J. gen. Virol. (1987), **68**, 1883–1890. *Printed in Great Britain*

Key words: *TGEV/peplomer protein E2/gene sequence*

The Predicted Primary Structure of the Peplomer Protein E2 of the Porcine Coronavirus Transmissible Gastroenteritis Virus

By DENIS RASSCHAERT AND HUBERT LAUDE*

Institut National de la Recherche Agronomique, Station de Recherches de Virologie et d'Immunologie, F-78850 Thiverval-Grignon, France

(*Accepted 31 March 1987*)

SUMMARY

The complete nucleotide sequence of cloned cDNAs containing the E2 glycoprotein-encoding region of the genome of transmissible gastroenteritis virus (TGEV) has been determined. A single large translatable frame of 4·3 kb starting at 8·2 kb from the 3′ end of the genome was identified. Its deduced amino acid sequence contains the characteristic features of a coronavirus peplomer protein: (i) the precursor polypeptide of TGEV E2 is 1447 residues long (i.e. 285 longer than the avian infectious bronchitis coronavirus spike protein); (ii) partial N-terminal sequencing demonstrated that a putative secretory signal sequence of 16 amino acids is absent in the virion-associated protein; (iii) the predicted mol. wt. of the apoprotein is 158K; most of the 32 potential *N*-glycosylation sites available in the sequence are presumed to be functional to account for the difference between this and the experimentally determined value (200K to 220K); (iv) a typical hydrophobic sequence near the C terminus is likely to be responsible for anchoring the peplomer to the virion envelope.

INTRODUCTION

Transmissible gastroenteritis virus (TGEV), a highly enteropathogenic virus of pigs, belongs to the family Coronaviridae, a group of enveloped viruses with a large, positive-stranded RNA genome (see Siddell *et al.*, 1983). Studies on the organization and expression of the TGEV genome (Dennis & Brian, 1982; Hu *et al.*, 1984; Jacobs *et al.*, 1986; Kapke & Brian, 1986; Rasschaert *et al.*, 1987) tend to confirm the findings reported for two other members of the Coronaviridae, murine hepatitis (MH) and avian infectious bronchitis (IB) viruses (see Laude *et al.*, 1987). Three functional classes of polypeptides have been identified in TGEV virions: a nucleocapsid protein, a matrix protein and a peplomer protein forming the characteristic surface projections (Garwes *et al.*, 1976). The peplomer protein E2, a highly glycosylated polypeptide of 200K to 220K, has been shown to elicit the production of neutralizing antibodies (Laude *et al.*, 1986; Jimenez *et al.*, 1986; Garwes *et al.*, 1987) which are able to confer protection on suckling piglets (Garwes *et al.*, 1978/79). At least four main antigenic sites have been defined on the E2 protein by means of topographical and functional mapping with monoclonal antibody probes (Delmas *et al.*, 1986). Most of the neutralization-mediating determinants appeared to be grouped in two related sites, both conserved between virus strains. The majority of the epitopes critical in neutralization appeared to be sensitive to denaturation (Jimenez *et al.*, 1986; Garwes *et al.*, 1987). In addition, it is anticipated that the peplomer bears virulence-modulating determinants, as has been suggested in the case of MHV-JHM (Fleming *et al.*, 1986).

Substantial information on peplomer functional organization has been accumulated for other coronaviruses. The spike protein of IBV comprises two or three copies each of two glycopeptides S1 (90K) and S2 (84K). S2 anchors the peplomer to the viral envelope through a short C-terminal hydrophobic domain and S1 is non-covalently attached to S2 (Cavanagh, 1983; Binns *et al.*, 1985). Neutralizing and haemagglutinating antibodies bind to the S1 subunit (Mockett *et al.*,

1984). Removal of S1 abolished infectivity but not attachment to cells (Cavanagh & Davis, 1986). Recently, comparison of the amino acid sequences of two IBV strains, Beaudette (Binns *et al.*, 1985) and M41 (Niesters *et al.*, 1986), led to the proposal that two candidates for neutralization epitopes are located near the S1 N terminus (Niesters *et al.*, 1986). In MHV, the 180K E2 protein is cleavable by host cell proteases or by trypsin to form two comigrating products of 90K. Palmitic acid has been found to be covalently attached to one of the 90K species, probably defining the membrane-anchored subunit. Proteolytic cleavage of E2 may be required for membrane fusion activities (Sturman *et al.*, 1985; Frana *et al.*, 1985). Three independent studies on MHV (Talbot & Buchmeier, 1985) and TGEV (Delmas *et al.*, 1986; Jimenez *et al.*, 1986) characterized an epitope resistant to antibody selection which might be essential for productive infection.

