The Sequence of the M RNA of an Isolate of La Crosse Virus

By LEO J. GRADY,* MYRNA L. SANDERS AND WAYNE P. CAMPBELL
Wadsworth Center for Laboratories and Research, New York State Department of Health,
Empire State Plaza, Albany, New York 12201, U.S.A.

(Accepted 6 August 1987)

SUMMARY

The middle-size (M) genomic RNA of a New York State, U.S.A. isolate of La Crosse
(LAC) virus has been cloned by a random priming procedure and its nucleotide
sequence determined by the dideoxy method. The RNA was found to be 4526
nucleotides in length and to have a base composition of 34.2% U, 27.8% A, 20.6% C
and 17.4% G. There is a single, long open reading frame in the viral complementary
RNA that contains sufficient information to code for a protein of 1441 amino acids. In
these respects, as in many others, the LAC virus M RNA and its encoded protein were
very similar, if not identical, to those previously reported by other investigators for the
closely related snowshoe hare virus. The M RNAs of the two viruses show 79%
nucleotide sequence homology and 89% homology in the amino acid sequence of their
encoded proteins. Several algorithms for predicting surface residues, as well as the
Chou–Fasman rules for predicting secondary structure, were used to compare the LAC
virus and snowshoe hare virus M gene proteins. These analyses identified 39 sites on
the proteins as those most likely to be linear antigenic determinants that might
contribute to the differences between the two viruses.

INTRODUCTION

La Crosse (LAC) virus is an arbovirus belonging to the California encephalitis (CE) complex
of the family Bunyaviridae (Bishop & Shope, 1979). At the molecular level, all of the viruses of
the CE complex are very similar. The virions contain four polypeptides: the large (G1) and the
small (G2) envelope glycoproteins, a nucleocapsid protein and a minor, large protein also
associated with the nucleocapsid (Obijeski et al., 1976a; Gentsch & Bishop, 1978, 1979; Cash et
al., 1979; Fuller & Bishop, 1982; Fuller et al., 1983). Their genomes consist of three single-
stranded RNA molecules of negative polarity, each with a distinctive size (Obijeski et al.,
1976b). The gene products of the middle-size (M) RNA (the two glycoproteins and a non-
structural protein; Gentsch & Bishop, 1979; Fuller & Bishop, 1982) have been implicated in
many of the important biological characteristics of these viruses, such as (i) neutralization
(Kingsford & Hill, 1981; Gonzalez-Scarano et al., 1982; Grady et al., 1983; Kingsford et al.,
1983) including host-dependent neutralization (Grady & Kinch, 1985), (ii) haemagglutination
(Gonzalez-Scarano et al., 1982; Grady et al., 1983; Kingsford et al., 1983), (iii) cell fusion
(Gonzalez-Scarano et al., 1984), (iv) neuropathogenicity in a mouse model (Rozhon et al., 1981;
Shope et al., 1981a, b; Tignor et al., 1983; Gonzalez-Scarano et al., 1985; Janssen et al., 1986)
and (v) transmissibility by mosquitoes (Beatty et al., 1981a, b). Several of these phenomena have
been further localized to the G1 glycoprotein (Kingsford & Hill, 1981; Gonzalez-Scarano et al.,
1982, 1984, 1985; Grady et al., 1983; Kingsford et al., 1983; Grady & Kinch, 1985; Sundin et al.,
1987). Since much of this information was obtained using monoclonal antibodies against LAC
virus, the cloning and sequencing of the M genomic segment of this virus was undertaken as
another step toward further association of structure and function in this group of viruses.

At present, the sequence of the M RNA has been reported for five other members of the
family Bunyaviridae: Rift Valley fever virus and Punta Toro virus of the genus Phlebovirus
(Collett et al., 1985; Ihara et al., 1985), Hantaan virus of the genus Hantavirus (Schmaljohn et
All, 1987) and Bunyamwera virus and snowshoe hare (SSH) virus of the genus Bunyavirus (Eshita & Bishop, 1984; Lees et al., 1986). SSH virus is also a member of the CE complex and it is so similar to LAC virus serologically that the two have been considered as varieties of the same virus (Carlshier, 1983). Consequently, with the availability of the present data on LAC virus it became possible to make a detailed comparison both of the M RNA sequences of two closely related bunyaviruses and of the deduced amino acid sequences of the polyproteins which they encode. In the latter case, this has included application of methods for predicting secondary structure, hydrophilicity, accessibility and flexibility in an attempt to identify regions of the proteins likely to play a role in the antigenic differences between these two viruses.

