Sequence Determination of the Sendai Virus Fusion Protein Gene

By BENJAMIN M. BLUMBERG,† COLOMBA GIORGI, KEITH ROSE AND DANIEL KOLAKOFSKY

Département de Microbiologie, Université de Genève, 64 Avenue de la Roseraie, 1205 Geneva, Switzerland, Laboratorio di Virologia, Istituto Superiore di Sanità, Rome, Italy and Département de Biochimie Medicale C.M.U., 9 Avenue de Champel, 1211 Geneva 4, Switzerland

(Accepted 30 October 1984)

SUMMARY

From a genomic DNA library of Sendai virus, we have identified and sequenced clones corresponding to the F glycoprotein gene. The limits of the F gene region were defined by mapping the 5' and 3' ends of the mRNA with S1 nuclease. The Sendai virus F gene is 1821 nucleotides long. The predicted primary translation product of the single long open reading frame would code for a protein of 565 amino acids, containing a putative signal peptide, three carbohydrate addition sites, a hydrophobic region corresponding to the known cleavage/activation site of F0, and a long, very hydrophobic region near the C-terminus which probably represents the transmembrane region of the protein. The signal peptide cleavage site of the mature protein was determined by mass spectrometry. Interestingly, the amino acid sequence surrounding the cleavage/activation site of the Sendai virus F protein shows significant homology to the same region of the influenza B and C virus HA proteins, suggesting that these genes may have evolved from a common ancestor. The ability of the Sendai virus F protein to fuse membranes relative to its primary structure is discussed.

INTRODUCTION

Sendai virus is the prototype of the parainfluenza subgroup of the family Paramyxoviridae. Parainfluenza virions are composed of a helical nucleocapsid core which contains the genomic RNA of negative polarity plus the viral NP, P and L proteins, and a surrounding envelope which contains the viral M protein on the inside layer of the membrane, and the viral F and HN proteins which are transmembrane proteins and appear as spikes on the outside surface of the envelope (Choppin & Compans, 1975). Besides the six structural proteins listed above, the Sendai virus genome is also known to code for a pair of non-structural proteins (C and C') expressed from an overlapping reading frame of the P gene (Giorgi et al., 1983).

Unlike other enveloped RNA viruses, Sendai virus appears to enter its host cell by fusion through the cytoplasmic membrane rather than through endocytic vesicles (Choppin & Compans, 1975; Fan & Sefton, 1978; Nagai et al., 1983). This ability to fuse membranes at neutral pH has long been recognized for Sendai virus and is due to the action of viral F protein. Interestingly, the F protein is synthesized as an inactive precursor (F0) which is activated by cleavage to generate two disulphide-linked polypeptides (F1 and F2), of which only F1 has an unblocked N-terminus (Homma & Ohuchi, 1973; Scheid & Choppin, 1974, 1977). Since the uncleaved F0 protein also has a blocked N-terminus, this suggested that the order of the F protein was NHX–F2–F1–COOH (Scheid & Choppin, 1977; Gething et al., 1978). In order to examine the properties of this protein more carefully, we have determined its primary structure by nucleotide sequencing of DNA clones that cover the F gene region.

†Present address: Department of Neurosciences, University of Medicine and Dentistry of New Jersey, Newark, New Jersey 07103, U.S.A.
**METHODS**

