Nucleotide Sequence of the Genome Region Encoding the 26S mRNA of Eastern Equine Encephalomyelitis Virus and the Deduced Amino Acid Sequence of the Viral Structural Proteins

By G WONG-JEN J. CHANG 1 AND DENNIS W. TRENT 2

1 Department of Microbiology, Colorado State University, Fort Collins, Colorado 80523 and 2 Division of Vector-Borne Viral Diseases, Center for Infectious Diseases, Centers for Disease Control, Public Health Service, U.S. Department of Health and Human Services, P.O. Box 2087, Fort Collins, Colorado 80522, U.S.A.

(Accepted 21 April 1987)

SUMMARY

The 26S mRNA and most of the nsP4 encoding regions of the eastern equine encephalomyelitis (EEE) viral genome have been cloned. Excluding the poly(A) tail, the 26S mRNA region was determined to be 4139 nucleotides long and to share the same general organization as that of other alphaviruses. A highly conserved region of 19 nucleotides, the putative transcriptase recognition site for 26S mRNA synthesis, was present at the 26S/42S junction region of the 42S genomic RNA. Translation of the 26S mRNA began at the first AUG (positions 59 to 61) initiation codon and continued with an open reading frame that coded for a polyprotein of 1258 amino acids ending at a UAA ochre termination codon (positions 3776 to 3778). All four putative post-translational cleavage sites used to generate the capsid, E3, E2, 6K and E1 proteins were conserved. Transmembrane domains present in the EEE virus structural polyprotein have been identified and their functions discussed. Pairwise comparison of the deduced amino acid sequences of the polyproteins of five alphaviruses (EEE, Venezuelan equine encephalitis, Sindbis, Semliki Forest and Ross River viruses) revealed EEE virus to be more closely related to VEE virus than to the other three viruses.

INTRODUCTION

Eastern equine encephalomyelitis (EEE) virus was first isolated in New Jersey by TenBroeck & Merrill (1933) from infected horse brain. Since then, the virus has been isolated from mosquitoes, birds, horses and humans along the freshwater marshes of the Atlantic and Gulf coasts of the United States, among the islands and along South and Central American coastlines of the Caribbean and in foci near the Great Lakes (Shope, 1985). Outbreaks in man are generally sporadic, with 152 human cases reported between 1938 and 1971 in the United States. The largest recorded equine outbreak in the United States occurred in 1947 in the bayous of coastal Louisiana and adjacent Texas (Monath & Trent, 1981), where 14344 equine cases with 11722 deaths and 10 human cases were attributed to this epidemic.

EEE virus belongs to the alphavirus genus of the family Togaviridae (Westaway et al., 1985). Casals (1964) determined that there are two antigenic variants of the virus. North American viruses are serologically identical but differ from the Panamanian and South American strains, which differ among themselves and are distinct from the North American isolates. EEE virus strains from the Dominican Republic (Eklund et al., 1951) and Jamaica are indistinguishable from the North American type (Casals, 1964).

Alphaviruses replicate in the cytoplasm of vertebrate and invertebrate cells and mature by budding through the plasma membrane (Murphy, 1980). Two species of viral mRNA are
present in alphavirus-infected cells: 42S mRNA, which is packaged into mature virions and functions as the message for the non-structural proteins, and a 26S mRNA, which encodes the structural polypeptides. The 26S mRNA is homologous to the 3' third of the 42S mRNA and is translated into a 130K polyprotein that is cotranslationally cleaved and processed into the capsid protein and two glycosylated membrane proteins, E1 and E2 (reviewed in Strauss & Strauss, 1983). The nucleotide sequences of 26S RNA of four alphaviruses [Semliki Forest (SF), Sindbis (SIN), Ross River (RR), and Venezuelan equine encephalitis (VEE) viruses] have been determined (Garoff et al., 1980a, b; Rice & Strauss, 1981a, b; Dalgarno et al., 1983; Kinney et al., 1986). Comparative data for the 3' terminal, 5' terminal and 26S/42S junction regions for a number of alphaviruses have also been reported (Ou et al., 1981, 1982a, b, 1983). We have cloned and sequenced the genome regions that encode the structural genes and most of the nsP4 gene of EEE virus strain 82V-2137. The nucleotide sequence of the structural genes and the deduced amino acid sequence have been analysed to (i) identify the potential glycosylation sites of the glycoproteins, (ii) determine the possible major antigenic determinants for the E1 and E2 envelope proteins, and (iii) recognize the postulated functional domains in the viral structural proteins.

