VIROLOGY 174, 608–612 (1990)

Sequence Analysis of the Membrane Protein Gene of Human Coronavirus 229E

PATRICIA JOUVENNE,* CHRISTOPHER D. RICHARDSON,† STEVEN S. SCHREIBER,‡
MICHAEL M. C. LAI,‡ AND PIERRE J. TALBOT*,[1]

Institut Armand–Frappier, Université du Québec, Virology Research Center, Laval, Québec, Canada H7N 4Z3; †Biotechnology Research Center, National Research Council of Canada, Montréal, Québec, Canada H4P 2R2; and ‡University of Southern California, School of Medicine, Departments of Neurology and Microbiology, Los Angeles, California 90033

Received June 22, 1989; accepted October 18, 1989

Human coronaviruses (HCV) are ubiquitous pathogens which cause respiratory, gastrointestinal, and possibly neurological disorders. To better understand the molecular biology of the prototype HCV-229E strain, the complete nucleotide sequence of the membrane protein (M) gene was determined from cloned cDNA. The open reading frame is preceded by a consensus transcriptional initiation sequence UCUAAACU, identical to the one found upstream of the N gene. The M gene encodes a 225-amino acid polypeptide with a molecular weight (MW) of 25,822, slightly higher than the apparent MW of 19,000–22,000 observed for the unprocessed M protein obtained after *in vitro* translation and immunoprecipitation. The M amino acid sequence presents a significant degree of homology (38%) with its counterpart of transmissible gastroenteritis coronavirus (TGEV). The M protein of HCV-229E is highly hydrophobic and its hydropathicity profile shows a transmembranous region composed of three major hydrophobic domains characteristic of a typical coronavirus M protein. About 10% (20 amino acids) of the HCV-229E M protein constitutes a hydrophilic and probably external portion. One *N*-glycosylation and three potential *O*-glycosylation sites are found in this exposed domain.

Human coronaviruses (HCV) belong to either one of two antigenic groups, represented by the prototype strains 229E and OC43 (*1*). They are responsible for as much as 25% of common colds (*2, 3*) and have been associated with gastrointestinal disorders (*4*). Their possible involvement in neurological diseases was suggested by the observation of coronavirus-like particles in the brain of one multiple sclerosis (MS) patient (*5*), the isolation of coronaviruses from two MS brain tissues passaged in mice (*6*), and the detection of intrathecal antibodies to HCV-OC43 and HCV-229E in MS patients (*7*). However, the association of human coronaviruses with neurological diseases has not yet been confirmed.

HCV-229E possesses a single-stranded, positive-sense RNA genome with a molecular weight of 5.8 $\times 10^6$ and a poly(A) tail of about 70 nucleotides at the 3′ end (*8*). As with other coronaviruses, six subgenomic RNAs are synthesized in infected cells (*9*). These appear to have lower molecular weights than viral RNAs synthesized in cells infected with murine hepatitis virus (MHV). At least four polypeptides have been found in purified HCV-229E virions: 160- to 200-kDa and 88- to 105-kDa glycoproteins which may be analogous to the spike glycoprotein S (previously designated E2) of MHV (*10*); a 47- to 53-kDa polypeptide corresponding to the nucleocapsid protein N and a 17- to 26-kDa M protein (previously designated E1) observed in both glycosylated and nonglycosylated forms (*11–14*). One author also reported glycoproteins of 31 and 65 kDa (*11*).

The nucleotide sequence of the genes encoding the nucleocapsid proteins as well as the mRNA leader sequences of HCV-229E and HCV-OC43 have recently been determined (*15, 16*). As a continuation of these studies, we report the nucleotide sequence of the gene encoding the membrane protein M of HCV-229E. Its predicted amino acid sequence is compared with sequences determined for other coronaviruses.

