Coronaviruses: Structure and Genome Expression

By W. SPAAN, D. CA VANAGH AND M. C. HORZINEK

Institute of Virology, Veterinary Faculty, State University of Utrecht, Yalelaan 1, 3508 TD Utrecht, The Netherlands and AFRC Institute for Animal Health, Houghton Laboratory, Huntingdon, Cambridgeshire PE17 2DA, U.K.

INTRODUCTION

Progress in coronavirology is illustrated by the number of workshops convened and reviews written. International meetings have been held in Germany (1980), the Netherlands (1983) and the U.S.A. (1986), and the Fourth Coronavirus Symposium will be organized by one of us (D.C.) in Cambridge, U.K. in July 1989. In addition, reviews have appeared which highlighted particularly interesting characteristics of the family, e.g. the replication strategy (Lai, 1986) and the glycoproteins (Sturman & Holmes, 1985). As the last general accounts were published some 5 years ago (Siddell et al., 1983; Sturman & Holmes, 1983) an update is timely. The present article is based on the large amount of sequence data accumulated in these years and focuses on the viral nucleic acids and proteins and their function.

Coronaviruses cause infections in man, other mammals and birds. Most experimental data have been obtained from studies of mouse hepatitis virus (MHV) and infectious bronchitis virus of chickens (IBV). Additional representatives of the family reviewed in this article are the human (HCV) and bovine (BCV) coronaviruses, transmissible gastroenteritis virus (TGEV), haemagglutinating encephalitis virus (HEV) and feline infectious peritonitis virus (FIPV).

VIRION PROTEINS

Coronaviruses possess three major structural proteins: a nucleocapsid protein (N), a small integral membrane glycoprotein (M, E1) and a large spike glycoprotein (S, E2); for the sake of uniformity we use the letters N, M and S in this review. While all coronaviruses contain these proteins, a subset (HEV, HCV-OC43 and BCV) is now recognized to possess an additional glycopolypeptide (gp65), which is unrelated to S or M.

N protein

The number of amino acids in the N protein has been determined by cloning and sequencing for MHV strains A59 (Armstrong et al., 1983) and JHM (Skinner & Siddell, 1983), IBV strains Beaudette and M41 (Boursnell et al., 1985a), for TGEV (Kapke & Brian, 1986) and BCV (Lapps et al., 1987; Table 1). These proteins are basic, the basic residues occurring in clusters; the C terminus is acidic. Serine residues account for 8 to 10% of the total number of amino acids; their clustering may correlate with the fact that N is phosphorylated specifically on serines. The homology of the BCV N protein with that of MHV is 70% (72% base homology), of TGEV 29% (37% base homology) and of IBV 29% (43% base homology; Kapke & Brian, 1986; Lapps et al., 1987). One prominent region of homology is a stretch of about 68 amino acids, which exhibits between 51 and 79% similarity depending on the pairs of viruses compared (Kapke & Brian, 1986; Lapps et al., 1987).

M glycoprotein

As is the case with the N protein, the M glycoprotein of the various coronaviruses also exhibits different M, values in polyacrylamide gels (see review by Siddell et al., 1983; also Resta et al., 1985; Hogue & Brian, 1986; Sugiyama et al., 1986; Cavanagh & Davis, 1987). These variations are not only due to differences in the number of amino acid residues, but also to the extent of
Table 1. Properties of virion proteins