In this paper we present the complete sequence of the E2 gene of TGEV. The main features predicted from the primary structure of the encoded protein are described and compared with those previously reported for the IBV spike protein.

METHODS

cDNA cloning and sequencing. The strategy and protocol were as reported previously (Laude *et al.*, 1987). Briefly, purified genomic RNA was copied by reverse transcriptase using either oligo-d$(T)_{12-18}$ or a specific 30-mer primer (pE2: 5′ CATCATCCTTAACAAAATTCTCTAGCAGAA). RNase T2-treated cDNA–RNA hybrids were dC-tailed and inserted in *Pst*I-cut dG-tailed pBR322. Transfection of *Escherichia coli* RR1 and selection of recombinant clones were performed following standard methods. 'Shotgun' DNA sequencing by Sanger's chain termination method and sequence analysis were accomplished as described previously (Laude *et al.*, 1987); part of the 6.47 clone was sequenced using a 15-mer oligonucleotide (p47) instead of the M13mp18 universal primer. Synthetic oligonucleotides were obtained by the beta amidite method using a Biosearch 8600 DNA synthesizer. DNA has been sequenced at least twice on each strand.

N-terminal sequencing of protein E2. Virion polypeptide E2 resolved by SDS–PAGE was purified by electroelution as described (Laude *et al.*, 1987), and about 100 pmol were analysed in a 'gas phase' Applied Biosystems 470A apparatus.

RESULTS

The coordinates of the four sequenced cDNA clones on the restriction map of the genome are given in Fig. 1. (The pTG clones 6.3 and 6.47 were derived using pE2 as a primer.) In Northern blot analyses, the pTG2.26 insert was shown to contain sequences that hybridized only with the two largest RNA species detected in TGEV-infected cells: the genomic RNA and subgenomic RNA 2 (data not shown). DNA sequencing led to the identification of a 4341 base open reading frame (ORF), which was an obvious candidate for the E2 gene. The sequences encompassing the E2 ORF are presented in Fig. 2. A characteristic feature is that it is flanked by an identical sequence 5′ **ACTAAACTT** 3′ at each end. A homologous sequence has been identified within each intergenic junction (Rasschaert *et al.*, 1987). The first consensus sequence is located 25 bases upstream from the potential ATG initiation codon, and maps at 8·25 kb from the 3′ end of the genome, which is in agreement with the size estimates for RNA 2 (Hu *et al.*, 1984; Jacobs *et al.*, 1986; Rasschaert *et al.*, 1987). The second consensus sequence is followed 22 bases downstream by an ORF putatively located at the 5′ end of RNA 3 (partly shown in Fig. 2). The deduced sequence of the 1447 amino acid primary translation product and its hydrophilicity profile are shown in Fig. 2 and 3 respectively. A hydrophobic stretch with the characteristics of an eukaryotic signal peptide is predicted at the N terminus (Von Heijne, 1986). Indeed, partial N-terminal microsequencing demonstrated that the first 16 residues were absent from the virion-associated E2 protein, as its N-terminal sequence was found to be XXFPCSKLTXRTIGNQ. Accordingly, the mature product would be 1431 amino acids long. It has a predicted mol. wt. of 158316, comprising 126 acidic and 91 basic residues. There are 33·5% hydrophobic residues. Thirty-two sites for *N*-glycosylation (Asn–X–Ser or Asn–X–Thr) occur in the sequence, involving as many as 27·5% of the available Asn residues. Most of them are associated with a hydrophilic segment of the E2 polypeptide (Fig. 3).