Clones containing the sequence of the M protein gene were obtained from a cDNA library constructed with mRNA isolated from HCV-229E-infected L132 cells, and identified using a genome-specific probe (*15*). One clone, designated L8, was selected for sequencing since it contained a large 3.6-kb insert overlapping by 1.2 kb the 5′ end of the N protein gene. The remaining 2.4-kb fragment was excised from an internal *Pst*I site of clone L8 and subcloned into the pBluescript II vector (Stratagene). Unidirectional deletions of the 2.4-kb insert were created using exonuclease III, mung bean nuclease, and deoxythionucleotide derivatives (Stratagene). The sequencing of both strands was

[1] To whom requests for reprints should be addressed.

0042-6822/90 $3.00

```
5'- CTACTAGTGTGTATTACAATAATTAAACTAACTAAGCTTTGTTTCACTTGCCATATGTTTTGTACTAGAACAATT      75

TATGGCCCCATTAAAAATGTGTACCACATTTACCAATCATATATGCACATAGACCCTTTCCCTAAACGAGTTATTGATC     154

TCTAAACTAAACGACA ATG TCA AAT GAC AAT TGT ACG GGT GAC ATT GTC ACC CAT TTG AAG AAT     218
                 M   S   N   D   N   C   T   G   D   I   V   T   H   L   K   N       16
                     ★       ●       ★               ★
TGG AAT TTT GGT TGG AAT GTT ATT CTA ACC ATA TTC ATT GTT ATT CTT CAG TTT GGA CAC     278
W   N   F   G   W   N   V   I   L   T   I   F   I   V   I   L   Q   F   G   H       36

TAT AAA TAC TCC AGA TTG CTT TAT GGT TTG AAG ATG CTT GTA CTG TGG CTT CTT TGG CCA     338
Y   K   Y   S   R   L   L   Y   G   L   K   M   L   V   L   W   L   L   W   P       56

CTC GTA CTT GCT TTG TCA ATC TTT GAC ACC TGG GCT AAT TGG GAT TCT AAT TGG GCC TTT     398
L   V   L   A   L   S   I   F   D   T   W   A   N   W   D   S   N   W   A   F       76

GTT GCA TTT AGC CTT CTT ATG GCC GTA TCA ACA CTC GTT ATG TGG GTG ATG TAC TTC GCA     458
V   A   F   S   L   L   M   A   V   S   T   L   V   M   W   V   M   Y   F   A       96

AAC AGT TTC AGA CTT TTC CGA CGT GCT CGA ACT TTT TGG GCA TGG AAT CCT GAG GTC AAT     518
N   S   F   R   L   F   R   R   A   R   T   F   W   A   W   N   P   E   V   N      116

GCA ATC ACT GTC ACA ACC GTG TTG GGA CAG ACA TAC TAT CAA CCC ATT CAA CAA GCT CCA     578
A   I   T   V   T   T   V   L   G   Q   T   Y   Y   Q   P   I   Q   Q   A   P      136

ACA GGC ATT ACT GTG ACC TTG CTG AGC GGC GTG CTT TAC GTT GAC GGA CAT AGA TTG GCT     638
T   G   I   T   V   T   L   L   S   G   V   L   Y   V   D   G   H   R   L   A      156

TCA GGT GTT CAG GTT CAT AAC CTA CCT GAA TAC ATG ACA GTT GCC GTG CCG AGC ACT ACT     698
S   G   V   Q   V   H   N   L   P   E   Y   M   T   V   A   V   P   S   T   T      176

ATA ATT TAT AGT AGA GTC GGA AGG TCC GTA AAT TCA CAA AAT AGC ACA GGC TGG GTT TTC     758
I   I   Y   S   R   V   G   R   S   V   N   S   Q   N   S   T   G   W   V   F      196
                                                    ●
TAC GTA CGA GTA AAA CAC GGT GAT TTT TCT GCA GTG AGC TCT CCC ATG AGC AAC ATG ACA     818
Y   V   R   V   K   H   G   D   F   S   A   V   S   S   P   M   S   N   M   T      216
                                                            ●
GAA AAC GAA AGA TTG CTT CAT TTT TTC TAA ACTGAACGAAAAG ATG -3'                       864
E   N   E   R   L   L   H   F   F                                                   225
```

FIG. 1. Complete nucleotide sequence of the M protein gene of HCV-229E and its predicted amino acid sequence. The leader sequences are underlined and the potential *N*-glycosylation (●) and *O*-glycosylation (★) sites at the putatively external N-terminus of the polypeptide are also indicated. Two other *N*-glycosylation sites are found at the C-terminus of the protein. The nucleotide sequence from position 792 is from Ref. (*15*).

performed by the plasmid sequencing technique (*17*), using T7 DNA polymerase. *In vitro* translation of poly(A)$^+$ mRNAs isolated from HCV-229E-infected L132 cells was carried out in order to determine the molecular mass of the unprocessed viral polypeptides.

