Proc. Natl. Acad. Sci. USA
Vol. 81, pp. 7683–7687, December 1984
Biochemistry

Nucleotide sequence of the gene encoding the fusion (F) glycoprotein of human respiratory syncytial virus

(glycoprotein structure/mRNA consensus sequences/comparisons among paramyxoviruses)

PETER L. COLLINS*, YUNG T. HUANG†, AND GAIL W. WERTZ

Department of Microbiology and Immunology, School of Medicine, University of North Carolina at Chapel Hill, Chapel Hill, NC 27514

Communicated by Purnell W. Choppin, August 9, 1984

ABSTRACT **The nucleotide sequence of the mRNA encoding the F protein of respiratory syncytial (RS) virus (strain A2) was determined from cDNA clones that contain the complete mRNA sequence. The mRNA is 1899 nucleotides long exclusive of polyadenylylate. The single major open reading frame encodes a protein of 574 amino acids, with a calculated molecular weight of 63,453. Major structural features predicted from the amino acid sequence include an NH_2-terminal signal sequence (residues 1–22), hydrophobic transmembrane anchor sequence (residues 525–550), five potential acceptor sites for asparagine-linked carbohydrate, and a potential site (residues 131–136) for the proteolytic cleavage that generates the disulfide-linked F_1 and F_2 subunits, which, by analogy to other paramyxoviruses, constitute the biologically active form of the F protein. The sequence also contains an internal hydrophobic domain (residues 137–154) that, as a consequence of the activating proteolytic cleavage described above, would become the NH_2 terminus of the larger, F_1 subunit. The amino acid sequence of the hydrophobic terminus of the F_1 subunit is known to be highly conserved among several paramyxoviruses but is markedly dissimilar for RS virus. The F_2 subunit is relatively hydrophilic and contains four of the five potential carbohydrate acceptor sites. The subunit order is NH_2-F_2-F_1-COOH. The nucleotide sequences at the 5′ and 3′ mRNA termini are conserved among the eight RS viral mRNAs sequenced to date. The conserved sequences are: 5′ G-G-G-G-C-A-A-A-U ... A-G-U-$\frac{A}{U}$-A-$(N)_{0-2}$-$\frac{A}{U}$-U-poly(A). These are candidates to be signals for viral transcription. The nucleotide and amino acid sequences described further define the relationship between RS virus and other paramyxoviruses.**

Human respiratory syncytial virus (RS virus) is an important agent of serious respiratory tract disease in infants and children (1). A safe and effective vaccine for RS virus is unavailable, and the factors involved in the unusual pathogenesis of RS virus and the development of host immunity are poorly understood (1).

RS virus has general similarities with the paramyxoviruses and is classified provisionally within that family (2). As with other paramyxoviruses (3), the envelope of RS virus contains two surface glycoproteins. The M_r 84,000 glycoprotein (G) mediates attachment to target cells (4, 5), and the second, M_r 68,000 glycoprotein is the viral fusion (F) glycoprotein (4). Work with Sendai virus and other prototypic paramyxoviruses demonstrated that the F protein directs fusion of viral and cellular membranes resulting in viral penetration and can direct fusion of infected cells with adjoining cells resulting in the formation of syncytia (3, 6). Syncytia formation is both a prominent cytopathic effect and an additional mechanism of viral spread (6). Neutralization of fusion activity is important in host immunity (6). The active paramyxovirus F protein consists of two disulfide-linked units, F_1 and F_2, which are generated from an inactive precursor, F_0, by a specific internal cleavage by cellular proteases (7–11). The availability of an appropriate cellular protease and the susceptibility of the F_0 protein to the requisite cleavage are important parameters in paramyxovirus pathogenesis (9, 10). However, the sequence of the F protein was not known for either the prototypic paramyxoviruses or for RS virus.

By the use of cDNA cloning, we recently showed that the RS viral genome, a single negative strand of RNA (12, 13), encodes 10 unique mRNAs (12–16). The proteins encoded by the 10 viral mRNAs were identified by hybrid-selection and *in vitro* translation (16). One of the *in vitro* translation products, a M_r 59,000 species, was identified by peptide mapping as the unglycosylated and uncleaved form of the F protein (refs. 12, 17; unpublished data). These data identified mRNA 5, M_r 0.74×10^6 (13, 15), as the F mRNA.

In the work described here, we constructed complete cDNA clones of the F mRNA and deduced, by DNA sequencing, the sequences of the mRNA and the encoded F protein.

MATERIALS AND METHODS

cDNA Clones. cDNAs were synthesized with intracellular viral mRNA as template, inserted into the *Pst* I site of pBR322 by homopolymer tailing, and cloned in *Escherichia coli* HB101. cDNAs 5-1 and 5-2 were isolated from a library constructed in previous work (15). A second library was constructed essentially as described by Land *et al.* (18), except that size selection on a 1.5% agarose gel followed the final poly(dC) tailing step.

DNA Sequence Analysis. DNA sequences were determined by the chemical method of Maxam and Gilbert (19).

