Sequence Analysis of the P and C Protein Genes of Human Parainfluenza Virus Type 3: Patterns of Amino Acid Sequence Homology Among Paramyxovirus Proteins

By MELANIE K. SPRIGGS AND PETER L. COLLINS*

Laboratory of Infectious Diseases, Building 7, Room 100, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, Maryland 20892, U.S.A.

(Accepted 28 August 1986)

SUMMARY

The complete nucleotide sequence of the P + C mRNA of human parainfluenza virus type 3 (PF3) was determined by sequencing cDNA, viral genomic RNA and mRNA. The P + C mRNA is 2009 nucleotides in length, exclusive of poly(A), and contains two overlapping open reading frames (ORFs). The P + C mRNA encodes two proteins, the 602 amino acid nucleocapsid phosphoprotein P and the 199 amino acid non-structural protein C. Peptide mapping confirmed that the two proteins are unrelated. Hybrid-arrest translation experiments assigned each of the two proteins to its respective ORF. These studies showed that the coding strategy of the PF3 P + C mRNA is similar to that of Sendai virus. Amino acid sequence alignment showed that the P and C proteins of PF3 and Sendai virus represent homologous pairs. However, these homologies are represented by high contents of accepted amino acid substitutions and by similarity in hydropathy profiles rather than by high contents of exact amino acid matches. Homology with the P and C proteins of measles, canine distemper and respiratory syncytial viruses was at the threshold of significance. The patterns of amino acid sequence homology among the paramyxovirus HN, F, NP, P and C proteins are compared.

INTRODUCTION

Human parainfluenza virus type 3 (PF3) is second in importance only to respiratory syncytial virus (RSV) as a major aetiological agent of respiratory tract infections in infants (Chanock & McIntosh, 1985). The genome (vRNA) of PF3 is a single strand of negative-sense RNA of approximately 15000 nucleotides (Storey et al., 1984). Six major structural proteins have been described for PF3: a nucleocapsid (NP) protein, mol. wt. 66000 to 68000 (66K to 68K); a large nucleocapsid-associated protein (L), mol. wt. 185K to 250K; a nucleocapsid phosphoprotein (P), mol. wt. 83K to 90K; a matrix protein (M), mol. wt. 35K to 40K; a surface glycoprotein responsible for cell fusion (F), mol. wt. 60K to 63K; a second surface glycoprotein, the haemagglutinin–neuraminidase protein (HN), mol. wt. 69K to 72K (Storey et al., 1984; Jambou et al., 1985; Sanchez & Banerjee, 1985a; Wechsler et al., 1985a, b).

The PF3 P protein, like the P proteins of other paramyxoviruses, is phosphorylated and nucleocapsid-associated. The P proteins of Sendai virus (SV) and Newcastle disease virus (NDV) occur in nucleocapsids in both monomers and disulphide-linked trimers (Smith & Hightower, 1981; Hightower et al., 1984; Deshpande & Portner, 1985, and references therein), although similar observations have not been described for PF3. For NDV and SV, nucleocapsid reconstitution experiments (Hamaguchi et al., 1983), protease-mapping experiments (Chinchar & Portner, 1981b), and monoclonal antibody-inhibition experiments (Deshpande & Portner, 1985) established that the P protein is a component of the active transcriptase. For SV, protease-mapping experiments indicated that the C-terminal half of the P protein alone is sufficient for transcription (Chinchar & Portner, 1981a; Deshpande & Portner, 1985). Intracellularly and in purified virus, the P protein exists in multiple electrophoretically distinct forms, and it is...
tempting to speculate that the functions of the P protein might be modulated by structural modifications (for example, see Hightower et al., 1984).

In addition to the six structural proteins of PF3, a seventh, non-structural protein (C), mol. wt. 22K, has been detected in extracts of PF3-infected cells (Sanchez & Banerjee, 1985a; Spriggs & Collins, 1986). All paramyxoviruses appear to encode one or two non-structural proteins (see Discussion); to date, no functions have been assigned to any paramyxovirus non-structural protein.

