THE COMPLETE NUCLEOTIDE SEQUENCE OF AVIAN INFECTIOUS BRONCHITIS VIRUS: ANALYSIS OF THE POLYMERASE-CODING REGION

M. E. G. Boursnell, T. D. K. Brown, I. J. Foulds,
P. F. Green, F. M. Tomley and M. M. Binns

Houghton Poultry Research Station
Houghton
Huntingdon
Cambridgeshire PE17 2DA
England, UK

INTRODUCTION

Avian infectious bronchitis virus (IBV) is the type species of the family Coronaviridae (Siddell et al., 1983). It has a large positive-stranded RNA genome which has been estimated at 20-24 kilobases (Lomniczi & Kennedy, 1977). As with other coronaviruses, a number of subgenomic messenger RNA species are produced in infected cells which form a 3'-coterminal nested set (Stern & Kennedy, 1980a; 1980b). In the case of IBV there are six mRNA species in total, which are designated mRNAs A-F, mRNA A being the smallest, and mRNA F being of genome size. mRNAs A, C and E encode the three main structural components of the virion, the nucleocapsid polypeptide, the membrane polypeptide and the precursor polypeptide to the spike (Stern & Sefton, 1984). mRNA D encodes at least one product, a 12.4 kilodalton polypeptide of unknown function (Smith et al., this volume), but no product has yet been detected for mRNA B. The coding regions of mRNAs A-E are situated in the 3'-most 7.3 kilobases of the IBV genome, the nucleotide sequence of which has been determined previously from cDNA clones (Boursnell et al., 1984, 1985a, 1985b; Boursnell & Brown, 1984; Binns et al., 1985b).

RNase T_1 fingerprint analysis reveals no difference between messenger RNA F, the genome-sized mRNA present in infected cells, and the virion RNA (Stern & Kennedy, 1980a), although the possibility of minor differences between them cannot be ruled out. If they are taken as being identical, then the remainder of the IBV genome constitutes the 'unique' region of mRNA F, in other words that part of mRNA F not present in the smaller mRNAs. Because the genome is infectious (Lomniczi, 1977) and because there is no evidence for a virion-associated RNA polymerase (Schochetman et al., 1977) this region of the genome is thought to encode a polymerase or polymerases which carry out the necessary replication and transcription functions of the virus. We have now determined the nucleotide sequence, from cDNA clones, of the 'unique' region of mRNA F. This completes the sequence of the IBV genome.

M. M. C. Lai et al. (eds.), *Coronaviruses*

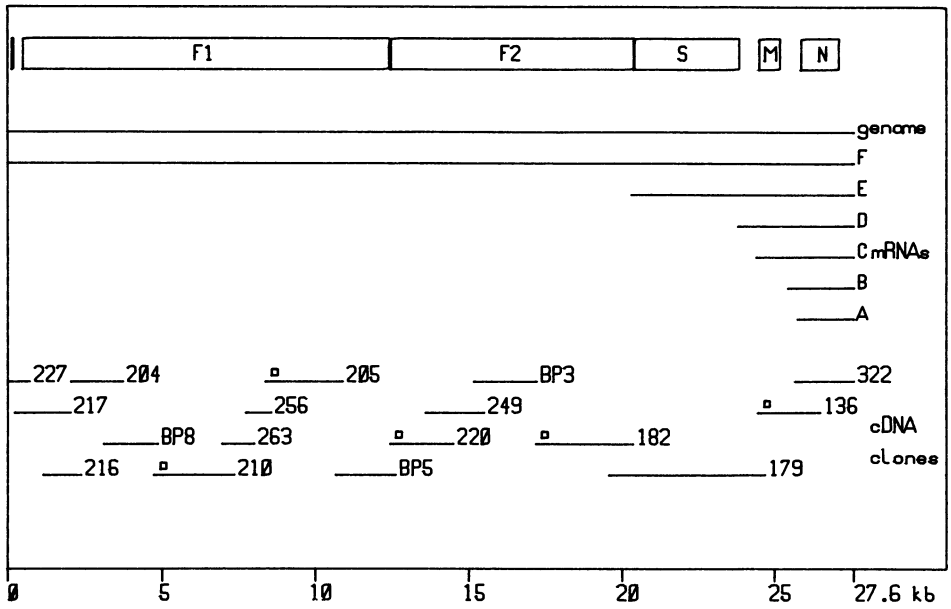

Fig. 1. Diagram showing the positions of all the cDNA clones used in obtaining the nucleotide sequence. The squares at the end of some of the clones show the positions of oligonucleotide primers used to prime synthesis of cDNA for adjacent clones. Above the clones are shown mRNAs A-F. The positions of the main open reading frames are marked with boxes. The small open reading frame at position 131 is also shown.

RESULTS

cDNA Cloning

17 cDNA clones have been used to obtain the complete sequence of the genome of the Beaudette strain of IBV. These cover the 3'-most 27569 kb of the genome. The 39 nucleotides at the 5' end of the genome have not been obtained in cDNA clones, but the sequence has been determined by Maxam & Gilbert (1980) sequencing of primer-extended products from virion RNA. These clones are shown in Figure 1. The majority of the clones used to obtain the sequence of the unique region of mRNA F were obtained by a random priming method, using calf thymus DNA primers (Binns et al., 1985a). These clones, numbers 217, 216, 204, 210, 205, 220 and 249 were mapped by identifying overlaps using Southern blotting (Southern, 1975). Clone 227 was identified as coming from the 5' end of the genome by probing a random library of cDNA clones with a leader-specific oligonucleotide (Brown et al., 1986). Clone 182 was produced by priming with a specific oligonucleotide primer. At this stage, the clones did not form a single contiguous block but fell into five groups of overlapping clones, with four gaps remaining. Oligonucleotide primers synthesised using sequence information from clones on the 3' side of the gaps were used to obtain cDNA clones in the region of the gaps. Clones spanning the gaps were identified by using either 'prime-cut' probes (Biggin et al., 1984) made from M13 subclones of cDNA clones on either side of the gaps or by using Southern blotting. The five clones 256,

263, BP3, BP5 and BP8 were identified in this way and the overlaps confirmed by sequencing.