Primer Extension. Recombinant plasmid 5-26 (100 μg) was digested with *Hpa* II and *Dde* I, and the 228-base-pair (bp) fragment spanning the cDNA terminus (bp 3550 in pBR322 to bp 149 in the 5-26 sequence, inclusive) was isolated, 5′ end-labeled (19), digested with *Pst* I, and electrophoresed on a 6% sequencing gel. The 89-nucleotide (nt), antimessage sense strand (nt 61–149 in cDNA 5-26), which was resolved from its complement because it was 7 nt longer, was recovered and used as a primer. Each primer extension reaction (20 μl) contained 0.5 pmol of primer (5×10^5 dpm), 5 μg of intracellular viral mRNA (16), 100 mM Tris·HCl (pH 8.3), 140 mM KCl, 8 mM magnesium acetate, 30 mM 2-mercaptoethanol, 80 μg of actinomycin D per ml, 200 μM dATP, and 100 μM (each) dCTP, dTTP, and dGTP. Reactions were preincubated at 42°C for 30 min, 20 units of reverse transcriptase (Life Sciences, St. Petersburg, FL) was added, and

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviations: RS virus, respiratory syncytial virus; F protein, fusion protein; F mRNA, fusion protein mRNA; nt, nucleotide(s); bp, base pair(s).

*Present address: Laboratory of Infectious Diseases, National Institutes of Health, Bethesda, MD 20205.
†Present address: Department of Pathology, Case Western University School of Medicine, Cleveland, OH 44106.

incubation was continued for 1 hr. Chain-termination reactions contained, in addition, 33 μM ddGTP, 50 μM ddATP, 33 μM ddTTP, or 50 μM ddCTP. Nucleic acids were precipitated with ethanol and analyzed on a 6% sequencing gel.

Chemical Sequencing of the Extended Primer. Four picomoles of 5′ end-labeled primer (4×10^6 dpm) were extended as described above, precipitated once with ethanol, resuspended in 100 μl of water, dissociated from the RNA template by boiling for 2 min, cooled, immediately used to prepare a 150-μl reaction containing 25 mM potassium cacodylate (pH 7.2), 1 mM $CoCl_2$, 0.2 mM dithioerythritol, 0.5 mM dTTP, and 75 units of terminal transferase (P-L Biochemicals), and incubated for 4 min at 15°C. Nucleic acids were purified with phenol, precipitated with ethanol, and electrophoresed on a 6% sequencing gel. The extended, tailed primer was recovered and sequenced by the chemical method.

RESULTS AND DISCUSSION

cDNA Sequencing. The strategy for sequencing cDNAs of the F mRNA is shown in Fig. 1. Every position in the complete nucleotide sequence was determined from a minimum of two independent cDNAs. Much of the sequence was obtained from two nearly complete cDNAs, 5-1 and 5-2; sequencing was completed on seven approximately full-length cDNAs represented by cDNA 5-26 (Fig. 1). The cDNA end that is 5′ relative to the mRNA was identified by primer extension experiments described below. The 3′ end was identified by the detection of terminal poly(dA) (8–20 residues) in five of the seven approximately full-length cDNAs.

Mapping and Sequencing of the 5′ mRNA End. To examine the sequence of the 5′ mRNA end, an 89-nt DNA primer was hybridized to mRNA and extended to the 5′ terminus with reverse transcriptase (Fig. 2*a*). If the 5-26 cDNA insert exclusive of the poly(dG) tract contained the complete 5′ mRNA sequence, the primer would have been increased in length by 56 nt. Instead, the major extension product was increased by 60 nt. A second, minor extension product that was longer than the major product by 1 nt was also obtained. In previous work, analogous longer extension products were obtained and appeared to be artifactual, possibly generated by inefficient copying of the mRNA cap (20). We therefore considered the major extension product to represent the authentic mRNA end.

Primer extension under conditions of partial chain termination (Fig. 2*a*) identified the additional 4 nt at the mRNA terminus as, written in mRNA sense, 5′ N-G-G-G. To precisely identify the terminal nucleotide, an extended primer (5′ end-labeled) was prepared, a poly(dT) tail was added to the 3′ end, and the product was sequenced by the chemical method (Fig. 2*b*). The deduced sequence of the mRNA end was 5′ (G)-*G-G-G-G*. The G in parentheses represents the additional nucleotide contributed by the minor extension product described above and might represent a cap structure. We concluded that the italicized sequence represents the authentic, exact 5′ mRNA end. In the cloned cDNA, these sequences are indistinguishable from the adjoining poly(dG) tail added during cloning. Thus, five of the cDNAs analyzed in this work contain the complete mRNA sequence.

Nucleotide Sequence of the F mRNA. The complete sequence of the F mRNA exclusive of poly(A) is 1899 nt long

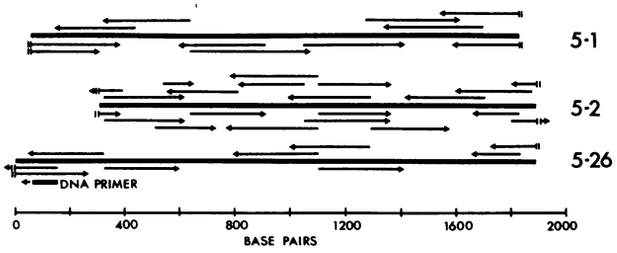

FIG. 1. Strategy for DNA sequencing. cDNAs 5-1, 5-2, and 5-26 are drawn to scale with the poly(dG)·poly(dC) tails omitted. The orientation from left to right is 5′ to 3′ relative to the mRNA. The arrows indicate the directions and extents of sequence determination in the message-sense strand (arrows above the line) and antimessage strand (below the line). Double slash marks in certain arrows indicate that sequencing spanned the F cDNA boundary. Seven additional, confirmatory fragments analyzed in 5-2 are not shown. The DNA fragment that was used in primer extension experiments is shown. Although not shown, a second complete cDNA, 5-7, was sequenced to an extent similar to 5-26, and the sequences at the termini of 5-26 and 5-7 were confirmed in five additional cDNAs.

