Sequence determination of the nucleocapsid protein gene of the porcine epidemic diarrhoea virus confirms that this virus is a coronavirus related to human coronavirus 229E and porcine transmissible gastroenteritis virus

Anne Bridgen,[1]† Mariela Duarte,[2] Kurt Tobler,[1] Hubert Laude[2] and Mathias Ackermann[1]*

[1] *Institut für Virologie, Veterinärmedizinische Fakultät, Universität Zürich, Winterthurerstrasse 266a, CH-8057 Zürich, Switzerland and* [2] *I.N.R.A., Unité de Virologie et Immunologie Moléculaires, CR de Jouy-en-Josas, 78352 Jouy-en-Josas Cedex, France*

The nucleotide sequence of 1·7 kbp cDNA, comprising the region nearest the 3′ end of the genome of the porcine epidemic diarrhoea virus (PEDV), has been independently determined for two European isolates of PEDV. Almost identical results were obtained for the two isolates, which were derived from cases of PEDV infection in Belgium and Britain in 1977 and 1987, respectively. The sequences contained a 1323 nucleotide (nt) open reading frame (ORF), which showed moderate identity to the nucleocapsid (N) gene of other coronaviruses. The greatest similarity at both the nucleic acid and protein levels was to the human coronavirus 229E. The PEDV N gene was, however, notably larger than that of the human 229E and porcine transmissible gastroenteritis viruses. This reflects the presence of a putative insertion of approximately 135 nt located towards the middle of the N gene. A second 336 nt ORF, which might encode a leucine-rich protein similar to, but shorter than, the bovine coronavirus internal protein was found within the PEDV N gene. Several RNA motifs typical of coronaviruses were also observed. These results confirm the earlier provisional classification of PEDV as a coronavirus.

Introduction

Porcine epidemic diarrhoea virus (PEDV) causes diarrhoea in pigs, particularly in neonates.The disease has been recognized for approximately 20 years, but the causative virus was first described only in 1978 (Pensaert & Debouck, 1978) and another 10 years elapsed before a method was developed for propagation of the virus in cell culture (Hofmann & Wyler, 1988). During this time, outbreaks of the disease have been reported from numerous European countries as well as Korea, China and Japan. The epidemiology and pathogenesis of the disease have been well described by Pensaert (1989). The biological behaviour, electron microscopic appearance and polypeptide structure of PEDV resulted in its provisional classification as a coronavirus (Egberink *et al.*, 1988; Hofmann & Wyler, 1989). This designation has not, until now, been confirmed since many groups have failed to detect any antigenic cross-reactivity between PEDV and other coronaviruses (Pensaert *et al.*, 1981; Kusanagi *et al.*, 1992). However, some cross-reactivity has been detected with feline infectious peritonitis virus (FIPV; Yaling *et al.*, 1988). Most recently, mink sera have been shown to react with the PEDV membrane (M) protein, as well as with transmissible gastroenteritis virus (TGEV) and canine coronavirus (CCV) M and nucleocapsid (N) proteins (Have *et al.*, 1992). In the study described here a part of the PEDV cDNA was sequenced in order to assess the nature of this virus.

The work described in this paper was carried out by two groups working with different European isolates of PEDV: the Zürich group used the Belgian CV777 isolate obtained from Dr M. B. Pensaert (Pensaert & Debouck, 1978), whereas the Paris group worked with the British 1/87 isolate (Br1/87) provided by Dr P. Have (Have *et al.*, 1992). Both these isolates were derived from acute outbreaks of the disease. The two groups, which each independently determined the sequence of one isolate and contributed equally to this work, used different techniques for the cDNA synthesis and screening. The Zürich group used degenerate primers with sequences based on conserved regions near the 3′ end of the coronavirus genome to amplify the cDNA by the PCR.

† Present address: Institute of Virology, Church Street, Glasgow G11 5JR, U.K.

The DNA sequence has been deposited with the GenEmbl databases with the EMBL accession number Z14976. The protein sequence has also been deposited in the SWISSPROT database.

At present the only reported examples of PCR amplification of coronaviral cDNA of unknown sequence have involved amplification with primers derived from closely related coronaviruses. For example, porcine respiratory coronavirus (PRCV), the sequence of which shows a 96 to 98% amino acid identity to that of TGEV (Britton *et al.*, 1991; Rasschaert *et al.*, 1990), was amplified using TGEV primers (Britton *et al.*, 1991), and feline enteric coronavirus (FECV) was amplified using FIPV primers (Vennema *et al.*, 1992). The approach used in the experiments described in this paper should, however, be applicable to the cloning of any coronaviral genome.

In contrast, the Paris group constructed a cDNA library in a phage λ expression vector and used PEDV polyclonal antisera to screen the library. Despite the differences in viral isolate and the experimental approach used to construct the cDNA clones, almost identical results were obtained by the two groups and so these are presented together.

Methods

Cloning of the Belgian CV777 isolate

Viral culture. The viral culture was performed as described elsewhere (Knuchel *et al.*, 1992; Hofmann & Wyler, 1988). Virions were isolated from cells disrupted by three cycles of freezing and thawing, cellular debris was removed by low speed centrifugation and virions were pelleted by centrifugation using a Beckman SW28 rotor at 100000 ***g*** at 4 °C for 2 h.

RNA isolation. RNA was isolated both from PEDV-infected Vero cells and from virions using guanidinium thiocyanate according to Kingston (1991). For the primer extension experiments poly(A) mRNA was purified from total RNA using the PolyATtract mRNA system (Promega).

Primer extension and cDNA synthesis. First strand cDNA was synthesized in two steps, comprising primer binding and extension (both of which were performed at 42 °C), according to Wirth *et al.* (1991). Poly(dT) (Pharmacia) or the degenerate primers P24 and P25(dT) used for the PCR reactions (see below) were used to prime the cDNA synthesis. The cDNA was extracted with (1:1) phenol:chloroform then chloroform prior to ethanol precipitation. Primer extension experiments were done in a similar manner, but the primers were labelled with [γ-^{33}P]ATP (Amersham) at their 5′ ends using polynucleotide kinase (New England Biolabs) prior to annealing to the poly(A)-selected RNA from PEDV-infected cells and extension. The extended products were phenol-extracted, heat-denatured and loaded onto a denaturing urea gel (as for the sequencing reactions). Radioactive markers for the primer extension were phosphatase-treated *Hin*fI digestion products of plasmid pGEM4, which were labelled with [γ-^{33}P]ATP as described above.