Selection of large F protein-specific clones. Molecular cloning of Sendai virus using random pentamers from calf thymus DNA as primers for reverse transcriptase on the viral 50S genome resulted in a library of five Grunstein–Hogness filters containing about 1200 clones (Dowling et al., 1983). In order to reduce this library to a single filter containing mostly the largest glycoprotein-specific clones, we first used fragments of viral 50S genome, degraded by treatment with 0.2 M-Na2CO3 at 55 °C for 10 min to about 2S size and then 5' end-labelled with T4 polynucleotide kinase and [γ-32P]ATP, to probe the filters. With short probes, the intensity of the signal should be proportional to the size of the virus-specific insert, and those colonies giving the most intense signals were noted. Next, 32P-labelled 18S Sendai virus mRNA was isolated by velocity sedimentation on sucrose gradients and radio-labelled after alkali treatment as above. Unlabelled Sendai virus RNA from the 33S region of the gradient was added to this material to reduce the signal due to L mRNA sequences. When this mixture was used as a probe, clones containing viral 18S mRNA sequences were highlighted. Finally, the PstI inserts of a series of clones sufficient to cover the NP, P/C and M gene regions were excised, pooled and nick-translated to probe the filters for clones containing sequences from these genes. The 120 colonies that were most strongly positive to the first two probes and negative to the third were plated onto a single Grunstein filter for further examination. This filter was first probed with the nick-translated insert of clone 12/1-10, which was known to lie at the end of the M gene region (Blumberg et al., 1984a, c). One positive clone, 12/1-14, was shown by restriction analysis to have a 1050 bp insert and to contain viral sequences well beyond the end of 12/1-10. A synthetic 17-base oligodeoxynucleotide (5'-TAAGAAAAACTTAGGGT) complementary to Sendai virus intercistronic sequences (Giorgi et al., 1983; Gupta & Kingsbury, 1984) was next 5' end-labelled and used to probe a dot blot made from minilysates of the first 80 of the above 120 clones. Two colonies (12/12-5 and 15/12-12) positive to this oligonucleotide were also positive to a probe made from the small PstI fragment that lies at the right side of 12/1-14, suggesting that these clones contained the sequence from the end of 12/1-14 to the next intercistronic junction. Restriction analysis showed that each of these clones contained an approximately 1 kb insert, sufficient to span this region, but failed to confirm the overlap with 12/1-14, which was finally proven by sequencing. The methods used above are fully described in our cited publications (Blumberg et al., 1984a, b, c; Dowling et al., 1983; Giorgi et al., 1983), and in Maniatis et al. (1982). The oligodeoxynucleotide was the kind gift of A. Chollet, Biogen SA, Geneva, Switzerland.

SI mapping. Mapping with SI nuclease was done by annealing end-labelled dsDNA probes with CsCl pellet RNA from Sendai virus-infected or mock-infected BHK cells at 53 °C in buffer containing 80% formamide (Giorgi et al., 1983; Blumberg et al., 1984b, c).

DNA sequencing. Chemical sequencing reactions were performed according to Maxam & Gilbert (1980). Wherever possible, sequencing was done from restriction sites with 5' overhanging ends, so that the fragments could be labelled simultaneously, in separate reactions, at the 3' end (by fill-in with Klenow fragment of DNA polymerase from Boehringer) and at the 5' end (with T4 polynucleotide kinase from P-L Biochemicals after treatment with alkaline phosphatase from Worthington). These complementary sequences were run side-by-side on 14% and 8% polyacrylamide gels containing 7 M-urea, so that any ambiguities in reading one sequence could be immediately cross-checked by reference to the other. Some sequences were obtained by 3' labelling at PstI sites (with terminal deoxynucleotide transferase from P-L Biochemicals and deoxy ATP). Isotopes were purchased from Amersham.

Preparation and analysis of the Sendai virus F2 protein. Ten ml of purified egg-grown Sendai virus in water (approximately 4 mg/ml) was solubilized with 1% Triton X-100 for 15 min at 30 °C and the insoluble material was removed by centrifugation (90 min, 39000 r.p.m., SW41 rotor). The supernatant containing the HN and F proteins in 1% Triton X-100 was adjusted to 20 mM-phosphate, 0.15 M-NaCl, pH 7.2 and applied batchwise 2 ml at a time to a MonoQ ion-exchange column essentially as described previously (Welling et al., 1983). Under these conditions, protein F emerges as a triple peak and the whole region was pooled. SDS–PAGE showed the protein to be homogeneous. The protein was reduced with dithiothreitol and applied directly to a reversed phase C18 HPLC column essentially as described by Van der Zee et al., (1983). Under these conditions, fragment F2 is obtained before the strongly absorbing Triton peak. A portion of the F2 fragment was subjected to N-terminal peptide analysis using techniques already described (Rose et al., 1983, 1984). The enzymes employed for the digestion were chymotrypsin and proline-specific endopeptidase.