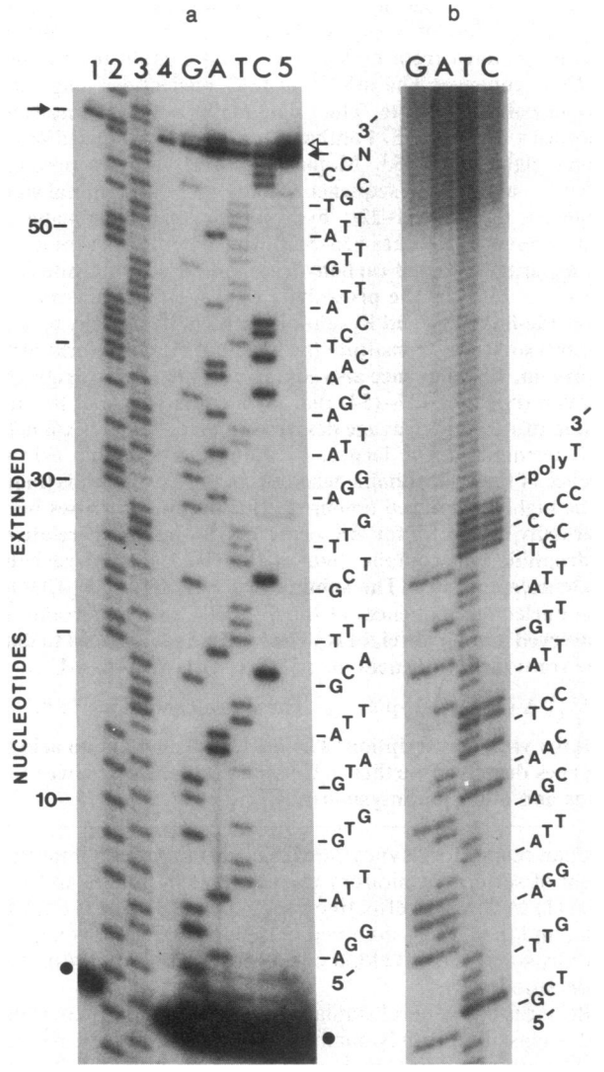

FIG. 2. Mapping the 5′ end of the F mRNA. (*A*) An 89-nt, 5′ end-labeled primer was hybridized to mRNA and extended to the 5′ mRNA end: 10^4 dpm (lanes 1 and 4) and 2×10^5 dpm (lane 5) of the reaction products were analyzed on a 6% sequencing gel in parallel with G + A (lane 2) and C + T (lane 3) chemical sequencing ladders prepared from an unrelated DNA. The positions of the unreacted primer (closed circle), major extension product (closed arrow), and minor extension product (open arrow) are marked. Lanes G, A, T, and C represent partial chain-termination reactions containing the appropriate ddNTP. The deduced nucleotide sequence (antimessage sense) is shown. (*B*) The 5′-labeled primer was elongated on mRNA, extended further at the 3′ end with poly(dT), sequenced by the chemical method, analyzed on a 6% sequencing gel, and visualized with an intensifying screen. The lane headings denote the reaction specificities: lane G, G; lane A, G + A; lane T, C + T; and lane C, C. The deduced nucleotide sequence (antimessage sense) is shown.

(Fig. 3). The mRNA contains a single long open reading frame, which begins with the first ATG (nt 14–16) and accounts for 91% of the sequence. The other reading frames, including the three in the antimessage strand, are blocked extensively throughout the sequence (not shown). The proposed major open reading frame encodes a polypeptide of 574 amino acids with a calculated M_r of 63,453. This is in good agreement (8% larger) with the M_r of 59,000 estimated by gel electrophoresis for unprocessed F protein synthesized *in vitro* (12, 14, 16).

The next longest open reading frame initiates with the triplet ATG (nt 39–41), ends with the triplet TGA (nt 156–158), and would encode a polypeptide of 39 amino acids (Fig. 3). This open reading frame is noteworthy because of its proximity to the 5′ mRNA end: the initiator codon is the second ATG in the sequence. Translation of secondary open reading frames has been reported in cases in which the initiating codons are 5′ proximal (21, 22). However, we have no evidence that this second open reading frame in the F mRNA is utilized.

Preliminary comparison of the sequence of the F mRNA with those of the 1A, 1B, and 1C mRNAs (unpublished results) and the mRNAs that encode the major nucleocapsid protein (ref. 23; unpublished results), matrix protein (24), 24-kilodalton protein, and G protein (unpublished data) identified consensus sequences at both mRNA termini (Fig. 4). These conserved sequences were overlooked in published work (23, 24). The mRNAs of Sendai virus (Fig. 4) and the rhabdovirus vesicular stomatitis virus (26) contain similar consensus termini (Fig. 4) that are transcribed from the conserved 22- or 23-nt sequences that extend across each intergenic region in the viral genomes (20, 22, 25, 26). For Sendai virus and vesicular stomatitis virus, these sequences are thought to have functional roles in viral transcription and have been proposed to have evolutionary significance (20, 22, 25, 26). The identification of similar sequences in RS viral mRNAs suggests that some fundamental similarities exist in the structure, function, and possibly origin of these viral RNAs. As shown in Fig. 4, the 5′ conserved RS viral sequence, 5′ G-G-G-G-C-A-A-A-U ..., is generally similar in length and sequence to that of Sendai virus. The 3′ termini of several RS viral mRNAs resembled the Sendai viral 3′ consensus sequence. However, the 3′ sequence that was conserved among all eight RS viral mRNAs, 5′ ... A-G-U-$\frac{A}{U}$-A-$(N)_{0\text{–}2}$-$\frac{A}{U}$-U-poly(A), differed in length, position, and sequence from the Sendai viral consensus sequence. It will be of interest to extend this preliminary comparison to the RS viral intergenic sequences as well as to the sequences of other paramyxoviruses as they become available.