<table>
<thead>
<tr>
<th>Protein</th>
<th>IBV</th>
<th>MHV</th>
<th>FIPV</th>
<th>TGEV</th>
<th>BCV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nucleocapsid protein (N)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( M_r \times 10^3 )</td>
<td>45</td>
<td>50</td>
<td>-</td>
<td>43</td>
<td>49</td>
</tr>
<tr>
<td>No. of amino acids</td>
<td>409</td>
<td>455</td>
<td>-</td>
<td>382</td>
<td>448</td>
</tr>
<tr>
<td>Matrix glycoprotein (M)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( M_r \times 10^3 )</td>
<td>25</td>
<td>26</td>
<td>-</td>
<td>30</td>
<td>26</td>
</tr>
<tr>
<td>No. of amino acids</td>
<td>224</td>
<td>227</td>
<td>-</td>
<td>245</td>
<td>229</td>
</tr>
<tr>
<td>Total, mature protein</td>
<td>21</td>
<td>24</td>
<td>-</td>
<td>29</td>
<td>23</td>
</tr>
<tr>
<td>Hydrophilic N terminus</td>
<td>76</td>
<td>81</td>
<td>-</td>
<td>88</td>
<td>81</td>
</tr>
<tr>
<td>Membrane-embedded domain</td>
<td>127</td>
<td>122</td>
<td>-</td>
<td>128</td>
<td>125</td>
</tr>
<tr>
<td>Luminal C terminus</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>No. potential glycosylation sites</td>
<td>1, 2</td>
<td>3</td>
<td></td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>Spike protein (S)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( M_r \times 10^3 )</td>
<td>128</td>
<td>137</td>
<td>159</td>
<td>158</td>
<td>-</td>
</tr>
<tr>
<td>No. of amino acids†</td>
<td>1162</td>
<td>1235†</td>
<td>1452</td>
<td>1447</td>
<td>-</td>
</tr>
<tr>
<td>Total, with signal</td>
<td>537</td>
<td>628‡</td>
<td>NA‡</td>
<td>NA‡</td>
<td>-</td>
</tr>
<tr>
<td>S1, including signal sequence</td>
<td>717§</td>
<td>607†</td>
<td>NA§</td>
<td>NA§</td>
<td>-</td>
</tr>
<tr>
<td>S2</td>
<td>625</td>
<td>607†</td>
<td>NA§</td>
<td>NA§</td>
<td>-</td>
</tr>
<tr>
<td>Signal sequence</td>
<td>18</td>
<td></td>
<td>16</td>
<td></td>
<td>-</td>
</tr>
<tr>
<td>No. potential glycosylation sites</td>
<td>28</td>
<td>21</td>
<td>35</td>
<td>32</td>
<td>-</td>
</tr>
<tr>
<td>gp65/130</td>
<td>-</td>
<td>JHM +/A59 -</td>
<td>-</td>
<td>-</td>
<td>+</td>
</tr>
</tbody>
</table>

* Information not available.
† Approximate, because of strain variation.
‡ MHV-JHM.
§ MHV-A59.
∥ NA, Not applicable; spike not cleaved.

glycosylation, the type of linkage of the glycans (O-linked or N-linked) and the degree to which N-linked high-mannose (simple) glycans have been converted to complex glycans (Table 1). Nucleotide sequencing of the M gene of MHV-A59 and MHV-JHM (Armstrong et al., 1984; Pfleiderer et al., 1986), IBV-Beaudette and IBV-6/82 (Boursnell et al., 1984; Binns et al., 1986a), BCV (Lapps et al., 1987) and TGEV (Laude et al., 1987) has revealed many interesting features of the M glycoprotein. Computer predictions of its secondary structure have led to a model in which approximately 10% of the N-terminal part of the molecule is exposed on the outer surface of the virus membrane (see references above and Rottier et al., 1986). This view is supported by experimental evidence (Rottier et al., 1984; Cavanagh et al., 1986a). The next 80 or so residues, approximately one-third of the molecule, form three hydrophobic \( \alpha \)-helices, which span the membrane three times. The C-terminal half of the protein has neither strong hydrophobic nor hydrophilic properties and is located in the interior of the virus particle.

Although the M protein of MHV, IBV and BCV does not possess an N-terminal signal sequence, M needs the signal recognition particle for membrane insertion, like many secretory and membrane proteins (Rottier et al., 1985). The first (amino-terminal) and/or third membrane-spanning helices can function as signal sequences (Machamer & Rose, 1987; Mayer et al., 1988). However, this feature is not universal among the coronaviruses; a putative N-terminal signal peptide of 17 residues has been identified for the TGEV M protein (Laude et al., 1987). The MHV M protein shares 86% homology with BCV but only 35% and 38% with IBV and TGEV, respectively (Lapps et al., 1987; Laude et al., 1987).

