VIROLOGY 157, 47–57 (1987)

Sequence Analysis of the Bovine Coronavirus Nucleocapsid and Matrix Protein Genes

WILLIAM LAPPS, BRENDA G. HOGUE, AND DAVID A. BRIAN[1]

Department of Microbiology, The University of Tennessee, Knoxville, Tennessee 37996-0845

Received July 17, 1986; accepted October 22, 1986

The 3′ end of the 20-kb genome of the Mebus strain of bovine enteric coronavirus (BCV) was copied into cDNA and cloned into the *PstI* site of the pUC9 vector. Four clones from the 3′ end of the genome were sequenced either completely or in part to determine the sequence of the first 2451 bases. Within this sequence were identified, in order, a 3′-noncoding region of 291 bases, the gene for a 448-amino acid nucleocapsid protein (N) having a molecular weight of 49,379, and the gene for a 230-amino acid matrix protein (M) having a molecular weight of 26,376. A third large open reading frame is contained entirely within the N gene sequence but is positioned in a different reading frame; it potentially encodes a polypeptide of 207 amino acids having a molecular weight of 23,057. A higher degree of amino acid sequence homology was found between the M proteins of BCV and MHV (87%) than between the N proteins (70%). For the M proteins of BCV and MHV, notable differences were found at the amino terminus, the most probable site of O-glycosylation, where the sequence is N-Met–Ser–Ser–Val–Thr–Thr for BCV and N-Met–Ser–Ser–Thr–Thr for MHV. BCV apparently uses two of its six potential O-glycosylation sites.

INTRODUCTION

The bovine enteric coronavirus (BCV) is one cause of severe enteritis in calves and may be responsible for as much as one-quarter of all deaths due to this disease (House, 1978). Vaccines produced from cell culture-attenuated strains of virus have failed to be completely protective. Before attempting to develop vaccines by recombinant DNA that may have improved usefulness, it is imperative that the genes and gene products responsible for inducing protective immunity be thoroughly characterized. Toward this end, and for the purpose of determining the function of individual proteins in coronavirus replication, we have begun to clone and sequence the BCV genome.

BCV is known to possess a single-stranded, nonsegmented, polyadenylated RNA genome of approximately 20 kb (Guy and Brian, 1979; Lapps and Brian, 1985). The total number of genes encoded by the genome is not known, but presumably, because of its close antigenic relatedness to the well-characterized mouse hepatitis coronavirus, BCV will be similar to MHV in the number and arrangement of genes on its genome. One striking dissimilarity between BCV and MHV, however, is the possession by BCV of a fourth major structural protein, the 140-kDa hemagglutinin protein that comprises two disulfide-linked subunits of 65 kDa (Hogue *et al.*, 1984; King and Brian, 1982; King *et al.*, 1985). Questions therefore arise concerning not only its origin, function, and role in inducing protective immunity, but also the location of the hemagglutinin gene on the genome and the resulting divergence from the MHV genome structure.

In this paper we describe experiments that begin to examine the BCV genome by cDNA cloning and DNA sequencing. Within the 3′ 2451-base sequence we find a gene map that parallels that for MHV. We report the primary structure for the N and M genes and their deduced amino acid sequences. Structural comparisons with other coronavirus N and M sequences are made and some conserved structural domains are identified.

MATERIALS AND METHODS

Virus and cells

The Mebus strain of bovine coronavirus (BCV) was plaque purified and grown on the human rectal tumor (HRT) cell line as previously described (Hogue *et al.*, 1984; Lapps and Brian, 1985).

Radiolabeling of viral proteins and purification of virus

Confluent monolayers of cells grown in 150-cm^2 flasks were infected with a multiplicity of approximately 1 PFU per cell. After 1.5 hr adsorption at 37°, inoculum was removed and 15 ml of the appropriate medium and radioisotope was added. Viral polypeptides were labeled by adding 400 μCi ^{3}H-labeled essential amino acids (150–200 mCi/mg; ICN) per flask in medium containing 10% normal essential amino acid concentration and 2% fetal calf serum (Sterile Systems, Inc.). Viral glycoproteins were labeled by adding 400 μCi of [^{3}H]glucosamine (5–15 Ci/mmol, ICN) per flask to me-

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

0042-6822/87 $3.00

dium containing 5% fetal calf serum. Virus was harvested and purified by isopycnic sedimentation in continuous sucrose gradients as previously described (Hogue *et al.*, 1984; Lapps and Brian, 1985).

Polyacrylamide gel electrophoresis and immunoblotting

The discontinuous buffer gel system of Laemmli (1970) was used as previously described (Hogue *et al.*, 1984). For examining intracellular proteins, whole cell lysates were prepared by sonication. Cells in 60-mm petri dishes were washed twice with cold phosphate-buffered saline (PBS), scraped into cold PBS, and pelleted by centrifugation at 2000 rpm. The cell pellet was suspended in 100 μl sterile distilled water, sonicated for 10 sec in a bath sonicator, and stored at −80°. For inhibitor studies, tunicamycin (Sigma) was used at a final concentration of 1.2 or 12 μ*M* and monensin (Calbiochem) was used at a final concentration of 1.0 μ*M*. Tunicamycin or monensin was added to cells immediately after virus adsorption and was incubated with the cells for a total of 24 hr, the time of cell lysate preparation. For electrophoresis, equal volumes of cell lysate and double-strength sample treatment buffer were mixed and heated at 100° for 5 min prior to electrophoresis. Unit strength sample treatment buffer is 0.125 *M* Tris–HCl (pH 6.8)–4% sodium dodecyl sulfate–5 *M* urea. For immunoblotting, a modified method of Towbin *et al.* (1979) was used as previously described (Hogue *et al.*, 1984). The preparation of rabbit antiserum against individual BCV proteins was previously described (Hogue *et al.*, 1984). Monoclonal antiserum to human coronavirus OC43 M protein, which also recognizes BCV M protein, was a gift from J. Fleming, University of Southern California.

Purification of genomic RNA

Virus was purified from clarified supernatant fluids as described above. One-tenth of the virus preparation was labeled with [^{3}H]uridine (400 Ci/mmol, ICN), 20 μCi/ml, in order to follow RNA purification. Viral RNA was extracted using the proteinase K–SDS method (Lapps and Brian, 1985) and phenol–chloroform–isoamyl alcohol extraction and was ethanol precipitated after adding sodium acetate. Because subgenomic RNA species are incorporated into BCV virions (Lapps and Brian, 1985), full-length genomic RNA to be used for cDNA cloning and for making probe for colony screening was selected by rate-zonal sedimentation on preformed 5-ml linear gradients of 15 to 30% sucrose (w/w) made up in TNE (10 m*M* Tris–HCl (pH 7.5), 100 m*M* NaCl, 1 m*M* EDTA)–0.1% SDS. RNA was dissolved in water and sedimented for 1.5 hr at 110,000 *g*, 25°. RNA sedimenting faster than mammalian 28 S ribosomal RNA was recovered by ethanol precipitation.

cDNA cloning of the 3′ end of the BCV genome

BCV genomic RNA was cloned using a modified method of Gubler and Hoffman (1983). First-strand synthesis was carried out in a volume of 50 μl containing 50 m*M* Tris–HCl (pH 8.1), 148 m*M* KCl, 8 m*M* $MgCl_2$, 1 m*M* DTT, 2 m*M* each of the four dNTPs, 10 μCi [^{32}P]dCTP (3000 Ci/mmol, ICN), 15 units RNasin (Promega), 50 pmol oligo(dT_{12-18}), 3 μg BCV RNA, 20 U reverse transcriptase (Seikagaku) for 1 hr at 37°, and the reaction was stopped by adding 2.5 μl 0.5 *M* EDTA. BCV RNA was heated to 100° for 5 min and quick cooled to 37° immediately before its addition to the reaction. Reaction products were extracted with phenol–chloroform–isoamyl alcohol and ethanol precipitated after adding ammonium acetate.

