Key words: flavivirus/Kunjin virus/non-structural proteins

Gene Mapping and Positive Identification of the Non-structural Proteins NS2A, NS2B, NS3, NS4B and NS5 of the Flavivirus Kunjin and Their Cleavage Sites

By G. SPEIGHT, G. COIA, M. D. PARKER† AND E. G. WESTAWAY*

Department of Microbiology, Monash University, Clayton, Victoria 3168, Australia

(Accepted 7 August 1987)

SUMMARY

Partial N-terminal amino acid analyses of five radiolabelled non-structural (ns) proteins specified by Kunjin (KUN) virus provided positive identification of NS3, NS5 and three previously hypothetical ns proteins of flaviviruses, ns2a, ns2b and ns4b. Their correct gene order was obtained from their deduced amino acid sequences. Thus the gene order for KUN virus relative to that proposed for yellow fever (YF) virus was as follows: KUN 5'...GP44·P19·P10·P71·(?)·P21·P98-3', YF 5'...NS1·ns2a·ns2b·NS3·ns4a·ns4b·NS5-3'. The identity of GP44 as NS1 was assumed from the known nucleotide and deduced amino acid sequences; ns4a was not identified. The cleavage sites in the polyprotein for KUN NS2B, NS3 and NS5 were identical, Lys-Arg-Gly, similar in form to the sequence Arg-Arg-Ser defined at the cleavage sites of YF NS3 and NS5. A new consensus cleavage site for NS1, NS2A and NS4B in the form Val-X-Ala where X is any one of several uncharged amino acids, was found at corresponding sites homologous to those of KUN virus in all published flavivirus sequences (a total of 18 sites). NS1 and NS4B, but not NS2A, were preceded by a putative signal sequence.

INTRODUCTION

Flaviviruses have only recently been raised to the status of a new family of enveloped RNA viruses, the Flaviviridae (Westaway et al., 1985), comprising about 65 species. The virions are 45 to 50 nm in diameter, and the plus sense RNA genome of about 4 x 10⁶ Mr functions as the only viral mRNA. Kunjin (KUN) virus is a member of the serological complex represented by Japanese encephalitis (JE), Murray Valley encephalitis (MVE), St Louis encephalitis (SLE) and West Nile (WN) viruses (Westaway et al., 1977); it is distributed widely in Australia often in association with MVE virus (Marshall et al., 1982) but is only rarely associated with clinical infection (Muller et al., 1986).

Complete nucleotide sequences of the flavivirus coding region are available only for yellow fever (YF) and WN viruses. These sequences established the presence of one long open reading frame of over 10 kilobases (Rice et al., 1985; Castle et al., 1985, 1986; Wengler et al., 1985). The gene order was described by Rice et al. (1985) as 5'·C-prM-E-NS1·ns2a·ns2b·NS3·ns4a·ns4b·NS5·3'. PrM is the precursor of the membrane protein M. N-terminal amino acid analyses of envelope protein (E), core protein (C) and M, and of the non-structural (ns) proteins NS1, NS3 and NS5 established their cleavage sites in the YF virus polyprotein (Rice et al., 1985; Bell et al., 1985). NS1, NS3 and NS5 correspond to the ns proteins originally described as NV3, NV4 and NV5 respectively (Shapiro et al., 1971; Westaway, 1973).

The four gene products of YF virus described by Rice et al. (1985) as ns2a, ns2b, ns4a and ns4b have not been identified. Their existence was postulated on the basis of sizes of several ns proteins radiolabelled in infected cells and on putative cleavage sites in the polyprotein. Despite

† Present address: Veterinary Infectious Diseases Organization, University of Saskatchewan, 124 Veterinary Road, Saskatoon, Saskatchewan, Canada S7N 0W0.
this uncertainty, cleavage sites and nucleotide and amino acid sequences purporting to represent such genes have been described also for ns2a and ns2b of MVE virus (Dalgarno et al., 1986) and of SLE virus (Trent et al., 1987) and for ns2a, ns2b and ns4b of dengue-2 (DEN-2) virus (Yaegashi et al., 1986; Rice et al., 1986b). Candidate proteins for such genes for KUN virus would be P10, P19 and P21 which were shown to be unique products by tryptic peptide mapping (Wright & Westaway, 1977). Previously P19 was assumed incorrectly to be identical with the glycoprotein NV2 which migrated similarly in some polyacrylamide gels (Westaway & Shew, 1977; Wright & Westaway, 1977), but was recently found to be separable and distinct from NV2 by use of concanavalin A-Sepharose 4B and peptide mapping (Crawford & Wright, 1987). NV2 was positively identified only recently as prM for WN and YF viruses (Castle et al., 1985; Rice et al., 1985). The fourth candidate protein, of Mr about 30 000 (originally described as NVX; Westaway, 1973), is often labelled poorly in flavivirus-infected cytoplasm and no peptide maps are available. A further flavivirus product in infected cells of Mr about 14 000 (NV1½) was shown to represent the core protein C with a methionine-containing tryptic peptide deleted (Wright & Westaway, 1977; Wengler et al., 1979).