METHODS

**Virus.** The virus employed was an isolate of LAC virus obtained by Dr M. Grayson of this laboratory in 1974 from Aedes triseriatus mosquitoes collected in Albany County, N.Y., U.S.A. It has been designated 74-32813 and its characterization and passage history in our laboratory have been described (Grady et al., 1983). The virus was grown and plaque-purified on BHK-21 cells using the Glasgow modification (Macpherson & Stoker, 1962) of Eagle's medium plus 10% foetal bovine serum.

**Isolation of RNA for cloning.** Plaque-purified virus was used to infect monolayers of BHK-21 cells at an m.o.i. of 0-1 p.f.u./cell. At the first signs of c.p.e., the cell culture supernatants were collected and the virions precipitated with polyethylene glycol (PEG) 6000. After resuspension of the PEG precipitate, the virus was purified by two rounds of centrifugation; the first employed a potassium tartrate-glycerol gradient and the second a linear sucrose gradient (Obijeski et al., 1976a). Only material having the density of whole virions (1.2 g/ml in the former, and 1.17 to 1.19 g/ml in the latter gradient) was collected. The virions were then disrupted with 1% SDS and the three genomic RNAs separated on linear sucrose gradients as described by Obijeski et al. (1976b). The peak fraction for each RNA species was collected, precipitated with ethanol and stored at -20°C until used.

**Cloning of the M RNA.** The M RNA of LAC virus was cloned by a random priming method similar to that used by Binns et al. (1985) for infectious bronchitis virus. Complementary DNA was synthesized in 50 μl reactions which contained 10 μCi of [3H]dCTP (sp. act. 25-7 Ci/mmol, New England Nuclear), 0.5 μg purified M RNA, 126 μg random oligonucleotide primer (Maniatis et al., 1982), 30 units RNAse (Promega Biotech), 200 units avian myeloblastosis virus transcriptase (Molecular Genetic Resources), 0.5 mm-dCTP, 1 mm each of dATP, dGTP, dTTP, 50 mm-Tris-HCl pH 8.1, 40 mm-KCl, 5 mm-MgCl₂, and 2 mm-dithiothreitol. After incubation for 1 h at 37°C, the reaction was stopped by adding EDTA to 40 mm. To remove unincorporated triphosphates the sample was passed over a column of Sephadex G-50 and the material eluting in the void volume collected. The products from several such reactions were pooled, brought to 0.25 M with sodium acetate and precipitated with ethanol. The precipitate was resuspended in 50 μl Tris-HCl pH 8.1 containing 40 mm-KCl, 5 mm-MgCl₂, and 2 mm-dithiothreitol and then brought to 0.15 M with NaOH. After 1 h at 65°C, the solution was neutralized with HCl. Next, the cDNA was sized on an alkaline sucrose gradient (Imaizumi et al., 1973) and those molecules > 400 nucleotides in length were pooled, extracted with phenol/chloroform and precipitated with ethanol. Second strand synthesis was accomplished with the Klenow fragment of Escherichia coli DNA polymerase I (Pharmacia P-L) primed by the hairpin loop at the 3' terminus of the first strand (Maniatis et al., 1982). The loop was then cleaved with mung bean nuclease. Aliquots from before and after nuclease treatment were centrifuged through linear alkaline sucrose gradients to verify that cleavage had occurred. The size distribution of the material used to make was calculated to determine the number of free ends. Terminal transferase (Pharmacia P-L) was then used to add 15 to 20 dC residues to the 3' ends of the double-stranded cDNA (Maniatis et al., 1982). Next, DNA from the plasmid pUC9 (Vieira & Messing, 1982) that had been cut at the PstI site and dG-tailed (Pharmacia P-L) was annealed with the dC-tailed cDNA. Finally, the plasmid with the cDNA insert was used to transform E. coli JM83 (Messing, 1979) by the CaCl₂ technique (Cohen et al., 1972). Medium containing ampicillin and X-gal was used to select recombinants. Antibiotic-resistant, white colonies were screened for M RNA-specific inserts as described by Maniatis et al. (1982). The 32P-labelled cDNA probe was prepared from M RNA using random oligonucleotide primer and reverse transcriptase (Maniatis et al., 1982).

**Sequencing.** The method of Holmes & Quigley (1981) was employed to obtain a mini-preparation of plasmid DNA from each of the positive colonies identified by the above screening procedure. Sequencing was carried out by a variation of the dyeoxy method of Sanger et al. (1977) that employs reverse transcriptase and elevated temperature to allow reading through dG-dC homopolymer regions (Graham et al., 1986). This permitted the use of the universal M13 primers to obtain sequence from both ends of the cloned inserts directly, without any need for purification of either plasmid or insert. Some clones were also sequenced by the chemical method of Maxam & Gilbert (1977) using the approach described by Ruther et al. (1981).

