Sequence of the major nucleocapsid protein gene of pneumonia virus of mice: sequence comparisons suggest structural homology between nucleocapsid proteins of pneumoviruses, paramyxoviruses, rhabdoviruses and filoviruses

J. Barr, P. Chambers, C. R. Pringle and A. J. Easton*

Department of Biological Sciences, University of Warwick, Coventry CV4 7AL, U.K.

The complete nucleotide sequence of gene 3 of pneumonia virus of mice has been determined, and the 5' end of the mRNA mapped using a modification of the polymerase chain reaction technique. The gene contains a single open reading frame, beginning with a 5'-proximal AUG initiation codon, encoding a polypeptide with a predicted Mr of 43141. Expression of the gene 3 protein in Escherichia coli and in vitro showed that it reacted with virus-specific antiserum and comigrated with the major nucleocapsid (N) polypeptide. The predicted amino acid sequence has extensive identity with that of the N protein of human respiratory syncytial virus. Comparisons with the amino acid sequences of N proteins of other paramyxoviruses, vesicular stomatitis virus and Ebola virus suggest that these proteins may have retained much of the same structure. These regions of conserved structure would most likely have the common functions of RNA binding and protein/protein interactions in the virus nucleocapsid.

Introduction

Pneumonia virus of mice (PVM) is classified in the family Paramyxoviridae on the basis of general structural characteristics (Compans et al., 1967; Berthiaume et al., 1974). Within this family, PVM is grouped in the Pneumovirus genus together with human respiratory syncytial (RS) virus, bovine RS virus and a recently described virus causing rhinotracheitis in turkeys (Ling & Pringle, 1988; Collins & Gough, 1988); all three pneumoviruses are associated with infection of the respiratory tract (Pringle, 1987). Human RS virus is an important pathogen and is the principal cause of pneumonia and bronchiolitis in infants under the age of 2 years (Pringle, 1987). PVM was initially identified as the agent responsible for a latent or inapparent infection of mice (Horsfall & Hahn, 1939, 1940); however, recent studies suggest that the human population is frequently exposed to infection by PVM or an antigenically related virus (Pringle & Eglin, 1986). No disease has yet been associated with the human infection and the clinical relevance is unclear. Antigenic cross-reactions involving the nucleocapsid (N) protein and phosphoprotein of PVM and human RS virus have been described (Gimenez et al., 1984; Ling & Pringle, 1989).

Whereas the molecular biology of human RS virus has been studied extensively and understanding of this virus has increased rapidly, the molecular cloning of PVM mRNA species has only recently been reported (Chambers et al., 1990). Studies of the genetic organization of PVM and RS virus demonstrate that these two viruses are similar to one another and distinct from other members of the family Paramyxoviridae (Chambers et al., 1990, 1991).

We describe here the determination of the nucleotide sequence of PVM gene 3 and demonstrate considerable identity with the gene encoding the human RS virus N protein (Collins et al., 1985) and lower levels of amino acid identity with the N proteins of other non-segmented negative-strand viruses. Expression of PVM gene 3 in vitro and in Escherichia coli confirms that it encodes the virus N protein.

Methods

cDNA cloning and nucleotide sequence analysis. cDNA clones representing PVM mRNAs have been described previously (Chambers et al., 1990). Many cDNA clones representing gene 3 were obtained, one of which contained an insert of approximately 1300 bp, very close to the size of the mRNA detected by Northern blot analysis (Chambers et al., 1990). This was designated 6.19.

The insert within clone 6.19 was excised, purified and digested with a variety of restriction endonucleases to generate several smaller fragments which were subcloned into bacteriophage M13mp18 and M13mp19 sequencing vectors using standard techniques (Maniatis et
The nucleotide sequence was determined using the chain-termination method of Sanger et al. (1980). The sequence of another clone, 6.42, representing approximately 60% of the 5'-terminal region of gene 3 was determined also and found to be identical to that of 6.19.