Dideoxynucleotide sequencing of intergenic and flanking gene regions in vRNA established that the PF3 gene order is 3'-NP-P+C-M-F-HN-L-5' (Spriggs & Collins, 1986). These results exclude the existence of additional intervening genes. Complete nucleotide and amino acid sequences have been determined for the F, HN and NP mRNAs and proteins (Spriggs et al., 1986; Elango et al., 1986; Galinski et al., 1986; Jambou et al., 1986). These studies showed that a high degree of homology exists between PF3 and SV with regard to genome map, sequences of the intergenic, gene-start and gene-end regions, and the sequences of the F, HN and NP mRNAs and proteins. The lack of evidence for a separate mRNA coding for the non-structural C protein suggested that the nucleotide sequence encoding this protein existed as an overlapping open reading frame (ORF), most likely in the P mRNA in a manner analogous to that found in SV (Giori et al., 1983) and the two morbilliviruses, measles virus (MV) and canine distemper virus (CDV) (Bellini et al., 1985; Barrett et al., 1985). This possibility was supported by (i) hybrid-arrest translation experiments, which showed that the P and C proteins are encoded at approximately the same genomic locus, and (ii) RNA blot hybridization experiments, which showed that this genetic locus is expressed as a single species of mRNA (Sanchez & Banerjee, 1985b; Spriggs & Collins, 1986). Here we present the nucleotide and predicted amino acid sequences for the P + C mRNA and P and C proteins, demonstrate that the P and C proteins are unrelated, show that the two proteins are encoded by separate, overlapping ORFs in the P + C mRNA, and describe the patterns of amino acid homology among the P, C and other proteins for several paramyxoviruses.

METHODS

Virus and cells. PF3, strain 47885, was plaque-purified three times and propagated in monolayer cultures of HEp-2 cells. Preparation of virus as a source of vRNA was as described by Spriggs et al. (1986).

cDNA clones. Double-stranded cDNAs of intracellular PF3 mRNAs were synthesized and cloned into pBR322 as described previously (Elango et al., 1986). The cDNA library was screened by conventional methods of colony hybridization, reciprocal cross-hybridization, restriction enzyme digestion and agarose gel electrophoresis (Maniatis et al., 1982). The strategy for identifying cDNAs as virus-specific has been described elsewhere (Spriggs & Collins, 1986).

Nucleotide sequence analysis. cDNA clones were sequenced by a combination of the chemical method (Maxam & Gilbert, 1980), and dideoxynucleotide sequencing using synthetic oligonucleotides as primers (Zagursky et al., 1985; Spriggs et al., 1986). Dideoxynucleotide sequencing of vRNA was as described by Spriggs et al. (1986). To confirm the sequence derived from vRNA, synthetic oligonucleotides complementary to sequences in the PF3 P + C mRNA were 5' end-labelled with [γ-32P]ATP by polynucleotide kinase, hybridized to intracellular mRNA and extended with reverse transcriptase in the presence of dideoxynucleotides (Collins et al., 1984a; Spriggs & Collins, 1986).

Oligonucleotides. Oligonucleotides were synthesized on an Applied Biosystems synthesizer. The oligonucleotides used in the hybrid-arrest experiments, numbered according to the distance of the hybridized primer from the 5' mRNA end, were P + C-61, GCACTCCTTTCCAATCAACTCTTATGATTGAG, and P + C-704, CATTTATTTTCTTCTCATTCTCGTGCC. P + C-61 and P + C-704 were purified prior to use by electrophoresis on a 12% polyacrylamide gel containing 5 M-urea. Primers used in sequencing cDNA and vRNA were as follows, identified by gene, polarity and position in the complete 5' to 3' sequence: NP-1603, P + C-1748, P + C-1448, P + C-1206, P + C-1180, P + C-945, P + C-783, P + C-546, P + C-394, P + C-194, P + C-1696, P + C-1396, P + C-1161, P + C-916, P + C-625, P + C-350, P + C-110. Two primers that consisted of PBR322 sequences flanking either side of the PstI site were used to sequence the ends of cDNA inserts; they were pBR+3589, TGGCGAAGTGTGTTG, and pBR-3636, AAGGACAGCCTGAC.