PCR amplification and cloning. The cDNA was amplified with the following degenerate primers: P23, 5′ AAGCTTTTACTA(C/T)-TT(A/G/T)GG(A/C/T)ACAGGACC 3′ (27mer, 18-fold degeneracy); P24, 5′ CTCGAGCGACCCAGA(A/C)GAC(A/T)CC(G/T)TC 3′ (25mer, eight-fold degeneracy); P25, 5′ GACTAGTTGGTGGAG-(A/T)TTTAA(C/T)CC(A/T)GA 3′ (27mer, eight-fold degeneracy), the sequences of which were based on conserved regions of coronaviral genomes (Bridgen *et al.*, 1993; Tobler *et al.*, 1993). In brief, P23 was based on the consensus peptide sequence FY(Y/F)LGTGP, P24 on the sequence (D/E)G(V/I)(V/F)WVA and P25 on the sequence S(W/F)WS(F/W)NPE. P25 was also tailed with T residues [designated P25(dT)] with terminal deoxynucleotidyl exotransferase (Boehringer) as recommended by the manufacturer. PCR amplifications were performed using a Hybaid Intelligent Heating Block (Model IHB 2024) and *Taq* polymerase from Perkin-Elmer Cetus. P24 and P25 were used together to amplify a 0·7 kbp DNA fragment using cDNA primed with oligo(dT) or with P24, and P23 and P25 were used to amplify a 1·6 kbp DNA fragment using P25(dT)-primed cDNA in a modification of the 3′ RACE (rapid amplification of cDNA ends) technique of Frohman *et al.* (1988). The P24/P25 product was amplified with 38 cycles of 50 s at 94 °C, 60 s at 48 °C, and 60 s at 72 °C, whereas that of P23/P25 was amplified with 40 cycles of 50 s at 94 °C, 60 s at 47 °C and 150 s at 72 °C. A final 5 min extension was made at 72 °C for both reactions. Digoxigenin dUTP-containing P32/P35 probe (P32 is 5′ ATCTTTAATTACTCGTGCAA 3′ and P35 is 5′ CAGTGTAG-TTGAGATTGTT 3′) was prepared by PCR amplification of cloned PEDV cDNA in the presence of this nucleotide. The PCR products were phosphorylated and cloned as blunt-ended fragments into digested and dephosphorylated pBluescript II KS$^+$ vector (Stratagene) using standard procedures (Sambrook *et al.*, 1989) as described by Tobler *et al.* (1993).

Sequencing. At least two clones from independent PCR reactions were sequenced on both strands using the Sequenase 2 kit (United States Biochemicals). Three sequence ambiguities were resolved by use of the Vent polymerase sequencing kit (New England Biolabs) or by replacing dGTP by dITP in the Sequenase reaction. Sequences were analysed with the IntelliGenetics PC/GENE program or with the University of Wisconsin Genetics Computer Group programs (UWGCG; Devereux *et al.*, 1984).

Cloning of the Br1/87 isolate

The methods used to obtain PEDV-specific clones from this isolate will be reported in more detail in a subsequent publication. Briefly, poly(A)-selected RNA prepared from virus-infected cells was used to construct a cDNA expression library in the λ Zap II vector (Stratagene) according to the manufacturer's instructions. The library was screened with an anti-PEDV hyperimmune serum obtained from C.N.E.V.A. (Laboratoire de Pathologie Porcine, Ploufragan, France). The anti-*Escherichia coli* activity of the serum was depleted before use. Subsequent screening was done by DNA hybridization using the first cDNA clone as the probe.

Shotgun DNA sequencing was performed using sonicated plasmid fragments from two cDNA clones (see Results) subcloned into *Sma*I-digested M13mp18 phage vector, such that the PEDV sequence was determined at least twice from each cDNA strand. Sequencing reactions were done using the Sequenase kit (United States Biochemicals) and the M13 universal primers. The resulting data were analysed using the Microgenie sequencing program (Queen & Korn, 1984) and the UWGCG programs (Devereux *et al.*, 1984).

Results

Amplification of the CV777 cDNA using degenerate primers

CV777 cDNA could be amplified using the three degenerate primers designed from conserved sequences in the coronaviral M (primer P25) and N genes (P23 and P24) (Bridgen *et al.*, 1993). In addition, the cDNA could be amplified from the poly(A) tail using the 3′ RACE technique of Frohman *et al.* (1988). This technique is

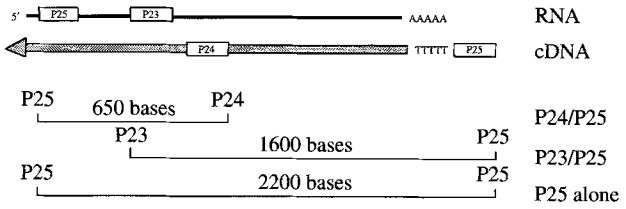

Fig. 1. Polymerase chain amplification of PEDV CV777 cDNA using the RACE technique of Frohman *et al.* (1988). Diagrammatic representation of the RACE products showing the viral RNA (upper strand) and first strand cDNA (lower strand). The three PCR products amplified from this cDNA (P24/P25, P23/P25 and P25) are also shown. The 5′ end of the N gene is located between P23 and P25.

illustrated in Fig. 1 and requires first strand cDNA synthesis to be initiated using a sequence-specific dT-containing primer in order to allow specific amplification from the poly(A) tail to the region of known or predicted sequence.

Derivation and characterization of the Br1/87 cDNA clones

From 3×10^5 plaques representing recombinant phage, 10% produced a signal with the antibody; of these 20 were purified to homogeneity with one or two successive

