Nucleotide sequence analysis of the large (L) genes of phocine distemper virus and canine distemper virus (corrected sequence)

M. A. Mclhatton, M. D. Curran and B. K. Rima

School of Biology and Biochemistry, The Queen's University of Belfast, Belfast BT9 7BL, Northern Ireland, UK

This paper corrects the previously published sequence of the L gene of canine distemper virus (CDV). Errors in the published sequence (M. S. Sidhu et al., 1993, Virology 193, 50–65) led to frame shifts between residues 1021–1032, 1190–1219 and 1645–1650; a deletion of 21 amino acids between residues 1684–1705, and a single residue deletion at residue 1478. Residue 237 is now found to be glycine rather than tryptophan and residue 1626 proline instead of threonine. The sequence of the L gene of phocine distemper virus (PDV) was also determined. Alignment of the morbillivirus L proteins showed that PDV and CDV are more closely related to each other than to rinderpest virus and measles virus. Two regions of low identity are proposed to function as hinge regions between three highly conserved domains (I–III) in the morbillivirus L proteins. New sequence motifs have been identified on the basis of conservation in the morbilliviruses and the Paramyxovirinae.

The morbilliviruses form a serologically cross-reactive closely related genus in the family Paramyxoviridae with each species having a different host range. Before 1988, only four species were recognized; namely the human measles virus (MV), rinderpest virus (RPV), which infects large ruminants and artiodactyls, peste-des-petits ruminants virus (PPRV), which infects small ruminants, and canine distemper virus (CDV), which affects all terrestrial carnivores (Appel et al., 1981). Since 1988, new morbilliviruses have been found which have their own host ranges (Osterhaus & Vedder, 1988; Domingo et al., 1990). Phocine distemper virus (PDV; Cosby et al., 1988) infects primarily pinnipeds and a second has been found to infect cetaceans (Kennedy et al., 1988; Domingo et al., 1990) such as dolphins (dolphin morbillivirus or DMV) and porpoises (porpoise morbillivirus or PMV). Recently, a paramyxovirus has been isolated from horses which does not cross-react with the other morbilliviruses, but which has a low level of sequence identity (Gould, 1996).

Morbilliviruses contain six structural proteins (reviewed in Barrett et al., 1991). Three internal proteins are associated with the negative-strand genome: the nucleocapsid protein (N), phosphoprotein (P) and large protein (L). On the inner leaflet of the envelope there is a matrix protein (M) which probably plays a role in budding and the structural integrity of the virus. Protruding from the envelope are two glycoproteins, haemagglutinin (H) protein and fusion (F) protein. Two additional non-structural proteins are generated from the P gene either by translation of an overlapping ORF for the C protein or, by insertion of an extra G residue in the P mRNA, mRNA is generated for the V protein, which is amino-coterminal with P protein but which terminates in 70 residues C-terminal with a zinc-finger motif unique to this protein. The genes for the N, P/V/C, M, F and L proteins are ordered consecutively 3’ to 5’ on a negative-stranded genomic template of about 15–15·9 kb (reviewed in Barrett et al., 1991).

The morbilliviruses are a unique set of closely related viruses in which comparative analysis of the protein sequences can give information about the extent to which functionally related proteins can diverge. Genome sequences of morbilliviruses have often been completed with the publication of the L gene, which is the largest. The deduced amino acid sequence of the L protein of MV has been reported for the vaccine Edmonston strain by Blumberg et al. (1988). The genome sequence of RPV has been completed recently by Baron & Barrett (1995), and the sequence of CDV was completed by the analysis of the L protein gene and the genome termini by Sidhu et al. (1993). However, alignments of the published CDV L protein with those of paramyxoviruses, MV and RPV indicated that its sequence may have contained errors. In some areas the +1 and −1 frames aligned better with the other L proteins. This prompted us to re-analyse the sequence of the L gene of...
Table 1. Alignment of 5' UTRs of morbilliviruses and lengths of L mRNAs