In this paper we positively identify KUN virus-specified proteins which correspond to the postulated genes ns2a, ns2b and ns4b, and show that two of the identified cleavage sites differ from those assumed by others.

METHODS

Cells and virus. Details of KUN virus strain MRM61C and Vero cells have been described previously (Westaway, 1973).

Preparation of radiolabelled KUN virus ns proteins. Confluent Vero cells were infected with KUN virus (m.o.i. > 30), or were mock-infected. At 24 h post-infection the medium was replaced with Eagle's MEM deficient in the radiolabelling amino acid and supplemented with 0·1% bovine serum albumin and 3 μg/ml actinomycin D (generously supplied by Merck, Sharpe & Dohme, Sydney, Australia). After a further 2 h the medium was replaced with the above medium plus the radioactively labelled amino acid, and the cells were incubated for a further 6 h. At the end of this time the medium was removed and the cells were washed with phosphate-buffered saline. The labelled virus-specified proteins were recovered from the cell monolayers either by extraction in 2% SDS or by resuspending the cells in lysis buffer [0·01 M-NaCl, 0·01 M-Tris-HCl pH 7·4, 1·5 mM-MgCl₂, 2 mM-phenylmethylsulphonyl fluoride (PMSF)]. The cells in lysis buffer were disrupted by 10 passages through a 26-gauge needle and the nuclei removed by low speed centrifugation (2000 g for 2 min). The supernatant was then pelleted at 16 000 g for 10 min and the pellet dissolved in 2% SDS. This fraction contained most of the virus-specified proteins with relatively little radiolabelled cell-specified protein.

[35S]Methionine-labelled, virus-infected cytoplasm was added to 3H-labelled cytoplasm to enable visualization of the viral protein bands after electrophoresis in SDS–polyacrylamide gels or phosphate gels (Laemmli, 1970; Westaway, 1973) and exposure of the wet gels to Kodak XAR-5 X-ray film. Gel slices corresponding to the viral proteins were excised and eluted with 0·1% SDS containing PMSF. Homogeneity of the labelled viral proteins was confirmed by analysis on 10% polyacrylamide gels together with appropriate markers (Fig. 1). The amino acids used to radiolabel the viral proteins were purchased from Amersham Australia: L-[4,S-3H]isoleucine (93 Ci/mmol), L-[4,5-3H]leucine (133 Ci/mmol), L-[4,5-3H]lysine (79 Ci/mmol), L-[35S]methionine (1300 Ci/mmol), L-[2,3-3H]valine (48 Ci/mmol), or from New England Nuclear: L-[3-3H]alanine (70 Ci/mmol).

N-terminal amino acid sequencing. Labelled viral proteins were analysed by Edman degradation (Edman & Begg, 1967) using an Applied Biosystems model 470A gas phase sequenator. After the appropriate number of cycles the fractions were counted in a Packard 300CD liquid scintillation spectrometer.

Cloning and nucleotide sequencing of the KUN virus genome. Details are published elsewhere (Coia et al., 1988) but a brief description follows. RNA from purified virions was copied by reverse transcriptase using a synthetic oligonucleotide 12-mer primer, complementary to the 5' terminus of the RNA of the closely related WN virus (Wengler & Wengler, 1981). Second strand cDNA synthesis was achieved by the method of Okayama & Berg (1982). The ds cDNA was tailed by an extension of about 20 residues of dC to the Y ends and ligated to the vector pBR322 cut at the PstI site and tailed with dG. After transformation in Escherichia coli strain DH1, plasmids containing viral cDNA inserts were identified and restriction maps prepared which showed that almost the entire genome was represented in the cDNA library. Clones of 1200 to 2800 bases were sequenced by the dideoxy method (Sanger et al., 1977) after subcloning of fragments into the phage M13. An open reading frame was identified which extended over all of the ns genes discussed in this paper.
New cleavage sites of flavivirus proteins