Structure of the F Protein. The amino acid sequence presented in Fig. 3 provided the basis for predicting major struc-

```
                                G GGG CAA ATA ACA ATG GAG TTG CTA ATC CTC AAA GCA AAT GCA ATT ACC ACA ATC CTC ACT GCA GTC ACA TTT TGT TTT GCT TCT  85
                                              Met Glu Leu Leu Ile Leu Lys Ala Asn Ala Ile Thr Thr Ile Leu Thr Ala Val Thr Phe Cys Phe Ala Ser  24

GGT CAA AAC ATC ACT GAA GAA TTT TAT CAA TCA ACA TGC AGT GCA GTT AGC AAA GGC TAT CTT AGT GCT CTG AGA ACT GGT TGG TAT ACC AGT GTT ATA ACT ATA  190
Gly Gln Asn Ile Thr Glu Glu Phe Tyr Gln Ser Thr Cys Ser Ala Val Ser Lys Gly Tyr Leu Ser Ala Leu Arg Thr Gly Trp Tyr Thr Ser Val Ile Thr Ile  59

GAA TTA AGT AAT ATC AAG GAA AAT AAG TGT AAT GGA ACA GAT GCT AAG GTA AAA TTG ATA AAA CAA GAA TTA GAT AAA TAT AAA AAT GCT GTA ACA GAA TTG CAG  295
Glu Leu Ser Asn Ile Lys Glu Asn Lys Cys Asn Gly Thr Asp Ala Lys Val Lys Leu Ile Lys Gln Glu Leu Asp Lys Tyr Lys Asn Ala Val Thr Glu Leu Gln  94

TTG CTC ATG CAA AGC ACA CCA CCA ACA AAC AAT CGA GCC AGA AGA GAA CTA CCA AGG TTT ATG AAT TAT ACA CTC AAC AAT GCC AAA AAA ACC AAT GTA ACA TTA  400
Leu Leu Met Gln Ser Thr Pro Pro Thr Asn Asn Arg Ala Arg Arg Glu Leu Pro Arg Phe Met Asn Tyr Thr Leu Asn Asn Ala Lys Lys Thr Asn Val Thr Leu  129

AGC AAG AAA AGG AAA AGA AGA TTT CTT GGT TTT TTG TTA GGT GTT GGA TCT GCA ATC GCC AGT GGC GTT GCT GTA TCT AAG GTC CTG CAC CTA GAA GGG GAA GTG  505
Ser Lys Lys Arg Lys Arg Arg Phe Leu Gly Phe Leu Leu Gly Val Gly Ser Ala Ile Ala Ser Gly Val Ala Val Ser Lys Val Leu His Leu Glu Gly Glu Val  164

AAC AAG ATC AAA AGT GCT CTA CTA TCC ACA AAC AAG GCT GTA GTC AGC TTA TCA AAT GGA GTT AGT GTC TTA ACC AGC AAA GTG TTA GAC CTC AAA AAC TAT ATA  610
Asn Lys Ile Lys Ser Ala Leu Leu Ser Thr Asn Lys Ala Val Val Ser Leu Ser Asn Gly Val Ser Val Leu Thr Ser Lys Val Leu Asp Leu Lys Asn Tyr Ile  199

GAT AAA CAA TTG TTA CCT ATT GTG AAC AAG CAA AGC TGC AGC ATA TCA AAT ATA GAA ACT GTG ATA GAG TTC CAA CAA AAG AAC AAC AGA CTA CTA GAG ATT ACC  715
Asp Lys Gln Leu Leu Pro Ile Val Asn Lys Gln Ser Cys Ser Ile Ser Asn Ile Glu Thr Val Ile Glu Phe Gln Gln Lys Asn Asn Arg Leu Leu Glu Ile Thr  234

AGG GAA TTT AGT GTT AAT GCA GGT GTA ACT ACA CCT GTA AGC ACT TAC ATG TTA ACT AAT AGT GAA TTA TTG TCA TTA ATC AAT GAT ATG CCT ATA ACA AAT GAT  820
Arg Glu Phe Ser Val Asn Ala Gly Val Thr Thr Pro Val Ser Thr Tyr Met Leu Thr Asn Ser Glu Leu Leu Ser Leu Ile Asn Asp Met Pro Ile Thr Asn Asp  269

CAG AAA AAG TTA ATG TCC AAC AAT GTT CAA ATA GTT AGA CAG CAA AGT TAC TCT ATC ATG TCC ATA ATA AAA GAG GAA GTC TTA GCA TAT GTA GTA CAA TTA CCA  925
Gln Lys Lys Leu Met Ser Asn Asn Val Gln Ile Val Arg Gln Gln Ser Tyr Ser Ile Met Ser Ile Ile Lys Glu Glu Val Leu Ala Tyr Val Val Gln Leu Pro  304