<table>
<thead>
<tr>
<th>Virus</th>
<th>Length of 5' UTR</th>
<th>5' UTR sequence</th>
<th>Length of coding sequence</th>
<th>Length of 3' UTR</th>
<th>Total length</th>
</tr>
</thead>
<tbody>
<tr>
<td>MV</td>
<td>22</td>
<td>AGGGUCAGAACCUUGCUCGTTTAUGG</td>
<td>6552</td>
<td>69</td>
<td>6643</td>
</tr>
<tr>
<td>RPV</td>
<td>22</td>
<td>AGGGUCAGAACGUCGUCGUACCAGAG</td>
<td>6555</td>
<td>65</td>
<td>6642</td>
</tr>
<tr>
<td>CDV</td>
<td>22</td>
<td>AGGAUCCAGAAACCUUGCUCGAGAUGG</td>
<td>6555</td>
<td>65</td>
<td>6642</td>
</tr>
<tr>
<td>PDV</td>
<td>22</td>
<td>AGGAUCCCAGGAAUCCGUGGAGAUGG</td>
<td>6555</td>
<td>65</td>
<td>6642</td>
</tr>
<tr>
<td>DMV*</td>
<td>22</td>
<td>AGGGACCAGGAUUAUGCAGAG</td>
<td>6555</td>
<td>65</td>
<td>6642</td>
</tr>
<tr>
<td>Consensus</td>
<td></td>
<td>AGGRNCCARGNNNNNYYYNRRRAUGG</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* Reference Blixenkrone-Möller et al. (1996).

the Onderstepoort strain of CDV. Here we compare a corrected sequence with the L gene sequence of the closely related morbillivirus PDV and those of MV and RPV.

The sequence of the L gene of the Onderstepoort strain of CDV was determined by direct sequencing of PCR amplicons derived from RT–PCR of total CDV-infected Vero cell RNA with primers flanking the ambiguous sequences. All sequences were determined in both directions on at least two individual RT–PCR reactions by direct PCR cycle sequencing using standard methodology. Amino acid residue 237 was found to be glycine, conserved in all morbillivirus L proteins, rather than a tryptophan; residue 1626 was proline, conserved in all the L proteins, instead of threonine. Frame shifts appeared to have occurred in the originally published sequence between residues 1021–1032, 1190–1219 and 1645–1650. A deletion of 21 nucleotides in length, the same as that found for CDV and one

The alignment of the L proteins of morbilliviruses (Fig. 1) reveals two areas (boxed) of low identity between the various morbilliviruses. They may form domain boundaries in the L protein which due to its size is unlikely to be a single globular protein. The first boxed area extends from residues 607–650, the second from 1695–1717. Preliminary studies on the L genes of DMV and PMV support the delineation of these regions. These may be hinge regions between three domains conserved in the L protein, designated here as I, II and III.

Hinge region I (607–650) was previously identified by Poch et al. (1990) as a region of restricted identity and variable length in the L proteins of two paramyxoviruses (Sendai virus and Newcastle disease virus) and two rhabdoviruses (rabies virus and vesicular stomatitis virus). Hinge I has also been found to be variable in a comparison of the L proteins of five different MV strains (Komase et al., 1995). The second hinge between residues 1695–1712 has not been identified before.

In domain I (residues 1–606) Poch et al. (1990) identified two blocks of conserved sequence. In the first (residues
Fig. 1. The L proteins of four morbilliviruses are aligned from residue 1–2184. An asterisk (*) below the alignment indicates a residue conserved throughout the *Paramyxovirinae*; a dash (–) indicates a position at which only conservative amino acid replacements occur in the subfamily. Underlined residues in the MV sequence differ between strains (Komase et al., 1995).
M. A. McIlhatton, M. D. Curran and B. K. Rima

The alignment of the Paramyxovirinae L proteins was made including Newcastle disease virus (accession no. P11205), human parainfluenza virus type 2 (P26676), mumps virus (P30929), SV5 (M81721), Sendai virus (P06447) and human parainfluenza virus type 3 (P12577), as well as the morbilliviruses (MV) (the consensus sequence from Komase et al., 1995), RPV (Baron & Barrett, 1995), CDV and PDV (this work).
Table 2. Percentage identities between the L genes and proteins of morbilliviruses

Nucleotide identities between the L genes are given in the top right-hand corner of the table; amino acid identities between the L proteins are given in the bottom left-hand corner.

<table>
<thead>
<tr>
<th></th>
<th>MV</th>
<th>RPV</th>
<th>CDV</th>
<th>PDV</th>
</tr>
</thead>
<tbody>
<tr>
<td>MV</td>
<td>100</td>
<td>72:6</td>
<td>65:6</td>
<td>65:3</td>
</tr>
<tr>
<td>RPV</td>
<td>81:5</td>
<td>100</td>
<td>65:2</td>
<td>66:0</td>
</tr>
<tr>
<td>CDV</td>
<td>71:4</td>
<td>72:9</td>
<td>100</td>
<td>78:8</td>
</tr>
<tr>
<td>PDV</td>
<td>71:9</td>
<td>72:9</td>
<td>89:7</td>
<td>100</td>
</tr>
</tbody>
</table>

