NUCLEOTIDE SEQUENCE OF THE PORCINE TRANSMISSIBLE GASTROENTERITIS CORONAVIRUS MATRIX PROTEIN GENE

Paul A. Kapke[1], Frank Y. C. Tung[2], David A. Brian[2],
Roger D. Woods[1], and Ronald Wesley[1]

[1]USDA-ARS, National Animal Disease Center
P.O. Box 70
Ames, Iowa 50010
and
[2]Department of Microbiology
The University of Tennessee
Knoxville, Tennessee 37996-0845

ABSTRACT

cDNA clones mapping within the first 2601 bases of the 3' end of the TGEV genome were sequenced completely or in part by the method of Maxam and Gilbert and open reading frames were examined. One reading frame yielding a protein having properties of the matrix (M) protein was identified. It is positioned at the immediate 5' side of the nucleocapsid (N) gene but is separated by an intergenic region of 12 bases. The deduced M protein is comprised of 262 amino acids, has a molecular weight of 29,544, is moderately hydrophobic, and has an amino acid sequence homology of approximately 36% with the mouse hepatitis coronavirus, 37% with the bovine enteric coronavirus, and 28% with the avian infectious bronchitis virus. Judging from an alignment with MHV and IBV proteins, the amino terminus of the TGEV M protein extends 54 amino acids from the virion envelope which compares with 26 for MHV and 21 for IBV.

INTRODUCTION

The porcine transmissible gastroenteritis coronavirus (TGEV) is comprised of 3 major structural proteins, an internal nucleocapsid phosphoprotein (N) of 43 Kd, and two glycosylated envelope proteins, one of 29 Kd (a matrix-like protein, M or E1) and one of 200 kd (the peplomeric, P, or spike protein) (Brian et al., 1983; Garwes and Pocock, 1975; Kapke and Brian, 1986; Wesley and Woods, 1986). While the 200 Kd P glycoprotein is demonstrably important in stimulating neutralizing antibody (Garwes et al., 1978), the 29 Kd M glycoprotein may also be important, especially if complement is part of the virus-antibody reaction (R. Wood et al., this volume).

To investigate the role of individual viral proteins in virus replication and in induction of immunity, we have prepared cDNA clones beginning from the polyadenylated 3' end of the TGEV genome and examined

M. M. C. Lai et al. (eds.), *Coronaviruses*
© Plenum Press, New York 1987

the sequences of potential genes (Kapke and Brian, 1986). Within the first (3') 2000 bases we deduced, from an examination of open reading frames, a noncoding region of 276 bases, and genes for a 9101 mol. wt. hypothetical hydrophobic polypeptide, a 43,426 mol. wt. nucleocapsid protein, and part of a matrix protein, arranged in that order from the 3' end of the genome (Kapke and Brian, 1986). Assuming that a conserved intergenic sequence would be found in TGEV as has been found in the mouse hepatitis coronavirus (MHV) (Budzilowicz et al., 1985), and the avian infectious bronchitis coronavirus (IBV) (Brown and Boursnell, 1984), we prepared a synthetic oligonucleotide that is complementary to the TGEV intergenic sequence and used it as a primer for first-strand DNA synthesis for the preparation of additional genomic cDNA clones. Several cDNA clones were thus prepared and seven that mapped within the first (3') 2601 bases were sequenced in part and another clone was sequenced completely to derive a potential gene sequence for the M protein.

MATERIALS AND METHODS

Cells and Virus

The Purdue strain of TGEV was grown on swine testicle (ST) cells as previously described (Kapke and Brian, 1986).

cDNA Cloning of TGEV Genomic RNA

cDNA cloning was accomplished by the method of Gubler and Hoffman (1983) essentially as described (Kapke and Brian, 1986) except that the synthetic oligonucleotide 5' TTAGAAGTTTAGTTA 3' was used as primer for first-strand cDNA synthesis. The primer was synthesized by the phosphoramadite method and was purified by polyacrylamide gel electrophoresis. Clones were selected by colony hybridization to random-primed cDNA prepared from size-selected genomic RNA (Kapke and Brian, 1986). Clones were initially mapped by a matrix cross-hybridization method using purified inserts that were labeled by nick-translation.

