins (NV3, NV4, and NV5 by the old nomenclature) (*35*) have been located by $NH_2$-terminal amino acid sequence analysis (*36*). As previously suggested by peptide mapping of the corresponding nonstructural proteins from other flaviviruses (*9, 15, 34*), the sequence data show that these proteins map to nonoverlapping segments in the yellow fever virus nonstructural region (Figs. 1 and 2).

In an attempt to simplify the description of flavivirus encoded nonstructural polypeptides, in particular the smaller proteins, we suggest a modified nomenclature (*35*) (Table 1) based on the linear order of these proteins in the yellow fever virus genome to complement designations based on their apparent molecular weights (*37*). In taking this approach we assume that members of Flaviviridae will have similar genome organization and express homologous proteins from homologous regions of their genomes. This assumption has been partially verified by an extensive sequence comparison of yellow fever virus with another member of the flavivirus genus, Murray Valley encephalitis virus (*23*).

Several features of the yellow fever virus nonstructural region are apparent from the localization of NS1, NS3, and NS5 (formerly NV3, NV4, and NV5). First, NS1 immediately follows the putative transmembrane segment of the E

protein. It should be noted that NS1 is glycosylated (*27*), and monoclonal antibodies against NS1 are capable of mediating complement-dependent lysis of yellow fever virus–infected cells, suggesting its presence at the plasma membrane (*38*). Thus, the COOH-terminal uncharged hydrophobic sequence of E could function as a signal sequence for translocation of NS1 across the endoplasmic reticulum. NS1 contains two sites of the type Asn-X-Ser/Thr which could serve as glycosylation sites. The probable COOH terminus of NS1 from estimates of molecular weight could contain a hydrophobic sequence for anchoring the protein in the membrane (Fig. 3). Thus the three glycoproteins of yellow fever virus, prM, E, and NS1, are adjacent to one another in the genome and are possibly inserted into the membrane one after another during synthesis. The sequence data support the hypothesis that each has the usual membrane protein topology of an $NH_2$ terminus outside and a COOH-terminal hydrophobic anchor. However, additional experiments are required to rigorously establish their orientation with respect to the lipid bilayer and exact COOH termini. The function of NS1 is unknown, but it could be involved in virus assembly rather than RNA replication. In this regard, it is of interest that NS1 has been shown to be the soluble complement-fixing antigen for dengue 2 (*28*) and suggestive evidence exists for a comparable role of NS1 in yellow fever virus infection (*8, 27*). Thus, this protein may exist in alternative membrane-associated and soluble forms, perhaps because of the presence or absence of the COOH-terminal hydrophobic domain.

NS3 begins at residue 1485 in the polyprotein sequence and is produced by cleavage at the site Gly-Ala-Arg-Arg ↓ Ser; the $NH_2$-terminus of NS5 has been tentatively identified as residue 2507 after cleavage at Thr-Gly-Arg-Arg ↓ Gly. Since no host proteases with this specificity (which are active in the cytosol) have been characterized and animal viruses often encode proteases active in the processing of their cytoplasmic polyprotein precursors, yellow fever virus may encode a protease that cleaves after two Arg residues (or two basic residues) surrounded by amino acids with short side chains, often Gly (Table 1 and footnote asterisk).

These assignments leave two regions in the polyprotein for which polypeptide products have not yet been identified. Assuming that other nonstructural proteins will be produced from these regions by the same protease responsible for $NH_2$-terminal cleavage of NS3 and NS5, we have scanned the remaining sequences for additional cleavage sites. Estimates of molecular weight (*27*) have positioned the COOH terminus of NS1 near residue 1187. The next potential cleavage sequence, Gly-Arg-Arg ↓ Ser, at residue 1355 would produce two small nonstructural polypeptides of approximately 18 kD (ns2a) and 14 kD (ns2b) located between NS1 and NS3 (Fig. 2 and Table 1). Both of these polypeptides would be extremely hydrophobic (Fig. 3) with ns2b containing a short internal charged domain. The putative cleavage at the sequence Glu-Gly-Arg-Arg ↓ Gly (residue 2108) would produce a polypeptide whose calculated mass agrees well with the observed size of NS3 on polyacrylamide gels (*27, 29*). Between this site and the $NH_2$ terminus of NS5 a single potential cleavage site (Ala-Gln-Arg-Arg ↓ Val) is found preceding residue 2395. Cleavage here would result in two methionine-rich, hydrophobic polypeptides of 31 kD (ns4a) and 12 kD (ns4b) (see Figs. 2 and 3 and Table 1). Polypeptides of these approximate sizes (10, 14, 18, and 30 kD) do exist in yellow fever–infected cells, but definitive mapping of these polypeptides as well as other minor species await additional $NH_2$-terminal sequence data. Similarly in the absence of COOH-terminal sequence data we cannot be sure of the exact terminal residues. Some heterogeneity in flavivirus polypeptides may result from variable exopeptidase digestion of the COOH-terminal residues or alternative internal cleavages. The predicted size of NS5, if the protein encompasses the remainder of the open reading frame, agrees well with its observed size (*27*).

