Molecular characterization of the S protein gene of human coronavirus OC43

Samir Mounir and Pierre J. Talbot*

Virology Research Center, Institut Armand-Frappier, Université du Québec, 531 boulevard des Prairies, Laval, Québec, Canada H7N 4Z3

The gene encoding the spike protein of the OC43 strain of human coronavirus (HCV-OC43) was cloned and sequenced. The complete nucleotide sequence revealed an open reading frame of 4062 nucleotides encoding a protein of 1353 amino acids with a predicted M_r of 150078. Structural features include 22 *N*-glycosylation sites, an N-terminal hydrophobic signal sequence of 17 amino acids, an hydrophilic cysteine-rich sequence of 35 amino acids near the C terminus, and a potential proteolytic cleavage site (RRSR) between amino acid residues 758 and 759, yielding S1 and S2 segments of 84730 and 65366 M_r, respectively. The predicted amino acid sequence of the spike protein of HCV-OC43 has 91% identity with that of the Mebus strain of bovine coronavirus, revealing more sequence divergence in the putative bulbous part (S1) than in the predicted stem region (S2).

Human coronaviruses (HCV) are enveloped positive-stranded RNA viruses that cause respiratory infections and have been associated with gastrointestinal and neurological disorders (Jouvenne *et al.*, 1992; McIntosh, 1974; Macnaughton & Davies, 1981; Murray *et al.*, 1992; Resta *et al.*, 1985; Stewart *et al.*, 1992; Talbot & Jouvenne, 1992; Tyrrell, 1986). They are categorized into two major antigenic groups, represented by the prototype strains 229E and OC43 (Macnaughton *et al.*, 1981; Wege *et al.*, 1982).

The HCV-OC43 virion is composed of four structural proteins. Three of them are transmembrane proteins: spike (S), membrane (M) and haemagglutinin–esterase (HE). The fourth protein is an internal nucleocapsid (N) protein, possibly associated with the internal portion of the M protein (Sturman *et al.*, 1980). The N protein binds to the virion RNA, forming the nucleocapsid of the virion (Baric *et al.*, 1988). Both coronavirus glycoproteins (S and M) are synthesized in the endoplasmic reticulum on membrane-bound ribosomes (Nieman *et al.*, 1982). The integral membrane M protein interacts with the viral nucleocapsid and is believed to play a role in determining the intracellular site of virus budding. The S glycoprotein mediates binding of virions to the host cell receptor (Williams *et al.*, 1991; Delmas *et al.*, 1992; Yeager *et al.*, 1992), possesses a fusogenic activity, is the major target for antiviral neutralizing antibodies (Spaan *et al.*, 1988; Daniel & Talbot, 1990) and can also be recognized by T lymphocytes (Körner *et al.*, 1991). During maturation and intracellular transport, some S molecules are cleaved by host cell proteases probably located in the Golgi apparatus to yield two large subunits called S1 and S2 (Frana *et al.*, 1985). Primary sequence analysis suggested that the bulbous part of the S protein is formed by the N-terminal half of the molecule, S1 (Cavanagh, 1983; de Groot *et al.*, 1987*a*). The C-terminal half of the S molecule, S2, is anchored in the virion envelope and is predicted to form an intrachain coiled-coil structure via heptad repeat patterns which would give it an elongated stem-like structure (de Groot *et al.*, 1987*a*).

HCV-OC43-infected cells contain a genomic-sized viral mRNA plus eight subgenomic viral mRNA species (Mounir & Talbot, 1993). These mRNAs are arranged in a 3′-coterminal nested-set structure, in which the sequence of every mRNA is contained within the sequence of the next larger mRNA (Lai, 1990) and each mRNA possesses a leader sequence identical to the 5′ end of the genome (S. Mounir & P. J. Talbot, unpublished).

The nucleotide and deduced amino acid sequences of structural and non-structural proteins as well as the leader sequence of HCV-OC43 have been determined (Kamahora *et al.*, 1989; Zhang *et al.*, 1992; Mounir & Talbot, 1992, 1993). Given the key biological importance of the S glycoprotein in coronavirus pathogenesis and its