CTA TAT GGT GTT ATA GAT ACA CCC TGT TGG AAA CTA CAC ACA TCC CCT CTA TGT ACA ACC AAC ACA AAA GAA GGG TCC AAC ATC TGT TTA ACA AGA ACT GAC AGA  1030
Leu Tyr Gly Val Ile Asp Thr Pro Cys Trp Lys Leu His Thr Ser Pro Leu Cys Thr Thr Asn Thr Lys Glu Gly Ser Asn Ile Cys Leu Thr Arg Thr Asp Arg  339

GGA TGG TAC TGT GAC AAT GCA GGA TCA GTA TCT TTC TTC CCA CAA GCT GAA ACA TGT AAA GTT CAA TCA AAT CGA GTA TTT TGT GAC ACA ATG AAC AGT TTA ACA  1135
Gly Trp Tyr Cys Asp Asn Ala Gly Ser Val Ser Phe Phe Pro Gln Ala Glu Thr Cys Lys Val Gln Ser Asn Arg Val Phe Cys Asp Thr Met Asn Ser Leu Thr  374

TTA CCA AGT GAA ATA AAT CTC TGC AAT GTT GAC ATA TTC AAC CCC AAA TAT GAT TGT AAA ATT ATG ACT TCA AAA ACA GAT GTA AGC AGC TCC GTT ATC ACA TCT  1240
Leu Pro Ser Glu Ile Asn Leu Cys Asn Val Asp Ile Phe Asn Pro Lys Tyr Asp Cys Lys Ile Met Thr Ser Lys Thr Asp Val Ser Ser Ser Val Ile Thr Ser  409

CTA GGA GCC ATT GTG TCA TGC TAT GGC AAA ACT AAA TGT ACA GCA TCC AAT AAA AAT CGT GGA ATC ATA AAG ACA TTT TCT AAC GGG TGC GAT TAT GTA TCA AAT  1345
Leu Gly Ala Ile Val Ser Cys Tyr Gly Lys Thr Lys Cys Thr Ala Ser Asn Lys Asn Arg Gly Ile Ile Lys Thr Phe Ser Asn Gly Cys Asp Tyr Val Ser Asn  444

AAA GGG ATG GAC ACT GTG TCT GTA GGT AAC ACA TTA TAT TAT GTA AAT AAG CAA GAA GGT AAA AGT CTC TAT GTA AAA GGT GAA CCA ATA ATA AAT TTC TAT GAC  1450
Lys Gly Met Asp Thr Val Ser Val Gly Asn Thr Leu Tyr Tyr Val Asn Lys Gln Glu Gly Lys Ser Leu Tyr Val Lys Gly Glu Pro Ile Ile Asn Phe Tyr Asp  479

CCA TTA GTA TTC CCC TCT GAT GAA TTT GAT GCA TCA ATA TCT CAA GTC AAC GAG AAG ATT AAC CAG AGC CTA GCA TTT ATT CGT AAA TCC GAT GAA TTA TTA CAT  1555
Pro Leu Val Phe Pro Ser Asp Glu Phe Asp Ala Ser Ile Ser Gln Val Asn Glu Lys Ile Asn Gln Ser Leu Ala Phe Ile Arg Lys Ser Asp Glu Leu Leu His  514

AAT GTA AAT GCT GGT AAA TCC ACC ACA AAT ATC ATG ATA ACT ACT ATA ATT ATA GTG ATT ATA GTA ATA TTG TTA TCA TTA ATT GCT GTT GGA CTG CTC TTA TAC  1660
Asn Val Asn Ala Gly Lys Ser Thr Thr Asn Ile Met Ile Thr Thr Ile Ile Ile Val Ile Ile Val Ile Leu Leu Ser Leu Ile Ala Val Gly Leu Leu Leu Tyr  549

TGT AAG GCC AGA AGC ACA CCA GTC ACA CTA AGC AAA GAT CAA CTG AGT GGT ATA AAT AAT ATT GCA TTT AGT AAC TAA ATA AAA ATA GCA CCT AAT CAT GTT CTT  1765
Cys Lys Ala Arg Ser Thr Pro Val Thr Leu Ser Lys Asp Gln Leu Ser Gly Ile Asn Asn Ile Ala Phe Ser Asn ***                                      574

ACA ATG GTT TAC TAT CTG CTC ATA GAC AAC CCA TCT GTC ATT GGA TTT TCT TAA AAT CTG AAC TTC ATC GAA ACT CTC ATC TAT AAA CCA TCT CAC TTA CAC TAT  1870

TTA AGT AGA TTC CTA GTT TAT AGT TAT AT                                                                                                  1899
```