*Implications for flavivirus replication.* It has been suggested that flavivirus RNA is translated by multiple internal initiation events (*17, 18*) which would make flaviviruses atypical among eukaryotic viruses and eukaryotic genes. The presence of a single long open reading frame in yellow fever virus RNA, the

Table 1. Flavivirus polypeptides.

| Protein nomenclature (*35*) | | $NH_2$-terminal cleavage site* | $M_r$† | $M_{pred.}$‡ | Glycosylated? | Comments |
|---|---|---|---|---|---|---|
| Proposed | Old | | | | | |
| | | | *Structural region* | | | |
| C | V2 (NV1½) | M ↓ S | 13,000 to 16,000 | 11,320 | No | Nucleocapsid protein |
| prM | (NV2) (NV2½) | ? | 19,000 to 23,000 | 20,925 | Yes | Precursor to M |
| M | V1 | SRR ↓ A | 8,000 to 8,500 | 8,526 | No | Virion envelope protein |
| E | V3 | AYS ↓ A | 51,000 to 60,000 | 53,712 | Both forms§ | Major virion envelope protein |
| | | | *Nonstructural region* | | | |
| NS1 | NV3 | VGA ↓ D | 44,000 to 49,000 | 45,869 | Yes | Soluble complement-fixing antigen |
| ns2a | (NV2½) (NV2) | (TVA ↓ V) | 16,000 to 21,000 | 18,086 | No | Hydrophobic; function unknown |
| ns2b | (NV1½) | (GRR ↓ S) | 12,000 to 15,000 | 13,823 | No | Hydrophobic; function unknown |
| NS3 | NV4 | ARR ↓ S | 67,000 to 76,000 | 69,319 | No | Replicase component ? |
| ns4a | (NVX) (NV2½) | (GRR ↓ G) | 24,000 to 32,000 | 31,196 | No | Hydrophobic; function unknown |
| ns4b | (NV1) | (QRR ↓ V) | 10,000 to 11,000 | 12,159 | No | Hydrophobic; function unknown |
| NS5 | NV5 | GRR ↓ G | 91,000 to 98,000 | 104,079 | No | Replicase component ? |

*Cleavage sites predicted by $NH_2$-terminal protein sequence data (*20, 36*). Tentative sites (indicated by parenthesis) are based on homology with confirmed cleavage sites and the sizes of yellow fever–specific polypeptides observed in infected cells (*27, 29*). Alternative cleavage sites in the nonstructural region occur after residue 1946 (Gln-Arg-Arg ↓ Gly), residue 2548 (Ala-Arg-Arg ↓ His), residue 2707 (Gln-Arg-Arg ↓ Phe), and residue 3104 (Ser-Arg-Arg ↓ Asp). †Range of flavivirus protein sizes estimated from acrylamide gel electrophoresis. Some of these proteins have not yet been identified for all flaviviruses thus far examined [for comparative analyses see (*33, 34*)]. In particular, definitive comparisons between NV3, NV2, NV2½, NVX, and NV1½ are difficult because of the complexity of flavivirus protein patterns in the 8,000 $< M_r <$ 45,000 size range. Alternative pathways of posttranslational cleavage may be used by different viruses for the production of the smaller nonstructural polypeptides. However, given the relatively consistent pattern of structural and larger nonstructural proteins, it seems likely that the small nonstructural proteins are also conserved, but their apparent migration on acrylamide gels and labeling efficiency are influenced by differences in amino acid composition. [For more complete discussion of this subject see (*49*).] ‡Polypeptide molecular weights calculated according to the cleavage sites shown in Fig. 1, with the C-prM cleavage site between residues 101 and 102. §Both glycosylated and nonglycosylated forms of E have been identified for yellow fever virus (*27*) and Kunjin virus (*7, 30*).