DNA Sequencing and Sequence Analyses

DNA sequencing and sequence analyses were done as previously described (Kapke and Brian, 1986).

RESULTS

Six clones named C4, F5, E2, FT36, FT35, and FT43, mapping in the positions illustrated in Fig. 1, were sequenced in part to extend the TGEV genomic sequence that was known from clones FG5 and J21 (Kapke and Brian, 1986). Clone FG5 maps at the extreme 3' end of the genome and contains the sequence for the hypothetical hydrophobic protein gene, the N gene and part of the M gene. Identification of the third open reading frame as the M gene sequence was based on regions of extensive amino acid homology with the M proteins of MHV and IBV. The sequencing strategy we used is described in Fig. 1.

The molecular weight of the glycosylated M protein has been estimated to be approximately 29 Kd to 30 Kd (Brian et al., 1983; Garwes and Pocock, 1975; Wesley and Woods, 1986). We therefore anticipated that we would be able to deduce from the gene sequence a molecular weight of 29 Kd or less for the unglycosylated protein. The extended sequence of what we identified earlier as part of the open reading frame for the M gene

(Kapke and Brian, 1986) has not yielded an unequivocal demarcation for the amino terminus of the M protein (Fig. 2). The nucleotide sequence derived from the 5' end of clone FT36 yields a continuous open reading frame beginning at base position 56 and continuing through the postulated carboxy terminus of the M protein identified as base number 922 in Fig. 2 (Fig. 3). A protein produced by this open reading frame would contain 289 amino acids and have a molecular weight of greater than 32 Kd. Although possible, it is unlikely that this polypeptide represents the species identified earlier in protein analyses because of its large size. At least three possibilities exist. 1. There is an error in our sequence that disguises a stop codon. This is entirely possible especially early in the sequence since the first 210 bases come from only one clone (FT36) and need yet to be confirmed by further sequencing. 2. A precursor polypeptide of greater than 29-30 Kd is made and rapidly processed by proteolytic cleavage to yield a 29-30 Kd product. 3. There is, in fact, an open reading frame that is larger than necessary in the genome, but a message of the proper size for the M protein is generated by a transcriptional initiation signal.

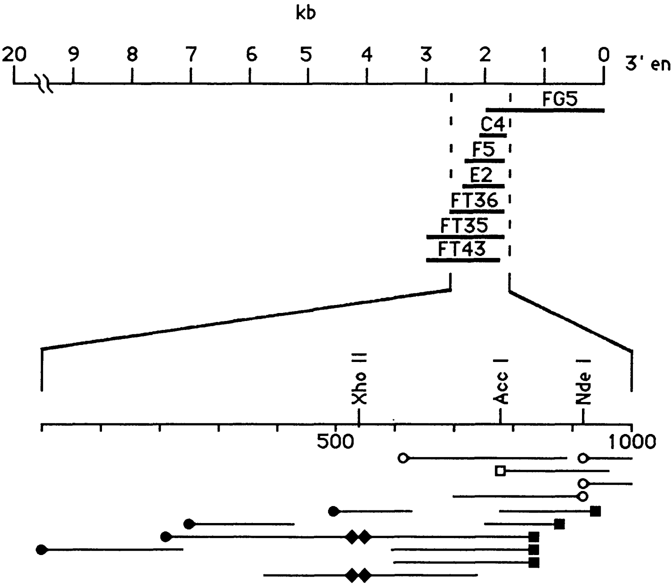

Fig. 1. Sequencing strategy used to derive the TGEV M gene sequence. cDNA clones FG5, C4, F5, E2, FT36, and FT35 were cloned into the PST I site of vector pUC9 and were all found to be in the same orientation with respect to the virus genomic RNA illustrated at the top of the figure. FT43 was likewise cloned but was found to be in the opposite orientation. Nucleotide position #1 on the restriction map sequence is the first base at the 5' end (virus-sense) of the FT36 insert. ○ and □ indicate sites labeled on fragments of clone FG5 at the 3' end of DNA with reverse transcriptase and at the 5' end with polynucleotide kinase, respectively (Kapke and Brian, 1986). ● indicates 3' end labeling with reverse transcriptase at the Sal I site in the multiple cloning region of clones C4, F5, E2 and FT36. ■ indicates 3' end labeling with reverse transcriptase at the HindIII site in the multiple cloning region of clones C4, F5, E2, FT36 and FT35. ◆ indicates 3' end labeling with reverse transcriptase at the Xho II site in clones E2 and FT43.