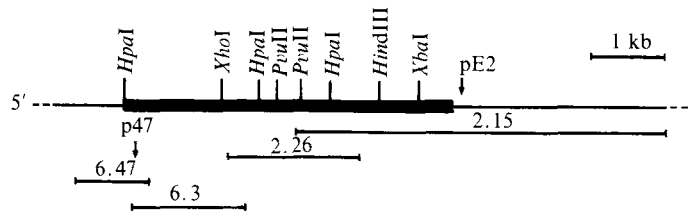


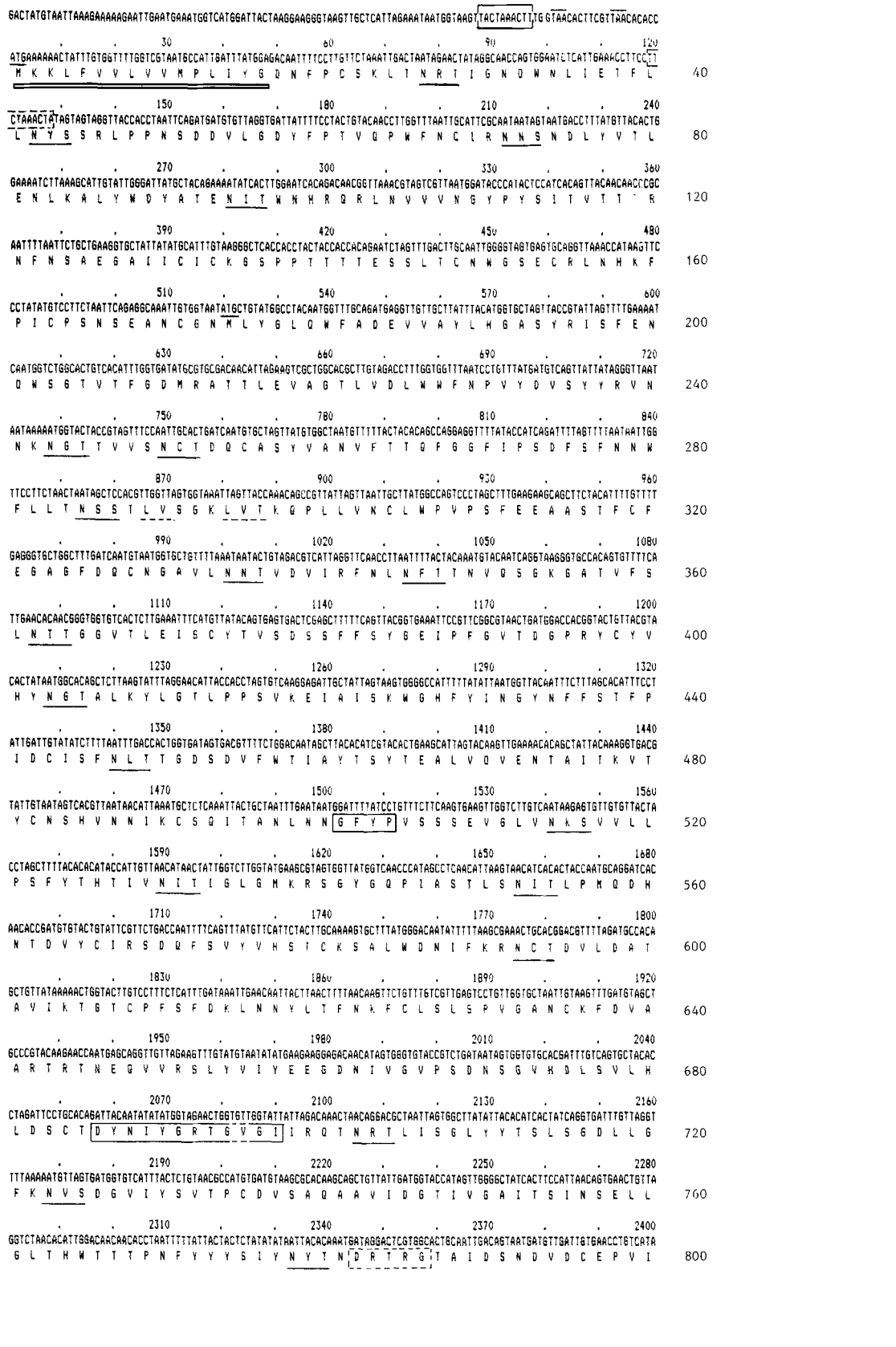
Fig. 1. Restriction map of the region of TGEV genome encoding the E2 gene (bar). The positions of the four cDNA clones and of the two primers used are shown.