METHODS

Preparation of EEE genomic 42S RNA. The 82V-2137 strain of EEE virus used was a North American subtype of the virus isolated in 1982 from mosquitoes collected in an endemic area of Florida. Virus was grown in BHK-21 cells in medium containing 12.5 μCi/ml of [3H]uridine and was purified as previously described (Kinney et al., 1986). Genomic RNA, extracted from gradient-purified virus (Trent et al., 1979), was ethanol-precipitated, resuspended in 100 μl of RNA gradient buffer (10 mM-Tris-HCl pH 7.4, 10 mM-NaCl, 1 mM-Na2 EDTA, 0.2% w/v SDS) and centrifuged in a 15% to 30% sucrose density gradient in RNA gradient buffer for 3-5 h at 20 °C and 40000 r.p.m. in an SW40 rotor. Gradient fractions containing high molecular weight RNA were pooled, ethanol-precipitated, and used as the template for the cDNA synthesis.

Synthesis of cDNA. EEE virus cDNA was synthesized as described by Gubler & Hoffman (1983) with some modifications. Two synthetic primers, oligo(dT)30 (P-1) and 5'-ATTATGCGCTGCCTGTAGTG-3' (P-2), were synthesized on an Applied Biosystems Model 380A DNA synthesizer and used to prime the first-strand cDNA synthesis. P-2 is complementary to a sequence of 20 bases beginning 72 nucleotides from the 3' poly(A) tail of EEE 42S RNA (Ou et al., 1982b). Five μg of 42S RNA in 20 μl of distilled water was denatured by heating (93 °C for 3 min), chilled quickly and 3.0 μl of 14 μM-methylmercuric hydroxide (Alpha, Danvers, Mass., U.S.A.) was added. After the sample had been incubated at room temperature for 5 min, 10 μl of 2-mercaptoethanol (1.85 M) was added and incubated for an additional 5 min. To this was added 36 μl of transcriptase solution to give final concentrations of 50 mM-Tris-HCl pH 8.3, 70 mM-KCl, 10 mM-MgCl2, 30 mM-2-mercaptoethanol, 700 μM-dNTPs, 600 units/ml of RNase inhibitor (Promega Biotechnologies Inc., Madison, Wis., U.S.A.), 20 μCi of [α-32P]dATP (New England Nuclear, 3000 Ci/mmol), 60 μg/ml of P-1 or P-2 synthetic DNA primer and 1500 units/ml of avian myeloblastosis virus reverse transcriptase (Promega Biotechnologies Inc.). The reverse transcriptase reaction was carried out at 42 °C for 2 h, phenol/chloroform-extracted, ether-washed and ethanol-precipitated.

Second-strand cDNA was synthesized by the RNase H-DNA polymerase I method of Gubler & Hoffman (1983). After incubation at 12 °C for 18 h, the ds cDNA was extracted with phenol/chloroform, ether-washed and ethanol-precipitated. The ds cDNA was then fractionated by column chromatography using Sephacryl S-1000 (Pharmacia) in 20 mM-Tris-HCl pH 9.0, 100 mM-NaCl and 1 mM-Na2 EDTA. Column fractions containing the high molecular weight cDNA were pooled and precipitated with ethanol.

Cloning of EEE cDNA. PstI-cut pUC18 plasmid (Norlander et al., 1983) and ds cDNA were tailed with poly(dG) and poly(dC), respectively, using terminal transferase (Roychoudhury & Wu, 1980). Tailed vector and cDNA were annealed and used to transform Escherichia coli cells, strain TB1 (a derivative of strain JM83; Yanisch-Perron et al., 1985). The TB1 cells were made competent by the low pH-MnCl2 method described by Clark-Curtiss & Curtiss (1983). Ampicillin-resistant colonies containing recombinant plasmids were grown, and the plasmids were extracted by the rapid boiling method of Holmes & Quigley (1981). Recombinant plasmids were digested with PstI and analysed on 1% agarose gels to determine the size of the cDNA insert.

Hybridization screening of recombinant plasmids for relevant cDNA sequences. Recombinant plasmids with viral cDNA inserts larger than 3 kb were screened for relevant cDNA sequences by hybridization using three synthetic DNA probes (P-2, P-E1 and P-E2) synthesized on an Applied Biosystems Model 380A DNA synthesizer. The P-2 probe was identical to the primer used during cDNA synthesis and the P-E1 and P-E2 probes were mixed 17-mer DNA, each containing all possible permutations of a sequence of 17 nucleotides coding for contiguous amino acids at the amino terminal of the EEE virus E1 and E2 glycoproteins (Bell et al., 1984). The probes were labelled at their 5' termini with [γ-32P]dATP using T4 DNA kinase and used for dot blot hybridization.
as described by Wallace et al. (1981). Two clones, designated pEE-14 and pEE-72, that hybridized with these probes were purified for subcloning and restriction map analysis (Garger et al., 1983).