**Identification of the 5' end of gene 3 mRNA.** A 125 bp region representing the 5' end of gene 3 mRNA was amplified by the polymerase chain reaction (PCR) technique using a method based on that of Land et al. (1981) and Saiki et al. (1988). PVM-infected BSC-1 cell RNA, prepared by the method of Kumar & Lindberg (1972), was used as a template for cDNA synthesis using oligo(dT) as the primer (Maniatis et al., 1982). The resulting double-stranded cDNA–RNA hybrid molecules were tailed with dGTP using terminal transferase to create an oligo(dG) tract at the 3' ends (Maniatis et al., 1982). The DNA segment homologous to the 5' end of the mRNA was amplified between two oligonucleotide primers, one of which (designated oli C; CCCCCAGCTCTGACAGGATCCCCCCCCCCCC) was designed to hybridize to the poly(dG) tract, the other (CATGCGCTCTGGAATCTGG) was designed to hybridize to a region spanning nucleotides 108 to 125 of the sequence shown in Fig. 1.

Amplification of this segment was achieved by 35 cycles of DNA denaturation for 1·5 min at 95 °C, and oligonucleotide annealing and extension of the primer oligonucleotides using the thermostable Taq DNA polymerase (Amplitaq; Cetus) for 3 min at 65 °C. The amplification was performed in a reaction mixture containing cDNA generated from 2 μg of infected cell RNA, each deoxyribonucleoside triphosphate at 500 mM, 10 mM-Tris-HCl pH 8·3, 50 mM-KCl, 1·5 mM-MgCl₂, 0·01% gelatin and 1 mM-dithiothreitol. Taq DNA polymerase was added to give a total of 2·5 units in a 100 μl reaction volume. To prevent volume reduction by evaporation, 30 μl of mineral oil was overlaid on the reaction mix. The amplified DNA was recovered from the reaction mix by phenol/chloroform extraction and ethanol precipitation. The reaction products were separated by electrophoresis in a low melting point agarose gel and the DNA fragment was isolated (Maniatis et al., 1982). The fragment was digested with PstI, which has a recognition site in oligo C, and XbaI, which has a recognition site at position 37 in the gene 3 sequence, and was then ligated into the bacteriophage M13mp19 vector for sequencing (Maniatis et al., 1982).

**Expression in E. coli.** A DNA fragment containing the entire coding sequence of PVM gene 3 was isolated by digestion of clone 6.19 with FokI and BamHI, followed by incubation with nuclelease S1 to generate blunt ends (Maniatis et al., 1982). The fragment, extending from position 20 in gene 3 (Fig. 1) through the whole gene to a BamHI site in the pUC13 cloning vector downstream of the Smal insertion site, was purified following agarose gel electrophoresis.

The gene 3 DNA fragment was inserted into the vector pKQV4, which had been digested with EcoRI followed by treatment with nuclease S1. Plasmid pKQV4 was a kind gift of Dr J. Hoch (Scripps Clinic and Research Foundation) and was constructed by cloning the intact lacI gene into the expression vector pKK223 (Pharmacia). In this way, PVM gene 3 was placed under the control of the synthetic tac promoter, which can be induced by isoprropyl-β-D-thiogalactopyranoside (IPTG).

Expression was achieved as described by Stark (1987). Bacteria were grown to an OD₆₀₀ of 0·5 and expression was induced by the addition of IPTG to a final concentration of 0·5 mM. Following a 2 h incubation at 37 °C, the polypeptides were analysed by Western blotting. PVM-specific polypeptides were detected with polyclonal mouse anti-PVM antisera, prepared by Dr R. Ling using virus-infected cells as antigen, together with the Amersham mouse antibody detection kit which uses alkaline phosphatase as the indicator, as described by the manufacturer.

In vitro transcription/translation. The FokI–BamHI fragment of clone 6.19 described above, containing the entire coding sequence of PVM gene 3, was inserted into the Smal site in the multiple cloning site of the transcription vector pGEM1 (Promega Biotech). Gene 3-specific mRNA was then transcribed from the T7 promoter using the reaction conditions recommended by the manufacturer. The RNA was used to direct the in vitro translation of [³⁵S]methionine-labelled protein in a rabbit reticulocyte lysate system (Amersham). Total cytoplasmic RNA from PVM-infected cells was also translated in vitro as described previously. The protein products were analysed by electrophoresis on a 10% acrylamide gel containing 50% (w/v) sucrose in the resolving gel (Chambers & Samson, 1982).

