Nucleotide sequence comparisons of the fusion protein gene from virulent and attenuated strains of rinderpest virus

S. A. Evans, M. D. Baron, R. W. Chamberlain, L. Goatley and T. Barrett*

AFRC Institute for Animal Health, Pirbright Laboratory, Pirbright, Woking, Surrey GU24 0NF, U.K.

We have cloned and sequenced the entire fusion (F) protein gene of the RBOK vaccine strain of rinderpest virus and the coding regions for the F genes of two mild field isolates of the virus from Africa. Analysis of the nucleotide and the predicted amino acid sequences showed that the vaccine virus was more than 99% identical in the protein coding region to the virulent Kabete O strain from which it was derived, whereas the field isolates differed by 10 to 12% from each other and from the vaccine strain. No changes were found in the F protein which could explain attenuation of the vaccine; however, each of the mild field isolates had amino acid changes in important functional areas which may be related to their attenuated phenotype.

Rinderpest virus (RPV) is a member of the genus Morbillivirus in the family Paramyxoviridae. The morbilliviruses are a group of antigenically related viruses each distinguished by a uniquely restricted mammalian host range, e.g. RPV mainly causes disease in large ruminants and most often infects domestic cattle and buffalo (Scott, 1959). Measles virus (MV) infects man and other primates; dogs and other wild carnivores are infected by canine distemper virus (CDV); sheep, goats and other small ruminants by peste-des-petits-ruminants virus (PPRV); seals by phocid distemper virus (PDV) and dolphins and porpoises by other recently described morbilliviruses (Barrett et al., 1993; Visser et al., 1993). There is only one serotype of each virus, yet despite their antigenic stability, different morbillivirus isolates can exhibit different pathogenicities in the same host species and the factors which determine this variation remain unknown (Plowright, 1963; Taylor, 1986).

The F (fusion) protein of all paramyxoviruses is responsible for fusion of the virus envelope with the host cell membrane (reviewed by Lamb, 1993) and also causes cell-to-cell fusion during the later stages of infection with the formation of syncytia (a primary cytopathic effect) allowing the virus to spread from cell to cell even in the presence of anti-H neutralizing antibodies (Choppin & Scheid, 1980; Norrby et al., 1975). The F protein is synthesized as an inactive precursor (F₀) which is subsequently cleaved by endogenous cell proteases to the biologically active form, a disulphide-linked heterodimer composed of the F₁ and F₂ subunits (Hardwick & Bussell, 1978). The F protein is an important determinant of pathogenicity in Newcastle disease virus (Glickman et al., 1988; Collins et al., 1993; Morrison et al., 1993) and so it was of interest to compare the sequence of this gene in the attenuated vaccine strain with that of the virulent parental virus and with the F gene of field isolates of the virus. In this report we present the complete nucleotide sequence of the F protein gene of the RPV vaccine strain (Plowright & Ferris, 1962) and compare it with the published sequence of the virulent parent (Hsu et al., 1988). The F protein coding regions of two more recent virus isolates (RBT-1 and Egypt/84), which are of moderate virulence in cattle (Taylor, 1986), were also determined and compared with these sequences.

A recombinant plasmid (p6/35-2) containing a complete cDNA copy of the RPV F gene was obtained by conventional cloning methods using EcoRI linkers and an EcoRI-cut pUC13 vector. Recombinant plasmids were selected from the cDNA library (produced using oligo d(T)-primed total polyadenylated RNA from infected Vero cells) by differential hybridization with labelled RNA from infected and uninfected cells. The F-specific clone was identified by its specific hybridization to the correct size mRNA in Northern blots (Barrett & Underwood, 1985). Another F-specific clone (L2) was subsequently isolated from a Agt11 cDNA library which had been prepared from a different batch of infected cell mRNA (Baron et al., 1994). Both p6/35-2 and L2 were found to be derived from M–F bicistronic messages. The complete F cDNA portion of each was cloned into...
were derived from bicistronic M-F mRNAs they provided sequence data for the intergenic region and the plasmid p6/35-2 was sequenced on each strand using digestion to produce nested deletions using the Erase-a-pKS(+) prior to sequencing. The two separate RPV F clones were sequenced by subcloning restriction digests assembled into a contiguous stretch of nucleotide only in three nucleotides at positions 850, 926 and 2208. (RPV-R) F cDNA clones were almost identical, differing sequence spanning the entire RPV F gene; the nucleotide (13)

\[
\begin{align*}
\text{RPPS} & \quad \text{GAC} \\
\text{FAC} & \quad \text{CTT}
\end{align*}
\]