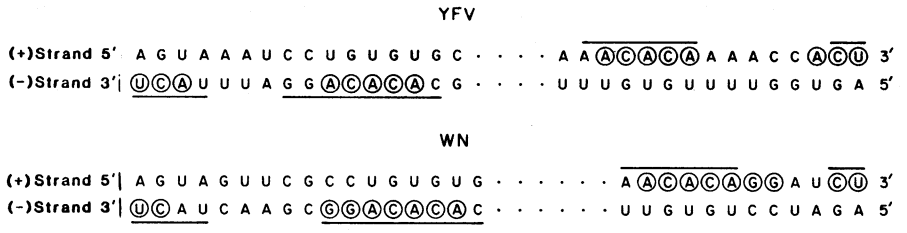



Fig. 4. Nucleotide homology between yellow fever virus (YFV) and West Nile virus (WN) [WN data are from (*21*)] at the 3′ termini of the genomic (+) strand and complementary (−) strand RNA's. Nucleotide identities in the 3′-terminal sequences of (+) and (−) strands are circled; those which are homologous between yellow fever and West Nile RNA's are underlined [(−) strand] or overlined [(+) strand].

fact that the final proteins found do not initiate with methionine but appear to arise from a consistent set of proteolytic cleavages, the gene order deduced from the pactamycin runoff experiments of Westaway (*17*), the in vitro translation data (*15, 16*), and recent evidence for polyprotein precursors (*39*) all support the view that translation of the flavivirus genome in vivo initiates with the capsid protein near the 5′ end of the genome and proceeds sequentially through the genome to produce one precursor polyprotein.

Cleavage of this precursor is rapid and occurs during translation so that the precursor is not seen in its entirety. The location and frequency of characteristic cleavage sites in this precursor suggest that processing involves both virus encoded and cellular organelle bound proteases. Although internal translation initiation cannot be formally excluded, the 5′ terminal location of the structural genes and the 3′ terminal replicase genes implies that the relative amounts of structural and nonstructural gene products could also be regulated by premature termination as well as by nonuniform rates of translation (*40*) or differential stability of the final products. A potential secondary structure in yellow fever RNA just past the structural protein genes could possibly be active in the former mechanism. It is unclear why gene mapping experiments with ultraviolet light to inactivate translation (*18*) or high salt to synchronize initiation of translation (*17*) suggest multiple independent sites of initiation and do not allow prediction of the correct gene order. Possible explanations are that ribosomes might have slow transit velocities in some areas, due to RNA secondary structures or the presence of rare codons (*40*), or that it might be necessary to translate a functional protease to produce the final products.

Several features potentially important in RNA replication or packaging (or both) can be identified in the genomic sequence. First, the extreme 5′- and 3′-terminal sequences are homologous to those found for another flavivirus, West Nile virus (*21*) (Fig. 4), and the complement of the 5′-terminal sequence [equivalent to the 3′ terminus of the (−) strand] is related to the 3′-terminal sequence of the (+) strand. This suggests that the viral replicase may have similar recognition sites for (+) and (−) strand synthesis. In addition, a stable secondary structure ($\Delta G$ = −40 to 45 kcal) can be formed from the 3′-terminal 87 nucleotides of the yellow fever genomic RNA (Fig. 5). This may be involved in RNA

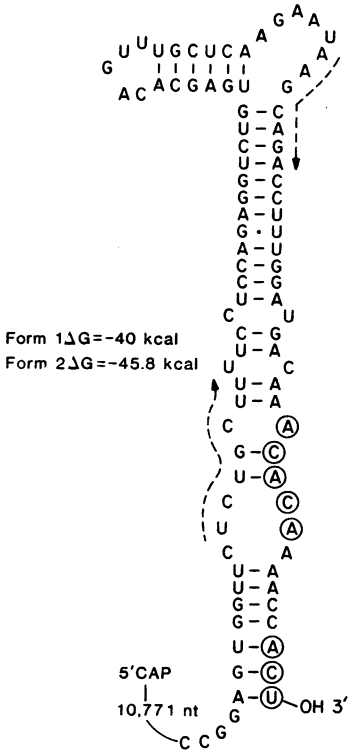



Fig. 5. Possible secondary structures at the 3′ terminus of yellow fever virus genomic RNA. Circled nucleotides are shared with the 3′ terminus of the yellow fever (−) strand (see Fig. 4). $\Delta G$ values were calculated according to Tinoco *et al.* (*55*). A more stable conformation than the one shown (form 1) can be formed if the two overlined sequences are base paired (form 2).

