The Fusion Glycoproteins of Human Respiratory Syncytial Virus of Subgroups A and B: Sequence Conservation Provides a Structural Basis for Antigenic Relatedness

By PHILIP R. JOHNSON1,2 AND PETER L. COLLINS1 *

1Laboratory of Infectious Diseases, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda Maryland 20892 and 2Department of Pediatrics, Vanderbilt University School of Medicine, Nashville, Tennessee 27232, U.S.A.

(Accepted 30 June 1988)

SUMMARY

Two major antigenic subgroups (A and B) have been described for human respiratory syncytial virus. The complete nucleotide sequence was determined for the fusion (F) mRNA of the subgroup B strain 18537 and the amino acid sequence of the F protein deduced, for comparison with the previously described sequences for the A2 strain of subgroup A. The F proteins (excluding the cleaved signal peptide) were 91% identical between the subgroups, consistent with the previously described high degree of antigenic relatedness. The greatest divergence occurred within the F2 subunit immediately preceding the cleavage activation site.

Human respiratory syncytial virus (RSV) is an important cause of severe respiratory tract illness in young children world-wide (McIntosh & Chanock, 1985). RSV is an enveloped paramyxovirus (genus Pneumovirus) containing a genome of single-stranded negative sense RNA of approximately 15 000 nucleotides (Collins et al., 1986). The host cell-derived envelope contains at least two virus-encoded structural glycoproteins, the attachment (G) protein and the fusion (F) protein. The G protein has a number of unusual structural features, including a large number of proline, serine and threonine residues and a significant content of O-linked oligosaccharides (Wertz et al., 1985; Johnson et al., 1987b). The F protein is synthesized as a precursor (F0) that is cleaved into disulphide-linked polypeptides (F1 and F2) (Collins et al., 1984; Elango et al., 1985; Norrby et al., 1986; Olmsted et al., 1986). By analogy to Sendai virus (Scheid & Choppin, 1974), this cleavage is a prerequisite for fusion activity and is carried out by a cellular protease(s). Both of these integral membrane proteins have been shown to be major viral neutralization antigens, although studies in cotton rats indicated that the F protein has a greater role than the G protein in host immunity (Olmsted et al., 1986; Johnson et al., 1987a).

Infants and children are commonly reinfected by RSV in successive yearly seasons, and the possibility exists that RSV antigenic variation may play a role in this phenomenon. Two distinct antigenic subgroups (A and B) of RSV have been recognized based on in vitro neutralization assays and reactions with monoclonal and polyclonal antibodies (Coates et al., 1966; Anderson et al., 1985; Mufson et al., 1985; Gimenez et al., 1986; Johnson et al., 1987a). Analysis by glycoprotein-specific ELISA and by cross-protection studies using RSV strain A2–vaccinia virus recombinants (Johnson et al., 1987a) showed that the G proteins of the two subgroups are antigenically quite distinct, whereas the F proteins are more closely related, although epitopic variation does occur (Anderson et al., 1985; Mufson et al., 1985).

To delineate the molecular basis for variation between subgroups, we undertook cDNA cloning and nucleotide sequencing of an RSV strain (18537) of subgroup B for comparison with the published data for strain A2 (subgroup A). We previously reported (Johnson et al., 1987b) that the G proteins of subgroup A and B have only 53% identical amino acids, but contain a highly conserved cysteine-rich region that is presumably important for structural and functional
Fig. 1. Alignment of the nucleotide and deduced amino acid sequences of the 18537 (subgroup B) and A2 (subgroup A) F mRNAs and proteins. The alignments were performed using the NUCALN and PRTALN programs of Wilbur & Lipman (1983). For the A2 sequence, only nucleotides and amino acids that differ from the 18537 sequences are shown. Gaps introduced during alignment of the sequences are boxed; the additional site in the 18537 sequence is in a shaded box. Cysteine residues are indicated by solid triangles over the sequence; a single additional cysteine residue in the A2 signal sequence is indicated by an open triangle under the sequence. The hydrophobic domains of the FI subunit are shown by solid triangles over the sequence; a single additional cysteine residue in the A2 signal sequence is indicated by an open triangle under the sequence. The hydrophobic domains of the FI subunit are...
**Table 1. Summary of nucleotide and amino acid identities between the A2 and 18537 F proteins**

<table>
<thead>
<tr>
<th>18537 F domains</th>
<th>Nucleotide</th>
<th>Amino acid</th>
</tr>
</thead>
<tbody>
<tr>
<td>Signal peptide*</td>
<td>61</td>
<td>36</td>
</tr>
<tr>
<td>F2</td>
<td>81</td>
<td>83</td>
</tr>
<tr>
<td>F1</td>
<td>83</td>
<td>93</td>
</tr>
<tr>
<td>3' Non-coding</td>
<td>47</td>
<td>–</td>
</tr>
<tr>
<td>Overall</td>
<td>79</td>
<td>89 (91)†</td>
</tr>
</tbody>
</table>

* Here, the signal peptide is defined as amino acids 1 to 22, but it has not been mapped by direct methods.
† Ninety-one% without the signal peptide (F1 + F2).