Second-strand synthesis was carried out as described by Gubler and Hoffman in 100 μl containing 20 m*M* Tris–HCl (pH 7.5), 5 m*M* $MgCl_2$, 10 m*M* $(NH_4)_2SO_4$, 10 m*M* KCl, 0.15 m*M*β-NAD, 40 μ*M* dNTPs, 8.5 U/ml *Escherichia coli* RNase H, 230 U/ml DNA polymerase I, 10 U/ml DNA ligase, and all of the product from first-strand reaction. Free nucleotides were removed by three cycles of ethanol precipitation of the reaction product and the total quantity of product was estimated from the amount of radiolabeled first strand that remained.

Double-stranded cDNA was homopolymer tailed essentially by the method of Roychoudhury and Wu (1980). The following were added to the dried DNA in order: 20 μl 10× cacodylate–Tris buffer (1.4 *M* K-cacodylate, 0.3 *M* Tris–HCl (pH 7.6)), 4 μl 5 m*M* DTT, 3 μl 10 m*M* dCTP, 162 μl H_2O, 2 μl 100 m*M* $CoCl_2$, 50 μCi[α – ^{32}P]dCTP (>3000 Ci/mol, ICN) in 5 μl, and 16 units of TdT in 2 μl. The reaction was carried out at 37° for 1 min and then stopped by adding 10 μl 0.5 *M* EDTA and 2 μl 10% SDS, and the product was phenol–chloroform–isoamyl alcohol extracted and ethanol precipitated. This reaction assumed an average size of 1 kbp for the ds cDNA and was designed to give an average of 15 dCMP residues per 3′ end of dsDNA.

C-tailed ds cDNA was annealed to G-tailed, *Pst*I-linearized pUC9 vector (PL Biochemicals) and *E. coli* strain JM103 was transformed by the method of Hanahan (1983) using a total concentration of DNA of less than 0.1 μg/ml.

Identification of large inserts containing 3′-specific BCV sequences

Cells containing recombinant plasmids were observed as white colonies on YT agar plates that con-

tained 100 μg ampicillin/ml, 1 m*M* IPTG, 0.004% X-gal and were transferred to nitrocellulose (Millipore, HATF) and probed with random-primed cDNA copied from BCV genomic RNA (Maniatis *et al.*, 1982). ^{32}P-labeled, random-primed cDNA was synthesized as described above for the oligo(dT)-primed reaction except that dNTP concentrations were 2.5 μ*M* each, 0.2 μg RNA was used, and oligo(dT) was replaced by 20 μg fragmented calf thymus DNA. Colonies yielding strong signals were analyzed for plasmid size and inserts of 1.0 to 4.1 kb (the largest) were further analyzed by Southern hybridization with ^{32}P-labeled poly(dT) to detect poly(dA). ^{32}P-labeled poly(dT) probe was prepared as described above for the oligo(dT)-primed reaction except that 50 pmol oligo(dT) • poly(rA) (PL Biochemicals) replaced the RNA. Alkali-treated [^{32}P]poly(dT) probe was incubated for hybridization at 37° for 12 hr, then at 20° for 36 hr, and blots were washed in 2× SSC, 0.1% SDS at 20°.

DNA sequencing and sequence analysis

Plasmids were purified by alkaline lysis and cesium chloride centrifugation as described by Maniatis *et al.* (1982), and restriction endonuclease mapping was done as described by Smith and Birnstiel (1976) using plasmids that were labeled at the *Sal*I site within the multiple cloning linker region. Restriction fragments end-labeled with ^{32}P were isolated and sequenced by the method of Maxam and Gilbert (1980). Many end-labeled fragments of less than 700 bases were first strand-separated before sequencing (James and Bradshaw, 1984). Sequences were analyzed with the aid of the program developed by Queen and Korn (1984) marketed as part of the Beckman Microgenie program, March 1986 version (Beckman Instruments, Inc.).

RESULTS

cDNA cloning and sequencing of four clones from the 3′ end of the genome

Starting material for cDNA cloning was approximately 3 μg of rate-zonally purified genomic RNA obtained from 500 ml of tissue culture medium. An estimated 70 ng of ds cDNA was generated and from this 670 white colonies were obtained. By colony screening, 89 colonies gave a strong signal to [^{32}P]cDNA prepared from genomic RNA, and of these, 9 had inserts ranging from 1.2 to 4.1 kb as determined by agarose gel electrophoresis of linearized plasmids. The 9 clones were further analyzed to determine their restriction enzyme maps and poly(A) content. Only one of the clones, an insert of 1.2 kb identified as clone CB9, reacted strongly under hybridization conditions by Southern blotting to ^{32}P-labeled oligo(dT). Three other clones identified as MN3 (2.1 kb), MA5 (2.8 kb), and MA7 (4.1 kb) were found to contain sequences that overlap with CB9 on the basis of hybridization and restriction endonuclease maps (data not shown).

The orientation of all four clones in reference to the 20-kb virus genome and the restriction enzyme sites used for sequencing are illustrated in Fig. 1. Our orientation presumes polyadenylation at only the 3′ end of the genome and this is based on the documented 3′ polyadenylation site in the avian infectious bronchitis virus and mouse hepatitis virus genomes (Lai *et al.*, 1981; Stern and Kennedy, 1980). The strategy used for sequencing is described in the legend to Fig. 1. Initially, clone MN3 was sequenced completely and was found to contain all of the N and part of the M sequences. Greater than 98% of the sequence containing the complete N gene was determined either by sequencing both strands of clone MN3 or by repeated sequencing of the same strand using different methods of end labeling. Some of the sequences were confirmed from subclones of MN3. To complete the sequencing of the M gene, clone MA5 was sequenced from its second *Dde*I site and parts of MA7 were sequenced as described in Fig. 1. The total sequence of the M gene was determined by sequencing both strands of DNA and by repeated sequencing of some fragments using different methods of end labeling.

The total nucleotide sequence of 2451 bases from the 3′ end of the genome and the deduced amino acid sequences for the three largest open reading frames contained in this sequence are illustrated in Fig. 2.