Subcloning and sequencing of clones pEE-14 and pEE-72. Single and multiple restriction enzyme digestions (Lawn et al., 1978) were used to map the restriction sites of the pEE-14 and pEE-72 clones. Restriction fragments suitable for sequence analysis were separated by low melting point agarose electrophoresis, isolated from the gel (Kinney et al., 1986) and subcloned into bacteriophage M13 mp18 or mp19 (Barnes & Bevan, 1983; Norrander et al., 1983). E. coli strain JM101 (Yanisch-Perron et al., 1985) cells were made competent for transfection by the magnesium chloride/calcium chloride method (Messing, 1983). Recombinant phages were tested for insert cDNA identity and orientation using the C-test (Messing, 1983). Recombinant phage clones containing cDNA inserts larger than 700 bp were manipulated further by the rapid deletion subcloning procedure of Dale et al. (1985) to facilitate sequencing across the entire cDNA. DNA sequencing was performed by the dideoxy chain termination method (Sanger et al., 1977, 1980).

Data management and analysis. Computer analysis of DNA sequence information was accomplished with the aid of a Fortran version 77 program, RKSEQ, running on a Control Data Corporation mainframe Cyber computer (Kinney et al., 1986). A commercial software program, DNASIS (Hitachi) run on an IBM PC/XT H personal computer was also used to analyse the nucleotide sequence.

RESULTS

Molecular cloning and sequencing of viral cDNA

Twenty clones containing cDNA of 2700 to 4900 bp were isolated and the cDNA sequences relative to EEE viral RNA determined by dot blot hybridization. Two clones, pEE-14 and pEE-72, which should contain the 26S mRNA sequence, were further characterized by restriction enzyme mapping (Fig. 1).

Our strategy for sequencing clones pEE-14 and pEE-72 is presented in Fig. 1. Both strands of the cDNA in these clones were fully sequenced and all the restriction enzyme junctions were confirmed by sequencing other recombinant clones which contained cDNA sequences that spanned the junctions.

The nucleotide and deduced amino acid sequences of EEE virus genome encoding the 26S mRNA

The sequence of clone pEE-72 begins 18 nucleotides beyond the opal termination codon UGA that is located between nsP3 and nsP4 in the SIN virus sequence (Strauss et al., 1984), extends through the junction region of the 26S mRNA, and contains three-quarters of the length of the entire 26S mRNA. The pEE-14 clone overlaps clone pEE-72 by about 1200 nucleotides and extends to the poly(A) tract (Fig. 1).

The cap site of the EEE virus 26S mRNA was identified by homology alignment with other alphaviruses (Fig. 2). Excluding the poly(A) tract, the EEE virus 26S mRNA was 4139 nucleotides long with a base composition of 27.9% A, 25.5% C, 23.3% G and 23.3% U. The 5' and 3' non-translated regions were 58 and 362 nucleotides in length, respectively. Translation started at the first AUG initiation codon (position 59 to 61) and extended for 3717 nucleotides until an ochre termination codon (UAA) was reached. The open reading frame coded for a polyprotein of 1239 amino acids which, by homology with other alphaviruses, comprises five proteins: capsid, E3, E2, 6K and E1 (Fig. 3). Codon usage in the structural region of the genome was not random (Table 1), which may permit efficient translation of viral proteins in both vertebrate and invertebrate cells (Dalgarno et al., 1983; Kinney et al., 1986; Strauss & Strauss, 1986).