In vitro protein synthesis. mRNA was translated in the presence of [35S]methionine in reticulocyte lysates (Bethesda Research Laboratories) and used according to the supplier's protocols. Hybrid-arrest reactions were performed as described previously (Paterson et al., 1977). Proteins were analysed by gel electrophoresis on 10% SDS–polyacrylamide gels (Laemmli et al., 1970).
Peptide mapping. Limited proteolytic digestions were performed in general by the method of Cleveland et al. (1977) with minor modifications (Collins et al., 1982). Papain and V8 *Staphylococcus aureus* protease were obtained from Boehringer Mannheim Biochemicals.

Protein sequence alignments. Alignments were made using the global search program ALIGN (Dayhoff, 1978). To provide an optimal alignment, the program aligns and scores both matches and mismatches according to empirically determined patterns of amino acid conservation and substitution among protein families in the sequence data base. As shown in Table 1, scores are expressed in standard deviations relative to a mean score of alignments of 100 randomized permutations of the two sequences, which corrects for sequence content. This scoring matrix was designed to detect and quantify distant relationships; scores of 3.0 standard deviations and above are considered indicative of an authentic relationship. The mutational data matrix used in scoring was 250 PAMs, and the gap penalty was 12 except where noted (Dayhoff, 1978).

RESULTS

Nucleotide sequence analysis

RNA blot hybridization showed that the mRNA that encodes the P protein, designated here the P+C mRNA, contains approximately 2200 nucleotides including polyadenylate (Sanchez & Banerjee, 1985b; Spriggs & Collins, 1986). cDNA clones of the P+C mRNA were identified in previous work (Sanchez & Banerjee, 1985b; Spriggs & Collins, 1986). The complete sequence of the P+C mRNA was determined by sequencing cDNAs, vRNA and mRNA by the dideoxynucleotide method using synthetic oligonucleotide primers (Methods). Two overlapping cDNA clones, pP4 (1151 bp excluding homopolymer tails) and pP6 (945 bp), were sequenced in both strands, yielding a sequence of 1437 nucleotides excluding homopolymer tails. pP6 contained, in addition, a terminal poly(dA) tract, suggesting that it represented the 3' end of the mRNA. This presumptive 3'-terminal mRNA sequence and adjoining poly(dA) tail were confirmed in five additional, independently derived cDNA clones (not shown). The remainder of the sequence of the P+C mRNA was determined by sequencing vRNA and mRNA such that the entire sequence was determined in both strands. In all, the 5'-proximal 921 nucleotides of the P+C mRNA were obtained or confirmed from vRNA, proof that this portion of the sequence was authentic and representative of the virus population. The 5' end of the mRNA was mapped and sequenced by primer extension in previous work (Spriggs & Collins, 1986). Finally, the sequences of both ends of the P+C mRNA were confirmed in previous work by sequencing intergenic and flanking gene sequences (Spriggs & Collins, 1986).

Sequence of the P+C mRNA

The complete nucleotide sequence of the P+C mRNA contains 2009 bases exclusive of poly(A) (Fig. 1). The 5' end of the mRNA, 5'-AGGAUUAAAG-3', is consistent with the semi-conserved 10 nucleotide sequence 5'-AGGANNAAAG found at the 5' ends of the NP, M, F, HN and L mRNAs of PF3 (Spriggs & Collins, 1986). The 3' end of the P+C mRNA, 5'... AAAUAAG-poly(A), was identified by analysis of several cDNAs, and sequence analysis of vRNA showed that five residues of the poly(A) tail are represented directly in vRNA (Spriggs & Collins, 1986). The corresponding, complementary sequence in vRNA, 3'... UUUAUU-CUUUUU, is consistent with the semi-conserved gene-end sequence 3'... UU1AUU2-UUUU identified for the NP, F and HN genes of PF3 (Spriggs & Collins, 1986).