The complete nucleotide sequence of the M protein gene of HCV-229E and its predicted amino acid sequence are presented in Fig. 1. The AUG codon is preceded by the consensus intergenic sequence UCUAAACU, which is identical to that upstream of the nucleocapsid protein-coding sequence (*15;* and Fig. 1). This sequence is conserved among coronaviruses of various species and represents the binding site of the leader RNA which mediates a discontinuous transcription of mRNAs (*18*). The longest open reading frame extends from base 171 through base 848 and encodes a 225-amino acid polypeptide with a calculated molecular weight of 25,822. The products of *in vitro* transla-

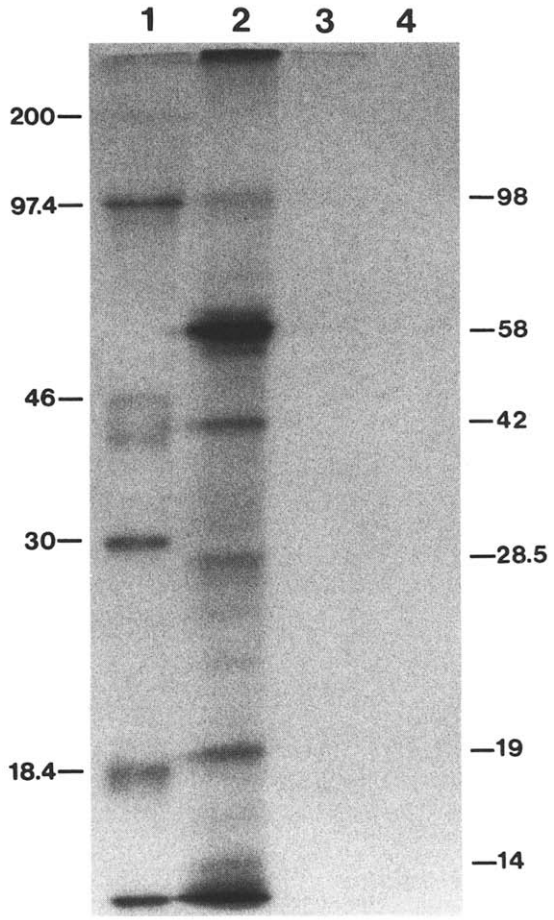


FIG. 2. Immunoprecipitation of *in vitro* translation products from HCV-229E mRNAs. Poly(A)$^+$ mRNAs were translated in the presence of [^{35}S]methionine, using a rabbit reticulocyte lysate (Promega Biotec). The viral polypeptides were immunoprecipitated and separated by SDS–PAGE (13% acrylamide). Lane 1, molecular mass standards; lane 2, mRNAs from HCV-229E-infected cells; lane 3, mRNAs from noninfected cells; lane 4, translation without exogenous mRNA. Molecular mass standards (kDa) are indicated on the left. The calculated molecular masses of relevant viral proteins (kDa) are indicated on the right.

tion of poly(A)$^+$ mRNAs from HCV-229E-infected cells were precipitated with a polyclonal antiserum prepared against purified HCV-229E virions. As shown in Fig. 2, six viral polypeptides were observed, which migrated with apparent molecular masses of 98, 58, 42, 28.5, 19, and 14 kDa, respectively. Although the identity of these proteins has not been firmly established, by comparing with other coronaviruses, p98 probably corresponds to S, p58 to N, and p19 to M. The nature of p42 and p28.5 is not known at this time. Thus, the molecular mass of M predicted from the nucleotide sequence is slightly higher than the molecular mass estimated by SDS–PAGE. Other studies have shown that the mature M protein has a molecular mass of 23- to 26-kDa (*12–14*) and that virions also incorporate a nonglycosylated 20- to 22-kDa precursor of the M protein (*12, 14*). The latter observation is consistent with the identification of *in vitro* translated p19 as M. The lower apparent molecular mass of M estimated by SDS–PAGE is consistent with the unusual electrophoretic behavior of this and other hydrophobic proteins, as was observed for MHV (*19*).