Previous studies (see Siddell et al., 1983) have shown that the glycans of MHV (Niemann et al., 1984), and of BCV are of the O-linked type, in contrast to the N-linked glycans of IBV and TGEV (Stem & Sefton, 1982a; Cavanagh, 1983a; Garwes et al., 1984; Jacobs et al., 1986; Cavanagh & Davis, 1987, 1988). Most M protein glycan molecules of IBV are of the simple type but a proportion are converted to complex glycans (Stem & Sefton, 1982a; Cavanagh, 1983a).
S protein

The S protein gene has been sequenced for several strains of IBV (Binns et al., 1985; Niesters et al., 1986; Binns et al., 1986b), MHV-JHM (Schmidt et al., 1987) and MHV-A59 (Luytjes et al., 1987), TGEV (Rasschaert & Laude, 1987; Jacobs et al., 1987) and FIPV (de Groot et al., 1987b). These S glycoproteins possess an overall hydrophobic hydrophobicity profile, an N-terminal signal sequence, a C-terminal hydrophilic sequence preceded by a membrane-spanning domain, and a large number of potential N-linked glycosylation sites, most of which appear to be glycosylated (Table 1). In contrast, there are large differences in the number of amino acids of S between different coronaviruses and different isolates of a given virus (see MHV in Table 1; also Taguchi et al., 1985).

The extent to which S is cleaved into S1 and $2 depends on the virus in question and the cells infected. Most authors have reported the S protein of IBV (Stern & Sefton, 1982b; Cavanagh, 1983b, c; Cavanagh & Davis, 1987) and BCV (Hogue et al., 1984; Deregt et al., 1987) to be in a cleaved form, but no S protein of TGEV, FIPV or canine coronavirus (Garwes & Reynolds, 1981; Horzinek et al., 1982) and only a little of HCV (Hogue & Brian, 1986) occurs cleaved. Cleavage of MHV S protein varies from 0 to 100%, depending on the virus strain and cell type (Sturman et al., 1985; Sugiyama et al., 1986).

The order of S1 and $2 within the S0 protein of IBV and MHV is: N terminus–S1–S2–C terminus (in MHV 90B and 90A are equivalent to S1 and S2, respectively; Binns et al., 1985; Cavanagh et al., 1986b; Luytjes et al., 1987). For IBV and MHV the cleavage site is adjacent to the amino acid sequence RRFRR, RRSRR or RRHRR (IBV, eight strains sequenced; Cavanagh et al., 1986b; Binns et al., 1986b; J. G. Kusters et al., unpublished data), RRAHR (MHV-A59; Luytjes et al., 1987) and RRARR (MHV-JHM; Schmidt et al., 1987). As would be expected from such basic sequences, S0 of both IBV and MHV can be cleaved in vitro by trypsin and, in the case of S0 with the connecting peptide RRFRR, by chymotrypsin (Sturman et al., 1985; Frana et al., 1985; Cavanagh et al., 1986a). The S0 protein of TGEV and FIPV lacks such pairs of basic residues.

Sedimentation studies have indicated that the peplomer of IBV is an oligomer comprising two or three molecules of S1 + S2 (Cavanagh, 1983c). Interpeptide disulphide bonds are not involved in maintaining quaternary structure, permitting S1 to be removed from virions by urea treatment, with S2 left in place. This led to the proposal that the outer, bulbous part of S might be formed largely by S1, with S2 being anchored by its C terminus in the virus envelope (Cavanagh, 1983c). Computer-aided analysis of the S protein primary structure of IBV, MHV and FIPV (de Groot et al., 1987a) has identified two heptad repeats in the C-terminal domain, which indicate an intra-chain coiled coil structure. These results have been confirmed for TGEV (Rasschaert & Laude, 1987). The major repeat suggests a helix occupying more than half the length of the peplomer. In the oligomer the major helices are probably involved in an inter-chain coiled coil, reminiscent of structures in the haemagglutinin glycoprotein trimer of influenza virus. The $2 protein of IBV has been found to be susceptible to hydrolysis by several proteases within a region adjacent to the N-terminal side of the membrane-spanning hydrophobic domain (Cavanagh et al., 1986a).

