Respiratory syncytial virus envelope glycoprotein (G) has a novel structure

Masanobu Satake¹, John E. Coligan², Narayanasamy Elango³, Erling Norrby⁴ and Sundararajan Venkatesan^{5*}

Laboratory of Infectious Diseases, ²Laboratory of Immunogenetics, ³Laboratory of Viral Diseases, ⁵Laboratory of Molecular Microbiology, National Institute of Allergy and Infectious Diseases, Bethesda, MD 20205, ¹Basic Research Program, Litton Bionetics, Inc., FCRF, Frederick, MD 21701, and ⁴Department of Virology, State Bacteriological Laboratory and Karolinska Institute, S-105 21, Stockholm, Sweden

Received 6 May 1985; Revised and Accepted 7 October 1985

ABSTRACT

Amino acid sequence of human respiratory syncytial virus envelope glycoprotein (G) was deduced from the DNA sequence of a recombinant plasmid and confirmed by limited amino acid microsequencing of purified 90K G protein. The calculated molecular mass of the protein encoded by the only long open reading frame of 298 amino acids was 32,588 daltons and was somewhat smaller than the 36K polypeptide translated *in vitro* from mRNA selected by this plasmid. Inspection of the sequence revealed a single hydrophobic domain of 23 amino acids capable of membrane insertion at 41 residues from the N-terminus. There was no N-terminal signal sequence and the hydrophilic N-terminal 20 residues probably represent the cytoplasmic tail of the protein. The N-terminally oriented membrane insertion was somewhat analogous to paramyxovirus hemagglutinin-neuraminidase (HN) and influenza neuraminidase (NA). The protein was moderately hydrophilic and rich in hydroxy-amino acids. It was both N- and O-glycosylated with the latter contributing significantly to the net molecular mass 90K.

INTRODUCTION

Enveloped negative strand RNA viruses attach to receptors of susceptible cells via one or more glycoproteins that are required for entry and cell to cell spread (1,2). Additionally these proteins determine host tropism and antigenic variation. Among the several paramyxoviruses, these functions are mediated by two vectorially inserted membrane glycoproteins, namely the HN (hemagglutinin-neuraminidase) and F (fusion) protein (3,4). Biochemical studies with the purified glycoproteins and genetic analysis of natural variants or those derived *in vitro* have firmly established a receptor binding function for HN and a role in cell penetrance by membrane fusion for the F protein (3-6). Unlike the segmented influenza viruses, both the receptor binding (hemagglutinin) and the neuraminidase (receptor destroying) activities reside on the same polypeptide, HN of the paramyxoviruses. The F protein is synthesized as a precursor F₀ that is activated by a specific proteolytic cleavage by host cells during cytoplasmic transport (6). Although antibodies against the F protein inhibit cell to cell spread following infection (7), the

HN protein is of primary importance since it is the attachment protein of these viruses (8,9). In light of this, somatic mutations in the HN can afford the virus an opportunity to undergo variation and thereby evade immune surveillance.

Human respiratory syncytial (RS) virus, a Pnuemovirus, is a pleomorphic negative strand RNA virus. Although similar in many respects to the paramyxoviruses, it is distinct in several biochemical and morphological characteristics (10). In particular, RS virus possesses nucleocapsids of distinct morphology (11) and lacks both the hemagglutinating and neuraminidase activities (12). The negative strand genome encodes at least 10 genes and has a novel transcriptional order in that two nonstructural (NS) protein genes rather than the major nucleocapsid (NP) protein gene are 3' proximal on the linear genetic map (13,14). Of the two surface glycoproteins, the identity of a 68K protein has been determined to be F_0 by cDNA cloning and sequencing (15,16). The character of the other glycoprotein (G) thought to be the equivalent of HN has been sketchy so far. Its molecular mass has been estimated to be 84-90 kilodaltons (K) and it is heavily glycosylated (17-21). Unlike the paramyxovirus HN (22), it does not exist in multimeric forms but apparently undergoes processing as reflected by the somewhat smaller size of the extracellular form (18).

Notwithstanding its inherent novel molecular aspects, RS virus is an important human pathogen in early childhood for which no effective immunoprophylaxis is presently available (10). Therefore, knowledge of the primary structure of the surface glycoproteins was sought to hopefully gain insights into the understanding of the nature of the virus-host cell interaction and the mechanism of antigenic variation. Such information is important for devising a strategy of vaccination. Viral gene products have been identified by a combination of immunological and recombinant DNA techniques and several viral genes sequenced and their reading frames established (13-16, 21, 23-26,28,29). The primary structure of the RS virus fusion protein deduced in this manner had several similarities to other paramyxovirus F_0 proteins and influenza virus hemagglutinin (HA) (15,16). In this communication, we present the primary sequence of RS virus G protein deduced initially from the cloned cDNA sequence and further confirmed by limited amino acid microsequencing CNBr cleaved G protein. The biological implications of its unusual structure are discussed.