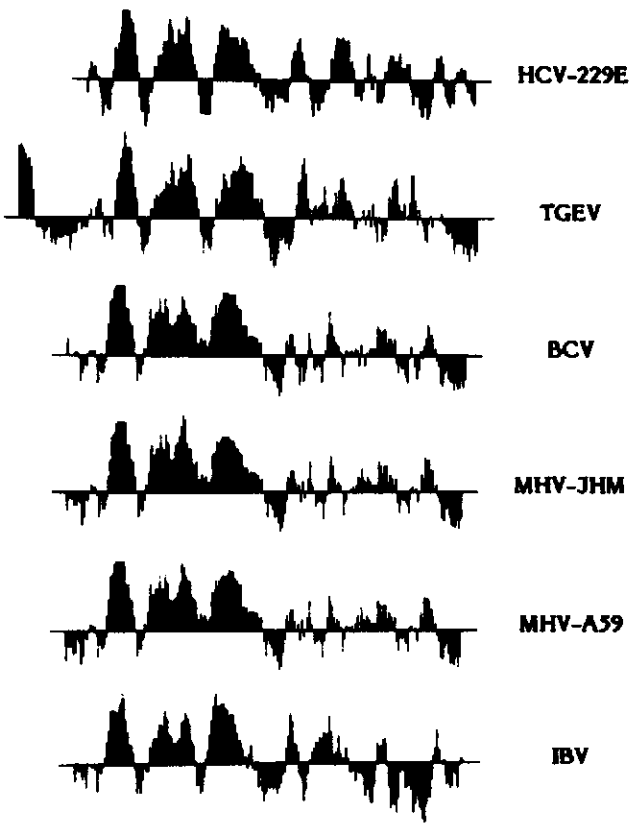
Like TGEV (*20*), there are three amino acid sequences characteristic of *N*-glycosylation sites in the predicted M protein sequence (Asn-5; Asn-190; and Asn-214), although only one (Asn-5) is found near the N-terminus, as compared to two for TGEV. Moreover, three potential *O*-glycosylation sites are located in the putatively external N-terminus of the polypeptide (Ser-2; Thr-7; and Thr-12). In addition, there is only one cysteine residue (Cys-6). Other coronavirus M proteins contain two (bovine coronavirus, BCV; Ref. (*21*)), four (MHV-A59 and JHM; Refs. (*19*) and (*22*), respectively), eight (TGEV; Ref. (*20*)), or nine (infectious bronchitis virus, IBV; Ref. (*23*)) cysteine residues. This cysteine residue is probably important in forming interchain disulfide bridges, since M of HCV-229E has been shown to form oligomers under nonreducing conditions (*14*).

No significant nucleotide sequence homology exists between the M genes of HCV-229E and other coronaviruses. The highest M amino acid homology (38% or 100 of 262 residues) occurs between HCV-229E and TGEV, which was reported to be antigenically related (*24*). Antigenically distinct BCV, MHV, and IBV show amino acid homologies of 32, 30, and 28%, respectively. In contrast, a homology of 87% was found between the M proteins of BCV and MHV-A59 (*21*), which belong to another antigenic subgroup (*24*). On the other hand, a homology of 34% was found between the M protein of TGEV and BCV (*25*), which belong to two different antigenic subgroups. Figure 3 illustrates the M regions common to both HCV-229E and TGEV.