The P+C mRNA contains two significant ORFs. The longer ORF initiates at the first, 5'-proximal AUG triplet (nucleotides 80 to 82) and is a candidate to encode a P protein of 602 amino acids with a calculated unmodified mol. wt. of 67 541. This is substantially less than the estimate of 83K to 90K mol. wt. obtained by SDS-PAGE, but the magnitude of the difference is the same as observed in analogous comparisons for the calculated versus the observed molecular weights of the P proteins of SV, MV and CDV (Giorgi et al., 1983; Bellini et al., 1985; Barrett et al., 1985). The second, shorter ORF initiates at the second AUG of the mRNA (nucleotides 90 to 92) and is a candidate to encode a C protein of 199 amino acids with a calculated mol. wt. of 23297, in good agreement with the *M* of 22K obtained by gel electrophoresis. Evidence that the P and C proteins are unrelated and are encoded by the predicted separate ORFs in the P+C mRNA is presented below.
TTG ATG GAA ACC GAT CAG TAT AAA AAC ATG CTG TAA TTA CAG CTT GCT TCT TAC ATG TCT TGG GAA GAA CAA CAG CTA GAG AAA TCA AGT ATG ATC TGG AAC ATG AAC AGG

Fig. 1. Complete 5' to 3' nucleotide sequence of the P-t-C mRNA, exclusive of poly(A). The predicted amino acid sequences for the P and C proteins are shown with the reading frames identified on the right.
Paramyxovirus protein sequence homologies

Comparison by peptide mapping of the P and C proteins

To show that the C protein is a unique polypeptide and not a subset of the larger P protein, the two proteins were translated in vitro and compared by limited-digest peptide mapping (Fig. 2). The resulting oligopeptide patterns for the two proteins were clearly distinct. Significantly, the papain-resistant oligopeptides of the C protein were substantially larger than those of the P protein, eliminating the possibility that the differences in the oligopeptide patterns were due to the smaller size of the C protein. These results established that the C protein is a seventh unique protein encoded by PF3.

Coding assignments within the P+C mRNA

The coding assignments of the two overlapping ORFs in the P+C mRNA were examined by hybrid arrest of translation in vitro using synthetic oligonucleotides. Oligonucleotides were constructed to be complementary to nucleotides 61 to 93 (oligonucleotide P-61) and nucleotides 704 to 736 (oligonucleotide P-704) of the P+C mRNA; hybridization with P-61 would block access to the translational start sites of both ORFs, whereas P-704 would block access to the 3' proximal half of the larger ORF while leaving the smaller ORF unobstructed.
Fig. 3. SDS-PAGE of [35S]methionine-labelled PF3 proteins showing assignment of the P and C proteins to ORFs in the P+C mRNA by hybridization arrest of in vitro translation using specific oligonucleotides. Proteins labelled during a 1 h pulse were extracted from (a) PF3-infected and (b) uninfected HEp-2 cells and compared to proteins synthesized in reticulocyte lysates in response to mRNA from (c, e) PF3-infected and (d) uninfected cells. To identify ORFs by hybridization arrest, additional samples of mRNA from PF3-infected cells were hybridized to oligonucleotides P+C-61 (f) and P+C-704 (g), translated in vitro and the products were electrophoresed in parallel. The NP, P and C proteins are marked. The bands marked 1 and 2 are putative truncated P proteins resulting from premature translational termination due to hybridized oligonucleotide P+C-704. Bands 3 and 4 are putative artefacts of translation in vitro and are described in the text.