Of these possibilities, we have for the purpose of our present analysis chosen the third one to explain our data. The most probable site for initiation of transcription of the M message is suggested by the sequence CTAAAC beginning at base 128 in Fig. 2, which may be part of a conserved intergenic sequence in the TGEV genome. It is found in total and again in part between the M and N genes beginning at base 926 in Fig. 2, and also between the N and hypothetical hydrophobic protein genes (Kapke and Brian, 1986). It is also part of the intergenic sequence found in the MHV genome (Budzilowicz et al., 1985). If CTAAAC is an intergenic sequence that directs leader-primed synthesis and thereby defines the start of the M transcript for TGEV, then the M protein coding sequence could start with the first available methionine which begins at base 137 in Fig. 2. Using this as the amino terminus, the deduced M protein is comprised of 262 amino acids and has a molecular weight of 29,544. The protein is moderately hydrophobic with 44% of its amino acids being hydrophobic, and is basic since it carries a net charge of +7 at neutral pH.

DISCUSSION

Assuming that the TGEV M protein begins with the methionine codon starting at base position 137 in Fig. 2 and ends with the stop codon starting at base position 926, then it has several features that are shared with the M proteins of MHV and IBV, and also some that are in striking contrast. By inspection, regions of high amino acid homology can be found among TGEV, MHV, BCV and IBV proteins. For example, within a 21 amino acid stretch (beginning with amino acid number 132 in the TGEV

```
         .         .        30         .         .        60         .         .        90         .         .       120
CTATGCATGGTGTGTTGCAATTTAGCAAGGACAGTTATTATTGTTCCAGCGCAACATGCTTACGATGCCTATAAGAATTTTATGCGAATTAAAGCATACAACCCCGATGGAGCACTCCTT
                                                        M  L  T  M  P  I  R  I  L  C  E  L  K  H  T  T  P  M  E  H  S  L

         .         .       150         .         .       180         .         .       210         .         .       240
GCTTGAACTAAACAAAATGAAGATTTTGTTAATATTAGCGTGTGTGATTGCATGCGCATGTGGAGAACGCTATTGTGCTATGAAATCCGATACAGATTTGTCATGTCGCAATAGTACAGC
  L  E  L  N  K  M  K  I  L  L  I  L  A  C  V  I  A  C  A  C  G  E  R  Y  C  A  M  K  S  D  T  D  L  S  C  R  N  S  T  A

         .         .       270         .         .       300         .         .       330         .         .       360
GTCTGATTGTGAGTCATGCTTCAACGGAGGCGATCTTATTTGGCATCTTGCAAACTGGAACTTCAGCTGGTCTATAATATTGATCGTTTTTATAACTGTGCTACAATATGGAAGACCTCA
  S  D  C  E  S  C  F  N  G  G  D  L  I  W  H  L  A  N  W  N  F  S  W  S  I  I  L  I  V  F  I  T  V  L  Q  Y  G  R  P  Q

         .         .       390         .         .       420         .         .       450         .         .       480
ATTCAGCTGGTTCGCGTATGGCATTAAAATGCTTATAATGTGGCTATTATGGCCCGTTGTTTTGGCTCTTACGATTTTTAATGCATACTCGGAATACCAAGTGTCCAGATATGTAATGTT
  F  S  W  F  A  Y  G  I  K  M  L  I  M  W  L  L  W  P  V  V  L  A  L  T  I  F  N  A  Y  S  E  Y  Q  V  S  R  Y  V  M  F

         .         .       510         .         .       540         .         .       570         .         .       600
CGGCTTTAGTATTGCAGGTGCAATTGTTACATTTGTACTCTGGATTATGTATTTTGTAAGATCCATTCAGTTGTACAGAAGGACTAAGTCTTGGTGGTCTTTCAACCCTGAAACTAAAGC
  G  F  S  I  A  G  A  I  V  T  F  V  L  W  I  M  Y  F  V  R  S  I  Q  L  Y  R  R  T  K  S  W  W  S  F  N  P  E  T  K  A

         .         .       630         .         .       660         .         .       690         .         .       720
AATTCTTTGCGTTAGTGCATTAGGAAGAAGCTATGTGCTTCCTCTCGAAGGTGTGCCAACTGGTGTCACTCTAACTTTGCTTTCAGGGAATTTGTACGCTGAAGGGTTCAAAATTGCAGG
  I  L  C  V  S  A  L  G  R  S  Y  V  L  P  L  E  G  V  P  T  G  V  T  L  T  L  L  S  G  N  L  Y  A  E  G  F  K  I  A  G

         .         .       750         .         .       780         .         .       810         .         .       840
TGGTATGAACATCGACAATTTACCAAAATACGTAATGGTTGCATTACCTAGCAGGACTATTGTCTACACACTTGTTGGCAAGAAGTTGAAAGCAAGTAGTGCGACTGGATGGGCTTACTA
  G  M  N  I  D  N  L  P  K  Y  V  M  V  A  L  P  S  R  T  I  V  Y  T  L  V  G  K  K  L  K  A  S  S  A  T  G  W  A  Y  Y

         .         .       870         .         .       900         .         .       930         .         .       960
TGTAAAATCTAAAGCTGGTGATTACTCAACAGAGGCAAGAACTGATAATTTGAGTGAGCAAGAAAAATTATTACATATGGTATAACTAAACTTCTAAATGGCCAACCAGGGACAACGTGT
  V  K  S  K  A  G  D  Y  S  T  E  A  R  T  D  N  L  S  E  Q  E  K  L  L  H  M  V                 M  A  N  Q  G  Q  R  V

         .         .       990         .
CAGTTGGGGAGATGAATCTACCAAAACACGTGGTCGTTCC
  S  W  G  D  E  S  T  K  T  R  G  R  S
```