**Results**

**cDNA sequencing**

The DNA sequence and the predicted polypeptide sequence of PVM gene 3 are presented in Fig. 1. The 5'-terminal 42 nucleotides of this sequence were derived from the sequence of the PCR-amplified segment which defined the precise 5' end of the mRNA as described below. The remainder of the sequence was derived from cDNA clones 6.19 and 6.42. The 3'-terminal 73 nucleotides were confirmed by analysis of cDNA clones spanning the gene 3 to gene 4 intergenic region (Chambers et al., 1991).

PVM gene 3 is 1215 nucleotides long and contains a single open reading frame, with a methionine initiation codon nucleotide positions 32 to 34, which has the potential to encode a protein of 393 amino acids. The predicted Mr of this protein is 43141, in close agreement with the observed Mr, (41000) of the PVM N protein on SDS–polyacrylamide gels (Cash et al., 1977; Ling & Pringle, 1989). We have shown that the 41000 Mr protein is encoded by gene 3 using hybrid-arrested translation in vitro (Chambers et al., 1990).

**Determination of the 5' end of PVM gene 3 mRNA**

A variation on the PCR technique was used to amplify the 5' end of the PVM mRNA representing gene 3 from a known internal position in the gene to the 5' end of the mRNA. Thirty independent M13mp19 phage clones containing PCR-amplified DNA were selected; 10 of these were sequenced individually. Aliquots of the remaining 20 phage DNA preparations were pooled and sequenced collectively. The sequences obtained were antisense with respect to the mRNA.

The nucleotide sequences of the 10 individually sequenced clones were identical until the sequence TCCT preceding the poly(G) tract was reached. The nucleotide incorporated after this sequence was either a C residue (in seven clones) or the first G residue of the tail (three clones). An autoradiograph of the sequencing gel on which four of these samples were analysed is
Fig. 1. Nucleotide sequence of PVM gene 3 shown in mRNA sense. The sequence of the predicted polypeptide encoded by the single large open reading frame is shown underneath. The numbers refer to the nucleotide positions in the gene.

shown in Fig. 2 (lanes 1 to 4). In one of the clones (lane 1) the poly(G) tract was preceded by GA. The reasons for this are unclear but the A nucleotide may represent a misincorporation of A for G by Taq polymerase. The sequence analysis of the pooled sample is shown in Fig. 2 (lane 5); this showed the presence of a faint band in the C lane coincident with the first base of the poly(G) tract. The presence of this faint band suggests that the variable base position (G or C) may be due to an attempt by the reverse transcriptase to form a base-pairing match with the 5' cap residue on some template molecules. This phenomenon has also been observed in primer extension reactions (Gupta & Kingsbury, 1984; Collins et al., 1984). Nucleotide sequencing of cDNA clones from virus genomic RNA has shown that the nucleotide preceding the start of gene 3 is a U in the mRNA sense strand (Chambers et al., 1991), confirming that the G residue which is frequently observed in this position in PCR products is due to misincorporation. Because of these considerations, we believe that the PVM gene 3
mRNA begins with the sequence AGGAUAAAU. This or a closely related sequence is found at the beginning of the nine PVM genes sequenced to date (Chambers et al., 1991).