DNA Sequencing

Random M13 subclones of the cDNA clones were made, and sequenced by the dideoxy method, as previously described (Boursnell et al., 1985a, 1985b). All sequence information has been obtained from both strands of the DNA and the majority of the sequence has been obtained several times from different M13 clones. For the 24,765 bases of sequence contained in the 14 cDNA clones used to obtain the unique region of mRNA F, 203,113 bases have been sequenced, so that each base has on average been sequenced 8.2 times.

Open Reading Frames

The size of the unique region of mRNA F is 20,298 bases. The total size of the IBV genome, excluding the polyA tail, is 27,608 bases. The complete sequence is not presented here, but will be published elsewhere (Boursnell et al., 1986). Figure 1 shows the positions of the largest open reading frames (ORFs) in the IBV genome. It can be seen that most of the sequence of the unique region of mRNA F codes for two very large ORFs. These two large ORFs have been designated F1 and F2, and could code for polypeptides of predicted molecular weights 441 kilodaltons and 300 kilodaltons. Although the unique region of mRNA F is dominated by these two large ORFs, the first AUG codon to occur in the genome is not at the start of F1, but is at position 131, at the start of a very small ORF of 11 amino acids. The sequence of the first 600 bases at the 5' end of the genome is shown in Figure 2, with translations of the small ORF and the NH_2-terminal of F1. The sequence context around the first AUG, at position 131, does not conform well to the Kozak consensus for functional initiation codons (Kozak, 1984). The second initiation codon

```
   1  ACTTAAGATAGATATTAATATATATCTATTACACTAGCCTTGCGCTAGATTTTTAACTTA   60

  61  ACAAAACGGACTTAAATACCTACAGCTGGTCCTCATAGGTGTTCCATTGCAGTGCACTTT  120

                  M  A  P  G  H  L  S  G  F  C  Y  *
 121  AGTGCCCTGGATGGCACCTGGCCACCTGTCAGGTTTTTGTTATTAAAATCTTATTGTTGC  180

 181  TGGTATCACTGCTTGTTTTGCCGTGTCTCACTTTATACATCTGTTGCTTGGGCTACCTAG  240

 241  TGTCCAGCGTCCTACGGGCGTCGTGGCTGGTTCGAGTGCGAGGAACCTCTGGTTCATCTA  300

 301  GCGGTAGGCGGGTGTGTGGAAGTAGCACTTCAGACGTACCGGTTCTGTTGTGTGAAATAC  360

 361  GGGGTCACCTCCCCCCACATACCTCTAAGGGCTTTTGAGCCTAGCGTTGGGCTACGTTCT  420

 421  CGCATAAGGTCGGCTATACGACGTTTGTAGGGGGTAGTGCCAAACAACCCCTGAGGTGAC  480

                                                  M  A  S  S
 481  AGGTTCTGGTGGTGTTTAGTGAGCAGACATACAATAGACAGTGACAACATGGCTTCAAGC  540

       L  K  Q  G  V  S  P  K  P  R  D  V  I  L  V  S  K  D  I  P
 541  CTAAAACAGGGAGTATCTCCCAAACCACGGGATGTCATTCTTGTGTCCAAAGACATCCCT  600
```

Fig. 2. The 600 bases from the 5' end of the IBV genome. A translation in single letter amino acid code is shown for the first small open reading frame and for the start of F2. The homology region at position 57 is underlined.

is at the start of F1, and the sequence context around this AUG, with a purine at -3, conforms well to the Kozak consensus. The second large ORF, F2, extends into the 'unique' region of mRNA E, and overlaps the coding sequence for the spike precursor gene by 16 amino acids.

The possibility has occurred to us that the presence of two ORFs in the sequence as obtained by us is due to either a sequencing error or a mutant cDNA clone. Accordingly the sequence in this region has been checked extremely carefully. The sequence on both strands in this region is perfectly clear, with no signs of compressions or any other hidden artefacts (see Figure 3a,b). We have, however, attempted to reveal any cryptic compressions by running the sequence reactions on highly denaturing gels, either 40-50% formamide gels or high-temperature (80°C) thermostatted gels. Other techniques which can resolve compressions have also been used, namely the use of deoxyinosine triphosphate (Bankier & Barrell, 1983) or deoxy-7-deazaguanosine triphosphate (Mizusawa et al., 1986) to replace deoxyguanosine triphosphate, and cytosine modification of the sequence reaction products (Ambartsumyan & Mazo, 1980). All of these methods are effective at revealing compressions produced by DNA secondary structure, but in every case the sequence in the region between F1 and F2 appeared exactly the same. To rule out the possibility of a cDNA clone synthesised from a mutant virus, or a reverse transcriptase error, the sequence has also been obtained from two additional independent cDNA clones, both by sequencing directly from the