All possible translation products were deduced for both virus-sense RNA and virus complementary-sense RNA (Fig. 3) because of the precedent that some single-strand RNA virus genomes are of ambisense polarity (Auperin *et al.*, 1984). RNA complementary to coronavirion-sense RNA could therefore theoretically function as mRNA. Figure 3 illustrates the three largest open reading frames found in the virus-sense sequence, each having the proper initiation codon and each being preceded by a termination codon. These are labeled N (for nucleocapsid protein), M (for matrix protein), and IORF (for "internal" open reading frame). Within the virus complementary-sense RNA, two open reading frames of greater than 250 bases exist, beginning at approximate nucleotide positions 1220 and 1880 in the first reading frame. The significance of these is unknown.

The largest open reading frame predicts a protein having properties of the nucleocapsid protein

The largest open reading frame extends from base 817 through base 2160 and predicts a 448-amino acid protein of 49,379 mol wt. We conclude this to be the coding sequence for the nucleocapsid protein (N) for the following reasons: (i) The only BCV protein de-

scribed to date that approaches this size is the 52-kDa phosphorylated nucleocapsid protein (King and Brian, 1982). (ii) The predicted protein is basic, a property expected of nucleic acid-binding proteins. Fifty-nine (13%) of the amino acids are basic whereas 43 (10%) are acidic, giving the protein a net charge of +16 at neutral pH. (iii) The amino acids encoded by this sequence share extensive (70%) sequence homology with the N protein of the closely related mouse hepatitis virus strains A59 and JHM (Armstrong *et al.*, 1983; Skinner and Siddell, 1983).

The N gene for BCV shares other properties with the N gene of MHV. (i) It is rich in serine. Forty-two residues of serine make it the most abundant amino acid. (ii) It is flanked on its 5′ side by the gene for the M protein, and it is flanked on its 3′ side by a noncoding region of 291 bases, only 3 bases fewer than that for MHV JHM. (iii) The intergenic sequence between the M and N

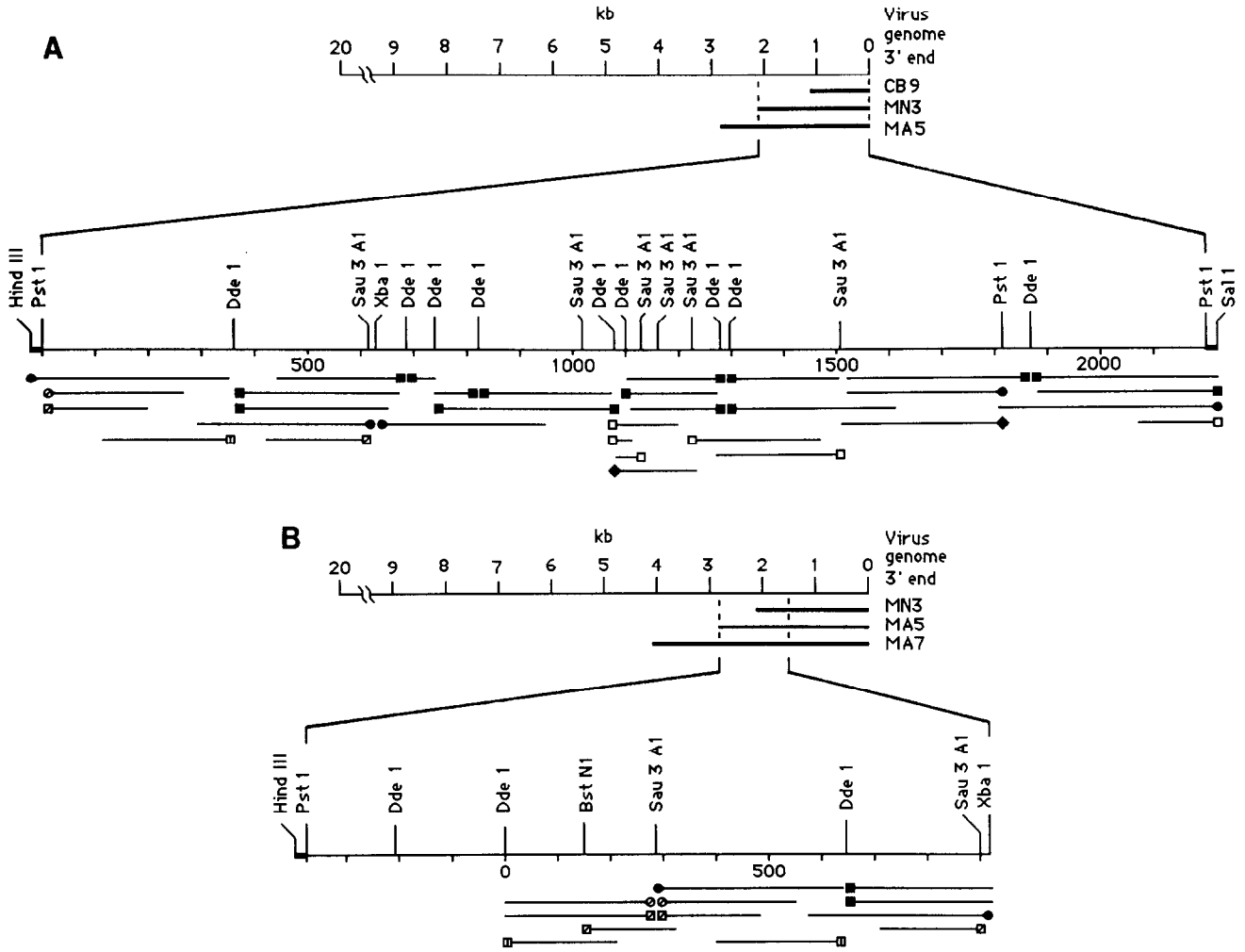


FIG. 1. Sequencing strategy used to obtain BCV genomic sequences containing the N and M genes. (A) Strategy for obtaining the N gene sequence. Clone MN3 was sequenced completely and clones CB9 and MA5 were sequenced in part. The internal *Dde*I, *Pst*I, *Sau*3A I, and *Xba*I sites derived by restriction endonuclease mapping and the *Hind*III and *Sal*I sites in the multiple cloning region of the pUC9 vector were the sites used for DNA sequencing. ■, □ and ⊠ indicate sites labeled at the 5′ end using polynucleotide kinase for clones MN3, CB9, and MA5, respectively. ●, O, and ∅ indicate sites labeled at the 3′ end using reverse transcriptase, and the appropriate labeled deoxynucleotide triphosphate, for clones MN3, CB9, and MA5, respectively. ♦ indicates sites labeled at the 3′ end using radiolabeled cordycepin and terminal transferase for a subclone of MN3. (B) Strategy for obtaining the M gene sequence. Parts of clones MN3, MA5, and MA7 were sequenced beginning with the second *Dde*I site from the 5′ end of clone MA5. ■, ⊠, and ⊡ indicate restriction sites that are 5′ end labeled with polynucleotide kinase for clones MN3, MA5, and MA7, respectively. ● and ∅ indicate sites that were 3′ end labeled using reverse transcriptase and the appropriate labeled deoxynucleotide triphosphate for clones MN3 and MA5, respectively. Uniquely labeled molecules for sequencing were obtained from gels after electrophoretic separation of restriction endonuclease-treated, end-labeled fragments or after strand separation. The orientation of the clones in the pUC9 vector are as shown except for CB9 which is inverted; i.e., the poly(A) end is next to the *Hind*III site of the vector. pUC9 sequences are indicated as a bold line at the end of the restriction endonuclease maps.

genes is very similar. Beginning with the first base following the M gene termination codon, the sequence is UAUCUAAACUUUAAGG for BCV, and UCUAAACUUUAAGG for MHV. (iv) The consensus sequence surrounding the initiation codon for the N gene, AGGAUGU, is the same, and is a preferred sequence for translation initiation among eukaryotic messenger RNAs (Kozak, 1983).