DISCUSSION

cDNA copies of the TGEV genome covering the 5′ coding region of mRNA 2 were sequenced. A single large translation frame was found, yielding a 1447 amino acid product with the characteristics of a coronavirus peplomer glycoprotein. The other identified ORFs did not exceed 200 bases and are within the E2 gene (not shown). The first ATG of the E2 ORF is positioned 24 bases downstream from a consensus sequence which is assumed to be the start of the mRNA 2 transcript (Fig. 2). The sequence upstream from the ATG codon (CACCATGA) is optimal for initiation by eukaryotic ribosomes [(CC)ACCATGG; Kozak, 1986]. Moreover the first Met is followed by a leader sequence that has been shown to be removed from the mature protein. The deduced cleavage site for signal peptidase is located after the 16th residue, between Gly and Asp. Inspection of the nucleotide data revealed the occurrence, 120 bases downstream, of an additional consensus sequence, TTCTAAACTA, which could function in the initiation of transcription. However, the next ATG in frame with the E2 ORF occurs only at position 520 and is not followed by a peptide sequence likely to translocate E2 into the membrane. Comparison of our results with previously reported partial nucleic acid data indicates a discrepancy at the 3′ end of the E2 ORF, where the ORF is 3·9 kb long with GCCATGA at the 3′ terminus (Hu *et al.*, 1984). Our data prove that the sequence GCCTAGA occurs instead, at 3·8 kb from the initiation codon. Hence, the ORF extends up to a double stop sequence CCATTAAATTTAA occurring at 4·3 kb, and thereby includes the sequences predicting the anchor structure of the protein (see below).

The deduced mol. wt. of the virion-associated E2 is 158K (aproprotein), a value in close agreement with the M_r 160K determined for the TGEV E2 unglycosylated form in tunicamycin-treated cells (Jacobs *et al.*, 1986; B. Delmas & H. Laude, unpublished results). The 130K M_r species detected by translation of mRNA 2 in reticulocyte lysates might thus correspond to an incomplete translation product (Jacobs *et al.*, 1986). In the mature polypeptide, the carbohydrate moiety should approach 27% of the total weight, implying that a large proportion of the 32 potential sites for *N*-glycosylation are functional. An equivalent high sugar content has been reported for the IBV spike protein (Binns *et al.*, 1985).

The hydrophilicity profile predicts that E2 is hydrophobic overall (Fig. 3), reflecting the spatial importance of the tightly packed core in the peplomer. The amino half of the E2 chain shows few highly hydrophilic (virtually exposed) segments, whereas several prominent peaks are visible in the carboxy half. Examination of the sequence near the C terminus reveals the presence of a highly hydrophobic segment, comprising 45 unpolar residues, including 11 cysteines (Fig. 2). A similar structure has been previously noted in the IBV spike protein (44 hydrophobic residues including six Cys; Binns *et al.*, 1985). Such a high ratio of cysteine residues (24·5% as compared to 3·4% for the whole molecule) in the presumptive anchor region of the peplomer seems so far to be a distinctive feature of the coronaviruses. In both the viruses, the Cys residues cluster mainly in the carboxy distal half part of the hydrophobic domain. Hypothetically, these residues may serve as a site for covalent linkage of fatty acid chains (see Schmidt, 1983), as one E2 subunit of MHV has been reported to be acylated (Sturman *et al.*, 1985). Moreover, an eight residue segment, KWPWYVWL, probably corresponding to the site of entry into the membrane, is perfectly conserved in TGEV and IBV (Fig. 2). In both cases, this



