J. gen. Virol. (1987), **68**, 57–77. *Printed in Great Britain*

Key words: *IBV/coronavirus/nucleotide sequence*

Completion of the Sequence of the Genome of the Coronavirus Avian Infectious Bronchitis Virus

By 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, U.K.

(*Accepted 19 September 1986*)

SUMMARY

The nucleotide sequence determination of the genome of the Beaudette strain of the coronavirus avian infectious bronchitis virus (IBV) has been completed. The complete sequence has been obtained from 17 overlapping cDNA clones, the 5′-most of which contains the leader sequence (as determined by direct sequencing of the genome) and the 3′-most of which contains the poly(A) tail. Approximately 8 kilobases at the 3′ end of this sequence have already been published. These contain the sequences of mRNAs A to E within which are the genes for the spike, the membrane and the nucleocapsid polypeptides: the main structural components of the virion. The remainder of the sequence, equivalent to the 'unique' region of mRNA F, is some 20 kilobases in length and is thought to code for a polymerase or polymerases which are involved in the replication of the genome and the production of the subgenomic messenger RNAs. This sequence contains two large open reading frames, potentially coding for polypeptides of molecular weights 441000 and 300000. Unlike other large open reading frames in the virus, the 300000 open reading frame appears to have no subgenomic RNA associated with it which would allow it to be at the 5′ end of an mRNA species. Because of this, and because of the characteristics of the sequence in the region immediately upstream of its start codon, other mechanisms of translation, such as ribosome slippage, must be postulated.

INTRODUCTION

Avian infectious bronchitis virus (IBV) is the type species of the family Coronaviridae (Siddell *et al.*, 1983*a*). Coronaviruses are enveloped, pleomorphic particles with a distinctive 'corona' of club-shaped surface projections, and a large single-stranded RNA genome of positive polarity (Siddell *et al.*, 1983*b*). In infected cells, in addition to genome-sized RNA, a number of subgenomic RNAs can be detected which have a common 3′ terminus, but extend for different lengths in the 5′ direction, forming a nested set (Stern & Kennedy, 1980*a*, *b*; Leibowitz *et al.*, 1981). In the case of IBV these are designated mRNAs A to F, mRNA A being the smallest and mRNA F being of genome length. *In vitro* translation studies have demonstrated that mRNAs A, C and E code for the nucleocapsid polypeptide, the membrane polypeptide and the precursor polypeptide to the spike or surface projection respectively (Stern & Sefton, 1984). These three polypeptides form the three known structural proteins of coronavirus virions (Cavanagh, 1981). Sequencing of cDNA clones derived from IBV genomic RNA has shown that, in the case of mRNAs A, C and E, only the 5′ region of each mRNA which is not present in the next smallest mRNA is translated (Boursnell *et al.*, 1985*a*, 1984; Binns *et al.*, 1985*b*). This region is often referred to, for convenience, as the 'unique' region of the particular mRNA. For mRNAs B and D the situation is more complicated in that each mRNA has more than one open reading frame (ORF) and also has ORFs overlapping the next smallest mRNA (Boursnell & Brown, 1984; Boursnell *et al.*, 1985*b*).

The genome of IBV is infectious (Lomniczi, 1977) indicating that it has a messenger function. There is also no evidence for a virion-associated RNA polymerase (Schochetman *et al.*, 1977).

On entry into the cell therefore the virion RNA probably codes for a polymerase, the gene for which must lie in the large 5′ region of the genome, the 'unique' region of mRNA F, which does not contain the genes for the structural polypeptides. This polymerase would then be used to synthesize a negative-stranded template. The negative strand could then be used by another polymerase, or a modified form of the same polymerase, to produce the subgenomic mRNAs and virion RNA. Both the negative strand and two distinct polymerase activities have been detected in cells infected with the coronavirus mouse hepatitis virus (MHV) (Lai *et al.*, 1982; Brayton *et al.*, 1982). Translation of MHV virion RNA in reticulocyte lysates produced three structurally related polypeptides of molecular weights greater than 200000 (200K) (Leibowitz *et al.*, 1982).

In this paper we present the nucleotide sequence, obtained from cDNA clones, of the 'unique' region of mRNA F, the genome-sized mRNA. The sequence of approximately 8 kilobases from the 3′ end of the genome, containing the genes for the major structural polypeptides, has already been published (Boursnell & Brown, 1984; Boursnell *et al.*, 1984, 1985*a*, *b*; Binns *et al.*, 1985*b*). The 20500 bases of sequence reported here complete the sequence of the IBV genome, which is, as far as we are aware, the first complete sequence of a coronavirus and the largest RNA virus sequenced to date.

METHODS

cDNA cloning. Seventeen cDNA clones covering the 3′-most 27569 kb of the genome have been obtained. These are shown in Fig. 1. They have been derived from RNA isolated from gradient-purified virus of the Beaudette strain (Beaudette & Hudson, 1937; Brown & Boursnell, 1984). cDNA has been obtained by three methods: oligo(dT) priming (Brown & Boursnell, 1984), priming with specific oligonucleotides (Boursnell *et al.*, 1984) and random priming with calf thymus DNA oligonucleotides (Binns *et al.*, 1985*a*). The Southern blotting technique was used to identify overlapping clones (Southern, 1975). Specific cDNA clones were identified using 'prime-cut' probes. These are made by synthesizing labelled DNA from selected M13 clones using the normal sequencing primer, cutting with a restriction enzyme, and eluting the labelled, single-stranded probe from denaturing acrylamide gels (Biggin *et al.*, 1984).

Subcloning for M13 sequencing. Random subclones of each cDNA clone were generated by sonication (Deininger, 1983) and subcloning into *Sma*I-cut, phosphatase-treated M13mp10 (Amersham). Bacterial colonies containing M13 with inserts were grown, transferred to nitrocellulose filters, and probed with nick-translated purified viral insert DNA from the cDNA clone. Single-stranded templates were prepared from M13 clones identified as viral in this way.

DNA sequencing. Sequencing was carried out by the dideoxy method (Sanger *et al.*, 1977; Bankier & Barrell, 1983). [α-^{35}S]dATP was used in the sequencing reactions and the products were analysed on buffer gradient gels (Biggin *et al.*, 1983). Additional sequencing information was obtained by reverse sequencing (Hong, 1981). For regions containing compressions due to DNA secondary structure, sequencing samples were run on hot (80 °C) gels or gels containing 42% formamide. For some regions cytosine residues were modified by the method of Ambartsumyan & Mazo (1980) prior to separating on gels, to reduce GC base pairing. Deoxyinosine triphosphate (Bankier & Barrell, 1983) and deoxy-7-deazaguanosine triphosphate (Mizusawa *et al.*, 1986) were used in place of deoxyguanosine triphosphate in some cases, again to reduce GC base pairing. For sequencing directly from the viral RNA the method used was essentially as described by Caton *et al.* (1982).

Computer analysis of the sequence data. Sequence data were read directly into a BBC microcomputer using a sonic digitizer (Graf/Bar, Science Accessories Corporation) and data were analysed on a VAX 11/750 using the programs of Staden (1982*a*, *b*, 1984*a*, *b*). Comparisons with the National Biomedical Research Foundation (NBRF) protein identification resource was made using the programs SEARCH and FASTP (George *et al.*, 1986; Lipman & Pearson, 1985) and SEQHP (Kanehisa, 1982).

RESULTS

Selection of cDNA clones

The majority of the cDNA clones which have been used to obtain the sequence of the 'unique' region of mRNA F were produced by a random priming method (Binns *et al.*, 1985*a*). Clone 182 was produced by priming with a specific oligonucleotide from existing sequence at the 5′ end of mRNA D. Clone 227 was identified as coming from the 5′ end of the genome by probing a random library with leader-specific probes. The randomly primed clones 217, 216, 204, 210, 205, 220 and 249 were mapped by identifying overlaps using Southern blotting. The nine clones were

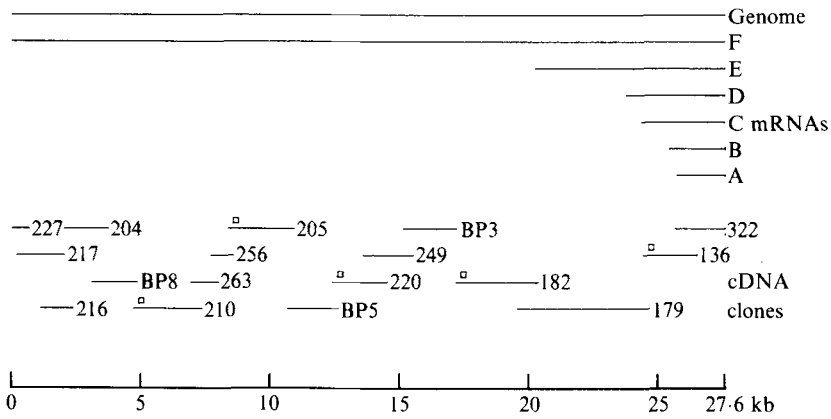


Fig. 1. Diagram showing the positions of all the cDNA clones used in obtaining the nucleotide sequence. The squares at the end of some of the clones show the positions of oligonucleotide primers used to prime synthesis of cDNA for adjacent clones. Above the clones are shown mRNAs A to F.

not contiguous but formed four blocks. cDNA clones in the region of the three remaining gaps were obtained using specific oligonucleotide primers. Clones spanning the gaps were identified using either 'prime-cut' probes (Biggin *et al.*, 1984) made from M13 subclones of cDNA clones on either side of the gap or by using Southern blotting. Five clones, 256, 263, BP3, BP5 and BP8 were identified in this way and the overlaps confirmed by sequencing. Fig. 1 shows the positions of all the cDNA clones used in obtaining the complete sequence of the virus, and the positions of the oligonucleotide primers.

DNA sequencing

Fourteen cDNA clones have been sequenced to obtain the complete sequence of the 'unique' region of mRNA F, the genome-sized messenger RNA. The 20500 bases of sequence presented here stretch from the 5′ end of the genome to an arbitrary position 190 bases 3′-wards of the end of the body of mRNA E. The 39 nucleotides at the very 5′ end of the genome have not been obtained in cDNA clones from the Beaudette strain, and the sequence here is derived from Maxam & Gilbert (1980) sequencing of primer-extended products from Beaudette virion RNA (Brown *et al.*, 1986). Fig. 2 shows the DNA sequence obained from the cDNA clones, with a translation in single-letter amino acid code of the main ORFs.

Sequence analysis

Fig. 3 shows the positions of ORFs in this region. Most of the sequence encodes two very large ORFs which could code for polypeptides of predicted molecular weights 441K and 300K. These two large ORFs have been designated F1 and F2.