The second largest open reading frame predicts a protein having properties of the matrix protein and also identifies potential O-glycosylation sites

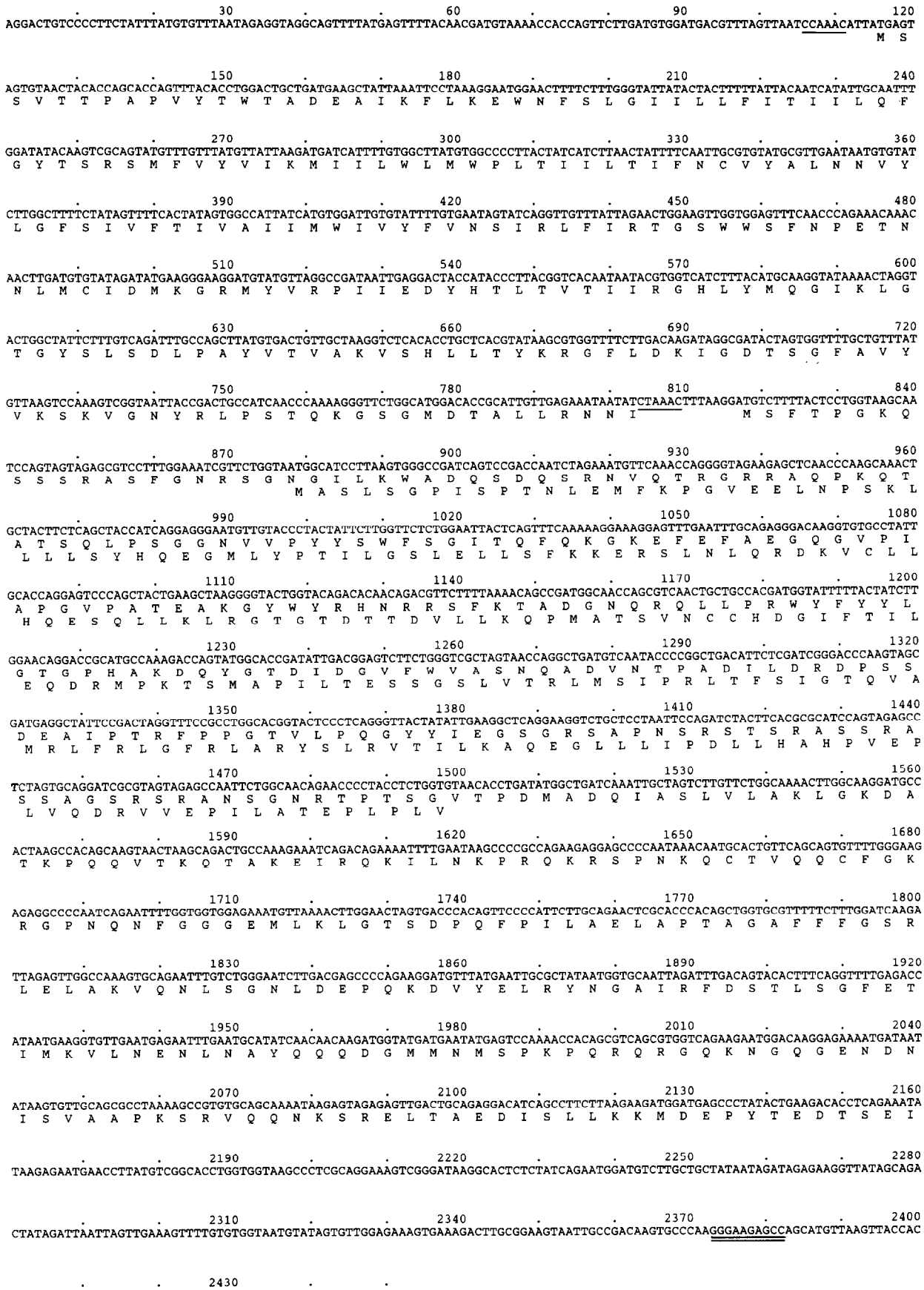
The second largest open reading frame extends from base 112 through base 804 and predicts a 230-amino acid protein having a molecular weight of 26,376 (Fig. 2). This protein has extensive amino acid homology with the M protein of MHV A59 strain (Armstrong *et al.*, 1984), as expected from its close antigenic relatedness (Hogue *et al.*, 1984), and is therefore the apparent BCV counterpart (Fig. 4). By maximum alignment of the proteins, 200 of the amino acids (>86%) are the same as 200 of the 228 in the MHV sequence, and another 16 (7%) represent conservative changes. Because of the strong similarity in structure between the BCV and MHV M proteins, the BCV M protein can be expected to have a similar topology with respect to the virion envelope (Rottier *et al.*, 1986). Namely, the central portion of the molecule would be expected to span the membrane three times, with approximately 28 amino acids (26 for MHV) at the amino terminus being external to the virion and approximately 100 amino acids at the carboxy terminus being internal to the virion. The protein is slightly basic, having a net charge of +9 at neutral pH. The basic amino acids are clustered in the carboxy terminal 40% of the protein. Within the carboxy terminal 100 amino acids are 14 of the 20 basic amino acids and 6 of the 14 acidic amino acids, giving this region of the molecule a net charge of +8. It is therefore reasonable to expect that this part of the molecule might be interacting with the negatively charged RNA as suggested (Sturman *et al.*, 1980) or possibly with an acidic portion of the N protein to contribute to a direct interaction between the M and N molecules. We predict the latter occurs on the basis of a 1:1 molar ratio between the M and N proteins in BCV (King and Brian, 1982). One hundred thirteen (49%) of the amino acids are hydrophobic and the distribution of hydrophobic amino acids is nearly identical to that for the MHV M protein.