The nucleotide sequence data reported in this paper have been submitted to the EMBL and GenBank Nucleotide Sequence Databases under accession number L14643.

```
        10        20        30        40        50        60        70        80        90       100
GGCTGCATGATGCTTAGACCATAATCTAAACATGTTTTTGATACTTTTAATTTCCTTACCAACGGCTTTTGCTGTTATAGGAGATTTAAAGTGTACTTCA 100
          *                   M  F  L  I  L  L  I  S  L  P  T  A  F  A  V  I  G  D  L  K  C  T  S   23

GATAATATTAATGATAAAGACACCGGTCCTCCTCCTATAAGTACTGATACTGTTGATGTTACTAATGGTTTGGGTACTTATTATGTTTTAGATCGTGTGT 200
D  N  I  N  D  K  D  T  G  P  P  P  I  S  T  D  T  V  D  V  T  N  G  L  G  T  Y  Y  V  L  D  R  V  Y  57

ATTTAAATACTACGTTGTTTCTTAATGGTTATTACCCTACTTCAGGTTCCACATATCGTAATATGGCACTGAAGGGAAGTGTACTATTGAGCAGACTATG 300
  L  N  T  T  L  F  L  N  G  Y  Y  P  T  S  G  S  T  Y  R  N  M  A  L  K  G  S  V  L  L  S  R  L  W   90
     °

GTTTAAACCACCATTTCTTTCTGATTTTATTAATGGTATTTTTGCTAAGGTCAAAAATACCAAGGTTATTAAAGATCGTGTAATGTATAGTGAGTTCCCT 400
F  K  P  P  F  L  S  D  F  I  N  G  I  F  A  K  V  K  N  T  K  V  I  K  D  R  V  M  Y  S  E  F  P  123

GCTATAACTATAGGTAGTACTTTTGTAAATACATCCTATAGTGTGGTAGTACAACCACGTACAATCAATTCAACACAGGATGGTGATAATAAATTACAAG 500
A  I  T  I  G  S  T  F  V  N  T  S  Y  S  V  V  V  Q  P  R  T  I  N  S  T  Q  D  G  D  N  K  L  Q  G  157
                          °                             °

GTCTTTTAGAGGTCTCTGTTTGCCAGTATAATATGTGCGAGTACCCACAAACGATTTGTCATCCTAACCTGGGTAATCATCGCAAAGAACTATGGCATTT 600
  L  L  E  V  S  V  C  Q  Y  N  M  C  E  Y  P  Q  T  I  C  H  P  N  L  G  N  H  R  K  E  L  W  H  L  190

GGATACAGGTGTTGTTTCCTGTTTATATAAGCGTAATTTCACATATGATGTGAATGCTGATTATTTGTATTTTCATTTTTATCAAGAAGGTGGTACTTTT 700
D  T  G  V  V  S  C  L  Y  K  R  N  F  T  Y  D  V  N  A  D  Y  L  Y  F  H  F  Y  Q  E  G  G  T  F  223
                          °

TATGCATATTTTACAGACACTGGTGTTGTTACTAAGTTTTTGTTTAATGTTTATTTAGGCATGGCGCTTTCACACTATTATGTCATGCCTCTGACTTGTA 800
Y  A  Y  F  T  D  T  G  V  V  T  K  F  L  F  N  V  Y  L  G  M  A  L  S  H  Y  Y  V  M  P  L  T  C  N  257

ATAGTAAGCTTACTTTAGAATATTGGGTTACACCTCTCACTTCTAGACAATATTTACTCGCTTTCAATCAAGATGGTATTATTTTTAATGCTGAAGATTG 900
  S  K  L  T  L  E  Y  W  V  T  P  L  T  S  R  Q  Y  L  L  A  F  N  Q  D  G  I  I  F  N  A  E  D  C  290

TATGAGTGATTTTATGAGTGAGATTAAGTGTAAAACACAATCTATAGCGCCACCTACTGGTGTTTATGAATTAAACGGTTACACTGTTCAGCCAATCGCA 1000
M  S  D  F  M  S  E  I  K  C  K  T  Q  S  I  A  P  P  T  G  V  Y  E  L  N  G  Y  T  V  Q  P  I  A  323

GATGTTTACCGACGTAAACCTAATCTTCCCAATTGCAATATAGAAGCTTGGCTTAATGATAAGTCGGTGCCCTCTCCATTAAATTGGGAACGTAAGACAT 1100
D  V  Y  R  R  K  P  N  L  P  N  C  N  I  E  A  W  L  N  D  K  S  V  P  S  P  L  N  W  E  R  K  T  F  357

TTTCAAATTGTAATTTTAATATGAGCAGCCTGATGTCTTTTATTCAGGCAGACTCATTTACTTGTAATAATATTGATGCTGCTAAGATATATGGTATGTG 1200