The first large ORF, F1, is not the first ORF to occur after the homology region. At position 131 there is an AUG codon followed by a small ORF which could code for a polypeptide of 11 amino acids. This AUG is the first initiation codon to occur on the genome. The second initiation codon is at the start of F1. Both the large ORFs have a codon usage (Staden & McLachlan, 1982) very similar to that of the genes for the structural polypeptides S, M and N. The small ORF also appears to have the same codon usage, insofar as that is significant for such a short sequence. After the end of the small ORF the reading frame is open, in the other two possible frames, for a further 232 or 73 bases but the codon usage of the predicted amino acids for these sections of ORF is not similar to that previously found for IBV. The sequence context around the first AUG codon is not similar to that used by most eukaryotic mRNAs (Kozak, 1983) in that it has a pyrimidine at position −3. The context around the second AUG on the other hand has a purine at −3, in addition to a C at positions −1 and −4, both of which mean that it conforms well to the consensus for functional initiation codons.

```
   1 ACTTAAGATAGATATTAATATATATCTATTACACTAGCCTTGCGCTAGATTTTTAACTTAACAAAACGGACTTAAATACCTACAGCTGGTCCTCATAGGT 100

                                 M  A  P  G  H  L  S  G  F  C  Y  *
 101 GTTCCATTGCAGTGCACTTTAGTGCCCTGGATGGCACCTGGCCACCTGTCAGGTTTTTGTTATTAAAATCTTATTGTTGCTGGTATCACTGCTTGTTTTG 200

 201 CCGTGTCTCACTTTATACATCTGTTGCTTGGGCTACCTAGTGTCCAGCGTCCTACGGGCGTCGTGGCTGGTTCGAGTGCGAGGAACCTCTGGTTCATCTA 300

 301 GCGGTAGGCGGGTGTGTGGAAGTAGCACTTCAGACGTACCGGTTCTGTTGTGTGAAATACGGGGTCACCTCCCCCCACATACCTCTAAGGGCTTTTGAGC 400

 401 CTAGCGTTGGGCTACGTTCTCGCATAAGGTCGGCTATACGACGTTTGTAGGGGGTAGTGCCAAACAACCCCTGAGGTGACAGGTTCTGGTGGTGTTTAGT 500

                                 M  A  S  S  L  K  Q  G  V  S  P  K  P  R  D  V  I  L  V  S  K  D  I  P
 501 GAGCAGACATACAATAGACAGTGACAACATGGCTTCAAGCCTAAAACAGGGAGTATCTCCCAAACCACGGGATGTCATTCTTGTGTCCAAAGACATCCCT 600

     E  Q  L  C  D  A  L  F  F  Y  T  S  H  N  P  K  D  Y  A  D  A  F  A  V  R  Q  K  F  D  R  S  L  Q  T
 601 GAACAACTTTGTGACGCTTTGTTTTTCTATACGTCACATAACCCTAAGGATTACGCTGATGCTTTTGCAGTTAGGCAGAAGTTTGACCGTAGTCTCCAGA 700

     G  K  Q  F  K  F  E  T  V  C  G  L  F  L  L  K  G  V  D  K  I  T  P  G  V  P  A  K  V  L  K  A  T
 701 CTGGGAAACAGTTCAAATTTGAAACTGTGTGTGGTCTCTTCCTCTTGAAGGGAGTTGACAAAATAACACCTGGCGTCCCAGCAAAAGTTTTAAAAGCCAC 800

     S  K  L  A  D  L  E  D  I  F  G  V  S  P  L  A  R  K  Y  R  E  L  L  K  T  A  C  Q  W  S  L  T  V
 801 TTCTAAGTTGGCAGATTTAGAAGACATCTTTGGTGTCTCTCCTTTAGCGCGGAAGTACCGTGAATTGTTGAAAACAGCGTGTCAGTGGTCTCTTACTGTA 900

     E  A  L  D  V  R  A  Q  T  L  D  E  I  F  D  P  T  E  I  L  W  L  Q  V  A  A  K  I  H  V  S  S  M  A
 901 GAAGCACTGGATGTTCGTGCACAAACTCTCGATGAAATTTTTGACCCCACTGAAATACTTTGGCTTCAGGTGGCTGCAAAAATTCATGTTTCATCTATGG 1000

     M  R  R  L  V  G  E  V  T  A  K  V  M  D  A  L  G  S  N  L  S  A  L  F  Q  I  V  K  Q  Q  I  A  R
1001 CAATGCGCAGGCTTGTTGGAGAAGTAACTGCAAAAGTCATGGATGCTCTGGGCTCAAACTTGAGTGCTCTTTTTCAAATTGTTAAACAACAAATAGCCAG 1100

     I  F  Q  K  A  L  A  I  F  E  N  V  N  E  L  P  Q  R  I  A  A  L  K  M  A  F  A  K  C  A  R  S  I
1101 AATCTTTCAAAAGGCACTGGCTATTTTTGAGAATGTGAATGAATTACCACAGCGTATTGCAGCACTTAAGATGGCTTTTGCCAAGTGTGCTAGGTCAATT 1200

     T  V  V  V  V  E  R  T  L  V  V  K  E  F  A  G  T  C  L  A  S  I  N  G  A  V  A  K  F  F  E  E  L  P
1201 ACTGTTGTGGTTGTTGAAAGAACTCTAGTTGTTAAAGAGTTCGCAGGAACTTGTCTTGCAAGCATTAATGGTGCTGTCGCAAAATTCTTTGAAGAGTTGC 1300

     N  G  F  M  G  S  K  I  F  T  T  L  A  F  F  K  E  A  A  V  R  V  V  E  N  I  P  N  A  P  R  G  T
1301 CAAACGGCTTCATGGGTTCTAAGATTTTCACAACACTTGCCTTCTTTAAAGAGGCAGCTGTGAGAGTTGTGGAGAACATACCAAATGCACCGAGAGGTAC 1400

     K  G  F  E  V  V  G  N  A  K  G  T  Q  V  V  V  R  G  M  R  N  D  L  T  L  L  D  Q  K  A  D  I  P
1401 TAAGGGATTTGAAGTTGTTGGCAATGCCAAAGGCACACAGGTAGTTGTGCGCGGCATGCGAAATGACTTAACATTGCTTGACCAAAAAGCTGATATTCCT 1500

     V  E  P  E  G  W  S  A  I  L  D  G  H  L  C  Y  V  F  R  S  G  D  R  F  Y  A  A  P  L  S  G  N  F  A
1501 GTTGAACCAGAAGGTTGGTCTGCAATTTTGGATGGACATCTTTGCTATGTCTTTAGGAGTGGTGATCGCTTTTATGCTGCACCTCTTTCAGGAAATTTTG 1600

     L  S  D  V  H  C  C  E  R  V  V  C  L  S  D  G  V  T  P  E  I  N  D  G  L  I  L  A  A  I  Y  S  S
1601 CTTTGAGTGATGTTCATTGCTGTGAGCGTGTAGTCTGTCTATCTGATGGTGTAACACCGGAGATAAATGATGGACTCATTCTAGCTGCAATCTACTCTTC 1700

     F  S  V  S  E  L  V  T  A  L  K  K  G  E  P  F  K  F  L  G  H  K  F  V  Y  A  K  D  A  A  V  S  F
1701 TTTTAGTGTCTCTGAGCTTGTAACAGCTCTTAAAAAGGGTGAACCATTCAAGTTCTTGGGCCATAAATTCGTGTATGCGAAGGATGCAGCAGTGTCTTTT 1800

     T  L  A  K  A  A  T  I  A  D  V  L  R  L  F  Q  S  A  R  V  I  A  E  D  V  W  S  S  F  T  E  K  S  F
1801 ACTTTAGCGAAGGCTGCCACTATTGCAGATGTCTTGAGGCTGTTTCAATCAGCTCGTGTGATAGCAGAAGATGTTTGGTCTTCATTTACTGAAAAGTCTT 1900

     E  F  W  K  L  A  Y  G  K  V  R  N  L  E  E  F  V  K  T  Y  V  C  K  A  Q  M  S  I  V  I  L  A  A
1901 TTGAATTCTGGAAGCTTGCATATGGAAAAGTGCGCAACCTTGAAGAATTTGTGAAGACCTATGTTTGTAAGGCTCAAATGTCGATTGTGATTCTAGCAGC 2000

     V  L  G  E  D  I  W  H  L  V  S  Q  V  I  Y  K  L  G  V  L  F  T  K  V  V  D  F  C  D  K  H  W  K
2001 AGTGCTTGGAGAGGACATTTGGCATCTTGTCTCACAAGTCATCTATAAATTAGGTGTTCTTTTTACTAAAGTCGTTGACTTTTGTGACAAACACTGGAAA 2100

     G  F  C  V  Q  L  K  R  A  K  L  I  V  T  E  T  F  C  V  L  K  G  V  A  Q  H  C  F  Q  L  L  L  D  A
2101 GGTTTTTGTGTACAGTTGAAAAGAGCTAAGCTCATTGTCACCGAAACCTTCTGTGTTTTAAAAGGAGTTGCACAGCATTGTTTTCAACTGCTGCTAGATG 2200

     I  H  S  L  Y  K  S  F  K  K  C  A  L  G  R  I  H  G  D  L  L  F  W  K  G  G  V  H  K  I  V  Q  D
2201 CAATACACTCTTTGTACAAGAGTTTTAAGAAGTGTGCACTTGGTAGAATCCATGGAGATTTGCTCTTCTGGAAAGGAGGTGTGCATAAAATTGTTCAAGA 2300

     G  D  E  I  W  F  D  A  I  D  S  V  D  V  E  D  L  G  V  V  Q  E  K  S  I  D  F  E  V  C  D  D  V
2301 TGGCGATGAAATATGGTTTGACGCCATTGATAGTGTTGATGTTGAAGATCTGGGTGTTGTTCAGGAAAAATCGATTGATTTTGAGGTTTGCGATGACGTG 2400

     T  L  P  E  N  Q  P  G  H  M  V  Q  I  E  D  D  G  K  N  Y  M  F  F  R  F  K  K  D  E  N  I  Y  Y  T
2401 ACACTTCCAGAAAACCAACCTGGTCATATGGTTCAAATAGAGGATGATGGTAAGAACTACATGTTCTTCCGTTTTAAAAAGGATGAGAACATTTATTATA 2500
```

```
      P  M  S  Q  L  G  A  I  N  V  V  C  K  A  G  G  K  T  V  T  F  G  E  T  T  V  Q  E  I  P  P  P  D
2501  CACCAATGTCTCAACTTGGTGCTATTAATGTGGTTTGCAAAGCAGGCGGTAAGACTGTCACCTTTGGAGAAACTACAGTACAAGAGATACCACCACCTGA  2600

      V  V  P  I  K  V  S  I  E  C  C  G  E  P  W  N  T  I  F  K  K  A  Y  K  E  P  I  E  V  D  T  D  L
2601  TGTCGTGCCTATTAAGGTTAGCATAGAATGTTGTGGTGAACCATGGAATACGATCTTCAAGAAGGCTTATAAAGAGCCTATAGAAGTAGATACAGACCTC  2700

      T  V  E  Q  L  L  S  V  I  Y  E  K  M  C  D  D  L  K  L  F  P  E  A  P  E  P  P  P  F  E  N  V  A  L
2701  ACAGTAGAACAATTGCTCTCTGTGATCTATGAGAAAATGTGTGACGACCTTAAATTGTTTCCAGAGGCACCAGAGCCTCCACCATTTGAGAATGTCGCAC  2800

      V  D  K  N  G  K  D  L  D  C  I  K  S  C  H  L  I  Y  R  D  Y  E  S  D  D  D  I  E  E  E  D  A  E
2801  TTGTTGATAAGAACGGTAAAGATTTGGATTGTATAAAATCTTGCCATTTGATCTATCGTGACTATGAGAGCGATGATGACATCGAGGAGGAAGATGCTGA  2900

      E  C  D  T  D  S  G  E  A  E  E  C  D  T  N  S  E  C  E  E  E  D  E  D  T  K  V  L  A  L  I  Q  D
2901  GGAGTGTGACACAGACTCAGGTGAAGCTGAGGAGTGTGACACTAATTCAGAATGTGAAGAAGAGGATGAGGATACTAAAGTGTTGGCTCTTATACAAGAC  3000

      P  A  S  I  K  Y  P  L  P  L  D  E  D  Y  S  V  Y  N  G  C  I  V  H  K  D  A  L  D  V  V  N  L  P  S
3001  CCGGCAAGTATTAAATACCCTCTGCCTCTTGATGAAGATTATAGCGTCTATAATGGATGTATTGTACACAAGGACGCTCTTGATGTTGTGAATTTACCAT  3100

      G  E  E  T  F  V  V  N  N  C  F  E  G  A  V  K  P  L  P  Q  K  V  V  D  V  L  G  D  W  G  E  A  V
3101  CTGGTGAAGAAACTTTTGTTGTCAATAACTGTTTTGAGGGAGCTGTTAAACCACTTCCACAGAAGGTAGTTGATGTTCTTGGTGACTGGGGAGAGGCTGT  3200

      D  A  Q  E  Q  L  C  Q  Q  E  P  L  Q  H  T  F  E  E  P  V  E  N  S  T  G  S  S  K  T  M  T  E  Q
3201  TGATGCGCAAGAACAACTGTGTCAACAAGAGCCTCTGCAACATACCTTTGAAGAACCAGTCGAAAATTCTACTGGTAGTTCTAAGACAATGACTGAACAA  3300

      V  V  V  E  D  Q  E  L  P  V  V  E  Q  D  Q  D  V  V  V  Y  T  P  T  D  L  E  V  A  K  E  T  A  E  E
3301  GTCGTTGTAGAAGATCAAGAACTACCTGTTGTTGAACAAGATCAGGATGTAGTTGTTTATACACCTACAGATCTTGAAGTTGCAAAAGAAACAGCAGAAG  3400

      V  D  E  F  I  L  I  F  A  V  P  K  E  E  V  V  S  Q  K  D  G  A  Q  I  K  Q  E  P  I  Q  V  V  K
3401  AGGTTGATGAGTTTATTCTCATTTTTGCTGTTCCTAAAGAAGAAGTTGTGTCCCAGAAAGATGGGGCACAGATTAAACAAGAGCCTATTCAAGTTGTTAA  3500

      P  Q  R  E  K  K  A  K  K  F  K  V  K  P  A  T  C  E  K  P  K  F  L  E  Y  K  T  C  V  G  D  L  T
3501  ACCACAACGTGAGAAGAAGGCTAAAAAGTTCAAAGTTAAACCAGCCACATGTGAGAAACCTAAATTTTTGGAGTATAAAACATGTGTGGGTGATTTGACT  3600

      V  V  I  A  K  A  L  D  E  F  K  E  F  C  I  V  N  A  A  N  E  H  M  T  H  G  S  G  V  A  K  A  I  A
3601  GTTGTAATTGCCAAAGCATTGGATGAGTTTAAAGAGTTCTGCATTGTAAATGCTGCAAATGAGCATATGACTCATGGTAGTGGCGTTGCAAAGGCAATTG  3700

      D  F  C  G  L  D  F  V  E  Y  C  E  D  Y  V  K  K  H  G  P  Q  Q  R  L  V  T  P  S  F  V  K  G  I
3701  CAGACTTTTGTGGACTGGATTTTGTTGAATATTGTGAGGACTATGTTAAGAAACATGGGCCACAACAGAGACTTGTTACACCTTCGTTTGTCAAAGGCAT  3800

      Q  C  V  N  N  V  V  G  P  R  H  G  D  N  N  L  H  E  K  L  V  A  A  Y  K  N  V  L  V  D  G  V  V
3801  TCAATGTGTGAATAATGTTGTAGGACCCCGCCATGGAGACAACAACTTGCATGAGAAGCTTGTTGCTGCCTACAAGAATGTGCTTGTAGATGGCGTAGTC  3900

      N  Y  V  V  P  V  L  S  L  G  I  F  G  V  D  F  K  M  S  I  D  A  M  R  E  A  F  E  G  C  T  I  R  V
3901  AATTATGTTGTGCCAGTTCTTTCATTAGGAATTTTTGGTGTAGATTTTAAAATGTCAATAGACGCAATGCGTGAAGCTTTTGAAGGTTGCACCATACGCG  4000

      L  L  F  S  L  S  Q  E  H  I  D  Y  F  D  V  T  C  K  Q  K  T  I  Y  L  T  E  D  G  V  K  Y  R  S
4001  TTCTTTTGTTTTCTCTGAGCCAAGAACACATCGATTATTTCGATGTAACTTGCAAACAGAAGACAATTTATCTTACGGAGGATGGTGTTAAATACCGCTC  4100

      I  V  L  K  P  G  D  S  L  G  Q  F  G  Q  V  Y  A  K  N  K  I  V  F  T  A  D  D  V  E  D  K  E  I
4101  CATTGTTCTAAAACCTGGTGACTCATTGGGTCAATTTGGACAGGTTTATGCTAAAAACAAGATAGTTTTTACAGCCGATGATGTTGAGGACAAAGAAATT  4200

      L  Y  V  P  T  T  D  K  S  I  L  E  Y  Y  G  L  D  A  Q  K  Y  V  I  Y  L  Q  T  L  A  Q  K  W  N  V
4201  CTCTACGTCCCCACGACTGATAAAAGCATTCTTGAATACTATGGTTTAGATGCGCAAAAGTATGTAATATATTTGCAAACGCTTGCGCAGAAATGGAATG  4300

      Q  Y  R  D  N  F  L  I  L  E  W  R  D  G  N  C  W  I  S  S  A  I  V  L  L  Q  A  A  K  I  R  F  K
4301  TCCAATATAGGGACAATTTTCTTATACTAGAGTGGCGCGATGGAAATTGTTGGATTAGTTCAGCAATAGTTCTCCTTCAAGCTGCTAAAATTAGGTTTAA  4400

      G  F  L  T  E  A  W  A  K  L  L  G  G  D  P  T  D  F  V  A  W  C  Y  A  S  C  T  A  K  V  G  D  F
4401  AGGTTTTCTAACAGAAGCGTGGGCTAAACTGTTAGGTGGAGATCCTACAGACTTTGTTGCCTGGTGTTATGCAAGTTGTACTGCTAAAGTAGGTGATTTC  4500

      S  D  A  N  W  L  L  A  N  L  A  E  H  F  D  A  D  Y  T  N  A  F  L  K  K  R  V  S  C  N  C  G  I  K
4501  TCAGATGCTAATTGGCTTTTAGCGAATTTAGCAGAACATTTTGACGCAGATTACACAAATGCGTTTCTTAAGAAGCGCGTTTCGTGTAACTGTGGTATTA  4600

      S  Y  E  L  R  G  L  E  A  C  I  Q  P  V  R  A  T  N  L  L  H  F  K  T  Q  Y  S  N  C  P  T  C  G
4601  AGAGCTATGAGCTTAGAGGCCTTGAAGCTTGTATTCAGCCAGTTCGGGCAACTAATCTGCTACATTTTAAGACGCAATATTCAAATTGCCCAACCTGTGG  4700

      A  N  N  T  D  E  V  I  E  A  S  L  P  Y  L  L  L  F  A  T  D  G  P  A  T  V  D  C  D  E  D  A  V
4701  CGCAAATAATACGGATGAAGTAATAGAAGCTTCGTTACCGTACTTATTGCTTTTTGCTACTGATGGTCCTGCTACAGTTGATTGTGATGAAGATGCTGTG  4800

