Nucleotide sequence of the glycoprotein S gene of bovine enteric coronavirus and comparison with the S proteins of two mouse hepatitis virus strains

Pascal Boireau,1 Catherine Cruciere1 and Jaques Laporte2*


The gene encoding the spike glycoprotein (S) of bovine enteric coronavirus (BECV) was cloned and its complete sequence of 4092 nucleotides was determined. This sequence contained a single long open reading frame with a coding capacity of 1363 amino acids (Mr 150747). The predicted protein had 19 N-glycosylation sites. A signal sequence comprising 17 amino acids was observed starting from the first methionine residue. A potential peptidase cleavage site was located between amino acids 763 and 767. These cleavages explain the maturation of the primary product of the S gene to S1 (M, 104692) and S2 (M, 84175) spike structural proteins. Two amphipathic α-helices (amino acids 1007 to 1077 and 1269 to 1294) which may constitute the 12 nm stalk of the viral spike were also observed; another α-helix (amino acids 1305 to 1335) may be involved in the anchorage of the spike in the viral membrane. Comparison of this protein sequence to the described homologous mouse hepatitis (MHV) strain A59 and MHV-JHM S protein sequences led us to suggest that MHV-A59 and MHV-JHM S genes could be derived from a deletion of the BECV S gene.

Bovine enteric coronavirus (BECV) was first identified by electron microscopy in faecal samples of calves suffering from acute enteritis (Mebus et al., 1973a). The involvement of BECV in the aetiology of diarrhoeal diseases has been suggested in several studies (Mebus et al., 1973b; Bridger et al., 1978; Gouet et al., 1978). During the acute stage of infection, virus particles are excreted in large amounts and have been identified within brush border cells of the small intestine and in differentiated colonic epithelial cells. Although propagation of BECV is difficult in conventional cell lines (Mebus et al., 1973a), it has been grown successfully in HRT18 cells (human rectal tumour cell line; Laporte et al., 1980).

As a member of the Coronaviridae family BECV is a pleiomorphic, enveloped spherical particle (120 nm in diameter) surrounded by a fringe of 20 nm long club-shaped spikes. The coronavirus genome is a positive and single-stranded capped RNA with a polyadenylated 3′ end (Siddell et al., 1982; Sturman & Holmes, 1983). The structural and non-structural proteins of the virus are translated from a 3′-coterminal nested set of mRNAs, each having a common 5′ leader sequence (Lai et al., 1984). Only the unique 5′-terminal sequence is translated; this sequence is absent from the next smaller RNA of the set.

BECV possesses five main structural proteins, i.e. a phosphorylated nucleocapsid protein (N; 50K), a transmembrane matrix glycoprotein (M; 28K) and three peplomer glycoproteins [S1, S2 (S; spike) and haemagglutinin (HA)]. Glycoproteins S1 (105K) and S2 (95K) are the cleavage products of the S gene-encoded primary product (180K; J. F. Vautherot, unpublished results; Deregt & Babiuk, 1987). HA (M, 125K) is split by reducing agents into two subunits of equal size with an Mr of 65K (Laporte & Bobulesco, 1981; King & Brian, 1982); the neutralizing epitopes are located on S1 and HA (Vautherot et al., 1984).

The cloning and sequencing of the gene encoding the S protein of BECV, a necessary first step for the production of a genetically engineered vaccine against BECV, is reported in this paper.

BECV strain F15 (BECV-F15) was grown in HRT18 cells; virus and genomic RNA purifications were performed as described previously (Cruci6re & Laporte, 1988). The virus genome was used as a template for cDNA synthesis. Poly(dC)-tailed RNA–cDNA heteroduplexes were inserted into a dG-tailed PstI-linearized pBR322 plasmid. Complementary DNAs were then cloned in Escherichia coli RR1; tetracycline-resistant, ampicillin-susceptible colonies were then transferred onto nitrocellulose filters and lysed in situ (Maniatis et al., 0000-8981 © 1990 SGM)
By Northern blot analysis (data not shown), we established that the large cDNA insert G7 (2.4 kb) covered a part of the S gene (Fig. 1); sequence analysis of that clone and comparison with the sequence of the gene coding for the E2 protein of coronavirus mouse hepatitis virus strain JHM (MHV-JHM) (Schmidt et al., 1987) showed that this insert mapped in the middle of the S gene. G7 cDNA was used to screen the cDNA library and to obtain clones. Clones P G7 8 and P 27 40 were used also as probes to identify cDNA clones located at the 5' and the 3' ends of the S gene, respectively (Fig. 1).