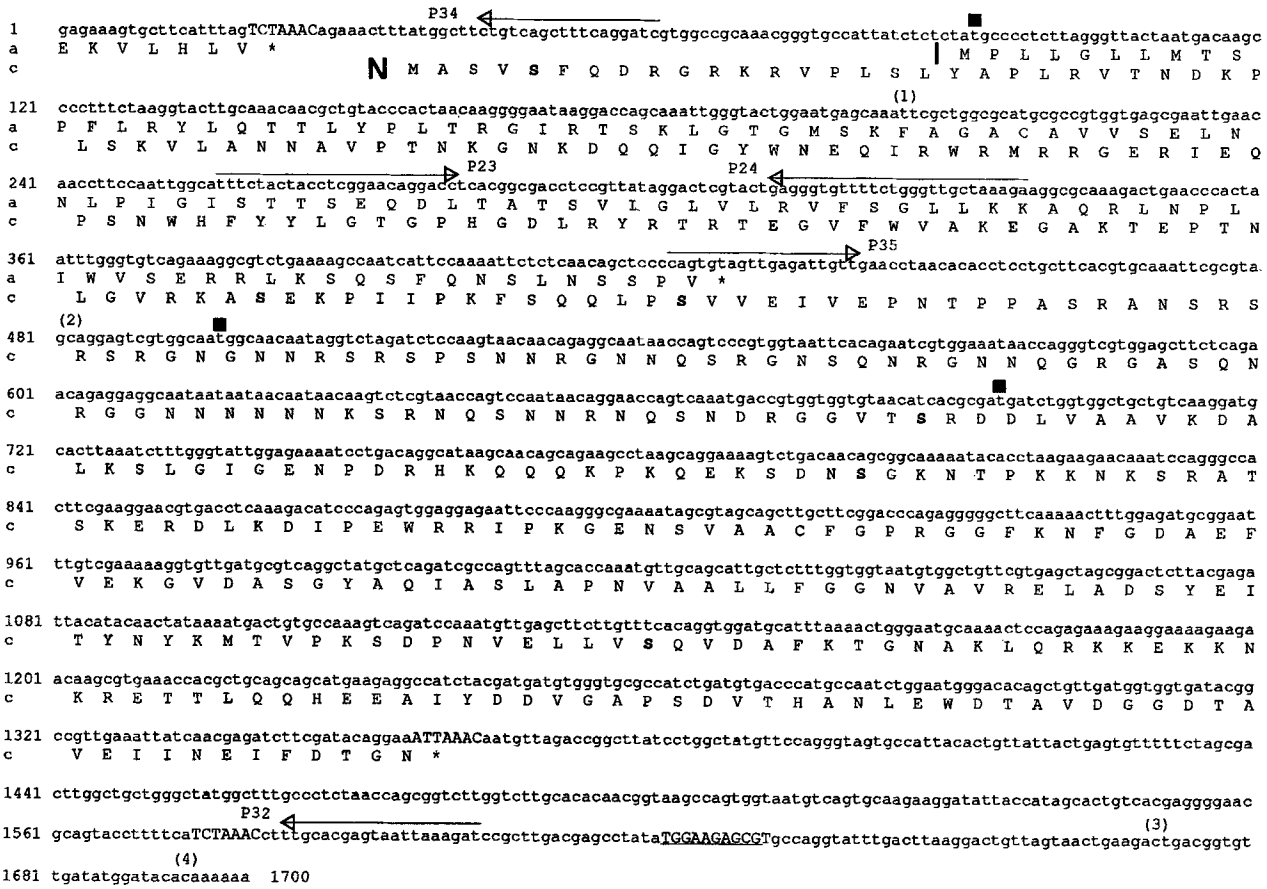

Fig. 2. Nucleotide sequence of the 1700 bases of the PEDV genome nearest to the poly(A) tail. The longest ORF, extending from nucleotides 36 to 1361 and present within the (c) frame of the nucleotide sequence comprises the PEDV N gene. A second, short ORF was also present in this reading frame downstream of the N gene, from nucleotides 1365 to 1472. Numerous ORFs were present in the (a) frame of the sequence; these include the 3′ end of the membrane gene (nucleotides 1 to 24) and the putative I gene within the +1 reading frame of the N gene (bases 91 to 429). Several further ORFs were present within this frame, for example from bases 496 to 684 and 694 to 909; the start codons for the three largest ORFs in the (a) frame internal to the N gene are indicated by squares above the sequence. In contrast, only one very small ORF was present in the third frame, from positions 1514 to 1579. Other features shown are sequences similar to coronavirus intergenic regions (indicated in upper case), the conserved coronavirus sequence at position 1622 to 1632 (indicated by upper case and underlining) and the 5′ end of the poly(A) tail (positions 1695 to 1700). Serine residues which are possible phosphorylation sites for protein kinase C or casein kinase II are indicated in bold. Differences between the groups or between different clones generated by the PCR are indicated by the numbers above the sequence: 1, C not T in one of three PCR-generated clones, causing I–T amino acid change in the N gene, F–L change in the putative I gene product (CV777); 2, CG not GC in isolate Br1/87, causing S–T amino acid change; 3, one C in isolate Br1/87, two in CV777; 4, T not C in one of two PCR clones (CV777). The positions of primers referred to in the text are shown by lines above the sequence and are labelled with arrowheads at their 3′ ends. Primers P23 and P24 are degenerate, and P34 has additional bases to create a *Bam*HI site.

rescreens. One clone named pBE1, with an insert size of 1692 nucleotides (nt), was shown to be positive for PEDV cDNA by Northern hybridization using total RNA extracted from PEDV-infected cells. Additional PEDV-specific cDNA clones were selected from the library using a 1150 nt *Pst*I fragment of pBE1 as a probe. One of these additional clones, pBE5, had an insert of 3356 nt extending from the 3′ end of the genome to the 3′ end of the putative S gene (Duarte *et al.*, 1993).

Sequence comparison of the two isolates and open reading frame analysis

Sequence data obtained for the two PEDV isolates were remarkably similar, with only three base differences in the entire sequence between all clones analysed (Fig. 2). Coding region analysis revealed one large open reading frame (ORF) capable of encoding a 441 amino acid protein, with a predicted M_r of 49K. Comparison of the translation product of the large ORF with all the proteins in the SWISSPROT database showed that it was very similar in both length and sequence to coronavirus N proteins and is therefore likely to represent the PEDV N gene. A second ORF of 336 nt, which could encode a leucine-rich (18%) protein with a predicted M_r of just over 12K, was present in the reading frame +1 to the N gene. Such a protein might be expressed since the PEDV N ORF has a poor site for ribosomal binding proving a suboptimal context for ribosome initiation (Kozak, 1986). This reading frame also contains the 3′ end of another ORF resembling the M gene of coronaviruses. The third reading frame, in contrast, contained a large number of stop codons. The codon usage of the two longest ORFs, assessed using the PC/GENE COD_FICK and COD_RNY programs, suggested that they were both likely to be expressed. The nucleotide sequence, together with the putative translation products of the three ORFs, is shown in Fig. 2. The product of the internal ORF showed no close similarity to any protein in the SWISSPROT databank, but the high leucine content was reminiscent of the ORFs found internal to the N gene in some other coronaviruses.

Analysis of the viral RNA

Hybridization with a digoxigenin-labelled P32/P35 PCR fragment probe to a blot of mRNAs from PEDV-infected cells revealed several hybridizing species (not shown), which were thought to correspond to the subgenomic mRNA species known to be transcribed from coronavirus genomes (Lai, 1990). The major hybridizing species was 1900 bases in length and probably represents the N mRNA comprising the N gene and downstream untranslated sequences, a poly(A) tail of 100 to 200 bases and a leader sequence (see below).