      G  T  V  V  F  V  G  S  T  N  S  G  H  C  Y  T  Q  A  A  G  Q  A  F  D  N  L  A  K  D  R  K  F  G  K
4801  GGGACTGTCGTGTTTGTTGGTTCTACTAATAGTGGCCATTGTTATACACAAGCTGCAGGGCAAGCTTTTGATAATCTTGCTAAAGATAGAAAATTTGGAA  4900
```

```
      K  S  P  Y  I  T  A  M  Y  T  R  F  A  F  K  N  E  T  S  L  P  V  A  K  Q  S  K  G  K  S  K  S  V
4901  AGAAGTCGCCTTACATTACTGCAATGTATACGCGATTCGCTTTTAAGAATGAAACCTCTTTGCCTGTTGCTAAACAGAGCAAGGGTAAGTCTAAGTCGGT  5000

      K  E  D  V  S  N  L  A  T  S  S  K  A  S  F  D  N  L  T  D  F  E  Q  W  Y  D  S  N  I  Y  E  S  L
5001  AAAGGAAGATGTTTCTAACCTTGCTACTAGTTCTAAGGCCAGTTTTGATAATCTTACTGACTTCGAACAGTGGTATGATAGTAACATCTATGAAAGTCTT  5100

      K  V  Q  E  S  P  D  N  F  D  K  Y  V  S  F  T  T  K  E  D  S  K  L  P  L  T  L  K  V  R  G  I  K  S
5101  AAAGTGCAGGAATCACCTGATAACTTTGATAAATATGTGTCATTCACAACAAAGGAAGATTCTAAGTTGCCATTGACACTTAAGGTTAGAGGTATTAAAT  5200

      V  V  D  F  R  S  K  D  G  F  I  Y  K  L  T  P  D  T  D  E  N  S  K  A  P  V  Y  Y  P  V  L  D  A
5201  CAGTTGTTGACTTTAGATCGAAGGATGGTTTTATTTATAAGTTAACACCTGATACTGATGAAAATTCAAAAGCACCAGTCTACTACCCAGTCTTGGACGC  5300

      I  S  L  K  A  I  W  V  E  G  N  A  N  F  V  V  G  H  P  N  Y  Y  S  K  S  L  H  I  P  T  F  W  E
5301  TATTAGTCTTAAGGCAATATGGGTGGAAGGTAATGCTAACTTTGTTGTTGGTCATCCAAATTATTATAGTAAGTCTCTTCATATTCCTACTTTTTGGGAA  5400

      N  A  E  N  F  V  K  M  G  D  K  I  G  G  V  T  M  G  L  W  R  A  E  H  L  N  K  P  N  L  E  R  I  F
5401  AATGCTGAGAATTTTGTTAAAATGGGTGATAAAATTGGTGGTGTAACTATGGGACTTTGGCGTGCAGAACACCTTAATAAACCTAATTTGGAGAGAATTT  5500

      N  I  A  K  K  A  I  V  G  S  S  V  V  T  T  Q  C  G  K  L  I  G  K  A  A  T  F  I  A  D  K  V  G
5501  TCAACATTGCTAAGAAAGCCATTGTTGGATCTAGTGTTGTTACTACACAATGCGGTAAATTAATAGGTAAAGCAGCTACATTCATTGCTGATAAAGTAGG  5600

      G  G  V  V  R  N  I  T  D  S  I  K  G  L  C  G  I  T  R  G  H  F  E  R  K  M  S  P  Q  F  L  K  T
5601  TGGTGGTGTAGTTCGCAATATTACAGATAGCATTAAGGGTCTTTGTGGAATTACACGAGGGCATTTTGAAAGAAAAATGTCTCCACAATTCCTAAAGACG  5700

      L  M  F  F  L  F  Y  F  L  K  A  S  V  K  S  V  V  A  S  Y  K  T  V  L  C  K  V  V  L  A  T  L  L  I
5701  CTTATGTTCTTTTTATTCTATTTCTTGAAGGCTAGTGTTAAGAGTGTTGTCGCTAGCTATAAGACCGTGTTATGTAAGGTGGTACTTGCTACTTTACTTA  5800

      V  W  F  V  Y  T  S  N  P  V  M  F  T  G  I  R  V  L  D  F  L  F  E  G  S  L  C  G  P  Y  K  D  Y
5801  TAGTTTGGTTTGTCTACACAAGTAACCCAGTAATGTTTACAGGAATACGTGTGTTAGATTTTCTATTCGAGGGTTCTTTGTGTGGTCCTTATAAAGACTA  5900

      G  K  D  S  F  D  V  L  R  Y  C  A  D  D  F  I  C  R  V  C  L  H  D  K  D  S  L  H  L  Y  K  H  A
5901  TGGTAAAGATTCTTTTGATGTGTTACGATATTGTGCAGATGATTTTATTTGTCGTGTGTGTTTACATGACAAAGATTCACTTCATTTGTACAAACACGCT  6000

      Y  S  V  E  Q  V  Y  K  D  A  A  S  G  F  I  F  N  W  N  W  L  Y  L  V  F  L  I  L  F  V  K  P  V  A
6001  TATAGTGTAGAGCAGGTCTATAAAGATGCAGCTTCTGGTTTTATTTTTAATTGGAATTGGCTTTATTTGGTCTTTCTAATATTATTTGTTAAACCAGTGG  6100

      G  F  V  I  I  C  Y  C  V  K  Y  L  V  L  N  S  T  V  L  Q  T  G  V  C  F  L  D  W  F  V  Q  T  V
6101  CAGGTTTTGTTATTATTTGCTATTGTGTTAAGTATTTGGTATTGAATTCAACTGTGCTGCAAACTGGTGTTTGTTTTTTAGATTGGTTTGTACAAACAGT  6200

      F  S  H  F  N  F  M  G  A  G  F  Y  F  W  L  F  Y  K  I  Y  I  Q  V  H  H  I  L  Y  C  K  D  V  T
6201  TTTTAGTCACTTTAATTTTATGGGAGCAGGGTTTTATTTCTGGCTCTTTTACAAGATATATATACAGGTGCATCATATACTGTATTGTAAGGATGTAACA  6300

      C  E  V  C  K  R  V  A  R  S  N  R  Q  E  V  S  V  V  V  G  G  R  K  Q  I  V  H  V  Y  T  N  S  G  Y
6301  TGTGAAGTGTGCAAAAGGGTTGCACGCAGCAACAGGCAAGAGGTTAGCGTGGTTGTTGGTGGACGCAAGCAGATAGTGCATGTTTACACTAACTCTGGCT  6400

      N  F  C  K  R  H  N  W  Y  C  R  N  C  D  D  Y  G  H  Q  N  T  F  M  S  P  E  V  A  G  E  L  S  E
6401  ATAACTTTTGTAAGAGACATAATTGGTATTGTAGAAATTGTGATGATTATGGTCACCAAAATACATTTATGTCTCCTGAAGTTGCTGGCGAGCTCTCTGA  6500

      K  L  K  R  H  V  K  P  T  A  Y  A  Y  H  V  V  D  E  A  C  L  V  D  D  F  V  N  L  K  Y  K  A  A
6501  AAAGCTTAAGCGCCATGTTAAACCTACAGCATACGCTTACCACGTTGTGGATGAGGCATGCTTAGTTGATGATTTTGTCAATTTAAAATATAAAGCTGCA  6600

      T  P  G  K  D  S  A  S  S  A  V  K  C  F  S  V  T  D  F  L  K  K  A  V  F  L  K  E  A  L  K  C  E  Q
6601  ACTCCTGGTAAGGATAGTGCATCTTCAGCTGTTAAGTGTTTCAGTGTTACAGATTTCTTGAAGAAAGCTGTTTTTCTTAAGGAAGCACTGAAATGTGAAC  6700

      I  S  N  D  G  F  I  V  C  N  T  Q  S  A  H  A  L  E  E  A  K  N  A  A  I  Y  Y  A  Q  Y  L  C  K
6701  AAATATCTAATGATGGTTTTATAGTGTGTAATACACAGAGTGCTCATGCATTAGAGGAAGCAAAGAATGCAGCCATCTATTATGCGCAATATCTGTGTAA  6800

      P  I  L  I  L  D  Q  A  L  Y  E  Q  L  V  V  E  P  V  S  K  S  V  I  D  K  V  C  S  I  L  S  S  I
6801  GCCAATACTTATACTTGACCAGGCACTTTATGAGCAATTAGTAGTAGAGCCTGTGTCTAAGAGTGTTATAGATAAAGTGTGTAGCATTTTGTCTAGTATA  6900

      I  S  V  D  T  A  A  L  N  Y  K  A  G  T  L  R  D  A  L  L  S  I  T  K  D  E  E  A  V  D  M  A  I  F
6901  ATATCTGTAGATACTGCAGCTTTAAATTATAAGGCAGGCACACTTCGTGATGCTCTGCTTTCTATTACTAAAGACGAAGAGGCCGTAGATATGGCTATAT  7000

      C  H  N  H  D  V  D  Y  T  G  D  G  F  T  N  V  I  P  S  Y  G  I  D  T  G  K  L  T  P  R  D  R  G
7001  TCTGTCATAATCATGATGTGGATTACACTGGTGATGGTTTTACTAATGTGATACCGTCATATGGTATAGACACTGGCAAGTTAACACCTCGTGATAGAGG  7100

      F  L  I  N  A  D  A  S  I  A  N  L  R  V  K  N  A  P  P  V  V  W  K  F  S  E  L  I  K  L  S  D  S
7101  GTTTTTGATAAATGCAGATGCTTCTATTGCTAACTTAAGAGTTAAAAATGCTCCGCCGGTAGTATGGAAGTTTTCTGAGCTTATTAAGTTGTCTGACAGT  7200

      C  L  K  Y  L  I  S  A  T  V  K  S  G  V  R  F  F  I  T  K  S  G  A  K  Q  V  I  A  C  H  T  Q  K  L
7201  TGTCTTAAATATTTAATTTCGGCTACTGTTAAGTCAGGTGTTCGTTTCTTTATAACAAAGTCTGGTGCTAAACAAGTTATTGCTTGTCATACACAGAAGT  7300
```

```
      L  V  E  K  K  A  G  G  I  V  S  G  T  F  K  C  F  K  S  Y  F  K  W  L  L  I  F  Y  I  L  F  T  A
7301  TGTTAGTAGAGAAAAAGGCAGGTGGTATTGTTAGCGGCACCTTTAAGTGTTTTAAGAGTTATTTTAAATGGCTCTTGATCTTTTACATACTTTTTACAGC  7400

      C  C  S  G  Y  Y  Y  M  E  V  S  K  S  F  V  H  P  M  Y  D  V  N  S  T  L  H  V  E  G  F  K  V  I
7401  ATGTTGTTCGGGTTATTACTATATGGAGGTGAGTAAAAGTTTTGTTCACCCCATGTATGATGTAAACTCCACACTGCATGTTGAAGGTTTTAAAGTTATA  7500

      D  K  G  V  L  R  E  I  V  P  E  D  T  C  F  S  N  K  F  V  N  F  D  A  F  W  G  R  P  Y  D  N  S  R
7501  GATAAAGGTGTTCTTAGGGAAATTGTACCAGAAGATACATGTTTCTCTAATAAATTTGTTAATTTTGATGCTTTTTGGGGCAGACCATATGATAATAGTA  7600

      N  C  P  I  V  T  A  V  I  D  G  D  G  T  V  A  T  G  V  P  G  F  V  S  W  V  M  D  G  V  M  F  I
7601  GAAACTGTCCAATTGTCACAGCTGTTATAGATGGTGATGGGACAGTAGCTACAGGTGTTCCTGGTTTTGTGTCCTGGGTTATGGATGGTGTTATGTTTAT  7700

      H  M  T  Q  T  E  R  K  P  W  Y  I  P  T  W  F  N  R  E  I  V  G  Y  T  Q  D  S  I  I  T  E  G  S
7701  ACATATGACACAGACTGAGAGAAAACCGTGGTACATTCCTACTTGGTTTAATAGAGAAATTGTCGGTTACACTCAGGATTCAATTATTACTGAGGGTAGT  7800

      F  Y  T  S  I  A  L  F  S  A  R  C  L  Y  L  T  A  S  N  T  P  Q  L  Y  C  F  N  G  D  N  D  A  P  G
7801  TTTTATACATCTATAGCGTTATTTTCCGCTAGGTGTTTATATTTAACAGCCAGCAATACACCTCAATTGTATTGCTTTAATGGTGATAATGATGCACCTG  7900

      A  L  P  F  G  S  I  I  P  H  R  V  Y  F  Q  P  N  G  V  R  L  I  V  P  Q  Q  I  L  H  T  P  Y  V
7901  GGGCTTTGCCATTTGGTAGTATTATTCCTCATAGAGTTTATTTCCAACCCAATGGTGTTAGGCTTATAGTTCCACAACAAATACTGCACACACCCTACGT  8000

      V  K  F  V  S  D  S  Y  C  R  G  S  V  C  E  Y  T  R  P  G  Y  C  V  S  L  N  P  Q  W  V  L  F  N
8001  AGTAAAGTTTGTATCAGACAGCTATTGTAGGGGTAGTGTGTGTGAGTACACTAGACCAGGTTACTGTGTGTCATTAAACCCACAATGGGTTTTGTTTAAT  8100

      D  E  Y  T  S  K  P  G  V  F  C  G  S  T  V  R  E  L  M  F  S  M  V  S  T  F  F  T  G  V  N  P  N  I
8101  GACGAATACACAAGTAAACCCGGTGTTTTCTGTGGTTCTACTGTTAGAGAACTTATGTTTAGTATGGTTAGTACATTCTTTACTGGTGTTAACCCCAATA  8200

      Y  M  Q  L  A  T  M  F  L  I  L  V  V  V  V  L  I  F  A  M  V  I  K  F  Q  G  V  F  K  A  Y  A  T
8201  TCTATATGCAATTAGCAACTATGTTTTTAATACTAGTTGTTGTTGTATTAATCTTTGCAATGGTTATAAAGTTTCAAGGTGTTTTTAAAGCTTATGCAAC  8300

      T  V  F  I  T  M  L  V  W  V  I  N  A  F  I  L  C  V  H  S  Y  N  S  V  L  A  V  I  L  L  V  L  Y
8301  CACTGTTTTTATAACAATGTTAGTTTGGGTAATTAACGCATTTATTTTGTGTGTACATAGTTACAACAGTGTTTTAGCTGTTATATTACTAGTACTCTAT  8400

      C  Y  A  S  L  V  T  S  R  N  T  V  I  I  M  H  C  W  L  V  F  T  F  G  L  I  V  P  T  W  L  A  C  C
8401  TGCTATGCGTCATTGGTTACAAGTCGCAATACTGTTATAATAATGCATTGTTGGCTTGTTTTTACCTTTGGTTTAATAGTACCCACATGGTTGGCTTGTT  8500

      Y  L  G  F  I  I  Y  M  Y  T  P  L  F  L  W  C  Y  G  T  T  K  N  T  R  K  L  Y  D  G  N  E  F  V
8501  GCTACCTGGGATTTATTATTTATATGTATACACCGTTGTTTTTATGGTGTTATGGTACTACAAAAAACACTCGTAAGCTGTATGATGGCAATGAGTTTGT  8600

      G  N  Y  D  L  A  A  K  S  T  F  V  I  R  G  S  E  F  V  K  L  T  N  E  I  G  D  K  F  E  A  Y  L
8601  TGGTAATTATGATCTTGCTGCGAAGAGCACTTTTGTTATTCGCGGCTCTGAATTTGTTAAGCTTACTAATGAGATAGGTGATAAATTTGAGGCCTACCTT  8700

      S  A  Y  A  R  L  K  Y  Y  S  G  T  G  S  E  Q  D  Y  L  Q  A  C  R  A  W  L  A  Y  A  L  D  Q  Y  R
8701  TCAGCGTATGCTAGATTAAAGTACTATTCAGGCACTGGCAGTGAACAAGATTATTTGCAAGCTTGTCGTGCATGGTTAGCTTATGCTTTGGACCAATATA  8800

      N  S  G  V  E  I  V  Y  T  P  P  R  Y  S  I  G  V  S  R  L  Q  S  G  F  K  K  L  V  S  P  S  S  A
8801  GAAATAGTGGTGTGGAAATTGTTTATACTCCGCCACGTTACTCTATTGGTGTTAGTAGATTACAATCTGGTTTTAAGAAACTGGTTTCTCCTAGTAGTGC  8900

      V  E  K  C  I  V  S  V  S  Y  R  G  N  N  L  N  G  L  W  L  G  D  T  I  Y  C  P  R  H  V  L  G  K
8901  TGTTGAAAAGTGCATTGTTAGTGTCTCTTATAGAGGTAATAATCTTAATGGACTGTGGCTAGGTGACACTATCTACTGTCCTCGTCATGTATTGGGTAAG  9000

      F  S  G  D  Q  W  N  D  V  L  N  L  A  N  N  H  E  F  E  V  T  T  Q  H  G  V  T  L  N  V  V  S  R  R
9001  TTTTCAGGTGACCAATGGAATGATGTACTTAATCTTGCTAATAATCATGAGTTTGAAGTTACAACTCAACATGGTGTTACTTTGAATGTTGTCAGTAGGC  9100

      L  K  G  A  V  L  I  L  Q  T  A  V  A  N  A  E  T  P  K  Y  K  F  I  K  A  N  C  G  D  S  F  T  I
9101  GTTTAAAAGGTGCAGTTTTAATTTTACAAACTGCTGTTGCTAATGCTGAAACTCCAAAGTATAAGTTTATTAAAGCTAATTGTGGTGATAGTTTCACTAT  9200

      A  C  A  Y  G  G  T  V  V  G  L  Y  P  V  T  M  R  S  N  G  T  I  R  A  S  F  L  A  G  A  C  G  S
9201  AGCTTGTGCTTATGGTGGTACAGTTGTAGGACTCTACCCTGTTACTATGCGTTCTAATGGTACTATTAGAGCATCTTTTCTTGCGGGAGCCTGTGGTTCA  9300

      V  G  F  N  I  E  K  G  V  V  N  F  F  Y  M  H  H  L  E  L  P  N  A  L  H  T  G  T  D  L  M  G  E  F
9301  GTTGGTTTTAATATAGAAAAGGGTGTAGTTAATTTCTTTTATATGCACCATCTTGAGTTACCTAATGCATTACACACTGGAACTGACCTAATGGGTGAAT  9400

      Y  G  G  Y  V  D  E  E  V  A  Q  R  V  P  P  D  N  L  V  T  N  N  I  V  A  W  L  Y  A  A  I  I  S
9401  TCTATGGTGGTTATGTTGATGAAGAGGTTGCACAAAGAGTGCCACCAGATAATTTAGTTACTAACAATATTGTAGCATGGCTCTATGCCGCAATTATTAG  9500

      V  K  E  S  S  F  S  L  P  K  W  L  E  S  T  T  V  S  V  D  D  Y  N  K  W  A  G  D  N  G  F  T  P
9501  TGTTAAGGAGAGTAGTTTCTCGCTGCCTAAATGGTTGGAGAGTACTACTGTTAGTGTTGATGATTATAATAAGTGGGCTGGTGACAATGGTTTTACACCA  9600

      F  S  T  S  T  A  I  T  K  L  S  A  I  T  G  V  D  V  C  K  L  L  R  T  I  M  V  K  N  S  Q  W  G  G
9601  TTTTCTACTAGTACCGCTATTACTAAATTAAGTGCTATAACTGGAGTTGATGTTTGTAAGCTCCTTCGCACTATTATGGTAAAAAATAGCCAGTGGGGTG  9700
```

```
          D P I L G Q Y N F E D E L T P E S V F N Q I G G V R L Q S S F V R
9701  GTGACCCCATTTTAGGGCAATATAATTTTGAAGATGAATTGACACCGGAGTCTGTATTTAATCAGATTGGTGGTGTTAGATTACAATCTTCTTTTGTAAG  9800

          K A T S W F W S R C V L A C F L F V L C A I V L F T A V P L K F Y
9801  AAAAGCTACATCTTGGTTTTGGAGTAGATGTGTGTTAGCTTGCTTCTTATTTGTGTTGTGTGCTATTGTCTTGTTTACGGCAGTGCCACTTAAATTTTAT  9900

          V Y A A V I L L M A V L F I S F T V K H V M A Y M D T F L L P T L I