EXPERIMENTAL PROCEDURES

Cells and Virus:

RS virus (strain A2) was grown on Hep-2 cell monolayers. Actinomycin-D (1 μ g/ml) was routinely present both for virus purification and mRNA isolation by guanidine isothiocyanate method (30). Different batches of mRNAs were screened by cell free translation (24). For protein labeling cells were infected at a multiplicity of 5 plaque forming units/cell and the appropriate (3 H)-amino acid or (35 S)-methionine was added 2h post infection. Preparation of cell extracts for immunoprecipitation has been described (16).

DNA sequencing and related procedures:

Characterization of recombinant cDNA plasmids by several standard procedures has been described (21, 23-26). DNA sequencing was according to Maxam and Gilbert (31). Sequence-related plasmids were further confirmed by comparing the cloned sequences with the results of dideoxysequencing of poly(A) RNA from infected cells using restriction fragment primers from individual plasmids. In this manner plasmids containing noncontiguous sequences (due to cloning artifacts or transcripts from DI genomes etc.) were eliminated. The 5' end(s) of the transcript(s) were similarly deduced by primer extension (24). Queen and Korn program (32) was used for computer analysis. Hydropathicity diagrams of the protein was according to Kyte and Doolittle (33) and the Wilbur and Lipman algorithm (34) was used for homology searches against protein sequence databases.

Immunoprecipitation:

Preparation and characterization of mouse anti-RS virus monoclonal antibodies will be described elsewhere. Cytosolic protein from 10^8 infected cells in 4 ml was incubated with 0.1 volume of 1:10 dilution of pooled ascites fluids specific for G protein for 4h and the Ag/Ab complexes were recovered by binding to and elution from protein-A Sepharose beads (35). The G protein was subsequently purified by SDS (sodium dodecyl sulfate)-polyacrylamide gel electrophoresis, extracted from gel and dialyzed exhaustively against 0.001% SDS prior to CNBr cleavage as described (36).

Amino acid sequencing:

The G protein was digested with cyanogen bromide (CNBr) as described by Ewenstein et al (37). CNBr cleaved protein fragments were resolved by gel filtration on Sephacryl-200 column. The column (2 x 195 cm.) was equilibrated in 6M guanidine-HCl plus 0.5% acetic acid. Fraction size was 2 ml and the flow rate was 0.2 ml/min. Automated N-terminal microsequencing was by Beckman 890C sequenator and the Beckman 0.1 M Quadrol program (38). For each sequence

step, 25% of the butyl chloride extract was removed for determination of radioactivity. The amino acid derivatives of those fractions showing radioactivity above background levels were converted to PTH (phenylthiohydantoin) derivatives and mixed with PTH-amino acid standards. Individual PTH derivatives were resolved by HPLC (38).

Immunofluorescence:

Cell monolayers on coverslips were infected at a multiplicity of 0.05 plaque forming units/cell, and processed 24h later. Live cell indirect immunofluorescence using monoclonal antibodies and fluorescein conjugated anti-mouse IgG has been described (39). Where indicated tunicamycin (0.2 μ g/ml) or monensin (0.2 μ M) were added 2h post infection.

RESULTS

Identification of a RS viral cDNA plasmid encoding a 33K protein: During initial characterization of a recombinant cDNA library constructed using viral mRNAs, three nonoverlapping classes of recombinants were detected (21,40). They could not be readily assigned to definite viral genes since the translation products of mRNAs hybrid selected by these plasmids did not correspond to any of the viral proteins identified till then (40). Serial RNA filter hybridizations by the Northern blotting procedures demonstrated these plasmids to be complementary to three viral mRNAs of ca. 2000, 1050 and 1000 nucleotides respectively (40). Recombinant pRSC₂₂ reacted with a unique viral RNA of ca. 1000 bases, that yielded a ca. 36K protein on cell free translation (Fig. 1). A protein of similar size previously shown by others to be derived from a ca. 1kb mRNA designated as 2b (14) had not been visualized either in infected cells or purified virus till then. Various attempts such as *in vitro* translation in alternate systems, addition of selective RNA denaturants during translation, oocyte microinjection of hybrid-selected mRNA, addition of microsomal membranes to promote core glycosylation or nascent translation products were unsuccessful in effecting conversion of the 36K protein to a species of higher molecular weight.

Since pRSC₂₂ had a long poly(A) tail at one end of the viral insert and was close to the size of its presumptive template, it was chosen for complete sequence analysis. To eliminate possible DNA sequencing and cDNA cloning artifacts additional recombinants hybridizing with pRSC₂₂ were identified. One of these (pRSC₁₀) lacked 70 nucleotides corresponding to the 5' end of the mRNA but had a perfect nucleotide match with pRSC₂₂ throughout the remainder of the coding sequence and upto the poly(dA) tail. Two other recombinants had smaller inserts (ca. 600 bp) and represented sequences corresponding to the 5'

Fig. 1. In vitro translation of mRNA hybrid selected by recombinant pRSC₂₂. Conditions for preparation of DNA filters, hybridization and cell free translation have been described (22). (³⁵S) labeled translation products were analyzed by SDS/polyacrylamid (15%) gel electrophoresis and visualized by fluorography. Results of translation using when total infected cell RNA (lane A), infected cell RNA hybrid selected by pRSC₂₂ (lane B) or uninfected cell RNA selected by pRSC₂₂ (lane C) are shown. Commercial radioactive protein standards were run in lane D.