  S  N  C  N  F  N  M  S  S  L  M  S  F  I  Q  A  D  S  F  T  C  N  N  I  D  A  A  K  I  Y  G  M  C  390
                    °

TTTTTCCAGCATAACTATAGATAAGTTTGCTATACCCAATGGCAGGAAGGTTGACCTACAATTGGGTAATTTGGGCTATTTGCAGTCATTTAACTATAGA 1300
F  S  S  I  T  I  D  K  F  A  I  P  N  G  R  K  V  D  L  Q  L  G  N  L  G  Y  L  Q  S  F  N  Y  R  423

ATTGATACTACTGCAACAAGTTGTCAGTTGTATTATAATTTACCTGCTGCTAATGTTTCTGTTAGCAGGTTTAATCCTTCTACTTGGAATAAGAGATTTG 1400
I  D  T  T  A  T  S  C  Q  L  Y  Y  N  L  P  A  A  N  V  S  V  S  R  F  N  P  S  T  W  N  K  R  F  G  457
                                               °                       °

GTTTTATAGAAGATTCTGTTTTTAAGCCTCGACCTGCAGGTGTTCTTACTAATCATGATGTAGTTTATGCACAACACTGTTTCAAAGCTCCTAAAAATTT 1500
  F  I  E  D  S  V  F  K  P  R  P  A  G  V  L  T  N  H  D  V  V  Y  A  Q  H  C  F  K  A  P  K  N  F  490

CTGTCCGTGTAAATTGAATGGTTCGTGTGTAGGTAGTGGTCCTGGTAAAAATAATGGTATAGGCACTTGTCCTGCAGGTACTAATTATTTAACTTGTGAT 1600
C  P  C  K  L  N  G  S  C  V  G  S  G  P  G  K  N  N  G  I  G  T  C  P  A  G  T  N  Y  L  T  C  D  523
                   °

AATTTGTGCACTCCTGATCCTATTACATTTACAGGTACTTATAAGTGCCCCCAAACTAAATCTTTAGTTGGCATAGGTGAGCACTGTTCGGGTCTTGCTG 1700
N  L  C  T  P  D  P  I  T  F  T  G  T  Y  K  C  P  Q  T  K  S  L  V  G  I  G  E  H  C  S  G  L  A  V  557

TTAAAAGTGATTATTGTGGAGGCAATTCTTGTACTTGCCGACCACAAGCATTTTTGGGTTGGTCTGCAGACTCTTGTTTACAAGGAGACAAGTGTAATAT 1800
  K  S  D  Y  C  G  G  N  S  C  T  C  R  P  Q  A  F  L  G  W  S  A  D  S  C  L  Q  G  D  K  C  N  I  590

TTTTGCTAATTTTATTTTGCATGATGTTAATAGTGGTCTTACTTGTTCTACTGATTTACAAAAAGCTAACACAGACATAATTCTTGGTGTTTGTGTTAAT 1900
F  A  N  F  I  L  H  D  V  N  S  G  L  T  C  S  T  D  L  Q  K  A  N  T  D  I  I  L  G  V  C  V  N  623

TATGACCTCTATGGTATTTTAGGCCAAGGCATTTTTGTTGAGGTTAATGCGACTTATTATAATAGTTGGCAGAACCTTTTATATGATTCTAATGGTAATC 2000
Y  D  L  Y  G  I  L  G  Q  G  I  F  V  E  V  N  A  T  Y  Y  N  S  W  Q  N  L  L  Y  D  S  N  G  N  L  657
                                   °

TCTACGGTTTTAGAGACTACATAATAAACAGAACTTTTATGATTCGTAGTTGCTATAGCGGTCGTGTTTCTGCGGCCTTTCACGCTAACTCTTCCGAACC 2100
  Y  G  F  R  D  Y  I  I  N  R  T  F  M  I  R  S  C  Y  S  G  R  V  S  A  A  F  H  A  N  S  S  E  P  690
                         °                                                   °

AGCATTGCTATTTCGGAATATTAAATGCAACTACGTTTTTAATAATAGTCTTACACGACAGCTGCAACCCATTAACTATTTTGATAGTTATCTTGGTTGT 2200
A  L  L  F  R  N  I  K  C  N  Y  V  F  N  N  S  L  T  R  Q  L  Q  P  I  N  Y  F  D  S  Y  L  G  C  723
                                      °

GTTGTCAATGCTTATAATAGTACTGCTATTTCTGTTCAAACATGTGATCTCACAGTAGGTAGTGGTTACTGTGTGGATTACTCTAAAAACAGACGAAGTC 2300
V  V  N  A  Y  N  S  T  A  I  S  V  Q  T  C  D  L  T  V  G  S  G  Y  C  V  D  Y  S  K  N  R  R  S  R  757
                °
```