9901  GTATATGCAGCTGTTATTTTGTTAATGGCTGTACTTTTTATTTCTTTTACTGTTAAACATGTTATGGCATATATGGATACTTTTCTATTGCCAACATTGA  10000

          T V I I G V C A E V P F I Y N T L I S Q V V I F L S Q W Y D P V V
10001 TTACAGTTATTATTGGAGTTTGTGCTGAAGTGCCTTTCATCTACAATACTCTAATTAGTCAAGTTGTTATTTTCTTAAGTCAATGGTATGACCCAGTAGT  10100

          F D T M V P W M F L P L V L Y T A F K C V Q G C Y M N S F N T S L
10101 CTTTGATACTATGGTACCATGGATGTTCTTGCCACTAGTGTTGTATACTGCTTTTAAGTGTGTACAAGGTTGCTATATGAATTCTTTCAATACTTCTTTG  10200

          L M L Y Q F V K L G F V I Y T S S N T L T A Y T E G N W E L F F E L
10201 TTAATGCTGTATCAGTTTGTGAAGTTAGGTTTTGTTATTTACACCTCTTCTAATACTCTTACTGCATACACAGAAGGTAATTGGGAGTTATTCTTCGAGT  10300

          V H T T V L A N V S S N S L I G L F V F K C A K W M L Y Y C N A T
10301 TGGTGCACACTACTGTGTTGGCTAATGTTAGTAGTAATTCTTTAATTGGTTTATTTGTTTTTAAGTGTGCTAAATGGATGTTGTATTATTGTAATGCAAC  10400

          Y L N N Y V L M A V M V N C I G W L C T C Y F G L Y W W V N K V F
10401 ATACTTAAACAATTATGTACTAATGGCAGTTATGGTTAACTGCATTGGCTGGCTCTGCACTTGTTACTTTGGGTTGTATTGGTGGGTTAATAAGGTTTTT  10500

          G L T L G K Y N F K V S V D Q Y R Y M C L H K I N P P K T V W E V F
10501 GGTTTAACCTTAGGTAAATACAATTTTAAAGTTTCAGTAGATCAATATAGGTATATGTGTTTGCACAAGATAAACCCACCTAAAACTGTGTGGGAAGTCT  10600

          S T N I L I Q G I G G D R V L P I A T V Q A K L S D V K C T T V V
10601 TTTCGACAAATATACTTATACAAGGAATTGGTGGTGACCGTGTGTTGCCTATTGCTACAGTTCAAGCTAAATTGAGTGATGTAAAGTGTACAACTGTTGT  10700

          L M Q L L T K L N V E A N S K M H V Y L V E L H N K I L A S D D V
10701 TTTAATGCAGCTTTTGACTAAGCTTAATGTTGAAGCAAATTCAAAAATGCATGTTTATCTTGTTGAGTTACACAATAAAATTCTTGCTTCTGATGATGTT  10800

          G E C M D N L L G M L I T L F C I D S T I D L S E Y C D D I L K R S
10801 GGAGAGTGCATGGATAATTTGTTGGGTATGCTTATAACACTATTTTGTATAGATTCTACTATTGATTTGAGTGAGTATTGTGATGACATACTTAAGAGGT  10900

          T V L Q S V T Q E F S H I P S Y A E Y E R A K N L Y E K V L V D S
10901 CAACTGTATTACAATCGGTTACTCAAGAATTCTCACATATACCCTCTTATGCTGAATATGAAAGGGCTAAGAATCTTTATGAAAAGGTTTTAGTTGATTC  11000

          K N G C V T Q Q E L A A Y R K A A N I A K S V F D R D L A V Q K K
11001 TAAAAATGGTGGTGTTACACAGCAAGAGCTTGCTGCATATCGTAAAGCTGCCAATATTGCAAAGTCAGTTTTTGATAGAGACTTGGCTGTCCAAAAGAAG  11100

          L D S M A E R A M T T M Y K E A R V T D R R A K L V S S L H A L L F
11101 TTAGATAGCATGGCAGAGCGTGCTATGACAACAATGTATAAAGAGGCGCGTGTAACAGATAGACGAGCAAAATTAGTCTCATCACTACATGCGTTACTTT  11200

          S M L K K I D S E K L N V L F D Q A S S G V V P L A T V P I V C S
11201 TCTCAATGCTTAAGAAAATAGATTCTGAAAAGCTTAATGTCTTGTTTGACCAGGCTAGTAGTGGTGTTGTGCCCCTAGCGACTGTTCCAATTGTTTGTAG  11300

          N K L T L V I P D P E T W V K C V E G V H V T Y S T V V W N I D T
11301 TAATAAGCTTACACTTGTAATACCAGACCCAGAAACGTGGGTCAAGTGTGTGGAAGGTGTGCATGTTACATATTCAACAGTTGTTTGGAATATAGACACT  11400

          V I D A D G T E L H P T S T G S G L T Y C I S G A N I A W P L K V N
11401 GTTATTGATGCCGATGGCACAGAGTTACACCCAACTTCTACAGGTAGTGGATTGACATACTGTATAAGTGGTGCTAATATAGCATGGCCTTTAAAGGTTA  11500

          L T R N G H N K V D V V L Q N N E L M P H G V K T K A C V A G V D
11501 ACTTGACTAGGAATGGGCATAATAAGGTTGATGTTGTTTTGCAAAATAATGAGCTTATGCCACATGGTGTTAAAACAAAGGCTTGCGTAGCAGGTGTAGA  11600

          Q A H C S V E S K C Y Y T N I S G N S V V A A I T S S N P N L K V
11601 TCAAGCACATTGTAGCGTAGAGTCTAAATGTTATTATACAAATATTAGTGGCAATTCAGTTGTAGCTGCTATTACTTCTTCAAATCCAAATCTGAAAGTA  11700

          A S F L N E A G N Q I Y V D L D P P C K F G M K V G V K V E V V Y L
11701 GCTTCGTTTTTGAATGAGGCAGGCAATCAGATTTATGTAGACTTAGACCCACCATGTAAATTTGGCATGAAAGTGGGTGTCAAGGTTGAGGTTGTTTACT  11800

          Y F I K N T R S I V R G M V L G A I S N V V V L Q S K G H E T E E
11801 TGTATTTTATAAAGAATACAAGGTCGATTGTTAGGGGTATGGTACTTGGTGCTATATCTAATGTTGTTGTCTTACAGTCTAAAGGGCATGAAACAGAGGA  11900

          V D A V G I L S L C S F A V D P A D T Y C K Y V A A G N Q P L G N
11901 AGTGGATGCTGTTGGCATTCTTTCACTATGTTCATTTGCAGTAGATCCCGCGGACACATATTGTAAATATGTGGCAGCAGGTAATCAACCTTTAGGTAAC  12000

          C V K M L T V H N G S G F A I T S K P S P T P D Q D S Y G G A S V C
12001 TGTGTTAAAATGTTGACAGTGCATAATGGTAGTGGTTTTGCTATAACTTCAAAGCCAAGTCCTACTCCTGACCAGGATTCTTATGGAGGAGCTTCTGTGT  12100
```

```
        L  Y  C  R  A  H  I  A  H  P  G  S  V  G  N  L  D  G  R  C  Q  F  K  G  S  F  V  Q  I  P  T  T  E
12101   GTCTCTATTGTAGAGCACACATAGCACATCCAGGAAGTGTAGGAAATTTAGATGGACGTTGTCAATTTAAAGGTTCTTTTGTGCAAATACCTACTACGGA  12200

        K  D  P  V  G  F  C  L  R  N  K  V  C  T  V  C  Q  C  W  I  G  Y  G  C  Q  C  D  S  L  R  Q  P  K
12201   GAAAGACCCCGTTGGATTCTGTCTACGTAATAAGGTTTGCACTGTTTGCCAGTGTTGGATTGGTTATGGATGTCAGTGTGATTCACTTAGACAACCAAAA  12300

        S  S  V  Q  S  V  A  G  A  S  D  F  D  K  N  Y  L  N  G  Y  G  V  A  V  R  L  G  *
12301   TCTTCTGTTCAATCAGTTGCTGGAGCATCTGATTTTGATAAGAATTATTTAAACGGGTACGGGGTAGCAGTGAGGCTCGGCTGATACCCCTTGCTAGTGG  12400

                                                                    M  F  Q  N  L  K  R  N  C  A  R  F  Q  E
12401   ATGTGATCCTGATGTTGTAAAGCGAGCCTTTGATGTTTGTAATAAGGAATCAGCTGGTATGTTTCAAAATTTGAAGCGTAACTGCGCTAGATTCCAGGAA  12500

        L  R  D  T  E  D  G  N  L  E  Y  L  D  S  Y  F  V  V  K  Q  T  T  P  S  N  Y  E  H  E  K  S  C  Y  E
12501   CTACGCGATACTGAAGATGGAAATCTTGAGTATCTTGATTCTTACTTTGTAGTTAAACAAACCACTCCTAGTAATTATGAACATGAAAAATCTTGTTACG  12600

        D  L  K  S  E  V  T  A  D  H  D  F  F  V  F  N  K  N  I  Y  N  I  S  R  Q  R  L  T  K  Y  T  M  M
12601   AAGACTTAAAGTCAGAAGTAACAGCTGACCATGACTTCTTTGTGTTCAATAAGAACATTTACAATATTAGTAGGCAACGGCTTACTAAATATACTATGAT  12700

        D  F  C  Y  A  L  R  H  F  D  P  K  D  C  E  V  L  K  E  I  L  V  T  Y  G  C  I  E  D  Y  H  P  K
12701   GGACTTCTGCTATGCTTTGAGACATTTCGACCCAAAGGATTGTGAAGTTCTTAAAGAAATACTTGTCACTTATGGTTGTATAGAAGACTATCACCCTAAG  12800

        W  F  E  E  N  K  D  W  Y  D  P  I  E  N  S  K  Y  Y  V  M  L  A  K  M  G  P  I  V  R  R  A  L  L  N
12801   TGGTTTGAGGAGAATAAGGATTGGTACGACCCAATAGAAAACTCAAAATATTATGTCATGTTGGCTAAAATGGGACCTATTGTACGACGTGCTTTATTGA  12900

        A  I  E  F  G  N  L  M  V  E  K  G  Y  V  G  V  I  T  L  D  N  Q  D  L  N  G  K  F  Y  D  F  G  D
12901   ATGCTATTGAGTTCGGAAACCTTATGGTTGAAAAAGGTTATGTTGGTGTTATTACACTCGATAACCAAGACCTTAATGGCAAATTTTATGATTTTGGTGA  13000

        F  Q  K  T  A  P  G  A  G  V  P  V  F  D  T  Y  Y  S  Y  M  M  P  I  I  A  M  T  D  A  L  A  P  E
13001   TTTTCAGAAGACGGCACCTGGTGCTGGTGTTCCTGTTTTTGATACGTATTATTCTTACATGATGCCCATCATAGCCATGACGGATGCTTTAGCACCTGAG  13100

        R  Y  F  E  Y  D  V  H  K  G  Y  K  S  Y  D  L  L  K  Y  D  Y  T  E  E  K  Q  E  L  F  Q  K  Y  F  K
13101   AGGTACTTTGAATATGATGTGCACAAGGGTTATAAATCTTATGATCTCCTCAAGTATGATTATACTGAGGAGAAACAAGAATTGTTTCAGAAGTACTTTA  13200

        Y  W  D  Q  E  Y  H  P  N  C  R  D  C  S  D  D  R  C  L  I  H  C  A  N  F  N  I  L  F  S  T  L  I
13201   AGTACTGGGATCAAGAGTATCATCCTAACTGCCGTGACTGTAGTGATGACAGGTGTTTGATACATTGTGCAAACTTCAATATCTTGTTTTCTACACTTAT  13300

        P  Q  T  S  F  G  N  L  C  R  K  V  F  V  D  G  V  P  F  I  A  T  C  G  Y  H  S  K  E  L  G  V  I
13301   ACCGCAGACTTCTTTCGGTAATTTGTGTAGAAAAGTTTTTGTTGATGGTGTACCATTTATAGCTACTTGTGGCTATCATTCTAAGGAACTTGGTGTTATT  13400

        M  N  Q  D  N  T  M  S  F  S  K  M  G  L  S  Q  L  M  Q  F  V  G  D  P  A  L  L  V  G  T  S  N  N  L
13401   ATGAATCAAGATAACACCATGTCTTTTTCAAAAATGGGTTTAAGTCAACTCATGCAGTTTGTTGGAGATCCTGCTTTGTTAGTGGGAACATCCAATAATT  13500

        V  D  L  R  T  S  C  F  S  V  C  A  L  T  S  G  I  T  H  Q  T  V  K  P  G  H  F  N  K  D  F  Y  D
13501   TAGTTGATCTTAGAACGTCTTGTTTTAGTGTTTGTGCGTTAACATCTGGTATTACTCATCAAACGGTAAAGCCAGGTCACTTTAACAAGGATTTCTATGA  13600

        F  A  E  K  A  G  M  F  K  E  G  S  S  I  P  L  K  H  F  F  Y  P  Q  T  G  N  A  A  I  N  D  Y  D
13601   TTTTGCAGAGAAGGCTGGTATGTTTAAGGAGGGTTCGTCTATACCACTTAAACATTTTTTCTATCCTCAAACTGGTAATGCTGCTATAAACGATTATGAT  13700

        Y  Y  R  Y  N  R  P  T  M  F  D  I  C  Q  L  L  F  C  L  E  V  T  S  K  Y  F  E  C  Y  E  G  G  C  I
13701   TATTATCGTTATAACAGGCCTACCATGTTTGACATATGTCAACTTCTATTTTGTTTAGAAGTGACTTCTAAATACTTTGAGTGTTATGAAGGCGGCTGTA  13800

        P  A  S  Q  V  V  V  N  N  L  D  K  S  A  G  Y  P  F  N  K  F  G  K  A  R  L  Y  Y  E  M  S  L  E
13801   TACCAGCTAGCCAAGTTGTAGTTAACAACTTAGATAAGAGTGCAGGCTATCCATTTAATAAGTTTGGAAAAGCCCGCCTCTATTATGAAATGAGTCTAGA  13900

        E  Q  D  Q  L  F  E  I  T  K  K  N  V  L  P  T  I  T  Q  M  N  L  K  Y  A  I  S  A  K  N  R  A  R
13901   GGAACAGGACCAACTCTTCGAGATTACGAAGAAGAATGTCCTACCCACTATAACTCAAATGAATTTAAAATATGCCATATCCGCGAAAAATAGAGCGCGT  14000

        T  V  A  G  V  S  I  L  S  T  M  T  N  R  Q  F  H  Q  K  I  L  K  S  I  V  N  T  R  N  A  S  V  V  I
14001   ACAGTGGCAGGTGTGTCTATCCTTTCTACTATGACTAATAGGCAGTTTCATCAGAAGATTCTTAAGTCTATAGTCAACACTAGAAATGCTTCTGTAGTTA  14100

        G  T  T  K  F  Y  G  G  W  D  N  M  L  R  N  L  I  Q  G  V  E  D  P  I  L  M  G  W  D  Y  P  K  C
14101   TTGGAACAACCAAGTTTTATGGCGGTTGGGACAACATGTTGAGAAACCTGATTCAGGGTGTTGAAGACCCAATTCTTATGGGTTGGGATTATCCTAAGTG  14200

        D  R  A  M  P  N  L  L  R  I  A  A  S  L  V  L  A  R  K  H  T  N  C  C  S  W  S  E  R  I  Y  R  L
14201   TGATAGAGCAATGCCTAATTTGTTGCGTATAGCAGCATCCTTAGTACTTGCTCGCAAACACACTAACTGTTGTAGTTGGTCTGAACGCATTTATAGGTTG  14300

        Y  N  E  C  A  Q  V  L  S  E  T  V  L  A  T  G  G  I  Y  V  K  P  G  G  T  S  S  G  D  A  T  T  A  Y
14301   TATAATGAATGCGCCCAGGTCTTATCTGAAACTGTACTTGCTACAGGTGGTATTTATGTTAAACCTGGTGGCACTAGCAGTGGTGATGCTACTACTGCTT  14400

        A  N  S  V  F  N  I  I  Q  A  T  S  A  N  V  A  R  L  L  S  V  I  T  R  D  I  V  Y  D  N  I  K  S
14401   ATGCAAACAGTGTTTTTAACATAATACAAGCCACATCTGCTAATGTTGCGCGTCTTTTGAGTGTTATAACGCGTGATATTGTCTATGATAATATTAAGAG  14500
```

```
      L  Q  Y  E  L  Y  Q  Q  V  Y  R  R  V  N  F  D  P  A  F  V  E  K  F  Y  S  Y  L  C  K  N  F  S  L
14501 CTTGCAGTATGAATTGTATCAGCAGGTCTACAGGCGAGTTAATTTTGACCCAGCCTTTGTTGAAAAGTTTTATTCTTACTTATGTAAGAATTTTTCGTTG 14600

      M  I  L  S  D  D  G  V  V  C  Y  N  N  T  L  A  K  Q  G  L  V  A  D  I  S  G  F  R  E  V  L  Y  Y  Q
14601 ATGATCTTGTCTGACGACGGTGTTGTTTGTTACAACAACACATTAGCCAAACAAGGTCTTGTAGCAGATATTTCTGGTTTTAGAGAGGTTCTCTACTATC 14700

      N  N  V  F  M  A  D  S  K  C  W  V  E  P  D  L  E  K  G  P  H  E  F  C  S  Q  H  T  M  L  V  E  V
14701 AGAATAATGTTTTTATGGCTGATTCTAAATGTTGGGTTGAACCAGATTTAGAAAAAGGCCCACATGAGTTTTGTTCACAACACACAATGCTAGTGGAGGT 14800

      D  G  E  P  K  Y  L  P  Y  P  D  P  S  R  I  L  G  A  C  V  F  V  D  D  V  D  K  T  E  P  V  A  V
14801 TGATGGTGACCCTAAGTATTTGCCATACCCAGACCCTTCACGCATTTTGGGTGCATGTGTTTTTGTAGATGACGTGGATAAGACAGAACCTGTGGCTGTT 14900

      M  E  R  Y  I  A  L  A  I  D  A  Y  P  L  V  H  H  E  N  E  E  Y  K  K  V  F  F  V  L  L  A  Y  I  R
14901 ATGGAGCGTTATATAGCTCTTGCCATAGATGCTTATCCACTAGTACATCATGAAAATGAAGAGTACAAGAAGGTATTCTTTGTTCTCCTTGCATATATCA 15000

      K  L  Y  Q  E  L  S  Q  N  M  L  M  D  Y  S  F  V  M  D  I  D  K  G  S  K  F  W  E  Q  E  F  Y  E
15001 GAAAACTCTATCAAGAGCTTTCTCAGAATATGCTTATGGACTACTCTTTTGTAATGGATATAGACAAGGGTAGTAAATTTTGGGAACAGGAGTTCTATGA 15100

      N  M  Y  R  A  P  T  T  L  Q  S  C  G  V  C  V  V  C  N  S  Q  T  I  L  R  C  G  N  C  I  R  K  P
15101 GAATATGTATAGAGCTCCTACGACTTTACAATCTTGTGGCGTTTGTGTAGTTTGTAATAGTCAAACTATACTACGCTGCGGTAATTGTATTCGTAAACCG 15200

      F  L  C  C  K  C  C  Y  D  H  V  M  H  T  D  H  K  N  V  L  S  I  N  P  Y  I  C  S  Q  L  G  C  G  E
15201 TTTTTGTGTTGTAAGTGTTGCTATGACCACGTCATGCATACGGACCACAAAAATGTTTTATCTATAAATCCTTATATTTGCTCACAGCTAGGTTGCGGTG 15300