FIG. 3. Nucleotide sequence of the F mRNA and the predicted protein sequence. The NH_2-terminal and COOH-terminal hydrophobic domains are underlined. The hydrophobic domain at the predicted NH_2 terminus of the F_1 subunit is marked with double underlining. The arrow marks the predicted site of the proteolytic cleavage that activates fusion activity. The positions of cysteine residues (circles) and potential asparaginyl acceptor sites for carbohydrate (closed triangles) are marked below the amino acid sequence. Nucleotide variability was detected at seven positions (marked with ticks above the nucleotide sequence): position 303, shown as T, was C in one of three cDNAs (changing Met to Thr); position 317, C, was T in two of four cDNAs (Pro to Ser); position 342, A, was C in one of four cDNAs (Glu to Ala); position 573, T, was C in one of four cDNAs (Val to Ala); position 915, T, was C in one of four cDNAs (Val to Ala); position 1148, A, was G in two of four cDNAs (Ile to Val); position 1352, A, was G in two of four cDNAs (Met to Val).

F mRNA	5'GGGGCAAAUAACA	...	UUAUAGUUAUAU-polyA
G mRNA[a]	5'GGGGCAAAUGCAA	...	CAGUAGUUACUU-polyA
1A mRNA[a]	5'NGGGCAAAUAAUC	...	AAGUAGUUAAUU-polyA
1B mRNA[a]	5'(nd)CAAAUAAAU	...	UUAUAGUAAUUU-polyA
1C mRNA[a]	5'(nd)CAAAUAAGA	...	AUUAGUUAAUAU-polyA
24kdal mRNA[a]	5'GGGGCAAAUAUGU	...	CGAUAGUUAUUU-polyA
N mRNA[a]	5'GGGGCAAAUACAA	...	CUUUGAGUUAAU-polyA
N mRNA[b]	5'(nd)CAAAUACAA	...	CAGUAGUUACUU-polyA
M mRNA[c]	5'NGGGCAAAUAUGG	...	AAAUAAGUUAAU-polyA
SENDAI mRNAs[d]	5'AGGG$_A^U$NAA$_A^U$G	...	UAAG-polyA

FIG. 4. Conserved sequences at the termini of RS viral mRNAs. The RS viral sequences were compiled from our unpublished work (a) and the work of Elango and Venkatesan (23) (b) and Satake and Venkatesan (24) (c). As shown, the N mRNA sequence that we obtained (unpublished work) is different at the 3' end than the published sequence (23). This discrepancy will be discussed elsewhere. For comparison, Sendai viral consensus sequences were compiled from published work (20, 22, 25). Symbols in the nucleotide sequences: N, unidentified nucleotide for the RS viral mRNAs and variable nucleotide for the Sendai viral mRNAs; (nd), mRNA terminus has not been sequenced or mapped exactly.

tural features of the F_0 protein. These are discussed below in the context of information available from other sources.

Glycosylation of the F_0 protein was inhibited completely in the presence of tunicamycin (12, 17), indicating that carbohydrate side chains are attached exclusively through N-linkage to asparagine residues. The glycosylated and nonglycosylated forms of F_0 protein differed in apparent size by 11,000 daltons. Assuming a contribution of 2100 daltons per carbohydrate chain (27), five chains are predicted. The amino acid sequence contains five potential acceptor sites for N-linked carbohydrate (Asn-X-Ser/Thr), suggesting that all potential sites are utilized. The placement of four of the acceptor sites in the F_2 subunit (Figs. 3 and 5) is consistent with the electrophoretic mobilities of the glycosylated and nonglycosylated, reduced and nonreduced forms of the F_0 and F protein (12, 17).

The biosynthesis and membrane orientation of glycoproteins is mediated by characteristic hydrophobic "signal" sequences in the polypeptide chain (28, 29). To search for possible hydrophobic signals in the sequence of F_0 protein, a plot of local hydrophobicity versus sequence position was

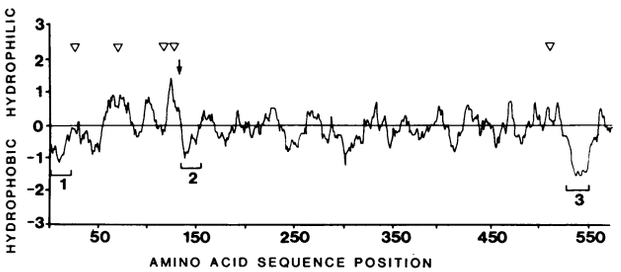

FIG. 5. Plot of hydrophilic (areas above the line) and hydrophobic (below the line) regions in the unprocessed F protein. A window of 14 amino acids was used to calculate local hydrophobicity of each position based on published hydrophobicity values (30) (local hydrophobicity equals the average of the values for each group of 14 amino acids). The deduced positions of (*i*) the NH_2-terminal signal peptide, (*ii*) the proposed NH_2 terminus of the F_1 subunit, and (*iii*) the transmembrane anchor are shown. The proposed site for the activating proteolytic cleavage is marked with an arrow. Potential acceptor sites for N-linked carbohydrate are marked with open triangles.

prepared (Fig. 5). The NH_2 terminus of the F_0 protein (amino acids 1–22) is strongly hydrophobic and is similar in length and amino acid content to known NH_2-terminal signal peptides (28, 29, 31). By analogy we presume that this sequence is a signal peptide that initiates translocation of nascent F_0 protein across the membrane and into the lumen of the rough endoplasmic reticulum. Most known NH_2-terminal signal peptides are cleaved by a signal peptidase during or after translocation (28, 29, 31). In the case of the RS virus F protein, unglycosylated F_0 protein synthesized in tunicamycin-treated cells was 2000 daltons smaller than its unprocessed counterpart synthesized *in vitro* in the absence of membranes (12). We attribute this difference to a cleaved signal sequence (29). Thus, the amino acid sequence shown in Fig. 3 probably represents pre-F_0 protein; F_0 protein would be generated by cleavage of the signal peptide. Based on the size estimate for the signal peptide and the observed preference of the signal peptidase for cleavage following glycine, serine, alanine, and cysteine (31), cleavage might occur following amino acids 17, 21, or 23.

The most strongly hydrophobic region of the F_0 protein is near the COOH terminus (amino acids 525–550, Figs. 3 and 5). The length and amino acid content of this sequence is characteristic of known "stop-transfer" signals (28, 29). These have been shown to terminate passage of the nascent polypeptide through the membrane, with the stop-transfer signal remaining embedded in the membrane and the polypeptide positioned in a transmembrane orientation (28, 29). For the F_0 protein, the region predicted to be external includes the NH_2 terminus, 91% of the amino acid sequence, all five potential carbohydrate acceptor sites, and 15 of 16 cysteine residues (Fig. 3). The COOH-terminal region, which consists of 24 amino acids and is moderately hydrophilic and moderately basic, would remain in the cytoplasm. This is consistent with the membrane orientation of the Sendai virus F_0 protein (32). The single cysteine residue at the cytoplasmic membrane face is a potential acceptor site for palmitate (33).