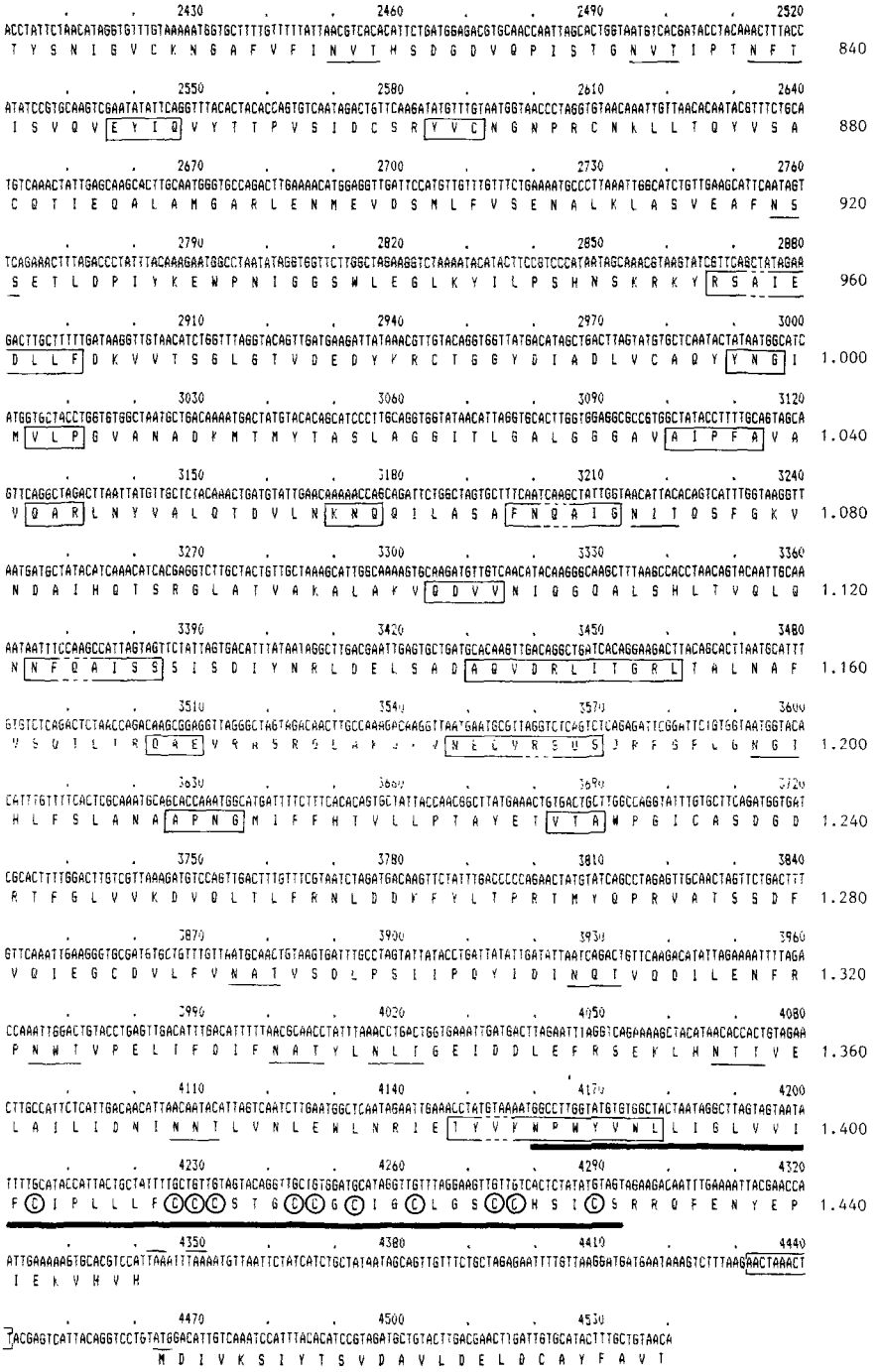


Fig. 2. The nucleotide sequence and the predicted amino acid sequence of the peplomer protein of the Purdue-115 strain of TGEV. The consensus decanucleotides (see text) are boxed. Proximal ATG codons are underlined, stop codons are overlined. Amino acids are numbered at the right. Potential sites for *N*-glycosylation (NXT or NXS) are underlined. The potential signal peptide and membrane-anchoring domain are indicated by open and closed lines respectively. Homology regions (at least three consecutive matches) with the spike sequence of the Beaudette strain of IBV are boxed.

