J. gen. Virol. (1985), **66**, 2253–2258. *Printed in Great Britain*

Key words: *infectious bronchitis/coronavirus IBV/mRNA D/nucleotide sequence*.

Sequencing of Coronavirus IBV Genomic RNA: Three Open Reading Frames in the 5′ 'Unique' Region of mRNA D

By M. E. G. BOURSNELL,* M. M. BINNS AND T. D. K. BROWN

Houghton Poultry Research Station, Houghton, Huntingdon, Cambs. PE17 2DA, U.K.

(*Accepted 2 July 1985*)

SUMMARY

The nucleotide sequence of a genomic cDNA clone corresponding to the 5′ terminal domain of mRNA D of the Beaudette strain of infectious bronchitis virus (IBV) has been determined. This region contains three open reading frames which predict polypeptides of molecular weights 6700 (6·7K), 7·4K and 12·4K. The predicted 12·4K polypeptide has a codon usage very similar to that predicted for the products of the IBV nucleocapsid, membrane and spike genes. The sequence also predicts a hydrophobic, potentially membrane-anchoring, region in the N terminal half of the 12·4K polypeptide, and a hydrophilic C terminus.

Coronaviruses are enveloped viruses with a single-stranded RNA genome of positive polarity (Siddell *et al.*, 1983; Sturman & Holmes, 1983). The genome of infectious bronchitis virus (IBV) is about 20 kilobases in length (Stern & Kennedy, 1980*a*; Siddell *et al.*, 1983). In IBV-infected cells six major mRNA species are produced. These mRNAs, designated A to F, range in length from about 2 kb to genome length, and have been shown to share a common 3′ terminus and form an overlapping or 'nested' set (Stern & Kennedy, 1980*a*, *b*) (see Fig. 1). Translation studies *in vitro* have demonstrated that mRNAs A, C and E encode the three major viral proteins: the nucleocapsid protein, the membrane glycoprotein and the precursor to the spike or surface projection glycoprotein, respectively (Stern & Sefton, 1984). Sequencing of the IBV genome has shown that the coding sequences for these polypeptides lie largely within the 'unique' 5′ terminal region of each mRNA species which is not present in the next smallest mRNA (Boursnell *et al.*, 1984, 1985; Binns *et al.*, 1985). However, no specific translation products have, to date, been detected from mRNAs B and D (Stern & Sefton, 1984; Boursnell & Brown, 1984). Sequencing studies of genomic RNA in the regions of the 5′ terminal domains of mRNAs B and D have been carried out to determine whether these contain potential coding sequences. The 5′ terminal sequence of mRNA B contains two open reading frames (ORFs) which potentially code for polypeptides of 7·5K and 9·5K (Boursnell & Brown, 1984). In this paper, we present the sequence, obtained from genomic cDNA clones, of the 'unique' 5′ terminal region of mRNA D.

The isolation of the cDNA clone, pMB179, which contains these sequences, has already been described (Binns *et al.*, 1985). Briefly, a 13 base oligonucleotide primer, complementary to sequences at the 5′ end of mRNA C (Boursnell *et al.*, 1984), was used to prime cDNA synthesis from purified IBV Beaudette (Beaudette & Hudson, 1937) viral genomic RNA. One of the clones obtained, pMB179, contained a 5·3 kb insert which DNA sequence analysis subsequently showed had a 3′ end 12 bases from the 5′ end of the primer sequence. Prior to dideoxy sequencing (Sanger *et al.*, 1977; Bankier & Barrell, 1983), *Pst*I and *Rsa*I digests of pMB179 were subcloned into *Pst*I-digested M13mp11 and *Sma*I-digested, phosphatase-treated M13mp10, respectively. DNA sequence data were also obtained in the region of mRNA D by sequencing of DNase I-treated (Anderson, 1981) or sonicated (Deininger, 1983) fragments of pMB179 which had been subcloned into M13mp10 as described by Binns *et al.* (1985). Fig. 1 shows the position of clone pMB179 and marks the region of sequence presented in this paper.