There were four regions of the M RNA that were not spanned by any of the clones and for which it was thus necessary to obtain sequence data directly by using 15-mer oligonucleotide primers and M RNA as template.
These regions were (i) the first 69 nucleotides at the 3' end of the M RNA, where the data of Clerx-van Haaster et al. (1982) provided the basis for the synthesis of an oligonucleotide complementary to the 3' terminus (AGTAGTGTACTACCA), (ii) nucleotides 1046 to 1068, obtained using an oligonucleotide complementary to 953 to 967 (GCCAGAGTCATGTGC), (iii) nucleotides 1431 to 1474, sequenced using an oligonucleotide complementary to 1366 to 1380 (TCATAGGAAAACCTA) and (iv) the last 32 nucleotides at the 5' end of the viral M segment obtained using an oligonucleotide complementary to 4425 to 4459 (TATTCTATTTGTAGT). Oligonucleotides were synthesized with an Applied Biosystems 381A DNA synthesizer in the Oligonucleotide Synthesis Facility of the Wadsworth Laboratories, operated by Drs A. Lobo and S. Santas. Dideoxy sequencing was carried out as described above.

**RESULTS**

**Nucleotide sequence of the M RNA of LAC virus**

The M RNA of LAC virus was cloned by a random priming procedure as described in Methods. This procedure yielded several hundred clones with cDNA inserts ranging from 100 to 600 nucleotides in length. Universal sequencing primers were employed to obtain the sequence of both strands of these clones. Overlapping regions were used to align the clones, which finally yielded 96 to 97% of the M gene sequence. More than 80% of the sequence was obtained from two or more independent clones. The remainder of the sequence was determined by direct sequencing of the RNA using synthetic oligonucleotide primers. Four regions were sequenced by the latter method: nucleotides 1 to 69 at the 3' end of the genomic RNA [a 15-mer oligonucleotide complementary to the first 15 bases was synthesized based on the data of Clerx-van Haaster et al. (1982)], nucleotides 1047 to 1068, nucleotides 1430 to 1474 and the last 32 nucleotides at the 5' end of the RNA. The clones used in determining the sequence are shown in Fig. 1.

The nucleotide sequence of the M RNA of LAC virus is given in Fig. 2 in the form of the viral complementary strand. The LAC virus M RNA was found to consist of 4526 nucleotides and, in terms of the genomic strand, to have a base composition of 34.2% U, 27.8% A, 20.6% C and 17.4% G. In length, it is only one nucleotide smaller than the M RNA of SSH virus (Eshita & Bishop, 1984). With respect to base composition, it corresponds closely to the values that have been reported for SSH (Eshita & Bishop, 1984) and Bunyamwera (Lees et al., 1986) viruses.

As might be expected from the available data for other bunyaviruses, the LAC virus M RNA exhibits complementarity between its 3' and 5' ends. In this case, except for a U-G mismatch at positions 9 and 4518, there is complete homology over the first 23 nucleotides at each end. The U-G mispairing which occurs has also been detected in the M RNAs of SSH (Eshita & Bishop, 1984) and Bunyamwera (Lees et al., 1986) viruses, as well as in the small genomic RNAs of a number of bunyaviruses (Akashi & Bishop, 1983; Akashi et al., 1984; Cabradilla et al., 1983). Its significance, if any, is still unknown.

In the viral complementary strand, Clerx-van Haaster et al. (1982) found earlier that the M RNAs of SSH virus and two isolates of LAC virus possessed non-coding regions of 61 nucleotides at their 5' ends. It was also evident from their results that the nucleotides flanking the AUG initiation codon were identical in these viruses. Both observations have been confirmed by the present data. Furthermore, when compared over the first 200 nucleotides at the 5' end of the viral complementary RNA, which corresponds to the region sequenced by Clerx-van Haaster et al. (1982), the only difference between the N.Y. isolate used in their study and the N.Y. isolate 74-32813 employed in the present investigation was a single base change at position 94.

At the 5' end of the viral complementary strand, a 141 nucleotide non-coding region was detected in LAC virus, whereas the corresponding tract in SSH virus consists of 142 nucleotides.
Fig. 1. Clones used to obtain the sequence of the M RNA of LAC virus. Sequence data were obtained from both strands of each clone. D is used to indicate where information was obtained by direct sequencing of the RNA using synthetic oligonucleotides as primers.

(Eshita & Bishop, 1984). The purine-rich portion of this segment, located at approximately 4453 to 4470 and suggested as a possible transcription termination signal by Eshita & Bishop (1984), is highly conserved between SSH and LAC viruses.