The nucleotide sequence of the PCR-generated segment was not identical to the corresponding sequence of clone 6.19. A 20 nucleotide region of the PCR-amplified segment occurred in inverted sense at the 5' end of 6.19. Nucleotide sequencing of PCR-amplified segments representing the 5' regions of other PVM mRNAs has revealed the presence of similar rearrangements in many original cDNA clones (data not shown). This appears to correlate with the presence of inverted repeat sequences at the 5' termini of many PVM genes which could be significant for mRNA secondary structure formation. The consequence for cDNA cloning of PVM mRNAs is that, in the PVM cDNA clones where an inverted sequence is present, characteristically there are sequences flanking the inverted sequence which may form the stem of a hairpin loop structure at the 3' end of the first-strand cDNA. Nuclease S1 was not used to resolve hairpin loops in the method of cDNA synthesis employed to construct these clones but DNA polymerase I was used in the second-strand synthesis has 5'-3' exonuclease activity which may do so. Sequence inversion could have occurred by the following process which is shown diagrammatically in Fig. 3; a hairpin loop formed in the first strand of cDNA and primed synthesis of the second strand, DNA polymerase I-directed DNA synthesis replaced the first strand using the 5'-3' exonuclease and 5'-3' polymerase activities. Such a hairpin loop structure may present a physical barrier to further DNA synthesis. Since the polypeptide domain specifying the 5'-3' exonuclease activity of DNA polymerase I precedes that specifying the 5'-3' polymerase activity (Joyce & Steitz, 1987), several nucleotides would not be replaced, thus creating a gap in the duplex. If the remaining few base pairs between the gap and the loop in the molecule were subsequently to unwind, loop and stem would then be replicated by the 5'-3' polymerase activity and an inverted form of part of the original sequence, flanked by a short inverted repeat section, would be created. For clone 6.19, the FokI site used to excise the N protein gene is created by the inversion event (Fig. 3). Furthermore, there are similar rearrangements in many of the original mRNA-derived cDNA clones representing the other eight PVM genes (unpublished data), which can be rationalized readily by means of the explanation given above.

Thus we believe that the correct sequence of the 5' terminus of gene 3 is represented by the PCR-derived DNA. All 10 of the individually sequenced, PCR-generated DNAs had sequences identical to that derived from the pooled samples.
Expression of the PVM N protein in *E. coli* and in vitro

In order to confirm that PVM gene 3 encoded the virus N protein, a DNA fragment containing the entire open reading frame was inserted into an *E. coli* expression vector under the control of the synthetic *tac* promoter (Amman *et al.*, 1983; De Boer *et al.*, 1983). Following induction with IPTG, the bacteria were lysed and the bacterial proteins analysed by Western blotting using a polyclonal anti-PVM antiserum in conjunction with an alkaline phosphatase-based detection system. As can be seen in Fig. 4, the bacteria induced for expression of the heterologous protein directed the synthesis of a novel polypeptide, of approximate M₇, 40000, which reacted with the anti-PVM antibody confirming that gene 3 is a genuine virus gene. The many other bands which can be seen were common to both induced and uninduced bacteria and may be due to the presence of anti-bacterial antibodies in the polyclonal serum used. The M₇ of the induced polypeptide was similar to that predicted from the major open reading frame in the nucleotide sequence of gene 3 and to the polypeptide previously identified as the gene 3 product and the PVM N protein (Ling & Pringle, 1989; Cash *et al.*, 1977; Chambers *et al.*, 1990).
Confirmation that PVM gene 3 encodes the major N protein was obtained by *in vitro* transcription and translation of gene 3-specific RNA using the pGEM system. Translation of gene 3 RNA directed the synthesis of two products with Mr's of approximately 41000 and 22000 (Fig. 4). The larger product comigrated on sucrose/acrylamide gels with the polypeptide previously identified as the N protein (Cash *et al.*, 1977; Chambers *et al.*, 1990). The origin of the smaller 22000 Mr product is unclear as it comigrates with material translated from virus-infected cell RNA whose level of synthesis was apparently unchanged in hybrid-arrested translation assays (Chambers *et al.*, 1990; unpublished data).

**Discussion**

The complete nucleotide sequence of PVM gene 3 contains only one large open reading frame, which begins with a 5' proximal AUG initiation codon, encoding a polypeptide with a calculated Mr of 43141. Expression of the gene product in *E. coli* showed that the polypeptide product is recognized by anti-PVM antibody. The *in vitro* synthesized polypeptide comigrated with the major N protein on SDS–polyacrylamide gels.

A high level of amino acid identity (60%) is seen in a comparison of the predicted amino acid sequences of the RS virus and PVM N proteins (Fig. 7). Amino acid residues 1 to 150 and 150 to 393 contain 38% and 74% identity respectively, whereas residues 245 to 315 contain 68 identical amino acids out of 71 (96% identity). The high degree of conservation of these proteins is consistent with observations that the N proteins of RS virus and PVM are serologically related (Gimenez *et al.*, 1984; Ling & Pringle, 1989).