Fig. 2. Nucleotide sequence of the TGEV M gene and deduced amino acid sequence for the protein. The nucleotide sequence comes from the part of the virus genome illustrated in Figure 1. A continuous open reading frame beginning at nucleotide position 56 and continuing through nucleotide 922 is identified. The CTAAAC intergenic sequences are underlined. The proposed amino terminus for the M protein is identified by an underlined methionine residue near base position 137.

sequence) there are regions of 1 to 8 amino acids showing perfect homology among all four viruses, the longest being the sequence SWWSFNPE. When M amino acid sequences are aligned for maximum homology by computer assistance, an amino acid sequence homology of approximately 36% between TGEV and MHV, 37% between TGEV and BCV, and 28% between TGEV and IBV are found (data not shown). Similarly, inspection of hydrophobic amino acid positions suggests that the hydrophobicity patterns conserved between MHV and IBV (Boursnell et al., 1984) are also conserved for TGEV. That is, from its entrance into the virion membrane and as it extends toward its carboxy terminus, the TGEV M protein has three regions of high hydrophobicity that are apparently transmembrane and a relatively hydrophilic carboxy terminal region that is intravirion (Rottier et al., 1986).

External to the virion envelope, however, the TGEV M sequence contrasts with those of MHV and IBV. Assuming a parallel structure for the M proteins of the three viruses and assuming the MHV M protein enters the virion envelope at position 26, then the external amino terminal portion is 21 amino acids for IBV and 54 for TGEV. Within the 54 amino acids there are three asparagine residues but only one at position 32 has the proper surrounding sequence for glycosylation (Hubbard and Ivatt, 1981). There are 5 serine residues within the first 54 amino acids and these are potential O-glycosylation sites. Only asparagine-linked glycosylation has been reported for the TGEV M protein, however (Jacobs et