When the total nucleotide sequence of the M RNA of LAC virus is compared to that of SSH virus it is found to contain 909 base substitutions, 20 insertions and 21 deletions. The overall sequence homology is 79%. Six of the substitutions, but none of the insertions or deletions, occur in the 61 non-coding nucleotides at the 5' end of the viral complementary RNA. On the other hand, the 3' non-coding region contains 33 substitutions, two deletions and one insertion.

Gene products of the M RNA

Consistent with other bunyaviruses for which information is available (Eshita & Bishop, 1984; Collett et al., 1985; Ihara et al., 1985; Lees et al., 1986; Schmaljohn et al., 1987), the LAC virus M RNA contains a single, long open reading frame in the viral complementary strand. The situation in LAC virus parallels that in SSH virus (Eshita & Bishop, 1984), i.e. translation begins with an AUG codon at residues 62 to 64, ends with an ochre termination codon at residues 4385 to 4387, and the encoded polyprotein consists of 1441 amino acids. The amino acid composition of the LAC virus protein is given in Table 1. It has a mol. wt. of 162,598 and a net charge of +31.5, the latter based on the assumption that at neutral pH Arg and Lys are +1, His is +0.5 and Asp and Glu are -1.

When the deduced amino acid sequences of the M gene products of the LAC and SSH viruses were compared, only 156 amino acid substitutions were found, indicating that they are 89% homologous (Fig. 3). The M gene proteins of both viruses are cysteine-rich, a feature that appears to be common among the bunyaviruses (Eshita & Bishop, 1984; Collett et al., 1985; Ihara et al., 1985; Lees et al., 1986; Schmaljohn et al., 1987). In the present case, all 69 cysteine residues in the LAC virus protein also occur in SSH virus. Conversely, 69 of the 71 cysteine residues reported for SSH virus (Eshita & Bishop, 1984) are likewise found in LAC virus. One of the two differences between these viruses involves a Cys to Arg change at position 3 of LAC virus and falls within a region that has the characteristics of a signal sequence. The second
"M RNA sequence of La Crosse virus

MET Arg MET Leu Val Leu Ile Val Val

Thr Ala Ala Ser Pro Val Tyr Gln Arg Cys Phe Gin Asp Gly Ala Ile Val Lys Gin

GGT GCA ACC CCA GTG TAT CAT AGA TCT TCG ATC GTA GCT ATT GTA GGT

Thr Glu Val Cys Leu Asp Asp Val Ser MET Ile Lys Thr Glu Arg Arg Tyr Val Lys

Asn Ala Thr Gly Phe

ACA GAA TCT TCA AAA CAT GAT ATC AAA ACA GAG GGC AGG ATT GTA AAA

Ser Asn Asn Val Ala Ile Arg Lys Trp Leu Arg Ser Asp Trp His Asp Cys Pro Lys Ile Val Gly Gly His

TCA AAC ATG CAA GAA TTT TTA GTA CTT TAT CAT GAT TGG TCA AGA AAC TCA ACC

Ile Asn Val Ile Glu Val Gly Asp Leu Ser Leu His Thr Glu Ser Tyr Val Cys Ser Asp Cys Thr Ile Gly

ATC AAT GTA AAG GTC GTG CAT TGG CTT CAT ACT GAA TCA TAT TGC ACT ACC ATA GGT

Val Asp Lys Glu Thr Ala Gin Val Arg Leu Gin Thr Asp His Phe Glu Ile Ala Gly Thr Thr Val Lys

GTA GAC AAG CAA CCA GCA GTC AGA CTT CAA ACA ACG ACC ACT CAT GCT GGT AGG

Ser Gly Trp Phe Ser Thr Thr Tyr Ile Thr Leu Asp Gin Thr Cys Glu His Leu Lys Val Ser Cys Gly Pro Lys

TCA GGA TCG TCC ARA ACC ACA ACA ATC ACT CTA CAT GAC ACC TTA CAT CCA AAA

Ser Val Gin His Cys Phe Asn Gin His MET Ser Cys Val Arg Phe Leu His Arg Thr Ile Pro Gly His

TCT GTA CAC TCC GTC TGG GCC TCA AAC TCC AAA GAA ACC ATC TGT GCG

Ile Ala Asn Ser Ile Cys Gin Asn Ile Glu Ile Ile Leu Val Thr Leu Thr Leu Leu Ile Phe Leu Leu Ser

ATA GCT ATT GCC TCA AAA CAT GAT ATC AAA ACA GAT TTC GGT TAC GAA

Ile Leu Ser Thr Tyr Ile Cys Tyr Leu Leu MET Pro Ile Phe Ile Pro Ile Ala Tyr MET Tyr Gly Ile Tyr