Infected cell mRNA was hybridized with P-61 and P-704 individually and analysed by translation in vitro (Fig. 3). As described previously (Spriggs & Collins, 1986), the P and C proteins are readily detected as major species both in extracts of infected cells (Fig. 3a) and among the products of translation in vitro (Fig. 3c, e). Hybridization with P-61 blocked the synthesis of both the P and C proteins (Fig. 3f), confirming that both proteins are encoded by the P+C mRNA. In contrast, hybridization with P-704 (Fig. 3g) blocked the synthesis of the P protein, had no effect on the synthesis of the C protein, and resulted in the appearance of two new, heterodisperse polypeptide species of approximately 25K and 28K (labelled 1 and 2, respectively, in Fig. 3). These results assigned the P protein to the longer ORF and the C protein to the shorter. The additional 25K and 28K species obtained in response to hybridization by P-704 presumably were P-related products generated by premature translational termination due to the hybridized oligonucleotide; analogous truncation products have been described previously in hybrid-arrest experiments (Paterson et al., 1984). The experiment shown in Fig. 3
Table 1. Relationships among the amino acid sequences of paramyxovirus P and C proteins*

<table>
<thead>
<tr>
<th></th>
<th>PF3 P versus SV</th>
<th>Identity (%)†</th>
<th>S.D.‡</th>
<th>PF3 C versus SV</th>
<th>Identity (%)†</th>
<th>S.D.‡</th>
</tr>
</thead>
<tbody>
<tr>
<td>SV</td>
<td>24.6</td>
<td>30-3</td>
<td></td>
<td>22.7</td>
<td>25.2</td>
<td></td>
</tr>
<tr>
<td>MV</td>
<td>15.8</td>
<td>4.5</td>
<td></td>
<td>14.9</td>
<td>2.9</td>
<td></td>
</tr>
<tr>
<td>CDV</td>
<td>11.5</td>
<td>2.8</td>
<td></td>
<td>15.0</td>
<td>3.3</td>
<td></td>
</tr>
<tr>
<td>RSV</td>
<td>12.7</td>
<td>1.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* Alignments were scored using the global search program ALIGN (see Methods, Dayhoff, 1978). The sources for the published sequences were as follows: SV (Giorgi et al., 1983); CDV (Barrett et al., 1985); RSV (Satake et al., 1984).
† Number of identities divided by the total length of the aligned sequences, including gaps.
‡ Each alignment score is expressed as the number of standard deviations from the mean of scores for 100 randomized permutations of the two sequences (Dayhoff, 1978). For proteins that are thought to serve similar functions, scores of 3 and above are considered indicative of an authentic relationship (Barker & Dayhoff, 1982).
The unmodified PF3 C protein contains 199 amino acids with a calculated mol. wt. of 23 293. The PF3 C protein, like the C proteins of SV, MV and CDV, is strongly basic, with an estimated pI of 10.4 and a calculated charge of +14 at pH 7.
Fig. 5. Comparison of the hydropathy profiles of the P and C protein of PF3 and SV. The profiles were generated using the values of Kyte & Doolittle (1982). Regions of relatively low sequence homology in the P protein sequences are indicated by solid bars. Gaps were introduced into the profiles for the P proteins to maximize similarity.
The C proteins of PF3, SV, MV and CDV were compared by amino acid sequence alignment (Table 1). The results were analogous to those described above for the P protein: (i) the number of exact amino acid matches in each alignment was insufficient to be of clear significance; (ii) the ALIGN scoring matrix showed that the C proteins of PF3 and SV share a moderate amount of homology; (iii) homology between the C protein of PF3 and the C proteins of MV and CDV was detectable, but was at the lower limit of significance. The sequence of the PF3 C protein was also compared with the sequence of the two non-structural proteins of RSV, the 1C/NS1 and 1B/NS2 proteins (Collins & Wertz, 1985), and the NS proteins of the Indiana and New Jersey serotypes of VSV (Gill & Banerjee, 1985) (data not shown). In each case, the alignments contained only scattered amino acid matches, and the ALIGN program assigned a negative value to each alignment, indicating that no significant homology was evident.

The alignment of the sequences of the PF3 and SV C proteins is shown in Fig. 6. Sequence homology was distributed throughout the molecule, although sequence identity was greatest within positions 116 to 151 of the PF3 C sequence and the region of the alignment representing positions 16 to 57 contained a series of alternating gaps and homologous regions.

