SEQUENCE ANALYSIS OF THE NUCLEOCAPSID PROTEIN GENE OF PORCINE EPIDEMIC DIARRHOEA VIRUS

Kurt Tobler, Anne Bridgen and Mathias Ackermann

Institute for Virology
Veterinary Medical Faculty
University of Zürich
Winterthurerstrasse 266a
CH-8057 Zürich

ABSTRACT

The nucleotide (nt) sequence of 1.7 kbp cDNA representing the 3' end of the PEDV genome has been determined. Viral RNA was reverse transcribed and the cDNA was amplified by polymerase chain reaction using degenerate primers. The sequences of the primers were based on conserved regions of coronaviral genomes. A 1323 nt open reading frame (ORF) showed good homology to the nucleocapsid (N) gene of other coronaviruses. The greatest homologies at the amino acid and the nt levels were observed with Human Coronavirus 229E. A second 336 nt ORF, which might encode a leucine-rich protein, was found within the N gene. Between the 3' end of the N gene and the poly(A) tail was a sequence of eleven nt, which is conserved among the other sequenced coronaviruses. Finally, a seven base sequence similar to the conserved intergenic sequences was present 5' to the N gene. These results confirm the classification of PEDV as a coronavirus.

INTRODUCTION

Porcine epidemic diarrhoea virus (PEDV) causes diarrhoea in pigs, particularly in neonates. The biological behaviour and electron microscopic appearance of PEDV resulted in its provisional classification as a coronavirus. However, inconsistent cross-reactivity with antibodies directed against other coronaviruses[1] did not permit the classification of PEDV into a coronavirus antigenic group. This provisional classification of PEDV did, however, allow a genomic amplification strategy based on anticipated similarity to other, sequenced members of the coronavirus family. The amplified products were cloned and sequenced and the results are described in this article.

MATERIALS AND METHODS

The virus strain used in this study was the CV 777 strain of PEDV isolated by Pensaert and Debouck[2] and adapted to culture in Vero cells by Hofmann and Wyler.[3] Vero cells (ATCC 1587) were grown and after three days infected with PEDV at an moi of 0.03 as described previously.[4] RNA was extracted from both semi-purified virions and from infected cells. RNA from infected cells was harvested when the cells showed 85-90%

Coronaviruses, Edited by H. Laude and J.F. Vautherot
Plenum Press, New York, 1994

cytopathic effect (cpe) and processed by the caesium chloride method as described elsewhere.[5] For virion RNA, cells showing 95-100% cpe were disrupted by three cycles of freeze-thaw, cellular debris were removed by low speed centrifugation, and virus was pelleted by centrifugation in a Beckman SW28 rotor at 100,000 g at 4°C for 2 h. The viral pellets were washed, and the RNA was extracted.[5]

First strand cDNA was synthesized in two steps, comprising primer binding and extension, according to Wirth et al.[6] Firstly, ng amounts of purified viral RNA or μg amounts of total infected-cell RNA were annealed to 500 ng primer (see below) for 1 h at 42°C. This was followed by 2 h extension reactions, also at 42°C. The cDNA was extracted with phenol:chloroform (1:1), then with chloroform, before being precipitated with ethanol and resuspended in 20 μl of sterile distilled water.

The cDNA was amplified with the following degenerate primers, P23: 5' AAGCTTTT-ACTA(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, 8 fold degeneracy) and P25: 5' GACTAGTTGGTGGAG(A/T)TTTAA(C/T)CC(A/T)GA 3' (27mer, 8 fold degeneracy), the sequences of which were based on conserved regions of coronaviral genomes.[7] P25 was also tailed with T residues (designated P25(dT)) with terminal deoxynucleotidyl exotransferase (Boehringer) for 30 min at 37°C. The buffer supplied by the manufacturer was supplemented with 1.5 mM cobalt chloride and a 50:1 molar ratio of dTTP to DNA template. First strand cDNA was primed with oligo(dT) primer (Pharmacia), P24 or P25(dT).