Comparison of the peplomer protein amino acid sequences of IBV, MHV and FIPV has shown homologies in the S2 half of the molecule of 35, 30 and 29% for IBV–FIPV, IBV–MHV and MHV–FIPV, respectively (de Groot et al., 1987a). In contrast, S1 (or the equivalent region in FIPV) exhibits very little conservation. In S2 several regions of more than 30% homology, including sequences of seven to ten identical amino acids, have been identified (Schmidt et al., 1987; Rasschaert & Laude, 1987). The unusually high number of cysteine residues in the vicinity of the transmembrane domain is also conserved in the S proteins sequenced so far. Some of these residues may be involved in S2 acylation (Sturman et al., 1985) which can occur in the absence of glycosylation (Van Berlo et al., 1987).

gp65 glycopolypeptide

Early studies have shown that HCV-229E, HCV-OC43, HEV, MHV-JHM and BCV possess a glycopolypeptide in addition to S and M (see Siddell et al., 1983; Makino et al., 1983). More
recent studies with BCV (King et al., 1985; Deregt et al., 1987), HCV-OC43 (Hogue & Brian, 1986) and diarrhoea virus of infant mice (Sugiyama et al., 1986) have confirmed that these viruses contain a glycopolypeptide of about 65K (gp65) which, in the absence of mercaptoethanol, runs as a dimer of 130K to 140K (gpl30-140) in PAGE. Siddell (1982) has shown by tryptic peptide fingerprinting that gp65 is structurally unrelated to S.

**Functions of the virion proteins**

In addition to its role in encapsidating genomic RNA and facilitating its incorporation into virions (by the formation of ribonucleoprotein, RNP), the N protein has been implicated in the process of RNA replication. Addition of antisera raised against N (but not against S or M) to an in vitro replication system inhibited the synthesis of genome-sized RNA by 90% (Compton et al., 1987). In the presence of tunicamycin, MHV and IBV formed virions which lacked S but contained M and RNP (Holmes et al., 1981; Rottier et al., 1981; Stern & Sefton, 1982a). In infected cells virion budding occurs at the site of M accumulation (Tooze et al., 1984). These data indicate that M is necessary for virus maturation and that it determines the site at which virus particles are assembled. The intracellular accumulation of the M protein is a property of the protein itself (Rottier & Rose, 1987; Machamer & Rose, 1987). A mutant M protein of IBV possessing only the first transmembrane domain accumulated intracellularly, while another one with only the third domain was transported to the plasma membrane (Machamer & Rose, 1987). The domain of M that interacts with the RNP is not known; it has been shown that M itself has an affinity for RNA (Sturman et al., 1980).

It has been assumed that attachment of virions to cells is mediated by the S protein or by both S and gp65 (in HEV, BCV and HCV-OC43, which possess the latter). Recently, a 110K plasma membrane glycoprotein with an affinity for MHV-A59 was described, and its presence was correlated with the virus susceptibility of target cells (Boyle et al., 1987). Neuraminidase and glycosidase treatment did not destroy its S protein-binding properties. The host cell protein probably serves as a virus receptor as cells were protected against infection after incubation with homologous polyclonal and monoclonal antibody (MAb) (K. Holmes, personal communication). Spikeless virions of IBV and MHV are non-infectious (Cavanagh, 1981; Stern & Sefton, 1982a; Holmes et al., 1981; Rottier et al., 1981). Removal from IBV of S1 (but not of S2) by urea abolished both infectivity and haemagglutinating activity (HA; Cavanagh & Davis, 1986). Although this result indicated that attachment to erythrocytes had been affected qualitatively, the amount of virus that attached to red blood cells and chicken embryo kidney cells was not reduced. The inference that HA is mediated by S1 is supported by the finding that MAbS to S1 inhibit HA (Mockett et al., 1984). In those coronaviruses which possess gp65 it is this protein, probably as a dimer, which is associated with HA (see Siddell et al., 1983; King et al., 1985). Haemagglutination studies indicated that HCV-OC43 and BCV recognize O-acetylated sialic acid or a similar derivative as the red blood cell receptor, as do influenza C viruses (Vlasak et al., 1988). It will be necessary to examine whether viral binding in vivo also involves modified sialic acids. BCV also exhibits an acetylesterase receptor-destroying activity similar to the enzymic activity found in influenza C viruses (Vlasak et al., 1988). This activity is associated with the gp65 spike protein (R. Vlasak, W. Luytjes, W. Spaan & P. Palese, unpublished observations).