and 3' regions of the mRNA respectively. Other potentially related recombinants were sequenced to a limited extent but later abandoned since they had significant sequence divergence from pRSC₁₀ and pRSC₂₂ notably at either end of their inserts. During this analysis, cDNA plasmids containing

Fig. 2. Dideoxysequencing of 5' end of G protein mRNA. The strategy is schematically illustrated. Conditions of hybridization and sequencing have been described (14,22). Lanes G,A,T, and C represent reactions using ddGTP, ddATP, ddTTP and ddCTP inhibitors.

sequences of nonadjacent transcripts were detected. In particular, one recombinant (pRSB₁₁) contained cDNA sequence of viral NP transcript covalently linked to the 3'-284 nucleotides of the 36K transcript. pRSB₁₁ was previously reported to encode the viral NP protein (23). Of the three plasmids (pRSB₅, pRSB₁₀ and pRSB₁₁) encoding the NP gene and possessing a poly(dA) tail, it had the largest insert and its open reading frame extended beyond those found in pRSB₅ and pRSB₁₀. All of them hybridized to a single mRNA of ca. 1400 bases and the selected mRNA yielded NP as the exclusive translation product. Since the 5' ends of all three inserts were within a few nucleotides of one another, and since the cloned viral sequence in pRSB₁₁ was nearest in size to the putative NP mRNA, it was erroneously deemed to be the candidate clone of the NP transcript. The sequence divergence near the poly (dA) string of pRSB₅ and

CAAATGCAAAAC

```

      10
ATG TCC AAA AAC AAG GAC CAA CGC ACC GCT AAG ACA TTA GAA AGG ACC TGG GAC ACT 20
MET SER LYS ASN LYS ASP GLN ARG THR ALA LYS THR LEU GLU ARG THR TRP ASP THR CTC
      30
AAT CAT TTA TTA TTC ATA TCA TCG TGC TTA TAT AAG TTA AAT CTT AAA TCT GTA GCA CAA
ASN HIS LEU LEU PHE ILE SER SER CYS LEU TYR LYS LEU ASN LEU LYS SER VAL ALA GLN
      50
ATC ACA TTA TCC ATT CTG GCA ATG ATA ATC TCA ACT TCA CTT ATA ATT GCA GCC ATC ATA
ILE THR LEU SER ILE LEU ALA MET ILE ILE SER THR SER LEU ILE ILE ALA ALA ILE ILE
      70
TTC ATA GCC TCG GCA AAC CAC AAA GTC ACA CCA ACA ACT GCA ATC ATA CAA GAT GCA ACA
PHE ILE ALA SER ALA ASN HIS LYS VAL THR PRO THR THR THR ALA ILE ILE GLN ASP ALA THR
      90
AGC CAG ATC AAG AAC ACA ACC CCA ACA TAC CTC ACC CAG AAT CCT CAG CTT GGA ATC AGT
SER GLN ILE LYS ASN THR THR PRO THR TYR LEU THR GLN ASN PRO GLN LEU GLY ILE SER
      110
CCC TCT AAT CCG TCT GAA ATT ACA TCA CAA ATC ACC ACC ATA CTA GCT TCA ACA ACA CCA
PRO SER ASN PRO SER GLU ILE THR SER ILE THR THR THR LEU ALA SER THR THR THR PRO
      130
GGA GTC AAG TCA ACC CTG CAA TCC ACA ACA GTC AAG ACC AAA AAC ACA ACA ACA ACT CAA
GLY VAL LYS SER THR LEU GLN SER THR THR VAL LYS THR LYS ASN THR THR THR THR GLN
      150
ACA CAA CCC AGC AAG CCC ACC ACA AAA CAA CGC CAA AAC AAA CCA CCA AGC AAA CCC AAT
THR GLN PRO SER LYS PRO THR THR LYS GLN ARG GLN ASN LYS PRO PRO SER LYS PRO ASN
      170
AAT GAT TTT CAC TTT GAA GTG TTC AAC TTT GTA CCC TGC AGC ATA TGC AGC AAC AAT CCA
ASN ASP PHE HIS PHE GLU VAL PHE ASN PHE VAL PRO CYS SER ILE CYS SER ASN ASN PRO
      190
ACC TGC TGG GCT ATC TGC AAA AGA ATA CCA AAC AAA AAA CCA GGA AAG AAA ACC ACT CCA
THR CYS TRP ALA ILE CYS LYS ARG ILE PRO ASN LYS LYS PRO GLY LYS LYS THR THR THR
      210
AAG CCC ACA AAA AAA CCA ACC CTC AAG ACA ACC AAA AAA GAT CCC AAA CCT CAA ACC ACT
LYS PRO THR LYS LYS PRO THR LEU LYS THR THR LYS LYS ASP PRO LYS PRO GLN THR THR
      230
AAA TCA AAG GAA GTA CCC ACC ACC AAG CCC ACA GAA GAG CCA ACC ATC AAC ACC ACC AAA
LYS SER LYS GLU VAL PRO THR THR LYS PRO THR GLU GLU PRO THR ILE ASN THR THR LYS
      250
ACA AAC ATC ATA ACT ACA CTA CTC ACC TCC AAC ACC ACA GGA AAT CCA GAA CTC ACA AGT
THR ASN ILE ILE THR THR LEU LEU THR SER ASN THR THR THR GLY ASN PRO GLU LEU THR SER
      270
CAA ATG GAA ACC TTC CAC TCA ACT TCC TCC GAA GGC AAT CCA AGC CCT TCT CAA GTC TCT
GLN MET GLU THR PHE HIS SER THR SER SER GLU GLY ASN PRO SER PRO SER GLN VAL SER
      290
ACA ACA TCC GAG TAC CCA TCA CAA CCT TCA TCT CCA CCC AAC ACA CCA CGC CAG TAG
THR THR SER GLU TYR PRO SER GLN PRO SER SER PRO PRO ASN THR PRO ARG GLN END