As has been demonstrated with Sendai virus and other prototypic paramyxoviruses, the NH_2 terminus of the paramyxovirus F_1 subunit is generated by the cleavage that activates the F protein. Amino acid sequencing determined that this newly exposed terminus is strongly hydrophobic (34, 35). For RS virus, the NH_2 terminus of the F_1 subunit has not been identified directly. On the basis of the predicted amino acid sequence (Fig. 3), the best candidate to be the site of the activating cleavage is the sequence Lys-Lys-Arg-Lys-Arg-Arg (amino acids 131–136). This sequence would have a marked sensitivity to trypsin-like proteases and cleavage would generate F_1 and F_2 subunits with molecular weights in agreement with estimates made by gel electrophoresis (12, 36, 37). The predicted subunit order, NH_2-F_2-F_1-COOH, is the same as for Sendai virus (11, 32). More important, the NH_2 terminus that would be created by cleavage at this site is strongly hydrophobic (Figs. 3 and 5).

For other paramyxoviruses, several lines of evidence suggested that the hydrophobic NH_2 terminus of the F_1 polypeptide has an important role in fusion: (*i*) the cleavage that activates fusion activity exposes a new hydrophobic domain in the native protein (38), (*ii*) the amino acid sequence of the F_1 subunit terminus is highly conserved among several paramyxoviruses (35), and (*iii*) synthetic peptides that mimic this sequence inhibit fusion (35). It therefore was of interest to compare the predicted NH_2-terminal sequence of the F_1 polypeptide of RS virus with that of Sendai virus (Fig. 6). Surprisingly, the sequences are dissimilar: 4 of 20 positions are in agreement, compared to 14 of 20 amino acids in common among Sendai virus, Newcastle disease virus, and simian virus 5, three paramyxoviruses that have different avian or mammalian hosts (35). A search of the complete amino

```
                1                                   10                                  20
RS VIRUS  NH2-Phe-Leu-Gly-Phe-Leu-Leu-Gly-Val-Gly-Ser-Ala-Ile-Ala-Ser-Gly-Val-Ala-Val-Ser-Lys ...

SENDAI    NH2-Phe-Phe-Gly-Ala-Val-Ile-Gly-Thr-Ile-Ala-Leu-Gly-Val-Ala-Thr-Ala-Ala-Gln-Ile-Thr ...
```

FIG. 6. Comparison of the amino acid sequence at the proposed NH_2 termini of the F_1 polypeptide of RS virus with that of Sendai virus. The Sendai viral sequence is that of Richardson *et al.* (35); a similar sequence was determined by Gething *et al.* (34), except that position 8 is Ile and position 13 is Pro. The underlined positions in the Sendai viral sequence are identical in simian virus 5 and Newcastle disease virus (35). The double-underlined positions in the RS viral sequence are identical in Sendai virus.

acid sequence of the RS viral F_0 protein for tripeptides resembling the NH_2 termini of the F_1 polypeptides of the other paramyxoviruses provided no evidence of an alternate, more homologous candidate to be the NH_2 terminus of the RS viral F_1 polypeptide. As has been noted previously (3, 34), the NH_2 termini of the F_1 polypeptide of Sendai virus and the HA_2 polypeptides of influenza virus types A and B are similar in sequence for the first 12 or 13 amino acids, having identity at 6 positions. This similarity in structure was thought to reflect similarity in function. However, the corresponding sequences of RS virus and human influenza virus types A, B, and C (39) share only 1–4 amino acids out of 20 (not shown). The two common features of these RS viral, Sendai viral, and influenza viral sequences are (*i*) overall hydrophobicity and (*ii*) the occurrence of glycine residues at intervals in each sequence, which, as destabilizers of α-helical and β-sheet conformations (30), might contribute to similar secondary structures (40). Among RS virus and the other paramyxoviruses, a third similarity is the sequence NH_2-Phe-X-Gly at the exact F_1 subunit terminus. Otherwise, the general dissimilarity of the RS viral sequence suggests that fusion activity depends upon general features of hydrophobicity and possibly secondary structure rather than upon a common amino acid sequence.

In conclusion, this report describes the sequences and structural features of the F mRNA and F protein of RS virus. The general structural features described here for the F protein were consistent with the available data for other paramyxoviruses. However, specific nucleotide and amino acid sequences that might be expected to have been conserved among closely related viruses were dissimilar in RS virus in comparison with Sendai virus. These observations are suggestive of evolutionary distance within the paramyxovirus family between RS virus and other paramyxoviruses exemplified by Sendai virus.

Note Added in Proof. Further work has identified the 5′ termini of the 1A, 1B, and 1C mRNAs as 5′ G-G-G-G, 5′ N-G-G-G, and 5′ G-G-G-G, respectively, with N representing an unidentified nucleotide. Comparison of the sequence of the phosphoprotein mRNA with the sequences shown in Fig. 4 indicates that the 3′ terminal sequence that is conserved among nine RS viral mRNAs is 5′ A-G-U-$\overset{A}{U}$-A-$(N)_{1-4}$-poly(A).