**RESULTS**

Selection, ordering and SI mapping of F protein-specific clones

In previous communications, we described the molecular cloning of the Sendai virus genome and the ordering and sequencing of 12 clones corresponding to the leader, NP, P/C and M gene regions (Dowling et al., 1983; Giorgi et al., 1983; Blumberg et al., 1984a, b, c). In order to select F protein-specific clones from our library, we first triple screened our Grunstein filters with probes
Sendai virus fusion protein gene

Fig. 1. Genomic map of the Sendai virus F protein region. On the top is a partial genomic map showing the distance in kb from the 3' end of the (−) genome. The open bar underneath indicates the single long ORF contained within the gene, and the flanking vertical bars show the limits of the mRNA. The assignment of HN as the following gene is tentative (see Discussion). On the bottom are shown the clones used to determine the sequence and a partial restriction map showing those sites used in the sequence determination. Lines underneath each clone which contain complete arrowheads represent sequences obtained on both strands from sites giving 5' overhangs. Lines that contain only half an arrowhead represent sequences on one strand either from a PstI site or a HinfI site used to cross major sequencing sites or to determine the end sequences of the clones. The bars just underneath the clones 12/1-14 and 15/12-12 show the 5' and 3' ends of the F mRNA respectively as determined by S1 mapping (Fig. 2). The letters above clones refer to the restriction enzymes used as follows: B (BamHI); G (BglII); C (ClaI); P (PstI); T (TaqI); D (DdeI); F (HinfI); A (AccI).

for clones containing large inserts, for clones containing sequences corresponding to Sendai virus 18S mRNAs, and for clones containing intercistronic sequences. After eliminating from consideration all clones corresponding to the NP, P/C and M genes, a combination of dot blotting and restriction analysis identified a series of overlapping clones running from the intercistronic junction at the end of M to the next intercistronic region (data not shown; see Methods). Three of these clones are shown in Fig. 1, along with a gene map, a partial restriction map, and strategies for sequencing and S1 nuclease mapping.

The F gene region appeared to be spanned by two long overlapping clones 12/1-14 and 15/12-12. To determine the limits of the F gene, we mapped both ends of the mRNA with S1 nuclease relative to the DNA clones which were obtained from genome RNA. As these clones appeared to have no unique restriction site in common, we adopted the strategy shown in Fig. 1. Plasmid 12/1-14 was cleaved with ClaI, 5' end-labelled and redigested with PstI. The approximately 640 bp ClaI–PstI virus-specific insert fragment was recovered from a non-denaturing polyacrylamide gel, and annealed with Sendai virus-infected or mock-infected CsCl pellet RNA from BHK cells to probe for the 5' end of the mRNA (Fig. 2a). As shown in lanes 5 and 6 a single strong band between the BamHI site and the tail of the insert was protected by the Sendai virus mRNA, placing the 5' end of the mRNA approximately 500 bp to the left of the ClaI site. Inspection of the sequence in this region shows the termination signal (TAAGAAAAA, nucleotides 13 to 21, Fig. 3), the intercistronic trinucleotide (CTT, nucleotides 22 to 24) and the sequence (AGGGNNAAAG, nucleotides 25 to 34) which is found at all the Sendai virus intercistronic boundaries (Giorgi et al., 1983; Gupta & Kingsbury, 1984). Further, this same region has also been determined by S1 mapping to contain the 3' end of the M gene mRNA.
Fig. 2. S1 nuclease mapping of the Sendai virus F gene mRNA. 

(a) Mapping of the 5' end of the F mRNA. Plasmid 12/1-14 was cut with ClaI, 5' end-labelled and recut with PstI (lane 2). A portion of this material was also recut with BamHI to serve as internal markers (lane 3). The approximately 640 bp ClaI-PstI fragment (arrow, lane 2) was isolated from a non-denaturing 6% polyacrylamide gel and used as a probe. Lane 4, the 640 bp ClaI-PstI probe annealed with 1.0 absorbance units of mock-infected BHK cell RNA, incubated under S1 conditions but without S1 nuclease. Lanes 5 and 6, the 640 bp probe annealed respectively with 0.5 or 1.5 absorbance units of Sendai virus-infected BHK cell RNA and digested with S1 nuclease. Lane 7, the 640 bp probe annealed with 1.5 absorbance units of mock-infected BHK cell RNA and digested with S1 nuclease. Lane 1 shows HindIII-digested pBR322 as markers: the lengths of the marker fragments are shown on the left side. 