Evidence for two O-linked oligosaccharides per M molecule

The M proteins of BCV and MHV were together the first viral glycoproteins shown to possess O-linked oligosaccharides (Holmes *et al.*, 1981; Niemann and Klenk, 1981). The character of the oligosaccharides, however, has been described only for MHV A59 (Niemann *et al.*, 1984). Our data suggest there may be up to two O-linked oligosaccharides per BCV M molecule. First, three separate species of M (gp26) molecules were identified from purified BCV when radiolabeled proteins were resolved by electrophoresis (Fig. 5). These were also observed, but less clearly resolved, when identified by immunoblotting with M-specific polyvalent antiserum (Fig. 5) or with M-specific monoclonal antiserum (data not shown). They have apparent molecular weights of 22K, 24K, and 26K and their appearance is consistent with the notion that the 22K species is the unglycosylated precursor and 1 or 2 oligosaccharide chains, each contributing approximately 2 kDa toward the molecular weight (Klenk and Rott, 1981), are added to assemble a 24- and 26-kDa species, respectively. Second, only three species of M protein were resolved in lysates of infected cells by immunoblotting and neither their sizes nor relative amounts were altered by tunicamycin, an inhibitor of N-glycosylation but not O-glycosylation. Tunicamycin does, however, inhibit the glycosylation of gp190 (the peplomeric protein for which the virion-associated subunits are gp120 and gp100) and gp140 (the hemagglutinn) (Hogue and Brian, in preparation). In the presence of radiolabeled glucosamine only the 24- and 26-kDa species were labeled (Fig. 5). The fact that monensin, an inhibitor of Golgi function and hence O-glycosylation, diminishes the amount of the 24- and 26-kDa species and enhances the relative abundance of the 22-kDa species strengthens the notion the M glycosylation is O-linked (Niemann *et al.*, 1982).

Assuming the BCV M protein is glycosylated in the region external to the virion envelope, i.e., within the first 28 amino acids of the N terminus, then the serine residues at positions 2 and 3 or the threonine residues at positions 5, 6, 12, and 14 are potential sites for O-glycosylation (Fig. 4). If, as presumed for MHV, the glycosylation sites are primarily within the N-terminal N-Met–Ser–Ser–Thr–Thr sequence, a region identical to the glycosylated region of glycophorin A, then the sequence per se may not be an absolute requirement for glycosylation since the N terminal sequence for BCV is N-Met–Ser–Ser–Val–Thr–Thr.

The discrepancy between the observed molecular weight of 22 kDa for the unglycosylated polypeptide and the molecular weight of 26,376 deduced from sequence data could be explained by a strong tendency of the hydrophobic regions of the M protein to remain in close proximity, even in the presence of SDS, giving rise to more rapidly migrating globular molecules. Certainly such behavior would explain the self-aggregation



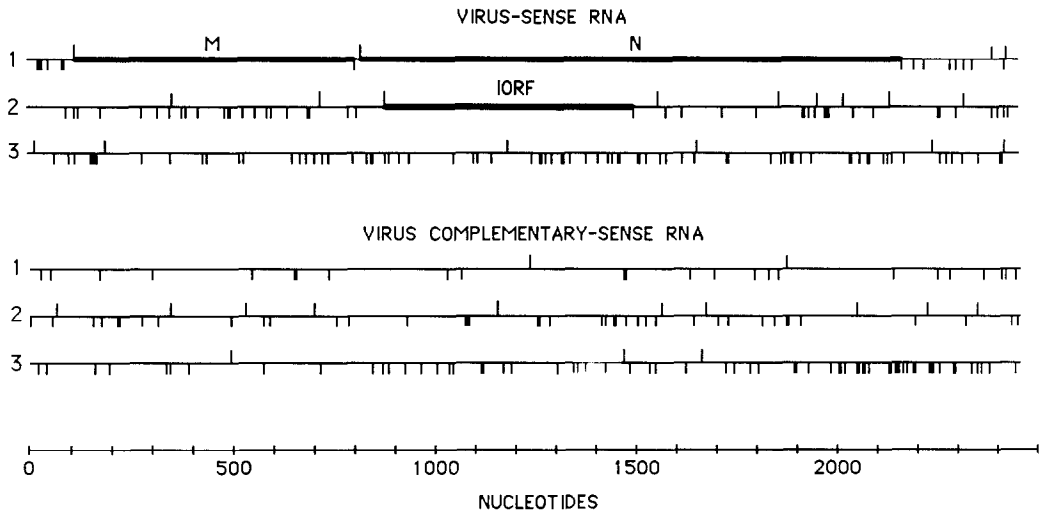


FIG. 3. Schematic diagram of open reading frames obtained when translating the 3′ 2451-base sequence of the BCV genome as either virus-sense RNA or virus complementary-sense RNA. Vertical bars above the line represent the first methionine codon that could serve as the initiation site for translation. Vertical bars below the line represent termination codons. M: sequence of the matrix protein gene. N: sequence of the nucleocapsid protein gene. IORF: sequence of an internal open reading frame within the N gene. 5′ to 3′ orientation is always left to right.

of the M protein observed so frequently (Fig. 5; Hogue *et al.*, 1984).

A large open reading frame found internal to the N gene, but in a different reading frame, predicts a 207-amino acid protein having a molecular weight of 23,057

The nucleotide sequence from base 878 through base 1498 in the second reading frame encodes a 207-amino acid protein of 23,057 (Fig. 2). This protein is hypothetical since we have no proof yet of its existence. The protein has a net charge of +1 at neutral pH and is moderately hydrophobic since 79 (38%) of its amino acids are hydrophobic. The hydrophobic amino acids are spread somewhat evenly throughout the protein except at the carboxyl terminus, where there are enough to make this part of the protein a potential membrane anchor region. Twenty-seven of the terminal 41 amino acids (66%) are hydrophobic. The existence of the protein cannot be ruled out on the basis of the consensus sequence (GUAAUGG) surrounding its initiation codon since it is one commonly used, being found at the initiation site of 18% of all eukaryotic mRNAs catalogued (Kozak, 1983), nor can it be ruled out on the basis of codon usage since it is similar to that used for the N and M proteins.

DISCUSSION

We present the first nucleotide sequence data available for BCV or for any member of the hemagglutinating mammalian coronavirus subgroup which includes the human respiratory coronavirus OC43 and the porcine hemagglutinating encephalitis virus. Despite the fact that BCV has the hemagglutinin structural protein that is missing on MHV A59 (Hogue *et al.*, 1984; King *et al.*, 1985), it shares membership with MHV in one of the four major antigenic subgroups of coronaviruses (Pedersen *et al.*, 1978). Both the gene map and the primary sequence for that part of the BCV genome described in this paper reflect a close relatedness to MHV, consistent with patterns of shared antigenicity between the two viruses (Hogue *et al.*, 1984). Genome sequence divergence with regard to the hemagglutinin gene must therefore lie 5′-ward of this sequence. Both gene arrangement and primary sequence at the 3′ end of the genome, however, suggest a greater degree of evolutionary divergence from both the porcine transmissible gastroenteritis virus (TGEV) and avian infectious

