Sequence Analysis of the Matrix Protein Gene of Human Parainfluenza Virus Type 3: Extensive Sequence Homology among Paramyxoviruses

By MELANIE K. SPRIGGS,1 PHILIP R. JOHNSON1,2 AND PETER L. COLLINS1*

1 Laboratory of Infectious Diseases, Building 7, Room 100, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, Maryland 20892 and 2 Department of Pediatrics, Vanderbilt University School of Medicine, Nashville, Tennessee 37232, U.S.A.

(Accepted 10 February 1987)

SUMMARY

The sequences of the human parainfluenza virus type 3 (PIV3) matrix (M) mRNA [1150 nucleotides exclusive of poly(A)] and predicted M protein (353 amino acids) were determined by sequence analysis of cloned cDNA and viral genomic RNA. The gene-end sequence of the M gene differed from the semi-conserved gene-end sequence of the other PIV3 genes by an apparent insertion of eight nucleotides. The PIV3 M protein shared high sequence homology with Sendai virus and moderate homology with measles virus and canine distemper virus. Statistical analysis of the available sequences showed that the M protein was the most highly conserved parainfluenza viral protein.
consequences in virion assembly, the establishment of persistent infection, and viral pathogenesis.

To determine the structures of the M gene and M protein of PIV3 (strain 47885), the entire M gene and flanking gene regions were sequenced directly from genomic RNA by the dideoxy method using synthetic oligonucleotides as primers (Spriggs et al., 1986). Additionally, the 3' proximal 902 nucleotides of the M mRNA were determined by sequencing both strands of an M cDNA clone, PM14, by the dideoxy method using synthetic oligonucleotides as primers (Zagursky et al., 1985).

The nucleotide and predicted amino acid sequences of the M mRNA are given in Fig. 1. The mRNA is 1150 nucleotides in length, exclusive of poly(A). The first ATG occurs at positions 33 to 35 and initiates the only significant open reading frame which terminates at positions 1092 to 1094. Thus, the PIV3 M mRNA resembles that of SV (Blumberg et al., 1984; Hidaka et al., 1984), in that it does not possess a long 3' untranslated region such as that found in MV and canine distemper virus (CDV) (Bellini et al., 1986). Precise 5' mapping and sequencing of the M mRNA by primer extension was reported previously (Spriggs & Collins, 1986a). The sequence of the 5' end of the M mRNA, 5' AGGATTAAAG, is consistent with the semi-conserved 10 nucleotide sequence 5' AGGANNAAAG, found at the 5' ends of the PIV3 NP, P+C, F, HN and L genes (Spriggs & Collins, 1986a). The 3' end of the M mRNA, 5' AAATAAGAGATAAT-CAAAAA, differs from the semi-conserved gene-end sequence for the NP, P+C, F and HN mRNAs, 5' AA~TA~AAAAA, by the insertion of eight nucleotides (bold type) just upstream of the five templated A residues that mark the beginning of the poly(A) tail. As described previously, this aberrant gene-end sequence correlated with an increased incidence of readthrough transcription, resulting in a greater accumulation of the M-F bicistronic mRNA than was observed for any other polycistronic mRNA (Spriggs & Collins, 1986a).

The M mRNA encodes a protein of 353 amino acids with a calculated mol. wt. of 39507. The hydropathy profile of the PIV3 M protein (not shown) was almost identical to that of the SV M protein. The sequence did not appear to contain terminal membrane-spanning domains, but did contain a structurally conserved hydrophobic region (amino acids 254 to 298) that Bellini et al. (1986) had noted previously for SV, MV and CDV and suggested might be involved in membrane interaction. As is the case with the other paramyxoviral M proteins cited above, the PIV3 M protein is highly basic, with an estimated pI of 10.6 and a calculated charge of +21.5 at neutral pH.

In previous work, we used amino acid alignments and the ALIGN program scoring matrix (Table 1; Dayhoff, 1978) to identify patterns of sequence homology for several paramyxoviral proteins. This scoring matrix assigns a value to each amino acid pair based on their identity or similarity. The score of the optimally aligned sequences is compared to the average score of alignments of 100 randomized permutations of the two sequences and is reported as the number of standard deviations from the randomized mean. This showed that the proteins of PIV3 and SV are closely related and that these protein families can be listed in order of decreasing relatedness as HN, NP, F, P and C (Spriggs & Collins, 1986b). Alignment of the sequences of the M proteins of PIV3 and SV is shown in Fig. 2. The lengths of the proteins differ by only five amino acids and are optimally aligned with no internal gaps in either sequence. There is 61% amino acid identity overall and eight stretches of six or more exact amino acid matches, five of which occur in the carboxy-terminal third of the molecule.