(b) Mapping of the 3' end of the F mRNA. Plasmid 15/12-12 was cut with BamHI, 3' end-labelled and recut with PstI (lane 1). A portion of the end-labelled material was further cut with AccI to serve as internal markers (lane 2). The approximately 800 bp BamHI-PstI fragment (arrow, lane 1) was isolated as above and used as a probe. Lane 5, the 800 bp BamHI-PstI fragment annealed with 1 absorbance unit of mock-infected BHK CsCl pellet RNA and incubated under S1 conditions but without S1 nuclease. Lanes 6 and 7, the 800 bp probe annealed with 1 absorbance unit of Sendai virus-infected and mock-infected CsCl pellet RNA respectively, and treated with S1 nuclease. Lanes 3 and 4, HindIII-digested pBR322. All digestions were run on a 7 M-urea-8% polyacrylamide sequencing gel.

These results thus demonstrate that the Sendai virus F gene immediately follows the M gene and that the 5' end of the F gene mRNA probably starts with the AGGG listed as nucleotides 25 to 28 in Fig. 3.

The clone 15/12-12, which had been shown by dot blotting to overlap 12/1-14 and also contained an intercistronic region, was next used to probe for the 3' end of the mRNA. Plasmid
Table 1. Codon usage of the Sendai virus F gene

<table>
<thead>
<tr>
<th>Codon</th>
<th>Usage</th>
<th>Codon</th>
<th>Usage</th>
<th>Codon</th>
<th>Usage</th>
<th>Codon</th>
<th>Usage</th>
<th>Codon</th>
<th>Usage</th>
</tr>
</thead>
<tbody>
<tr>
<td>TTT Phe 3 0-5%</td>
<td>TCT Ser 14 2-4%</td>
<td>TAT Tyr 9 1-5%</td>
<td>TGT Cys 7 1-2%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TTC Phe 9 1-5%</td>
<td>TCC Ser 3 0-5%</td>
<td>TAC Tyr 7 1-2%</td>
<td>TGC Cys 4 0-7%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TTA Leu 7 1-2%</td>
<td>TCA Ser 8 1-4%</td>
<td>TAA --- 0 0-0%</td>
<td>TGA --- 0 0-0%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TTG Leu 7 1-2%</td>
<td>TCG Ser 6 1-0%</td>
<td>TAG --- 0 0-0%</td>
<td>TGG Trp 3 0-5%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CTT Leu 11 1-9%</td>
<td>CCT Pro 4 0-7%</td>
<td>CAT His 4 0-7%</td>
<td>CGT Arg 2 0-3%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TTC Leu 9 1-5%</td>
<td>CCC Pro 7 1-2%</td>
<td>CAC His 4 0-7%</td>
<td>CGC Arg 3 0-5%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TTA Leu 5 1-4%</td>
<td>CCA Pro 5 0-8%</td>
<td>CAA Gin 6 1-0%</td>
<td>CGA Arg 2 0-3%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TTG Leu 16 2-8%</td>
<td>CCG Pro 3 0-5%</td>
<td>CAG Gin 16 2-8%</td>
<td>CGG Arg 2 0-3%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ATT Ile 13 2-3%</td>
<td>ACT Thr 8 1-4%</td>
<td>AAT Asn 9 1-5%</td>
<td>AGT Ser 3 0-5%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ATC Ile 22 3-8%</td>
<td>ACC Thr 10 1-7%</td>
<td>AAC Asn 13 2-3%</td>
<td>AGA Arg 12 2-1%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ATA Ile 21 3-7%</td>
<td>ACA Thr 22 3-8%</td>
<td>AAG Lys 11 1-9%</td>
<td>AGG Arg 12 2-1%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ATG Met 9 1-5%</td>
<td>GCT Ala 16 2-8%</td>
<td>GAT Asp 17 3-0%</td>
<td>GGT Gly 13 2-3%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GTT Val 12 2-1%</td>
<td>GCC Ala 7 1-2%</td>
<td>GAC Asp 11 1-9%</td>
<td>GGC Gly 4 0-7%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GTC Val 14 2-4%</td>
<td>GCA Ala 13 2-3%</td>
<td>GAA Glu 11 1-9%</td>
<td>GGA Gly 6 1-0%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GTA Val 8 1-4%</td>
<td>GCG Ala 6 1-0%</td>
<td>GGU Gln 17 3-0%</td>
<td>GGG Gly 17 3-0%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