The accumulated data from the above strategies were cleavage points for the signal peptide and F1/F2 junction are indicated by a grey and a black arrow respectively. The hydrophobic signal; fusion and trans-membrane regions; and cleavage sites for generating the Fx/F2 proteins and for the signal peptidase.

The MV and RPV F mRNAs, including the poly(A) tails, are approximately 2400 nucleotides long (Barrett & Underwood, 1985; Tsukiyama et al., 1988), which corresponds closely to the size of the F gene sequence reported for MV (2377 bases) by Richardson et al. (1986), for the lapinized vaccine strain of RPV (2359 bases) by Tsukiyama et al. (1988), and for the RBOK vaccine strain of RPV reported here (2366 bases). The sequence is shorter than the 2620 nucleotides reported for the F protein mRNA of the virulent parental virus vaccine strain of RPV reported here (2366 bases). These authors failed to recognize the conserved intergenic triplet at positions 252 to 254 and the semi-conserved morbillivirus stop-start signals at the end of the M gene and beginning of the F gene; the first 251 nucleotides of this sequence are derived from the

pKS(+) prior to sequencing. The two separate RPV F clones were sequenced by subcloning restriction digests of the gene into M13 and by direct sequencing of double stranded plasmid DNA, following exonuclease III digestion to produce nested deletions using the Erase-a-Base kit (Promega). The whole F gene cDNA from plasmid p6/35-2 was sequenced on each strand using M13-cloned restriction digests and on one strand of the L2 clone using exonuclease digestion. Since both clones were derived from bicistronic M-F mRNAs they provided sequence data for the intergenic region and the start of the F gene. The junction between the F and H genes was amplified by PCR from total infected cell RNA, and from these sequence data we determined the number of A bases at the 3' terminus of the gene.

The accumulated data from the above strategies were assembled into a contiguous stretch of nucleotide sequence spanning the entire RPV F gene; the nucleotide and deduced amino acid sequences are shown in Fig. 1. The nucleotide sequences from the two RBOK vaccine (RPV-R) F cDNA clones were almost identical, differing only in three nucleotides at positions 850, 926 and 2208.

In each case the consensus sequence was determined either by PCR from genomic RNA or by isolation of further cDNA clones. Features indicated in Fig. 1 include: the hydrophobic signal; fusion and trans-membrane regions; and cleavage sites for generating the Fx/F2 proteins and for the signal peptidase.

Fig. 1. Nucleotide sequence of the F gene of RPV. The entire sequence of the gene from the vaccine RBOK virus is shown in positive sense cDNA format. The amino acids encoded by a single large ORF are indicated using the single-letter notation. The two ATG codons are highlighted in bold. Three hydrophobic regions are shown: the potential signal sequence in a black bordered box; the fusion peptide sequence in a grey-filled box; the transmembrane domain in a black-filled box. Potential glycosylation sites are underlined and cleavage points for generating the Fx/F2 proteins and for the signal peptidase.
preceding M gene (Baron et al., 1994). When this is taken into account the two F genes differ by 20 nucleotides, including five gaps, in the GC-rich 5′ region (nucleotides 1 to 588) and by only three nucleotides, leading to one amino acid change, in the protein coding region of the gene. It is important to note that a difference (GC instead of CG at positions 1717 to 1718) was found between the published Kabete O F gene sequence and that obtained from the EMBL database (M21514); we have used the database version for all the sequence comparisons reported here. In the published sequence (Hsu et al., 1988) there is only one nucleotide change, compared to the vaccine in the F coding region, which does not result in an amino acid difference.