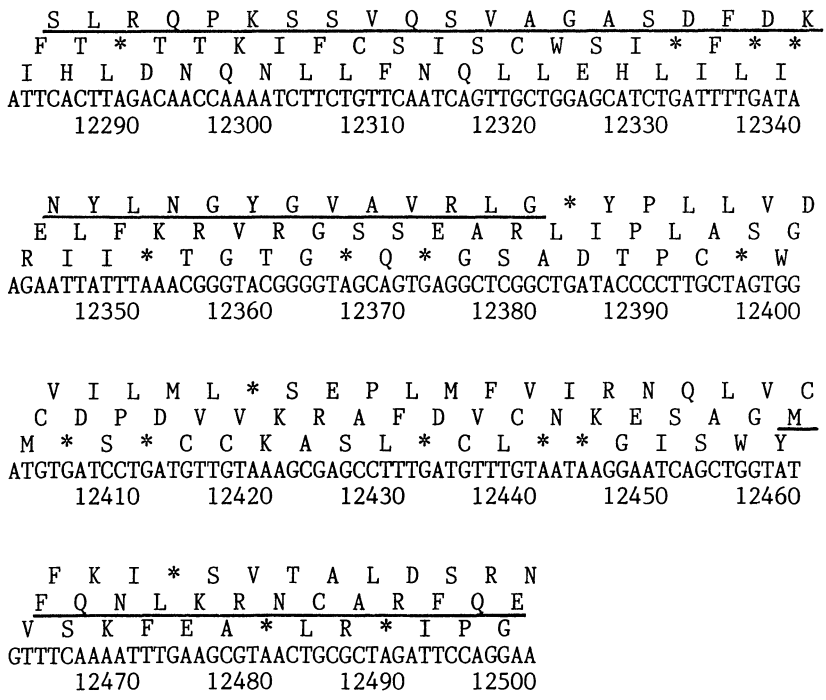

Fig. 3a. The nucleotide sequence in the region between F1 and F2, with a translation in single-letter amino acid code of the three reading frames. Stop codons are marked as asterisks. The open reading frames of F1 and F2 are shown underlined.

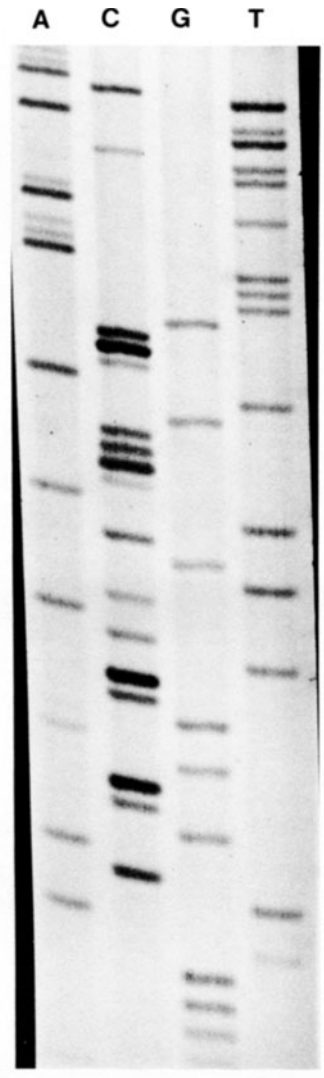

Fig. 3b. A DNA sequencing gel obtained by sequencing a double-stranded cDNA clone using an oligonucleotide primer. The sequence shown is from 12333-12390, and is the reverse complement of the sequence shown in 4a.

double-stranded DNA (Korneluk et al., 1985) and after subcloning into M13, and is again identical (see Figure 3b). In addition the sequence has been confirmed by sequencing directly from the virion RNA, of both the Beaudette and M41 strains of IBV, using an oligonucleotide primer (Caton et al., 1982).

Computer Analysis

Extensive computer analysis has been carried out on the amino acid sequences of F1 and F2. The predicted amino acid sequences have been compared against themselves and against each other, to search for

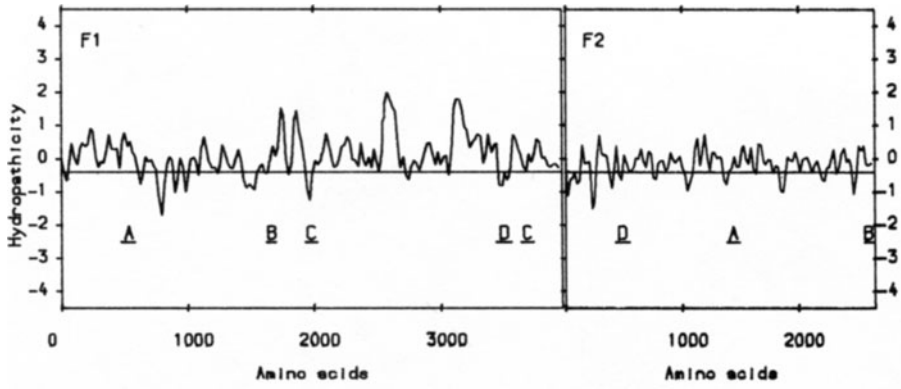

Fig. 4. Hydropathicity plots (Kyte & Doolittle, 1982) of the predicted amino acid sequences of ORFs F1 and F2. Values above the line are hydrophobic and values below the line are hydrophilic. The hydropathicity is calculated using a moving window of 41 amino acids, with a value plotted every 21 residues. The pairs of bars marked A, B, C and D show regions of partial homology.