15/12-12 was cleaved at the BamHI site, 3' end-labelled and redigested with PstI, and the approximately 800 bp PstI-BamHI fragment was isolated and annealed with Sendai virus-infected or mock-infected BHK cell CsCl pellet RNA (Fig. 2b). A single band just below the probe and estimated to be 750 bp in length was protected when infected cell CsCl pellet RNA was annealed to the probe (lane 6) but was absent when mock-infected CsCl pellet RNA was used (lane 7). Inspection of the sequence 750 bp to the right of the BamHI site again shows the canonical intercistronic sequence mentioned above (TAAGAAAAA-CTT-AGGGNNAAAG, nucleotides 1837 to 1858). Since we have previously shown by high resolution S1 mapping that the 3' end of the P/C mRNA ends within the A run of the TAAGAAAAA sequence (Giorgi et al., 1983), these results suggest that the F gene mRNA would end at the A residue at nucleotide 1845. Taken together, these results demonstrate that clones 12/1-14 and 15/12-12 are oriented as shown in Fig. 1, and contain the sequence of a Sendai virus mRNA of approximately 1800 bases in length.

**Nucleotide sequence of the Sendai virus F gene**

Having demonstrated by S1 mapping that clones 12/1-14 and 15/12-12 contained the limits of our gene, these clones plus clone 14/12-1 (which was shown by dot hybridization to cross both clones 12/1-14 and 15/12-12) were sequenced by the Maxam & Gilbert (1980) technique (Methods) according to the strategy outlined in Fig. 1. The sequence of the F gene region is shown in Fig. 3 as mRNA sense or (+) DNA. According to our sequencing results, the Sendai virus F gene mRNA is 1821 nucleotides long [exclusive of its non-templated poly(A) tail] beginning on the decanucleotide AGGGATAAAG (nucleotides 25 to 34) and ending on the TAAGAAAAA sequence (nucleotides 1837 to 1845), and is flanked on both sides by the intercistronic trinucleotide CTT. There is only a single long open reading frame (ORF) beginning with the ATG at position 78, which is not precisely in the most favoured context for initiation on eukaryotic ribosomes since it is followed by an A residue rather than a G (Kozak, 1983) and continues for 1695 nucleotides to position 1773 where it is interrupted by an opal termination codon followed closely by two additional in-frame terminators. The other two reading frames are blocked frequently. The coding region of the F mRNA contains 479 A (28-2%), 416 T (24-5%), 428 G (25-2%) and 372 C (21-9%) residues. Codon usage (Table 1) is not remarkable except for the bias against CG dinucleotides reported in eukaryotic genes (Rose & Gallione, 1981; Swartz et al., 1962); for example, of the codons listed in Fig. 3 that contain C in the first position, only 8-8% contain G in the second position.

The predicted amino acid sequence of the primary translation product of the ORF is shown in Fig. 3 beneath the nucleotide sequence of the gene. The unmodified protein would contain 565 amino acids, with a molecular weight of 61666 and have a charge of +9 at pH 7.
Fig. 3. Nucleotide sequence and predicted primary translation product of the Sendai virus F gene. In the gene sequence, shown as mRNA sense (+) DNA, the viral polymerase termination/polyadenylation signals are in bold type. In the protein sequence, the determined cleavage point of the signal peptide is shown by an arrow underneath the sequence, potential carbohydrate addition sites are underlined, and the hydrophobic N-terminus of F1 and the transmembrane region are in bold type. The 5' end of the F mRNA at position 25 is marked by an arrowed line above the sequence; the 3' end is marked by the termination/polyadenylation sequence in bold type (position 1837 to 1845).
DISCUSSION

For unknown reasons, we have consistently failed to detect the synthesis of the Sendai virus F protein in vitro using rabbit reticulocyte lysates. Thus, standard methods of gene identification such as hybrid selection of mRNA or hybrid arrest of translation have not been possible with this gene. Fortunately, the identification of our DNA clones as being that of the F gene is nonetheless unambiguous. As mentioned in the Introduction, the Sendai virus F₀ protein is activated by cleavage to produce the new N-terminus of F₁ which is unblocked and whose sequence has been determined by Edman degradation. Inspection of the amino acid sequence of our gene product (Fig. 3) shows that the sequence Phe–Phe–Gly ... beginning at position 117 is identical to the first 15 amino acids of the N-terminus of F₁ determined by Gething et al. (1978) except for two positions, and to the first 20 amino acids of F₁ as determined by Richardson et al. (1980) except for one position. These small differences presumably reflect strain variation. In addition, position 116 of our sequence is occupied by Arg, consistent with the fact that F₀ is cleaved by a protease with trypsin-like specificity. The Sendai virus F₀ protein is thus presumably cleaved between amino acids 116 and 117 during activation without further modification of the newly generated N-terminus of F₁.