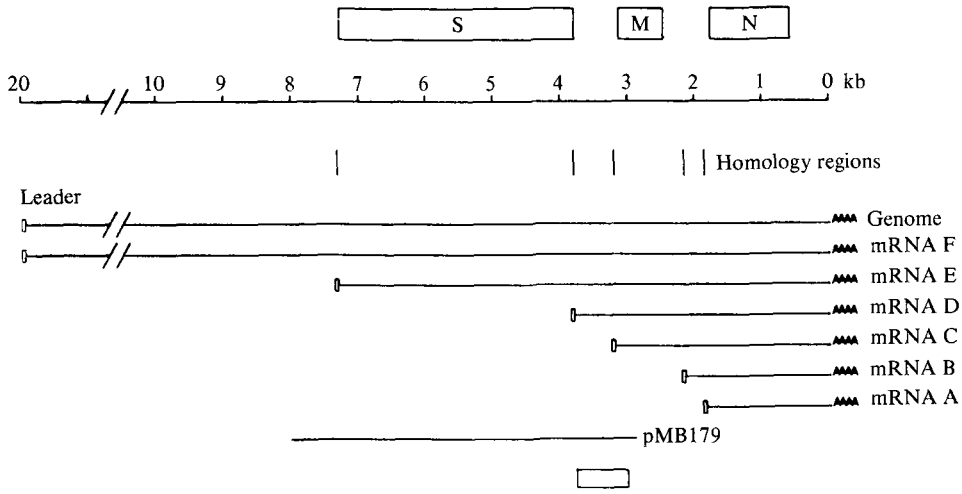

Fig. 1. Genomic organization of infectious bronchitis virus. The 3′ co-terminal 'nested' set of mRNAs is shown. At the top are shown the positions of the genes coding for the major structural components of the virion, the spike (S), membrane (M) and nucleocapsid (N) polypeptides. Also shown are the positions of the 'homology regions' which are sequences present in the genome at positions corresponding to the 5′ termini of the bodies of the mRNAs. The position of clone pMB179 is shown, with the region of sequence presented in Fig. 2 represented by a box.

Seven hundred and fifty-five bases of sequence are presented here. These are shown in Fig. 2 with a translation in single-letter amino acid code of the main ORFs. They extend from a sequence CTGAACAA at position 1, which differs by only one base from sequences which appear at the 5′ ends of the bodies of mRNAs A, B, C (CTTAACAA) and is identical to that found in mRNA E (Brown & Boursnell, 1984; Boursnell *et al.*, 1984, 1985; Boursnell & Brown, 1984; Binns *et al.*, 1985), to an arbitrary position within the sequence of mRNA C. At position 596 is the sequence CTTAACAA, which probably marks the 5′ end of the body of mRNA C. These two sequences lie 3783 and 3188 bases from the poly(A) tract at the 3′ end of the viral genome. These sizes would represent the lengths of the bodies of mRNAs D and C without either leader sequence (Brown *et al.*, 1984) or poly(A) tract, and therefore agree well with the estimated size of these mRNAs of 4·1 and 3·4 kilobases (Boursnell & Brown, 1984). Thus, bases 1 to 596 of this sequence appear to represent the 'unique' 5′ terminal domain of mRNA D which is not present in mRNA C. Bases 1 to 29 code for the COOH terminus of the spike gene and bases 681 to 755 code for the NH_2 terminus of the membrane protein gene (Binns *et al.*, 1985; Boursnell *et al.*, 1984).

There are three ORFs which lie in the 5′ region of mRNA D. The first two non-overlapping ORFs, from bases 32 to 202 and 205 to 396, potentially code for polypeptides of 6·7K and 7·4K. A third ORF, from bases 383 to 706, potentially coding for a polypeptide of 12·4K, overlaps the second ORF by six amino acids and overlaps the coding sequences for the membrane glycoprotein by nine amino acids. Examination of the potential polypeptides encoded by these ORFs shows the 6·7K polypeptide to be neutral and hydrophobic whereas the 7·4K polypeptide is acidic with an overall negative charge of 13. The 12·4K polypeptide would have a hydrophobic N terminal domain and a hydrophilic C terminal domain. The sequences around the initiation codons of the two small ORFs, UNNAUGA and CNNAUGU, are used extremely rarely in functional eukaryotic initiation codons, but are the most common sequences found around 'non-functional' upstream AUGs (22% and 44% of mRNAs surveyed by Kozak, 1983). The sequence flanking the initiation codon of the 12·4K ORF, GNNAUGA, is also fairly rare as a functional initiation codon (2% of mRNAs surveyed) but is not classified as a 'non-functional' upstream AUG (Kozak, 1983). Examination of the codon usage of these three potential polypeptides (Staden, 1984) shows that the 12·4K ORF has a codon usage very similar