```

TTACTTAAAAAAAAAAAA

MOLECULAR WEIGHT = 32588

Fig. 3. Nucleotide sequence of mRNA for G protein and the deduced amino acid sequence. The entire viral sequence in pRSC₂₂ is shown and therefore the four nucleotide NGGG sequence at the 5'-end of the transcript is omitted. All the methionine (MET) residues are underlined. The long hydrophobic stretch between residues 41 and 63 is enclosed. Wavy underlines highlight the potential N-glycosylation sites.

pRSB₁₀ from that of pRSB₁₁ was neglected as possible cloning artifacts. Our present efforts at sequencing the viral NP protein near its C-terminus should establish the 3' terminus of its mRNA. To avoid similar pitfalls at the 3' end of the G clone, restriction fragments lying just upstream of the poly(dA) tail of pRSC₂₂ were partially extended on viral mRNA by the dideoxysequencing method. The RNA sequence deduced in this manner agreed precisely with the cloned cDNA sequence.

To determine the 5' end of the cloned transcript a 139bp Fnu4HI/DdeI fragment lying within 50 nucleotides of the putative 5' end of the mRNA was asymmetrically labeled at the 5' end of the antimessenger strand and annealed to the viral poly(A) RNA. The DNA primer was then partially extended on the template RNA to obtain the complementary sequence of the 5' end of the RNA. In this manner it was found that the first four nucleotides of the 5' NGGGCAAAU sequence were lacking in this plasmid (Fig. 2). The cloned insert might be a complete copy of G mRNA since the 5'NGGG--- sequence cannot be distinguished from the string of G residues generated by the terminal transferase reaction during cloning. The 5'NGGGCAAAU initiation sequence is conserved in all of the eight viral transcripts examined to date (15,16,23-26). At the 3' end of the transcript, there is a AGUUA-3' sequence three nucleotides upstream of the poly(A) tail. The same sequence is found 1-4 nucleotides upstream of the poly(A) tail in seven of eight RS virus transcripts that we have sequenced (15,16,23-26 and Satake, M *et al* unpublished data). In the transcript encoding the NS₂ protein this sequence is replaced by a AGUAA-3' sequence (26). In this respect, RS virus resembles other negative strand unsegmented genomes (VSV, and Sendai virus) which have conserved 5' and 3' termini (27).

The sequence of the gene in the messenger sense and the deduced amino acid sequence for the longest open reading frame (ORF) encoding a 33K protein are presented in Fig. 3. Between the ATG codon at position 12 from the oligo(dG) tail and the termination codon 6 residues upstream of the poly(dA) was encoded a protein of 32588 daltons containing 298 amino acids. The calculated m.w. was somewhat smaller than the observed 36K. Neither this AUG nor the one five residues upstream occurred in the canonical eucaryotic translation initiation sequence PXXAUGG (41). Except for a small stretch of 41 residues in the middle of the transcript the other two reading frames were extensively blocked. The 33K protein is highly basic (35 vs. 15 acidic residues) and extremely rich in hydroxyamino acids (33 ser. and 58 thr.). Both the basic and hydroxy aminoacids were found predominantly in the C-terminal half of the molecule. There were moderate number of proline residues (30 in all) again mostly in the C-terminal portion. The protein is not overtly hydrophobic and there is no apparent signal sequence (42) at either the N or the C-terminus. In fact the only region of extreme hydrophobicity was internal; between 41 and 63 residues from the N-terminus (Figs. 3 and 4). Seven N-glycosylation sites of the type Asn-X-Thr/Ser were present on the C-terminal side of this hydrophobic domain.

Fig. 4. Hydropathy diagram of RS virus G protein. The relative hydrophilicity and hydrophobicity values were calculated for strings of seven residues. The average for most proteins is indicated by the central line.