Fig. 1. Autoradiograms of purified KUN virus ns proteins separated by gel electrophoresis in (a) a 10% SDS discontinuous gel and (b) a 10% SDS-phosphate gel. In each gel, lanes 1 to 7 contain KUN virus-infected cytoplasm, P98, P71, P21, P19, P10 and mock-infected cytoplasm respectively. The $M_r \times 10^{-3}$ of KUN virus ns proteins are indicated by P98 etc.; GP96 could not be resolved in infected cytoplasm. GP20 has recently been identified as prM, the precursor of membrane (M) protein (Rice et al., 1985; Castle et al., 1985). The structural proteins E, C and M are represented in cytoplasm by P51(E) and P14(C), and by the M precursor, prM or GP20 (Rice et al., 1985; Castle et al., 1985; Crawford & Wright, 1987). The ns proteins were radiolabelled in infected cells with $[^{35}S]$methionine and eluted from gels after electrophoresis, as described in Methods.

RESULTS

Cloning and sequencing of cDNA of KUN virus RNA

Seven overlapping clones were selected by colony blot hybridization assays (Maniatis et al., 1982) and by restriction enzyme mapping to determine their relative locations and polarities. Clone pKV62 (the largest) was one of several of similar size which all cross-hybridized, and it was assumed (correctly) to represent the distal 3' region of the genome. Subsequent nucleotide sequencing after pKV62 was subcloned into phage M13 showed that a single long open reading frame terminated in three closely spaced stop codons and continued in the 3' direction for approximately another 300 nucleotides. The other six clones were mapped and sequenced, and found to extend in a continuous sequence of over nine kilobases with one open reading frame (Coia et al., 1988).

Preparation and N-terminal amino acid analyses of KUN virus ns protein

Five ns proteins of KUN virus, namely P98, P71, P21, P19 and P10, were labelled with individual radioactive amino acids and the purified proteins subjected to Edman degradation to determine the position of the labelled residue with respect to the N terminus of the protein. We were unable to locate in gels any labelled protein bands in the expected positions of GP44 (NS1,
Fig. 2. Amino-terminal sequence analysis of KUN virus ns proteins radiolabelled and purified as described in Fig. 1 and in Methods. Each preparation was applied to a gas phase sequenator as a mixture of $[^{35}S]$methionine and single $^3$H-amino acid-labelled protein. Residues were collected sequentially and counted for radioactivity in a liquid scintillation spectrometer; the scale on the y axis represents c.p.m. × $10^{-3}$. The number of cycles used for analyses varied. The amino acid sequence in the single-letter code shown at the bottom of each frame was located in the polyprotein translated from the KUN nucleotide sequence (Coia et al., 1988) by comparison with the positions of the radioactive residues. The identified residues are underlined. No $[^{35}S]$methionine peak appeared in the first 16 amino acid residues of P71 and P98. The assigned gene for each ns protein (Rice et al., 1985) is shown in parentheses: (a) P10 (NS2B), (b) P71 (NS3), (c) P98 (NS5), (d) P19 (NS2A), (e) P21 (NS4B).

formerly NV3 or the soluble complement-fixing antigen; Smith & Wright, 1985) or of p32 (NVX) (Westaway, 1973). The radiolabelled ns proteins purified as in Fig. 1 were sequenced from their N termini in the gas phase sequenator, and the collected cleaved residues containing $[^{35}S]$methionine or specific $^3$H-labelled amino acids were identified by liquid scintillation counting (Fig. 2). In the initial analyses, co-migrating host proteins labelled with the same $^3$H-labelled amino acids prepared in the same way as KUN virus ns proteins were identified in autoradiograms by addition to mock-infected cell cytoplasm of $[^{35}S]$methionine-labelled, virus-infected cytoplasm. None of these $^3$H-labelled cell proteins when purified and analysed as for the viral ns proteins yielded peaks of radioactivity which might be confused with those from purified ns proteins (results not shown). Because infected cytoplasm $^3$H-labelled with amino acids had been mixed with corresponding $[^{35}S]$methionine-labelled cytoplasm (see Methods) it was possible to show that any methionine residue in the N-terminal sequence was reproducibly recovered in the correct position during sequencing.