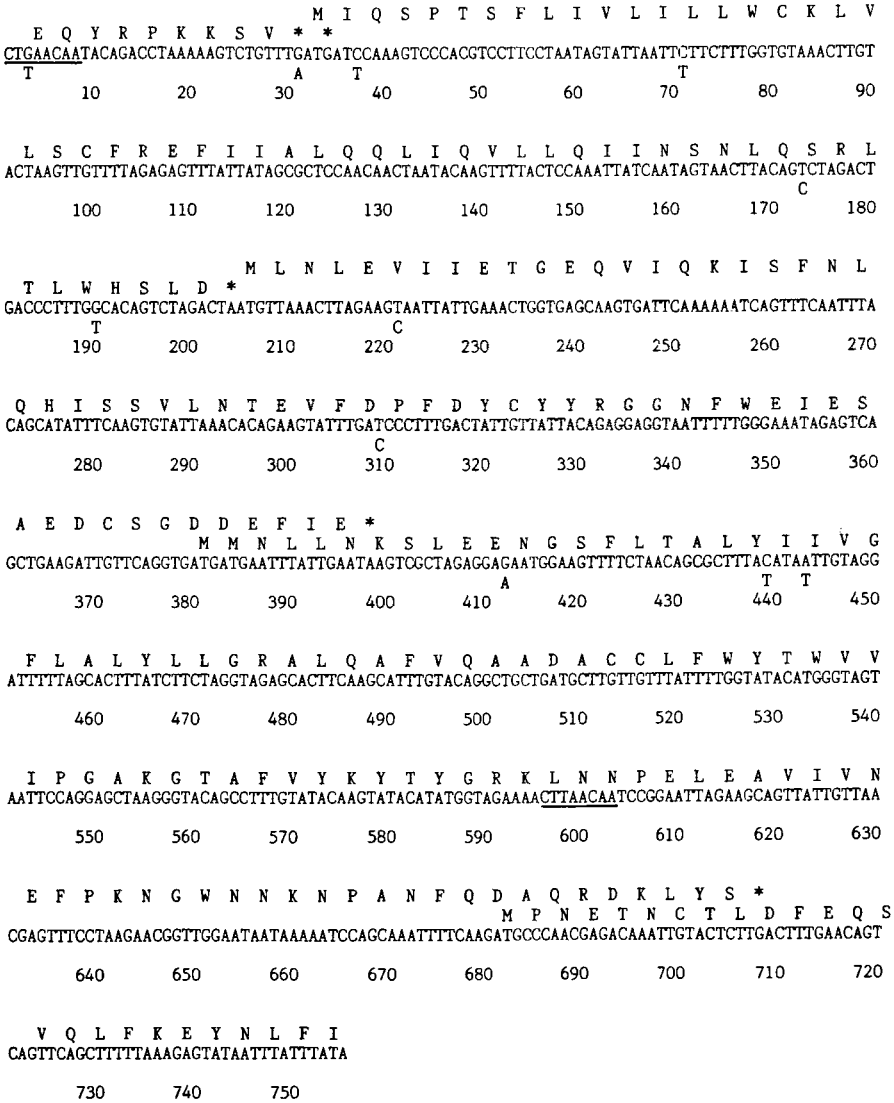

Fig. 2. 755 bases of DNA sequence from the IBV Beaudette genomic cDNA clone pMB179, representing the 5′ terminal domain of IBV mRNA D. A translation in single-letter amino acid code is shown above the three main open reading frames (ORFs). The 'homology regions' (see Fig. 1) are underlined. Where the M41 sequence obtained overlaps the Beaudette sequence (bases 1 to 560) the differences are shown beneath the Beaudette sequence. In all cases the sequence has been completely determined on both strands.

to that predicted for the other IBV polypeptides whose genes have been sequenced, but that the two smaller ORFs have not. These results suggest that the two small ORFs may not code for polypeptides *in vivo* but may only be chance ORFs.

To investigate whether the upstream ORFs are conserved between different IBV strains we have sequenced a cDNA clone from another strain, M41 (Geilhausen *et al.*, 1972), which covers the region of sequence where these small ORFs occur. The M41 clone, 169, was made as described by Boursnell *et al.* (1984) and overlaps the sequences presented here from positions 1 to 560. There are 12 base changes between the two strains. The bases altered in M41 in this region are shown beneath the Beaudette sequence in Fig. 2. The sizes and positions of the two