replication as well as encapsidation, and if conserved among flaviviruses could explain the observation that many flavivirus RNA's are poor substrates for 3′-terminal enzymatic modification including ligation and addition of poly(A) (polyadenylate). Similar transfer RNA–like secondary structures and conserved sequences have been identified at the 3′ end of many plant viral RNA's (*41*); in addition to serving as substrates for aminoacylation both in vivo and in vitro (*42*) they are important for initiation of (−) strand RNA synthesis (*43*). Last, the 3′-untranslated region contains a set of three closely spaced repeated sequences (underlined in Fig. 1) (located between nucleotides 10,374 and 10,520) each approximately 40 nucleotides long with an average of six changes between them in pairwise comparisons. The significance of these repeats in flavivirus replication is unknown.

*Evolution of flaviviruses.* It is becoming clear that the flaviviruses deserve their recent reclassification as a family separate from the alphaviruses. Although the mature virions are morphologically similar to alphaviruses in that they have a single-stranded RNA (+) sense genome encapsidated in an icosahedral nucleocapsid and surrounded by a lipid bilayer containing virus-specified polypeptides, they differ markedly in genome organization and replication strategy (*44*). The location of the genes encoding the structural proteins at the 5′ end of the genome, the single long reading frame, and the lack of a subgenomic message are all characteristics shared with picornaviruses rather than togaviruses.

In order to understand the evolutionary role of flaviviruses and their relation to other RNA viruses we have searched for homologies within the putative polymerase genes of various plant and animal viruses. Significant homologies have been found between alphaviruses and plant viruses (*45*) and less extensive homologies between picornaviruses and alphaviruses (*46*). Kamer and Argos (*46*) have aligned the polymerase gene of poliovirus with those of several viruses including alfalfa mosaic virus, bromegrass mosaic virus, tobacco mosaic virus, Sindbis virus, foot and mouth disease virus, encephalomyocarditis virus, and cowpea mosaic virus. The amino acid sequence of yellow fever virus NS5 between residues 3037 and 3181 can also be aligned with this collection of diverse RNA viruses (Fig. 1). These homologous regions are convincing but short and probably represent conserved functional domains for particular RNA-dependent polymerase functions. It is interesting to

speculate on the origin of this diverse group of viruses. Whether they arose from one or a few protoviruses (perhaps insect viruses) and have radiated to their current divergent hosts or whether the viruses have repeatedly cannibalized their hosts, obtaining their replicases from eukaryotic cellular functions cannot be resolved at present. However, one possible measure of host adaptation or origin of viral genes from host functions is the CG doublet frequency in the RNA. Insects, insect viruses, and alphaviruses (insect-borne with vertebrate hosts) have the expected CG doublet frequency predicted from their base compositions (*47*), whereas vertebrate DNA (*48*), viruses with exclusively vertebrate hosts, and yellow fever virus have low CG doublet frequencies (2.4 percent CG found in yellow fever compared to 6.1 percent predicted from the base composition). Given the rapid evolution of RNA genomes, it is unlikely that this difference applies directly to the question of evolutionary origin of alphaviruses and flaviviruses but rather reflects alternative strategies of adaptation to their arthropod and vertebrate hosts in ways which are not currently understood.

Comparative studies with other flaviviruses should help to define areas of commonality of function in the nonstructural proteins, to localize biologically important antigenic epitopes on the structural polypeptides (and NS1) and to ascertain whether certain features of the yellow fever sequence (like the putative secondary structure at the extreme 3′ terminus and repeated nucleotide sequences) are functionally significant landmarks conserved among flaviviruses. In addition, the construction of cDNA clones designed for expression of functional virus gene products or production of infectious virus should provide useful new approaches for studying flavivirus molecular biology and pathogenesis as well as for development of flavivirus vaccines.