similarities between the products of the two genes. DIAGON (Staden, 1982), a dot matrix comparison program, reveals no homologies either within or between F1 and F2. However small regions of homology can be detected using the program FASTP (Lipman & Pearson, 1985). These are not perfect homologies but are identified by scoring matches between similar amino acids as well as identical amino acids. Figure 4 shows these imperfect repeats marked as bars beneath hydropathicity plots of F1 and F2. Searches for homologies with other amino acid sequences, such as RNA polymerases from other viruses, have been carried out using the program FASTP and the NBRF protein identification resource (George et al., 1986). There is no extensive homology with any polymerase in the database although several short regions of homology with polymerases can be identified which do not rise significantly above the background of matches with apparently unrelated proteins. One region in F2, between amino acids 1248 and 1356, has a fairly good match with the nsP2 protein of sindbis virus, a protein which is involved in RNA replication (Strauss & Strauss, 1983), and also with the 1a protein of brome mosaic virus (see Figure 5a). One of the most interesting matches occurs at the 5' end of F1. The first 300 amino acids have a fairly low but extensive homology with a replication initiation protein from E.coli (Germino & Bastia, 1982). This match is shown in Figure 5b and suggests that this region of the polymerase protein may be involved in initiation of replication of either the positive or negative strands.

A conserved amino acid sequence has been found in RNA-dependent RNA polymerases of several viruses (Kamer & Argos, 1984). This segment consists of two aspartic acid residues flanked by hydrophobic residues and it is also found in several retroviral reverse transcriptases, suggesting that it is a possible active site or nucleic acid recognition sequence in RNA-dependent polymerase molecules. A consensus sequence from fifteen such segments from different viruses has been calculated using the program ANALYSEP (Staden, 1984) and compared against the amino acid sequences of F1 and F2. At position 3455 in F1 there is fairly good match which scores as well as at least three of the original fifteen segments. The most highly conserved residues in the original consensus

a)

```
BMV SCHRLLVDEAGLLHYGQLLVVAALSKCSQVLAF-GDTEQ-------ISFKSRDAGFKLLHGNLQYDRRDV
    ::. ::::::..:  .:  ...  ..  :.  ::..:       .. . .  ..... . . . ..:.
IBV SCDILLVDEVSMLTNYELSFINGKINYQYVVYV-GDPAQLPAPRTLLNGSLSPKDYNVVTNLMVCVKPDI
    . ..: :::.    .  :  . . .. .  : . ::: :      .   .  :.     .  .:.: ..
SV  AVEVLYVDEAFACHAGALLALIAIVRPRKKVVLCGDPMQ-----CGFFNMMQLKVHFNHPEKDICTK-TF

BMV -VHKTYRCPQDVIAAVNLLKRKCGNRDTKY
     . : ::::.....:. :
IBV FLAKCYRCPKEIVDTVSTLVYDGKFIANNP
    .   . ::.. ....:::: ::::. ..::
SV  YKYISRRCTQPVTAIVSTLHYDGKMKTTNP
```

b)

```
F1      61 QFKFETVCGLFLLKGVDKITPGVPAKVLKATSKLADLEDIFGVSPLARKYRELLKTACQW
                   : ...:::::... . ::...:: . ..:   .       . :. .
RIP     50 ERGRVFKIRAEDLAALAKITPSLAYRQLKEGGKLLGASKISLRGDDIIALAKELNLPFTA

F1     129 SLTVEALDVRAQTLDEIFDPTEILWLQVAAKIH--VSSM--AMRRLVGEVTAKVMDALGS
           . . :.::..  .  .  . .. : :. . .::  .::.   ........ .  .   .
RIP    110 KNSPEELDLNIIEWIAYSNDEGYLSLKFTRTIEPYISSLIGKKNKFTTQLLTASLRLSSQ

F1     179 NLSALFQIVKQQIARIFQKALAIFENVNELPQRIAALKMAFAKCARSITVVVVERTLVVK
             :.:.:.... .. :.:   .. .:.:: . . :   .:.: . .:.    . ..  .
RIP    170 YSSSLYQLIRKHYSN-FKKKNYFIISVDELKEELIAY--TFDK-DGNIEYKYPDFPIFKR

F1     240 EFAGTCLASINGAVAKFFEELPNGFMGSKIFTTLAFFKEAAVRVVENIPNAPRGTKGFEV
           .  .. .:.:.  ..  : ...   .... ...: :
RIP    230 DVLNKAIAEIKKKTEISFVGFTVHEKEGRKISKLKFEFVVDEDEFSGDKDDEAFFMNLSE
```

Fig. 5a. Comparison between amino acid sequences of brome mosaic virus (BMV) 1a protein residues 748-838, infectious bronchitis virus (IBv) F2 ORF residues 1248-1356, and sindbis virus (SV) nsP2 protein residues 785-878. A colon shows similar (Kanehisa, 1982) amino acids. The dashes in the sequences are padding characters inserted to achieve optimal alignment.

Fig. 5b. Comparison between the N-terminal of F1 and a replication initiation protein (RIP) from E.coli (Germino & Bastia, 1982). Between the two sequences there is 17.6% identity in 204 amino acids overlap. Colons, dots and dashes are as for 5a.

are immediately flanking the Asp-Asp pair, namely Tyr-Gly-Asp-Asp-Ile-Leu. The IBV sequence at this point is Tyr-Cys-Asp-Asp-Ile-Leu with a Cys for Gly substitution, both of which are uncharged, polar residues. The similarity of this site with that found in other viral RNA-dependent polymerases suggests that it may have some functional significance which is related to polymerase activity.

Codon Usage

There is a distinctive bias in codon usage in the predicted amino acid sequences of F1 and F2. This is shown in Table 1. Although the A-

Table 1. Codon usage table for the predicted amino acid sequences of F1 and F2.