FIG. 2. The primary nucleotide sequence of the 3′ 2451 bases of the BCV genome and the deduced amino acid sequences for the three largest open reading frames, M, N, and IORF. The M open reading frame extends from base positions 115 through 804, the N open reading frame from 817 through 2160, and the "internal" open reading frame from 878 through 1498. The conserved intergenic sequences, C$\overset{C}{T}$AAAC, are underlined with a solid line. A 10-base sequence highly conserved within the 3′-noncoding region among coronaviruses (Kapke and Brian, 1986) is identified with a double underline.

```
M S S V T T     P A P V Y T W T A D E A I K F L K E W N F S L G I I L L F I T
- - -   - - Q A - E - - - Q - - - - - - V Q - - - - - - - - - - - - - - - - -

I I L Q F G Y T S R S M F V Y V I K M I I L W L M W P L T I I L T I F N C V Y
- - - - - - - - - - - - - I - - V - - - - - - - - - - - - - V - C - - - - - -

A L N N V Y L G F S I V F T I V A I I M W I V Y F V N S I R L F I R T G S W W
- - - - - - - - - - - - - - - - S - V I - - M - - - - - - - - - - - - - - - -

S F N P E T N N L M C I D M K G R M Y V R P I I E D Y H T L T V T I I R G H L
- - - - - - - - - - - - - - - - T V - - - - - - - - - - - - - A - - - - - - -

Y M Q G I K L G T G Y S L S D L P A Y V T V A K V S H L L T Y K R G F L D K I
- - - - V - - - - - F - - - - - - - - - - - - - - - - - - - C - - - - A - - - - V

G D T S G F A V Y Y V K S K V G N Y R L P S T Q K G S G M D T A L L R N N I
D G V - - - - - - - - - - - - - - - - - N   - P - - A - - - - - -     -
```

FIG. 4. Comparison of the deduced amino acid sequences for the M proteins of the bovine coronavirus (top) and the mouse hepatitis virus A59 strain (bottom). In the MHV sequence, amino acids matching with those of the BCV sequence are indicated by a hyphen; unmatched amino acids are named. Gaps were introduced to maximize alignment. Potential O-glycosylation sites on the amino terminus of the BCV M proteins are indicated by solid circles.

bronchitis virus (IBV) than from MHV. TGEV has an open reading frame for a potential 9.1K protein positioned between the 3′-noncoding sequence and the N gene (Kapke and Brian, 1986); IBV has two open reading frames for proteins of 7.5K and 9.5K positioned between the N and M genes (Boursnell and Brown, 1984).

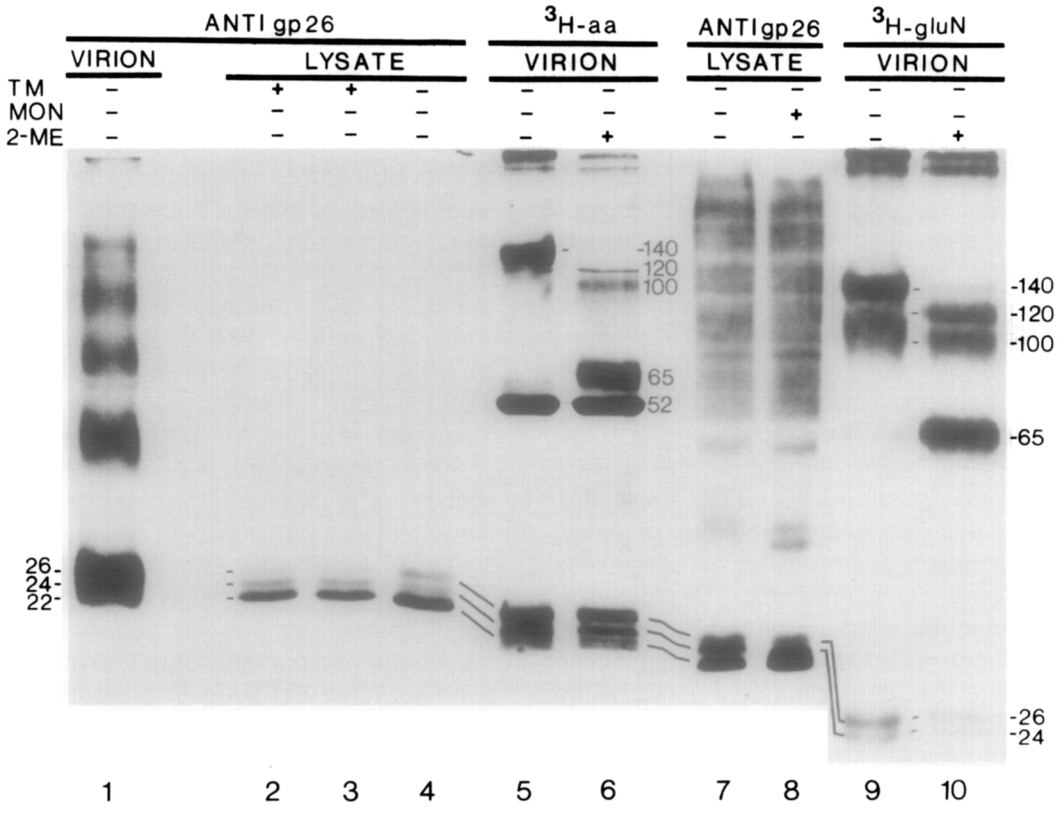


FIG. 5. Identification of different glycosylation states of the M protein. Lane 1: Purified virus immunoblotted with anti-gp26 serum. Lanes 2, 3, and 4: Lysates of infected cells grown respectively in 12, 1.2, or 0 μM tunicamycin (TM), electrophoresed on the same gel as lane 1, and immunoblotted with anti-gp26 serum. Lanes 5 and 6: Purified virus radiolabeled with ^{3}H-labeled amino acids. Lanes 7 and 8: Lysates of infected cells grown in 0 or 1 μM monensin (MON) and immunoblotted with anti-gp26 serum. Lanes 9 and 10: Purified virus radiolabeled with [^{3}H]glucosamine. Samples treated with 2-mercaptoethanol are indicated (2-ME). The polyacrylamide concentrations are 9% for lanes 1 through 4, and 8% for lanes 5 through 10.