Polymerase chain reaction (pcr) amplifications were performed overnight using a Hybaid Intelligent Heating Block (Model IHB 2024). The pcr buffer consisted of 50 mM potassium chloride, 10 mM Tris pH 8.3, 1 mM magnesium chloride, 10 μg/ml gelatin, 0.045% NP40 and 0.045% Tween 20. Primers were used at 10 μM each and dNTPs at 208 μM each, while 0.1u taq polymerase (Cetus) and 2 μl of the cDNA were used for each 30 μl reaction. P24 and P25 were used together to amplify cDNA primed with oligo dT or with P24, while P23 and P25 were used to amplify P25(dT)-primed cDNA in a modification of the 3' RACE technique of Frohman et al.[8] With primer pair P24/P25 the samples were processed with 38 cycles of 50 sec at 94°C, 60 sec at 48°C, and 60 sec at 72°C. With primers P23/P25 40 cycles of 50 sec at 94°C, 60 sec at 47°C and 150 sec at 72°C were run. A final 5 min extension was made at 72°C for both reactions.

For cloning, the two pcr products were blunt ended with 7.5 u T_4-DNA polymerase (NEB) in the presence of 100 μM dNTPs for 15 min at 12°C and then treated with polynucleotide kinase (NEB) for 30 min at 32°C. Using T_4-DNA ligase (NEB), the pcr products were inserted into pBluescript®II KS+ (Stratagene), which had previously been digested with EcoRV and dephosphorylated. Subclones were made by standard subcloning procedures. Competent *E. coli* DH5α cells, prepared by the calcium chloride method of Maniatis et al.[9] and stored frozen, were transformed by the constructs before plating out on LB plates containing 100 μg/ml ampicillin.

The sequencing reactions were performed by the dideoxynucleotide method using the Sequenase 2 kit (United States Biochemicals). Three sequence ambiguities were resolved by use of the Vent polymerase sequencing kit (NEB) or by replacing dGTP by dITP in the Sequenase reaction. The fragments were separated by electrophoresis through 6% polyacrylamide gels containing 7 M urea. Sequences were analysed with the IntelliGenetics PC/GENE programme or with the University of Wisconsin Genetics Computer Group programmes.[10]

RESULTS AND DISCUSSION

PEDV single stranded cDNA could be amplified with degenerate primers P24/25 and P23/P25(dT) to give products of 0.7 kbp and 1.6 kbp, respectively. The two cloned pcr products, together with their subclones, were sequenced on both strands. Two additional clones derived from independent pcr reactions were also sequenced on one strand in order to take account of possible errors introduced by the taq DNA polymerase; in fact only one base difference was observed between the two 1.6 kbp clones and none between the two 0.7 kbp clones. Three regions were resequenced with dITP using the Sequenase 2 kit and with Vent polymerase to resolve observed sequence ambiguities. In one position (1189-1261 bases from the poly (A) tail) the RNA had the potential to form a stem loop structure covering 72 bases.

Analysis of the sequenced cDNA, comprising the 1.7 kb nearest to the viral poly(A) tail, revealed several open reading frames (ORFs) as is illustrated in Figure 1. The most noticeable feature shown by this figure is the large ORF of 1323 nt in length, which has the capacity to encode a protein with a predicted molecular weight of 48,967 daltons. This figure is quite similar to the 55 to 58 kilodaltons observed for the PEDV N protein.[4,11]

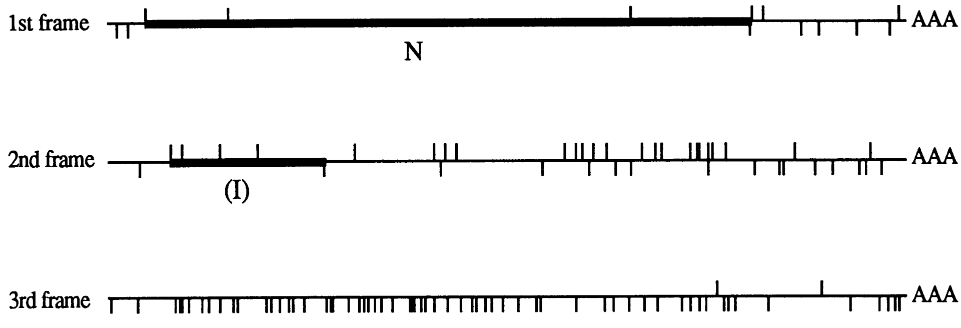

Figure 1. Representation of the 1.7 kb cDNA nearest to the 3' end of the PEDV genome. The codons for methionine (vertical lines above the baselines) and termination codons (vertical lines below the baselines) are shown in the three forward reading frames. The resulting open reading frames are depicted as bold lines and designated with N for nucleocapsid protein and (I) for internal ORF. AAA represents the poly(A) tail.