Scheid et al. (1978) and Gething et al. (1978) have suggested that the F₀ protein is anchored in the membrane by the C-terminus of F₁ because of the finding that the N-termini of both F₀ and F₂ were blocked and by comparison to the analogous influenza virus HA protein. The deduced amino acid sequence of the complete Sendai virus F protein would appear to confirm this. Amino acids 500 to 523 contain the strongest hydrophobic domain of the protein (Fig. 4) and nicely fit the definition of a transmembrane domain according to Kyte & Doolittle (1982): out of
Sendai virus fusion protein gene

24 residues, 20 are either Val, Ile or Leu, the three most hydrophobic amino acids; the hydrophobic region is flanked on either side by regions of high charge density, and there is a Trp residue just upstream at residue 494. The cytoplasmic tail of the F protein would then presumably begin with the highly charged Arg–Leu–Lys–Arg sequence at position 524 to 527 and extend to the C-terminus at position 565. This cytoplasmic tail is highly charged (16 out of the 42 amino acids) and contains an unusually high proportion of methionines (four out of the eight methionines present in F12 are clustered here), suggesting that the thioether side-chains of this amino acid may be important determinants of the F protein on the inside of the membrane.

The Sendai virus F protein should also contain a signal sequence near its N-terminus which initiates the export of the nascent chain across the rough endoplasmic reticulum. According to the signal hypothesis, the signal sequence is then cleaved from the protein at a specific site. The direct determination of this site has not been reported to date (the N-termini of both F0 and F1 are blocked). The cleavage site of the signal peptide can however be inferred. From the amino acid composition of F2, Gething et al. (1978) estimated that mature F2 was 90 amino acids in length, whereas from our sequence, the unprocessed F2 is 116 amino acids long, a difference of 26. This would place the cleavage site near the Gln at position 26 (Gln-26). Mature F2 was also found to contain only one His and one Cys (Gething et al., 1978). Inspection of our sequence shows a single His at position 51 and three Cys residues at positions 10, 25 and 70. Cleavage after the Cys-25 would thus eliminate the two extra cysteines. In addition, cleavage of the Cys-25/Gln-26 bond is the only site compatible with the rules of Von Heijne (1983) for predicting the cleavage site. Recently, Rose et al. (1983, 1984) have developed a micromethod involving gas–liquid chromatography and mass spectrometry (GLC/MS) for the determination of blocked N-termini. We have prepared a sample of the F2 protein from egg-grown virus and subjected it to GLC/MS analysis. The mass spectrum (Fig. 5) is clearly interpretable as the N-terminal blocked tripeptide PCA/leucine or isoleucine/proline [N-terminal glutamine spontaneously cyclizes to form pyrrolidone carboxylic acid (PCA)]. Inspection of our amino acid sequence shows that the tripeptide Gin–(Leu or Ile)-Pro exists only once, at position 26 to 28. These results therefore demonstrate that the Sendai virus F protein contains a signal sequence which is cleaved after the Cys-25. The mature F0 protein excluding the carbohydrate moiety would thus have a molecular weight of 58938 and the F2 and F1 proteins would have molecular weights of 9869 and 49069 respectively.

Scheid & Choppin (1977) have demonstrated that both the F2 and F1 polypeptides of Sendai virus are glycoproteins and by comparing the ratio of 14C-amino acids and 3Hglucosamine incorporated, noted that the F2 moiety contains approximately three times as much carbohydrate relative to protein as does F1. Inspection of our amino acid sequence shows three potential glycosylation sites (Asn–X–Thr/Ser) at position 104 to 106 in the F2 moiety, and positions 245 to 247 and 449 to 451 in the F1 moiety. Assuming that equal amounts of carbohydrate are added at each site, the sequence shown in Fig. 3 would predict that F2 would contain 2.5 times as much carbohydrate to protein as F1. The data of Scheid & Choppin (1977) would thus argue that all the potential glycosylation sites are filled.