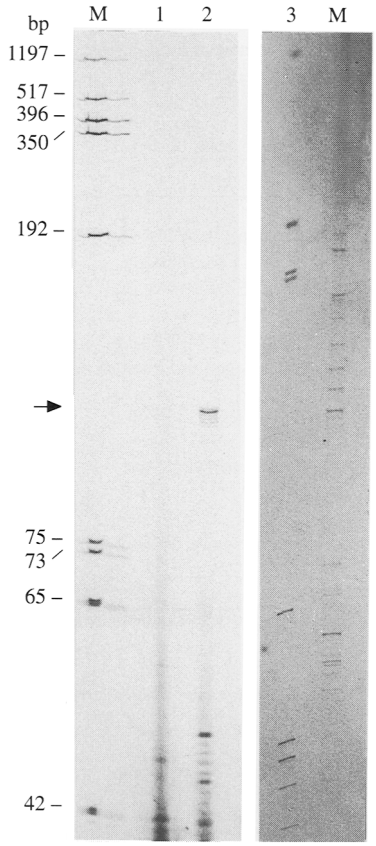

Fig. 3. Primer extension experiments using the [γ-^{32}P]ATP end-labelled primers P34 (complementary to the 5′ end of the N gene) and P32 (located near the 3′ end of the PEDV genome) to prime cDNA synthesis from poly(A)-selected RNA. Lane 1, P34, uninfected cell RNA; lane 2, P34, PEDV-infected cell RNA; lane 3, P32, PEDV-infected cell RNA; lane M shows the markers. The cDNA corresponding to the putative N gene leader sequence in lane 2 is indicated with an arrow.

The primer extension results illustrated in Fig. 3 are consistent with the presence of a PEDV leader sequence at the 5′ end of the viral mRNAs. Extension with both labelled primer P34 (5′ GAGGATCCTGAAAGCTG-ACAG 3′), of which the last 18 bases are complementary to the 5′ end of the N gene, and with P9, the sequence of which was based on the 5′ end of the PEDV M gene (unpublished), yielded identical results. In both cases one major product of 120 to 125 bases could be seen. Leader addition in most coronaviruses occurs at the consensus sequence TCTAAAC, and this sequence is seen upstream of the N gene, whereas the sequence TATAAAC is located upstream of the M gene (unpublished). Allowing for the length of the primers and their distance to the 5′ end of the intergenic sequence these results could be explained by the presence of a leader sequence of approximately 80 bases. This would then be intermediate

in length between the leader sequences of TGEV and human coronavirus (HCV) 229E, which are 91 and 60 bases, respectively (Page *et al.*, 1990; Schreiber *et al.*, 1989).

A second TCTAAAC sequence was also present in the sequenced cDNA, near the poly(A) tail. In contrast to the results obtained with P34, primer extension experiments using end-labelled P32 (5′ ATCTTTAATTACTCGTGCAA 3′), located immediately 3′ to this second TCTAAAC sequence, yielded no major extension product (Fig. 3). Instead, a variety of products were obtained which are likely to represent termination products of the primer extension owing to the secondary structure of the viral RNA. It therefore appears that this TCTAAAC sequence is not a site of leader sequence binding.

Discussion

The predicted properties of the product of the large ORF and its similarity to other coronaviral N proteins are consistent with the conclusion that the product of this ORF represents the PEDV N protein. It has previously been shown that PEDV possesses a non-glycosylated, structural protein of M_r 55K to 58K as indicated by PAGE (Egberink *et al.*, 1988; Jöhr, 1989; Knuchel *et al.*, 1992). This was thought to be the viral nucleocapsid protein on the basis of its abundance in virus-infected cells, size, high isoelectric point (pI), lack of glycosylation and RNA-binding properties (Egberink *et al.*, 1988; Knuchel *et al.*, 1992). Although the predicted M_r of the sequenced N gene product is lower than 55K to 58K such a discrepancy is commonly seen for coronaviral N proteins.

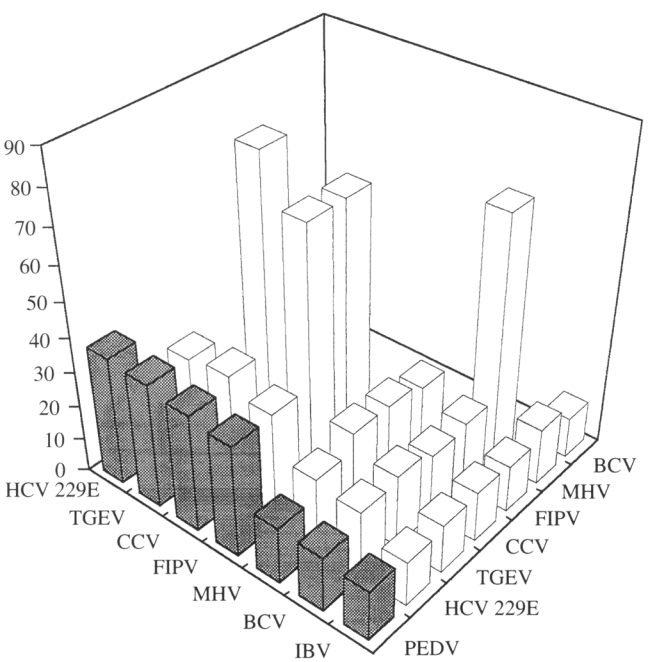

Fig. 4. Bar chart showing the percentage identity (vertical scale) of the N proteins of seven coronaviruses.

More than 10 coronavirus N genes have now been sequenced, including those of HCV 229E, TGEV, PRCV, FIPV, CCV, murine hepatitis virus (MHV), HCV OC43, bovine coronavirus (BCV) and infectious bronchitis virus (IBV) (Schreiber *et al.*, 1989; Kapke & Brian, 1986; Britton *et al.*, 1991; Vennema *et al.*, 1991; Horsburgh *et al.*, 1992; Parker & Masters, 1990; Kamahora *et al.*, 1989; Lapps *et al.*, 1987; Boursnell *et al.*, 1985). Many of these viruses have been sequenced by more than one group; the references quoted here were those used in the sequence analyses shown in Fig. 4 to 6. The percentage identity of the PEDV N protein with these coronaviruses ranged from 12 to 19% with MHV, IBV, HCV OC43 and BCV to 32 to 37% with FIPV, CCV, PRCV, TGEV and HCV 229E. Fig. 4 illustrates the percentage identity of these coronaviral N proteins to each other, and indicates the comparatively weak identity between the PEDV N protein and those of other coronaviruses. Since the N protein of PEDV appeared to be most closely related to the N proteins of TGEV and HCV 229E, the comparison of the PEDV protein with those of HCV 229E and the Purdue strain of TGEV is shown in Fig. 5. The three sequences have 94 residues in common (20%) in a consensus length of 470 amino acids. There are additionally 63 residues identical between PEDV and HCV 229E, and 49 residues identical between PEDV and TGEV. Considerably higher identity was observed in the N than in the C termini of these three coronaviruses.