Comparison of the PEDV protein encoded by the large ORF with all the proteins in the SWISS-PROT data base showed that it was highly homologous with coronavirus N proteins. At least ten coronavirus N genes have now been sequenced, including those of Human Coronavirus (HCV) 229E,[12] Transmissible Gastroenteritis virus (TGEV) strain Purdue,[13] Feline Infectious Peritonitis virus (FIPV),[14] Murine Hepatitis virus (MHV) A59,[15] HCV OC43,[16] Bovine Coronavirus (BCV)[17] and Infectious Bronchitis virus (IBV) strain Beaudette.[18] The percentage identity of the PEDV N protein with those of other coronaviruses, as measured using the Myers and Miller method (PC/GENE), ranged from 13-17% with MHV, IBV, HCV OC43 and BCV to 32-37% with FIPV, TGEV and HCV 229E. The highest homology (37%) was observed to the HCV 229E protein, closely followed by 36% for the TGEV protein.

The predicted properties of the PEDV N protein are also consistent with those of other coronaviruses. The isoelectric points (pI) of all the coronavirus N proteins are high because of the frequency of positively charged residues, their predicted values lying between 10.1 and 10.5. The estimated pI of the N protein of PEDV is 10.5. In contrast to the complete proteins, the C-termini of coronaviral N proteins are negatively charged. This is also true for PEDV: the 50 amino acids nearest to the C-terminus of the PEDV N protein have an estimated pI of only 3.4.

The N protein sequences of all the eight coronaviruses already mentioned contain four homologous regions, as is illustrated in Figure 2. Three are close together in the first third of the sequence, the last one occurring in the last third of the protein. The distances between the third and the fourth homologous regions in TGEV, FIPV, MHV, HCV OC43, BCV and IBV are all about 140 amino acids, but they differ in their distance from the N-terminus. In contrast, the third and fourth conserved regions are separated by about 160 amino acids in HCV 229E and by more than 190 amino acids in PEDV. The intervening sequences in all eight viruses contain two hydrophilic stretches interrupted by a hydrophobic region. The more N-terminal hydrophilic regions consist of a serine-rich sequence. The secondary structure analyses made using the methods of GARNIER and GGBSM (PC/GENE) predicted an α-helix in the amino acids following this serine-rich sequence.

Figure 1 shows an additional ORF within the N gene but in the second reading frame, indicated by (I); it starts 55 bases after the ATG of the N reading frame and could encode a leucine-rich (18%) 112 amino acid protein with a predicted molecular weight of 12,211 daltons. Such a protein might be expressed since the PEDV N gene, like that of other