Similar amino acid sequence comparisons (alignments not shown) demonstrated that the PIV3 M protein was also related, albeit less closely, to the M proteins of MV and CDV (34% and 32% sequence identity, respectively) (Bellini et al., 1986; Table 1). This is consistent with the general observation that, on the basis of comparisons of available amino acid sequences, SV and PIV3 represent one highly homologous pair and MV and CDV another, and that substantially less homology exists between the SV/PIV3 and MV/CDV pairs (Spriggs & Collins, 1986b; Rozenblatt et al., 1985). Comparison of all four M protein sequences in parallel revealed a remarkably high number (90) of amino acid matches (Fig. 2) and confirmed previous observations (Bellini et al., 1986) that homology is more extensive in the carboxy-terminal third of the molecules (residues 235 to 353 for PIV3, 231 to 348 for SV, and 217 to 335 for MV and
Fig. 1. Complete nucleotide and predicted amino acid sequences of the PIV3 M mRNA and protein.

CDV). Two additional, more localized areas of four-way homology were also apparent: residues 105 to 120 and 138 to 156 for PIV3, 101 to 116 and 134 to 152 for SV, and 89 to 109 and 122 to 140 for MV and CDV (Fig. 2).

The relationships between the PIV3 M protein and the M proteins of other non-segmented negative strand viruses are defined statistically in Table 1. Interestingly, the standard deviation (s.d.) value for the PIV3 and SV M comparison (89 s.d.) exceeded that for the corresponding HN (73), NP (60), F (38), P (30) or C (25) comparisons (Table 1; Spriggs & Collins, 1986b). Similarly, the s.d. values for the comparison of the M proteins of PIV3 and MV (35) and CDV (39) were greater than the values for the PIV3 NP and the MV (20) and CDV (9-5) NP proteins, the PIV3 HN protein and the MV H protein (4-6), the PIV3 C protein and the MV (2-9) and CDV (3-3) C proteins, and the PIV3 P protein and the MV (4-5) and CDV (2-8) P proteins.
Table 1. Relationships among the amino acid sequences of the M proteins of several non-segmented negative strand RNA viruses*

<table>
<thead>
<tr>
<th>PIV3 M compared to</th>
<th>Identity (%)†</th>
<th>Degree of homology (s.d.)‡</th>
</tr>
</thead>
<tbody>
<tr>
<td>SV M</td>
<td>61</td>
<td>89</td>
</tr>
<tr>
<td>MV M</td>
<td>34</td>
<td>39</td>
</tr>
<tr>
<td>CDV M</td>
<td>32</td>
<td>35</td>
</tr>
<tr>
<td>RSV M</td>
<td>10</td>
<td>0-8</td>
</tr>
<tr>
<td>VSV M</td>
<td>12</td>
<td>0-5</td>
</tr>
</tbody>
</table>

* Alignments were scored using the global search program ALIGN and the mutation data matrix which scores both matches and mismatches according to empirically determined patterns of amino acid conservation and substitution (Dayhoff, 1978). The gap penalty was 12 for all comparisons except those with RSV or VSV, where the penalty was 6. The sources for the published sequences were: SV (Blumberg et al., 1984), MV and CDV (Bellini et al., 1986), RSV (Satake & Venkatesan, 1984) and VSV (Rose & Gallione, 1981).
† 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 considered to share a common function, scores of 3 s.d. or greater are considered indicative of an authentic relationship (Barker & Dayhoff, 1982).

Thus, within the parainfluenza virus and morbillivirus genera, the M proteins exhibited relatively greater sequence conservation than the HN, NP, F, P and C proteins. As for the L protein, a complete L gene sequence is available only for SV (Rakestraw & Morgan, 1986; Shioda et al., 1986). By dideoxynucleotide sequencing of vRNA, we have determined a partial, 732 nucleotide sequence for the PIV3 L gene, including the predicted coding sequences for the N-terminal 236 amino acids of the L protein (Spriggs & Collins, 1986a; unpublished data). Comparison of this partial PIV3 sequence with the analogous segment of the SV L protein
identified a relationship with a S.D. value of 27 (not shown). If this value is representative of the homology between the complete L protein sequences, the M protein would be the most highly conserved of the paramyxoviral protein families. This somewhat unexpected finding implies that a more stringent requirement exists for conservation of the M protein sequence, presumably due to functional constraints. Interestingly, this differs from the situation reported for the rhabdoviruses vesicular stomatitis virus and rabies virus, which share significant homology for the G proteins but have M proteins that lack significant sequence homology and, surprisingly, have distinct hydrophobic profiles (Rayssiguier et al., 1986).