The codon from which translation is initiated is not clear. In the MV F gene, translation probably begins at one of two closely juxtaposed AUG codons at positions 574 and 583 (Buckland et al., 1987; Richardson et al., 1986). In the published sequences for the RPV Kabete O (RPV-K) F gene (Hsu et al., 1988) and that of the lapinized vaccine strain of RPV (RPV-L) (Tsukiyama et al., 1988) the first AUGs are found at different positions (316 and 151 respectively) and these codons are not in-frame with the main open reading frame (ORF) encoding the F protein. Translation of this must begin at AUG codons at positions 584 (RPV-K) or 587 (RPV-L), which is similar to the start position for the F gene sequence in MV. These codons are also in a more favourable context for translation initiation (Kozak, 1983, 1986). Clearly, some mechanisms must operate either to allow re-initiation of translation at the second AUG or the skipping of the first AUG. In RPV-R, however, the upstream AUG is in-frame with the main ORF and translation could conceivably begin here, resulting in an amino-terminal extension which would be cleaved off with the signal sequence. Alternatively, the same mechanism that allows translation initiation at a downstream AUG in RPV-K and RPV-L might direct translation to start from the second AUG in RPV-R. We have previously shown that deletion of the first AUG in RPV-R does not prevent the synthesis of apparently normal F protein both in vitro and in vivo (Evans et al., 1990), showing that this region is not essential for F protein synthesis and membrane insertion. The amino acid sequence between the first and second AUGs is therefore shown in italics in Fig. 1.

The long untranslated region (UTR) at the 5′ end of the F mRNA of the morbilliviruses is unique among the paramyxoviruses whose F protein gene nucleotide sequences have been reported to date. The matrix (M) gene of morbilliviruses precedes the fusion gene on the viral genome and nucleotide sequence analysis of the M gene of MV and CDV (Bellini et al., 1986) and RPV-R and RPV-K (Baron et al., 1994) revealed a stretch of approximately 400 nucleotides of non-coding sequence at the 3′ end of the M mRNA. Therefore, approximately 1000 nucleotides of apparently untranslated sequence exists between the end of the M coding region and the start of the F coding region. There is very high sequence conservation in the coding regions of the M and F genes between different morbilliviruses but the nucleotide sequences of the 5′ UTRs and 3′ UTRs of the mRNAs show little or no sequence identity (Fig. 2). It is probable that these sequences play some as yet unknown role in the virus replication strategy. Despite a lack of nucleotide sequence identity between these regions of the F mRNAs there may be conserved secondary structures present which could influence gene transcription and/or F protein translation. It has been noted in the past that an increase in secondary structure in the 5′ UTR of eukaryotic mRNA can reduce its translational efficiency (Pelletier & Sonenberg, 1985). In contrast, the role of the 5′ UTR of the F gene appears to enhance translation of the F protein, since removal of this sequence from RPV and CDV F genes results in a decrease in F protein

![Fig. 2. Nucleotide sequence comparison of the RBOK F gene with F genes from other morbilliviruses. The histogram shows sequence identity between the different virus F CDNAS for the 5′ UTR, the F ORF and the 3′ UTR. Sequence data obtained from the EMBL database included: RPV Kabete O (K, M21514); RPV lapinized strain (L, M20870); various strains of measles virus - Edmonston (MVED, K01711), IPSCA (X16566), measles inclusion body encephalitis (MIBE, X16569) virus and a subacute sclerosing panencephalitis (SSPE, X16567) virus; plus phocid and canine distemper viruses (PDV, D10371 and CDV, M21849). The GCG DNADISTANCES program was used to establish the sequence identity.

```
  5′ UTR  F coding  3′ UTR
RPV RBOK

Fig. 2
```
production in cells infected with vaccinia virus recombinants containing these truncated genes (Evans et al., 1990). However, it has not been possible to demonstrate internal ribosome binding by these 5' regions of RPV and CDV (Evans, 1991).