Both DNA strands were sequenced on five overlapping cDNA clones: P G7, P G7 8, P G7 8 12, P 27 40 and P 33 23 (Fig. 1a). After it had been established by restriction mapping and Southern blot analysis that these clones covered the whole length of the S gene, M13 dideoxynucleotide sequencing was carried out according to Sanger et al. (1977) by using sonicated cloned fragments subcloned into the SmaI site of the M13 mp10 vector (Deininger, 1983). Buffer gradient gels and [α-35S]dATP were used according to Biggin et al. (1983).

Sequence data were analysed and assembled with the aid of the program of Queen & Korn (1984), the Beckman Microgene program (March 1985 version, Beckman Instruments) adapted for the IBM PC-XT microcomputer. The nucleotide sequence obtained

Fig. 1. (a) Schematic diagram of part of the BECV genome and location of cDNA clones. A simplified restriction map is given; the inserts used as probes to screen the cDNA library are shaded grey; the sequence of the S gene was obtained with the clones that are boxed in the figure. (b) Alignment and comparison of amino acid sequences of MHV-A59, MHV-JHM and BECV S propolypeptides. The percentage homology is given; arrows indicate putative peptidase cleavage sites.
(Fig. 2) contains a single long open reading frame (ORF) in the mRNA sense which extends from the first ATG codon (nucleotides 31 to 33) to nucleotide 4122. This 4092 nucleotide sequence contains 28% A, 16.2% C, 19.7% G and 36.2% T, and has a coding capacity for 1363 amino acids (Fig. 2) with an $M_r$ of 150747 and a proposed pH of 5.6.

The length of the ORF is in the expected size range for the glycoprotein S gene sequence. Furthermore, comparison with the published S gene sequence of coronavirus MHV-A59 (Luytjes et al., 1987) shows homology which increases from 61% at the 5’ end of the ORF (BECV nucleotides 31 to 1513) to 73.8% at the 3’ end (BECV nucleotides 2711 to 4100).

Immediately upstream from the first initiation codon there is a sequence ATCTAAACAT very similar to the conserved intergenic sequences of BECV, MHV-JHM and MHV-A59 (Lapps et al., 1987; Cruciere & Laporte, 1988; Luytjes et al., 1987; Schmidt et al., 1987); it is also closely related to the conserved sequence ACTAAAC, reported for the transmissible gastroenteritis virus (TGEV) (Rasschaert & Laude, 1987). The sequence surrounding the translation initiation codon is in a suboptimal environment (Kozak, 1987). A similar situation has been observed for the initiation codon of the S gene of MHV (Luytjes et al., 1987).

Comparison of the first 400 nucleotides of the PG7 8 12 clone (P. Boireau & J. Laporte, unpublished observations) with the recently published HA-encoding gene sequence of bovine coronavirus (Parker et al., 1989) led us to the conclusion that the S gene is just downstream of the HA gene.

The deduced amino acid sequence of the BECV S protein contains 19 potential N-glycosylation sites. Assuming a mean $M_r$ value of 2100 per carbohydrate chain (Hunter et al., 1983), the $M_r$ of the mature S glycoprotein would be approximately 190600. It appears to be hydrophobic over most of its length (35% hydrophobic amino acids).

This protein shares some properties with S proteins described for other coronaviruses. After the first methionine residue, there is a potential signal sequence with a hydrophobic core of 13 amino acids and a helix-breaking residue, glycine 17 (Watson, 1984). According to the rule established for a potential cleavage site (von Heijne, 1984), a protease could act between amino acid residues 17 and 18. The consequence would be that the first amino acid in the S protein is aspartic acid 18.

Another potential peptidase cleavage site is located in the hydrophilic peak between residues 763 and 769. This sequence, Lys-Arg-Arg-Ser-Val-Arg, is collinear with the experimentally determined cleavage site of the MHV-A59 S protein (Luytjes et al., 1987), and very similar to the cleavage site of the infectious bronchitis virus (IBV) S protein (Binns et al., 1985). Furthermore, this series of basic amino acids resembles the tryptic cleavage sites of peptide prohormones or the F0 protein of Newcastle disease virus (MacGinnis & Morrision, 1986), which are processed in the trans-Golgi apparatus. The coronavirus S protein uses the same cellular metabolic pathway for maturation, and budding of the virus takes place in the Golgi apparatus and in the endoplasmic reticulum membrane.