**integrity of the G protein. In this communication, we present the nucleotide and predicted amino acid sequences for the F protein of RSV strain 18537 and demonstrate that the F proteins are highly conserved between subgroups A and B.**

A library of cDNA clones derived from RSV-18537-infected cell mRNAs was established as described previously (Johnson *et al.*, 1987b). Clones containing F mRNA nucleotide sequences were identified by two independent approaches: (i) heterologous hybridization with a cDNA probe representing the strain A2 (subgroup A) F mRNA and (ii) direct sequence analysis of cDNAs that had been identified as 18537-specific by differential hybridization with radiolabelled cDNAs of mRNAs from uninfected and 18537-infected cells. Dideoxynucleotide sequencing of recombinant plasmids and RSV mRNA using synthetic DNA primers was performed as previously described (Spriggs *et al.*, 1986; Johnson *et al.*, 1987b).

The sequence of the 18537 F mRNA (Fig. 1) was determined from five overlapping cDNA clones: pG27 (representing nucleotides 36 to 1986 of the complete mRNA sequence), pY58 (nucleotides 299 to 1986), pT39 (nucleotides 840 to 1986), pN66 (nucleotides 14 to 204) and pM32 (polycistronic clone starting at nucleotide 1085 in the F mRNA, spanning the F protein–22K intergenic region and ending at nucleotide 424 in the 22K mRNA). Three cDNA inserts (pG27, pY58 and pT39) were coterminal at one end and were followed by poly(A) tracts of various lengths, thereby identifying the 3' end of the 18537 F mRNA. The exact 5' end of the mRNA was mapped by primer extension with a synthetic oligonucleotide (complementary to the mRNA from nucleotides 101 to 124) under conditions for dideoxynucleotide sequencing (data not shown); nucleotide identities at positions 1, 2, 5 and 6 could not be determined unambiguously because of strong stop bands in each lane.

The nucleotide sequence of the strain 18537 F mRNA had 79% overall nucleotide identity with the previously published (Collins *et al.*, 1984; Elango *et al.*, 1985) strain A2 F mRNA sequence (Table 1 and Fig. 1). The overall lengths were nearly identical (1896 and 1899 nucleotides); the 3' non-translated regions of the 18537 F mRNA contained three areas of nucleotide insertion or deletion and was three nucleotides shorter overall than the corresponding A2 sequence. Regions of the F mRNAs that coded for protein were more highly conserved than were the long 3' non-translated regions (82% compared to 47% identity, respectively; Table 1 and Fig. 1). This is consistent with the observation that non-translated sequences, including intergenic regions, were relatively non-conserved (generally less than 50%) between viruses of the two RSV subgroups analysed thus far except for conserved gene-start and gene-end sequences (Johnson *et al.*, 1987b; P. R. Johnson *et al.*, unpublished data).