The ORF starting at the second AUG encodes 546 amino acids with a predicted $M_r$ of 58735 for the unglycosylated protein, which is identical to that predicted for the virulent parent virus if the corresponding AUG is used as the initiation codon. There is a potential hydrophobic signal sequence of 19 amino acids at the amino terminus and another hydrophobic region located near the carboxyl terminus (which probably acts as a membrane anchor), leaving a cytoplasmic domain of 33 amino acids. There is a third hydrophobic region located near the amino terminus of the polypeptide which exhibits a high degree of sequence conservation between all the paramyxoviruses, a feature commensurate with its proposed fusion function (reviewed by Morrison & Portner, 1991). This hydrophobic region immediately follows the double basic amino acid arrangement at the cleavage site (residues 85 to 89) necessary for cleavage by host trypsin-like proteases (Scheid & Choppin, 1974, 1976), which leaves the fusion peptide at the amino terminus of $F_1$.

To investigate whether there were any significant differences between the F protein of the virulent Kabete O virus and recent less virulent field isolates of RPV, we cloned and sequenced the coding region of the F gene from a mild (Tanzania/62; RPV-T) and an avirulent (Egypt/84; RPV-E) strain of RPV (Plowright, 1963; Taylor, 1986). Amplification via RT/PCR was carried out on total RNA from infected Vero cells using message sense primer F7 (nucleotides 528 to 552) and genome sense primer F8 (nucleotides 2218 to 2238). The resulting DNA products were purified by agarose gel electrophoresis and cloned into pBluescript KS(+) at the SmaI site by blunt end ligation. Two independently derived clones from each virus were sequenced to check for identical position to the first AUG in the RPV-R. This change, we prepared a new infected cell RNA sample and repeated the PCR reactions in this region with the same result. The implications of this for F gene expression in this strain and its effect on the pathogenicity of the virus have yet to be determined but it could clearly affect the efficiency of membrane insertion of the F protein. This finding is particularly important given that the RPV-R F protein may also be expressed from the upstream AUG. Although as yet there is no evidence for the use of non-AUG codons in morbilliviruses, numerous non-AUG codons (including AUA) differing from AUG by a single nucleotide have been shown to initiate protein translation in vitro and in vivo (Peabody, 1989). Hence, translation of RPV-E F protein from the AUA at position 589 cannot be ruled out.

All the morbillivirus F protein sequences were aligned from the predicted N-terminal amino acid of the $F_2$ protein, assuming the hydrophobic signal peptide is cleaved by the signal peptidase during translocation of the nascent polypeptide across the rough endoplasmic reticulum (Fig. 3). Comparisons of the F gene sequences from all the morbilliviruses revealed a difference of 10 to 12% between each of the RPV isolates (with the exception of RPV-R and RPV-K). There is a closer relationship with MV than with any other morbillivirus, consistent with RPV being the progenitor of MV (Norrby et al., 1985). There were three potential asparagine-linked glycosylation sites in the RPV $F_2$ polypeptides that are conserved within the genus, although CDV (Barrett et al., 1987) and PDV (Kövamees et al., 1991) have another potential site in $F_1$ at amino acids 382 to 384. Amino acid residues that could be involved in the folding of the fusion protein are also well conserved. For instance, the cysteine residues are conserved within the genus, with the exception of one in both the RPV-E (position 452) and RPV-T (position 339) strains and those found in the signal sequence and transmembrane regions where extra cysteine residues are found in the case of MV, CDV and PDV. Also, over 50% of the proline and glycine residues are maintained in all the morbilliviruses. A series of conserved leucine residues, each separated by five to eight variable residues, are positioned at the base of the external domain of the protein prior to the transmembrane anchor (amino acids 434 to 459). This motif could possibly act as a leucine zipper (Buckland & Wild, 1989), suggesting that the natural F protein forms a dimer. Although the hydrophobic transmembrane region and the first half of the cytoplasmic domain do not show sequence conservation between morbilliviruses, the cytoplasmic carboxy-terminal 15 amino acid residues are highly conserved in all morbilliviruses. However, we noted one conservative amino acid change (valine to isoleucine) in this region in the mild RPV-T isolate. The Japanese RPV-L vaccine strain has an arginine instead of a lysine residue at the beginning of this conserved region (Tsukiyama et al., 1988). This part of the F protein is involved in virus morphogenesis through its interaction with the M protein.
probably interfere with virus morphogenesis in these infections. (Parks & Lamb, 1990). The persistence of MV in the brain of patients with subacute sclerosing panencephalitis is associated with defective M protein in some cases (Cattaneo et al., 1986, 1988; Choppin et al., 1981) and with deletions in the cytoplasmic tail of the F protein in others (Cattaneo et al., 1989; Komase et al., 1990; Schmid et al., 1992). Alterations in the F and M proteins probably interfere with virus morphogenesis in these patients and explain the non-productive nature of these infections.