The failure of the virus lacking S1 to replicate despite attachment to cells suggested that some other function had been lost, most probably the fusion activity. The S protein induces membrane fusion, and this has been demonstrated in three ways: first, antiserum to S but not to M (Sturman et al., 1985) and MAbS specific for S (Collins et al., 1982; Wege et al., 1984) inhibited cell fusion after infection with MHV (fusion from within); second, MHV-A59 grown in the 17 Cl 1 line of spontaneously transformed BALB/c 3T3 cells did not cause fusion of L cells when added at high multiplicity (fusion from without) unless S had previously been cleaved by trypsin (Sturman et al., 1985); third, vaccinia virus recombinants containing either the MHV S gene or the FIPV S gene were able to induce cell fusion (H. Vennema, L. Heijnen & W. Spaan, unpublished observation). Although S is responsible for the induction of membrane fusion, its cleavage is not always a necessary precondition. In FIPV and TGEV, a cleaved S protein has not been observed, yet replication results in syncytium formation (de Groot et al., 1987c). Similarly,
HCV-OC43 and a small plaque variant of MHV-A59 (Sawicki, 1987) possess uncleaved S but nevertheless undergo multiple cycles of replication. The requirement of S cleavage for membrane fusion may therefore depend not only on the virus strain and host cell membrane properties (Frana et al., 1985) but also on the type of fusion in question: from within, from without or virus–endosome fusion. Concentrated MHV caused rapid fusion from without, the optimum pH being above 7 (Sturman et al., 1985; Frana et al., 1985). In contrast, agents that increased endosomal pH adversely affected MHV replication (Krzyżystyniak & Dupuy, 1984; Mizzen et al., 1985). Clearly more work is required to define the pH requirements for S protein activity in fusion.

**GENOMIC RNA**

*Genome organization*

The genomic RNA of coronaviruses is the largest among RNA viruses, approximately 27 to 30 kb. The genome is organized into six or seven regions, each containing one or more open reading frames (ORFs) which are separated by junction sequences that contain the signal(s) for the transcription of multiple subgenomic mRNAs.

The organization of the coronaviral genome was elucidated on the basis of sequence relationships between the subgenomic RNAs and of *in vitro* translation studies using the individual mRNAs (reviewed by Siddell et al., 1983; Siddell, 1983; Stern & Sefton, 1984; de Groot et al., 1987c). In recent years much sequence data have been obtained; amongst others, the complete sequence of IBV genomic RNA, and the sequences of approximately 12 kb and 8.3 kb (extending from the 3’ end) of the MHV and TGEV genomic RNAs, respectively (Fig. 1; for references see legend).

From the sequence data several ORFs could be deduced. The 5’ two-thirds of the IBV genome (region F) encode non-structural protein(s), probably the replicase/transcriptase. The ORFs encoding the structural viral proteins have been identified; from their location a consensus gene order 5’ S–M–N 3’ can be inferred. The number and location of the other ORFs are different between IBV, MHV and TGEV (Fig. 1).