Perhaps the most interesting property of the Sendai virus F protein, which has long been recognized, is its ability to fuse lipid membranes at neutral pH, a property required for viral penetration at the host cell cytoplasmic membrane. As mentioned in the Introduction, this activity requires cleavage of the F0 protein to generate a new hydrophobic N-terminus on F1. Concomitant with this cleavage/activation, Hsu et al. (1981) have shown conformational changes in the F protein both by an increase in α-helicity as measured by circular dichroism, and the appearance of extra hydrophobic domains as measured by the ability of the cleaved protein to bind increased amounts of non-ionic detergents. Inspection of the amino acid sequence and a computer-generated hydropathic plot of the polypeptide chain (Fig. 4) reveals two interesting points in this regard. Firstly, although the three domains that are thought to interact with lipid membranes (i.e. the signal sequence at the N-terminus of the primary translation product, the N-terminus of F1 and the transmembrane domain near the C-terminus of F1) are all clearly recognizable on the hydropathic plot, many other regions of F1 also appear to be hydrophobic. Thus, other regions of F1 besides its N-terminal domain may be involved in interacting with the
Sendai virus fusion protein gene

Table 2. Extent of homologies between the Sendai virus F protein and the influenza virus HA proteins at the cleavage-activation site*

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>24-3%</td>
<td>16-8</td>
<td>6-20</td>
<td>1-94</td>
</tr>
<tr>
<td>B</td>
<td>15-1%</td>
<td>18-4%</td>
<td>15-2%</td>
<td>14-5%</td>
</tr>
<tr>
<td>C</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* The percentage homologies were determined from the alignment depicted in Fig. 6 and are shown below the diagonal. These numbers do not take into account the gaps introduced to align the sequences. The numbers above the diagonal show the standard deviations above the comparison of the average of 100 randomized sequences by the method of Needleman & Wunsch (1970). Numbers above 3 are considered significant.

lipid bilayer during fusion. Secondly, inspection of the cysteine residues in the mature F12 protein shows a single Cys-70 in F2, and eight cysteines in F1 of which seven are clustered between positions 338 and 424. Since only one of these cysteines in F1 is involved in a disulphide linkage to the Cys-70 of F2, it is reasonable to assume that some of the other cysteines are involved in intrachain disulphide bonds within F1. This would lead to a ‘bunching’ of the polypeptide chain between positions 338 and 424. The conformational changes in F induced by cleavage to the F12 complex noted by Hsu et al. (1981) might then be expected to turn about this bunched region which would act as a hinge.

We have previously suggested the possible evolution of Sendai virus from an influenza virus-like ancestor, by concatenation of the separate gene segments into a single RNA strand (Giorgi et al., 1983). This suggestion was based on (i) the unexpected finding of functional overlapping reading frames within a Sendai virus gene, as also occurs in influenza virus (Lamb & Lai, 1980), (ii) the conservation of non-coding sequences at the start of the NP, P/C and M genes beyond the initial decanucleotide thought to represent the signal for the initiation of mRNA synthesis (Giorgi et al., 1983; Gupta & Kingsbury, 1984) and (iii) the many similarities in the envelope proteins of Sendai virus and influenza virus such as the distribution of three different activities (haemagglutination, neuraminidase and fusion) among two transmembrane proteins, and the amino acid homology between the new N-termini generated by cleavage activation of the Sendai virus F0 and influenza virus HA proteins. Availability of the entire Sendai virus F gene sequence and that of the recently reported influenza C virus glycoprotein gene (Nakada et al., 1984; Pfeifer & Compans, 1984) now allows further comparisons to be made. Fig. 6 shows a partial comparison of the Sendai virus F protein and that of the influenza A, B, and C HA proteins by aligning these proteins at the cleavage-activation sites. To maximize homology between the Sendai virus F protein and the influenza HA proteins, only a single additional gap was introduced beyond those already placed by Pfeifer & Compans (1984) in aligning the three influenza proteins. The maximum homology between the Sendai virus F and the influenza virus HA proteins is at the cleavage-activation site, as it is between the influenza HA proteins themselves, but homologies can be detected over the entire sequence shown of approximately 150 amino acids. The extent of the homologies between these regions of the four proteins is listed in Table 2, which shows the percentage homologies below the diagonal and the significance of the homologies as the number of standard deviations over the average of 100 randomized sequences by the method of Needleman & Wunsch (1970) above the diagonal. By these criteria,
Fig. 6. Amino acid homologies of the Sendai virus F protein and the influenza virus HA proteins at the cleavage-activation site. Lines A, B and C, which represent partial sequences of the A, B and C HA proteins respectively, were aligned by Pfeifer & Compans (1984) to maximize homology between the influenza virus proteins over their entire length. Only that portion which also shares homology to the Sendai virus F protein (line F) is shown here. The numbers above the sequences are arbitrary. In all cases, cleavage activation takes place after the arginine at position 19. Homologies between the Sendai virus F protein and the influenza virus HA proteins are indicated in bold type. Only a single additional gap (the valine at position 91) was introduced in aligning the Sendai virus F and the influenza virus HA proteins. The sequence of the hypothetical ancestor gene (line ANC) shows those amino acids shared by at least two of the four proteins. In those cases where two amino acids were shared, both amino acids are listed. Homologies between the Sendai virus F proteins and the hypothetical ancestor gene are also indicated in bold type. The dashed lines below the ANC sequence indicate the five regions of high (>48%) homology.
Fig. 7. Location of the regions of high homology on the three-dimensional HA structure. The schematic model of the structure of the HA1-HA2 monomer determined by Wilson et al. (1981) (published in Nature, London 289, 366-373) for the influenza A protein is shown. The five regions of high homology indicated in Fig. 6 which represent residues 312 to 329 of HA1 and 1 to 13 of HA2 (first), 35 to 43 (second), 53 to 80 (third), 103 to 109 (fourth) and 113 to 116 (fifth) of HA2 of the influenza A virus protein are indicated both by numbering the residues and filling in the regions.
the cleavage-activation region of the Sendai virus F protein shows as much significant homology to the influenza C glycoprotein as do the influenza A and B virus HA proteins, and also shows significant homology to the influenza B virus HA protein. Only the homology between the Sendai virus F protein and the influenza A virus HA protein is not significant by these criteria on a pair-wise basis, but taken together, the above evidence suggests that all the influenza virus HA proteins and the Sendai virus F protein are related in evolutionary terms. Interestingly, Rose et al. (1982) have also detected limited but significant homology between the M proteins of influenza virus and vesicular stomatitis virus (VSV), another non-segmented, negative-sense RNA virus which shares many features in common with Sendai virus.