F UUU	286	S UCU	149	Y UAU	238	C UGU	181
F UUC	85	S UCC	13	Y UAC	85	C UGC	42
L UUA	128	S UCA	81	* UAA	0	* UGA	1
L UUG	140	S UCG	24	* UAG	1	W UGG	87
L CUU	169	P CCU	98	H CAU	91	R CGU	58
L CUC	40	P CCC	21	H CAC	39	R CGC	33
L CUA	54	P CCA	99	Q CAA	140	R CGA	12
L CUG	41	P CCG	20	Q CAG	78	R CGG	5
I AUU	178	T ACU	173	N AAU	264	S AGU	119
I AUC	27	T ACC	30	N AAC	93	S AGC	29
I AUA	129	T ACA	155	K AAA	220	R AGA	68
M AUG	125	T ACG	28	K AAG	209	R AGG	37
V GUU	328	A GCU	205	D GAU	269	G GGU	237
V GUC	73	A GCC	35	D GAC	112	G GGC	60
V GUA	140	A GCA	163	E GAA	192	G GGA	74
V GUG	115	A GCG	34	E GAG	126	G GGG	19

and U-richness of the IBV sequence (A 27%, C 21%, G 23%, U 29%) means that there will be a predominance of A and U in the third base position, there is often a strong preference for U over and above this. For example the codon CUU is used more than three times as frequently as CUA for leucine. This codon bias can be used to predict which frame of the three possible reading frames is likely to be coding (see DISCUSSION).

Homology Regions

In IBV the conserved sequences which occur at the 5' end of the body of the mRNAs are often referred to as 'homology regions' (Brown & Boursnell, 1984). The sequence in these regions is CTTAACAA in the case of mRNAs A, B, C and F, and CTGAACAA in the case of mRNAs D and E. Other homologies between the different regions can be identified but this 'core' homology, which can be written as CT(A/G)AACAA, is the most highly conserved. The sequence CTGAACAA however also occurs at positions 599 and 3293 on the IBV genome. Any mRNA species associated with these regions might well not have been detected, either because it ran too near mRNA F or because it was of low abundance. Nevertheless, the position of these regions within the coding sequences of F1 suggests that they probably do not represent the 5' ends of the bodies of mRNA species. We have attempted to determine whether there is some feature of the sequence context surrounding these two homology regions which set them apart from homology regions which are known to occur at the 5' ends of mRNAs. We have therefore calculated a consensus from the sequences surrounding the known homology regions of mRNAs A-F. This consensus sequence is 18 bases in length with the first base of the core homology region falling at position 7. This consensus has been compared to the complete sequence using the computer program FITCONSENSUS, which assigns a score to each match it finds (Devereux et al., 1984). The program successfully identifies the known homology regions A-F with scores ranging from 74.6 to 64.1. The 34 next best regions identified have a range of scores well

Table 2. Computer search for homology regions using FITCONSENSUS (Devereux et al., 1984). The consensus sequence used for the search was made from the sequences at the known homology regions A-F.

FITCONSENSUS score	position	mRNA if known
74.63	51	mRNA F
72.00	23819	mRNA D
66.68	24414	mRNA C
65.95	25766	mRNA A
64.11/61.47	25454/25465	mRNA B
58.84	21236	
58.79	5537	
57.05	11203	
56.26	12547	
56.21	15683	
55.32	9116	
55.26	3287	
54.47	15488	
54.47	20527	
54.47	22876	
54.47	24679	
54.42	16300	
53.63	18171	

separated from those of the known homology regions, with a tight cluster of scores (51.8-58.8). Some of this data is shown in Table 2. The low scores of the two possible homology regions at 599 and 3293 suggest that they represent a chance match with the 'core' homology sequence, but that when the flanking sequences are considered the differences are enough to ensure that they are not major sites for the binding of the leader/polymerase complex.

Mutation Rates

The error rate of RNA polymerases is fairly high, being estimated at about 1 in 10,000 (Steinhauer & Holland, 1986). With such a mutation rate, over the 20kb of sequence in the unique region of mRNA F, there would be expected to be one or two changes at each round of replication. Mutant, and possibly defective, molecules might well accumulate and, as long as they were packaged into virions, would then be isolated as virion RNA, cloned and sequenced. When the number of rounds of replication that have occurred since the virus was originally plaque-purified are taken into account, one might expect that every cDNA clone isolated for a particular region of the RNA might be different. These predictions notwithstanding, there is no evidence for a very high mutation rate in the cDNA clones that we have sequenced. For the clones sequenced for mRNA F there are 4659 bases of overlap, which have been sequenced on two independent clones. In all the overlapping sequences there is 100% agreement between adjacent clones and no evidence for any mutant clones. This is in contrast to results found by Schubert et al., whilst

```
52 TTTAACTTAACAAAACGGACTTAAATACCTACAGCTGGTCCTCATAGGTGTTCCATTGCAGTGCACT 118
   :: :::::::: ::   ::: ::::   ::   ::   ::: :: :: :: : : ::: :: : ::
48 TTAAACTTAACTTAA---ACTAAAATT--TAGCTCTTCCCCTAATGGGCGTCCTAGTGCTGTACCCT 109
```

Fig. 6. Comparison between (top) the nucleotide sequence of the 5' end of the genome and (bottom) the reverse complement of the 3' end of the genome. Colons show identical bases. The dashes are padding characters inserted to achieve optimal alignment.

sequencing the polymerase gene of vesicular stomatitis virus (VSV). The gene of 6,380 nucleotides was sequenced on average from three independent cDNA clones. They found 20 nucleotide changes, including four insertions or deletions. This results in an overall mutation rate of approximately 10^{-3}. For the IBV sequence which can be checked on an independent cDNA clone, a mutation rate of 10^{-3} would lead to approximately 9 nucleotide differences. The fact that there are none prompts speculation that the presumably larger IBV polymerase has a lower intrinsic error rate than the VSV polymerase.