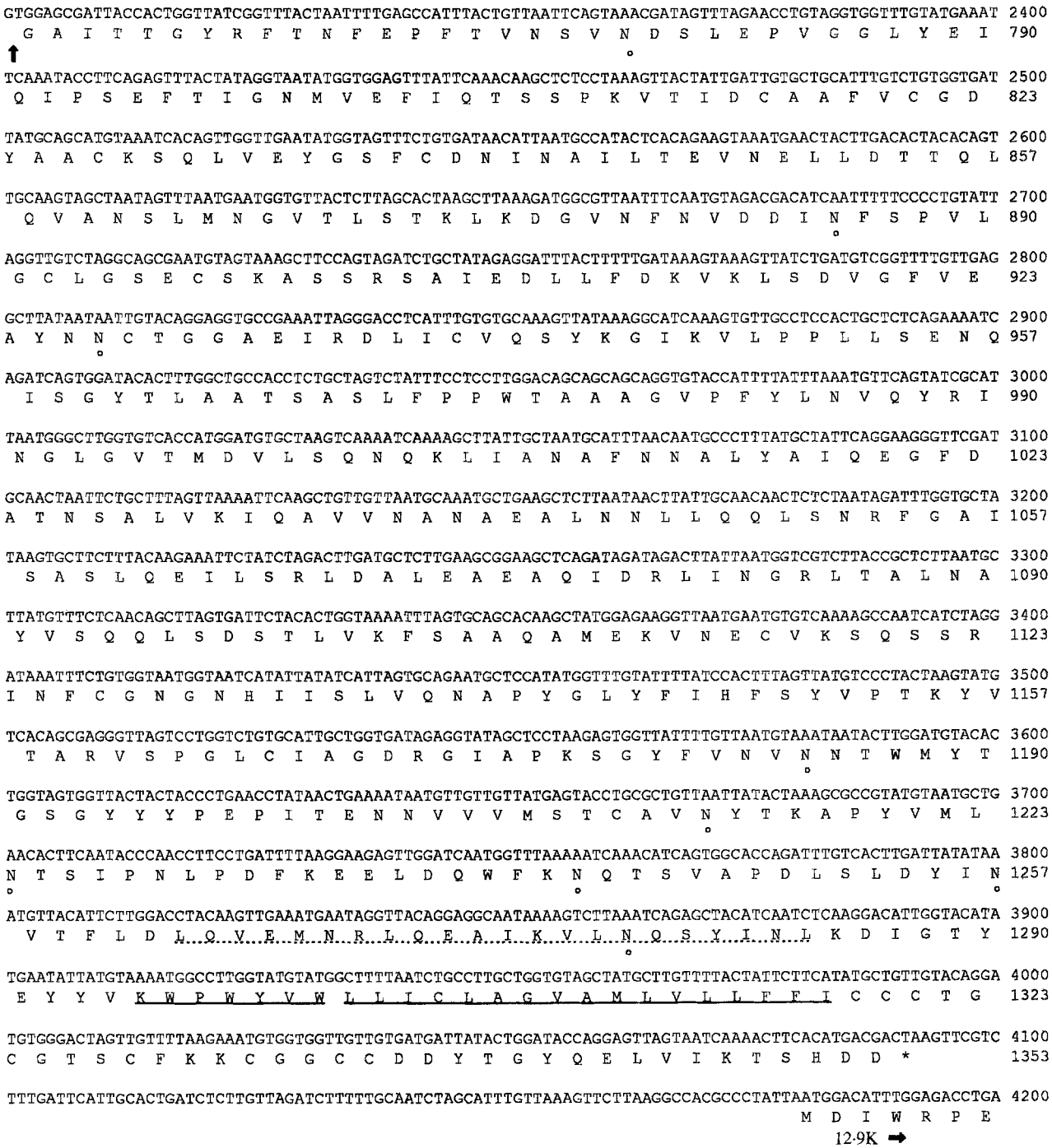

Fig. 1. Complete nucleotide sequence of the S protein gene of HCV-OC43 and its deduced amino acid sequence. The intergenic consensus sequence is doubly underlined. Potential *N*-glycosylation sites (°) are indicated. The N-terminal signal sequence and C-terminal anchor domain are singly underlined. The putative proteolytic cleavage site is indicated by an arrow. Dashes indicate the leucine zipper motif. Asterisks indicate termination codons. The conserved KWPWYVW motif preceding the transmembrane domain is thickly underlined.

interaction with the immune system, as well as the medical importance, both known and suspected, of human coronaviruses, structure–function studies of the S protein are highly important. As a first step, we now report the nucleotide sequence of the S protein gene of HCV-OC43 and compare it with the S protein gene of the closely related bovine coronavirus (BCV).

The origin and cultivation of the HRT-18 cells and the

OC43 strain of HCV as well as the preparation, reverse transcription and PCR amplification of viral RNA were described previously (Mounir & Talbot, 1992). Poly(A)-containing RNA was selected with the PolyATtract mRNA isolation system (Promega) according to the manufacturer's instructions.

Four S gene-specific primers were designed for cDNA synthesis and PCR amplification of the HCV-OC43 S gene, based on the high degree of genomic similarity between HCV-OC43 and BCV (Mounir & Talbot, 1992, 1993). The sense S1H primer represented the sequence 5′ GCTGCATGATGCTTAGAC 3′ (nucleotides 2 to 19, Fig. 1), the sense S2H primer was 5′ GCGATTACCACT-GGTTATCGG 3′ (nucleotides 2306 to 2326, Fig. 1) corresponding to the sequence downstream of the putative proteolytic cleavage site; the antisense S1I primer was 5′ CCGATAACCAGTGGTAATCGC 3′ (nucleotides 2306 to 2326, Fig. 1) and the antisense S2I primer 5′ GGGCGTGGCCTTAAGAAC 3′ (nucleotides 4158 to 4175, Fig. 1). Tandem *Eco*RI sites were added at the 5′ end of each oligonucleotide for cloning purposes.

Different purified PCR products were cloned into the pBluescript II SK$^+$ vector (Stratagene). Unidirectional deletions of the inserts were created using exonuclease III and mung bean nuclease (Pharmacia). Sequencing was performed on both strands of the PCR products by the dideoxyribonucleotide chain termination method (Sanger & Coulson, 1975), with universal, reverse or specific primers corresponding to various regions of the S gene, using T7 DNA polymerase (Pharmacia) and [^{35}S]dATP (Amersham). Sequence analyses were performed as described previously (Mounir & Talbot, 1992).

The complete nucleotide sequence of the HCV-OC43 S gene and its predicted amino acid sequence are shown in Fig. 1, together with some structural features. The sequence begins 15 nucleotides downstream of the termination codon of the HE gene (Zhang *et al.*, 1992). A single open reading frame (nucleotides 32 to 4093) encodes a polypeptide of 1353 amino acids (aa), with a predicted M_r of 150078. The sequence UCUAAAC at nucleotides 25 to 31 is identical to the conserved intergenic sequence of BCV (Abraham *et al.*, 1990), murine hepatitis virus (MHV) strain A59 (Luytjes *et al.*, 1987), MHV-JHM (Schmidt *et al.*, 1987), and almost identical to the ACUAAAC sequence found in transmissible gastroenteritis virus (Rasschaert & Laude, 1987), porcine respiratory coronavirus (Rasschaert *et al.*, 1990), and feline infectious peritonitis virus (de Groot *et al.*, 1987*b*). The deduced aa sequence of the HCV-OC43 S protein contains 22 potential *N*-glycosylation sites, 13 in S1 and nine in S2 (Fig. 1).