Hydropathy plots for the PF3 and SV C proteins were generally similar (Fig. 5), with the greatest difference being coincident with the region of the alignment (positions 16 to 57 in the PF3 sequence) that, as shown in Fig. 6, contained the series of gaps. This provided additional evidence of close similarity between the C proteins of PF3 and SV despite the absence of extensive sequence identity.

**DISCUSSION**

The complete sequence of the P + C mRNA of PF3 was determined by sequencing cDNAs, vRNA and mRNA. The mRNA contains two overlapping ORFs that, on the basis of hybrid-arrest and peptide mapping experiments, encode two unique proteins, the nucleocapsid-associated P protein and the non-structural C protein. Previous reports of non-structural proteins for PF3 have been inconsistent; in one report, several possible candidates were noted (Sanchez & Banerjee, 1985a), and a second report described an apparent lack of non-structural proteins (Wechsler et al., 1985a). However, the results described here, together with previous gene mapping experiments (Spriggs & Collins, 1986), establish that the genome of PF3 is transcribed into six mRNAs (NP, P + C, M, F, HN and L), that encode seven unique proteins (NP, P, C, M, F, HN and L).

Paramyxoviruses have three different strategies for encoding P and non-structural proteins. (i) The P mRNAs of NDV, simian virus 5 and mumps viruses encode relatively small P proteins (45K to 56K) and one or two non-structural proteins that are related to the P protein by peptide mapping and therefore presumably are encoded by the same ORF (Collins et al., 1982; Herrler & Compans, 1982; Paterson et al., 1984). (ii) The P mRNA of RSV encodes a single protein of 33K; the two RSV non-structural proteins are encoded by two separate mRNAs derived from...
different genomic loci (Collins et al., 1984b). (iii) The P+C mRNAs of SV, MV and CDV contain two overlapping ORFs that encode a relatively large P protein (79K to 84K) and one or two electrophoretically distinct forms of the unique, non-structural C protein (Giorgi et al., 1983; Bellini et al., 1985; Barrett et al., 1985). The results presented here show that the coding strategy for PF3 is similar to those for SV, MV and CDV. For all four viruses, the first, 5'-proximal AUG of the sequence initiates the P ORF and the second AUG, located 10 to 22 nucleotides downstream, initiates the C ORF.

Primer extension on PF3 mRNA (not shown) confirmed that the P+C mRNA does not contain sequence microheterogeneity at the 5' end, consistent with results for SV and MV (Giorgi et al., 1983; Bellini et al., 1985). Thus, the P and C proteins are not encoded by separate mRNA subpopulations which might have been generated by splicing or the use of alternate transcriptional start sites. Instead, as described previously (Giorgi et al., 1983), ribosomes presumably can gain access to the C ORF despite its internal location because the translational start site of the P ORF for each of these viruses is consistent with inefficient utilization (Kozak, 1986). Alternatively, Shioda et al. (1986) suggested that the SV P + C mRNA might assume two forms of secondary structure, one that would appear to favour expression of the P ORF and the other of the C ORF.

Alignments of the amino acid sequences of the PF3 P and C proteins with their counterparts in SV, MV and CDV revealed a relatively low content of exact amino acid matches. Nonetheless, consideration of both matches and mismatches provided unambiguous evidence that the P and C proteins of PF3 and SV are homologous pairs. Structural homology within each pair was also suggested by similarities in the hydrophobic profiles. An analogous situation of relatively low sequence identity and conserved hydrophathy profiles also exists for the NS proteins of the New Jersey and Indiana serotypes of VSV (Gil & Banerjee, 1985).

For the PF3 and SV P proteins, similarity in hydrophobic profile was the most striking in the C-terminal halves of the molecules, coincident with the region of greatest sequence homology. Previously reported protease mapping experiments showed that nucleocapsid-associated SV P protein could be reduced to a protease-resistant core (40K) without substantial loss of transcriptase activity (Chinchar & Portner, 1981a). This putative active core also was mapped to the C-terminal half of the molecule (Deshpande & Portner, 1985).