Ends of the Genome

Computer analysis has detected a homology between the non-coding region at the 5' end of the positive strand and the 5' end of the negative strand (i.e. the reverse complement of the sequence at the 3' end of the positive strand). These sequences, shown in Figure 6, are approximately the same distances from the ends of their respective strands, and may play some role in the replication of the positive and negative strands.

DISCUSSION

The 20,500 bases of sequence presented here complete the sequence of the Beaudette strain of infectious bronchitis virus, the type species of the family Coronaviridae. The best estimate of the size of the genome, in which an RNase T_1 digestion method was employed (Lomniczi & Kennedy, 1977) gave the complexity of the IBV genome as 23,170 $\pm$ 920 nucleotides. The genome size as determined by sequencing has proved to be somewhat larger than this, being 27,608 bases excluding the polyA tail.

Analysis of the sequence of the unique region of mRNA F shows it to contain two very large open reading frames. Because a sequencing error or mutant cDNA clone could mean the difference between one ORF and two, the sequence between the two ORFs F1 and F2 has been checked very carefully for any errors (see RESULTS). However the sequence appears perfectly clear in three independent clones, so, unless there is some bizarre form of undetectable sequencing artefact which we have not managed to resolve, the sequence must be taken to be correct.

The problem now arises as to how translation of the second ORF, F2, occurs. There is no homology region which might suggest the presence of an mRNA species, and indeed no species has been detected of this length. It is possible that the ribosomes reinitiate after translation of F1, or that internal initiation occurs, as appears to be the case in VSV (Herman, 1986). However Figure 7 shows some evidence that suggests that neither of these possibilities is the case. We have seen (RESULTS) that IBV coding sequences have a distinctive bias in favour of certain codons,

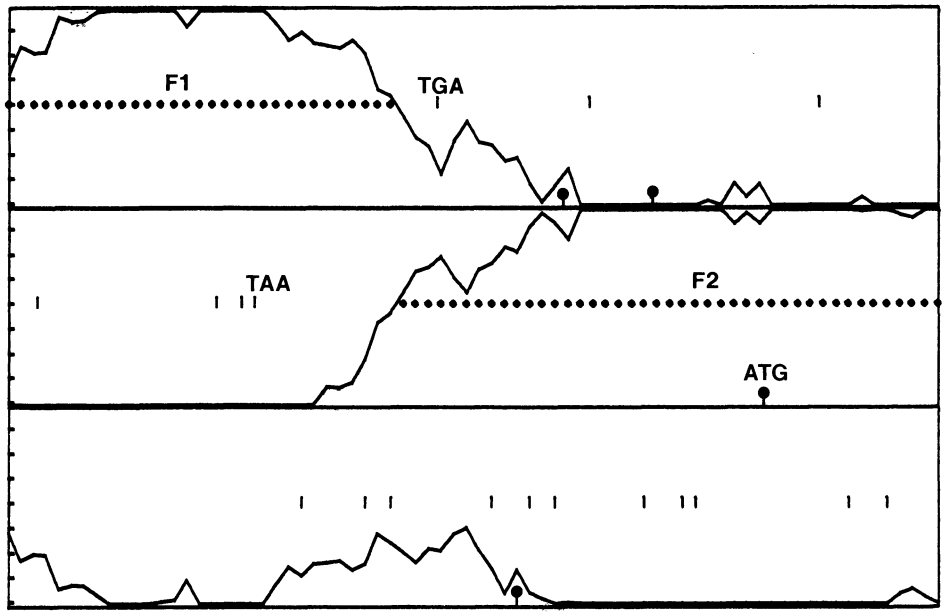

Fig. 7. The same region of sequence as that shown in Fig. 3a. Three reading frames are shown, with a graph for each showing the extent to which each frame conforms to the codon usage found for the amino acid sequence of F1 and F2. The frame which conforms best to the F1/F2 codon usage is marked with a series of dots and marked F1 or F2. Stop codons are marked as short vertical lines along the centre of each frame, and start codons bars with filled-in circles on top. The two stop codons at 12339 (TAA) and 12382 (TGA) are marked, as is the start codon at 12459.

and that this can be used to predict which of the three reading frames is likely to be translated. Figure 7 shows graphically the extent to which each of the three reading frames corresponds to the codon usage adopted by the two large ORFs F1 and F2. It can be seen that the codon usage immediately upstream of the putative initiation codon for F2 conforms extremely well to the IBV F1/F2 codon usage. If this region were merely a 5' non-coding region, similar to that upstream of all the other IBV genes, it would not be expected to have such a distinctive codon bias. It seems therefore that this region of the RNA is translated. There are several possible ways to explain how this may happen, some more baroque than others. For example the subgenomic mRNA F may not in fact be identical to the genome but may have a base missing in this region. This could possibly be achieved by some sort of controlled polymerase error, which would, however, have to occur during the synthesis of mRNA F but not during synthesis of virion RNA. A slightly simpler explanation would be that a 'ribosome slippage' can occur, which allows the ribosome to undergo a frame shift and continue translation of F1 directly through into F2. If the 'frame shift read-through' only occurred at a certain frequency then this could be conceived as a control mechanism to allow coordinated control of two polymerase genes, F1 being expressed at a higher level than F2. If it were necessary for the virus to express both