Location of each of the ns proteins in the polyprotein sequence

The locations of P71 and P98 within the polyprotein sequence were readily identified by comparison of the deduced amino acid sequence with that of WN and YF viruses (Rice et al., 1985; Castle et al., 1986). They were equivalent to NS3 and NS5 respectively (Rice et al., 1985). The N termini were identified by amino acid sequencing and found to be in identical positions within the deduced polyprotein sequence to the N termini of NS3 and NS5 of YF and DEN-2
New cleavage sites of flavivirus proteins

Fig. 3. Nucleotide sequences and the deduced amino acid sequences at the polyprotein cleavage sites of all the KUN virus ns proteins. The amino acid residues identified by N-terminal amino acid analyses (Fig. 2) are doubly underlined. The strongly enriched hydrophobic core of the amino acid sequence preceding a carboxy region of five residues (von Heijne, 1985) which includes the cleavage site of NS1 and of NS4B is underlined. The cleavage sites of NS1 and ns4a, indicated by open arrowheads, are assumed (see text).

viruses (Rice et al., 1985, 1986a; Biedrzycka et al., 1987). The matched nucleotide and amino acid sequences for KUN virus at the cleavage sites are shown in Fig. 2 and 3. The deduced amino acid sequence of NS1 (KUN GP44) was readily recognized from its close similarity to that of NS1 of WN virus (Wengler et al., 1985), and the KUN NS1 N terminus was located by its homology with the known NS1 N terminus of YF and SLE viruses (Rice et al., 1986a).

Although the locations in the flavivirus genome for the coding sequences of small unique ns products such as P10, P19 and P21 were unknown (Rice et al., 1985, 1986b), searches within the intergenic regions NS1-NS3 and NS3-NS5 located deduced amino acid sequences which matched their N-terminal amino acid analyses (Fig. 2, 3). The sizes of these three ns proteins could be accommodated within the available coding regions. P10, P19 and P21 were thus identified unambiguously as NS2B, NS2A and NS4B, respectively.

The gene order established for all the KUN ns proteins analysed is shown schematically in Fig. 4(a, b). Table 1 summarizes the gene products, including their measured and predicted sizes, the relationship of the three systems of nomenclature for flavivirus ns proteins, and the cleavage sites. These items are discussed below.

Characterization of cleavage sites

The cleavage site at the N terminus for each of NS2B, NS3 and NS5, and the putative site for ns4a (Fig. 3) included a pair of basic amino acids (Lys-Arg) in the -2, -1 positions relative to the point of cleavage, followed by a short side chain amino acid (Gly or Ser) in the +1 position. These sites were similar to those established by Rice et al. (1985) for NS3 and NS5 of YF virus, and it was proposed that sites of this form are recognized by a possibly virus-encoded protease for ns proteins, which is probably located in the cytosol. The amino acid preceding the Lys-Arg pair was variable, Arg for NS2B, Thr for NS3, Gly for ns4a and Leu for NS5.

Cleavage sites for NS1, NS2A and NS4B (Fig. 3) were all preceded by the sequence Val-X-Ala, where X is His, Asn, or Ala, respectively. This site conformed to the (-3, -1) rule for a sequence immediately adjacent to the signal peptidase cleavage site (Perlman & Halvorson, 1983; von Heijne, 1983, 1984, 1985). A small amino acid (Ala, Gly, Ser, Cys, Thr or Gln) must be at -1, the residue at -3 must not be an aromatic, charged, or large polar amino acid, and no Pro must occur in the -3 to +1 positions. NS1 was preceded by an enriched hydrophobic amino acid sequence bounded by Arg at -23, completing the criteria required for a signal sequence. Rice et al. (1985, 1986b) noted a similar sequence for NS1 of YF virus, and proposed that the hydrophobic sequence could act as a signal sequence for translocation of NS1 across the
Fig. 4. Locations of KUN virus ns proteins in the polyprotein sequence (a, b) and comparisons with YF virus (c, d). The hatched areas in (b) and (c) correspond to ns products which have not been identified. In (a) the number of amino acid residues translated from the nucleotide sequence of each gene (see Table 1) is plotted to scale, commencing with the first amino acid of NS1, through to the stop codon following ns amino acid residue 2642. In (b), the known N-terminal cleavage sites (Table 1) are shown, plus the postulated cleavage site for ns4a. The black boxes preceding the carboxyl termini of NS2A, NS2B, NS4B and NS5 represent the polypeptide sequences of 5000 to 6000 Mr, which appear to be missing from the respective gene products (see Table 1). In (c) the ns region of the protein of YF virus (amino acid residues 1 to 2633), the number of amino acid residues translatable from each gene, and the cleavage sites are shown, as specified by Rice et al. (1985). In (d), the YF sequence has been redrawn, with two new cleavage sites shown for NS2A and NS4B which appear to be homologous with those of KUN virus.