The N protein of BCV shows an overall amino acid sequence homology of 70% with both MHV A59 and MHV JHM (72% at the nucleotide level) (Skinner and Siddell, 1983; Armstrong *et al.*, 1983) but only 29% (37% at the nucleotide level) with the N protein of TGEV (Kapke and Brian, 1986), and 29% (43% at the nucleotide level) with the N protein of IBV (Boursnell *et al.*, 1985). The degree of homology between the N amino acid sequences of BCV and MHV is not evenly distributed throughout the gene, however. There are regions of up to 16 amino acid stretches, for example, that show less than 30% homology. Conversely there are regions of up to 69-amino acid stretches showing greater than 90% homology. A region of high homology among MHV (beginning at amino acid 86), IBV (beginning at amino acid 53), and TGEV (beginning at amino acid 53), and extending for 68 positions (Kapke and Brian, 1986), is also found in BCV (beginning at amino acid 83). Within this region there is 79% perfect homology between BCV and MHV. Such regions of conservation suggest that there are evolutionary pressures for retention of a specific function associated with this sequence. Other regions having similar chemical properties but little primary sequence homology also suggest conserved functional domains. These include clusters of serine residues and clusters of basic and acidic amino acids. Assuming all coronavirus N proteins are phosphorylated at only serine residues, as in the N protein of MHV (Stohlman and Lai, 1979), then "hot spots" for potential phosphorylation become apparent when the N protein sequences are compared (Fig. 6). By aligning the N proteins of MHV, BCV, IBV, and TGEV with the first amino acid of the conserved 68 amino acid region, three clusters of 3–12 serine residues in common among all viruses become apparent at BCV amino acid positions 40–70, 180–225, and 300–350. The major serine cluster region is at amino acid positions 180–225. Cluster groups of 5 to 26 basic amino acids can be seen within 50 residues of the amino terminus, within the 68-amino-acid conserved region, between amino acid positions 200 and 300, and in a region extending between amino acids 50 and 25 from the carboxy terminus, but not within six positions of the carboxy terminus (Fig. 6). Clustering of acidic amino acids is less striking but clusters of 10 to 12 are observed within the last 100 bases of the carboxy terminus (Fig. 6). Such regions may indicate sites for protein–nucleic acid or protein–protein interactions.

The high degree of amino acid sequence homology between the M proteins of BCV and MHV (86%) contrasts with the lower degree (70%) between the N proteins. The contrast becomes even more striking when amino acids of conserved nature are included, making the homology 93 and 79%, respectively, for the M and

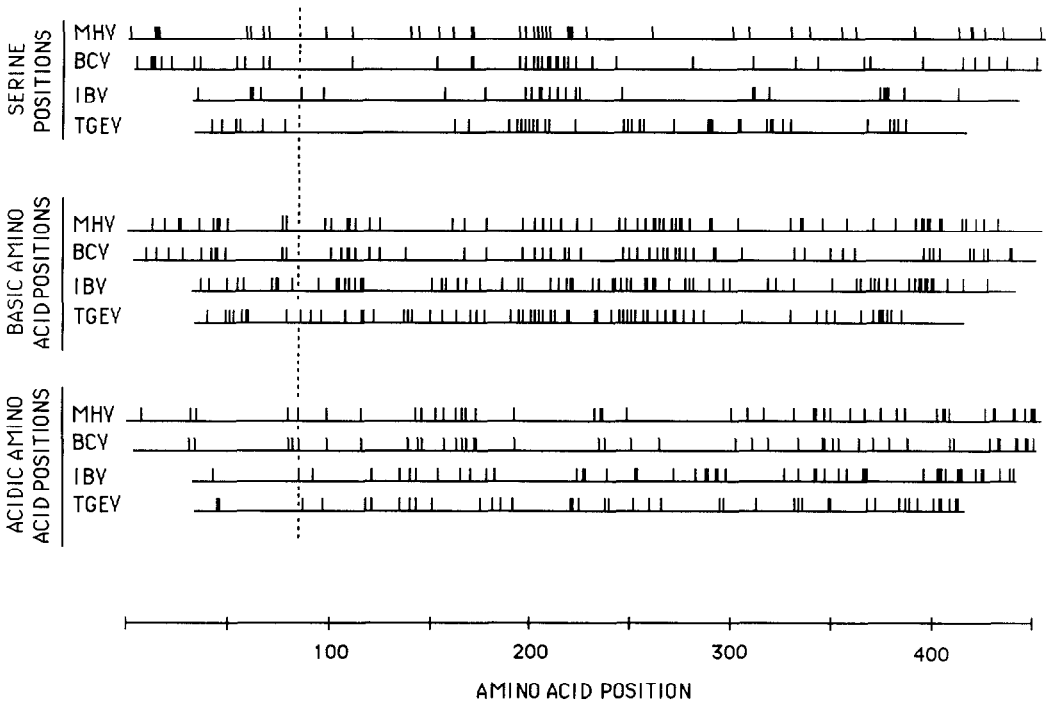


FIG. 6. Positions of serine residues, basic amino acids, and acidic amino acids on the N protein of MHV (A59), BCV (Mebus), IBV (Beaudette), and TGEV (Purdue), depicted in a way to show clustering. The amino terminus is positioned on the left and the proteins (depicted with no gaps in their sequence) are aligned by the first amino acid in the 68-amino-acid region of high homology which is indicated by the stippled line. Amino acid positions are indicated by vertical bars. Proteins are arranged in decreasing order of length.

N proteins. This contrast indicates either that structural constraints on the M protein are more rigid, resulting in a more limited evolution of this protein, or that there is a form of genetic exchange that has taken place between the two viruses. The notion that the M protein may be structurally constrained as a result of functional requirements is suggested by the conserved chemical features between the MHV and IBV M proteins in the absence of conserved primary structure. IBV is antigenically unrelated to MHV, and the IBV M protein shares an amino acid sequence homology of only 35% (perfect match only using the same method of alignment employed above) with that of MHV. Yet it shows an extremely similar hydrophobicity profile and thus an apparently similar membrane topology (Boursnell *et al.,* 1984). That is, amino acid changes were conservative. The notion of genetic exchanges, similar to those observed for RNA viruses with segmented genomes, must be seriously considered in light of recent evidence that coronaviruses undergo high-frequency recombination (Makino *et al.,* 1986). The mechanism giving rise to coronavirus recombinants is unknown but may involve displacement of nascent RNA polymerase complexes from the negative-strand template of one parent with subsequent attachment to the negative strand of a second parent (Makino *et al.,* 1986). Recombination might therefore be expected between the closely related BCV and MHV viruses if, by chance, they should replicate simultaneously in the same host. This most certainly would be expected if polymerase binding during the recombinational event involves the conserved intergenic sequences used to identify initiation sites for transcription (Baric *et al.,* 1985; Budzilowicz *et al.,* 1985; Makino *et al.,* 1986).

ACKNOWLEDGMENTS

We thank Paul Kapke for many helpful discussions. This work was supported by Public Health Service 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. W.L. was a predoctoral trainee on Grant T32-AI-07123 from the National Institutes of Health. B.H. was a predoctoral fellow supported by the Tennessee Center of Excellence Program for Livestock Diseases and Human Health.

REFERENCES

ARMSTRONG, J., NIEMANN, H., SMEEKENS, S., ROTTIER, P., and WARREN, G. (1984). Sequence and topology of a model intracellular membrane protein, E1 glycoprotein, from a coronavirus. *Nature (London)* **308,** 751–752.

ARMSTRONG, J., SMEEKENS, S., and ROTTIER, P. (1983). Sequence of the nucleocapsid gene from murine coronavirus MHV-A59. *Nucleic Acids Res.* **11,** 833–891.