(*a*) Amino acid sequence of 12·4K polypeptide from IBV mRNA D

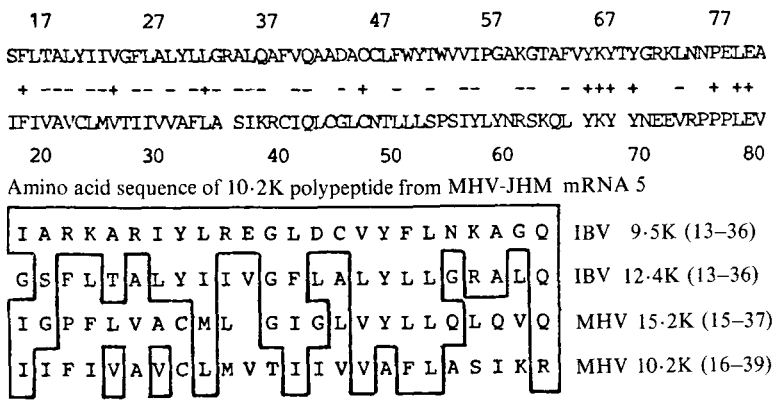

Fig. 3. (*a*) Amino acid homology between IBV 12·4K predicted polypeptide and MHV-JHM 10·2K predicted polypeptide. Plus signs show identical amino acids and minus signs show amino acids with similar (Kanehisa, 1982) properties. (*b*) Comparison of the predicted amino acid sequences of the IBV 12·4K, MHV-JHM 15·2K and MHV-JHM 10·2K putative polypeptides with the IBV 9·5K putative polypeptide. Amino acids boxed-in show residues identical or similar (Kanehisa, 1982) to those of the IBV 9·5K sequence. The distances of the amino acids from the predicted N termini of the polypeptides are shown in parentheses.

small ORFs are conserved in the M41 sequence, but the differences between the two strains at this point are not great enough to imply whether this is significant. However, the 'homology region' CTGAACAA, at position 1, is altered in M41 to CTTAACAA which is the form found in Beaudette at the 5′ ends of the bodies of mRNAs A, B and C. Interestingly, this single base change results in the introduction of a termination codon (UAA) in the coding sequences for the M41 spike protein, which predicts that the M41 spike precursor would lack nine amino acids at the C terminus which are present in the Beaudette polypeptide.

Two of the mRNAs from the mouse coronavirus MHV-JHM, mRNAs 4 and 5, also contain small ORFs which do not appear to code for any of the major structural components of the virion (Skinner & Siddell, 1985; Skinner *et al.*, 1985). The amino acid sequences predicted from the three ORFs in mRNA D and the two ORFs (7·5K and 9·5K) in mRNA B have therefore been compared with the sequences predicted from the three ORFs in mRNAs 4 and 5 from MHV-JHM using various computer programs (Staden, 1982; Kanehisa, 1982; Goad & Kanehisa, 1982). A homology was found between the 12·4K ORF in IBV mRNA D and the 10·2K ORF from MHV-JHM mRNA 5 (Fig. 3*a*). The match is statistically significant, the score being greater than four standard deviations away from that produced by comparing 100 random sequences of the same composition. The hydrophilicity plots (Kyte & Doolittle, 1982) of these two polypeptides are also similar, suggesting that they may be related or have a similar function. In addition there is some similarity between the N terminal regions of four of these putative small polypeptides. Fig. 3(*b*) shows these results.

The fact that the codon usage of the 12·4K putative polypeptide is very similar to that predicted for the nucleocapsid, membrane and spike polypeptides strongly suggests that the largest ORF in mRNA D does code for a product *in vivo*. It is not clear at the moment what, if any, is the function of the two smaller 'upstream' ORFs, but it is interesting to note that both mRNA B of IBV (Boursnell & Brown, 1984) and mRNA 5 of MHV-JHM (Skinner *et al.*, 1985) have 5′ terminal regions containing two overlapping ORFs, and thus may code for more than one polypeptide. At the moment it is not possible to say whether the 12·4K product of mRNA D might be a structural component of the virion, but if it were it must only be present at very low levels, since no polypeptide of this size has been detected in [^{3}H]leucine-labelled preparations of virus (Boursnell & Brown, 1984).

The hydrophobic N terminus of the 12·4K polypeptide has a stretch of 21 uncharged amino acids, enriched in hydrophobic residues, which could span the viral membrane, possibly acting as a membrane-anchoring region. Two of the small polypeptides (10·2K and 15·2K) of coronavirus MHV-JHM (Skinner *et al.*, 1985; Skinner & Siddell, 1985) have similar hydrophobic domains and it has been suggested that they may play a role in siting membrane-bound transcription or replication complexes (Skinner & Siddell, 1985). The 12·4K polypeptide of IBV may have a similar function but in view of the fact that these polypeptides are probably not translated until the subgenomic mRNAs have already been transcribed, an involvement with replication complexes, producing full-length viral RNA, seems the more likely of these two suggestions. Another possibility is that they could be involved in a switch from transcription to replication activities, which is suggested by the observation that, in MHV, late in infection the genomic RNA is synthesized at a faster rate than the subgenomic RNAs (Brayton *et al.*, 1984).