The deduced amino acid sequences of the 18537 and A2 F proteins were identical in length (574 amino acids) and had extensive amino acid identity (89% overall; Table 1 and Fig. 1). The strain 18537 F1 subunit had 93% amino acid identity with the strain A2 sequence, while the F2 subunit identity was somewhat lower at 83% (Table 1 and Fig. 1). The high degree of amino acid identity throughout the entirety of the F1 subunit was somewhat unexpected because the electrophoretic mobility of F1 differs significantly between the two subgroups (Norrby *et al.*, 1986; P. R. Johnson & P. L. Collins, unpublished observations). We presume that the few differences that do exist have an indirect, disproportionate effect on electrophoretic mobility, perhaps by altering protein conformation or detergent-binding properties.
All 15 cysteine residues in the mature F protein (lacking the signal peptide) were conserved, consistent with the extensive conservation of cysteine residues in general among the paramyxovirus F proteins (Spriggs et al., 1986). The five potential sites for \(N\)-linked glycosylation in the strain A2 F protein were totally conserved in the 18537 sequence; additionally, a sixth site occurred in the strain 18537 F sequence (residues 120 to 122 in F2). However, this sixth acceptor site also exists for the subgroup A strain RSS-2 (Baybutt & Pringle, 1987), indicating that variability at this site exists within as well as between subgroups. The F2 subunits of strains A2 and 18537 were almost identical in electrophoretic mobility (Norrby et al., 1986; P. R. Johnson & P. L. Collins, unpublished results), and therefore it is possible that only four of the five potential acceptor sites in strain 18537 F2 are utilized. The new potential glycosylation site was in a region of relatively high sequence divergence in F2 (in residues 111 to...
There are 10 mismatches) that immediately precedes the protease cleavage region (residues 130 to 136). This region is highly hydrophilic and is predicted to have a prominent secondary structure composed of a coil and turn (Fig. 2). Considering the new potential glycosylation site and the possible surface localization of this region, one possibility is that this area forms part of a strain-specific epitope that has been observed in reactions with monoclonal antibodies (Anderson et al., 1985; Mufson et al., 1985). To ascertain whether this region is of biological significance (e.g. a neutralization epitope) will require thorough analysis of monoclonal antibody escape mutants. Also, it will now be possible to use site-directed mutagenesis and restriction fragment substitution to construct F protein mutants and strain A2–18537 chimeras to characterize further monoclonal antibody reactivity.

The typical paramyxovirus F protein contains three prominent hydrophobic domains: (i) the N-terminal signal sequence, (ii) the F1 N terminus, which is thought to have a direct role in membrane fusion and (iii) the membrane anchor located near the F1 C terminus (Collins et al., 1984; Elango et al., 1985; Spriggs et al., 1986). The signal peptide, which would be predicted to be cleaved in the region of residues 21 to 25 (von Heijne, 1984), exhibited extensive sequence divergence between the subgroups (36% identity) although its hydrophobic nature was maintained. This level of sequence divergence was similar to that observed for two highly divergent regions (amino acids 71 to 163, 39% identity and amino acids 176 to 298, 44% identity) in the ectodomain of the G protein (Johnson et al., 1987b). The signal peptide presumably could accept extensive amino acid substitution because (i) it is cleaved cotranslationally and would not contribute to the structure of the mature protein, and (ii) signal function depends on features of hydrophobicity and local secondary structure rather than on a specific sequence. In contrast, the N terminus of the F1 subunit contained only a single, conservative substitution (Ile to Val at position 16 from the F1 N terminus), consistent with the high sequence conservation of this region among paramyxoviruses in general and its presumed importance in membrane fusion. Finally, the extensive sequence conservation of the membrane anchor suggested that the specific sequence of this domain, and not just its hydrophobic nature alone, is important to F protein structure.

The high degree of conservation of the F proteins of the two subgroups contrasted with the previously described extensive divergence of the extracellular domain of the G protein. The extensive sequence divergence of the F signal peptide and the G ectodomain illustrates that the potential has existed for extensive amino acid sequence divergence between the subgroups. Therefore, the high degree of conservation of the F protein is presumably due to strict functional or structural constraints on amino acid substitutions. The high degree of sequence identity between the F proteins of the two RSV subgroups is consistent with the finding that these proteins share a high degree of antigenic relatedness. In contrast, the highly divergent G protein exhibited a lower degree of antigenic relatedness (Johnson et al., 1987a).

Recently, Trudel et al. (1987) showed that a neutralizing epitope of the F protein of RSV strain Long (subgroup A) was contained, at least in part, in a 21 amino acid region (212 Cys to 232 Glu) of the F1 subunit. This region in the strain 18357 F protein was 86% identical (18 of 21 amino acids) to the strain A2 sequence and contained a central core of 12 exactly conserved residues (214 Ile to 225 Gin) flanked by residues exhibiting non-conservative substitutions. Whether this region comprises part of a cross-subgroup neutralization epitope will await more complete data on heterologous neutralization, the sequence analysis of neutralizing monoclonal antibody escape mutants, and the analysis of monoclonal antibody binding to mutant F proteins expressed from modified cDNAs.

The F protein is the major viral neutralization antigen, with the G protein having a secondary role (Olmsted et al., 1986; Johnson et al., 1987a). Additionally, the former, but not the latter, is structurally and antigenically highly conserved between the antigenic subgroups. These findings demonstrate that the F protein is the more important protective antigen and thus would be the most important component of a recombinant DNA or subunit vaccine for RSV.

This work was performed in the laboratories of Drs Brian R. Murphy and Robert M. Chanock. We thank them for their advice and interest.
REFERENCES


(Received 14 March 1988)