In Newcastle disease virus isolates, loss of basic amino acids at the cleavage site correlates with low virulence (Collins et al., 1993; Glickman et al., 1988) and similarly, in the case of avian influenzas, alterations in the cleavage site of the haemagglutinin protein correlate with pathogenicity (Li et al., 1990; McCauley, 1987; Wood et al., 1993). A similar comparison of the amino acid sequence around the F₁/F₂ protein cleavage site from a range of field isolates of RPV showed no alteration in the basic amino acid residues needed for cleavage by the endogenous proteases (Chamberlain et al., 1993). It is evident that factors which influence virulence will prove more elusive to determine in the case of RPV and other morbilliviruses. Even the highly conservative substitutions at the carboxy terminus of the F protein seen in the Tanzanian RPV isolate and the Japanese vaccine could be associated with virus attenuation, as could the genome may influence virulence, as in the case of polio

<table>
<thead>
<tr>
<th>Sequence</th>
<th>QIHW</th>
<th>NLS</th>
<th>IGV</th>
<th>YKVMT</th>
<th>SHQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>RPV-R</td>
<td>QIHW^</td>
<td>NLS</td>
<td>^IGV</td>
<td>YKVMT</td>
<td>SHQ^</td>
</tr>
<tr>
<td>RPV-K</td>
<td>QIHW</td>
<td>NLS</td>
<td>IGV</td>
<td>YKVMT</td>
<td>SHQ</td>
</tr>
<tr>
<td>RPV-T</td>
<td>QIHW</td>
<td>NLS</td>
<td>IGV</td>
<td>YKVMT</td>
<td>SHQ</td>
</tr>
<tr>
<td>RPV-L</td>
<td>QIHW</td>
<td>NLS</td>
<td>IGV</td>
<td>YKVMT</td>
<td>SHQ</td>
</tr>
<tr>
<td>MV</td>
<td>QIHW</td>
<td>NLS</td>
<td>IGV</td>
<td>YKVMT</td>
<td>SHQ</td>
</tr>
<tr>
<td>CDV</td>
<td>QIHW</td>
<td>NLS</td>
<td>IGV</td>
<td>YKVMT</td>
<td>SHQ</td>
</tr>
<tr>
<td>PDV</td>
<td>QIHW</td>
<td>NLS</td>
<td>IGV</td>
<td>YKVMT</td>
<td>SHQ</td>
</tr>
</tbody>
</table>

Fig 3. Alignment of morbillivirus F proteins. Protein sequences were translated from: MV (Richardson et al., 1986); CDV (Barrett et al., 1987); PDV (Kiyamae et al., 1991); RPV-K (Hsu et al., 1988); RPV-L (Tsukiyama et al., 1988); RPV-R, RPV-E (Egypt/84) and RPV-T (Tanzania/62) are the sequences presented in this paper. All sequences were aligned from the predicted amino-terminal amino acid of the F protein using the program PILEUP. In order to emphasize the regions of sequence variation, non-conserved residues are highlighted (white on black), conserved residues are shown as black on grey and identical residues are shown as normal text.
virus (Evans et al., 1985). Most of the proteins of the RBOK vaccine, except the L protein, have now been compared with those of the parental Kabete O strain and all are highly conserved (Baron et al., 1994; M. D. Baron & T. Barrett, unpublished). We are now carrying out a study of the extragenic sequences in the virulent virus to establish the extent of variation in these regions.

We thank Dr Paul Thomas for synthesizing the oligonucleotides and Miss Emma Hockley for excellent technical assistance. R. W. Chamberlain was supported by an AFRC studentship and the work was funded by the Wellcome Trust.

References


(Received 21 April 1994; Accepted 1 August 1994)