33K protein is the unglycosylated precursor to the viral envelope glycoprotein (G):

RS virus has been shown by us and other (1/-21) to possess two integral envelope glycoproteins of 84-90 and 68K. Of these, the former is extensively glycosylated and apparently undergoes processing as reflected by smaller size of the virus associated counterpart (18 and Venkatesan, unpublished observation). Our initial efforts at purifying either of these two glycoproteins were unsuccessful due to poor specificity of polyclonal antisera. We therefore utilized pools of monoclonal antibodies with high virus neutralization titers. Most of these monoclonal antibodies immunoprecipitated a single protein of 88K from infected cells individually labeled with (^3H)-leucine, isoleucine, phenylalanine or (^{35}S)-methionine. The 90K protein was further purified by SDS-polyacrylamide gel electrophoresis prior to direct amino acid microsequencing. Since initial direct sequencing attempts were unsuccessful, probably due to blocked N-terminus, the purified protein was cleaved by CNBr and the peptides resolved by gel filtration. Samples labeled with different amino acids were pooled and a large fragment selected as a candidate for the 214 residue middle fragment deduced from the 33K protein. N-terminal sequencing was carried out for 20 cycles and the radioactive PTH-amino acids were identified by HPLC. Technical considerations do not permit precise calls beyond 40 cycles in this method and the generally poor labeling of the G protein did not allow unambiguous assignments beyond 20 residues. The results are shown in Fig. 5. The cycles at which leucine, isoleucine and phenylalanine were detected precisely coincided with their expected positions from the methionine at 48th residue in the 33K protein. The random probability of finding these residues at their respective positions

Fig. 5. Automated N-terminal radiosequence analysis of the 214 residue CNBr fragment of RS virus G protein. The relevant deduced sequence starting after the Met residue at position 48 is written at the bottom (the Ile at position 49 would then correspond to residue #1 in the CNBr fragment). Proteins individually labeled with (^3H)-Ile, Leu and Phe were mixed together prior to CNBr cleavage. Radioactive phenylthiohydantoin derivatives extracted at each cycle were separated by HPLC runs with standard markers. The respective amino acids are identified by single letter codes.

by chance alone within any other protein was calculated to be 8.4×10^{-13} , far smaller than the natural complexity of eukaryotic coding sequences.

RS virus G protein is both N and O-glycosylated:

The experiments described so far validated our observations that the viral G protein was heavily glycosylated since the native m.w. of 90K would imply 60% contribution from carbohydrate residues. To determine the chemical nature of

Fig. 6. Immunoprecipitation of G protein by monoclonal antibodies from infected cells in the presence and absence of glycosylation inhibitors. Cell monolayers were infected at 1 pfu/cell and at the end of 24 hr the appropriate inhibitors were added. One hour later, the cells were metabolically labeled for 12 hr with (5,6- ^3H) glucosamine and cell lysates prepared for radioimmunoprecipitation using anti G monoclonal antibodies and *S. aureus* protein (A) sepharose beads. The immunoprecipitates were electrophoresed on 12.5% acrylamide gels in SDS and the radioactive bands were visualized by fluorography. Results in the absence of inhibitors (lane 1), in the presence of 0.5 $\mu\text{g/ml}$ of tunicamycin (lane 2), 0.05 μM of monensin (lane 3), 0.2 μM of monensin (lane 4) or in the presence of both tunicamycin (0.5 $\mu\text{g/ml}$) and monensin (0.2 μM) are shown. ^{14}C labeled protein molecular mass markers from Amersham Corp. were run alongside and the respective masses are indicated in kilodaltons (k) on the left.

Fig. 7. Indirect immunofluorescence of live infected cells using anti-RS virus F and G protein monoclonal antibodies with fluorescein conjugated goat anti-mouse IgG antibodies. Panels UI, C, T and M refer to uninfected cells, infected cells, infected cells treated with tunicamycin or monensin respectively.

glycosidic linkages in the protein, we studied the effect of glycosylation inhibitors on the synthesis and migration of G protein in infected cells by immunoprecipitation of G in infected cells and live cell immunofluorescence using anti-G monoclonal antibodies. In the presence of tunicamycin, an inhibitor of the first step in N-glycosylation pathway (43), there was a slight decrease in the size of the G protein (Fig. 6, lane 2). In the presence of the ionophore, monensin which is known to affect electrochemical gradients in the Golgi complex and secondarily influence the O-glycosylation pathways, there was a dramatic loss of the mature G protein and appearance of a diffuse band of ca. 50K sometimes resolvable as a doublet at low concentration of the inhibitor (lanes 3 and 4). With both inhibitors, only diffuse background staining was evident. We tried to correlate these observations with the effect on the egress of G protein to the cell surface. Tunicamycin treatment prevented surface expression of F protein but had no effect on the transport of G protein (Fig. 6, panel T). Monensin, even at high concentrations had minimal effects on the transport of either glycoprotein (panel M). Although, the experiment described in Fig. 7 utilized monoclonal antibodies, we observed essentially similar results for the F protein using polyclonal rabbit anti RS virus hyperimmune serum. The polyclonal serum, however, had very poor anti-G titers to be useful. However, a similar pattern of live cell immunofluorescence was observed in monensin treated cells using a mixture of different G specific monoclonal antibodies. The above observations are very similar to the recently published report of Gruber and Levine (44) thus suggesting a role for glycosylation in the vectorial transport of these glycoproteins.