## References and Notes


1. R. E. Shope, in *The Togaviruses*, R. W. Schlesinger, Ed. (Academic Press, New York, 1980), pp. 47–82.
2. R. W. Chamberlain, *ibid.*, pp. 175–227.
3. G. K. Strode, Ed., *Yellow Fever* (McGraw-Hill, New York, 1951).
4. W. G. Downs, *Yale J. Biol. Med.* **55**, 179 (1982).
5. W. Reed, *Med. Rec.* **60**, 201 (1901); _____ and J. Carroll, *Am Med.* **3**, 301 (1902); C. Norman, *Science* **223**, 1370 (1984).
6. M. Theiler and H. H. Smith, *J. Exp. Med.* **65**, 767 (1937); *ibid.*, p. 787.
7. P. J. Wright, *J. Gen. Virol.* **59**, 29 (1982).
8. P. K. Russell, W. E. Brandt, J. M. Dalrymple, in *The Togavirus*, R. W. Schlesinger, Ed. (Academic Press, New York, 1980), pp. 503–529.
9. E. G. Westaway, *ibid.*, pp. 531–581.
10. R. W. Boulton and E. G. Westaway, *Virology* **69**, 416 (1976); F. A. Murphy, in *The Togaviruses*, R. W. Schlesinger, Ed. (Academic Press, New York, 1980), pp. 241–316.
11. M. L. Ng, J. S. Pedersen, B. H. Toh, E. G. Westaway, *Arch. Virol.* **78**, 177 (1983).
12. G. Wengler, G. Wengler, H. J. Gross, *Virology* **89**, 423 (1978).
13. G. R. Cleaves and D. T. Dubin, *ibid.* **96**, 159 (1979); V. Deubel *et al.*, *Ann. Virol. (Inst. Pasteur)* **134E**, 581 (1983).
14. R. W. Boulton and E. G. Westaway, *Arch. Virol.* **55**, 201 (1977); C. W. Naeve and D. W. Trent, *J. Virol.* **25**, 535 (1978).
15. Y. V. Svitkin *et al.*, *FEBS Lett.* **96**, 211 (1978); G. Wengler, M. Beato, G. Wengler, *Virology* **96**, 516 (1979); Y. V. Svitkin *et al.*, *ibid.* **110**, 26 (1981); R. P. Monckton and E. G. Westaway, *J. Gen. Virol.* **63**, 227 (1982).
16. Y. V. Sitkin *et al.*, *Virology* **135**, 536 (1984).
17. E. G. Westaway, *ibid.* **80**, 320 (1977).
18. _____, G. Speight, L. Endo, *Virus Res.* **1**, 333 (1984).
19. M. Kozak, *Microbiol. Rev.* **47**, 1 (1983); we note a possible exception in the case of infectious pancreatic necrosis A RNA [P. P. C. Mertens and P. Dobos, *Nature (London)* **297**, 243 (1982)].
20. J. R. Bell *et al.*, *Virology* **143**, 224 (1985).
21. G. Wengler and G. Wengler, *ibid.* **113**, 544 (1981).
22. U. Boege, F. X. Heinz, G. Wengler, C. Kunz, *ibid.* **126**, 651 (1983).
23. L. Dalgarno, D. Trent, J. H. Strauss, C. M. Rice, in preparation.
24. D. Shapiro, W. E. Brandt, P. K. Russell, *Virology* **50**, 906 (1972).
25. H. Garoff, A. M. Frischauf, K. Simons, H. Lehrach, H. Delius, *Nature (London)* **288**, 235 (1980); C. M. Rice and J. H. Strauss, *Proc. Natl. Acad. Sci. U.S.A.* **78**, 2062 (1981).
26. K. Takio, T. Towatari, N. Katunuma, D. C. Teller, K. Titani, *Proc. Natl. Acad. Sci. U.S.A.* **80**, 3666 (1983).
27. J. J. Schlesinger, M. W. Brandriss, T. P. Monath, *Virology* **125**, 8 (1983).
28. G. W. Smith and P. J. Wright, *J. Gen. Virol.* **66**, 559 (1985).
29. J. Pata and C. M. Rice, unpublished data.
30. P. J. Wright, H. M. Warr, E. G. Westaway, *Virology* **109**, 418 (1981).
31. W. J. Welch and B. M. Sefton, *J. Virol.* **29**, 1186 (1979).
32. F. X. Heinz and C. Kunz, *Arch. Virol.* **60**, 207 (1979); F. Heinz, personal communication.
33. E. G. Westaway, *Virology* **51**, 454 (1973); _____, J. H. McKimm, L. G. McLeod, *Arch. Virol.* **53**, 305 (1977).
34. F. X. Heinz and C. Kunz, *J. Gen. Virol.* **62**, 271 (1982).