We are grateful to Penny Gatter, Bridgette Britton, Anne Foulds and Ian Foulds for excellent technical assistance. This research was carried out under Research Contract No. GBI-2-011-UK of the Biomolecular Engineering Programme of the Commission of the European Communities.

REFERENCES

ANDERSON, S. (1981). Shotgun DNA sequencing using cloned DNase I-generated fragments. *Nucleic Acids Research* **9**, 3015–3027.

BANKIER, A. & BARRELL, B. G. (1983). Shotgun DNA sequencing. In *Techniques in the Life Sciences (Biochemistry), vol. B5: Techniques in Nucleic Acid Biochemistry*, pp. B508, 1–34. Edited by R. A. Flavell. Ireland: Elsevier.

BEAUDETTE, F. R. & HUDSON, C. B. (1937). Cultivation of the virus of infectious bronchitis. *Journal of the American Veterinary Medical Association* **90**, 51–60.

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.

BOURSNELL, M. E. G. & 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. & BINNS, M. M. (1984). Sequence of the membrane protein gene from avian coronavirus IBV. *Virus Research* **1**, 303–313.

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

BRAYTON, P. R., STOHLMAN, S. A. & LAI, M. M. C. (1984). Further characterisation of mouse hepatitis virus RNA-dependent RNA polymerases. *Virology* **133**, 197–201.

BROWN, T. D. K. & BOURSNELL, M. E. G. (1984). Avian infectious bronchitis virus genomic RNA contains sequence homologies at the intergenic boundaries. *Virus Research* **1**, 15–24.

BROWN, T. D. K., BOURSNELL, M. E. G. & BINNS, M. M. (1984). A leader sequence is present on mRNA A of avian infectious bronchitis virus. *Journal of General Virology* **65**, 1437–1442.

DEININGER, P. L. (1983). Random subcloning of sonicated DNA: application to shotgun DNA sequence analysis. *Analytical Biochemistry* **129**, 216–223.

GEILHAUSEN, H. E., LIGON, F. B. & LUKERT, P. D. (1972). The pathogenesis of virulent and avirulent avian infectious bronchitis virus. *Archiv für die gesamte Virusforschung* **40**, 285–290.

GOAD, W. B. & KANEHISA, M. (1982). Pattern recognition in nucleic acid sequences. I. A general method for finding local homologies and symmetries. *Nucleic Acids Research* **10**, 247–263.

KANEHISA, M. I. (1982). Los Alamos sequence analysis package for nucleic acids and proteins. *Nucleic Acids Research* **10**, 183–196.

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

KYTE, J. & DOOLITTLE, R. F. (1982). A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology* **157**, 105–132.

SANGER, F., NICKLEN, S. & COULSON, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences, U.S.A.* **74**, 5463–5467.

SIDDELL, S., WEGE, H. & TER MEULEN, V. (1983). The biology of coronaviruses. *Journal of General Virology* **64**, 761–776.

SKINNER, M. A. & SIDDELL, S. G. (1985). Coding sequence of coronavirus MHV-JHM mRNA 4. *Journal of General Virology* **66**, 593–596.

SKINNER, M. A., EBNER, D. & SIDDELL, S. G. (1985). Coronavirus MHV-JHM mRNA 5 has a sequence arrangement which potentially allows translation of a second, downstream open reading frame. *Journal of General Virology* **66**, 581–592.

STADEN, R. (1982). An interactive graphics program for comparing and aligning nucleic acid and amino acid sequences. *Nucleic Acids Research* **10**, 2951–2961.

STADEN, R. (1984). Graphic methods to determine the function of nucleic acid sequences. *Nucleic Acids Research* **12**, 521–538.

STERN, D. F. & KENNEDY, S. I. T. (1980*a*). Coronavirus multiplication strategy. I Identification and characterisation of virus-specified RNA. *Journal of Virology* **34**, 665–674.

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

STERN, D. F. & SEFTON, B. M. (1984). Coronavirus multiplication: the locations of genes for the virion proteins on the avian infectious bronchitis virus genome. *Journal of Virology* **50**, 22–29.

STURMAN, L. S. & HOLMES, K. V. (1983). The molecular biology of coronaviruses. *Advances in Virus Research* **28**, 35–112.

(*Received 23 May 1985*)