The first 50 amino-terminal amino acids of E1 and the first 25 of E2 have been previously determined (Bell et al., 1984). The published amino acids at position 41 in E1 and position 23 in E2 are isoleucine and methionine, whereas the deduced amino acids at these positions were serine and glycine, respectively. This difference may be due to sequence differences between the NJO strain and the 82V-2137 strain EEE viruses studied. The amino-terminal sequences of other proteins were assigned on the basis of their homology with those of other alphaviruses (Kinney et al., 1986). Homologies around the presumptive proteolytic cleavage sites in the EEE viral structural polyproteins and other alphaviruses are illustrated in Fig. 4. The five proteins coded by EEE 26S mRNA (capsid, E3, E2, 6K and E1) were 259, 63, 420, 56 and 441 amino acids in length, respectively. The first amino acid in the amino-terminal end of the EEE virus E2
Fig. 1. Cloning strategy for dideoxy sequencing of EEE virus cDNA clones. Restriction maps of EEE virus cDNA clones are aligned with a schematic representation of the organization of the 3' half of the EEE viral genome. Regions included in the pEE-14 and pEE-72 clones are shown by double directional arrows. Restriction fragments of the cDNA were cloned into bacteriophages M13 mp18 or mp19. Single direction arrows indicate the sense of the cDNA sequenced and the extent of sequence obtained for relevant single-stranded M13 recombinants. A rapid deletion subcloning procedure was used to produce a series of shortened overlapping clones for those recombinants with a cDNA insert size larger than 700 bp.
EEE viral RNA sequence

```
-30 -20 -10
EEE ALAAGACUGCAACCUAUACGCAACG
       I R G H P I T L Y G #

VEE C:::GGC:::U:::U:;:::::::A:
       #

SIN A:::GG:::AU::AGAU:;::::::U:

SFV U:::A::UGU:;::::::A:;::::::U:

RRV C:GC:U:;A:CAU:G:CA:;::::::G:

MBV C:;A:CU:;UGAC:;::::::G:
```

Fig. 2. The junction region of EEE virus, strain 82V-2137, and several other alphaviruses. Flanking sequences around the putative transcriptase recognition site, as identified by Ou et al. (1982b), and the translation initiation site are aligned with the EEE virus sequence. Gaps are introduced between these two flanking sequences. Colons (:) represent nucleotides identical to the EEE virus sequence; (#) indicates termination codons. Data for SIN, SF, RR, MB and VEE viruses were obtained from Ou et al. (1982b) and Kinney et al. (1986).

glycoprotein was aspartic acid, which is different from other alphaviruses that have serine at this position. The last two amino acids at the carboxyl-terminal end of the EEE virus E3 glycoprotein were arginine, agreeing with the postulation that the proteolytic processing of PE2 to form E3 and E2 is carried out by the Golgi apparatus protease (Strauss & Strauss, 1986; Dean & Judah, 1980).

**EEE virus capsid protein**

The EEE capsid protein was 259 amino acids in length with a deduced molecular weight of 28,705. The amino-terminal 101 amino acids of the capsid protein were rich in alpha-helix-disrupting proline and showed little homology with other alphaviruses (Fig. 5) (Kinney et al., 1986). Amino acids 38 to 101 were highly hydrophilic (Fig. 6). Clusterings of arginine, lysine and proline residues in the amino-terminal half of the EEE virus capsid protein suggests that the highly basic amino-terminal half of the capsid protein stabilizes the interaction of capsid with virion RNA (Garoff et al., 1980a) or interacts with genomic RNA to initiate encapsidation (Strauss & Strauss, 1986).
Table 1. Codon usage of EEE virus 26S mRNA