DISCUSSION

Several recent studies have elucidated certain fundamental differences between RS virus and paramyxoviruses. They can be briefly summarized as the following: 1) unlike the paramyxoviruses which have six transcripts, RS virus encodes at least ten genes that are individually transcribed in a polar fashion, 2) two viral nonstructural protein genes rather than the nucleocapsid protein (NP) gene are 3' proximal on the linear genetic map, 3) at least three nonstructural proteins are encoded as distinct genetic units rather than being encoded by overlapping reading frame (e.g. Sendai virus P and C proteins) and 4) the transcribed genes may be separated on the genome by complex intercistronic sequences instead of the usual di or trinucleotide boundaries (13-15, 26,28,29). In spite of this notable evolutionary divergence, there are many similarities among these viruses in the organization of structural

proteins (15, 16, 23-25). This is especially evident in the case of the fusion protein whose general features are conserved among all these viruses and closely related to their distant cousin, namely influenza virus HA which has a similar function (15,16). However, this analogy does not extend to the other envelope glycoprotein (G). Although it was expected to be the functional equivalent of paramyxovirus attachment protein HN (albeit lacking hemagglutinin and neuraminidase activities), its structure had certain unusual features.

The primary structure of the G protein was largely deduced from cDNA cloning and sequencing. A unique recombinant plasmid identified by several subtractive screening procedures selected a viral mRNA that was translated into a 36K protein far smaller than the usual unglycosylated paramyxovirus HN proteins. The native glycosylated G protein was of variable size (84-90K) probably reflecting different degrees of glycosylation. The mRNA of 1000 nucleotides was adequate to encode a 36K protein. This mRNA had conserved 5' and 3' termini seen in all other RS viral transcripts. The cloned viral sequence had a single long ORF occupying almost the entire length of the message and encoding a protein of about 33K. As no protein of this size has been visualized either in the infected cells or extracellular virus (although recently Gruber and Levine (44) have demonstrated a 33K precursor of G protein in infected cells under pulse-chase conditions by enzymatic deglycosylation), positive identification required protein sequencing. This was far more definitive than immunoprecipitation of the sole 36K translation product of mRNA hybrid selected by this plasmid. Immunoprecipitation of translation products of total viral mRNA yielded ambiguous results since the G mRNA was inefficiently translated under these conditions and the prominent 36K phosphoprotein frequently overshadowed the G protein during gel electrophoresis. Also since the observed m.w. of 36K for the P protein was far above the actual 27.15K protein predicted from the mRNA sequence (24), we felt it unwise to rely on immunoprecipitation of translated G protein. Immune-affinity purified G protein was not amenable to direct amino acid sequencing probably due to blocked N-terminus. However, limited amino acid sequence of a large internal CNBr fragment agreed precisely with the deduced amino acid sequence in the 33K protein.

The G protein of RS virus is an integral membrane protein that is readily solubilized by detergent or trypsin treatment, and therefore would be expected to possess cardinal features of viral envelope glycoproteins such as a N-terminal signal sequence and N- or C-terminal hydrophobic anchorage domains

(1,2,15,16,45). The deduced sequence of 298 amino acids, however, is virtually devoid of these features. In fact, the only region of marked hydrophobicity is between residues 41 and 63 from the N-terminus (Fig. 4). This region might serve as the membrane anchorage site leaving the N-terminal 20 amino acids as the cytoplasmic tail. It is also significant that all the N-glycosylation sites occurred in the C-terminal half which forms the extracellular domain. This possible N-terminally oriented membrane attachment is somewhat analogous to what has been proposed for the HN of paramyxovirus SV5 (46) and to the influenza neuraminidase (NA) wherein an uncleaved hydrophobic sequence near the N-terminus serves both in translocation and membrane attachment (47). The presence of the hydrophilic sequence at the N-terminus of the G protein can then be thought of as a variation in the common theme of N-terminal orientation of all these glycoproteins. Since the identity of the N-terminus of the mature protein could not be determined, it may be enquired whether a portion of the hydrophilic N-terminus is cleaved during maturation. Although there is precedence for proteolytic processing of a precursor in the case Newcastle disease virus (NDV) HN (48-50), similar mechanisms are unlikely with RS virus in light of the recent report showing that under pulse-chase conditions of labeling, nascent G protein is readily converted to 33K protein following endoglycosidase H treatment (44).

For a glycoprotein composed of approximately 60% carbohydrates, there were only seven putative N-glycosylation sites but three of these are the Asn-Pro-Ser/Thr type, seldom utilized in all the glycoproteins studied so far (51,52). If all the remaining sites were fully glycosylated an effective contribution of 10-15K. might be expected from N-linked carbohydrates. Therefore additional increments in mass would have to be through O-glycosylation. This is consistent with the presence of extraordinary number of hydroxy-amino acids. Both the glycosylation processes may be presumed to occur following membrane insertion and vectorial migration in a manner analogous to the "needle pulling a string" model envisioned for the SV₅ virus HN protein (46). Since core N-glycosylation occurs in the rough endoplasmic reticulum (51), inhibition of subsequent O-glycosylation in the Golgi complex would be expected to yield a molecule of ca. 45K. This was borne out by the appearance of a broad band of 50K in the presence of monensin. If N-glycosylation were totally eliminated by tunicamycin treatment (43), the protein would still retain the option to be O-glycosylated in the Golgi apparatus. In either case, vectorial presentation at the cell surface was not drastically affected (Fig. 7).