As with other coronaviruses, the M protein of HCV-229E is a highly hydrophobic membrane protein. It contains 51% hydrophobic residues, compared to 45–51% for other coronaviruses (*19–23, 25*). The hydropathicity profiles of M proteins from HCV-229E, TGEV, BCV, MHV-JHM, MHV-A59, and IBV are presented in Fig. 4. The main features characterizing these M proteins include three large hydrophobic domains alternating with short hydrophilic regions. This suggests a selective pressure to maintain the potential transmembranous domains of this coronavirus protein. As with other coronaviruses (*26*), only about 10% (20 amino acids) of the HCV-229E M protein constitutes the hydrophilic putative external domain. On the other hand,

```
HCV-229E  ----------------------------------MSNDNC-T-GDIV
TGEV      MKILLILACVIACACGERYCAMKSDTDLSCRNSTASDC--ESCFNGGDLI      50

          THLKNWNFGWNVILTIFIVILQFGHYKYSRLLYGLKMLVLWLLWPLVLAL
          WHLANWNFSWSIILIVFITVLQYGRPQFSWFVYGIKMLIMWLLWPVVLAL     100

          SIFDTWANWD-SNWAFVAFSLLMAVSTLVMWVMYFANSFRLFRRARTFWA
          TIFNAYSEYQVSRYVMFGFSIAGAIVTFVLWIMYFVRSIQLYRRTNSWWS     150

          WNPEVNAITVTTVLGQTYYQPIQQAPTGITVTLLSGVLYVDGHRLASGVQ
          FNPETKAILCVSALGRSYVLPLEGVPTGVTLTLLSGNLYAEGFKIADGMN     200

          VHNLPEYMTVAVPSTTIIYSRVGRSVNSQNSTGWVFYVRVKHGDFSAVSS
          IDNLPKYVMVALPSRTIVYTLVGKKLKASSATGWAYYVKSKAGDYST-EA     250

          PMSNMTENERLLHFF
          RTDNLSEQEKLLH-MV                                       266
```

FIG. 3. Comparison of the predicted amino acid sequences of M proteins of HCV-229E (top row) and TGEV (bottom row) aligned for maximum homology. Regions common to both proteins are underlined. The analysis was performed on an Apple Macintosh Plus computer with the MacGene Plus program (Applied Genetic Technology Inc., Fairview Park, OH).



the large N-terminal putative signal sequence found only in the M protein of TGEV (*20, 25*) is not observed in HCV-229E, which is similar to the structure reported for BCV, MHV-JHM, MHV-A59, and IBV.

The coronavirus M protein is important for several reasons. This membrane protein is implicated in virus assembly and is believed to integrate the viral proteins prior to budding, most likely because of this protein's affinity for RNA (*26*). Moreover, some monoclonal antibodies against the M protein of MHV-JHM are protective *in vivo* and thus may influence the outcome of disease (*27*). We are currently pursuing the cloning and sequencing of other genes of HCV-229E, with emphasis on the gene coding for the spike protein S, which is potentially important in viral pathogenicity. The availability of molecular probes for human coronavirus genes opens new avenues for the verification of the potential involvement of these viruses in neurological diseases.

FIG. 4. Hydropathicity profiles of M proteins from HCV-229E, TGEV, BCV, MHV-JHM, MHV-A59, and IBV determined according to Kyte and Doolittle (*28*). The analysis was performed with the MacGene Plus program as described in the legend to Fig. 3. Each point is the mean hydropathicity of a span of seven residues. Peaks extending upwards correspond to hydrophobic regions and peaks extending downwards to hydrophilic areas.

ACKNOWLEDGMENTS

We thank François Fossiez for helpful discussions and Lucie Summerside for typing the manuscript. This work was supported by Grant MT-9203 from the Medical Research Council of Canada to P. J. Talbot, and U.S. Public Health Services Research Grant NS18146 to M. M. C. Lai. P. J. Talbot is also the recipient of a University Research Scholarship from the Natural Sciences and Engineering Research Council of Canada. P. Jouvenne acknowledges a studentship support from the Fonds de la recherche en santé du Québec. S. S. Schreiber was supported by a postdoctoral training fellowship from the U.S. National Institutes of Health Grant NS07149.