<table>
<thead>
<tr>
<th>Codon</th>
<th>Amino acid residues</th>
<th>Codon</th>
<th>Amino acid residues</th>
<th>Codon</th>
<th>Amino acid residues</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phe</td>
<td>20</td>
<td>Pro</td>
<td>26</td>
<td>Lys</td>
<td>42</td>
</tr>
<tr>
<td>UUU</td>
<td></td>
<td>CCU</td>
<td>11</td>
<td>AAA</td>
<td>42</td>
</tr>
<tr>
<td>UUC</td>
<td></td>
<td>CCC</td>
<td></td>
<td>AAG</td>
<td>32</td>
</tr>
<tr>
<td>UUA</td>
<td></td>
<td>CCA</td>
<td>40</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UUG</td>
<td>10</td>
<td>CCG</td>
<td>28</td>
<td>Asp</td>
<td>26</td>
</tr>
<tr>
<td>CUU</td>
<td>14</td>
<td>GAC</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CUC</td>
<td>18</td>
<td>ACU</td>
<td>16</td>
<td>Glu</td>
<td>24</td>
</tr>
<tr>
<td>CU A</td>
<td>8</td>
<td>ACC</td>
<td>32</td>
<td>GAA</td>
<td>24</td>
</tr>
<tr>
<td>CUG</td>
<td>32</td>
<td>ACA</td>
<td>23</td>
<td>GAG</td>
<td>24</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CCA</td>
<td>40</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ile</td>
<td>21</td>
<td>Ala</td>
<td>24</td>
<td>Arg</td>
<td>7</td>
</tr>
<tr>
<td>AUU</td>
<td></td>
<td>GCU</td>
<td></td>
<td>CGU</td>
<td>11</td>
</tr>
<tr>
<td>AUC</td>
<td>21</td>
<td>GCC</td>
<td>28</td>
<td>CGG</td>
<td>12</td>
</tr>
<tr>
<td>AUA</td>
<td>18</td>
<td>GCA</td>
<td>25</td>
<td>AGA</td>
<td>18</td>
</tr>
<tr>
<td>Met</td>
<td>23</td>
<td>UAU</td>
<td>22</td>
<td>AGG</td>
<td>12</td>
</tr>
<tr>
<td>AUG</td>
<td></td>
<td>UAC</td>
<td>28</td>
<td></td>
<td></td>
</tr>
<tr>
<td>V al</td>
<td>17</td>
<td>GCC</td>
<td>25</td>
<td>Gly</td>
<td>16</td>
</tr>
<tr>
<td>GUU</td>
<td>14</td>
<td>CAU</td>
<td>25</td>
<td>GGU</td>
<td>18</td>
</tr>
<tr>
<td>GUC</td>
<td>34</td>
<td>CAC</td>
<td>21</td>
<td>GGC</td>
<td>18</td>
</tr>
<tr>
<td>GUA</td>
<td></td>
<td>CCA</td>
<td>40</td>
<td>GGA</td>
<td>30</td>
</tr>
<tr>
<td>Ser</td>
<td></td>
<td>GCA</td>
<td>22</td>
<td>GGG</td>
<td>19</td>
</tr>
<tr>
<td>UCU</td>
<td>11</td>
<td>CAA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GCC</td>
<td>28</td>
<td>CAG</td>
<td>24</td>
<td>Trp</td>
<td>17</td>
</tr>
<tr>
<td>UCA</td>
<td></td>
<td>UAG</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UCG</td>
<td>6</td>
<td>AAU</td>
<td>17</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AGU</td>
<td>6</td>
<td>AAC</td>
<td>35</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AGC</td>
<td>14</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The carboxyl-terminal 158 amino acids of the capsid are highly conserved among alphaviruses (Fig. 5) and probably have a highly defined structure. The C-terminal region is postulated to have functions that include subunit–subunit interaction in assembling the nucleocapsid, capsid–glycoprotein interaction in virus budding and an autoproteolytic activity in processing the viral structural polyprotein (Strauss & Strauss, 1986).

The SIN virus capsid protein has been postulated to have an autoproteolytic serine protease activity formed by His-141, Asp-147 and Ser-215 (Hahn et al., 1985). These three residues were conserved in the EEE virus capsid protein at positions His-136, Asp-142 and Ser-210. A comparison of the EEE virus capsid protein with other alphavirus capsid protein amino acid sequences showed a 61.9% overall homology with the VEE virus capsid protein, 50.3% with that of SIN virus, 49.6% with that of SF virus and 48.7% with that of RR virus.

EEE virus E3 and E2 proteins

The EEE virus E3 protein was 63 amino acids in length with a deduced unglycosylated molecular weight of 7203. The first 16 amino acids, which were uncharged and fairly hydrophobic (Fig. 6), could serve as a signal sequence for the insertion of PE2 ( uncleaved E3
Fig. 4. Putative cleavage sites of the alphavirus structural polyproteins. The data for VEE, SIN, SF and RR viruses were obtained from Kinney et al. (1986), Rice & Strauss (1981a), Garoff et al. (1980a, b) and Dalgarno et al. (1983), respectively. The amino- and carboxyl-terminal amino acid sequence around the putative cleavage sites and the length (aa, amino acids) of each cleavage product are indicated.

Fig. 5. Comparison of EEE and VEE virus structural polyprotein sequences by the dot matrix method. A dot is placed in the matrix whenever four consecutive amino acid residues in the two strings, without introducing the gap between residues, are identical.