A striking difference between the PEDV and the TGEV and HCV 229E N proteins, clearly visible in Fig. 5(*a*), is the presence of approximately 40 extra residues, particularly asparagine, serine and arginine, in the central portion of the PEDV protein. Two possible explanations for such an additional sequence are recombination with other RNAs and stuttering of the viral polymerase. A search of the databank using the BLASTP program (Altschul *et al.*, 1990) revealed the presence of a sequence derived from the *Enterococcus faecium* initiation factor 2 with high similarity to the additional PEDV sequence (Fig. 5*b*). However, we think that stuttering of the PEDV polymerase is a more likely explanation because the additional sequence has a noticeable 12 amino acid periodicity, with a weaker six amino acid periodicity, which could arise from duplication of an initial six amino acid stretch followed by multiplication of the 12 amino acid stretch (Fig. 5*c*). The enterococcus sequence could have arisen in a similar manner, since it shows an 11 amino acid periodicity. PEDV is unique among coronaviruses in having these additional largely hydrophilic residues, as can be seen in Fig. 6. This figure

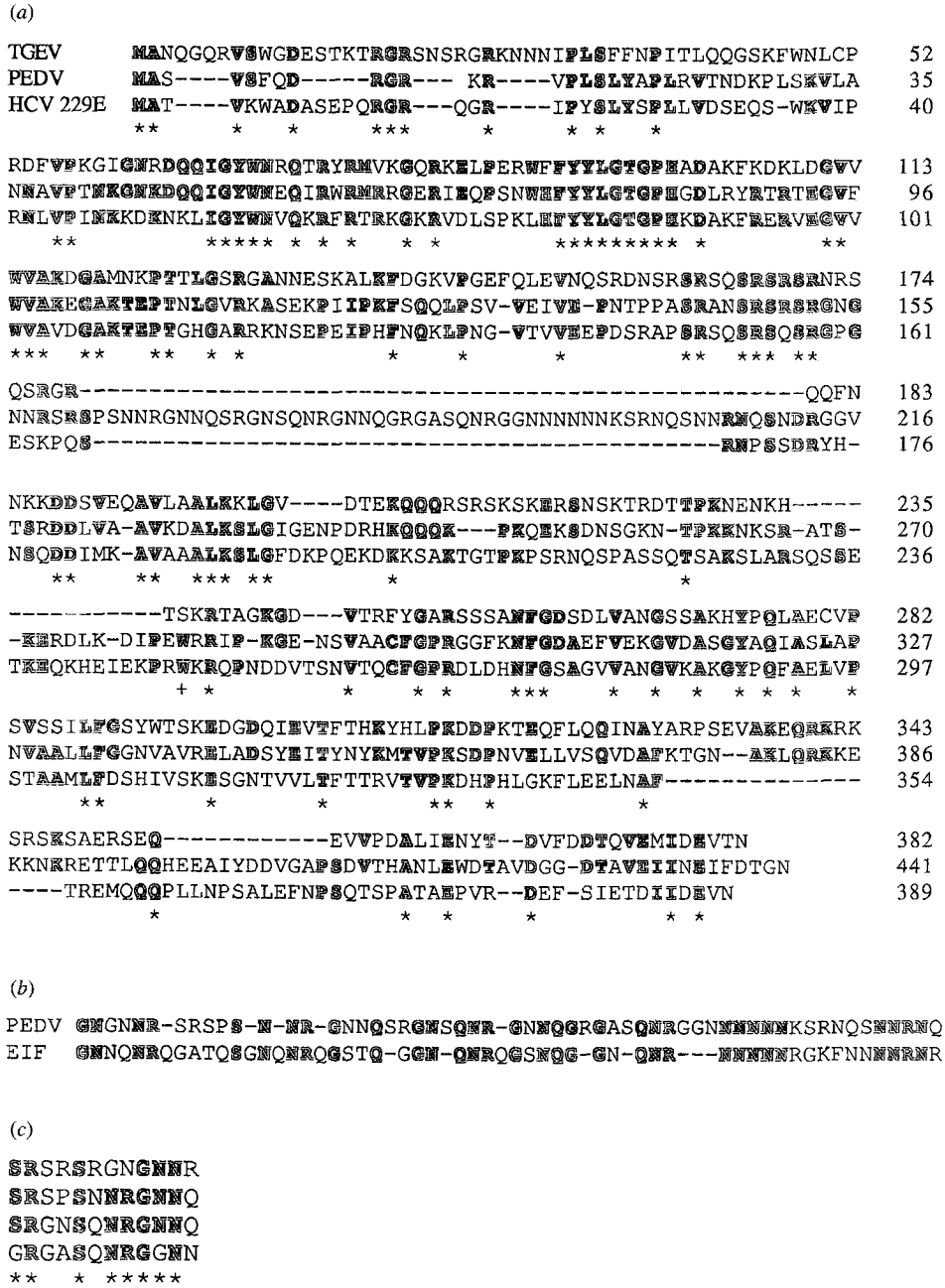

Fig. 5. (*a*) Alignment of the N proteins of PEDV, TGEV strain Purdue and HCV 229E. The alignment was made initially with the Pileup program of the UWGCG package (Devereux *et al.*, 1984) and was then refined manually. Identical amino acid residues between PEDV and at least one of the other two sequences are written in outlined letters. Amino acids common to all three are marked by asterisks. The single + represents a residue which would be common to all three viruses if the sequence of the Purdue 115 strain of TGEV determined by Rasschaert *et al.* (1987) had been used for the alignment. (*b*) Alignment of some of the PEDV extra residues and a portion of the *E. faecium* initiation factor (EIF) sequence (EMBL accession number M36878; Friedrich *et al.*, 1988). The PEDV sequence begins at residue 153. (*c*) Analysis of the central portion of the PEDV N protein showing the 12 residue periodicity of the additional PEDV residues compared with TGEV and HCV 229E and their relationship to the neighbouring residues. The sequence begins at residue 147. Residues present in at least three of the four sets of 12 amino acids are indicated with a star.

illustrates not only the sizes of the central hydrophilic regions of the proteins but also the positions of the conserved amino acid residues common to all coronaviruses.

The properties of the PEDV N protein, including the predicted pI, the number and distribution of charged residues and the high serine content (some of which are also illustrated in Fig. 6), are consistent with those of