ATT TAA GAT ACT TAT TAT TAT TAT TGG CAT CAA GCG CCT ACC GCG CAT TTC GTC

Asn Lys Ser Asp Arg MET Cys Lys Lys Cys Leu Cys Gly Leu Val Tyr His Pro Phe Thr Glu Cys Gly Thr His Cys Val Cys

GC CCG TAT GAT ACT TCT CAT GAT AGA CAT CAA CAT GC TCT CAT GCT GGT CAA

Gly Ala Arg Tyr Asp Thr Ser Asp Arg MET Lys His Arg Ala Ser Gly Leu Cys Pro Gly Lys Ser Leu Leu Arg

GGT GCT CCG TAT GAT ACT CAT GAA TCA AAG CAA GCA AGG ATT GGT TGA

Ala Asn Val Arg MET Ser Lys Leu Ser Gly Pro Ala Ser Ile Leu Ser Ile Thr Ala Val Leu Val Leu Thr Phe

GCT GCC AGA TCT TCC AAA CAA CAT GAC AGG AGA ATT TAC TTA GAA

Val Thr Pro Ile Asn Ser MET Val Leu Gly Glu Ser Lys Glu Thr Phe Glu Leu Glu Leu Pro Asp MET Leu

GTC CCA AAC ATC AAC GAA TCA AAA CAT GAT ATC AAA ACA GAT TTC GGA

Glu Met Ala Leu Arg Ile Asn Ser Tyr Tyr Phe Thr Cys Ile Leu Asn Tyr Ala Val Ser Trp Gly Leu Ile Ile Ala

GAA ATG TTA ACG ATC TAT TAT TAT TAT TAT TAT TAT ACG GAT TCT

Gly Leu Leu Val Gly Phe Leu Phe Leu Cys Lys Gin His Arg Pho Phe Leu Asn Ile Tyr Ala MET Tyr Cys Gly Glu Cys

GGA TTA TTA CAG CTA ATT AAA AAA AAC CAT GAC ATT TCT AGG CAT ATG TCT

Asn Met Tyr His Asp Lys Ser Gly Leu Lys Arg His Gly Phe Thr Asn Lys Cys Arg Gin Cys Thr Cys Gly Gin

AAC ATG TAT CAT GCC TGC CAG CTT GAA AGA CAC GGC GAC TCT AGC AAG CAA

Tyr Glu Asp Ala Thr Cys Leu Thr His Arg Lys Thr Tyr Asn Cys Leu Val Gin Lyr Lys Ala Lys Leu Trp MET MET

TAT GAA CAT GCT GCC TGG ATT GCA TCT CAT GAG AGA AAA ACC AAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT

Asn Phe Leu Ile Leu Ile Leu Ile Leu Asp Ser Ala Ile Val Gly Gin Ala Thr Gly Thr Asp Phe

ATT TTC CTA ATC ATT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT TAT
The nucleotide sequence of the M RNA of New York LAC virus isolate 74-32813 given in the viral complementary sense. The deduced amino acid sequence of the protein encoded by the long open reading frame is given below the line, each amino acid over the corresponding codon. Boxes show the locations of potential asparagine-linked glycosylation sites.

Table 1. Deduced amino acid composition of the M gene-encoded polyprotein of LAC virus isolate 74-32813

<table>
<thead>
<tr>
<th>Amino acid</th>
<th>Number of residues</th>
<th>Amino acid</th>
<th>Number of residues</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phe (F)</td>
<td>53</td>
<td>His (H)</td>
<td>35</td>
</tr>
<tr>
<td>Leu (L)</td>
<td>117</td>
<td>Gln (Q)</td>
<td>43</td>
</tr>
<tr>
<td>Ile (I)</td>
<td>121</td>
<td>Asn (N)</td>
<td>67</td>
</tr>
<tr>
<td>Met (M)</td>
<td>27</td>
<td>Lys (K)</td>
<td>116</td>
</tr>
<tr>
<td>Val (V)</td>
<td>86</td>
<td>Asp (D)</td>
<td>71</td>
</tr>
<tr>
<td>Ser (S)</td>
<td>105</td>
<td>Glu (E)</td>
<td>81</td>
</tr>
<tr>
<td>Pro (P)</td>
<td>49</td>
<td>Cys (C)</td>
<td>69</td>
</tr>
<tr>
<td>Thr (T)</td>
<td>97</td>
<td>Trp (W)</td>
<td>15</td>
</tr>
<tr>
<td>Ala (A)</td>
<td>88</td>
<td>Arg (R)</td>
<td>50</td>
</tr>
<tr>
<td>Tyr (Y)</td>
<td>75</td>
<td>Gly (G)</td>
<td>76</td>
</tr>
</tbody>
</table>