polymerase genes at these levels immediately on entry into the cell, i.e. before the subgenomic mRNAs were produced, then this might explain why the second polymerase gene does not have its own mRNA species. Ribosomal frameshifting has been described in bacteriophage (Kastelein et al., 1982), prokaryotic (Atkins et al., 1972) and eukaryotic (Fox & Weiss-Brummer, 1980; Jacks & Varmus, 1985) systems. In the case of Rous sarcoma virus (Jacks & Varmus, 1985) expression of a downstream gene (the pol gene) requires a frameshift of -1 (the same as would be needed in the case of IBV) to allow read-through from the gag gene. The authors have convincingly demonstrated that the frame-shifting is sequence-specific, and that the signals, whatever they are, appear to be recognised ten times as efficiently in a eukaryotic system than in a prokaryotic system. Similar experiments can now be performed with IBV to determine whether this frame-shifting can occur in in vitro systems and whether it appears to be sequence specific.

It is possible that the two open reading frames in the unique region of mRNA F represent genes coding for two different polymerases. Two polymerase activities, an early and a late, have been detected in MHV-infected cells (Brayton et al., 1982). These have different ionic requirements and different pH optima. Both polymerase activities are associated with membranes, but the late polymerase is associated with two different membrane fractions, a light fraction which appears to be involved in the synthesis of positive-stranded, genome size RNA and a heavy fraction which also synthesises subgenomic RNAs (Brayton et al., 1984). If F1 and F2 were in fact the genes for two different IBV RNA-dependent RNA polymerases, it might be expected that some relationship could be detected between them by examination of the amino acid sequences. Although there is no overall homology between them, some small regions of homology can be detected (see RESULTS). It is interesting to note that the spacing between the homologies marked A and B in Figure 3 is very similar in both genes, being 1157 amino acids in F1 and 1183 amino acids in F2. It is possible that these represent residual regions of homology between two polymerases which were at one time more closely related.

CONCLUSION

The complete sequence of infectious bronchitis virus illuminates some features of the organisation of the coronavirus genome, but, as is ever the way with sequence data, it leaves us in the dark in other ways. However the availability of the nucleotide sequence of the polymerase genes allows new and exciting experiments to be performed. For example antisera can be raised against products expressed from selected parts of the molecules which will prove useful in detecting the presence of the polymerase in coronavirus infected cells and in unravelling the relationship between the various different polymerase activities which have been detected.

ACKNOWLEDGEMENTS

We are grateful to Bridgette Britton, Penny Gatter, Neil Macey, Rona Chellew and Steve Laidlaw for excellent technical assistance. We would like to thank Dave Cavanagh and Phil Davis for help with the sequencing of the virion RNA. We would also like to thank Alan Bankier for general advice and encouragement during the DNA sequencing and Heather Thomson for typing the manuscript.

REFERENCES

Ambartsumyan, N. S. and Mazo, A. M., 1980, Elimination of the secondary structure effect in gel sequencing of nucleic acids. FEBS Letters, 114:265-268.

Atkins, J. F., Elseviers, D. and Gorini, L., 1972, Low activity of beta-galactosidase in frameshift mutants of Escherichia coli. Proc. Natl. Acad. Sci. USA, 69:1192-1195.

Bankier, A. and Barrell, B. G. Shotgun DNA sequencing, in: "Techniques in the Life Sciences (Biochemistry)" vol. B5, "Techniques in Nucleic Acid Biochemistry", pp. B508, 1-34 ed. R. A. Flavell, Elsevier Science Publishers, Ireland (1983).

Biggin, M., Farrell, P. J. and Barrell, B. G., 1984, Transcription and DNA sequence of the BamHl L fragment of B95-8 Epstein-Barr virus. EMBO J., 3:1083-1090.

Binns, M. M., Boursnell, M. E. G., Foulds, I. J. and Brown, T. D. K., 1985a, The use of a random priming procedure to generate cDNA libraries of infectious bronchitis virus, a large RNA virus, J.Virol., Meths 11:265-269.

Binns, M. M. Boursnell, M. E. G., Cavanagh, D., Pappin, D. J. C. and Brown, T. D. K., 1985b, Cloning and sequencing of the gene encoding the spike protein of the coronavirus IBV, J.Gen.Virol., 66:719-726.

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 Research, 1:303-313.

Boursnell, M. E. G., Binns, M. M., Foulds, I. J. and Brown, T. D. K., 1985a, Sequences of the nucleocapsid genes from two strains of avian infectious bornchitis virus, J.Gen.Virol., 66:573-580.

Boursnell, M. E. G., Binns, M. M. and Brown, T. D. K., 1985b, Sequencing of coronavirus IBV genomic RNA: three open reading frames in the 5' 'unique' region of mRNA D, J.Gen.Virol., 66:2253-2258.

Boursnell, M. E. G., Brown, T. D. K., Foulds, I. J., Green, P. F., Tomley, F. M. and Binns, M. M., 1986, The complete sequence of the genome of infectious bronchitis virus (IBV). J.Gen.Virol., (in press).

Brayton, P. R., Lai, M. M. C., Patton, C. D. and Stohlman, S. A., 1982, Characterisation of two polymerase activities induced by mouse hepatitis virus, J.Virol., 42:847-853.

Brayton, P. R., Stohlman, S. A. and Lai, M. M. C., 1984, Further characterisation of mouse hepatitis virus RNA-dependent RNA polymerases, Virology, 133:197-201.