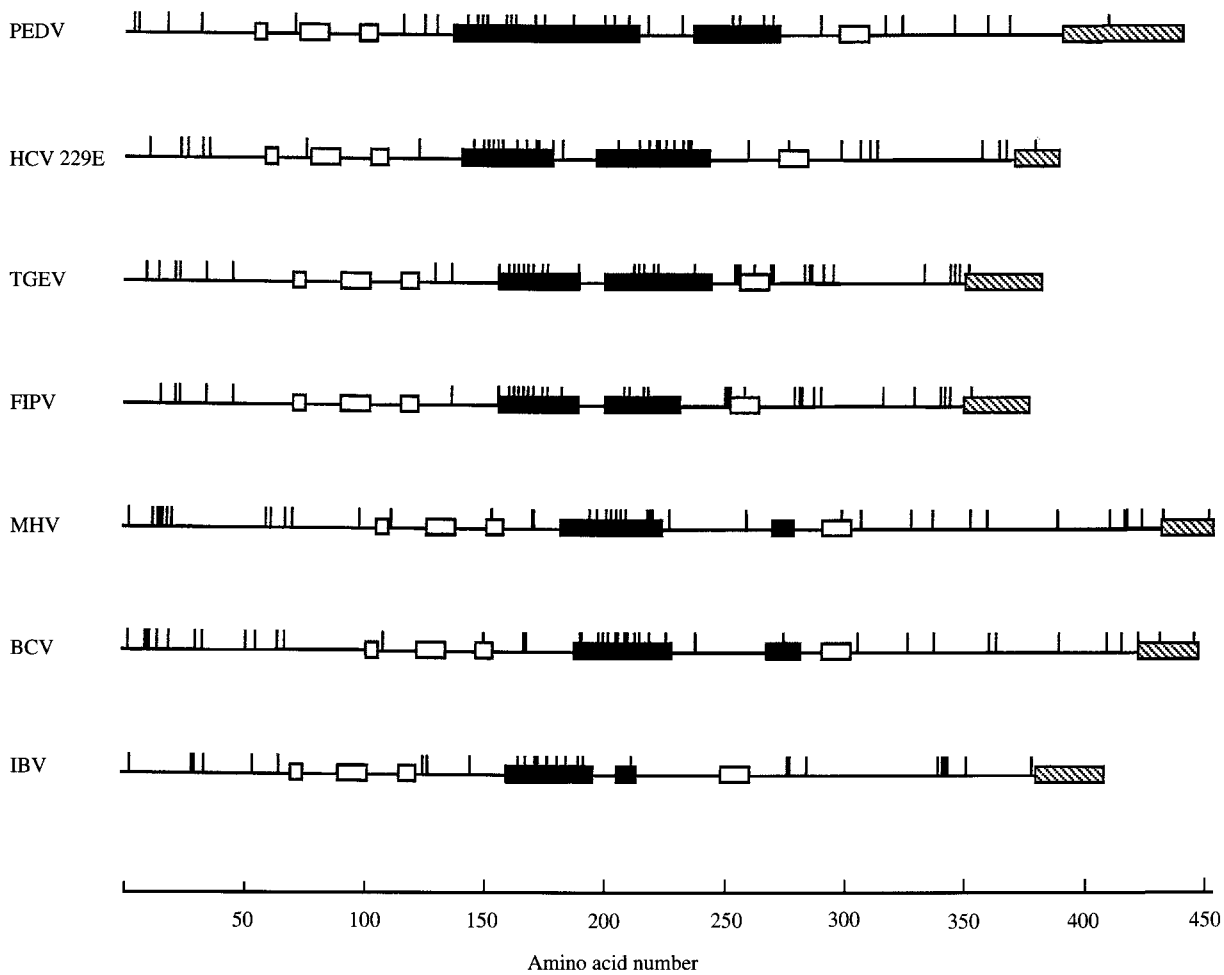

Fig. 6. Comparison of different coronavirus N proteins, showing the comparatively long hydrophilic region in the centre of the PEDV protein. Not all sequenced coronavirus N proteins are illustrated, only representative examples. Thus, the HCV OC43 and turkey coronavirus proteins are similar to that of BCV, whereas those of FECV and CCV are similar to the TGEV and FIPV proteins. Serine residues are shown as vertical lines above the base lines. Homologous regions of the sequences are indicated by open boxes and hydrophilic regions by filled-in boxes. The acidic C-terminal regions are indicated by striped boxes. The four homologous regions shown all have at least three consecutive amino acids conserved between all coronaviruses that have been sequenced and comprise the sequences GYW, FY(Y/F)(L/T)GTGP, (D/E)G(V/I)(F/Y/V)WVA and G(P/A/K)R(four to six residues)NFG for regions 1 to 4, respectively.

other coronavirus N proteins. All coronavirus N proteins have a high predicted pI of between 10·1 and 10·5 but a low pI at the C terminus of the protein where normally only negatively charged or neutral residues are observed. In good agreement with this, the predicted pI of the complete PEDV N protein is 10·5, and that of the C-terminal 50 residues only 3·4. This region of the protein could interact with the viral polymerase or possibly with cellular factors to alter host cell transcription. Apart from a structural function for the N protein additional roles have been suggested, for example the finding of PEDV N protein (M. Rosskopf, unpublished data) in the nuclei of virus-infected cells and the report by Baric *et al.* (1988) that the MHV N protein binds the leader sequences at the 5′ ends of the viral mRNAs. The possibility of additional roles for the N protein has been discussed most recently by Masters & Sturman (1990).

Several coronaviruses possess a second ORF enclosed within the N gene, but in another reading frame. BCV has a 621 nt ORF in this position, which has recently been shown to express a protein of M_r 23K in virus-infected cells and has therefore been called the internal (I) gene (Lapps *et al.*, 1987; Senanayake *et al.*, 1992). Amongst the coronaviruses antigenically related to BCV, four MHV strains possess a similar 621 nt ORF (Parker & Masters, 1990), whereas in MHV JHM and HCV OC43 two and three stop codons, respectively, reduce the size of the ORF (Skinner & Siddell, 1983; Kamahora *et al.*, 1989). PEDV, like HCV 229E, has a relatively short ORF of 336 nt which could, however, be 591 nt in length

were the first stop codon after the ORF initiation codon not present, or 816 nt long were two stop codons not present. Such stop codons could have been introduced by adaptation of the virus to cell culture. The PEDV ORF resembles the BCV I gene in coding for a high leucine content in the predicted products and having the ATG start codon in the same position with respect to that of the N gene. The viruses TGEV, PRCV, FIPV, FECV and CCV possess no such ORF, implying either that the I protein provides a non-essential function or that this function is provided by another protein, for example one encoded by a gene located downstream from the N gene in these viruses, which encodes a hydrophobic product of 78 amino acids (Garwes *et al.*, 1989; Tung *et al.*, 1992).

The PEDV sequences determined by the two groups were virtually identical, despite using different cloning techniques and working with different European isolates, isolated 10 years apart. This therefore justifies the use of PCR for the generation of the cDNA clones. Despite using a relatively large number of amplification cycles and the *Taq* polymerase, which is less accurate than currently available enzymes, only two base differences between CV777 clones were observed; these could also have been due to variation in the original RNA or to errors introduced by the reverse transcriptase. A further three nucleotide differences were consistently observed between clones derived from the two isolates. We propose that one PEDV isolate spread rapidly within the European population. This theory is supported by the observation that a monoclonal antibody raised against the M protein of the CV777 isolate of PEDV reacted with several European and one North Korean isolate of PEDV (A. Utiger, personal communication). It would however be interesting to obtain sequence data from more distant isolates, for example a recent Japanese isolate (Kusanagi *et al.*, 1992) and the North Korean isolate, to provide more information on the similarity between PEDV isolates that are geographically well separated.