*Translation strategy*

Expression of the genome of coronaviruses involves the production of seven (BCV), six (MHV and TGEV) or five (IBV and FIPV) subgenomic mRNAs, which together with the virion RNA form a 3’ cotermination nested set (reviewed by Siddell et al., 1983; Keck et al., 1988a; de Groot et al., 1987c; Rasschaert et al., 1987). Except for the smallest subgenomic mRNA, the virus-specific intracellular mRNAs are polygenic, but only the unique region of each mRNA is translationally active. The unique regions of the mRNAs encoding the N, M and S proteins of IBV, TGEV and MHV (see references in the section on virion proteins), the non-structural 15K protein of MHV (Skinner & Siddell, 1985) and mRNAs 4 and 7 of TGEV (Rasschaert et al., 1987) each comprise only one ORF. In contrast, two ORFs are present in the unique region of the genomic RNA (Boursnell et al., 1987) and of mRNA B (Boursnell & Brown, 1984) of IBV, mRNAs 2 (Luytjes et al., 1988) and 5 (Skinner et al., 1985; Budzilowicz & Weiss, 1987) of MHV and in mRNA 3 of TGEV (Rasschaert et al., 1987), while three ORFs have been identified at the unique 5’ end of IBV mRNA D (Boursnell et al., 1985b). These mRNAs may therefore be functionally polycistronic.

ORFs F1 and F2 which are present in the IBV genome-size mRNA have the capacity to encode 400K and 350K polypeptides, respectively; they overlap by 42 nucleotides, but F2 starts in a different reading frame (Boursnell et al., 1987). A cDNA fragment spanning the F1/F2 overlap was able to direct ribosomal frameshifting *in vitro* (Brierly et al., 1987); it remains to be determined whether the same mechanism operates *in vivo*.

The third ORF of IBV mRNA D has been expressed, and an antiserum was raised against its product, D3, a polypeptide of Mr 12.4K (Smith et al., 1987). Immunoprecipitation has shown that the D3 polypeptide occurs in IBV-infected cells, is unglycosylated, membrane-associated and co-fractionates with virions in sucrose gradients. A molar ratio of 1:2:2:7:11 has been estimated for D3:S1:S2:N:M in virion preparations (A. R. Smith, M. E. G. Boursnell, M. M.
Fig. 1. The organization and expression of the MHV, IBV and TGEV genomes. The genomes are presented from 5' (left) to 3' (right) by horizontal lines; short vertical lines show the location of the conserved junction sequences that contains the signal(s) for the transcription of multiple subgenomic mRNAs. The numbers (1 to 7) and letters (A to F) below the junction regions indicate the nomenclature of the corresponding mRNAs for MHV or TGEV and IBV, respectively. The mRNAs (not indicated) form a 3' coterminal nested set and their length can be deduced from the nucleotide kb map at the bottom of the figure. The symbol \( \text{A}_e \) indicates the poly(A) tract at the 3' end of the genome. ORFs are shown as open rectangles, and the position on the genome of the ORFs encoding the replicase/transcriptase, the peplomer, the small integral membrane and nucleocapsid protein are indicated by F1 and F2, S, M and N respectively. The rectangles which include numbers (1 to 3) indicate that the unique regions (enclosed by the junction sequences) of the corresponding mRNAs contain more than one ORF (for details see text). For IBV the data have been compiled from Boursnell et al. (1987), for MHV from Armstrong (1983, 1984), Skinner & Siddell (1983, 1985), Skinner et al. (1985), Budzilowicz & Weiss (1987), Schmidt et al. (1987), Luytjes et al. (1987, 1988), P. J. Bredenbeek & W. Spaan (unpublished data), and for TGEV from Rasschaert et al. (1987).

Binns, T. D. K. Brown & S. C. Inglis, personal communication). Whether the other two small ORFs of mRNA D remain silent is unknown, but the unusual codon usage and the presence of weak translation initiation codons (Kozak, 1987) suggest that they are translationally inactive.