Fig. 6. Hydrophobicity profile of EEE viral structural proteins. Moving average hydrophobicity indices, nine amino acids in length (obtained from Kyte & Doolittle, 1982), are plotted against amino acid positions along the EEE structural polyprotein. Positive values indicate hydrophobicity, whereas negative values indicate hydrophilic amino acid regions. Potential Asn-linked glycosylation sites are identified by arrows. Possible transmembrane domains are shadowed.
EEE viral RNA sequence

plus E2) into the endoplasmic reticulum during protein synthesis (Garoff et al., 1980b). A potential glycosylation site was found at position 11 (Fig. 3 and 6), which is conserved among alphaviruses. The SF virus and RR virus E3 proteins possess a second potential glycosylation site at Asn(60)-Gly-Thr and Asn(58)-Arg-Ser respectively. However, only one of the two sites in SF virus is glycosylated (Garoff et al., 1980a, b). The EEE virus E3 protein showed 57.81\%, 39.71\%, 57.58\% and 54.69\% overall amino acid sequence homology with the VEE, SIN, SF and RR virus E3 proteins, respectively.

The E2 protein of EEE virus had a deduced unglycosylated molecular weight of 26,979 and was 420 amino acids long. The E2 protein is the least conserved among the alphaviruses, especially the amino-terminal two-thirds of the protein (Fig. 5). To identify regions of the E2 glycoprotein which may contain major antigenic determinants, the amino acid sequence was analysed using the hydrophobicity plot of Hopp & Woods (1981). Results from this analysis, based on the identification of hydrophilicity and amino acid sequence in the region, revealed two stretches of amino acids as possible candidates for major antigenic determinants in the EEE virus E2 protein. One of these sequences was located at the extreme amino terminus (1-Asp-Leu-Asp-Thr-His-Phe-6) in a region of E2 which is highly variable and may contain alphavirus type-specific antigenic determinants (Kinney et al., 1986). Another predicted antigenic determinant was located in the amino acid sequence 232-Lys-Trp-Val-Tyr-Asn-Ser-237, which is shared with the VEE virus E2 protein but not with those of SIN, SF, or RR viruses (Kinney et al., 1986).

The E2 protein had two stretches of uncharged amino acids near its carboxyl terminus at amino acids 360 to 379 and 405 to 416 (Fig. 3 and 6). The sequence of the first peptide was highly variable, having only four conserved amino acids in the total of 20 residues. The second peptide had a higher homology, with nine conserved amino acids in a total of 11 residues. Both peptides span the membrane (Kyte & Doolittle, 1982) and function as a signal or stop-transfer sequence (Coleman et al., 1985).

The region between these two hydrophobic domains (residues 380 to 404) was relatively hydrophilic (Fig. 6) and is possibly exposed on the cytoplasmic side of the viral envelope. This region has five consecutive amino acids (393-Cys-Ile-Thr-Pro-Tyr-397) shared by all alphaviruses examined. Amino acids in this region may interact with the viral nucleocapsid during the morphogenesis of alphaviruses.

The EEE virus E2 protein has two potential glycosylation sites at amino acid positions 302 (Asn-Pro-Thr) and 315 (Asn-Phe-Thr) but its actual glycosylation pattern is still unknown. The EEE virus E2 protein had overall amino acid sequence homology of 45.86\%, 42.72\%, 39.96\% and 39.86\% with the E2 proteins of VEE, SIN, SF and RR virus, respectively.

EEE virus 6K and E1 proteins

EEE virus 6K protein was the most hydrophobic protein of the EEE virus structural proteins (Fig. 6). The only biological function assigned to this protein is that of a signal sequence involved in translocation of the E1 protein (Dalgarno et al., 1983; Kinney et al., 1986). The EEE virus 6K protein shares a 53.57\% overall sequence homology with the 6K protein of VEE virus but has only a 31.58\%, 35\% and 31.67\% with those of SIN, SF and RR viruses, respectively. The EEE virus 6K protein was 56 amino acids in length with a deduced molecular weight of 6,159. It could be separated into two domains (Fig. 6). The first was hydrophilic in nature and contained five charged amino acids, whereas the second was hydrophobic, thus fulfilling the criteria of Kyte & Doolittle (1982) to be a transmembrane domain and to function as a signal sequence (Cutler & Garoff, 1986).