The nature of the PEDV RNA also showed properties typical of coronaviruses, with a series of cross-hybridizing mRNAs in infected cells and conserved RNA motifs. A conserved 11 nucleotide sequence, thought to be a polymerase recognition site for the synthesis of the RNA negative strand during viral replication (Lai, 1990), was found near the 3′ end of the PEDV genome. The sequences from all the coronaviral genomes sequenced to date fall into two groups according to the base (G or T) before the completely conserved 5′ GGAAGAGC 3′ core sequence, and according to the distance of this sequence from the viral poly(A) tail (Schreiber *et al.*, 1989; Tobler *et al.*, 1993). Again, the PEDV sequence falls into the same group as the TGEV and HCV 229E viruses. The primer extension results suggest the presence of a leader sequence which is intermediate in length between those of TGEV and HCV 229E. Further work, however, is required to determine the exact site of leader sequence addition. If, as we anticipate, the TCTAAAC motif represents the leader binding site between the PEDV N and M genes, why does the downstream motif not also have this role? Possibly the separation between this sequence and the poly(A) tail has to be greater for subgenomic mRNA formation. Interestingly the PEDV primer extension products of P9 and P34, thought to represent the complements of the leader sequences of the M and N mRNAs, respectively, showed a series of minor products very similar in length to the major one(s). Leader sequence variation has so far been reported only for BCV, where variation in the first five bases of the leader sequence leads to generation of an AUG codon, allowing translation of an 11 amino acid peptide which results in the reduced expression of downstream genes (Hofmann & Brian, 1993). It would therefore be of interest to determine the sequence of the leader products to see whether minor species exist.

In conclusion, the nucleotide sequence described in this paper, containing a nucleocapsid gene and typical coronavirus conserved motifs, confirms that PEDV is a member of the coronavirus family, with the region of the genome sequenced showing the greatest homology to HCV 229E, TEGV, PRCV, FIPV, CCV, FECV and genetically related viruses. The sequence similarity observed between PEDV, HCV 229E and TGEV, together with the recent observation that HCV 229E and TGEV have the same cellular receptor (Yeager *et al.*, 1992; Delmas *et al.*, 1992), suggests a common origin of these viruses. Our data also suggest that different European strains of PEDV are very similar in sequence. The PCR technique developed for this study could be applied to other, presently uncharacterized coronaviruses. Further work is currently in progress to extend the sequence data and to study the functions of the PEDV N protein.

A. Bridgen, K. Tobler and M. Ackermann wish to thank M. Rosskopf and M. Schwyzer for advice during the course of this work. M. Duarte and H. Laude thank P. Have and A. Jestin for supplying the British PEDV isolate and the PEDV antiserum, respectively. They also thank J. Gelfi and P. Lambert (I.N.R.A.) for their expert assistance and D. Rasschaert (I.N.R.A.) for helpful advice. M. Duarte (C.I.V.E.T., Pando, Uruguay), holds a fellowship from the French Government. K. Tobler was supported by grant number 012.91.7 from the Swiss Federal Veterinary Services.

References

ALTSCHUL, S. F., GISH, W., MILLER, W., MYERS, E. W. & LIPMAN, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology* **215**, 403–410.

BARIC, R. S., NELSON, G. W., FLEMING, J. O., DEANS, R. J., KECK, J. G., CASTEEL, N. & STOHLMAN, S. A. (1988). Interactions between coronavirus nucleocapsid protein and viral RNAs: implications for viral transcription. *Journal of Virology* **62**, 4280–4287.

BOURSNELL, M. E. G., BINNS, M. M., FOULDS, I. J. & BROWN, T. D. K. (1985). Sequences of the nucleocapsid genes from two strains of avian infectious bronchitis virus. *Journal of General Virology* **66**, 573–580.

BRIDGEN, A., TOBLER, K. & ACKERMANN, M. (1993). Identification of coronaviral conserved sequences and application to viral genome amplification. *Advances in Experimental Biology and Medicine* (in press).

BRITTON, P., MAWDITT, K. L. & PAGE, K. W. (1991). The cloning and sequencing of the virion protein genes from a British isolate of porcine respiratory coronavirus: comparison with transmissible gastroenteritis genes. *Virus Research* **21**, 181–198.

DELMAS, B., GELFI, J., L'HARIDON, R., VOGEL, L. K., SJOSTROM, H., NOREN, O. & LAUDE, H. (1992). Aminopeptidase N is a major receptor for the entero-pathogenic coronavirus TGEV. *Nature, London* **357**, 417–420.

DEVEREUX, J., HAEBERLI, P. & SMITHIES, O. (1984). A comprehensive set of sequence analysis programs for the VAX. *Nucleic Acids Research* **12**, 387–395.

DUARTE, M., GELFI, J., LAMBERT, P., RASSCHAERT, D. & LAUDE, H. (1993). Genome organisation of porcine epidemic diarrhea virus (PEDV). *Advances in Experimental Biology and Medicine* (in press).

EGBERINK, H. F., EDERVEEN, J., CALLEBAUT, P. & HORZINEK, M. C. (1988). Characterization of the structural proteins of porcine epizootic diarrhea virus, strain CV 777. *American Journal of Veterinary Research* **49**, 1320–1324.

FRIEDRICH, K., BROMBACH, M. & PON, C. L. (1988). Identification, cloning and sequence of the *Streptococcus faecium infB* (translational initiation factor IF2) gene. *Molecular and General Genetics* **214**, 595–600.

FROHMAN, M. A., DUSH, M. K. & MARTIN, G. R. (1988). Rapid production of full-length cDNAs from rare transcripts: amplification using a single gene-specific oligonucleotide primer. *Proceedings of the National Academy of Sciences, U.S.A.* **85**, 8998–9002.

GARWES, D. J., STEWART, F. & BRITTON, P. (1989). The polypeptide of M_r 14000 of porcine transmissible gastroenteritis virus: gene assignment and intracellular location. *Journal of General Virology* **70**, 2495–2499.

HAVE, P., MOVING, V., SVANSSON, V., UTTENTHAL, A. & BLOCH, B. (1992). Coronavirus infection in mink (*Mustela vison*). Serological evidence of infection with a coronavirus related to transmissible gastroenteritis virus and porcine epidemic diarrhea virus. *Veterinary Microbiology* **31**, 1–10.

HOFMANN, M. A. & BRIAN, D. A. (1993). An intraleader open reading frame is selected from a hypervariable leader 5′ terminus during persistent infection by the bovine coronavirus. *Advances in Experimental Biology and Medicine* (in press).

HOFMANN, M. & WYLER, R. (1988). Propagation of the virus of porcine epidemic diarrhea in cell culture. *Journal of Clinical Microbiology* **26**, 2235–2239.

HOFMANN, M. & WYLER, R. (1989). Quantitation, biological and physicochemical properties of cell culture-adapted porcine epidemic diarrhea coronavirus (PEDV). *Veterinary Microbiology* **20**, 131–142.

HORSBURGH, B. C., BRIERLEY, I. & BROWN, T. D. K. (1992). Analysis of a 9·6 kb sequence from the 3′ end of canine coronavirus genomic RNA. *Journal of General Virology* **73**, 2849–2862.