      A  D  V  T  K  L  Y  L  G  G  M  S  Y  F  C  G  N  H  K  P  K  L  S  I  P  L  V  S  N  G  T  V  F
15301 AAGCAGATGTTACTAAATTGTACCTCGGGGGTATGTCGTACTTCTGTGGTAATCATAAACCGAAATTGTCAATACCGTTAGTATCTAATGGTACTGTTTT 15400

      G  I  Y  R  A  N  C  A  G  S  E  N  V  D  D  F  N  Q  L  A  T  T  N  W  S  I  V  E  P  Y  I  L  A
15401 TGGAATTTACAGGGCTAATTGTGCTGGTAGTGAAAATGTTGATGATTTTAATCAACTAGCTACTACTAATTGGTCCATTGTCGAACCTTATATTTTAGCA 15500

      N  R  C  S  D  S  L  R  R  F  A  A  E  T  V  K  A  T  E  E  L  H  K  Q  Q  F  A  S  A  E  V  R  E  V
15501 AATCGCTGTAGTGATTCATTGAGACGTTTTGCTGCAGAGACAGTAAAAGCCACAGAAGAATTACATAAGCAACAATTTGCTAGTGCAGAAGTGCGAGAAG 15600

      F  S  D  R  E  L  I  L  S  W  E  P  G  K  T  R  P  P  L  N  R  N  Y  V  F  T  G  Y  H  F  T  R  T
15601 TATTCTCAGATCGTGAATTGATTCTATCATGGGAACCAGGAAAAACCAGGCCGCCATTGAATAGAAATTATGTTTTCACAGGTTATCACTTTACAAGAAC 15700

      S  K  V  Q  L  G  D  F  T  F  E  K  G  E  G  K  D  V  V  Y  Y  K  A  T  S  T  A  K  L  S  V  G  D
15701 TAGTAAGGTGCAGCTTGGTGATTTTACATTTGAAAAAGGTGAAGGTAAGGATGTTGTCTATTATAAAGCAACGTCTACTGCTAAATTGTCTGTAGGAGAC 15800

      I  F  V  L  T  S  H  N  V  V  S  L  V  A  P  T  L  C  P  Q  Q  T  F  S  R  F  V  N  L  R  P  N  V  M
15801 ATTTTTGTTTTAACCTCACACAATGTTGTTTCTCTCGTAGCGCCAACATTGTGTCCACAACAAACCTTTTCTAGGTTTGTAAATTTAAGACCTAATGTAA 15900

      V  P  E  C  F  V  N  N  I  P  L  Y  H  L  V  G  K  Q  K  R  T  T  V  Q  G  P  P  G  S  G  K  S  H
15901 TGGTACCTGAATGTTTTGTAAATAACATTCCACTTTACCATTTAGTAGGTAAACAGAAGCGTACTACAGTACAAGGTCCTCCTGGCAGTGGTAAATCCCA 16000

      F  A  I  G  L  A  V  Y  F  S  S  A  R  V  V  F  T  A  C  S  H  A  A  V  D  A  L  C  E  K  A  F  K
16001 CTTTGCTATAGGCCTTGCAGTATACTTTAGTAGCGCTCGTGTTGTTTTTACTGCATGTTCTCATGCAGCTGTTGATGCTTTATGTGAAAAAGCTTTTAAG 16100

      F  L  K  V  D  D  C  T  R  I  V  P  Q  R  T  T  V  D  C  F  S  K  F  K  A  N  D  T  G  K  K  Y  I  F
16101 TTTCTTAAAGTTGATGATTGCACTCGTATAGTACCCCAAAGGACTACTGTCGATTGCTTCTCAAAATTTAAAGCTAATGACACAGGCAAAAAGTACATTT 16200

      S  T  I  N  A  L  P  E  V  S  C  D  I  L  L  V  D  E  V  S  M  L  T  N  Y  E  L  S  F  I  N  G  K
16201 TTAGTACTATTAATGCCTTGCCGGAAGTTAGTTGTGATATTCTTTTGGTTGACGAGGTTAGTATGTTGACCAATTACGAATTGTCCTTTATTAATGGTAA 16300

      I  N  Y  Q  Y  V  V  Y  V  G  D  P  A  Q  L  P  A  P  R  T  L  L  N  G  S  L  S  P  K  D  Y  N  V
16301 GATAAATTACCAATATGTTGTGTATGTAGGTGATCCGGCTCAATTACCGGCACCCCGCACTTTACTTAATGGTTCACTTTCTCCAAAGGATTATAATGTT 16400

      V  T  N  L  M  V  C  V  K  P  D  I  F  L  A  K  C  Y  R  C  P  K  E  I  V  D  T  V  S  T  L  V  Y  D
16401 GTCACAAACCTTATGGTTTGTGTTAAACCTGATATTTTCCTTGCAAAGTGTTATCGTTGTCCTAAGGAAATTGTAGACACTGTGTCTACTCTTGTTTATG 16500

      G  K  F  I  A  N  N  P  E  S  R  E  C  F  K  V  I  V  N  N  G  N  S  D  V  G  H  E  S  G  S  A  Y
16501 ATGGAAAGTTTATTGCAAATAACCCAGAATCACGTGAGTGTTTCAAGGTTATAGTTAATAATGGCAATTCTGATGTAGGACATGAAAGTGGTTCAGCCTA 16600

      N  T  T  Q  L  E  F  V  K  D  F  V  C  R  N  K  Q  W  R  E  A  I  F  I  S  P  Y  N  A  M  N  Q  R
16601 CAACACAACACAATTGGAATTTGTGAAAGACTTTGTTTGTCGCAATAAACAATGGCGGGAAGCAATATTTATTTCACCTTACAATGCTATGAACCAGAGA 16700

      A  Y  R  M  L  G  L  N  V  Q  T  V  D  S  S  Q  G  S  E  Y  D  Y  V  I  F  C  V  T  A  D  S  Q  H  A
16701 GCTTACCGTATGCTTGGACTTAATGTTCAAACAGTAGATTCTTCTCAAGGTTCAGAGTATGATTATGTCATCTTCTGTGTTACTGCAGATTCGCAGCATG 16800

      L  N  I  N  R  F  N  V  A  L  T  R  A  K  R  G  I  L  V  V  M  R  Q  R  D  E  L  Y  S  A  L  K  F
16801 CACTGAATATTAATAGATTTAATGTGGCGCTTACAAGAGCTAAGCGTGGTATACTAGTTGTCATGCGCCAGCGTGATGAATTGTATTCTGCTCTTAAGTT 16900
```

```
       T  E  L  D  S  E  T  S  L  Q  G  T  G  L  F  K  I  C  N  K  E  F  S  G  V  H  P  A  Y  A  V  T  T
16901  TACAGAGCTAGATAGTGAAACAAGTCTGCAAGGTACAGGTTTGTTTAAAATTTGCAACAAAGAATTTAGTGGTGTCCATCCTGCTTATGCAGTCACAACT  17000

       K  A  L  A  A  T  Y  K  V  N  D  E  L  A  A  L  V  N  V  E  A  G  S  E  I  T  Y  K  H  L  I  S  L  L
17001  AAGGCTCTTGCTGCAACCTATAAAGTTAATGATGAACTTGCTGCACTTGTTAATGTGGAAGCTGGTTCAGAAATAACATATAAACATCTTATTTCTCTGT  17100

       G  F  K  M  S  V  N  V  E  G  C  H  N  M  F  I  T  R  D  E  A  I  R  N  V  R  G  W  V  G  F  D  V
17101  TAGGATTCAAGATGAGTGTTAATGTTGAAGGCTGCCACAACATGTTTATAACACGTGATGAGGCAATCCGCAATGTAAGAGGTTGGGTAGGTTTTGATGT  17200

       E  A  T  H  A  C  G  T  N  I  G  T  N  L  P  F  Q  V  G  F  S  T  G  A  D  F  V  V  T  P  E  G  L
17201  AGAAGCAACACATGCTTGTGGCACTAACATTGGTACTAACCTGCCTTTTCAAGTAGGTTTCTCTACTGGTGCAGACTTTGTAGTCACGCCTGAGGGACTT  17300

       V  D  T  S  I  G  N  N  F  E  P  V  N  S  K  A  P  P  G  E  Q  F  N  H  L  R  V  L  F  K  S  A  K  P
17301  GTAGATACTTCAATAGGCAATAATTTTGAGCCTGTGAATTCTAAAGCACCTCCAGGTGAACAATTTAACCACTTGAGAGTGTTATTTAAAAGTGCTAAAC  17400

       W  H  V  I  R  P  R  I  V  Q  M  L  A  D  N  L  C  N  V  S  D  C  V  V  F  V  T  W  C  H  G  L  E
17401  CTTGGCATGTTATAAGACCAAGGATAGTGCAGATGTTAGCAGACAATCTATGCAACGTTTCAGATTGTGTAGTGTTTGTCACATGGTGTCATGGCCTAGA  17500

       L  T  T  L  R  Y  F  V  K  I  G  K  E  Q  V  C  S  C  G  S  R  A  T  T  F  N  S  H  T  Q  A  Y  A
17501  ACTAACTACTTTGCGCTATTTTGTTAAAATAGGCAAGGAACAAGTTTGTTCTTGTGGTTCTAGAGCTACAACTTTTAATTCTCATACTCAAGCTTATGCT  17600

       C  W  K  H  C  L  G  F  D  F  V  Y  N  P  L  L  V  D  I  Q  Q  W  G  Y  S  G  N  L  Q  F  N  H  D  L
17601  TGTTGGAAGCATTGTTTGGGTTTTGATTTTGTTTATAACCCACTTCTAGTGGATATTCAACAGTGGGGTTACTCGGGTAACCTACAGTTTAATCATGATT  17700

       H  C  N  V  H  G  H  A  H  V  A  S  V  D  A  I  M  T  R  C  L  A  I  N  N  A  F  C  Q  D  V  N  W
17701  TGCACTGTAATGTGCATGGCCACGCTCATGTAGCTTCTGTTGACGCTATAATGACTCGTTGTCTTGCAATTAACAATGCATTTTGTCAAGATGTCAACTG  17800

       D  L  T  Y  P  H  I  A  N  E  D  E  V  N  S  S  C  R  Y  L  Q  R  M  Y  L  N  A  C  V  D  A  L  K
17801  GGATTTGACATACCCTCACATTGCAAATGAGGATGAAGTCAATTCTAGTTGTAGATATCTACAACGCATGTATCTTAATGCGTGTGTTGATGCTCTTAAA  17900

       V  N  V  V  Y  D  I  G  N  P  K  G  I  K  C  V  R  R  G  D  V  N  F  R  F  Y  D  K  N  P  I  V  R  N
17901  GTTAATGTTGTCTATGATATAGGCAACCCTAAAGGTATTAAATGTGTTAGGCGTGGGGATGTTAATTTTAGATTCTATGATAAGAATCCAATAGTACGCA  18000

       V  K  Q  F  E  Y  D  Y  N  Q  H  K  D  K  F  A  D  G  L  C  M  F  W  N  C  N  V  D  C  Y  P  D  N
18001  ACGTCAAGCAGTTTGAGTATGACTATAATCAGCACAAAGATAAGTTTGCTGATGGTCTTTGTATGTTTTGGAATTGTAATGTGGATTGTTATCCTGATAA  18100

       S  L  V  C  R  Y  D  T  R  N  L  S  V  F  N  L  P  G  C  N  G  G  S  L  Y  V  N  K  H  A  F  Y  T
18101  TTCCTTGGTTTGTAGGTATGACACACGAAATTTGAGTGTGTTTAACCTACCAGGCTGTAATGGTGGTAGTCTGTACGTTAACAAACATGCATTCTACACA  18200

       P  K  F  D  R  I  S  F  R  N  L  K  A  M  P  F  F  F  Y  D  S  S  P  C  E  T  I  Q  V  D  G  V  A  Q
18201  CCTAAATTTGACCGCATTAGCTTCCGCAATTTGAAAGCTATGCCATTCTTTTTTTATGACTCATCGCCTTGTGAAACCATTCAAGTGGATGGAGTTGCGC  18300

       D  L  V  S  L  A  T  K  D  C  I  T  K  C  N  I  G  G  A  V  C  K  K  H  A  Q  M  Y  A  E  F  V  T
18301  AAGACCTTGTGTCTCTAGCTACGAAAGACTGTATCACAAAGTGCAACATTGGTGGCGCTGTTTGTAAGAAACATGCCCAGATGTATGCAGAATTTGTGAC  18400

       S  Y  N  A  A  V  T  A  G  F  T  F  W  V  T  N  K  L  N  P  Y  N  L  W  K  S  F  S  A  L  Q  S  I
18401  TTCTTACAATGCAGCTGTCACAGCTGGCTTTACTTTCTGGGTAACTAATAAACTTAACCCTTATAACTTATGGAAAAGTTTTTCAGCTCTCCAGTCTATC  18500

       D  N  I  A  Y  N  M  Y  K  G  G  H  Y  D  A  I  A  G  E  M  P  T  V  I  T  G  D  K  V  F  V  I  D  Q
18501  GACAATATTGCTTATAATATGTATAAGGGTGGTCATTATGATGCTATTGCTGGAGAAATGCCCACTGTCATAACTGGAGACAAAGTTTTTGTTATTGATC  18600

       G  V  E  K  A  V  F  V  N  Q  T  T  L  P  T  S  V  A  F  E  L  Y  A  K  R  N  I  R  T  L  P  N  N
18601  AAGGTGTAGAAAAGGCAGTTTTTGTTAATCAAACAACTCTACCTACATCTGTGGCGTTTGAGCTATATGCAAAGAGAAATATTCGCACACTGCCAAACAA  18700

       R  I  L  K  G  L  G  V  D  V  T  N  G  F  V  I  W  D  Y  A  N  Q  T  P  L  Y  R  N  T  V  K  V  C
18701  CCGTATTTTGAAAGGTTTAGGTGTAGACGTAACCAATGGATTTGTAATTTGGGATTATGCTAACCAAACACCATTGTATCGTAATACCGTCAAGGTATGT  18800

       A  Y  T  D  I  E  P  N  G  L  V  V  L  Y  D  D  R  Y  G  D  Y  Q  S  F  L  A  A  D  N  A  V  L  V  S
18801  GCATATACAGATATTGAGCCAAATGGCCTAGTAGTTCTGTATGATGATAGATATGGTGATTACCAGTCTTTTCTTGCTGCTGATAATGCTGTTCTAGTTT  18900

       T  Q  C  Y  K  R  Y  S  Y  V  E  I  P  S  N  L  L  V  Q  N  G  M  P  L  K  D  G  A  N  L  Y  V  Y
18901  CTACACAGTGTTATAAGCGATATTCATACGTAGAAATACCATCTAATTTGCTCGTTCAGAATGGTATGCCATTAAAAGATGGAGCGAACCTGTATGTTTA  19000

       K  R  V  N  G  A  F  V  T  L  P  N  T  I  N  T  Q  G  R  S  Y  E  T  F  E  P  R  S  D  I  E  R  D
19001  TAAGCGTGTTAATGGTGCGTTTGTTACACTACCTAACACAATAAACACCCAGGGTCGAAGTTATGAAACTTTTGAACCTCGTAGTGACATTGAGCGTGAT  19100

       F  L  A  M  S  E  E  S  F  V  E  R  Y  G  K  D  L  G  L  Q  H  I  L  Y  G  E  V  D  K  P  Q  L  G  G
19101  TTTCTCGCTATGTCAGAGGAGAGTTTTGTAGAAAGGTATGGTAAAGACTTAGGCCTACAACACATACTGTATGGTGAAGTTGATAAGCCCCAATTAGGTG  19200

       L  H  T  V  I  G  M  Y  R  L  L  R  A  N  K  L  N  A  K  S  V  T  N  S  D  S  D  V  M  Q  N  Y  F
19201  GTTTACACACTGTTATAGGTATGTACAGACTCTTACGTGCGAATAAGTTGAACGCAAAGTCTGTAACTAATTCGGATTCTGATGTCATGCAAAATTACTT  19300
```

```
      V L S D N G S Y K Q V C T V V D L L L D D F L E L L R N I L K E Y
19301 TGTATTGTCGGACAATGGTTCTTACAAGCAAGTGTGTACTGTTGTGGATTTACTGCTTGATGATTTCTTAGAACTTCTTAGAAACATACTTAAGGAGTAT 19400

      G T N K S K V V T V S I D Y H S I N F M T W F E D G S I K T C Y P Q
19401 GGTACTAATAAGTCAAAAGTTGTAACAGTGTCAATTGATTACCATAGCATAAATTTTATGACTTGGTTTGAAGATGGCAGTATTAAAACATGTTATCCAC 19500

      L Q S A W T C G Y N M P E L Y K V Q N C V M E P C N I P N Y G V G
19501 AGCTTCAATCAGCATGGACGTGTGGTTATAATATGCCTGAACTTTATAAAGTTCAGAATTGTGTTATGGAACCTTGCAACATTCCTAATTATGGTGTTGG 19600

      I T L P S G I L M N V A K Y T Q L C Q Y L S K T T I C V P H N M R
19601 AATAACGTTGCCTAGCGGTATTCTTATGAATGTGGCAAAGTATACACAACTTTGTCAATACCTTTCGAAAACAACAATTTGTGTACCGCATAACATGCGA 19700

      V M H F G A G S D K G V A P G S T V L K Q W L P E G T L L V D N D I
19701 GTAATGCATTTCGGAGCAGGAAGCGACAAAGGAGTGGCGCCAGGTAGTACTGTTCTTAAACAATGGCTCCCAGAAGGGACACTCCTTGTCGATAATGATA 19800

      V D Y V S D A H V S V L S D C N K Y N T E H K F D L V I S D M Y T
19801 TTGTAGACTATGTGTCTGATGCACATGTTTCTGTGCTTTCAGATTGCAATAAATATAATACAGAGCACAAGTTTGATCTTGTGATATCTGATATGTATAC 19900

      D N D S K R K H E G V I A N N G N D D V F I Y L S S F L R N N L A
19901 AGATAATGATTCAAAAAGAAAGCATGAAGGCGTGATAGCCAATAATGGCAATGATGACGTTTTCATATATCTCTCAAGTTTTCTTCGTAACAATTTGGCT 20000

      L G G S F A V K V T E T S W H E V L Y D I A Q D C A W W T M F C T A
20001 CTAGGTGGTAGTTTTGCTGTAAAAGTGACAGAGACAAGTTGGCACGAAGTTTTATATGACATTGCACAGGATTGTGCATGGTGGACAATGTTTTGTACAG 20100

      V N A S S S E A F L I G V N Y L G A S E K V K V S G K T L H A N Y
20101 CAGTGAATGCCTCTTCTTCAGAAGCATTCTTGATTGGTGTTAATTATTTGGGTGCAAGTGAAAAGGTTAAGGTTAGTGGAAAAACGCTGCACGCAAATTA 20200

      I F W R N C N Y L Q T S A Y S I F D V A K F D L R L K A T P V V N
20201 TATATTTTGGAGGAATTGTAATTATTTACAAACCTCTGCTTATAGTATATTTGACGTTGCTAAGTTTGATTTGAGATTGAAAGCAACGCCAGTTGTTAAT 20300

      L K T E Q K T D L V F N L I K C G K L L V R D V G N T S F T S D S F
                                                    M L V T P L L L V T L
20301 TTGAAAACTGAACAAAAGACAGACTTAGTCTTTAATTTAATTAAGTGTGGTAAGTTACTGGTAAGAGATGTTGGTAACACCTCTTTTACTAGTGACTCTT 20400

      V C T M *
      L C A L C S A V L Y D S S S Y V Y Y Y Q S A F R P P S G W H L Q G
20401 TTGTGTGCACTATGTAGTGCTGTTTTGTATGACAGTAGTTCTTACGTTTACTACTACCAAAGTGCCTTCAGACCACCTAGTGGTTGGCATTTACAAGGGG 20500
```