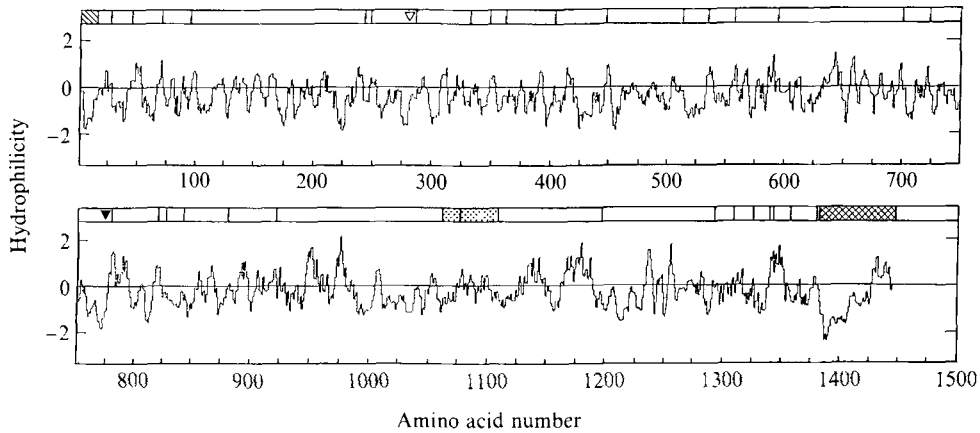


Fig. 3. Hydrophilicity plot of the TGEV E2 precursor polypeptide. Running average taken over an hexapeptide using the hydrophilicity values of Hopp & Woods (1981). Bars in the upper panels indicate the *N*-glycosylation sites; hatched areas represent the signal peptide and the anchoring domain respectively; dotted area indicates a predicted long amphipathic α-helix; the relative positions of the IBV spike protein signal (▽) and connecting (▼) peptides are indicated.

	a	*b*	*c*	*d*	*e*	*f*	*g*
Hydrophobic residues	5/7	7/7	3/7	3/7	6/7	2/7	0/7
1063-L	A	S	A	F	N	Q	
A	I	G	N	I	T	Q	
S	F	G	K	V	N	D	
A	I	H	Q	T	S	R	
G	L	A	T	V	A	K	
A	L	A	Q	V	Q	D	
V	V	N	I	Q	G	Q-1110	

Fig. 4. Search for a stable elongated structure in the TGEV peplomer. The amino acids are listed horizontally following a heptad pattern (two α-helix turns). Residues in the columns *b* and *e* may form the interface between the chains in a α-helical coiled-coil structure.

region of E2 is preceded by a cluster of *N*-glycosylation sites, starting at a markedly hydrophilic stretch of 20 amino acids (Fig. 3). The C-terminal hydrophilic segment of TGEV E2 (16 amino acids) is significantly shorter than that reported for both the Beaudette and M41 strains of IBV (Binns *et al.*, 1985; Niesters *et al.*, 1986). Recently de Groot *et al.* (1987) have described the presence of two heptad repeats in the peplomer proteins of MHV, IBV and feline infectious peritonitis virus, which are indicative of a coiled-coil structure. This structure could provide an explanation for the elongated shape of the peplomer. A Fourier transform of the distribution of hydrophobic residues in the TGEV E2 chain allowed us to characterize a segment of about 55 residues having a strong propensity to form an amphipathic structure with dominant periodicity of 100° ± 20° (De Lisi & Berzofsky, 1985). This segment is located in a region of E2 which is devoid of both Pro and Cys residues (1037 to 1184). In addition, few aromatic residues are present in the heptapeptide repeat (Fig. 4). This predicts an 8 nm long α-helical, possibly coiled-coil segment.

Three other features were noted while aligning optimally the TGEV and IBV E2 protein sequences (not shown). First, the overall highest homology obtained by Dayhoff's alignment is 32·3% (with 12·5% residues unmatched) which is consistent with the fact that the two viruses belong to separate antigenic groups. Most of the stringent homology regions (boxed in Fig. 2) cluster in the carboxy halves of the molecules. In particular, a hydrophilic stretch of 11 residues at position 1144 (TGEV precursor) is perfectly conserved. The sequences are markedly divergent in the amino part, except for one conserved region at positions 686 to 697. Second, a

basic sequence DRTRG occurs in TGEV E2 at position 782, at about the same distance from the C terminus as the sequence RRFRR in the IBV spike protein, where the S1/S2 cleavage site has been demonstrated (Cavanagh *et al.*, 1986). Third, the predicted TGEV E2 mature protein contains 287 residues more than the IBV protein, a difference expected from the comparison of their respective M_r values. The characteristic Lys–Val–Thr twofold repeat present in the IBV signal peptide is conserved in the TGEV E2 homologous sequence (positions 289 to 296). This, along with a tentative alignment of the sequences, suggests that the extra TGEV E2 sequence largely protrudes at the NH_2 terminus. Whether or not a specific function is associated with this sequence is not clear.