Brown, T. D. K. and 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. and Tomley, F. M., 1986, Cloning and sequencing of 5' terminal sequences from avian infectious bronchitis virus genomic RNA, J.Gen.Virol., 67:221-228.

Caton, A. J., Brownlee, G. G., Yewdell, J. W. and Gerhard, W., 1982, The antigenic structure of the influenza virus A/PR/8/34 hemagglutinin (Hl subtype), Cell, 31:417-427.

Devereux, J., Haeberli, P. and Smithies, O., 1984, A comprehensive set of sequence analysis programs for the VAX, Nucl. Acids Res., 12:387-395.

Fox, T. D. and Weiss-Brummer, B., 1980, Leaky +1 and -1 frameshift mutations at the same site in a yeast mitochondrial gene, Nature, 288:60-63.

George, D. G., Barker, W. C. and Hunt, L. T., 1986, The protein identification resource (PIR), Nucl. Acids Res., 14:11-15.
Germino, J. and Bastia, D., 1982, Primary structure of the replication initiation protein of plasmid R6K, Proc.Natl.Acad.Sci. USA, 79:5475-5479.
Herman, R. C., 1986, Internal initiation of translation on the vesicular stomatitis virus phosphoprotein mRNA yields a second protein, J.Virol., 58:797-804.
Jacks, J. and Varmus, H. E., 1985, Expression of the rous sarcoma virus pol gene by ribosomal frameshifting, Science, 230:1237-1242.
Kamer, G. and Argos, P., 1984, Primary structure comparison of RNA-dependent polymerases from plant, animal and bacterial viruses, Nucl. Acids Res., 12:7269-7282.
Kanehisa, M. I., 1982, Los Alamos sequence analysis package for nucleic acids and proteins, Nucl. Acids Res., 10:183-196.
Kastelein, R. A., Remaut, E., Fiers, W. and van Duin, J., 1982, Lysis gene expression of RNA phage MS2 depends on a frameshift during translation of the overlapping coat protein gene, Nature, 295:35-41.
Korneluk, R. G., Quan, F. and Gravel, R. A., 1985, Rapid and reliable dideoxy sequencing of double-stranded DNA, Gene, 40:317-323.
Kozak, M., 1983, Comparison of initiation of protein synthesis in procaryotes, eucaryotes and organelles, Microbiol.Revs., 47:1-45.
Kyte, J. and Doolittle, R. F., 1982, A simple method for displaying the hydropathic character of a protein, J.Mol.Biol., 157:105-132.
Lipman, D. J. and Pearson, W. R., 1985, Rapid and sensitive protein similarity searches, Science, 227:1435-1141.
Lomniczi, B., 1977, Biological properties of avian coronavirus RNA, J.Gen.Virol.., 36:531-533.
Lomniczi, B. and Kennedy, I., 1977, Genome of infectious bronchitis virus, J.Virol., 24:99-107.
Maxam, A. M. and Gilbert, W., 1980, Sequencing end-labelled DNA with base-specific chemical cleavages, in "Methods in Enzymology", L. Grossman and K. Moldave, Eds., Vol. 65, Part 1, Academic Press, New York, pp. 499-560.
Mizusawa, S., Nishimura, S. and Seela, F., 1986, Improvement of the dideoxy chain termination method of DNA sequencing by use of deoxy-7-deazaguanosine triphosphate in place of dGTP, Nucl. Acids Res., 14:1319-1324.
Schochetman, G., Stevens, R. H. and Simpson, R. W., 1977, Presence of infectious polyadenylated RNA in the coronavirus avian bronchitis virus, Virol., 77:772-782.
Schubert, M., Harmison, G. G. and Meier, E., 1984, Primary structure of the vesicular stomatitis virus polymerase (L) gene: evidence for a high frequency of mutations, J.Virol., 51:, 505-514.
Siddell, S. G., Anderson, R., Cavanagh, D., Fujiwara, K., Klenk, H. D., MacNaughton, M. R., Pensaert, M., Stohlman, S. A., Sturman, L. and van der Zeijst, B. A. M., 1983, Coronaviridae, Intervirology, 20:181-189.
Southern, E. M., 1975, Detection of specific sequences among DNA fragments separated by gel electrophoresis, J.Mol.Biol., 98:503-517.
Staden, R., 1982, An interactive graphics program for comparing and aligning nucleic acid and amino acid sequences, Nucl. Acids Res., 10:2951-2961.
Staden, R., 1984, A computer program to enter DNA gel reading data

into a computer, Nucl. Acids Res., 12:499-503.
Steinhauer, D. A. and Holland, J. J., 1986, Direct method for quantitation of extreme polymerase error frequencies at selected single base sites in viral RNA, J.Virol. 57:219-228.
Stern, D. F. and Kennedy, S. I. T., 1980a, Coronavirus multiplication strategy. I. Identification and characterisation of virus-specified RNA, J.Virol., 34:665-674.
Stern, D. F. and Kennedy, S. I. T., 1980b, Coronavirus multiplication strategy. II. Mapping the avian infectious bronchitis virus intracellular RNA species to the genome, J.Virol., 36:440-449.
Stern, D. F. and Sefton, B. M., 1984, Coronavirus multiplication: the locations of genes for the virion proteins on the avian infectious bronchitis virus genome, J.Virol., 50:22-29.
Strauss, E. G. and Strauss, J. H., 1983, Replication strategies of the single stranded RNA viruses of eukaryotes, Curr. Topics in Microbiol. and Immunol., 105:1-98.