Table 1. Cleavage sites and identity of each KUN virus protein in the translation sequence of the non-structural proteins

<table>
<thead>
<tr>
<th>NS1</th>
<th>NS2A</th>
<th>NS2B</th>
<th>NS3</th>
<th>ns4a</th>
<th>NS4B</th>
<th>NS5</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>E</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>352</td>
<td>231</td>
<td>131</td>
<td>619</td>
<td>149</td>
<td>255</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(b)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>E</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>GP46</td>
<td>P19</td>
<td>P10</td>
<td>P71</td>
<td>P21</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(c)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>E</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>409</td>
<td>167</td>
<td>130</td>
<td>623</td>
<td>267</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(d)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>E</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>352</td>
<td>224</td>
<td>130</td>
<td>623</td>
<td>149</td>
<td>250</td>
</tr>
</tbody>
</table>

* Nomenclature of ns proteins is based on 1, gene sequence (Rice et al., 1985), 2, relative size (Westaway, 1973) and 3, the measured size in Mr x 10^-3 (Westaway et al., 1980). GP indicates glycoprotein.
† Expressed as Mr x 10^-3 of the deduced amino acid sequence between the cleavage sites shown.
‡ Commencing with the N-terminal amino acid of NS1.
§ Based on correspondence of deduced amino acid sequence with N-terminal amino acid analyses (Fig. 2). The cleavage site for NS1 is based on homology with the location and the established N terminus of NS1 of YF and SLE viruses (Rice et al., 1986a). The cleavage site shown for ns4a is hypothetical, based on the measured size of the preceding product NS3 and on the presence at the approximate carboxy terminus of NS3 of a pair of basic amino acids flanked by short side chain amino acids (Rice et al., 1986a).
| The distinction between the glycoprotein NV2 (GP20 or prM, the precursor of the structural protein M) and the similarly migrating P19 has only recently become apparent (refer to the text).
¶ No KUN virus-specified protein corresponding approximately in size to the putative ns4a has been identified in infected cells, other than the intracellular equivalent of the virion core protein C (14000 Mr).
New cleavage sites of flavivirus proteins

**NS2B**

| KUN          | LVACDPNKRKR | GWPATEVMTA | LQYTKR | GGVL |
| WN           | M........    | ............ | ........ | ...... |
| MVE          | MI.N.K....... | ........... | K....... | F........ |
| SLE          | ILAH.G....... | S......S.... | L.G | GKHSS....... | A........ |
| YF           | FL.TRIFGR..... | SI.VN.A.A. | VRGAR. | S......DV |
| DEN-2        | TTLSERT.K..... | S......LN.AIM. | EVKKQ. | A...... |

**NS3**

| KUN          | LQYTKR | GGVL |
| WN           | ........ | ...... |
| MVE          | ........ | ...... |
| SLE          | ........ | ...... |
| YF           | ........ | ...... |
| DEN-2        | ........ | ...... |

**ns4a**

| KUN          | SFKDFASGKR | SQIGFIEVLG | KPGGLKR | GGAK |
| WN           | ............ | ............ | ........ | ...... |
| YF           | E.IK.E.R. | GAAEVLV.S | .T.-R. | .S.N |
| DEN-2        | ............ | ............ | TTNTR. | .TGN |

**NS5**

| KUN          | KPGGLKR | GGAK |
| WN           | ........ | ...... |
| YF           | ........ | ...... |
| DEN-2        | ........ | ...... |

Fig. 5 Comparison of cleavage sites of flavivirus ns proteins involving a pair of basic amino acids. The flanking amino acids are shown for NS2B, NS3, ns4a and NS5. Note that the sequence of the pair of basic amino acids preceding each cleavage site is Lys-Arg for the KUN-WN-MVE-SLE virus serological complex (Westaway et al., 1977), and Arg-Arg for YF virus. Where the amino acid residue in position −3 is not a short side chain amino acid, a cluster of three basic amino acids occurs. Sources of sequence data are as follows: WN virus (Castle et al., 1986), MVE virus (Dalgaro et al., 1986), JE virus (Sumiyoshi et al., 1986), SLE virus (Trent et al., 1987), YF virus (Rice et al., 1985) and DEN-2 virus (Biedrzycka et al., 1987; P. J. Wright, personal communication).

endoplasmic reticulum (ER). Hence KUN NS1 appears likely to be cleaved by signal peptidase.

NS4B was preceded by NS3 and the hypothetical ns4a, which are probably cleaved in the cytosol (see above). A strongly enriched hydrophobic sequence was present in the −6 to −23 amino acid domain preceding the N terminus of NS4B, which could function as a signal recognition sequence for translocation. Furthermore, all the residues in the −5 through to the +5 positions were compatible with a postulated signal peptidase site, the only caveat being that Ala in the −2 position is relatively rare (von Heijne, 1984, 1985). Hence signal peptidase may cleave NS4B as well as NS1.