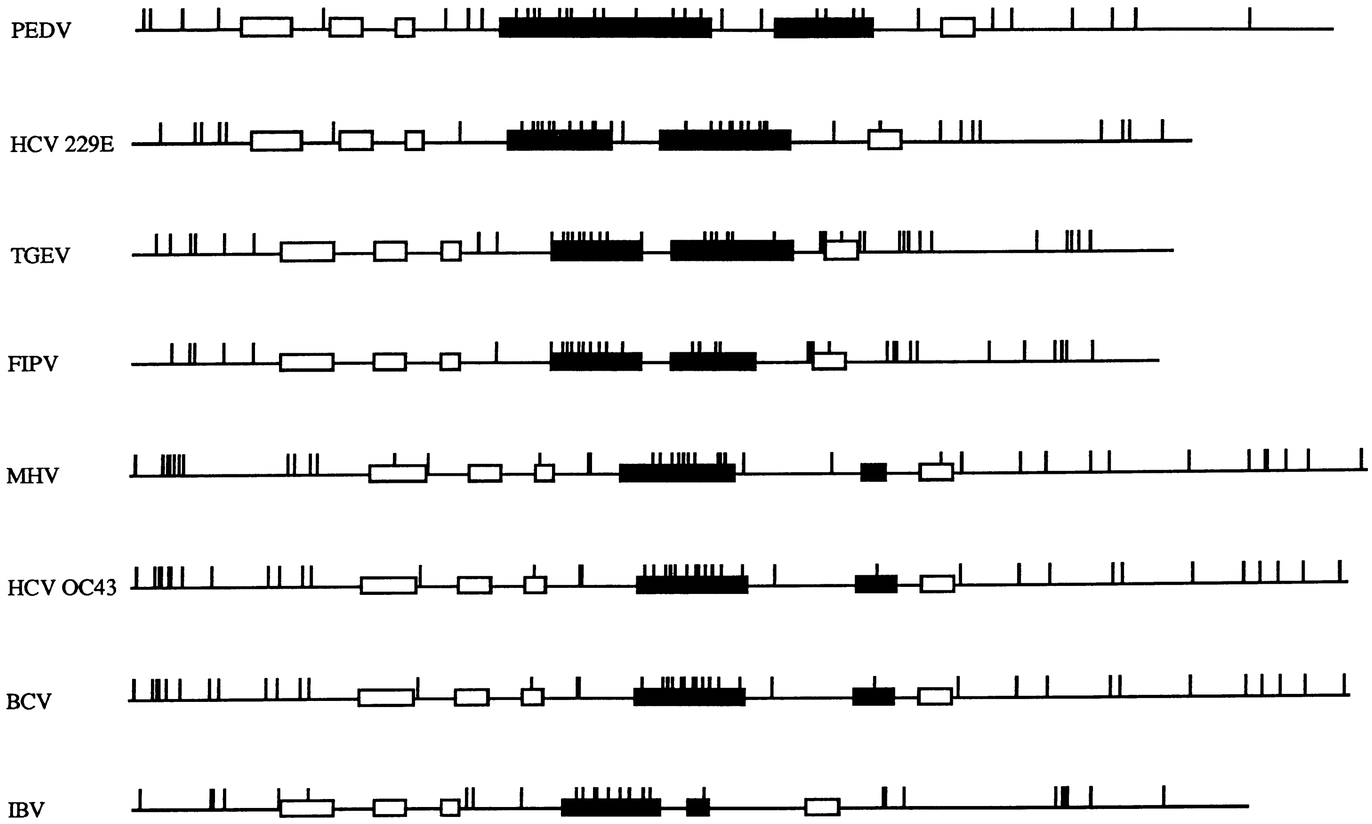

Figure 2. Comparison of different coronavirus N proteins. Serine residues are shown as vertical lines above the baselines. Homologous regions of the sequences are indicated by □ and hydrophilic regions by ■. Abbreviations of virus names are explained in the text. The consensus sequences in the homologous regions are: Gly-Tyr-Trp in the first, Gly-Thr-Gly-Pro in the second, Trp-Val-Ala in the third and Asn-Phe-Gly in the fourth conserved region.

coronaviruses, has only a poor site for ribosomal binding.[19] The site of the starting point relative to the ORF of the N protein and the frequency of leucine in the putative protein are characteristics also seen in the I protein of BCV[20] and the ORFs of MHV A59 and HCV 229E. The open reading frames in HCV 229E and PEDV are of a similar length, which is considerably shorter than that of MHV A59 or BCV.

The PEDV RNA also shows features typical of coronaviruses. The sequence 5' UCUAAAC 3', similar to the intergenic sequences of other coronaviruses, was found 16 bases upstream of the AUG of the N protein ORF. The intergenic sequence is thought to be the site of leader sequence addition during mRNA synthesis.[21] An eleven-nucleotide sequence, homologous to that found in other coronaviruses, was found near the 3' end of the PEDV genome; this sequence has been proposed to be a polymerase recognition site for synthesis of the RNA negative strand during viral replication.[21] These sequences, which are illustrated in Table 1, fall into two groups according to the base (G or U) before the completely conserved 5' GGAAGAGC 3' core sequence and according to the distance of this sequence from the viral poly(A) tail. Again, the PEDV sequence shows the highest similarity to the TGEV, FIPV and HCV 229E viruses.

Table 1. The 3' conserved sequences of eight coronaviruses

Virus	Sequence	Inclusive position from poly(A) tail (bases)
PEDV	UGGAAGAGCG U	74
HCV 229E	UGGAAGAGCC A	75
TGEV	UGGAAGAGCU A	76
FIPV	UGGAAGAGCU A	76
BCV	GGGAAGAGCC A	79
HCV OC43	GGGAAGAGCC A	79
IBV	GGGAAGAGCU A	81
MHV	GGGAAGAGCU C	82

modified from Schreiber et al.[12]

Thus, many features of the PEDV N protein are similar to those of other coronaviruses. One difference is the length of the protein compared with the related TGEV and HCV 229E proteins. This difference stems largely from additional amino acid residues, particularly asparagines, in the central region of the protein. The asparagine content of the protein is very high (10.9%), with a stretch of six adjacent asparagine residues occurring at one point. Although unusual, such amino acid runs do occur in nature; for example, the *Drosophila* mastermind gene contains numerous runs of asparagine, glutamine and glycine.[22] In this region of the PEDV N protein there is also an interesting motif comprising an arginine at every sixth residue, repeated five times. TGEV, but not HCV 229E, possesses a shorter such motif. Its significance is not known, but a computer search of the SWISS-PROT data base using the FASTA programme revealed that the motif is also present in numerous nucleic acid binding proteins including the Epstein Barr virus EBNA 1 and 2 nucleoproteins, yeast SNF2 nuclear protein and the Herpes Simplex Virus Us11 protein. The last is a small, abundant, basically charged regulatory protein which binds host 60S ribosomal subunits.[23] These possible homologies are interesting in that the N protein of PEDV has been observed in the nuclei of virus-infected cells (Rosskopf, unpublished data), where it could perform a regulatory function.