The HCV-OC43 S protein shares several properties with S proteins of other coronavirus S proteins. The first initiation codon at nucleotides 32 to 35 is followed by a potential signal peptide with a possible cleavage site (von Heijne, 1984) between aa residues 17 and 18. There are 17 hydrophobic residues near the C terminus (aa 1302 to 1318, Fig. 1) that represent the transmembrane domain. A stretch of eight aa (KWPWYVWL, aa 1295 to 1302; Fig. 1) of unknown function is found in all coronavirus S proteins sequenced to date (Britton, 1991). A leucine zipper motif terminates 10 amino acid residues upstream of this conserved KWPW motif located next to the transmembrane domain (aa 1270 to 1284, Fig. 1). It may be involved in the oligomerization of the S protein (Britton, 1991). A cysteine-rich hydrophilic C terminus of 35 aa (aa 1319 to 1353; Fig. 1), which is probably the intravirion domain, is also found in other coronavirus S proteins (Abraham *et al.*, 1990; Schmidt *et al.*, 1987; Binns *et al.*, 1985; Luytjes *et al.*, 1987; de Groot *et al.*, 1987*b*; Rasschaert & Laude, 1987).

The basic amino acid sequence RRSR (at positions 754 to 757, Fig. 1) is located in the hydrophilic region of the molecule (data not shown). This sequence resembles the BCV (Mebus strain) cleavage site RRSRR (Abraham *et al.*, 1990), the bovine enteric coronavirus F15 strain RRSVR (Boireau *et al.*, 1990) and the infectious bronchitis virus RRFRR (Binns *et al.*, 1985). Cleavage of the S protein would divide the molecule into an N-terminal segment S1 of 84730 M_r and a C-terminal segment S2 of 65366. Assuming a mean M_r of 2100 for addition at each *N*-glycosylation site (Hunter *et al.*, 1983) and the utilization of all sites, the mature S protein would comprise an S1 moiety of 112030 and S2 of 84266, for a total M_r of 196296. This corresponds to the observed sizes (Mounir & Talbot, 1992). Interestingly, most coronavirus S proteins, including that of the Mebus strain of BCV, possess two basic amino acids at the proteolytic cleavage site, whereas HCV-OC43 has only one (Abraham *et al.*, 1990; Cavanagh *et al.*, 1986*a*; Luytjes *et al.*, 1987; Schmidt *et al.*, 1987). Cleavage sites of other viral surface proteins all contain one or two basic residues (Bosch *et al.*, 1981; Dalgarno *et al.*, 1983; Garoff *et al.*, 1980; Paterson *et al.*, 1984; Porter *et al.*, 1979; Rice & Strauss, 1981; Schwartz *et al.*, 1983; Shinnick *et al.*, 1981). C.p.e. can be observed upon infection of HRT-18 cells by BCV but not with HCV-OC43 (data not shown). It is tempting to speculate that the number of basic amino acids at the cleavage site may be involved in an efficient viral infection.

As shown in Fig. 2, the S protein of HCV-OC43 is closely similar to the corresponding protein of the Mebus strain of BCV, with an identity of 91%. The S proteins of both strains of MHV (MHV-A59 and -JHM) show only 62 and 59% identity, respectively, with their HCV-OC43 counterpart (data not shown).