35. We propose an alternative nomenclature for flavivirus nonstructural proteins; our proposal is based on the yellow fever virus gene order determined by nucleic acid and protein sequence analysis (Fig. 1 and Table 1). The large nonstructural proteins (formerly NV3, NV4, and NV5) have been mapped and are numbered in order of appearance in the genome (5′ → 3′) with an upper case NS designation. We hypothesize that the remaining coding sequences in the nonstructural region encode several small flavivirus intracellular proteins (formerly NV1, NV1½, NV2, NV2½, and NVX), which are designated by a lower-case ns. Tentative identities with previously described flavivirus proteins are indicated by parentheses in Table 1. For alternative nomenclature see the text and (*37*). Minor virus-specific protein species that have been detected include two small glycoproteins ($M_r$ ~13,000 and 17,000) (*27*, *28*), and NV4½ (apparently related to NV4; perhaps equivalent to ns2b + NS3) (*5*).
36. J. Pata, J. Schlesinger, R. Aebersold, D. Teplow, S. Kent, J. H. Strauss, C. M. Rice, unpublished data.
37. E. G. Westaway *et al.*, *Intervirology* **14**, 114 (1980).
38. J. Schlesinger, personal communication.
39. G. Cleaves, personal communication.
40. R. Grantham, C. Gautier, M. Gouy, M. Jacobzone, R. Mercier, *Nucleic Acids Res.* **9**, 43 (1981).
41. T. C. Hall, in *International Review of Cytology*, (Academic Press, New York, 1979), vol. 60, pp. 1–26.
42. L. S. Loesch-Fries and T. C. Hall, *Nature (London)* **298**, 771 (1982); T. C. Hall, D. S. Shih, P. Kaesberg, *Biochem. J.* **129**, 969 (1972).
43. P. Ahlquist, J. J. Bujarski, P. Kaesberg, T. C. Hall, *Plant Mol. Biol.* **3**, 37 (1984).
44. E. G. Strauss and J. H. Strauss, *Curr. Top. Microbiol. Immunol.* **105**, 1 (1983).
45. J. Haseloff, J. Haseloff, P. Goelet, D. Zimmern, P. Ahlquist, R. Dasgupta, P. Kalsberg, *Proc. Natl. Acad. Sci. U.S.A.* **81**, 4358 (1984); P. Ahlquist *et al.*, *J. Virol.* **53**, 536 (1985).
46. G. Kamer and P. Argos, *Nucleic Acids Res.* **12**, 7269 (1984).
47. C. M. Rice and J. H. Strauss, *J. Mol. Biol.* **150**, 315 (1981).
48. G. J. Russell, P. M. B. Walker, R. A. Elton, J. H. Subak-Sharpe, *ibid.* **108**, 1 (1976); A. P. Bird, *Nucleic Acids Res.* **8**, 1499 (1980).
49. C. M. Rice, E. G. Strauss, J. H. Strauss, in *The Togaviruses and Flaviviruses*, S. Schlesinger and M. Schlesinger, Eds. (Plenum, New York, in press).
50. Obtained from Dr. Dennis Trent, Centers for Disease Control, Fort Collins, Colorado.
51. C. M. Rice, L. Dalgarno, D. W. Trent, J. H. Strauss; details of cloning and sequencing procedures are in preparation.
52. H. Okayama and P. Berg, *Mol. Cell. Biol.* **2**, 161 (1982).
53. A. M. Maxam and W. Gilbert, *Methods Enzymol.* **65**, 499 (1980).
54. J. Kyte and R. F. Doolittle, *J. Mol. Biol.* **157**, 105 (1982).
55. I. Tinoco *et al.*, *Nature (London)* **246**, 40 (1973).


56. We thank E. G. Strauss, L. Dalgarno, and C. Chang for helpful discussions and our many colleagues for critical comments on the manuscript; L. Hood and T. Hunkapiller for the use of their computer facilities; and C. S. Hahn for help with RNA secondary structure analysis. We also thank G. Cleaves, J. J. Schlesinger, and F. X. Heinz for allowing us to quote their unpublished results. Supported in part by grants AI 20612 and AI 10793 from NIH and by grant PCM 83-16856 from NSF.

17 May 1985; accepted 7 July 1985