It was of interest to compare the primary structure of this protein with other known eucaryotic or viral proteins. Computer assisted comparisons did not detect any significant homologies. Visual inspection of the sequence revealed a generalized similarity to chicken ovalbumin, antithrombin III, and α_1 -antitrypsin (53,54). Although glycoproteins containing both N- and O-glycosylations have been found in many virus systems (55-60) paramyxoviruses are not one of them. Except for the reported alteration of glycosylation pattern in the HA(-) vaccinia virus mutants (60), there has been no definite correlation of the type and degree of glycosylation with either antigenicity or other readily detectable viral functions. Now that a complete cDNA copy of RS virus G protein gene is available, such structure/function correlations could be addressed by site directed mutagenesis of cloned gene and subsequent functional expression in eucaryotic vector systems (61,62).

ACKNOWLEDGEMENTS

The authors thank Dr. R.M. Chanock of the Laboratory of Infectious Diseases, National Institute of Allergy and Infectious Diseases for his support. Technical assistance of E. Camargo in tissue culture work, virus purification and mRNA isolation is acknowledged. I. Cecil & G. Harbaugh were helpful in preparing this manuscript. After this manuscript was submitted, G. Wertz et al (Proc. Natl. Acad. Sci., 82, 4075-4079, 1985) have published an identical amino acid sequence of G protein deduced from DNA sequence of three full length cDNA clones.

*To whom correspondence should be addressed at LMM/NIAID, Bldg. 550, Frederick Cancer Research Facility, Frederick, MD 21701, USA

REFERENCES

1. Kingsbury, D.W. (1977) in *Molecular Biology of Animal Viruses*, ed., Nayak, D.P. (Marcel Dekker, New York) pp. 367-398.
2. Lamb, R.A., & Choppin, P.W. (1983) *Ann. Rev. Biochem.*, **52**, 467-506.
3. Scheid, A., Caligiuri, L.A., Compans, R.W. & Choppin, P.W. (1972) *Virology* **50**, 640-652.
4. Scheid, A. & Choppin, P.W. (1976) *Virology* **57**, 475-490.
5. Choppin, P.W. & Compans, R.W. (1975) in *Comprehensive Virology*, eds., Fraenkel-Conrat, H. & Wagner, R.R. (Plenum Press, New York) **4**, 95-178.
6. Scheid, A. & Choppin, P.W. (1977) *Virology* **80**, 54-66.
7. Merz, D.C., Scheid, A. & Choppin, P.W. (1980) *J. Exp. Med.* **151**, 275-288.
8. Orvell, C. & Norrby, E. (1977) *J. Immunol.* **119**, 1882-1887.
9. Merz, D.C., Scheid, A. & Choppin, P.W. (1981) *Virology* **109**, 94-105.
10. Chanock, R.M., Kim, H.W., Brandt, C.D., & Parrott, R.H. (1982) in *Viral Infections in Humans, Epidemiology & Control*, ed., Evans, A.S. (Plenum Press, New York) pp. 471-489.
11. Bloth, B. and Norrby, E. (1967) *Arch. Ges. Virusforsch.* **21**, 71-77.
12. Richman, A.V., Pedreira, F.A. and Taurago, N.M. (1971) *Appl. Microbiol.* **21**, 1099-1100.