In conclusion, the presented sequence data for PEDV confirms that this virus is a coronavirus since it is a polyadenylated RNA virus possessing RNA motifs distinctive to coronaviruses and a gene with high homology to the coronavirus N gene. The greatest homology was observed between the genomes of PEDV and the HCV 229E and TGEV viruses. Of these latter two viruses, the gene order of HCV 229E is more reminiscent of that of PEDV than is that of TGEV, since PEDV possesses no gene 3' to the N gene, but does possess a second ORF within the N gene with the capacity to code for a leucine rich protein. The PEDV sequence is also more homologous to that of HCV 229E at the nucleotide level. Thus, PEDV is a coronavirus with the greatest similarity to coronavirus HCV 229E.

ACKNOWLEDGEMENTS

This work was supported by Grant No 012.91.7 of the Swiss Federal Veterinary Services. We also wish to thank Dr M. Schwyzer for advice during the course of this work.

REFERENCES

1. Z. Yaling, J. Ederveen, H. Egberink, M. Pensaert, and M.C. Horzinek, *Arch. Virol.* 102:63 (1988).
2. M.B. Pensaert and P. Debouck, *Arch. Virol.* 58:243 (1978).
3. M. Hofmann and R. Wyler, *J. Clin. Microbiol.* 26:2235 (1988).
4. M. Knuchel, M. Ackermann, H.K. Müller, and U. Kihm, *Vet. Microbiol.* 32:117 (1992).
5. R.E. Kingston, "Current Protocols in Molecular Biology," section 4.1.2, Greene Publishing Associates, Brooklyn (1991).
6. U.V. Wirth, B. Vogt, and M. Schwyzer, *J. Virology* 65:195 (1991).
7. A. Bridgen, K. Tobler, and M. Ackermann, *Adv. Exp. Med. Biol.*, (this volume).
8. M.A. Frohman, M.K. Dush, and G.R. Martin, *Proc. Natl. Acad. Sci.* 85:8998 (1988).
9. T. Maniatis, E.F. Fritsch, and J. Sambrook. "Molecular Cloning: A Laboratory Manual (2nd edition)," Cold Spring Harbour Laboratory, Cold Spring Harbor NY (1989).
10. J. Devereux, P. Haeberli, and O. Smithies, *Nucl. Acids Res.* 12:387 (1984).
11. H.F.E. Egberink, J. Ederveen, P. Callebaut, and M.C. Horzinek, *Am. J. Vet. Res.* 49:1320 (1988)
12. S.S. Schreiber, T. Kamahora, and M.C. Lai, *Virology* 169:142 (1989).
13. P.A. Kapke and D.A. Brian, *Virology* 151:41 (1986).
14. H. Vennema, R.J. De Groot, D.A. Harbour, M.C. Horzinek, and W.J.M. Spaan, *Virology* 181:327 (1991).
15. M.M. Parker and P.S. Master, *Virology* 179:463 (1990).
16. T. Kamahora, L.H. Soe, and M.C. Lai, *Virus Research* 12:1 (1989).
17. W. Lapps, B.G. Hogue, and D.A. Brian, *Virology* 157:47 (1987).
18. M.E.G. Boursnell, M.M. Binns, I.J. Fould, and T.D.K. Brown, *J. Gen. Virol.* 66:573 (1985).
19. M. Kozak, *Cell* 44:283 (1986).
20. S.D. Senanayake, M.A. Hofmann, J.L. Maki, and D.A. Brian, *J. Virol.* 66:5277 (1992).
21. M.C. Lai, *Ann. Rev. Microbiol.* 44:303 (1990).
22. D. Smoller, C. Friedel, A. Schmid, D. Bettler, L. Lam, and D. Yedvobnick, *Genes Dev.* 4:1688 (1990).
23. R.J. Roller and B. Roizman, *J. Virol.* 66:3624 (1992).