JÖHR, L. (1989). *Induktion und Charakterisierung von monoklonalen Antikörpern gegen das Nukleokapsidprotein des Virus der Epidemischen Virusdiarrhoe der Schweine.* Inaugural Dissertation zur Erlangung der Doktorwürde der Veterinär-Medizinischen Fakultät der Universität Zürich.

KAMAHORA, T., SOE, L. H. & LAI, M. M. C. (1989). Sequence analysis of the nucleocapsid gene and leader RNA of human coronavirus OC43. *Virus Research* **12**, 1–9.

KAPKE, P. A. & BRIAN, D. A. (1986). Sequence analysis of the porcine transmissible gastroenteritis coronavirus nucleocapsid protein gene. *Virology* **151**, 41–49.

KINGSTON, R. E. (1991). *Current Protocols in Molecular Biology*. New York: Greene Publishing Associates.

KNUCHEL, M., ACKERMANN, M., MÜLLER, H. K. & KIHM, U. (1992). An ELISA for detection of antibodies against porcine epidemic diarrhoea virus (PEDV) based on the specific solubility of the viral surface glycoprotein. *Veterinary Microbiology* **32**, 117–134.

KOZAK, M. (1986). Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes. *Cell* **44**, 283–292.

KUSANAGI, K., KUWAHARA, H., KATOH, T., NUNOYA, T., ISHIKAWA, Y., SAMEJIMA, T. & TAJIMA, M. (1992). Isolation and serial propagation of porcine epidemic diarrhea virus in cell cultures and partial characterization of the isolate. *Journal of Veterinary Medical Science* **54**, 313–318.

LAI, M. M. C. (1990). Coronavirus: organization, replication and expression of genome. *Annual Review of Microbiology* **44**, 303–333.

LAPPS, W., HOGUE, B. G. & BRIAN, D. A. (1987). Sequence analysis of the bovine coronavirus nucleocapsid and matrix protein genes. *Virology* **157**, 47–57.

MASTERS, P. S. & STURMAN, L. S. (1990). Functions of the coronavirus nucleocapsid protein. *Advances in Experimental Medicine and Biology* **276**, 235–238.

PAGE, K. W., BRITTON, P. & BOURSNELL, M. E. G. (1990). Sequence analysis of the leader RNA of two porcine coronaviruses: transmissible gastroenteritis virus and porcine respiratory coronavirus. *Virus Genes* **4**, 289–301.

PARKER, M. M. & MASTERS, P. S. (1990). Sequence comparison of the N genes of five strains of the coronavirus mouse hepatitis virus suggests a three domain structure for the nucleocapsid protein. *Virology* **179**, 463–468.

PENSAERT, M. B. (1989). Porcine epidemic diarrhea virus. In *Virus Infections of Porcines*, pp. 167–176. Edited by M. B. Pensaert. Amsterdam: Elsevier.

PENSAERT, M. B. & DEBOUCK, P. (1978). A new coronavirus-like particle associated with diarrhea in swine. *Archives of Virology* **58**, 243–247.

PENSAERT, M. B., DEBOUCK, P. & REYNOLDS, D. J. (1981). An immunoelectron microscopic and immunofluorescent study on the antigenic relationship between the coronavirus-like agent, CV 777, and several coronaviruses. *Archives of Virology* **68**, 45–52.

QUEEN, C. & KORN, L. J. (1984). Microgenie sequence analysis program. *Nucleic Acids Research* **12**, 581–599.

RASSCHAERT, D., GELFI, J. & LAUDE, H. (1987). Enteric coronavirus TGEV: partial sequence of the genomic RNA, its organization and expression. *Biochimie* **69**, 591–600.

RASSCHAERT, D., DUARTE, M. & LAUDE, H. (1990). Porcine respiratory coronavirus differs from transmissible gastroenteritis virus by a few genomic deletions. *Journal of General Virology* **71**, 2599–2607.

SAMBROOK, J., FRITSCH, E. F. & MANIATIS, T. (1989). *Molecular Cloning: A Laboratory Manual*, 2nd edn. New York: Cold Spring Harbor Laboratory.

SCHREIBER, S. S., KAMAHORA, T. & LAI, M. M. C. (1989). Sequence analysis of the nucleocapsid protein gene of human coronavirus 229E. *Virology* **169**, 142–151.

SENANAYAKE, S. D., HOFMANN, M. A., MAKI, J. L. & BRIAN, D. A. (1992). The nucleocapsid protein gene of bovine coronavirus is bicistronic. *Journal of Virology* **66**, 5277–5283.

SKINNER, M. A. & SIDDELL, S. G. (1983). Coronavirus JHM: nucleotide sequence of the mRNA that encodes nucleocapsid protein. *Nucleic Acids Research* **11**, 5045–5054.

TOBLER, K., BRIDGEN, A. & ACKERMANN, M. (1993). Sequence analysis of the nucleocapsid protein gene of porcine epidemic diarrhoea virus. *Advances in Experimental Biology and Medicine* (in press).

TUNG, F. Y. T., ABRAHAM, S., SETHNA, M., HUNG, S.-L., SETHNA, P., HOGUE, B. G. & BRIAN, D. A. (1992). The 9-kda hydrophobic protein encoded at the 3′ end of the porcine transmissible gastroenteritis coronavirus genome is membrane-associated. *Virology* **186**, 676–683.

VENNEMA, H., DE GROOT, R. J., HARBOUR, D. A., HORZINEK, M. C. & SPAAN, W. J. M. (1991). Primary structure of the membrane and nucleocapsid protein genes of feline infectious peritonitis virus and

immunogenicity of recombinant vaccinia viruses in kittens. *Virology* **181**, 327–335.

VENNEMA, H., ROSSEN, J. W. A., WESSELING, J., HORZINEK, M. C. & ROTTIER, P. J. M. (1992). Genomic organization and expression of the 3′ end of the canine and feline enteric coronaviruses. *Virology* **191**, 134–140.

WIRTH, U. V., VOGT, B. & SCHWYZER, M. (1991). The three major immediate-early transcripts of bovine herpesvirus 1 arise from two divergent and spliced transcription units. *Journal of Virology* **65**, 195–205.

YALING, Z., EDERVEEN, J., EGBERINK, H., PENSAERT, M. & HORZINEK, M. C. (1988). Porcine epidemic diarrhea virus (CV 777) and feline infectious peritonitis virus (FIPV) are antigenically related. *Archives of Virology* **102**, 63–71.

YEAGER, C. L., ASHMUN, R. A., WILLIAMS, R. K., CARDELLICHIO, C. B., SHAPIRO, L. H., LOOK, A. T. & HOLMES, K. V. (1992). Human aminopeptidase N is a receptor for human coronavirus 229E. *Nature, London* **357**, 420–422.

(*Received 26 February 1993; Accepted 30 April 1993*)