The EEE virus E1 protein was 441 amino acids long with an unglycosylated molecular weight of 48,300. Biological functions assigned to the E1 protein include mediation of virus fusion and anchoring of the E1 protein in the virus envelope (Garoff et al., 1980b; Rice & Strauss, 1981a; Kinney et al., 1986). The fusion domain, a hydrophobic peptide, was present in the EEE virus E1 protein between positions 75 to 109. The second hydrophobic region, anchoring the E1 protein, was located at the carboxyl terminus, beginning at 413-Val and ending at 437-Phe. Only the last three charged amino acid residues (439-His-Arg-His-441) are located on the cytoplasmic
Table 2. Characterization of transmembrane domains in the EEE virus structural polyprotein precursor

<table>
<thead>
<tr>
<th>Peptide</th>
<th>Flanking basic amino acid</th>
<th>Average hydrophobicity index*</th>
<th>Presence of</th>
<th>Possible function</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N terminus</td>
<td>C terminus</td>
<td>Length</td>
<td>Gly</td>
</tr>
<tr>
<td>E3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2–17</td>
<td>–</td>
<td>–</td>
<td>16</td>
<td>–</td>
</tr>
<tr>
<td>E2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>362–379</td>
<td>359-Arg</td>
<td>–</td>
<td>18</td>
<td>+</td>
</tr>
<tr>
<td>405–416</td>
<td>398-Lys</td>
<td>417-Lys</td>
<td>12</td>
<td>–</td>
</tr>
<tr>
<td>6K</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>35–56</td>
<td>34-Arg</td>
<td>4-His§</td>
<td>22</td>
<td>+</td>
</tr>
<tr>
<td>E1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>413–438</td>
<td>412-Lys</td>
<td>439-His</td>
<td>25</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td></td>
<td>440-Arg</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>441-His</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* Based on the scale of Kyte & Doolittle (1982).
† SS, Signal sequence.
‡ STS, Stop-transfer sequence.
§ 4-His in the E1 protein.

side of the membrane. The EEE virus E1 glycoprotein has a single Asn-linked glycosylation site located at position 134 (Fig. 3 and 6).

Possible antigenic determinants on the EEE virus E1 are predicted by the hydrophilicity plot of Hopp & Woods (1981). One of the predicted regions has the amino acid sequence 104-Ser-Glu-Ala-Tyr-Val-Glu-109 and is located at the carboxyl terminus within the postulated fusion domain. Interaction of this peptide with the cellular receptor may induce conformational changes of the fusion domain and cause penetration of the virus.

DISCUSSION

The transcription start site of the EEE virus 26S RNA was identified by homology alignment with that of other alphaviruses, confirming the same general organization of the subgenomic mRNAs (Garoff et al., 1980a, b; Rice & Strauss, 1981a, b; Dalgarno et al., 1983; Kinney et al., 1986). The EEE virus gene encoding the 26S mRNA is 4139 nucleotides in length, excluding the poly(A) tract, 5' and 3' non-coding regions of 58 and 362 nucleotides, respectively. The 19 conserved nucleotides (Ou et al., 1982b), comprising the putative transcriptase recognition sequence for the alphavirus 26S RNA, was confirmed at the 26S/42S junction region of the EEE virus RNA (Fig. 2). Four in-phase termination codons are located in this region of the EEE virus genome. Both EEE and VEE viruses have opal termination codons (UGA) in the conserved sequence of the 26S junction. EEE virus also has an in-phase termination codon, similar to SIN, RR and Middleburg viruses, at the beginning of the 26S RNA (Fig. 2).

Translocation of the PE2 and E1 proteins through the rough endoplasmic reticulum membrane and anchoring proteins on the membrane requires at least two functional signal sequences and two stop-transfer sequences. These sequences have been identified unambiguously in the E3 and 6K protein region and at the carboxy-terminus of E2 and E1 (Table 2). Two proteolytic cleavages are required to separate PE2 and E1, both of which occur after alanine residues (Fig. 4). It has been suggested that signalase, which is found only on the luminal side of the rough endoplasmic reticulum, catalyses both these cleavages (Rice & Strauss, 1981a). If this is correct, then there must be another internal signal sequence located at the extreme carboxy terminus of the E2 protein. One transmembrane sequence (positions 405 to 416 of the EEE virus E2, Table 2), which is separated from the first transmembrane domain in E2 by 26 amino acids,
contains all the features necessary to function as an internal signal sequence. This region has an average hydrophobic index of 2.74 and is flanked by positively charged amino acids at both ends. The segment would facilitate the translocation of the 6K protein and also act as a second membrane anchor domain for the PE2 protein. Mutagenesis of viral cDNA in vitro around these two compatible domains has been carried out in the SF viral system to study the processing of the PE2 protein (Garoff et al., 1983; Cutler & Garoff, 1986; Cutler et al., 1986; Roman & Garoff, 1986). The results seem to be in favour of our prediction since deletion or base substitution mutations which do not change the first compatible domain (position 362 to 379 of the EEE virus E2) do not alter the behaviour of the PE2 protein in the SF system. The truncated PE2 protein associates with rough endoplasmic reticulum and does not reach the cell surface if the deletion affects the first transmembrane region in E2 (Garoff et al., 1983).