The LAC virus polyprotein contains five potential asparagine-linked glycosylation sites (N-X-S/T) occurring at amino acid residues 30 to 32, 57 to 59, 245 to 247, 490 to 492, and 1177 to 1179. These are identical both in location and in the amino acid occupying the X position to those in SSH virus (Eshita & Bishop, 1984). As would be expected for two such closely related viruses, the hydropathy profiles (data not shown) of the LAC virus and SSH virus proteins are so similar as to be almost indistinguishable. Thus, as in the case of SSH virus, the LAC virus protein has an N-terminal region of

difference results from a substitution of Tyr for Cys at residue 1017 in LAC virus. In addition to Cys, most of the Pro residues also remain unaltered in the two viruses. Of the 49 and 48 Pro residues found in the LAC virus and SSH virus proteins, respectively, 46 are conserved. The differences occur at positions 165, 701, 719, 725, 733, and 835.
hydrophobic residues that might function as a leader sequence for membrane translocation. In this regard, it is interesting to call attention to the work of Perlman & Halvorson (1983) who, in an examination of the presecretory signal peptides of 39 proteins from a variety of eukaryotic and prokaryotic sources, reported that application of the Chou & Fasman (1978) rules for predicting protein secondary structure led to prediction of a β-sheet for the core region preceding the cleavage site and often to a β-turn either just before or just after this site. As noted by Lees et al. (1986), the most likely cleavage site in the LAC virus and SSH virus proteins occurs between Ala and Ser at positions 13 and 14. It is therefore noteworthy that when the Chou & Fasman rules are applied to the amino-terminal end of the LAC virus and SSH virus proteins, the residues prior to 13 and 14 are predicted to form a β-sheet, and those immediately following (15 to 18) a β-turn. The conservation of the predicted structure in the two viruses, even though they
differ in five of the first nine amino acid residues, would appear to strengthen the idea that this region represents a signal sequence. It also raises the possibility that the sequence may be inserted into the membrane in the form of a hairpin.

Other regions of hydrophobicity conserved between the LAC virus and SSH virus polyproteins include that encompassed by residues 187 to 246 (interrupted by Glu at 202 and Lys at 222), two tracts at 306 to 376 that are separated by a sequence rich in Glu and Asp, and an area between residues 1390 and 1421 near the carboxy terminus. The latter region is followed by a sequence relatively abundant in charged amino acids and, when taken together, these tracts have been postulated to have a role in anchoring the protein within a membrane (Eshita & Bishop, 1984).

The M RNAs of LAC and SSH viruses are known to encode the G1 and G2 envelope glycoproteins, as well as a non-structural protein (Gentsch & Bishop, 1979; Fuller & Bishop, 1982). Although the size of the deduced polyprotein is sufficient to accommodate all three, at the present time their actual relationships within this putative precursor protein are unknown.

**Predicted linear antigenic sites in the LAC virus and SSH virus proteins**

Presented with the primary amino acid sequence for the proteins from two viruses as closely related as LAC virus and SSH virus, one of the first questions to arise concerns which of the amino acid substitutions plays a role in their antigenic differences. In this regard, attempts to use the information present in amino acid sequence data to identify linear antigenic determinants have largely focused on the prediction of those residues likely to be found on the protein surface. They have primarily employed data on hydrophilicity (Hopp & Woods, 1981: Kyte & Doolittle, 1982), accessibility (Janin, 1979), flexibility (Westhof et al., 1984; Karplus & Schultz, 1985) and contact surface (Novotny et al., 1987). Recently, Parker et al. (1986) introduced a method that represents a synthesis of several of these approaches. It makes use of a new hydrophilicity scale developed in their laboratory, as well as the results of accessibility and flexibility calculations, to produce a composite surface profile. Although it exhibited a slight tendency toward over-prediction, when this method was applied to a group of four proteins, among which was the haemagglutinin of influenza virus, a strong correlation was found between the predicted surface residues and known antigenic determinants (Parker et al., 1986). Consequently, as it appears least likely to overlook actual surface sites, this procedure was selected for use in the present study.

Because of the extent of homology between the polyproteins of LAC and SSH viruses it was expected that many, if not all, of their predicted surface sites would coincide. Therefore, since there is evidence that at least some antigenic sites are congruent with surface $\beta$-turns (Rose et al., 1985; Novotny et al., 1987), one way of dealing with this problem would be to determine which pairs of surface sites differ with respect to their predicted secondary structure, particularly as regards $\beta$-turns. Although it is recognized that there are limits to the accuracy of present predictive methods (Kabsch & Sander, 1983; however, see also Fasman, 1985), it is not unreasonable to expect that they might be extremely useful for the comparison of two highly homologous proteins. In this respect, several laboratories have observed that application of the Chou & Fasman rules to variant viruses selected by monoclonal antibodies resulted in predicted changes in secondary structure that involved the loss of a $\beta$-turn (Wunner et al., 1985; Pellett et al., 1985). Accordingly, residues identified as being on the surface in both the LAC virus and SSH virus proteins were further explored for potential differences by using the Chou & Fasman rules to determine which pairs differed in their predicted secondary structure.