Fig. 2. The sequence of the 'unique' region of mRNA F from the Beaudette strain of IBV. Translations of the ORFs are shown in single-letter amino acid code. The amino acid is shown above the first base of the appropriate codon. The translation starting at position 20368 is the NH_2 terminus of the spike precursor protein.

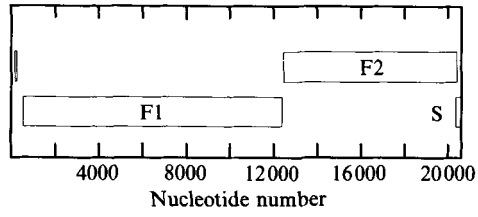


Fig. 3. Diagram showing the positions of the main ORFs in the 'unique' region of mRNA F. The two large ORFs, designated F1 and F2 are shown, as well as a small ORF at the 5′ end of the genome, and the start of the spike precursor gene, which overlaps with F2.

The second large ORF, F2, extends into the 'unique' region of mRNA E and in fact overlaps the coding sequences for the spike protein gene by 16 amino acids.

Potential sources of error

All the sequence information has been confirmed by sequencing M13 clones obtained from both strands of the DNA. In addition most of it has been sequenced several times from different M13 clones. The 14 cDNA clones used to obtain the sequence of mRNA F contain, including

overlaps, 24765 bases. During the shotgun sequencing of these clones 203113 bases have been sequenced, so that each base has, on average, been sequenced 8·2 times. However there are two regions we have checked more carefully. The first is at positions 12340 to 12390 where F1 ends and F2 begins. An error here leading to a frameshift could make the difference between two large ORFs and one very large ORF. The second is at position 167 where the very small 11 amino acid ORF ends. A frameshifting error here could mean that this first ORF can continue for another 77 amino acids until position 397. There are two possible sorts of error. The first is an artefact in the sequencing gels leading to a misreading. The sequence on both strands appears perfectly clear in both these regions. Both regions have been sequenced using formamide gels, high temperature gels, in addition to the use of deoxyinosine triphosphate (Bankier & Barrell, 1983) or deoxy-7-deazaguanosine triphosphate (Mizusawa *et al.*, 1986) to replace deoxyguanosine triphosphate and cytosine-modified sequence reaction products (Ambartsumyan & Mazo, 1980) to avoid gel compressions.

The second potential source of error is either a reverse transcriptase error during the synthesis of the cDNA or the occurrence of a mutant RNA molecule from which the cDNA was copied, both of which would lead to an incorrect cDNA clone. In the case of position 167 the sequence has been obtained from an equivalent clone from the M41 strain of IBV and is identical. In the case of the sequence between F1 and F2 the sequence has been confirmed from two additional independent cDNA clones, by sequencing directly from the double-stranded DNA using an oligonucleotide primer (Korneluk *et al.*, 1985). Fig. 4(*a*) shows the relevant sequence in this region and Fig. 4(*b*) shows a sequencing gel of bases 12333 to 12390 obtained directly from a cDNA clone using an oligonucleotide primer. In addition the sequence has been obtained directly from the virion RNA using specific oligonucleotide primers at both of these points and has confirmed the original gel readings. At positions 12333 to 12390 the sequence has also been obtained from virion RNA obtained from the M41 strain of IBV, and the sequence in this region is identical.

Gel compressions are thought to be caused by the presence of hairpin loops in the DNA migrating down the gel. Examination of the sequence in these regions shows that there are several possibilities for the formation of fairly large hairpins, including for example, at the position between F1 and F2, the sequence GGGGTA with its exact complement TACCCC 24 bases further on. At this position (12380), in the region where the reading frame changes between F1 and F2, the sequence has been determined from ten separate M13 clones. It is interesting to note that one of these clones gave a different sequence reading in that a CT dinucleotide, which appears in the other nine M13 readings, was not present. This is unusual as normally all independent M13 clones agree. It is possible that the secondary structure in this region has some effect on the fidelity of copying by polymerases.

Computer analysis

Extensive computer analysis has been carried out in an attempt to identify some salient features on the bleak landscapes of these large ORFs. Searches for homologies with other viral polymerases have been performed using the NBRF protein identification resource (George *et al.*, 1986). Short regions of fairly low homology with several viral polymerases can be identified but in general they do not rise significantly above the background of matches with proteins that are apparently unrelated. One region, between amino acids 1342 and 1350, has a fairly good match (8/9 amino acids) with the nsP2 protein of Sindbis virus, a protein which is known to be involved in RNA replication (Strauss & Strauss, 1983). This region also has a match with the 1a protein of brome mosaic virus. These matches are shown in Fig. 5. One of the most interesting matches is at the 5′ end of the first large ORF. The first 300 amino acids have a low-level but extensive homology with the replication initiation protein from *Escherichia coli* (Germino & Bastia, 1982). The homology is statistically significant and it may indicate that this region of the polymerase protein is involved in initiation of replication of either the positive or negative strands.