In the light of the extensive M protein sequence conservation among the parainfluenza viruses and morbilliviruses, it was surprising that no statistically significant global sequence homology could be demonstrated with the M protein of the pneumovirus RSV (Table 1). The use of a dot matrix analysis that employed the Dayhoff (1978) scoring parameters and was designed to detect areas of limited, local sequence homology also did not detect evidence of relatedness (not shown). A similar absence of detectable global homology exists for the RSV G and NP proteins, whose counterparts are highly conserved among other paramyxoviruses (Spriggs & Collins, 1986b). Indeed, unambiguous homology involving the available RSV sequences exists only in the case of the F protein (Spriggs et al., 1986). Thus, with regard to sequence homology, the paramyxoviruses appear to represent at least two distinct evolutionary pathways. Also, as described previously (Collins et al., 1986; Spriggs & Collins, 1986a, b), the paramyxoviruses examined to date appear to fall into three groups (the first containing SV, PIV3, MV and CDV; the second, simian virus 5 and Newcastle disease virus; and the third, RSV) based on differences in intergenic structures, gene maps, and numbers and types of mRNA and proteins.

If all present day paramyxoviruses originated in a single non-segmented progenitor, then (i) the differences in protein sequence described above indicate that extensive divergence has occurred between the different branches of the family, especially for RSV, and (ii) features of genome organization such as gene number, gene order and intergenic sequences must have been subject to greater evolutionary change than has been previously appreciated. An alternative model to reconcile the differences among paramyxoviruses is based on the previous suggestion that the non-segmented genome of SV originated by ligation of the gene segments of an influenza virus-like ancestor (Giorgi et al., 1983; Kolakofsky et al., 1986). If so, perhaps the present day paramyxoviruses evolved from several different non-segmented progenitors which in turn each arose from separate ligation events involving gene segments from more than one segmented ancestor. This speculative model of multiple, interrelated ancestors could account, for example, for the presence of a small hydrophobic protein gene in some paramyxoviruses (simian virus 5 and RSV) but not others (SV, PIV3, MV and CDV), for the presence of overlapping P and C genes in some viruses (SV, PIV3, MV and CDV) but not others (simian virus 5, Newcastle disease virus, RSV), for the novel 1B, 1C, 22K and G genes that are unique to RSV, for the reversed order of the fusion and attachment protein genes in the RSV map, and for the non-uniform pattern of sequence conservation and divergence within protein families. From the M protein sequence homology described here, the M proteins of SV, PIV3, MV and CDV would be predicted to have a common origin in one segmented ancestor, while the unrelated M protein of RSV might represent a different segmented ancestor. In contrast, the F proteins of all paramyxoviruses sequenced to date share significant homology and therefore would have originated from a common gene segment. All paramyxoviruses examined to date share the partial gene order 3' NP-P-M • • • F • • • L, which has been interpreted to be evidence that these viruses share a common non-segmented progenitor. However, this also would not be unexpected for viruses derived from separate linkage events, given the importance of gene order in the stoichiometry of gene expression, and postulating that ancestral non-segmented viruses with this appropriate gene order would enjoy a selective advantage.

This work was performed in the laboratories of Brian R. Murphy and Robert M. Chanock. We thank them for their interest and advice; Ena Camargo for virus, RNA and cells; Alicia Buckler-White for oligonucleotides; Debra Brunelle and Linda Jordan for editorial assistance and the Laboratory of Molecular Genetics, National Institute of Child Health and Human Development for the use of their VAX 11/750 computer.
Note added in proof. Following the submission of this manuscript, Luk et al. (Virology 156, 189–192) and Galinski et al. (Virology 157, 24–30) published sequences for the M gene of the same strain of PIV3 described here. The three sequences are in agreement except for nucleotide differences at positions 180, 271, 322, 323, 339, 340, 359, 447 and 841. Many of these result in amino acid changes, but these are clustered in a region that appeared to be relatively less conserved among paramyxoviruses and therefore perhaps less subject to functional or structural constraints. Thus, these differences could reflect naturally occurring sequence variation, and illustrate the potential for sequence diversity in PIV3.

REFERENCES


\textit{(Received 12 January 1987)}