Turning to a comparison of the M gene proteins of LAC and SSH viruses, it can be seen that five of the observed amino acid differences occur in the N-terminal region that corresponds to a signal sequence. As discussed previously, they had no effect on the predicted secondary structure. These residues have been omitted from subsequent analyses.

With respect to the remainder of the proteins, there were 109 sites predicted as lying on the surface. Among them, they contain 118 (76%) of the 156 deduced amino acid differences between the two viruses. These sites could be further subdivided into two groups. The first group
Table 2. Surface residue and/or secondary structure predictions which differ between the M gene-encoded polyproteins of LAC and SSH viruses

<table>
<thead>
<tr>
<th>Residues predicted on surface</th>
<th>Present in</th>
<th>Position of amino acid substitution</th>
<th>Predicted change in secondary structure</th>
<th>Change involves β-turn</th>
<th>Residues predicted on surface</th>
<th>Present in</th>
<th>Position of amino acid substitution</th>
<th>Predicted change in secondary structure</th>
<th>Change involves β-turn</th>
</tr>
</thead>
<tbody>
<tr>
<td>76–88</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>774–786</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>355–357</td>
<td>+</td>
<td>–</td>
<td>+</td>
<td>+</td>
<td>826–843</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>–</td>
</tr>
<tr>
<td>383–393</td>
<td>−</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>846–856</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>404–431</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>927–942</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>447–448</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>965–976</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>–</td>
</tr>
<tr>
<td>473–478</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>–</td>
<td>997–1006</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>482–487</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>1014–1020</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>518</td>
<td>−</td>
<td>+</td>
<td>+</td>
<td>–</td>
<td>1024–1041</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>534–538</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>1049–1051</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>557</td>
<td>−</td>
<td>+</td>
<td>+</td>
<td>−</td>
<td>1089</td>
<td>−</td>
<td>+</td>
<td>+</td>
<td>−</td>
</tr>
<tr>
<td>600</td>
<td>−</td>
<td>−</td>
<td>+</td>
<td>+</td>
<td>1093–1104</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>−</td>
</tr>
<tr>
<td>607–620</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>1123–1124</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>−</td>
</tr>
<tr>
<td>622–638</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>−</td>
<td>1159–1161</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>−</td>
</tr>
<tr>
<td>699–701</td>
<td>−</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>1174–1187</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>−</td>
</tr>
<tr>
<td>704–708</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>1251–1271</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>710–711</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>1293–1298</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>716–736</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>1321</td>
<td>−</td>
<td>+</td>
<td>+</td>
<td>−</td>
</tr>
<tr>
<td>750–753</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>−</td>
<td>1348–1349</td>
<td>−</td>
<td>+</td>
<td>+</td>
<td>−</td>
</tr>
<tr>
<td>760–765</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>−</td>
<td>1424–1438</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
</tbody>
</table>
Table 3. **Surface residue and secondary structure predictions which are similar between the M gene-encoded polyproteins of LAC and SSH viruses**

<table>
<thead>
<tr>
<th>Predicted surface residues</th>
<th>Position of amino acid substitutions</th>
<th>Predicted surface residues</th>
<th>Position of amino acid substitutions</th>
<th>Predicted surface residues</th>
<th>Position of amino acid substitutions</th>
<th>Predicted surface residues</th>
<th>Position of amino acid substitutions</th>
</tr>
</thead>
<tbody>
<tr>
<td>20–22</td>
<td>–</td>
<td>291–296</td>
<td>–</td>
<td>805</td>
<td>808</td>
<td>1143</td>
<td>–</td>
</tr>
<tr>
<td>72</td>
<td>–</td>
<td>398–402</td>
<td>401</td>
<td>901, 902</td>
<td>899</td>
<td>1235</td>
<td>1233</td>
</tr>
<tr>
<td>163–169</td>
<td>168</td>
<td>500–510</td>
<td>511</td>
<td>925</td>
<td>926</td>
<td>1285</td>
<td>–</td>
</tr>
<tr>
<td>189–193</td>
<td>–</td>
<td>570–583</td>
<td>570</td>
<td>956</td>
<td>959</td>
<td>1289–1291</td>
<td>–</td>
</tr>
<tr>
<td>221–222</td>
<td>219</td>
<td>652–654</td>
<td>–</td>
<td>983</td>
<td>985</td>
<td>1337–1339</td>
<td>1337, 1338</td>
</tr>
<tr>
<td>286–288</td>
<td>–</td>
<td>791–796</td>
<td>792, 793, 795, 797</td>
<td>–</td>
<td>–</td>
<td>1392</td>
<td>–</td>
</tr>
</tbody>
</table>
consists of 39 sites, accounting for 78 amino acid substitutions, which are predicted to differ between LAC and SSH viruses in terms either of surface location and/or secondary structure (Table 2). It should be noted that in approximately 60% of these sites the predicted alteration in secondary structure involved β-turns. Furthermore, this group included all of the predicted differences in β-turns between LAC and SSH viruses except for one located at residues 562 to 565 in SSH virus and at residues 563 to 566 in LAC virus. The second group comprises 70 sites, which encompass 40 amino acid changes, for which no differences were predicted between the viral proteins (Table 3).