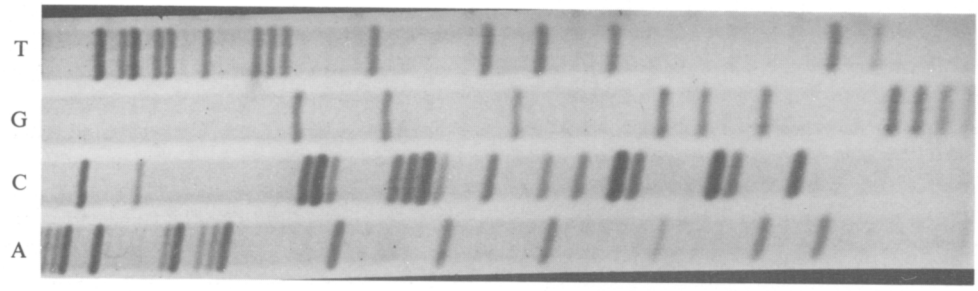
The predicted amino acid sequences of the large ORFs have been compared against themselves and against each other to see whether there are any repeats which might represent

(*a*)

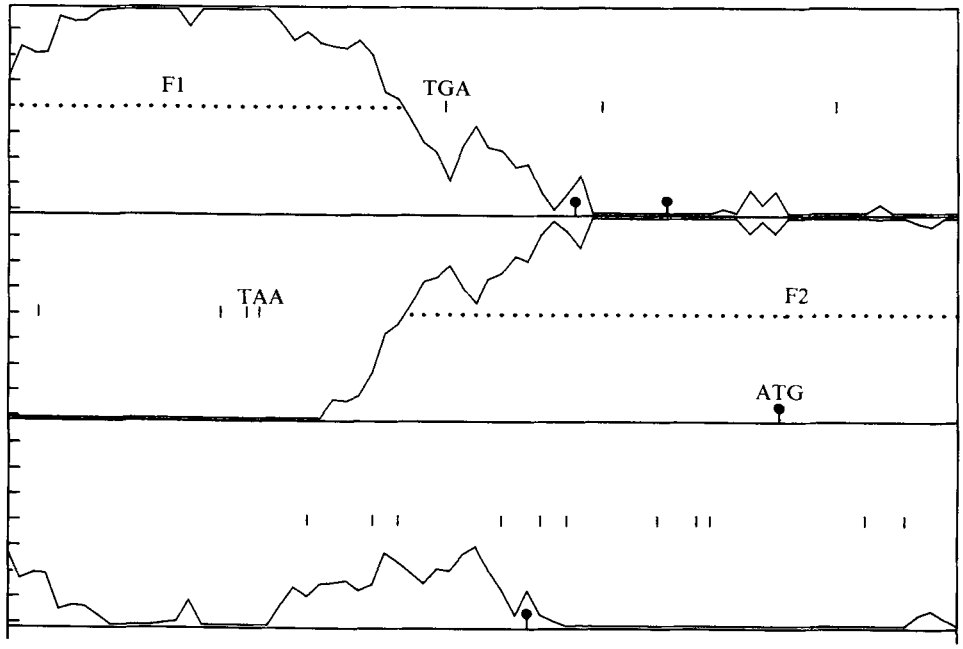
```
  S  L  R  Q  P  K  S  S  V  Q  S  V  A  G  A  S  D  F  D  K  N  Y  L  N  G  Y  G  V  A  V  R  L  G  *  Y  P  L
  F  T  *  T  T  K  I  F  C  S  I  S  C  W  S  I  *  F  *  *  E  L  F  K  R  V  R  G  S  S  E  A  R  L  I  P  L
  I  H  L  D  N  Q  N  L  L  F  N  Q  L  L  E  H  L  I  L  I  R  I  I  *  T  G  T  G  *  Q  *  G  S  A  D  T  P
ATTCACTTAGACAACCAAAATCTTCTGTTCAATCAGTTGCTGCAGCATCTGATTTTGATAAGAATTATTTAAACGGGTACGGGGTAGCAGTGAGGCTCGGCTGATACCCCT
    12290     12300     12310     12320     12330     12340     12350     12360     12370     12380     12390

  L  V  D  V  I  L  M  L  *  S  E  P  L  M  F  V  I  R  N  Q  L  V  C  F  K  I  *  S  V  T  A  L  D  S  R  N
  A  S  G  C  D  P  D  V  V  K  R  A  F  D  V  C  N  K  E  S  A  G  M  F  Q  N  L  K  R  N  C  A  R  F  Q  E
C  *  W  M  *  S  *  C  C  K  A  S  L  *  C  L  *  *  G  I  S  W  Y  V  S  K  F  E  A  *  L  R  *  I  P  G
GCTAGTGGATGTGATCCTGATGTTGTAAAGCGAGCCTTTGATGTTTGTAATAAGGAATCAGCTGGTATGTTTCAAAATTTGAAGCGTAACTGCGCTAGATTCCAGGAA
    12400     12410     12420     12430     12440     12450     12460     12470     12480     12490     12500
```

(*b*)



(*c*)



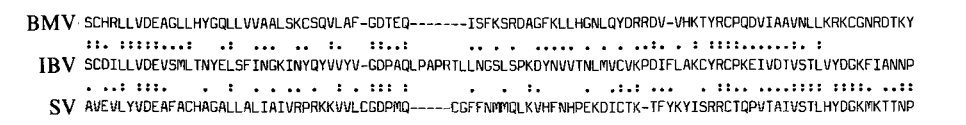


Fig. 5. Comparison between amino acid sequences of brome mosaic virus (BMV), infectious bronchitis virus (IBV) and Sindbis virus (SV). The BMV sequences are amino acids 748 to 838 of the 1a protein. The SV sequences are amino acids 785 to 878 of the nsP2 protein, The IBV sequences are amino acids 1248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

two separate but similar polymerases. A dot matrix comparison, such as DIAGON (Staden, 1982*a*), reveals no repeats. However several low homology repeats can be detected using the program FASTP (Lipman & Pearson, 1985). These are shown on Fig. 6(*a*) beneath a hydrophilicity plot (Kyte & Doolittle, 1982) of the amino acid sequences of F1 and F2. Fig. 6(*b* to *e*) shows the amino acid matches in these regions. The spacing between the repeats marked A and B is very similar in both cases, 1157 amino acids in F1 and 1183 amino acids in F2. It is possible that these represent residual domains of homology between two polymerases which were at one time more closely related. The areas marked C and D also show regions of homology. The diagram also shows several very hydrophobic regions in the first large ORF which represent potential membrane-spanning domains.

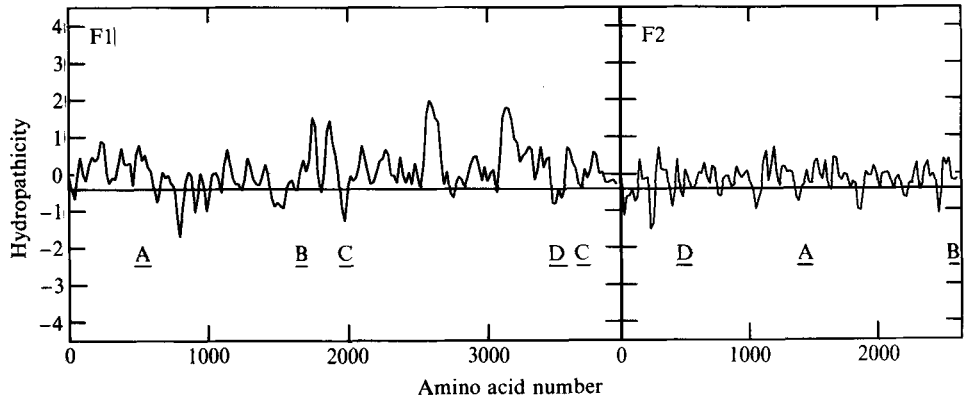
Computer analysis has also detected a homology between the non-coding region at the 5′ end of the positive strand, and the 5′ end of the negative strand (i.e. the reverse complement of the non-coding region at the 3′ end of the positive strand). This is shown in Fig. 7. These sequences, on the positive and negative strands, are approximately the same distance from their 5′ ends, 52 bases and 48 bases [excluding the poly(A) tail] respectively, and may play some role in the replication of the positive and negative strands.

Homology regions

At position 599 the sequence CTGAACAA occurs. This is identical to the sequence which occurs in the 'homology regions' at the 5′ ends of the bodies of mRNAs D and E (Boursnell *et al.*, 1985*b*; Binns *et al.*, 1985*b*). These sequences are thought to be recognition sites for binding of the polymerase/leader complex during the synthesis of the subgenomic RNAs (Baric *et al.*, 1983). The same sequence CTGAACAA occurs at position 3293. Neither of these positions are known to be situated at the 5′ end of an mRNA species as are all the other homology regions. We have attempted to determine whether there is some feature of the sequence context surrounding these homology regions which sets them apart from homology regions which are known to occur at the 5′ end of the bodies of mRNAs. Accordingly, a consensus sequence has been calculated from the sequences surrounding the known homology regions at the ends of mRNAs A to F. This consensus sequence includes six bases to the left of the core homology

Fig. 4. (*a*) The nucleotide sequence in the region between F1 and F2, with a translation in single-letter amino acid code of three reading frames. The amino acid is shown above the second base of the appropriate codon. Stop codons are marked as asterisks. The frames which are open in F1 and F2 are underlined and the methionine at the start of F2 is boxed in. (*b*). A DNA sequencing gel obtained by sequencing a double-stranded cDNA clone using an oligonucleotide primer. The sequence shown is from 12333 to 12390, and is the reverse complement of the sequence shown in (*a*). (*c*) The same three reading frames as shown in (*a*), with a graph for each showing the extent to which that reading 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 as 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. The program used is the 'codon usage' option from ANALYSEQ (Staden, 1984*b*, 1983*c*) and uses the method of Staden & McLachlan (1982). The parameters used were a window length of 25 and an output length of 1. (Codon usage analysis from the spike, membrane and nucleocapsid gene data gives a very similar result.)

(*a*)



(*b*) Repeat A

```
F1 484  EFVKTYVCKAQMSIVILAAVLGEDIWHLVSQVIYKLGVLFTKVVDFC---DKHWKGFCVQLKRAKLIVTE
        :::::..::...      .:..  . .. ..:  :..  : ...:: .   . ..  ::: .... . ..
F2 1387 EFVKDFVCRNKQW---REAIF-ISPYNAMNQRAYRMLGLNVQTVDSSQGSEYDYVIFCVTADSQHALNIN

F1      TFCVLKGVAQHCFQLLLDAIHSLYKSFKKCALGR---IHGDLLF
         : :  . :.. . ... . ..::...:  .:..  ..:. ::
F2      RFNVALTRAKRGILVVMRQRDELYSALKFTELDSETSLQGTGLF
```

(*c*) Repeat B

```
F1 1630 VKMGDKIGGVTMGLWRAEHLNKPNLERIFNIAKKAIVGSSVVTTQCGKLIGKAATFIADKVGGGVVRNITD
        ::...:.  ..  .::. .  ...  .::..:: ..  ... ...  .   .. .:   : :  .::....
F2 2570 VKVSGKTLHANYIFWRNCNYLQTSAYSIFDVAKFDLRLKATPVVNLKTEQKTDLVFNLIKCGKLLVRDVGN
```

(*d*) Repeat C

```
F1 3696 VKTKACVAGVDQAHCSVESKCYYTNISGNSVVAAITSSNPN------LKVASFLNEAGN--QI
        ::..:..  ::.: : :..    .  ....  .. .:.        :: : ::.:: .  ::
F2 1996 VKPTAYAYVVDEA-CLVDDFVNLKYKAATPGKDSASSAVKCFSVTDFLKKAVFLKEALKCEQI
```

(*e*) Repeat D

```
F1 3438 LFCIDSTIDLSE-YCDDILKRSTVLQSVTQEFSHIPSYAEYERAKNLYEKVLDSKNG--GVT
        :::.. : .  : : .. .  : :. .  .. .  : ......:.  ::. :....   ..:
F2 430  LFCLEVTSKYFECYEGGCIPASQVVVNNLDKSAGYP-FNKFGKARLYYEMSLEEQDQLFEIT
```

Fig. 6. (*a*) 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 [see Results and (*b*) to (*e*)]. (*b* to *e*) Amino acid sequences of the matches depicted by the bars in (*a*). A colon shows identical amino acids and a dot shows similar (Kanehisa, 1982) amino acids. The dashes in the sequences are padding characters inserted to achieve optimal alignment.

region CT(T/G)AACAA present in all the regions, the eight bases of the core homology itself, and four bases to the right. The consensus has been compared to the complete sequence using the computer program FITCONSENSUS (Devereux *et al.*, 1984). The program successfully identifies the known homology regions with scores ranging from 74·6 to 64·1. The 14 next best fitting regions identified have a range of scores well separated from those of the known

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

Fig. 7. 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 (i.e. the 5′ end of the negative strand). Colons show identical bases. The dashes in the sequences are padding characters inserted to achieve optimal alignment.

```
 G A S D F D K N Y L N G Y G V A V R L G *
GGAGCATCTGATTTTGATAAGAATTATTTAAACGGGTACGGGGTAGCAGTGAGGCTCGGCTGATACCCCTTGCTAGTG
: :::   : :  :  :  :      :: :  :   :  :  : :::::: :::       ::   :
GTAGCTATGGTTAGAGGGAGTATCCTAGGAAGAGATTGTCTGCAGGGCCTAGGGCTCCGCTTGACAAATTTATAGGGA
 V A M V R G S I L G R D C L Q G L G L R L T N L *
```

Fig. 8. Nucleotide and predicted amino acid sequences where ribosomal frameshifting may occur. The top sequence is at the F1/F2 junction of IBV, and the bottom sequence is at the *gag/pol* junction of Rous sarcoma virus. Colons show identical bases.

homology regions, with a tight cluster of scores (53·6 to 58·8). The CTGAACAA sequence at position 599 scores even lower. It seems probable, therefore, that the two CTGAACAA sequences at 599 and 3293 are chance matches with the core sequence, but when surrounding sequences are taken into account the differences are enough to ensure that they are not major sites for the binding of the leader/polymerase complex.

DISCUSSION

The 20500 bases of sequence presented in this paper complete the sequence of the Beaudette strain of avian infectious bronchitis virus, the type species of the Coronaviridae. The complete sequence, excluding the poly(A) tail at the 3′ end, is 27608 residues. This is somewhat larger than the previously estimated size of the viral RNA which had been put at 20 to 24 kilobases (Lomniczi & Kennedy, 1977). The sequence of the 'unique' regions of mRNAs A, B, C, D and E have already been published, covering some 8 kilobases at the 3′ end of the genome and including the genes for the major structural proteins of the virus. The 20 kilobases at the 5′ end of the viral RNA constitutes the 'unique' region of mRNA F, the genome-sized RNA. This is thought to code for a polymerase or polymerases which carry out all the necessary replication and transcription functions of the virus.

Sequence analysis shows that the main part of the 'unique' region of mRNA F appears to contain two large ORFs. Because of the importance of determining whether there are one or two ORFs, we have considered the possibility that mRNA F in fact contained one very large ORF, and that a sequencing error or a mutant cDNA clone had led to a frameshift. Because of this the sequence in the region between the two ORFs has been checked exceedingly carefully. The relevant sequence is shown, with translations in the three reading frames, in Fig. 4(*a*). Any frameshift error must occur within 43 bases between positions 12341 and 12383. Two independent cDNA clones and direct RNA sequences from virion RNA give the same result. There are no obvious signs of sequence artefacts such as compressions, and indeed several gel systems and sequencing methods which could resolve compressions (see Methods and Results) do not show any change in the sequence. Fig. 4(*b*) shows a sequencing gel representing this region, obtained by sequencing a cDNA clone directly using an oligonucleotide primer. It can be seen that the sequence appears clear and unambiguous. Unless, therefore, there is some singular form of unresolvable and undetectable sequencing artefact, we must accept that the sequence here is correct.