REFERENCES

1. MACNAUGHTON, M. R., MADGE, M. H., and REED, S. E., *Infect. Immun.* **33,** 734–737 (1981).
2. MCINTOSH, K., CHAO, R. K., KRAUSE, H. E., WASIL, R., MOCEGA, H. E., and MUFSON, M. A., *J. Infect. Dis.* **139,** 502–510 (1974).
3. WEGE, H., SIDDELL, S., and TER MEULEN, V., *Curr. Top. Microbiol. Immunol.* **99,** 165–200 (1982).
4. RESTA, S., LUBY, J. P., ROSENFELD, C. R., and SIEGEL, J. D., *Science* **229,** 978–981 (1985).
5. TANAKA, R., IWASAKI, Y., and KOPROWSKI, H., *J. Neurol. Sci.* **28,** 121–126 (1976).
6. BURKS, J. S., DEVALD, B. L., JANKOVSKY, L. D., and GERDES, J. C., *Science* **209,** 933–934 (1980).
7. SALMI, A., ZIOLA, B., HOVI, T., and REUNANEN, M., *Neurology* **32,** 292–295 (1982).
8. MACNAUGHTON, M. R., and MADGE, M. H., *J. Gen. Virol.* **39,** 497–504 (1978).
9. WEISS, S. R., and LEIBOWITZ, J. L., *In* "Biochemistry and Biology of Coronaviruses" (V. ter Meulen, S. Siddell, and H. Wege, Eds.), pp. 245–260. Plenum, New York, 1981.
10. STURMAN, L. S., RICARD, C. S., and HOLMES, K. V., *J. Virol.* **56,** 904–911 (1985).
11. HIERHOLZER, J. C., *Virology* **75,** 155–165 (1976).
12. MACNAUGHTON, M. R., *Arch. Virol.* **63,** 75–80 (1980).
13. SCHMIDT, O. W., and KENNY, G. E., *Infect. Immun.* **35,** 515–522 (1982).
14. ARPIN, N., and TALBOT, P. J., *Adv. Exp. Biol. Med.,* in press.
15. SCHREIBER, S. S., KAMAHORA, T., and LAI, M. M. C., *Virology* **169,** 142–151 (1989).
16. KAMAHORA, T., SOE, L. H., and LAI, M. M. C., *Virus Res.* **12,** 1–9 (1989).
17. HATTORI, M., and SAKAKI, Y., *Anal. Biochem.* **152,** 232–238 (1986).
18. SHIEH, C.-K., SOE, L. H., MAKINO, S., CHANG, M.-F., STOHLMAN, S. A., and LAI, M. M. C., *Virology* **156,** 321–330 (1987).
19. ARMSTRONG, J., NIEMANN, H., SMEEKENS, S., ROTTIER, P., and WARREN, G., *Nature (London)* **308,** 751–752 (1984).
20. LAUDE, H., RASSCHAERT, D., and HUET, J.-C., *J. Gen. Virol.* **68,** 1687–1693 (1987).
21. LAPPS, W., HOGUE, B. G., and BRIAN, D. A., *Virology* **157,** 47–57 (1987).
22. PFLEIDERER, M., SKINNER, M. A., and SIDDELL, S. G., *Nucleic Acids Res.* **14,** 6338 (1986).
23. BOURSNELL, M. E. G., BROWN, T. D. K., and BINNS, M. M., *Virus Res.* **1,** 303–313 (1984).
24. MACNAUGHTON, M. R., In "Biochemistry and Biology of Coronaviruses" (V. ter Meulen, S. Siddell, and H. Wege, Eds.), pp. 19–29. Plenum Press, New York, 1981.
25. KAPKE, P. A., TUNG, F. Y. T., HOGUE, B. G., BRIAN, D. A., WOODS, R. D., and WESLEY, R., *Virology* **165,** 367–376 (1988).
26. SPAAN, W., CAVANAGH, D., and HORZINEK, M. C., *J. Gen. Virol.* **69,** 2939–2952 (1988).
27. FLEMING, J. O., SHUBIN, R. A., SUSSMAN, M. A., CASTEEL, N., and STOHLMAN, S. A., *Virology* **168,** 162–167 (1989).
28. KYTE, J., and DOOLITTLE, R. F., *J. Mol. Biol.* **157,** 105–132 (1982).