220–410), no known sequence motif was identified. In the second (residues 495–602) an RNA binding motif, KEKxxRLxxKMxxKM, was identified at position 535–549. This motif is present in all the *Paramyxovirinae* and also in the *Filoviridae* L proteins (Mühlberger *et al.*, 1992). Our analysis of the morbilliviruses shows that there is as much conserved sequence between these blocks as within and thus does not support the division of the L protein into the blocks identified earlier. Domain I starts with a conserved sequence motif near the N terminus, NQILYPEVHLDSPIVTNK. Homologues of this are also found at the N termini of the L proteins of the *Filoviridae* Marburg virus (Mühlberger *et al.*, 1992) and Ebola virus. There are also weak but significant identities with parts of the 3D protein of picornaviruses, i.e. after residue 1912 in Coxsackie B4 virus, residue 1930 in Coxsackie A9 and residue 1922 in the single ORF encoding the polyprotein of echoviruses. The 3D protein of the picornaviruses also acts as an RNA polymerase. A second conserved sequence, IINGYRDHRGGSWPP, in this domain appeared to be present only in the paramyxovirus L proteins and no identities with other proteins in the databank were found. At position 357–359 the tripeptide GHP is conserved. This occurs very infrequently in the protein databank and the histidine residue is positioned so that it may be involved in enzymatic activity (Poch *et al.*, 1990).

In domain II (residues 650–1694) Poch *et al.* (1990) identified three blocks of conserved sequence. Again, the analysis of the morbillivirus alignment does not support the separation into three blocks as homologous regions and motifs are dispersed over the whole of domain II. For example, the sequence ALIGGDDDDINSFITEFL, which is strictly conserved among the morbilliviruses and which contains the GDDD motif associated with RNA polymerase activity, lies outside the blocks identified by Poch *et al.* (1990). Significant conserved motifs are found in this domain at positions 771–775, i.e. the QGDNQ motif flanked with hydrophobic residues associated with many RNA-dependent RNA polymerases (Poch *et al.*, 1990). This is preceded by the GGIEGxCQKLWTLI motif (residues 737–749), previously identified by Blumberg *et al.* (1988), which has homologues in the L proteins of the other *Mononegavirales* such as human respiratory syncytial virus (Stec *et al.*, 1991), Marburg virus (Mühlberger *et al.*, 1991) and Ebola virus (accession number U23458) and at position 651 in the L gene of infectious haematopoietic necrosis virus (Morzunov *et al.*, 1995). These motifs are embedded in a larger block (659–840) which starts with a motif of 15 residues, FoTTDLxKYCLNWRYE (ø indicates a small hydrophobic residue), and ends with the motif IGHxLKxxETioSxxFYxKxYxDG (residues 815–840). The former is present only in *Paramyxovirinae* L proteins. The latter has identity to motifs present in the L protein of Marburg virus (Mühlberger *et al.*, 1992). No functions have been proposed for either motif. Two other conserved sequence motifs in domain II, LRVPYIGSTTDER (residues 1217–1229) and SRLFVRNIGDPVTSSIADLKRM (residues 946–967), were found only in the L proteins of the *Paramyxovirinae*.

In domain III (residues 1717–2183) no long conserved sequence motifs are apparent and the percentage identity between the morbilliviruses is also very low. Poch *et al.* (1990) suggested that the motif around residues 1785–1799 may form part of an ATP binding site consisting of the sequence GxGxG followed by a lysine rich sequence and the conserved motif F/Y-Y-N, which is prevalent in protein kinases (Poch *et al.*, 1990). This motif is conserved in the paramyxovirus and filovirus L proteins. Its functional role in the L protein is not clear.

In conclusion, comparative analysis of the L proteins of four morbilliviruses indicates that there are three conserved domains. The first may have an RNA binding function, the second is probably the main domain with replicative and or transcripitive activity and the third domain may be acting as another nucleotide binding domain, if not as a protein kinase. A number of new conserved motifs present in all *Paramyxovirinae* have been identified. Two new motifs also present in the filoviruses further strengthen the suggested close evolutionary relationship between these and the paramyxoviruses (Mühlberger *et al.*, 1992). Interestingly, the implied three-domain structure of the morbillivirus L proteins may be functionally similar to the three-protein replication machinery of the orthomyxoviruses. However, apart from the recognized polymerase motif no others were found to be conserved between ortho- and paramyxoviruses.

We thank the National Environmental Research Council for support for this research (GR9/0938). We also thank the Wellcome Trust for supplying an automated sequencer to our laboratory and a specific project grant (42454/Z/94) to cover this work.

**References**


Received 20 September 1996; Accepted 15 November 1996