The problem now arises as to how translation of the second ORF, F2, is achieved. No mRNA has been detected at this point, and no homology region which might suggest the presence of one can be seen in the RNA sequence (see Results). It is possible that the ribosomes, having completed translation of the first ORF, F1, reinitiate translation at the first AUG of F2, or that internal initiation occurs, as appears to be the case with the phosphoprotein mRNA of vesicular

stomatitis virus (Herman, 1986). There is however one piece of evidence that suggests that neither of these alternatives is the case. If the second ORF is genuinely a separate gene, then the 70 or so bases preceding its initiation codon should be non-coding sequences, comparable to the 5′ non-coding sequences preceding other IBV genes. In fact, if translated, they exhibit a heavy codon bias (Staden & McLachlan, 1982; Staden, 1984*c*) similar to the bias found in other IBV genes. This is shown graphically in Fig. 4(*c*) where it can be seen that the frame with typical IBV codon bias switches from that of F1 to that of F2 exactly at the point where the ORF changes. This strongly suggests that the sequences before the AUG of F2 have a coding function. One way to resolve this problem is to postulate that on some occasions, during translation of mRNA F, a ribosome slippage occurs, which introduces a frameshift and allows translation to continue unhindered from F1 into F2. 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. Such a mechanism could be conceived in the case of IBV as a form of translational control designed to provide coordinated expression of two polymerases, with the protein from the first gene being produced at a higher level than that from the second gene. In the case of Rous sarcoma virus (Jacks & Varmus, 1985) expression of the *pol* gene requires a frameshift by the ribosome. Some well-controlled work by these authors, using cell-free translation systems, has demonstrated that the frameshifting is sequence-specific. Moreover it occurs ten times more efficiently in a eukaryotic system than in a prokaryotic system, indicating that there are specific eukaryotic signals to which the prokaryotic system responds poorly. The region of sequence responsible for the frameshifting has been narrowed down to 24 nucleotides. Both IBV and Rous sarcoma virus require a shift into the -1 frame to occur, and it may be that similar frameshifting signals are present in both sequences. Accordingly the 24 nucleotides of Rous sarcoma virus sequence have been compared to the 43 nucleotides of IBV sequence within which any frameshift must occur (see Fig. 4*a*). Interestingly a match of 8/9 nucleotides can be found, both sequences occurring in the same frame and both within 20 bases of the termination codon (see Fig. 8). Further work will be needed to determine whether this sequence forms part of any signals which may promote ribosomal frameshifting.

For each of the other IBV mRNAs, the first AUG to occur after the homology region either is used to initiate synthesis of a protein, as is the case for the spike and membrane proteins (Binns *et al.*, 1985*b*; Boursnell *et al.*, 1984), or is present at the start of a reasonable sized ORF which could code for a polypeptide of 7K or more. Thus it is surprising to find the first AUG, at position 131, at the start of a small, 11 amino acid, ORF. The sequence context around this first AUG does not conform to Kozak's consensus for functional initiation codons whereas the context round the second AUG does. A similar small ORF of 12 amino acids occurs at the 5′ end of RNA 1 of alfalfa mosaic virus (Cornelissen *et al.*, 1983), an RNA species encoding a 115K product thought to be involved in RNA replication. In this case also only the second AUG conforms to the Kozak consensus. Both these cases suggest the possibility that the ribosomes can bypass the first, non-functional, AUG and initiate translation at the second. It is likely that this also occurs in mRNA D of IBV to allow translation of the second and third ORFs (Boursnell *et al.*, 1985*b*).

It is not known for coronaviruses whether the sequences at the 5′ end of the genome produce a polyprotein which is subsequently cleaved into separate proteins, as is the case for alphaviruses (Strauss *et al.*, 1984), or whether the viral polymerase acts as an extremely large multifunctional enzyme. Whether or not it is cleaved post-translationally into separate proteins, such an enzyme would need to perform several functions. First it must synthesize the negative-stranded template. From this template it must synthesize the leader sequence and then the subgenomic mRNAs, for which it needs the ability to recognize highly conserved signal sequences (Baric *et al.*, 1983, 1985; Spaan *et al.*, 1983; Brown & Boursnell, 1984), a capping ability (Lai *et al.*, 1982) and probably the ability to reinitiate transcription at these points (Lai *et al.*, 1985; Makino *et al.*, 1986). If it is cleaved into separate proteins it may encode a protease function to do this. Two polymerase activities, early and late, have been identified in MHV-infected cells (Brayton *et al.*, 1982). These have different ionic requirements and different pH optima. Both polymerase activities are associated with two different membrane fractions, a light fraction which appears

to synthesize positive-stranded genome-size RNA and a heavy fraction which also synthesizes subgenomic RNAs (Brayton *et al.*, 1984). Some evidence for two polymerase-coding genes can be found in the nucleotide sequence of mRNA F, in that there are small regions of residual homology between the predicted amino acid sequences of F1 and F2 (see Results and Fig. 6).

The question of whether the cDNA clones sequenced in this study might derive from mutant, non-viable RNA molecules is an interesting one. The error rate of RNA polymerases is fairly high (Steinhauer & Holland, 1986) and many of the RNA molecules in an infected cell may be different from that in the original infecting virus. If the mutation rate is 1 in 10000 then over the 20 kilobases of sequence presented here, there may be one or two changes each time one strand was copied into another. While the viral RNA is replicating within the cell, it is likely that mutant, and possibly defective, virion RNA molecules will accumulate with little selection against them, and, unless they have gross structural defects, most of them will be packaged into virions. It is these virions, without any further selection for viability, which are used to extract the RNA which is used to synthesize cDNA. In addition the infecting virus will be a mixture of different RNA molecules, even though it has been plaque-purified. However, be that as it may, there is no evidence for very high mutation rates in the cDNA clones which we have sequenced here. For the clones covering the 20 kilobases there are 4659 bases of overlap between separate, independent clones (all made from the same RNA preparation). In the overlap regions there was not one difference, there being 100% agreement between the sequences from adjacent clones.

This is in contrast to results found by Schubert *et al.* (1984) while sequencing the polymerase gene of vesicular stomatitis virus. The gene spans 6380 nucleotides and each region was sequenced from approximately three cDNA clones, giving 19140 nucleotides of overlap. In these 19140 nucleotides they found 20 nucleotide changes, including four insertions or deletions, giving an overall mutation rate of approximately 10^{-3}. In the 9318 (4659 × 2) nucleotides of IBV cDNA clones which can be checked on another clone, there were no changes. Over 9318 nucleotides a mutation rate of $3{\cdot}2 \times 10^{-4}$ would give a 95% probability of at least one nucleotide change; thus, since there were no changes, the overall mutation rate is probably lower than this. Given the number of rounds of replication which will have occurred between the original plaque isolation and the production of the cDNA clones, the mutation rate per base incorporated is likely to be considerably lower than this. It is interesting to speculate on the disparity between the vesicular stomatitis virus and the IBV results in this case, and on whether the (presumably) very large IBV polymerase, or polymerases, has a lower intrinsic error rate than the VSV polymerase.

Sequencing of cDNA clones from the 'unique' region of mRNA F has revealed the rather unexpected presence of two large ORFs. Although the sequence in the region between these has been obtained from three independent cDNA clones and from the virion RNA, the possibility of some bizarre form of sequence artefact cannot be totally discounted. It will be interesting to see if a similar frameshift occurs in an equivalent position in the coronavirus MHV genome. Experiments can now be designed to confirm the reading frame switch by other means. For example *in vitro* translation of SP6 polymerase transcripts from this region can be performed and the sizes of the products determined. Although no mRNA has been detected with a 5′ end near the beginning of the second ORF, a search for a low abundance mRNA species can now be carried out by primer extension from mRNA preparations. In addition, the availability of sequence data from the IBV polymerase(s) allows antisera to be raised against products expressed from selected parts of the sequence. These will prove useful in determining the fate of the large polypeptides predicted from the nucleotide sequence, showing whether post-translational cleavage occurs, and attempting to unravel the relationship between the various polymerase activities which have been detected in coronavirus-infected cells.

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 the gift of some deoxy-7-deazaguanosine triphosphate and for general advice and encouragement during the DNA sequencing.

REFERENCES

AMBARTSUMYAN, N. S. & 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. & GORINI, L. (1972). Low activity of beta-galactosidase in frameshift mutants of *Escherichia coli*. *Proceedings of the National Academy of Sciences, U.S.A.* **69**, 1192–1195.

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

BARIC, R. S., STOHLMAN, S. A. & LAI, M. M. C. (1983). Characterisation of replicative intermediate RNA of mouse hepatitis virus: presence of leader RNA sequences on nascent chains. *Journal of Virology* **48**, 633–640.

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

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

BIGGIN, M. D., GIBSON, T. J. & HONG, G. F. (1983). Buffer gradient gels and ^{35}S label as an aid to rapid DNA sequence determination. *Proceedings of the National Academy of Sciences, U.S.A.* **80**, 3963–3965.

BIGGIN, M., FARRELL, P. J. & BARRELL, B. G. (1984). Transcription and DNA sequence of the *Bam*HI L fragment of B95-8 Epstein–Barr virus. *EMBO Journal* **3**, 1083–1090.

BINNS, M. M., BOURSNELL, M. E. G., FOULDS, I. J. & BROWN, T. D. K. (1985*a*). The use of a random priming procedure to generate cDNA libraries of infectious bronchitis virus, a large RNA virus. *Journal of Virological Methods* **11**, 265–269.

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

BOURSNELL, M. E. G., BINNS, M. M. & BROWN, T. D. K. (1985*b*). Sequencing of coronavirus IBV genomic RNA: three open reading frames in the 5′ 'unique' region of mRNA D. *Journal of General Virology* **66**, 2253–2258.

BRAYTON, P. R., LAI, M. M. C., PATTON, C. D. & STOHLMAN, S. A. (1982). Characterisation of two polymerase activities induced by mouse hepatitis virus. *Journal of Virology* **42**, 847–853.

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 genome RNA contains sequence homologies at the intergenic boundaries. *Virus Research* **1**, 15–24.

BROWN, T. D. K., BOURSNELL, M. E. G., BINNS, M. M. & TOMLEY, F. M. (1986). Cloning and sequencing of 5′ terminal sequences from avian infectious bronchitis virus genomeic RNA. *Journal of General Virology* **67**, 221–228.

CATON, A. J., BROWNLEE, G. G., YEWDELL, J. W. & GERHARD, W. (1982). The antigenic structure of the influenza virus A/PR/8/34 hemagglutinin (H1 subtype). *Cell* **31**, 417–427.

CAVANAGH, D. (1981). Structural polypeptides of coronavirus IBV. *Journal of General Virology* **53**, 93–103.

CORNELISSEN, J. C., BREDERODE, F. T., MOORMANN, R. J. M. & BOL, J. F. (1983). Complete nucleotide sequence of alfalfa mosaic virus RNA 1. *Nucleic Acids Research* **11**, 1253–1265.

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

DEVEREUX, J., HAEBERLI, P. & SMITHIES, O. (1984). A comprehensive set of sequence analysis programs for the VAX. *Nucleic Acids Research* **12**, 387–395.

FOX, T. D. & WEISS-BRUMMER, B. (1980). Leaky +1 and −1 frameshift mutations at the same site in a yeast mitochondrial gene. *Nature, London* **288**, 60–63.

GEORGE, D. G., BARKER, W. C. & HUNT, L. T. (1986). The protein identification resource (PIR). *Nucleic Acids Research* **14**, 11–15.

GERMINO, J. & BASTIA, D. (1982). Primary structure of the replication initiation protein of plasmid R6K. *Proceedings of the National Academy of Sciences, U.S.A.* **79**, 5475–5479.

HERMAN, R. C. (1986). Internal initiation of translation on the vesicular stomatitis virus phosphoprotein mRNA yields a second protein. *Journal of Virology* **58**, 797–804.

HONG, G. F. (1981). A method for sequencing single-stranded cloned DNA in both directions. *Bioscience Reports* **1**, 243–252.

JACKS, T. & VARMUS, H. E. (1985). Expression of the Rous sarcoma virus *pol* gene by ribosomal frameshifting. *Science* **230**, 1237–1242.

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

KASTELEIN, R. A., REMAUT, E., FIERS, W. & VAN DUIN, J. (1982). Lysis gene expression of RNA phage MS2 depends on a frameshift during translation of the overlapping coat protein gene. *Nature, London* **295**, 35–41.

KORNELUK, R. G., QUAN, F. & 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. *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.

LAI, M. M. C., PATTON, C. D. & STOHLMAN, S. A. (1982). Replication of mouse hepatitis virus: negative-stranded RNA and replicative form RNA are of genome length. *Journal of Virology* **44**, 487–492.

LAI, M. M. C., BARIC, R. S., MAKINO, S., KECK, J. G., EGBERT, J., LEIBOWITZ, J. L. & STOHLMAN, S. A. (1985). Recombination between nonsegmented RNA genomes of murine coronaviruses. *Journal of Virology* **56**, 449–456.

LEIBOWITZ, J. L., WILHELMSEN, K. C. & BOND, C. W. (1981). The virus-specific intracellular RNA species of two murine coronaviruses: MHV-A59 and MHV-JHM. *Virology* **114**, 39–51.

LEIBOWITZ, J. L., WEISS, S. R., PAAVOLA, E. & BOND, C. W. (1982). Cell-free translation of murine coronavirus RNA. *Journal of Virology* **43**, 905–913.

LIPMAN, D. J. & PEARSON, W. R. (1985). Rapid and sensitive protein similarity searches. *Science* **227**, 1435–1441.

LOMNICZI, B. (1977). Biological properties of avian coronavirus RNA. *Journal of General Virology* **36**, 531–533.

LOMNICZI, B. & KENNEDY, I. (1977). Genome of infectious bronchitis virus. *Journal of Virology* **24**, 99–107.

MAKINO, S., STOHLMAN, S. A. & LAI, M. M. C. (1986). Leader sequences of murine coronavirus mRNAs can be freely reassorted: evidence for the role of free leader RNA in transcription. *Proceedings of the National Academy of Sciences, U.S.A.* **83**, 4204–4208.

MAXAM, A. M. & GILBERT, W. (1980). Sequencing end-labeled DNA with base-specific chemical cleavages. *Methods in Enzymology* **65**, 499–560.

MIZUSAWA, S., NISHIMURA, S. & SEELA, F. (1986). Improvement of the dideoxy chain termination method of DNA sequencing by use of deoxy-7-deazaguanosine triphosphate in place of dGTP. *Nucleic Acids Research* **14**, 1319–1324.

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.

SCHOCHETMAN, G., STEVENS, R. H. & SIMPSON, R. W. (1977). Presence of infectious polyadenylated RNA in the coronavirus avian bronchitis virus. *Virology* **77**, 772–782.

SCHUBERT, M., HARMISON, G. G. & MEIER, E. (1984). Primary structure of the vesicular stomatitis virus polymerase (L) gene: evidence for a high frequency of mutations. *Journal of Virology* **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. & VAN DER ZEIST, B. A. M. (1983*a*). Coronaviridae. *Intervirology* **20**, 181–189.

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

SOUTHERN, E. M. (1975). Detection of specific sequences among DNA fragments separated by gel electrophoresis. *Journal of Molecular Biology* **98**, 503–517.

SPAAN, W., DELIUS, H., SKINNER, M., ARMSTRONG, J., ROTTIER, P., SMEEKENS, S., VAN DER ZEIJST, B. A. M. & SIDDELL, S. G. (1983). Coronavirus mRNA synthesis involves fusion of non-contiguous sequences. *EMBO Journal* **2**, 1839–1844.

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

STADEN, R. (1982*b*). Automation of the computer handling of gel reading data produced by the shotgun method of DNA sequencing. *Nucleic Acids Research* **10**, 4731–4751.

STADEN, R. (1984*a*). A computer program to enter DNA gel reading data into a computer. *Nucleic Acids Research* **12**, 499–503.

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

STADEN, R. (1984*c*). Measurements of the effects that coding for a protein has on a DNA sequence and their use for finding genes. *Nucleic Acids Research* **12**, 551–567.

STADEN, R. & MCLACHLAN, A. D. (1982). Codon preference and its use in identifying protein coding regions in long DNA sequences. *Nucleic Acids Research* **10**, 141–157.

STEINHAUER, D. A. & HOLLAND, J. J. (1986). Direct method for quantitation of extreme polymerase error frequencies at selected single base sites in viral RNA. *Journal of Virology* **57**, 219–228.

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.

STRAUSS, E. G. & STRAUSS, J. H. (1983). Replication strategies of the single stranded RNA viruses of eukaryotes. *Current Topics in Microbiology and Immunology* **105**, 1–98.

STRAUSS, E. G., RICE, C. M. & STRAUSS, J. H. (1984). Complete sequence of the genomic RNA of Sindbis virus. *Virology* **133**, 92–110.

(Received 21 August 1986)