The S protein of HCV-OC43 is composed of 1353 residues, whereas the S protein of BCV contains 1363

```
          10        20        30        40        50        60        70        80        90       100
OC43  MFLILLISLPTAFAVIGDLKCTSDNINDKDTGPPPISTDTVDVTNGLGTYYVLDRVYLNTTLFLNGYYPTSGSTYRNMALKGSVLLSRLWFKPPFLSDFI 100
BCV   .........M..........TVS...V...A.S....I.....................L..................TL..............  100

OC43  NGIFAKVKNTKVIKDRVMYSEFPAITIGSTFVNTSYSVVVQPRTINSTQDGDNKLQGLLEVSVCQYNMCEYPQTICHPNLGNHRKELWHLDTGVVSCLYK 200
BCV   ..............KG.........................H.T.L----.........I.....T.....H.........K.V....W.........  196

OC43  RNFTYDVNADYLYFHFYQEGGTFYAYFTDTGVVTKFLFNVYLGMALSHYYVMPLTCNSKLTLEYWVTPLTSRQYLLAFNQDGIIFNAEDCMSDFMSEIKC 300
BCV   ........................................TV......L....S.AM..........K..........V....V..K.........  296

OC43  KTQSIAPPTGVYELNGYTVQPIADVYRRKPNLPNCNIEAWLNDKSVPSPLNWERKTFSNCNFNMSSLMSFIQADSFTCNNIDAAKIYGMCFSSITIDKFA 400
BCV   ..L....S.....................I....D.....................................................................  396

OC43  IPNGRKVDLQLGNLGYLQSFNYRIDTTATSCQLYYNLPAANVSVSRFNPSTWNKRFGFIEDSVFKPRPAGVLTNHDVVYAQHCFKAPKNFCPCKLNGS-C 499
BCV   ..................................................R....T.QF....Q.V..F.H.............S.......D..L.  496

OC43  VGSGPG-----KNNGIGTCPAGTNYLTCDN------LCTPDPIT--FTGTYKCPQTKSLVGIGEHCSGLAVKSDYCGGNSCTCRPQAFLGWSADSCLQGD 586
BCV   ..N...IDAGY..S.............H.AAQCNC........SKS..P.......Y............I........P...Q........V.......  596

OC43  KCNIFANFILHDVNSGLTCSTDLQKANTDIILGVCVNYDLYGILGQGIFVEVNATYYNSWQNLLYDSNGNLYGFRDYIINRTFMIRSCYSGRVSAAFHAN 686
BCV   R..............T........S.................T.....................................LT.....................  696

OC43  SSEPALLFRNIKCNYVFNNSLTRQLQPINYFDSYLGCVVNAYNSTAISVQTCDLTVGSGYCVDYSKNRRSRGAITTGYRFTNFEPFTVNSVNDSLEPVGG 786
BCV   ..................T.S...................D...SSV.................TK....R.........T.....................  796
                                                                                  ↑
OC43  LYEIQIPSEFTIGNMVEFIQTSSPKVTIDCAAFVCGDYAACKSQLVEYGSFCDNINAILTEVNELLDTTQLQVANSLMNGVTLSTKLKDGVNFNVDDINF 886
BCV   ..............E..............S.....................................................................  896

OC43  SPVLGCLGSECSKASSRSAIEDLLFDKVKLSDVGFVEAYNNCTGGAEIRDLICVQSYKGIKVLPPLLSENQISGYTLAATSASLFPPWTAAAGVPFYLNV 986
BCV   .........D.N.V..........S.............................N..........V.................LS..V.........  996

OC43  QYRINGLGVTMDVLSQNQKLIANAFNNALYAIQEGFDATNSALVKIQAVVNANAEALNNLLQQLSNRFGAISASLQEILSRLDALEAEAQIDRLINGRLT 1086
BCV   ......I.....................D.....................................S.............Q.............  1096

OC43  ALNAYVSQQLSDSTLVKFSAAQAMEKVNECVKSQSSRINFCGNGNHIISLVQNAPYGLYFIHFSYVPTKYVTARVSPGLCIAGDRGIAPKSGYFVNVNNT 1186
BCV   ...V.....................................................................K.....................  1196

OC43  WMYTGSGYYYPEPITENNVVVMSTCAVNYTKAPYVMLNTSIPNLPDFKEELDQWFKNQTSVAPDLSLDYINVTFLDLQVEMNRLQEAIKVLNQSYINLKD 1286
BCV   ..F...........G.............D....I.T...H.............................D.....................  1296

OC43  IGTYEYYVKWPWYVWLLICLAGVAMLVLLFFICCCTGCGTSCFKKCGGCCDDYTGYQELVIKTSHDD                          1353
BCV   .................GF.....................I.........H..........                                  1363
```

Fig. 2. Amino acid sequence comparison of the HCV-OC43 S protein with that of BCV (Abraham *et al.*, 1990), by alignment for maximum identity. Dots indicate identical residues; hyphens represent gaps introduced into the sequence; the arrow indicates the putative proteolytic cleavage site. The analysis was performed with the GeneWorks 2.2.1 program (IntelliGenetics) using default settings.

residues. This difference appears in the N-terminal S1 region. The function of the additional sequence in BCV is not known. Sequence comparison between HCV-OC43 and BCV (Fig. 2) revealed more sequence divergence in S1 than in S2. This observation is consistent with the model which suggests that the S1 subunit forms the bulbous part of the S protein and S2 the stem region (Cavanagh, 1983; de Groot *et al.*, 1987*a*). Antigenic sites involved in virus neutralization have been identified in both S1 and S2 (Daniel *et al.*, 1993; Luytjes *et al.*, 1989, Stühler *et al.*, 1991; Talbot *et al.*, 1988; Takase-Yoden *et al.*, 1990; Vautherot *et al.*, 1992; Cavanagh *et al.*, 1986*b*). The comparison of the amino acid sequences of HCV-OC43 and BCV S proteins indicates that these viruses arose from a common progenitor.

Molecular studies of the HCV-OC43 S protein gene are important in the study of the interaction between the virus, the host cell and the immune system during infection. The study of the remainder of the genome of HCV-OC43 should provide important information on the replication, tropism, pathogenesis and evolution of this important human pathogen. Such studies are in progress.

We thank Francine Lambert for excellent technical assistance. This work was supported by grant MT-9203 from the MRC of Canada to P. J. Talbot, who also gratefully acknowledges salary support in the form of a University Research Scholarship from the National Sciences and Engineering Research Council of Canada. S. Mounir is the recipient of a Research Associate fellowship from the Institut Armand-Frappier.

References

Abraham, S., Kienzle, T. E., Lapps, W. & Brian, D. A. (1990). Deduced sequence of the bovine coronavirus spike protein and identification of the internal proteolytic cleavage site. *Virology* **176**, 296–301.

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.

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.

BOIREAU, P., CRUCIERE, C. & LAPORTE, J. (1990). Nucleotide sequence of the glycoprotein S gene of bovine enteric coronavirus and comparison with the S proteins of two mouse hepatitis virus strains. *Journal of General Virology* **71**, 487–492.