NS2A was preceded by NS1 which was presumably inserted into the ER via a translocation sequence, as noted above. No stop transfer sequence (see Sabatini et al., 1982) occurred within NS1 before the putative cleavage site Val-His-Ala (Fig. 3; Coia et al., 1988). Asn and Asp in the +2 and +4 positions relative to the putative cleavage site provided enrichment of charged and large polar amino acid residues which seemed to occur in the +1 to +5 sequence after a signal peptidase cleavage site (von Heijne, 1984). Because the enzyme is apparently located on the ER within the lumen (Sabatini et al., 1982) it may be able to recognize the Val-His-Ala sequence at the C terminus of NS1 as the nascent polypeptide emerges within the ER even though it is not immediately preceded by a hydrophobic (translocation) sequence. Alternatively, a virus-specific protease may cleave NS2A within the lumen.

**Comparisons with cleavage sites and identification with ns proteins of other flaviviruses**

The flanking amino acid sequences of the cleavage sites of the seven KUN virus ns proteins are compared with available sequences of other flaviviruses in Fig. 5 and 6. The latter sequences were chosen to represent the same regions of the genome as shown for KUN virus and modified for YF virus in Fig. 4(a), (b) and (d). This selection entailed movement of the cleavage site by 56 amino acid residues for NS2A and by 37 residues for NS4B, in both cases towards the N terminus, from the hypothetical sites proposed as a working hypothesis by Rice et al. (1985) which has been adopted by other authors. For the cleavage sites for NS2B, NS3, NS5 and the hypothetical ns4a, no changes were made. These four sites all involved a pair of basic amino acids (Fig. 5). However, the increase indicated in coding content of the NS2A and NS4B genes resulted in a reduction in size of the coding region for NS1 and ns4a, producing corresponding theoretical polypeptides of about 40000 and 16000 M_r respectively (Table 1). NS1 contains two
Fig. 6. Proposed cleavage sites of the flavivirus ns proteins NS1, NS2A and NS4B. The flanking amino acid sequences for each ns protein show the regions of greatest homology for each product, and the presence (underlined in NS1 and NS4B) of a strongly enriched hydrophobic sequence of 17 to 19 residues prior to the −5 position of the cleavage site and bounded at its N terminus by a basic residue (Arg). These are characteristic features of the hydrophobic core region of eukaryotic signal sequences (von Heijne, 1985). The consensus cleavage site Val-X-Ala appears to be displaced by three residues to the left for NS4B of DEN-2 virus (see text). Additional sources of data to those shown for Fig. 5 are for DEN-2 (Deubel et al., 1986) and DEN-4 (Zhao et al., 1986).

Inspection of Fig. 4 and Table 1 shows that all the flavivirus ns proteins could be accommodated between the proposed cleavage sites. The KUN proteins now identified as NS2A, NS2B, NS3, NS4B and NS5 were all found to be unique by tryptic peptide mapping (Wright et al., 1977; Wright & Westaway, 1977), and NS1 has been well characterized as a unique flavivirus product (Smith & Wright, 1985). A comparison was recently made of the sizes of ns proteins of eight flaviviruses (Westaway, 1987) and in almost all cases they can be matched to the sizes of the KUN virus ns proteins listed in Table 1. We believe that these represent all of the known unique products of the ns region of the flavivirus genome. The only significant gap in the coding region is represented by the hypothetical ns4a. Other products seen in some gels are (i) p79-82 or NV4½ which contains the amino acid sequences of NS3 (Svitkin et al., 1981) and hence is an aberrant (uncleaved) product, (ii) p30-32 (NVX) which is too large to represent ns4a and hence must also represent uncleaved products, or a degradation product, (iii) several small glycoproteins in the size range 13000 to 22000 Mₚ (Smith & Wright, 1985; Schlesinger et al., 1983) which are probably related to prM or the glycosylated residue of cleaved prM, and (iv) intracellular core protein C, Mᵦ about 14000.

Cleavage sites involving a pair of basic amino acids

Inspection of Fig. 5 shows that Lys-Arg was found at the cleavage site of NS2B, NS3, ns4a and NS5 of all members of the KUN–MVE–WN–SLE serological subgroup, and Arg-Arg at each such site of YF virus. Data are available only for NS3 and NS5 of DEN-2 virus; for NS5, the cleavage site is Arg-Arg↓, but for NS3 the site is Lys-Lys-Gln-Arg↓. Biedrzycka et al. (1987)
suggest that such a cluster may still be an acceptable site for the virus-encoded protease proposed by Rice et al. (1985). It is noteworthy that at the cleavage site for NS2B of KUN, WN, DEN-2 and MVE virus polyproteins, a cluster of three basic amino acids occurs, flanked by Asn at their N terminus. In all the previous studies, the residue at the −1 position with respect to the point of cleavage is Arg, and the residue in the +1 position is always Gly, Ala or Ser. The site proposed for cleavage of ns4a conformed to these criteria.