Translocation and proteolytic processing of the EEE virus polyprotein is summarized in Fig. 7. As indicated in this model, E3 could associate, at least temporarily, with the membrane. Cleavage of the PE2 protein to form the mature E2 protein requires the presence of E1 protein (Cutler & Garoff, 1986). Assembly of the heterodimeric spikes consisting of the E1 and E2 proteins may be initiated in the form of PE2 and E1 proteins in the rough endoplasmic reticulum and then cotransported to the Golgi apparatus. Association of PE2 and E1 proteins also exposes the cleavage site between E3 and E2 and may then be cleaved by the protease.

Several conserved domains present in the E1 protein may be important in assembly of the heterodimeric spikes consisting of the E1 and E2 proteins. The E1 protein is less hydrophilic than the E2 protein and may form the inner core of the heterodimeric spikes. Conserved domains found in the E1 protein may be involved in intrapeptide or interpeptide interactions that maintain the integrity of the alphavirus spike. The E2 protein, on the other hand, has less overall homology with other alphaviruses (42-1%) than the E1 protein (54-02%) and possesses the major antigenic determinants of the alphavirus. Most of the E2 protein regions may be exposed on the outer surface of the glycoprotein spike, which could explain why there is a more rapid evolution of E2 due to outside selection pressures such as host immune surveillance and vector competence. This structure–function relationship of E2 proteins in alphavirus evolution may explain why they are specifically involved in generation of strain diversity (Strauss & Strauss, 1986).
Bell et al. (1984) developed an evolutionary tree to relate eight alphaviruses, representing five antigenic complexes, based on the first 50 amino acids at the amino-terminal end of the E1 protein. This analysis imposes a potential bias in the comparisons. The 50 amino acid sequence of E1, which was used by Bell et al. (1984) for the comparisons, may be too short to represent the true divergence of the proteins in question. Our data show that VEE and EEE viruses are the most closely related alphaviruses when the entire E1 protein sequences of EEE, VEE, SIN, SF and RR viruses are compared (Table 3).

As an alternative approach, we compared the capsid protein sequences of five alphaviruses: EEE, VEE, SIN, SF and RR. We selected this protein for two reasons (Table 3). First, the capsid is an internal structural protein of the alphavirus, and selection pressures exerted on it should be the same for all alphaviruses. Second, the capsid protein possesses two domains, the first having less sequence homology than the other which is highly conserved. The evolutionary tree relating these five viruses was built by sequentially adding branches (Bell et al., 1984). EEE and VEE viruses have been placed in the same major branch and separated from SIN, SF and RR viruses on the other branch. The geographical distribution and the disease patterns of viruses on the two different branches also differ. EEE and VEE viruses occur only in the western hemisphere and cause encephalitis. On the other hand, SIN, SF and RR viruses are isolated only from the eastern hemisphere and cause rash, fever and polyarthritis (Shope, 1985). Western equine encephalitis (WEE) and SIN viruses are closely related serologically. However, WEE virus, like EEE and VEE viruses, is only found in the western hemisphere where it causes encephalitis. It is possible that alphaviruses were introduced into the western hemisphere through SIN. The tendency toward neurotropism may have occurred after the evolutionary divergence of the viruses in question.

### Table 3. Homology between capsid and E1 proteins of alphaviruses

<table>
<thead>
<tr>
<th>Capsid/capsid</th>
<th>EEE</th>
<th>VEE</th>
<th>SIN</th>
<th>RR</th>
<th>SF</th>
</tr>
</thead>
<tbody>
<tr>
<td>EEE</td>
<td>62*</td>
<td>50</td>
<td>49</td>
<td>50</td>
<td>50</td>
</tr>
<tr>
<td>VEE</td>
<td>57</td>
<td>50</td>
<td>43</td>
<td>43</td>
<td>46</td>
</tr>
<tr>
<td>SIN</td>
<td>51</td>
<td>50</td>
<td>54</td>
<td>57</td>
<td></td>
</tr>
<tr>
<td>RR</td>
<td>52</td>
<td>53</td>
<td>54</td>
<td>57</td>
<td>80</td>
</tr>
<tr>
<td>SF</td>
<td>55</td>
<td>51</td>
<td>69</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

* Percentage amino acid conservation.

**REFERENCES**


*(Received 19 November 1986)*