BOSCH, F. X., GARTEN, W., KLENK, H.-D. & ROTT, R. (1981). Proteolytic cleavage of influenza virus hemagglutinins: primary structure of the connecting peptide between HA1 and HA2 determines proteolytic cleavability and pathogenicity of avian influenza viruses. *Virology* **113**, 725–735.

BRITTON, P. (1991). Coronavirus motif. *Nature, London* **353**, 394.

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

CAVANAGH, D., DAVIS, P. J., PAPPIN, D. J. C., BINNS, M. M., BOURSNELL, M. E. G. & BROWN, T. D. K. (1986*a*). 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.

CAVANAGH, D., DAVIS, P. J., DARBYSHIRE, J. H. & PETERS, R. W. (1986*b*). Coronavirus IBV: virus retaining spike glycopolypeptide S2 but not S1 is unable to induce virus-neutralizing or haemagglutination-inhibiting antibody, or induce chicken tracheal protection. *Journal of General Virology* **67**, 1435–1442.

DALGARNO, L., RICE, C. M. & STRAUSS, J. H. (1983). Ross River virus 26S RNA: complete nucleotide sequences and deduced sequence of the encoded structural proteins. *Virology* **129**, 170–187.

DANIEL, C. & TALBOT, P. J. (1990). Protection from lethal coronavirus infection by affinity-purified spike glycoprotein of murine hepatitis virus, strain A59. *Virology* **174**, 87–94.

DANIEL, C., ANDERSON, R., BUCHMEIER, M. J., FLEMING, J. O., SPAAN, W. J. M., WEGE, H. & TALBOT, P. J. (1993). Identification of an immunodominant linear neutralization domain on the S2 portion of the murine coronavirus spike glycoprotein and evidence that it forms part of a complex tridimensional structure. *Journal of Virology* **67**, 1185–1194.

DE GROOT, R. J., LUYTJES, W., HORZINEK, M. C., VAN DER ZEIJST, B. A. M., SPAAN, W. J. M. & LENSTRA, J. A. (1987*a*). Evidence for coiled-coil structure in the spike proteins of coronaviruses. *Journal of Molecular Biology* **196**, 963–966.

DE GROOT, R. J., MADURO, J., LENSTRA, J. A., HORZINEK, M. C., VAN DER ZEIJST, B. A. M. & SPAAN, W. J. M. (1987*b*). cDNA cloning and sequence analysis of the gene encoding the peplomer protein of feline infectious peritonitis virus. *Journal of General Virology* **68**, 2639–2646.

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

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.

GAROFF, H., FRISCHAUF, A. M., SIMONS, K., LEHRACH, H. & DELIUS, H. (1980). Nucleotide sequence of cDNA coding for Semliki Forest virus membrane glycoproteins. *Nature, London* **288**, 236–241.

HUNTER, E., HILL, E., HARDWICK, M., BHOWN, A., SCHWARTZ, D. E. & TIZARD, R. (1983). Complete sequence of the Rous sarcoma virus *env* gene: identification of structural and functional regions of its product. *Journal of Virology* **46**, 920–936.

JOUVENNE, P., MOUNIR, S., STEWART, J. N., RICHARDSON, C. D. & TALBOT, P. J. (1992). Sequence analysis of human coronavirus 229E mRNAs 4 and 5: evidence for polymorphism and homology with myelin basic protein. *Virus Research* **22**, 125–141.

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

KÖRNER, H., SCHLIEPHAKE, A., WINTER, J., ZIMPRICH, F., LASSMANN, H., SEDGWICK, J., SIDDELL, S. & WEGE, H. (1991). Nucleocapsid or spike protein-specific $CD4^+$ T lymphocytes protect against coronavirus-induced encephalomyelitis in the absence of $CD8^+$ T cells. *Journal of Immunology* **147**, 2317–2323.

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

LUYTJES, W., STURMAN, L. S., BREDENBEEK, P. J., CHARITE, J., VAN DER ZEIJST, B. A. M., HORZINEK, M. & SPAAN, W. J. M. (1987). Primary structure of the glycoprotein E2 of coronavirus MHV-A59 and identification of the trypsin cleavage site. *Virology* **161**, 479–487.

LUYTJES, W., GEERTS, D., POSTHUMUS, W., MELOEN, R. & SPAAN, W. (1989). Amino-acid sequence of a conserved neutralizing epitope of murine coronaviruses. *Journal of Virology* **63**, 1408–1412.

MCINTOSH, K. (1974). Coronaviruses: a comparative review. *Current Topics in Microbiology and Immunology* **63**, 85–129.

MACNAUGHTON, M. R. & DAVIES, H. A. (1981). Human enteric coronaviruses. *Archives of Virology* **70**, 301–313.

MACNAUGHTON, M. R., MADGE, M. H. & REED, S. E. (1981). Two antigenic groups of human coronaviruses detected by using enzyme-linked immunosorbent assay. *Infection and Immunity* **33**, 734–737.

MOUNIR, S. & TALBOT, P. J. (1992). Sequence analysis of the membrane protein gene of human coronavirus OC43 and evidence for *O*-glycosylation. *Journal of General Virology* **73**, 2731–2736.

MOUNIR, S. & TALBOT, P. J. (1993). Human coronavirus OC43 RNA 4 lacks two open reading frames located downstream of the S gene of bovine coronavirus. *Virology* **192**, 355–360.