13. Collins, P.L., & Wertz, G.W. (1983) *Proc. Natl. Acad. Sci. USA* 80, 3208-3212.
14. Collins, P.L., Huang, Y.T. & Wertz, G.W. (1984) *J. Virol.* 49, 572-578.
15. Collins, P.W., Huang, Y.T. & Wertz, G.W. (1984) *Proc. Natl. Acad. Sci. USA* 81, 7683-7687.
16. Elango, N., Satake, M., Coligan, J.E., Norrby, E., Camargo, E. & Venkatesan, S. (1985) *Nucleic Acids Res.* 13, 1559-1574.
17. Peebles, M. & Levine, S. (1979) *Virology* 95, 137-145.
18. Bernstein, J.M. & Hruska, J.F. (1981) *J. Virol.* 38, 278-285.
19. Dubovi, E.J. (1982) *J. Virol.* 42, 372-378.
20. Gruber, C. & Levine, S. (1983) *J. Gen. Virol.* 64, 825-832.
21. Venkatesan, S., Elango, N. & Chanock, R.M. (1983) *Proc. Natl. Acad. Sci. USA* 80, 1280-1284.
22. Hardwick, J.M. & Busse, R.H. (1978) *J. Virol.* 25, 687-692.
23. Elango, N. & Venkatesan, S. (1983) *Nucleic Acids Res.* 11, 5941-5951.
24. Satake, M. & Venkatesan, S. (1984) *J. Virol.* 50, 92-99.
25. Satake, M., Elango, N. & Venkatesan, S. (1984) *J. Virol.* 52, 991-994.
26. Elango, N., Satake, M. & Venkatesan, S. (1985) *J. Virol.* 55, 101-110.
27. Gupta, K. and Kingsbury, D.W. (1984) *Nucleic Acids Res.* 12, 3829-3840.
28. Collins, P.L. & Wertz, G.W. (1985) *Virology* 141, 283-291.
29. Collins, P.L. & Wertz, G.W. (1985) *J. Virol.* 54, 65-71.
30. Maniatis, T., Fritsch, E.F. & Sambrook, J. (1982) *Molecular Cloning: A Laboratory Manual*. Cold Spring Harbor Press, New York.
31. Maxam, A.M. & Gilbert, W. (1980) *Methods in Enzymol.* 65, 499-560.
32. Queen, C.L. & Korn, L.J. (1980) *Methods in Enzymol.* 65, 595-609.
33. Kyte, J. & Doolittle, R.F. (1982) *J. Mol. Bio.* 157, 105-132.
34. Wilbur, W.J. & Lipman, D.J. (1983) *Proc. Natl. Acad. Sci. USA* 80, 726-730.
35. Lee, G.T.-Y., Para, M.F. & Spear, P.G. (1982) *J. Virol.* 43, 41-49.
36. Yang, Y.W.-H., Rechler, M., Nissley, S.P. & Coligan, J.E. (1985) *J. Biol. Chem.* 260, 2578-2582.
37. Ewenstein, B.M., Nisizawa, T., Uehara, H., Nathenson, S.G., Coligan, J.E. & Kindt, T.J. (1978) *Proc. Natl. Acad. Sci. USA* 75, 2909-2913.
38. Coligan, J.E., Gates, F.T.III, Kimball, E.S. & Maloy, W.L. (1983) *Methods in Enzymol.*, 91, 413-424.
39. Satake, M. & Luftig, R.B. (1983) *Virology* 124, 259-272.
40. Venkatesan, S., Elango, N., Satake, M., Camargo, E. & Chanock, R.M. (1984) in *Modern Approaches to Vaccines: Molecular and chemical basis of virus virulence and immunogenicity*, eds., Chanock, R.M. & Lerner, R.A. (Cold Spring Harbor Press, New York) pp. 37-44.
41. Kozak, M. (1984) *Nucleic Acids Res.* 12, 857-782.
42. Blobel, G. (1980) *Proc. Natl. Acad. Sci. USA* 77, 1496-1500.
43. Takatsuki, A., Arima, K. & Tamura, G. (1971) *J. Antibiot.* 24, 215-223.
44. Gruber, C. & Levine, M. (1985) *J. Gen. Virol.* 66, 417-432.
45. Paterson, R.G., Harris, T.J.R. & Lamb, R.A. (1984) *Proc. Natl. Acad. Sci. USA* 81, 6706-6710.
46. Hiebert, S.W., Paterson, R.G. & Lamb, R.A. (1985) *J. Virol.* 54, 1-6.
47. Markoff, L.J. & Lai, C.-J. (1982) *Virology* 119, 288-297.
48. Morrison, T.G. & Simpson, D. (1980) *J. Virol.* 36, 390-400.
49. Garten, W., Kohama, T. & Klenk, H.-D. (1980) *J. Gen. Virol.* 51, 207-211.
50. Schuy, W., Garten, W., Linder, D. & Klenk, H.-D. (1984) *Virus Res.* 1, 415-426.
51. Hart, G.W., Brew, K., Grant, G.A., Bradshaw, R.A. & Lennarz, W.J. (1979) *J. Biol. Chem.* 254, 9747-9753.
52. Ronin, C., Bouchilloux, S., Granier, C. & Van Rietschoten, J. (1978) *FEBS Letters* 96, 179-182.
53. Nisbet, A.D., Saundry, R.H., Moir, A.J.G., Fothergill, L.A. and

- Fothergill, J.E. (1981) Eur. J. Biochem. 115, 335-345.
54. Hunt, L.T. Dayhoff, M.O. (1980) Biochem. Biophys. Res. Commun. 95, 864-871.
55. Niemann, H., Boschek, B., Evans, D., Rosing, M., Tamura, T. & Klenk, H.-D. (1982) EMBO Journal 1, 1499-1504.
56. Johnson, D.C. & Spear, P.G. (1982) J. Virol. 43, 1102-1112.
57. Johnson, D.C. & Spear, P.G. (1983) Cell 32, 987-997.
58. Holmes, K.V., Doller, E.W. & Sturman, L.S. (1981) virology 115, 334-344.
59. Wenske, E.A. & Courtney, R.J. (1983) J. Virol. 46, 297-301.
60. Shida, H. & Dales, S. (1981) Virology 111, 56-72.
61. Sveda, M.M., Markoff, L.J. & Lai, C.-J. (1982) Cell 30, 649-656.
62. Markoff, L.J., Lin, B.-C., Sveda, M.M. & Lai, C.-J. (1984) Mol. Cell. Biol. 4, 8-16.