Although the observation is difficult to interpret at present, the two types of site are not distributed randomly among the polyproteins. This is particularly true for sites predicted as being similar in both viruses, which predominate between amino acid residues 22 to 350 (22/1), 790 to 983 (14/4) and 1130 to 1441 (18/8). The number of residues involved represent about 58% of the protein and account for 77% of the predicted shared sites. It is interesting also to note that the largest regions, 22 to 350 and 1139 to 1441, each encompass about 310 amino acids and are found at the amino and carboxy termini of the proteins. On the other hand, sites in which LAC and SSH viruses are predicted to differ are most common between residues 700 to 790 (1/7) and 997 to 1124 (3/7). This, in turn, is equivalent to 15% of the protein and 36% of the corresponding sites. Finally, there are 12 sites of each type that are more or less evenly distributed over residues 350 to 700.

**DISCUSSION**

The sequence of the LAC virus M RNA is 79% homologous to that of SSH virus and conforms to the general pattern that seems to be emerging as characteristic of the bunyaviruses (Eshita & Bishop, 1984; Collett et al., 1985; Ihara et al., 1985; Lees et al., 1986; Schmaljohn et al., 1987). The primary common features are complementary 3' and 5' ends, the existence of a longer non-coding region at the 3' end of the viral complementary RNA than at the 5' end, and the possession of a single long open reading frame in the viral complementary strand.

The polyprotein encoded by the viral complementary RNA likewise shows many similarities to those of other bunyaviruses. In this case, the major characteristics conserved include a size sufficient to account for the two envelope glycoproteins and a non-structural protein, a relatively high proportion of cysteine residues, and the presence of what are believed to be an N-terminal signal sequence and a C-terminal anchor sequence. There is 89% homology between the amino acid sequences of the LAC virus and SSH virus polyproteins and, not surprisingly, they show several additional similarities such as size, amino acid composition, and the number and location of potential asparagine-linked glycosylation sites.

Several laboratories have shown that many important aspects of the biology of LAC and SSH viruses are associated with the products of the M RNA (see Introduction). At present, it is known that some sites on the G1 glycoproteins of these viruses differ with respect to neutralization and haemagglutination (Gonzalez-Scarano et al., 1982; Grady et al., 1983; Kingsford & Ishizawa, 1984). There is also evidence that deviations in their M gene products are largely responsible for the disparity which exists in their efficiency of transmission by at least one species of mosquito (Beaty et al., 1981a, b). Inasmuch as there are only 156 amino acid substitutions between the proteins of LAC and SSH viruses, in one way or another they must underlie all of these differences, as well as any others which become evident in the future. In this context, the problem becomes one of determining which of the amino acid changes are important and which are not.

An initial attempt to determine which amino acid substitutions are significant with respect to differences between LAC and SSH viruses was made in the present study by applying methods for predicting surface residues and secondary structure to a comparison of their polyproteins. The outcome of such analyses was to reduce further the possibilities to 39 sites, containing 78 of the amino acid substitutions, that have the highest probability of being responsible for the biological differences associated with the M gene proteins, at least to the extent that they are defined by linear antigenic determinants. Although their ultimate validity must wait upon
experimental verification, in the absence of any other structural information such predictions are nevertheless valuable because they serve to call attention to regions of the molecule that might be best suited for further study.

We are indebted to Dr R. C. Herman, currently with Syntex Research Laboratories, for his assistance in the early stages of this study and for many helpful discussions throughout. We also wish to thank Dr R. Deibel of the Wadsworth Center for his encouragement and support during the course of the project.

REFERENCES


M RNA sequence of La Crosse virus


(Received 8 July 1987)