MURRAY, R. S., BROWN, B., BRIAN, D. & CABIRAC, G. F. (1992). Detection of coronavirus RNA and antigen in multiple sclerosis brain. *Annals of Neurology* **31**, 525–533.

NIEMANN, H., BOSCHEK, B., EVANS, D., ROSING, M., TAMURA, T. & KLENK, H.-D. (1982). Posttranslational glycosylation of coronavirus glycoprotein E1: inhibition by monensin. *EMBO Journal* **2**, 1499–1504.

PATERSON, R. G., HARRIS, T. J. R. & LAMB, R. A. (1984). Fusion protein of the paramyxovirus simian virus 5: nucleotide sequence of mRNA predicts a highly hydrophobic glycoprotein. *Proceedings of the National Academy of Sciences, U.S.A.* **81**, 6706–6710.

PORTER, A. G., BARBER, C., CAREY, N. H., HALLEWELL, R. A., THRELFALL, G. & EMTAGE, J. S. (1979). Complete nucleotide sequence of an influenza virus haemagglutinin gene from cloned DNA. *Nature, London* **282**, 471–477.

RASSCHAERT, D. & LAUDE, H. (1987). The predicted primary structure of the peplomer protein E2 of the porcine coronavirus transmissible gastroenteritis virus. *Journal of General Virology* **68**, 1883–1890.

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.

RESTA, S., LUBY, J. P., ROSENFELD, C. R. & SIEGEL, J. D. (1985). Isolation and propagation of a human enteric coronavirus. *Science* **229**, 978–981.

RICE, C. M. & STRAUSS, J. H. (1981). Nucleotide sequence of the 26S mRNA of Sindbis virus and deduced sequence of the encoded virus structural proteins. *Proceedings of the National Academy of Sciences, U.S.A.* **78**, 2062–2066.

SANGER, F. & COULSON, A. R. (1975). A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of Molecular Biology* **94**, 414–416.

SCHMIDT, I., SKINNER, M. & SIDDELL, S. (1987). Nucleotide sequence of the gene encoding the surface projection glycoprotein of coronavirus MHV-JHM. *Journal of General Virology* **68**, 47–56.

SCHWARTZ, D. E., TIZARD, R. & GILBERT, W. (1983). Nucleotide sequence of Rous sarcoma virus. *Cell* **32**, 853–869.

SHINNICK, T. M., LERNER, R. A. & SUTCLIFFE, J. G. (1981). Nucleotide sequence of Moloney murine leukaemia virus. *Nature, London* **293**, 543–548.

SPAAN, W., CAVANAGH, D. & HORZINEK, M. C. (1988). Coronaviruses: structure and genome expression. *Journal of General Virology* **69**, 2939–2952.

STEWART, J. N., MOUNIR, S. & TALBOT, P. J. (1992). Human corona-

virus gene expression in the brains of multiple sclerosis patients. *Virology* **191**, 502–505.

STÜHLER, A., WEGE, H. & SIDDELL, S. G. (1991). Localization of antigenic sites on the surface glycoprotein of mouse hepatitis virus. *Journal of General Virology* **72**, 1655–1658.

STURMAN, L. S., HOLMES, K. V. & BEHNKE, J. (1980). Isolation of coronavirus envelope glycoproteins and interaction with the viral nucleocapsid. *Journal of Virology* **33**, 449–462.

TAKASE-YODEN, S., KIKUCHI, T., SIDDELL, S. G. & TAGUCHI, F. (1990). Localization of major neutralizing epitopes on the S1 polypeptide of the murine coronavirus peplomer glycoprotein. *Virus Research* **18**, 99–107.

TALBOT, P. & JOUVENNE, P. (1992). Neurotropic potential of coronaviruses. *Médecine/Sciences* **8**, 119–125.

TALBOT, P. J., DIONNE, G. & LACROIX, M. (1988). Vaccination against lethal coronavirus-induced encephalitis with a synthetic decapeptide homologous to a domain in the predicted peplomer stalk. *Journal of Virology* **62**, 3032–3036.

TYRRELL, D. A. J. (1986). Common colds. *Intervirology* **25**, 177–189.

VAUTHEROT, J.-F., LAPORTE, J. & BOIREAU, P. (1992). Bovine coronavirus spike glycoprotein: localization of an immunodominant region at the amino-terminal end of S2. *Journal of General Virology* **73**, 3289–3294.

VON HEIJNE, G. (1984). How signal sequences maintain cleavage specificity. *Journal of Molecular Biology* **173**, 243–251.

WEGE, H., SIDDELL, S. & TER MEULEN, V. (1982). The biology and pathogenesis of coronaviruses. *Current Topics in Microbiology and Immunology* **99**, 165–200.

WILLIAMS, R. K., JIANG, G.-S. & HOLMES, K. V. (1991). Receptor for mouse hepatitis virus is a member of the carcinoembryonic antigen family of glycoproteins. *Proceedings of the National Academy of Sciences, U.S.A.* **88**, 5533–5536.

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.

ZHANG, X. M., KOUSOULAS, K. G. & STORZ, J. (1992). The hemagglutinin/esterase gene of human coronavirus strain OC43: phylogenetic relationships to bovine and murine coronaviruses and influenza C virus. *Virology* **186**, 318–323.

(*Received 25 February 1993; Accepted 13 May 1993*)