The genus *Pneumovirus* consists of viruses whose nucleocapsid morphology, genomic organization and nucleotide and amino acid sequences are rather different from those of members of either the *Paramyxovirus* or *Morbillivirus* genera. A striking difference between the amino acid sequences of N proteins of the *Paramyxoviridae* is the comparatively small size of the pneumovirus proteins, 391 or 393 residues compared to a typical length for most paramyxo- and morbillivirus N proteins of over 500 residues. We have tried to reconcile this size difference with conservation of functions of the N proteins.

Computer matrix comparisons of the amino acid sequences of the N proteins of PVM and members of various other virus groups reveal limited regions of apparent identity (Fig. 5). Most surprising is the similarity detected between the PVM N protein and the N protein of Ebola virus. A relationship between the N proteins of PVM and vesicular stomatitis virus (VSV) was only detected graphically when the comparison was biased. Gaps introduced by the CLUSTAL program (see below) were filled with X (for any residue), thus forcing an overall comparison between sections of pneumovirus and rhabdovirus sequences which were optimally spaced. Furthermore, the hydropathy profiles of N proteins from members of each of the distinct groups of non-segmented negative-strand viruses resemble each other in the region of greatest sequence similarity between paramyxoviruses and morbilliviruses (Galinski *et al.*, 1985). Approximately 180 amino acids, commencing 130 to 170 residues from the amino terminus of each protein form alternating hydrophobic and hydrophilic regions. For Ebola virus, however, there is more variation in the hydropathy profile in this region (Fig. 6). Secondary structure predictions for these amino acid sequences (Garnier *et al.*, 1978) suggest that this region of conserved hydropathy may commence with a high proportion of β-helix but that it terminates with a high proportion of α-helix.
procedure of Hopp & Woods (1981) was used with a window of 20 amino acids. Large arrows indicate the hydrophobic regions which flank the region of similar hydropathy. Arrowheads indicate additional hydrophobic regions within this general area.

proportion of \( \beta \)-sheet and reverse turn. We interpret these data taken together to suggest that these proteins may have a similar folded structure over the region of similar hydropathy. We have therefore aligned the amino acid sequences of representative N proteins in this region by means of the program CLUSTAL (Higgins & Sharp, 1988; Fig. 7).

Matrix comparisons (Fig. 5) suggest that Sendai virus and VSV N proteins are similar to PVM N in a region in the carboxy-terminal half of the PVM sequence (Fig. 7, box C). This region is also identified in DIAGON comparisons between Ebola virus and PVM with a window smaller than 99, which was used throughout Fig. 5 for consistency (not shown). Two short regions of similarity have also been detected within the sequences, each composed of hydrophobic regions interrupted by a single, conserved, basic amino acid (K or R: boxes A and B in Fig. 7). Corresponding regions are similarly spaced in other paramyxovirus and rhabdovirus N protein sequences (not shown). The levels of identity in these regions are described in Table 1. Although the highest proportions of identities are seen within a virus family, high levels of similarity are demonstrable in these regions between virus families when conservative substitutions are taken into account.

The parts of the N proteins located on the amino-terminal side of the regions aligned in Fig. 7 are similar in size and hydropathy (compare Fig. 6, b and d) but less similar in predicted secondary structure or amino acid sequence. In sharp contrast, the parts of the N proteins located on the carboxy-terminal side of the regions aligned in Fig. 7 vary greatly in size. Most of the sequences continue with a predicted \( \alpha \)-helix, however, and faint similarities between some of these sequences in this region are sufficient to continue the diagonals seen in Fig. 5 further towards the carboxy termini than the regions analysed in Table 1.

The variable carboxy termini of paramyxovirus N proteins may interact with the virus matrix protein during virus assembly and the amino termini may interact with the genomic RNA (Mountcastle et al., 1974; Rima, 1989). For Sendai virus, a fragment of
We thank Roger Ling for PVM-infected cell RNA, Dr B. Rima for providing paramyxovirus nucleocapsid protein sequence alignments and Dr J. Hoch for supplying the expression vector pKQV4. This work was funded by Medical Research Council grant G8514414 CA and a Wellcome Trust travel fellowship to A.J.E.