This paper, together with two papers reporting the sequences of the nucleocapsid N (Kapke & Brian, 1986) and the transmembrane E1 proteins (Laude *et al.*, 1987), provides a complete set of data on the major structural proteins of TGEV. The availability of cloned TGEV peplomer sequences, along with a panel of monoclonal antibodies and of neutralization-resistant mutants will allow localization of functionally important epitopes.

We thank J. Gelfi for her intelligent technical assistance, and J. C. Huet and J. C. Pernollet (Laboratoire d'Étude des Protéines, I.N.R.A., Versailles) for their helpful collaboration in N-terminal sequencing. The help of F. Borras-Cuesta (Laboratoire de Virologie, Thiverval-Grignon), who set up the program for prediction of the amphipathic α-helix, is gratefully acknowledged. Thanks are also due to A. Kumar for revising the English manuscript. Part of the results were presented at the Third International Coronavirus Symposium (Asilomar, September 1986).

REFERENCES

BINNS, M. M., BOURSNELL, M. E. G., CAVANAGH, D., PAPPIN, D. J. C. & BROWN, T. D. K. (1985). Cloning and sequencing of the gene encoding the spike protein of the coronavirus IBV. *Journal of General Virology* **66**, 719–726.

CAVANAGH, D. (1983). Coronavirus IBV: structural characterization of the spike protein. *Journal of General Virology* **64**, 2577–2583.

CAVANAGH, D. & DAVIS, P. J. (1986). Coronavirus IBV: removal of spike glycopolypeptide S1 by urea abolishes infectivity and haemagglutination but not attachment to cells. *Journal of General Virology* **67**, 1443–1448.

CAVANAGH, D., DAVIS, P. J., PAPPIN, D. J. C., BINNS, M. M., BOURSNELL, M. E. G. & BROWN, T. D. K. (1986). Coronavirus IBV: partial amino terminal sequencing of spike polypeptide S2 identifies the sequence Arg-Arg-Phe-Arg-Arg at the cleavage site of the spike precursor propolypeptide of IBV strains Beaudette and M41. *Virus Research* **4**, 133–143.

DE GROOT, R. J., LENSTRA, J. A., JACOBS, L., LUYTJES, W., NIESTERS, H. G. M., HORZINEK, M. C., VAN DER ZEIJST, B. A. M. & SPAAN, W. J. M. (1987). Structure and evolution of coronavirus peplomer proteins. In *Biochemistry and Biology of Coronaviruses*. Edited by M. M. C. Lai & S. Stohlman. New York & London: Plenum Press (in press).

DE LISI, C. & BERZOFSKY, J. (1985). T-cell antigenic sites tend to be amphipathic structures. *Proceedings of the National Academy of Sciences, U.S.A.* **82**, 7048–7052.

DELMAS, B., GELFI, J. & LAUDE, H. (1986). Antigenic structure of transmissible gastroenteritis virus. II. Domains in the peplomer glycoprotein. *Journal of General Virology* **67**, 1405–1418.

DENNIS, D. E. & BRIAN, D. A. (1982). RNA-dependent RNA polymerase activity in coronavirus-infected cells. *Journal of Virology* **42**, 153–164.

FLEMING, J. O., TROUSDALE, M. D., EL-ZAATARI, F. A. K., STOHLMAN, S. A. & WEINER, L. P. (1986). Pathogenicity of antigenic variants of murine coronavirus JHM selected with monoclonal antibodies. *Journal of Virology* **58**, 869–875.

FRANA, M. F., BEHNKE, J. N., STURMAN, L. S. & HOLMES, K. V. (1985). Proteolytic cleavage of the E2 glycoprotein of murine coronavirus: host-dependent differences in proteolytic cleavage and cell fusion. *Journal of Virology* **56**, 912–920.

GARWES, D. J., POCOCK, D. H. & PIKE, B. V. (1976). Isolation of subviral components from transmissible gastroenteritis virus. *Journal of General Virology* **32**, 283–294.