Tryptic cleavage would explain the maturation of the S protein of BECV: the primary gene product, P150 (i.e. the S polypeptide with an $M_r$ of 148967 without the signal peptide) is glycosylated in the endoplasmic reticulum and in the Golgi apparatus giving rise to a glycoprotein of 188867 (gp190), which is then cleaved into the S1 (104692) and S2 (84175) structural proteins. These results are in agreement with those published by Deregt & Babiuk (1987), using immunoelectrophoresis to study the biosynthesis of gp105 and gp95, and with our own observations (J. F. Vautheret et al., unpublished results).

Using the approach described by de Groot et al. (1987) for the S proteins of feline infectious peritonitis virus, MHV and IBV and by Rasschaert & Laude (1987) for the S protein of TGEV, we also demonstrated two amphipathic $\alpha$-helices for the S protein of BECV; they are located between amino acids 1007 to 1077 and between amino acids 1269 to 1294. These $\alpha$-helices may constitute the stalk of viral spikes, with a length of approximately 12 nm.

Using the method described by Rao & Argos (1986) a hydrophobic transmembrane $\alpha$-helix was also predicted between amino acids 1305 and 1335. Its C-terminal location, the presence of a potential myristylation site on glycine 1333 and comparison with other coronaviruses suggest that this helix is involved in the anchorage of the spike in the viral membrane. The myristylation site, surrounded by eight cysteine residues, would be located at the internal face of the viral membrane. Among coronaviruses this domain is highly conserved in structure, location and length. A stretch of eight amino acids, Lys-Trp-Pro-Trp-Tyr-Val-Trp-Leu (1305 to 1312; Fig. 2), is found in all coronavirus S protein sequences established so far; however its role is unknown.

Luytjes et al. (1987, 1988) have compared S amino acid sequences of the two MHV strains A59 and JHM. These sequences are very similar but the S protein of JHM was found to be shorter. As BECV belongs to the same antigenic group as these viruses, the present results enlarge upon this comparison. Dot matrix analyses (Beckman Microgenie program) of the deduced amino acid sequence of the S genes of BECV and MHV-A59 (data not shown), revealed that there is low homology (55%) between amino acids 488 to 593 of BECV and
There is a deletion in the MHV-A59 S protein sequence. BECV have no counterpart in MHV-A59. Furthermore, amino acids 481 to 545 of MHV-A59 and apparently counterpart in MHV-JHM (Fig. 1b). Amino acids 452 to 593 of the S protein of BECV have no homologous counterpart in MHV-A59 and does not exist in MHV-JHM (Luytjes et al., 1989) whereas only a related HA pseudogene was identified in MHV-A59 (Luytjes et al., 1988); this HA is most likely to occur if the BECV genome is more closely related to a possible common ancestor.

Fig. 2. Nucleotide sequence of the gene encoding the S protein and predicted amino acid sequence of the S protein. The sequence deduced from the cDNA inserts (Fig. 1a) is shown as positive sense DNA from 5' to 3' ends; the hatched bar indicates the proposed N-terminal signal sequence. An arrowhead indicates the potential tryptic cleavage site. A box surrounds the conserved intergenic sequence. Potential N-glycosylation sites are underlined (specific for the Asn-X-Thr/Ser sequence, with X different from Pro). Dots mark the proposed C-terminal membrane anchoring domain. The last eight cysteines are encircled.

Amino acids 481 to 545 of MHV-A59 and apparently there is a deletion in the MHV-A59 S protein sequence. More details emerged after alignment comparison of the two sequences (Fig. 1b). Amino acids 488 to 534 of BECV have no counterpart in MHV-A59. Furthermore, amino acids 452 to 593 of the S protein of BECV have no counterpart in MHV-JHM (Fig. 1b).

Luytjes et al. (1987, 1988) put forward two hypotheses to account for the difference between MHV-A59 and MHV-JHM in this domain of the S protein. The first hypothesis is that the MHV-JHM genome is deleted with respect to a nucleotide sequence corresponding to amino acids 453 to 545 of the S protein of MHV-A59; the second is that the MHV-A59 genome has acquired genomic material by non-homologous recombination. The comparisons presented above support the idea of a genetic instability in this area of the virus genome which would explain the difference in length of the S proteins of the three virus strains. Therefore we suggest an evolutionary progression from MHV-JHM to MHV-A59 to BECV or in reverse depending upon whether non-homologous recombination events or deletions have occurred. The observation that the nucleotide sequence of the S gene encoding amino acids 470 to 480 (data not shown) is highly homologous (81%) between BECV and MHV-JHM and Michel Brémont for helpful discussions. We would like to thank Jean François Vautherot, Denis Rasschaert and Michel Brémont for helpful discussions.
References


(Received 29 March 1989; Accepted 19 October 1989)