A partial sequence of the hypothetical ancestral gene is also shown in Fig. 6, where the homologies between the Sendai virus F proteins and the hypothetical ancestor protein are indicated in bold type. These homologies are not distributed evenly but are found in five groups of high homology (>48%) which range in length from four to 31 residues. Since the crystallographic structure of the influenza A HA protein has been determined, it is of interest to examine where these regions of high homology to the hypothetical ancestor gene are found on the three-dimensional structure. The schematic model of the influenza A HA₁-HA₂ monomer of Wilson et al. (1981) is shown in Fig. 7 with the high homology regions filled in. Note that the first (which includes the C-terminus of HA₁ and the N-terminus of HA₂), second, fourth and fifth regions are all found on the bottom half of the stalk of the protein which protrudes from the membrane and are in relatively close proximity, but the third region, which corresponds to residues 53 to 80 of HA₂, is found at the top half of the stalk and is quite separate from the others. It thus appears that regions which represent almost the entire length of the HA stalk may have been conserved throughout the evolution of the HA and F proteins. Further, the finding of conserved regions which are not in close proximity would also appear to argue against the possibility of functional convergence (Doolittle, 1981) rather than divergence from a common ancestor.

Finally, inspection of the F gene sequence (Fig. 3) also shows that after the CTT intercistronic trinucleotide which follows the F gene is the polypurine-rich sequence AGGGTGAAAG beginning at nucleotide 1849 which is similar to the decanucleotide found at the beginning of all the Sendai virus genes described to date (Giorgi et al., 1983; Gupta & Kingsbury, 1984). We have tentatively assigned the gene following the F gene as HN based on two assumptions, namely (i) that the Sendai virus genome contains only the six genetic domains that correspond to the six virion structural proteins, and (ii) that analogous to the VSV genome, the L gene is last. However, further work is required to demonstrate this point directly and to rule out the possibility of additional genes within the Sendai virus genome.

We thank Dr Marco G. Simona, Rosette Bandelier and Colette Pasquier for expert technical assistance, Rita Lenertz for help in typing the manuscript, François Veuthey for printing out Fig. 3 and Amos Bairoch for the hydropathy profile (Fig. 4). This work was supported by a grant from the Swiss National Fund for Scientific Research, and the Acute Respiratory Diseases of Children Program of the World Health Organization.

REFERENCES


Sendai virus fusion protein gene


(Received 2 August 1984)