GARWES, D. J., LUCAS, M. H., HIGGINS, D. A., PIKE, B. V. & CARTWRIGHT, S. F. (1978/79). Antigenicity of structural components from porcine transmissible gastroenteritis virus. *Veterinary Microbiology* **3**, 179–190.

GARWES, D. J., STEWART, F. & ELLEMAN, C. J. (1987). Identification of epitopes of immunological importance on the peplomer of porcine transmissible gastroenteritis virus. In *Biochemistry and Biology of Coronaviruses*. Edited by M. M. C. Lai & S. Stohlman. New York & London: Plenum Press (in press).

HOPP, T. P. & WOODS, K. R. (1981). Prediction of protein antigenic determinants from amino acid sequences. *Proceedings of the National Academy of Sciences, U.S.A.* **78**, 3824–3828.

HU, S., BRUSZEWSKI, J., BOONE, T. & SOUZA, L. (1984). Cloning and expression of the surface glycoprotein gp195 of porcine transmissible gastroenteritis virus. In *Modern Approaches to Vaccines*, pp. 219–223. Edited by R. M. Chanock & R. A. Lerner. New York: Cold Spring Harbor Laboratory.

JIMENEZ, G., CORREA, I., MELGOSA, M. P., BULLIDO, M. J. & ENJUANES, L. (1986). Critical epitopes in transmissible gastroenteritis virus neutralization. *Journal of Virology* **60**, 131–139.

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

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

LAUDE, H., CHAPSAL, J.-M., GELFI, J., LABIAU, S. & GROSCLAUDE, J. (1986). Antigenic structure of transmissible gastroenteritis virus. I. Properties of monoclonal antibodies directed against virion proteins. *Journal of General Virology* **67**, 119–130.

LAUDE, H., RASSCHAERT, D. & HUET, J.-C. (1987). Sequence and N-terminal processing of the transmembrane protein E1 of the coronavirus transmissible gastroenteritis virus. *Journal of General Virology* **68**, 1687–1693.

MOCKETT, A. P. A., CAVANAGH, D. & BROWN, T. D. K. (1984). Monoclonal antibodies to the S1 spike and membrane proteins of avian infectious bronchitis coronavirus strain Massachusetts M41. *Journal of General Virology* **65**, 2281–2286.

NIESTERS, H. G. M., LENSTRA, J. A., SPAAN, W. J. M., ZIJDERVELD, A. J., BLEUMINK-PLUYM, N. M. C., HONG, F., VAN SCHARRENBURG, G. J. M., HORZINEK, M. C. & VAN DER ZEIJST, B. A. M. (1986). The peplomer protein sequence of the M41 strain of coronavirus IBV and its comparison with Beaudette strains. *Virus Research* **5**, 253–263.

RASSCHAERT, D., DELMAS, B., CHARLEY, B., GROSCLAUDE, J., GELFI, J. & LAUDE, H. (1987). Surface glycoproteins of transmissible gastroenteritis virus: functions and gene sequence. In *Biochemistry and Biology of Coronaviruses*. Edited by M. M. C. Lai & S. Stohlman. New York & London: Plenum Press (in press).

SCHMIDT, M. F. G. (1983). Fatty acid binding: a new kind of post-translational modification of membrane proteins. *Current Topics in Microbiology and Immunology* **102**, 101–129.

SIDDELL, S. G., ANDERSON, R., CAVANAGH, D., FUJIWARA, K., KLENK, H. D., MACNAUGHTON, M. R., PENSAERT, M., STOHLMAN, S. A., STURMAN, L. & VAN DER ZEIJST, B. A. M. (1983). Coronaviridae. *Intervirology* **20**, 181–189.

STURMAN, L. S., RICARD, C. S. & HOLMES, K. V. (1985). Proteolytic cleavage of the E2 glycoprotein of murine coronavirus: activation of cell-fusing activity of virions by trypsin and separation of two different 90K cleavage fragments. *Journal of Virology* **56**, 904–911.

TALBOT, P. J. & BUCHMEIER, M. J. (1985). Antigenic variation among murine coronaviruses: evidence for polymorphism on the peplomer glycoprotein, E2. *Virus Research* **2**, 317–328.

VON HEIJNE, G. (1986). A new method for predicting signal sequence cleavage sites. *Nucleic Acids Research* **14**, 4683–4690.

(Received 18 December 1986)