We thank our friends and colleagues for their interest and encouragement. We thank Roh Swanstrom for introducing us to sequence analysis and John Glass and Andrew Ball for helping us compute. This work was supported by grants from the National Institute of Allergy and Infectious Diseases (A112464 and A120181) and a training grant from the National Cancer Institute (CA09156). P.L.C. received additional support from the National Institute of Allergy and Infectious Diseases (A106956).

1. Chanock, R. M., Kim, H. W., Brandt, C. D. & Parrott, R. H. (1982) in *Viral Infections of Humans: Epidemiology and Control*, ed. Evans, A. S. (Plenum, New York), pp. 471–489.
2. Kingsbury, D. W., Bratt, M. A., Choppin, P. W., Hanson, R. P., Hosaka, Y., ter Meulen, V., Norrby, E., Plowright, W., Rott, R. & Wunner, W. H. (1978) *Intervirology* **10**, 137–152.
3. Choppin, P. W. & Scheid, A. (1980) *Rev. Infect. Dis.* **2**, 40–61.
4. Walsh, E. E. & Hruska, J. (1983) *J. Virol.* **47**, 171–177.
5. Walsh, E. E., Schlesinger, J. J. & Brandriss, M. W. (1984) *J. Gen. Virol.* **65**, 761–767.
6. Merz, D. C., Scheid, A. & Choppin, P. W. (1980) *J. Exp. Med.* **151**, 275–288.
7. Homma, M. & Ohuchi, M. (1973) *J. Virol.* **12**, 1457–1465.
8. Scheid, A. & Choppin, P. W. (1974) *Virology* **57**, 475–490.
9. Nagai, Y., Klenk, H.-D. & Rott, R. (1976) *Virology* **72**, 494–508.
10. Scheid, A. & Choppin, P. W. (1976) *Virology* **69**, 265–277.
11. Scheid, A. & Choppin, P. W. (1977) *Virology* **80**, 54–66.
12. Huang, Y. T. (1983) Dissertation (Univ. of North Carolina at Chapel Hill, Chapel Hill, NC).
13. Huang, Y. T. & Wertz, G. W. (1982) *J. Virol.* **43**, 150–157.
14. Huang, Y. T. & Wertz, G. W. (1983) *J. Virol.* **46**, 667–672.
15. Collins, P. L. & Wertz, G. W. (1983) *Proc. Natl. Acad. Sci. USA* **80**, 3208–3212.
16. Collins, P. L., Huang, Y. T. & Wertz, G. W. (1984) *J. Virol.* **49**, 572–578.
17. Huang, Y. T., Collins, P. L. & Wertz, G. W. (1984) *Abstr. Annu. Meet. Am. Soc. Microbiol.*, abstr. T27, 233.
18. Land, H., Gretz, M., Hauser, H., Lindenmaier, W. & Schutz, G. (1981) *Nucleic Acids Res.* **9**, 2251–2266.
19. Maxam, A. M. & Gilbert, W. (1980) *Methods Enzymol.* **65**, 499–559.
20. Gupta, K. C. & Kingsbury, D. W. (1984) *Nucleic Acids Res.* **12**, 3829–3841.
21. Shaw, M. W., Choppin, P. W. & Lamb, R. A. (1983) *Proc. Natl. Acad. Sci. USA* **80**, 4879–4883.
22. Giorgi, C., Blumberg, B. M. & Kolakofsky, D. (1983) *Cell* **35**, 829–836.
23. Elango, N. & Venkatesan, S. (1983) *Nucleic Acids Res.* **11**, 5941–5951.
24. Satake, M. & Venkatesan, S. (1984) *J. Virol.* **50**, 92–99.
25. Gupta, K. C. & Kingsbury, D. W. (1982) *Virology* **120**, 518–529.
26. Rose, J. K. (1980) *Cell* **19**, 415–421.
27. Hunter, E., Hill, E., Hardwick, M., Bhown, A., Schwartz, D. E. & Tizard, R. (1983) *J. Virol.* **46**, 920–936.
28. Blobel, G. (1980) *Proc. Natl. Acad. Sci. USA* **77**, 1496–1500.
29. Sabatini, D. D., Kreibich, G., Morimoto, T. & Adesnik, M. (1982) *J. Cell Biol.* **92**, 1–22.
30. Levitt, M. (1976) *J. Mol. Biol.* **104**, 59–107.
31. Austen, B. M. (1979) *FEBS Lett.* **103**, 308–313.
32. Bowen, H. A. & Lyles, D. (1981) *J. Virol.* **37**, 1079–1082.
33. Rose, J. K., Adams, G. A. & Gallione, C. J. (1984) *Proc. Natl. Acad. Sci. USA* **81**, 2050–2054.
34. Gething, M. J., White, J. M. & Waterfield, M. D. (1978) *Proc. Natl. Acad. Sci. USA* **75**, 2737–2340.
35. Richardson, C. D., Scheid, A. & Choppin, P. W. (1980) *Virology* **105**, 205–222.
36. Fernie, B. F. & Gerin, J. L. (1982) *Inf. Immun.* **37**, 243–249.
37. Lambert, D. M. & Pons, M. W. (1983) *Virology* **130**, 204–214.
38. Hsu, M., Scheid, A. & Choppin, P. W. (1981) *J. Biol. Chem.* **256**, 3557–3563.
39. Nakada, S., Creager, R. S., Krystal, M., Aaronson, R. P. & Palese, P. (1984) *J. Virol.* **50**, 118–124.
40. Wilson, I. A., Skehel, J. J. & Wiley, D. C. (1981) *Nature (London)* **289**, 366–373.