AUPERIN, D. D., ROMANOWSKI, V., GALINSKI, M., and BISHOP, D. H. L. (1984). Sequence studies of Pichinde arenavirus S RNA indicate a novel coding strategy, an ambisense viral S RNA. *J. Virol.* **52,** 897–904.

BARIC, R. S., STOHLMAN, S. A., RAZAVI, M. K., and LAI, M. M. C. (1985). Characterization of leader-related small RNAs in coronavirus-infected cells: Further evidence for leader-primed mechanism of transcription. *Virus Res.* **3,** 19–33.

BOURSNELL, M. E. G., BINNS, M. M., FOULDS, I. J., and BROWN, T. D. K. (1985). Sequences of the nucleocapsid genes from two strains of avian infectious bronchitis virus. *J. Gen. Virol.* **66,** 573–580.

BOURSNELL, M. E. G., and BROWN, T. D. K. (1984). Sequencing of coronavirus IBV genomic RNA: A 195-base open reading frame encoded by mRNA B. *Gene* **29,** 87–92.

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,** 303–313.

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,** 834–840.

GUBLER, U., and HOFFMAN, B. J. (1983). A simple and very efficient method for generating cDNA libraries. *Gene* **25,** 263–269.

GUY, J. S., and BRIAN, D. A. (1979). Bovine coronavirus genome. *J. Virol.* **29,** 293–300.

HANAHAN, D. (1983). Studies on transformation of *Escherichia coli* with plasmids. *J. Mol. Biol.* **166,** 557–580.

HOGUE, B. G., KING, B., and BRIAN, D. A. (1984). Antigenic relationships among proteins of bovine coronavirus, human regulatory coronavirus OC43, and mouse hepatitis coronavirus A59. *J. Virol.* **51,** 384–388.

HOLMES, K. V., DOLLER, E. W., and STURMAN, L. S. (1981). Tunicamycin resistant glycosylation of a coronavirus glycoprotein: Demonstration of a novel type of viral glycoprotein. *Virology* **115,** 334–344.

HOUSE, J. A. (1978). Economic impact of rotavirus and other neonatal disease agents of animals. *J. Amer. Vet. Med. Assoc.* **173,** 573–576.

JAMES, R., and BRADSHAW, R. A. (1984). Strand separation of DNA fragments and their isolation from nondenaturing polyacrylamide gels. *Anal. Biochem.* **140,** 456–458.

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

KING, B., and BRIAN, D. A. (1982). Bovine coronavirus structural proteins. *J. Virol.* **42,** 700–707.

KING, B., POTTS, B. J., and BRIAN, D. A. (1985). Bovine coronavirus hemagglutinin protein. *Virus Res.* **2,** 53–59.

KLENK, H. D., and ROTT, R. (1981). Cotranslational and posttranslational processing of viral glycoproteins. *Curr. Top. Microbiol. Immunol.* **90,** 19–48.

KOZAK, M. (1983). Comparison of initiation of protein synthesis in procaryotes, eucaryotes, and organelles. *Microbiol. Rev.* **47,** 1–45.

LAEMMLI, U. K. (1970). Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *Nature (London)* **227,** 680–685.

LAI, M. M. C., BRAYTON, P. R., ARMEN, R. C., PATTON, C. D., PUGH, C., and STOHLMAN, S. A. (1981). Mouse hepatitis virus A59: mRNA structure and genetic localization in the sequence divergence from hepatotropic strain MHV-3. *J. Virol.* **39,** 823–834.

LAPPS, W., and BRIAN, D. A. (1985). Oligonucleotide fingerprints of antigenically related bovine coronavirus and human coronavirus OC43. *Arch. Virol.* **86,** 101–108.

MAKINO, S., KECK, J. G., STOHLMAN, S. A., and LAI, M. M. C. (1986). High-frequency RNA recombination of murine coronaviruses. *J. Virol.* **57,** 729–737.

MANIATIS, T., FRITSCH, E. F., and SAMBROOK, J. (1982). "Molecular Cloning: A Laboratory Manual." Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.

MAXAM, A. M., and GILBERT, W. (1980). Sequencing end-labeled DNA with base-specific chemical cleavages. *In* "Methods in Enzymology" (L. Grossman and K. Moldave, Eds.), Vol. 65, pp. 499–560. Academic Press, New York.

NIEMANN, H., BOSCHEK, B., EVANS, D., ROSING, M., TAMURA, T., and KLENK, H.-D. (1982). Post-translational glycosylation of coronavirus glycoprotein E1: Inhibition by monensin. *EMBO J.* **1,** 1499–1504.

NIEMANN, H., HEISTERBERG-MOUTSIS, G., GEYER, R., KLENK, H.-D., and WIRTH, M. (1984). Glycoprotein E1 of MHV-A59: Structure of the O-linked carbohydrates and construction of full length recombinant cDNA clones. *Adv. Exp. Med. Biol.* **173,** 201–213.

NIEMANN, H., and KLENK, H.-D. (1981). Coronavirus glycoprotein E1, a new type of viral glycoprotein. *J. Mol. Biol.* **153,** 993–1010.

PEDERSEN, N. C., WARD, J., and MENGELING, W. L. (1978). Antigenic relationship of the feline infections peritonitis virus to coronaviruses of other species. *Arch. Virol.* **58,** 45–53.

QUEEN, C., and KORN, L. J. (1984). A comprehensive sequence analysis program for the IBM personal computer. *Nucleic Acids Res.* **12,** 581–599.

ROTTIER, P. J., WELLING, G. W., WELLING-WEBSTER, S., NIESTERS, H. G. M., LENSTRA, J. A., and VAN DER ZEIJST, B. A. M. (1986). Predicted membrane topology of the coronavirus protein E1. *Biochemistry* **25,** 1335–1339.

ROYCHOUDHURY, R., and WU, R. (1980). Terminal transferase-catalyzed addition of nucleotides to the 3′ termini of DNA. *In* "Methods in Enzymology" (L. Grossman and K. Moldave, Eds.), Vol. 65, pp. 43–62. Academic Press, New York.

SKINNER, M. A., and SIDDELL, S. G. (1983). Coronavirus JHM: Nucleotide sequence of the mRNA that encodes nucleocapsid protein. *Nucleic Acids Res.* **11,** 5045–5054.

SMITH, H. O., and BIRNSTIEL, M. L. (1976). A simple method for DNA restriction site mapping. *Nucleic Acids Res.* **3,** 2387–2398.

STERN, D. F., and KENNEDY, S. I. T. (1980). Coronavirus multiplication strategy. II. Mapping the avian infectious bronchitis virus intracellular RNA species to the genome. *J. Virol.* **36,** 440–449.

STOHLMAN, S. A., and LAI, M. M. C. (1979). Phosphoproteins of murine hepatitis virus. *J. Virol.* **32,** 672–675.

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

TOWBIN, H., STAEHELIN, T., and GORDON, J. (1979). Electrophoretic transfer of proteins from polyacrylamide gels to nitrocellulose sheets: Procedure and some applications. *Proc. Natl. Acad. Sci. USA* **76,** 4350–4354.