One purpose in comparing amino acid sequences within paramyxovirus protein families is to search for areas of strong localized sequence homology, which might signify regions that are important to conserved structures or functions. For example, the NP proteins of PF3, SV, MV and CDV contain two sequences of 6 and 7 amino acids that are conserved exactly (Galinski et al., 1986; Jambou et al., 1986). Similarly, the N-terminal 25 amino acids of the F1 subunit for PF3, SV, NDV, simian virus 5 and MV contain approximately 70 to 88% sequence identity and are thought to be directly involved in membrane fusion (Spriggs et al., 1986; Varsanyi et al., 1985; Richardson et al., 1980). Also, the HN glycoprotein of PF3, SV and simian virus 5 contains several regions, 6 to 7 amino acids in length, that have almost complete identity (Kolakofsky et al., 1986). In contrast, the P and C protein families of PF3, SV, MV and CDV contain no analogous regions of conserved sequence. This was somewhat surprising because the P protein is known to be involved in transcription (see Introduction) and it has been speculated that the C protein may be involved in vRNA replication. These functions would be expected to be highly conserved. Whatever the roles of the P and C proteins are, they apparently do not involve active sites and/or structural features of conserved sequence.

Amino acid sequence alignments within protein families can also be used to estimate relative evolutionary distances between paramyxoviruses, and to compare the relative rates of change of the different protein families. Table 2 shows the scores for alignments of the PF3 HN, F and NP proteins with the available sequences of their counterparts in other paramyxoviruses. In all comparisons, PF3 was most closely related to SV (murine parainfluenza type 1), with simian virus 5 (canine parainfluenza virus type 2) being the next most closely related. Previously, the NP, P and C proteins of MV and CDV were shown to be closely related, indicating that MV and CDV represent a homologous pair analogous to PF3 and SV (Rozenblatt et al., 1985; Barrett et al., 1985). However, as shown in Tables 1 and 2, there was only low homology between the
Table 2. Relationships of the HN, F and NP proteins of PF3 with their counterparts in other parainfluenzaviruses

<table>
<thead>
<tr>
<th>Protein</th>
<th>Identity (%)</th>
<th>S.D.</th>
</tr>
</thead>
<tbody>
<tr>
<td>SV HN</td>
<td>45</td>
<td>72.7</td>
</tr>
<tr>
<td>SV5 HN</td>
<td>19</td>
<td>20.4</td>
</tr>
<tr>
<td>MV H</td>
<td>15.3</td>
<td>4.6</td>
</tr>
<tr>
<td>RSV G</td>
<td>12.3</td>
<td>-0.9</td>
</tr>
<tr>
<td>SV F</td>
<td>41.1</td>
<td>37.5</td>
</tr>
<tr>
<td>SV5 F</td>
<td>25.3</td>
<td>15.7</td>
</tr>
<tr>
<td>RSV F</td>
<td>19.2</td>
<td>8.5</td>
</tr>
<tr>
<td>SV NP</td>
<td>58.4</td>
<td>59.8</td>
</tr>
<tr>
<td>MV NP</td>
<td>ND§</td>
<td>19.7</td>
</tr>
<tr>
<td>CDV NP</td>
<td>20.3</td>
<td>9.5</td>
</tr>
<tr>
<td>RSV N</td>
<td>19.3</td>
<td>1.8</td>
</tr>
</tbody>
</table>

* Alignments were scored using the global search program ALIGN (Dayhoff, 1978). The sequences were obtained from published work (Alkhatib & Briedis, 1986; Collins et al., 1985; Galinski et al., 1986; Jambou et al., 1986; Morgan et al., 1984; Rozenblatt et al., 1985; Spriggs et al., 1986; Wertz et al., 1985; Kolakofsky et al., 1986; and references cited therein).

† Identity (%) and alignment scores were determined as described in the footnotes to Table 1.
‡ SV5, Simian virus 5.
§ ND, Not done.