Table 1. Comparison of homologous regions of various virus N proteins

<table>
<thead>
<tr>
<th>Region* (length)</th>
<th>Match†</th>
<th>Number of identities</th>
<th>No. of identities plus homologies‡ (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A (13)</td>
<td>PVM/SEN</td>
<td>2</td>
<td>9 (69)</td>
</tr>
<tr>
<td></td>
<td>PVM/EBO</td>
<td>5</td>
<td>9 (69)</td>
</tr>
<tr>
<td>B (13)</td>
<td>PVM/SEN</td>
<td>2</td>
<td>8 (62)</td>
</tr>
<tr>
<td></td>
<td>PVM/VSV</td>
<td>2</td>
<td>8 (62)</td>
</tr>
<tr>
<td></td>
<td>PVM/EO</td>
<td>1</td>
<td>8 (62)</td>
</tr>
<tr>
<td></td>
<td>PVM/VSV</td>
<td>1</td>
<td>6 (46)</td>
</tr>
<tr>
<td>C (50)</td>
<td>PVM/SEN</td>
<td>9</td>
<td>25 (50)</td>
</tr>
<tr>
<td></td>
<td>PVM/EBO</td>
<td>11</td>
<td>27 (54)</td>
</tr>
<tr>
<td></td>
<td>PVM/VSV</td>
<td>8</td>
<td>20 (40)</td>
</tr>
<tr>
<td></td>
<td>SEN/EBO</td>
<td>12</td>
<td>29 (58)</td>
</tr>
<tr>
<td></td>
<td>PVM/RS</td>
<td>44</td>
<td>46 (92)</td>
</tr>
<tr>
<td></td>
<td>SEN/MEA</td>
<td>19</td>
<td>32 (64)</td>
</tr>
<tr>
<td></td>
<td>VSV/RAB</td>
<td>17</td>
<td>225 (52)</td>
</tr>
<tr>
<td></td>
<td>PVM/SEN/EBO</td>
<td>5</td>
<td>18 (36)</td>
</tr>
<tr>
<td></td>
<td>PVM/SEN/EBO/VSV</td>
<td>0</td>
<td>10 (20)</td>
</tr>
<tr>
<td>All sequences</td>
<td>0</td>
<td>9 (18)</td>
<td></td>
</tr>
</tbody>
</table>

* A, B and C refer to boxes A, B and C in Fig.7.
† Virus abbreviations: Sendai virus (SEN); Ebola virus (EBO); measles virus (MEA); rabies virus (RAB).
‡ Homologous amino acid groups are: aliphatic (I, L, V, M and A); small neutral (G, P, S, T and A); acidic/amino (D, E, Q and N); aromatic (F, Y, W and H); basic (K and R); thiol (C).
§ Percentage identity between VSV and RAB does not consider gaps necessary to match with other sequences.

approximate M, 12000 can be removed from the N protein by trypsin digestion without affecting the integrity of the nucleocapsid structure (Heggeness et al., 1980). Similarly, for simian virus 5 (SV5) treatment with protease can reduce the M, of the N protein from 60000 to 43000 without removing the protein from nucleocapsid structures (Mountcastle et al., 1974). As the full-length and truncated N proteins of SV5 are both amino-terminally blocked, the removed material is likely to be the carboxy terminus. The pneumovirus N proteins are similar in size to these protease-cleaved forms of the paramyxovirus N proteins. The carboxy-terminal part of the N protein of Ebola virus, beyond the regions of identity that we have identified, is larger than those of the paramyxoviruses, so it is possible that the precise function of this region varies among groups of viruses.

We suggest that approximately the first 350 to 400 amino acids of these virus N proteins may have retained the same structure, that of their putative common ancestor, as they have diverged in sequence. These regions of conserved structure would most likely have the common functions of RNA binding and the protein/protein interactions which form the virus nucleocapsid.

References


(Received 18 September 1990; Accepted 6 December 1990)