### Cleavage sites not involving basic amino acids

The cleavage site for NS1, and the two newly characterized sites for NS2A and NS4B, have already been discussed for KUN virus, and include Val-X-Ala in the −1 to −3 position. When the three corresponding locations in the polyprotein are examined for other flaviviruses (Fig. 6) they all have the same sequence i.e. Val-X-Ala, except for DEN-2 NS4B where it is located three residues upstream. The residue X includes the uncharged amino acids Ala, Asn, Gln, Gly, His, Thr and Val. The flanking regions at the proposed cleavage sites of NS1 and NS4B conformed to the prescribed signal recognition and signal peptidase sites (Perlman & Halvorson, 1983; von Heijne, 1983, 1984, 1985). A charged residue (Arg or Lys) was followed immediately by an enriched hydrophobic core sequence of 17 to 19 residues plus a carboxyl or c-sequence of five residues (−5 to −1) which included at least one polar amino acid and conformed to the (−3, −1) rule noted earlier. The five residues beyond the cleavage site (+1 to +5) exhibited 80% or higher homology, the only exceptions being NS1 of DEN-2 virus and NS4B of YF virus with 60% homology. Again these residues appeared to be acceptable within the context of signal sequences.

For NS2A of all the flaviviruses, a possible cleavage site Val-X-Ala may be exposed to signal peptidase or a virus-specified protease within the lumen of the ER, as discussed earlier for KUN virus NS2A. In all cases there is no stop transfer sequence or translocation sequence within NS1 before the proposed Val-X-Ala site (see Fig. 6 and sequence comparisons in Trent et al., 1987) and charged or large polar residues occur in the +2 and +4 positions, except for NS2A of SLE virus which has three small polar residues (all Gly) in the +1 to +5 domain.

### Apparent truncation of NS2A, NS2B, NS4B and NS5

The size of four of the five identifiable KUN virus ns proteins, as measured by comparison with migration in gels of molecular weight standards, appeared to be significantly smaller than their predicted sizes obtained by translation of the relevant nucleotide sequences (Table 1, and see black boxes in Fig. 4b). Furthermore, NS1 of KUN virus and some other flaviviruses is present only in trace or negligible amounts in the cytoplasm (Westaway, 1973; Westaway et al., 1977; Wright & Westaway, 1977; Wengler et al., 1979), and the apparently correct size of the NS3 gene product is based on the assumption that the hypothetical gene ns4a commences where shown in this and other flavivirus reports. A search of the deduced KUN virus amino acid sequence for additional cleavage sites in the form of a pair of basic amino acids followed by Gly, Ser or Ala revealed six such sites but none of these was near the estimated position of the C terminus of any of the apparently truncated ns proteins. A similar search for sequences in the form Val-X-Ala in these regions was also unproductive. Furthermore we searched for radiolabelled proteins smaller than 10000 Mr in infected cytoplasm without success (results not shown).

### DISCUSSION

These results modify the working hypothesis of Rice et al. (1985) for the gene map of the flavivirus ns proteins by redefining the bounds of several of the ns genes. Even more importantly, positive identification of the gene products NS2A, NS2B and NS4B was achieved. The sequence of proteins shown in Fig. 4(b) appears to represent all the unique products of the ns region of the flavivirus genome (the hypothetical protein ns4a remains unidentified). All the KUN virus ns proteins described previously (Westaway, 1973; see Table 1) have now been allocated to individual genes, except for p30-32 (NVX). However, the latter is radiolabelled in only small amounts in infected cells, and until a peptide map is obtained its origin remains
obscure. It could not be accommodated within the genetic map as a unique or single gene product. Additional small ns proteins or glycoproteins described for several flaviviruses are probably related to the established products.