PF3/SV pair and the MV/CDV pair. All four viruses, however, share a common strategy for encoding the P and C proteins and have homologous genetic maps and intergenic sequences (Spriggs & Collins, 1986). Although PF3 and simian virus 5 share a moderate amount of sequence homology, simian virus 5 contains an additional gene (encoding the SH protein), has apparent differences in intergenic structure, and has a different strategy for encoding the P and non-structural proteins (Hiebert et al., 1985; Paterson et al., 1984). These comparisons delineate three interrelated groups among the parainfluenza and morbilliviruses: (i) PF3 and SV, (ii) simian virus 5 and (iii) MV and CDV.

The pneumovirus RSV appears to represent a fourth group on the basis of differences in its genetic map, gene products and intergenic sequences (Collins et al., 1986). Amino acid sequence alignments provided unambiguous evidence that the RSV F protein is related to its paramyxovirus counterparts (Table 2; Spriggs et al., 1986). Somewhat surprisingly, a similar global relationship for the RSV N protein could not be demonstrated (Table 2; Galinski et al., 1986; Jambou et al., 1986). Alignment of the sequence of the P protein of RSV with the P proteins of PF3, SV, MV and CDV (Table 1, unpublished data) also did not demonstrate a statistically significant global relationship. However, because the RSV P protein is substantially smaller (241 amino acids) than its parainfluenza virus and morbillivirus counterparts (602 and 507 amino acids, respectively), we also compared the RSV P protein with amino acid segments of similar size taken from these larger P proteins. This analysis failed to demonstrate a statistically significant relationship of RSV P with the P proteins of the animal viruses, SV and CDV, but did reveal significant homology with amino acids 243 to 499 of the PF3 P protein (2.83 s.d.) and amino acids 200 to 440 of the MV P protein (3.54 s.d.). These data suggested that the RSV P protein is related to a protein domain that is located at similar positions in the PF3 and MV P proteins. The significance of this homology is weakened by the lack of similar homology with SV and CDV, but the greater homology between RSV, PF3 and MV might be a consequence of their shared natural host.

The data in Tables 1 and 2 also show that the paramyxovirus protein families appear to have different rates of change. For example, comparison of protein sequences available for PF3 and SV show that the HN proteins share the greatest homology (based on the Dayhoff scoring matrix) followed in order by the NP, F, P and C proteins. Additional sequence information will
be needed to determine whether this is a general pattern. One exception is that the MV H protein appears to have less homology to the HN proteins of PF3/SV than do their corresponding NPs (Alkhatib & Briedis, 1986). However, this might reflect the lack of neuraminidase activity by the MV H protein. A second exception is that the RSV F protein is related unambiguously to its paramyxovirus counterparts, whereas relationships for the N and other RSV proteins cannot be demonstrated or are at the threshold of significance. The pattern of sequence diversity among the paramyxovirus protein families described here contrasts with the general pattern for influenza virus, where the sequences for the internal virion components are relatively more conserved and the sequences for the surface glycoproteins are relatively more divergent.

This work was performed in the laboratory of Dr Brian R. Murphy and Dr Robert M. Chanock. We thank them for their support and helpful discussions. We thank Alicia Buckler-White for the synthesis of oligonucleotides, A. B. W. and Charles E. Buckler for help with computer work, Ena Camargo for the preparation of virus, cells and mRNAs, and Linda Jordan for editorial assistance.

Note added in proof: While this paper was in press, Luk et al. (Virology 153, 318-325) published a nucleotide sequence of the P + C mRNA for the identical strain of PF3. The sequence reported by Luk et al. differs from that presented here by (i) an insertion of six nucleotides at position 804 and (ii) a deletion of a T residue at position 1883 which alters the location of the termination codon. We have addressed these differences by (i) confirming the sequence surrounding the termination codon in six independently derived cDNA clones and (ii) confirming the sequence of both regions directly from genomic RNA. From this, we conclude that the sequence reported here is authentic.

References


Paramyxovirus protein sequence homologies


(Received 16 June 1986)