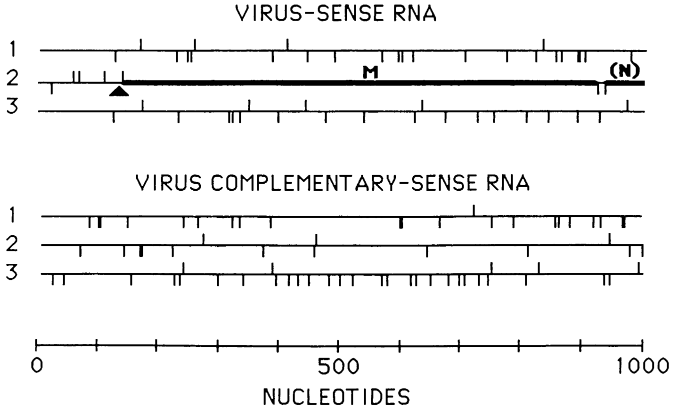

Fig. 3. Schematic diagram of possible open reading frames obtained when translating the nucleotide sequence illustrated in Fig. 2 as either virus-sense RNA or virus complementary-sense RNA in all three reading frames. Vertical bars above the line represent the first methionine codon that could serve as the initiation site for translation. In the M open reading frame the initiating methionine residue identified is the first one to follow the putative CTAAAC intergenic sequence. The CTAAAC intergenic sequence is identified with an arrowhead. Vertical bars below the line represent termination codons. M, open reading frame for the matrix protein. (N), partial open reading frame for the nucleocapsid protein.

al., 1986). The external portion of the protein is mostly hydrophilic except for the amino terminal region which is hydrophobic for a distance of 15 amino acids.

If the external portion of the TGEV M protein is in fact 54 amino acids, then the M protein may well take part in inducing immunity since there is ample exposure for interaction with antibody. The role of the M protein in virus replication and in immunity induction is the subject of continuing examination in our laboratories.

ACKNOWLEDGEMENTS

This work was supported by grant AI-14367 from the National Institute of Allergy and Infectious Diseases, by grant 82-CRSR-2-1090 from the U.S. Department of Agriculture, and in part by a grant from the National Foundation for Ileitis and Colitis, Inc.

REFERENCES

Armstrong, J., Smeekens, S., and Rottier, P., 1983, Sequence of the nucleocapsid gene from murine coronavirus MHV-A59, Nucl. Acids Res., 11:833.

Boursnell, M. E. G., Brown, T. D. K., and Binns, M. M., 1984, Sequence of the membrane protein gene from avian coronavirus IBV, Virus Res., 1:303.

Brian, David A., Brenda G. Hogue, William Lapps, Barbara J. Potts, and Paul A. Kapke. 1983. Comparative structure of coronaviruses. In "Proceedings from the Fourth International Symposium on Neonatal Diarrhea." (Ed. S. D. Acres) University of Saskatoon, Saskatoon, Saskatchewan, Canada.

Brown, T. D. K. and Boursnell, M. E. G., 1984, Avian infectious bronchitis virus genomic RNA contains sequence homologies at the intergenic boundaries, Virus Res., 1:15.

Budzilowicz, C. J., Wilczynski, S. P., and Weiss, S. R., 1985, Three intergenic regions of coronavirus mouse hepatitis virus strain A59 genome RNA contain a common nucleotide sequence that is homologous to the 3' end of the viral mRNA leader sequence, J. Virol., 53:834.

Garwes, D. J., Lucas, M. H., Higgens, D. A., Spike, B. V., and Cartwright, S. F., 1978, Antigenicity of structural components from porcine transmissible gastroenteritis virus, Vet. Microbiol., 3:179.

Garwes, D. J. and Pocock, D. H., 1975, The polypeptide structure of transmissible gastroenteritis virus, J. Gen. Virol., 29:25.

Gubler, U. and Hoffman, B. J., 1983, A simple and very efficient method for generating cDNA libraries, Gene, 25:263.

Hubbard, S. C. and Ivatt, R. J., 1981, Synthesis and processing of asparagine-linked oligosaccharides, Ann. Rev. Biochem., 50:555.

Jacobs, L., Van der Zeijst, B. A. M., and Horzinek, M. C., 1986, Characterization and translation of transmissible gastroenteritis virus mRNAs, J. Virol., 57:1010.

Kapke, P. A. and Brian, D. A., 1986, Sequence analysis of the porcine transmissible gastroenteritis coronavirus nucleocapsid protein gene, Virology, 151:41.

Wesley, R. D., and R. D. Woods. 1986. Identification of a 17,000 molecular weight antigenic peptide in transmissible gastroenteritis virus. J. Gen. Virol. 67:1419-1425.