Although the gene map is now almost completely defined, some puzzling features remain, e.g. the identity of ns4a, and the apparent truncation of NS2A, NS2B, NS4B and NS5, for which we can offer no explanation at present. The truncated products have been observed too consistently to be accounted for by abnormal electrophoretic migration in sizing gels. It would require very active carboxypeptidases to remove such large tracts of their C termini. Steiner et al. (1980) have suggested that the signal peptidase which cleaves preproinsulin must be associated with a variety of membrane-associated endo- and exopeptidases because the prepeptide of proinsulin is rapidly fragmented and degraded. However, the C termini of NS2A, NS2B and NS4B deduced from translation of the nucleotide sequence are probably cleaved in the cytosol, being characterized by a pair of basic amino acids rather than by signal sequences (Fig. 4). Interestingly, proteins NS2A and NS2B of YF, MVE and KUN viruses are strongly hydrophobic and appear to be truncated, whereas NS3 and NS5 are fairly hydrophilic (Rice et al., 1985; Dalgarno et al., 1986; Coia et al., 1988) but only NS3 is not truncated. The possibility should be noted that the missing product ns4a could represent a large truncated (and degraded) terminus of NS3.

The presented modifications in the gene map have significant consequences in regard to the identity of the protease(s) cleaving the flavivirus polyprotein in the ns region. Of the three cleavage sites preceded by the sequence Val-X-Ala, that for NS1 appears to be a legitimate signal peptidase site (von Heijne, 1984, 1985). The cleavage site for NS4B may be in the same category, but that for NS2A is conjectural. There appears to be no precedent for sequential cleavage within the lumen of the ER, i.e. in the absence of a stop transfer and translocation sequence before the next cleavage site is translated. Alternatively, a virus-specified protease may recognize the sequence Val-X-Ala preceding the N terminus of NS2A and possibly NS4B; also it should be emphasized that the Val-X-Ala sequence appears to precede the cleavage site at the N terminus of NS1 for seven flaviviruses in addition to KUN virus (Fig. 6). Although the NS4B cleavage site for DEN-2 virus does not appear to conform exactly (Fig. 6), the preceding sequence is in tandem for the other viruses, and exists in a similar form for DEN-2 NS4B (Val-Gly-Ala-Thr-Met-Ala). Hence in all cases for NS2A, NS4B and indeed NS1, Val-X-Ala appears to be a consensus sequence at the cleavage site.

It seems most probable that a virus-encoded protease cleaves NS2B, NS3 and NS5 after a pair of basic amino acids as proposed by Rice et al. (1985). However our data indicate that some flexibility in the requirement for small side chain amino acids is required to accommodate the variation observed in the cleavage sites for NS2B of KUN, WN and MVE viruses. Furthermore, the cleavage site for NS3 of DEN-2 virus varies even more (Val-Lys-Lys-Gln-Arg-Ala; Fig. 5). A consistent feature is that Arg is preferred to Lys on the carboxy side of the pair (Fig. 5) as noted for the cleavage sites of proproteins (Steiner et al., 1980), suggesting an evolutionary relationship with the similar cellular enzyme. These results thus imply that two virus-specified proteases are required for cleavage of the ns proteins, unless a signal peptidase in cell membranes cleaves NS2A and NS4B as well as NS1.

The new start point 56 amino acid residues upstream from the originally proposed site for NS2A (Fig. 4) means that the hydrophobic anchor region assumed to be near the C terminus of NS1 (Rice et al., 1985) does not exist as such, being incorporated into the N-terminal region of NS2A. This change has implications for the postulated role of NS1. Rice et al. (1985) suggested that membrane-associated and soluble forms of NS1 could conceivably differ by the presence or absence of the C-terminal hydrophobic segment. However, other hydrophobic regions in NS1 may be sufficient to produce the apparent membrane association of NS1 observed by indirect immunofluorescence in YF virus-infected cells using an appropriate monoclonal antibody (Gould et al., 1985), and in DEN-2 or KUN virus-infected cells (Westaway & Goodman, 1987) using the anti-GP46 antiserum of Smith & Wright (1985). The immunofluorescence results have not provided any evidence for an intracellular soluble form of NS1 in infected cells in vitro, unless it remains sequestered within the lumen of the ER.
New cleavage sites of flavivirus proteins

The introduction by Rice et al. (1985) of a system of nomenclature for the ns proteins based on the gene sequence is a logical development. Application has proved difficult hitherto for the small ns proteins because of the lack of identification of their gene locations. The rigorous identification obtained for the KUN ns proteins, and its application to several other flaviviruses, should remove the ambiguity in assignments for YF virus (Rice et al., 1985) and correct probable errors in arbitrary assignments for other flaviviruses. When questions of function for the small ns proteins are addressed, it should now be clear which gene is being analysed.

Valuable advice on amino acid sequencing was provided by Dr H. Niall. This work was supported by grants from the National Health and Medical Research Council of Australia and the William Buckland Foundation.

REFERENCES


G. SPEIGHT AND OTHERS


(Received 15 June 1987)