In vitro translation of the MHV genomic RNA has revealed the synthesis of a 250K polyprotein which is subsequently cleaved into a p28 and a p220 (28K and 220K respectively) protein (Denison & Perlman, 1986). The p28 protein could be labelled with \( N \)-formyl-S\[^{[35]S}\]methionyl tRNA indicating its N-terminal location; the protein was also identified in infected cells (Denison & Perlman, 1987). Tryptic peptide maps of p28 synthesized \textit{in vitro} by translation of genomic RNA and of RNA transcribed from a cDNA clone covering 1-1 kb of the 5' end of the MHV genome were similar, confirming that the p28 protein is the N-terminal cleavage product of the putative MHV RNA polymerase (Soe et al., 1987). Except for this cleavage and that of some spike protein precursors into two similarly sized subunits no processing of polyproteins has so far been observed in coronaviruses.

In vitro translation of MHV mRNA 2 gave rise to a 30K polypeptide (Siddell, 1983). Sequence analysis of the unique region of mRNA 2 of MHV strain A59 has revealed two ORFs (ORF 1 and ORF 2; Luytjes et al., 1988). A 30K protein has been predicted on the basis of the ORF 1 sequence and was detected in MHV-infected cells by antiserum raised against an expression product of ORF 1 (P. Bredenbeek, A. Noten & W. Spaan, unpublished observations). The second reading frame has the potential to encode a 43K protein, but it is unlikely to be expressed.
since in the first 109 triplets an AUG codon is lacking. Strikingly, the predicted amino acid sequence of ORF 2 has 30% homology with the haemagglutinin subunit 1 of influenza C virus. In strain JHM an ORF of which the predicted amino acid sequence is almost identical to ORF 2 of MHV-A59 and which includes an AUG translation start codon has been identified (E. Routledge & S. Siddell, personal communication).

An antiserum raised against the carboxy-terminal region of the mRNA 4 ORF product of MHV-JHM reacted specifically with a 15K protein synthesized in JHM-infected cells (Ebner et al., 1988). In vitro translation of RNA transcribed from cDNA containing both ORFs of MHV-A59 mRNA 5 indicated that the downstream reading frame is preferentially translated (Budzilowicz & Weiss, 1987). This ORF was shown to be expressed in infected cells (Leibowitz et al., 1988).

In conclusion, coronaviruses have developed more than one strategy to produce their proteins. Subgenomic mRNAs are synthesized to position internal genes at the unique 5' end of an mRNA. Though most of these mRNAs are functionally monocistronic, several do contain more than one ORF within the unique region. In these mRNAs internal initiation occurs, e.g. at the AUG codon of the second ORF of MHV mRNA 5 and the third ORF of IBV mRNA D. Finally, the expression of the 5' two-thirds of the IBV and MHV genomes involves ribosomal frameshifting and post-translational cleavage of a polyprotein, respectively.

Transcription and replication

The coronavirus genome serves as a template for the synthesis of a full-length negative-strand RNA (Lai et al., 1982). This RNA was found exclusively in the viral replicative intermediates (RI) and the rate of its synthesis declined 5 to 6 h after infection (Sawicki & Sawicki, 1986). Continuing protein synthesis is a prerequisite for both negative and positive-strand RNA synthesis, although the former is comparatively more sensitive to cycloheximide (Sawicki & Sawicki, 1986). The negative-strand RNA serves as a template for transcription of genomic RNA and subgenomic mRNAs. The viral proteins involved in positive- and negative-strand RNA synthesis have not yet been identified. Six complementation groups are involved in RNA synthesis (Leibowitz et al., 1982; B. A. M. van der Zeijst, personal communication). An RNA-dependent RNA polymerase activity has been detected in TGEV- and MHV-infected cells (Dennis & Brian, 1982; Brayton et al., 1982; Mahy et al., 1983). Brayton et al. (1982, 1984) have described two enzymically distinct RNA polymerase activities (acting early and late during infection); the early polymerase was involved in negative-strand RNA synthesis, whereas the late polymerase synthesized positive-stranded RNA. In contrast, only one polymerase activity was detected by Mahy et al. (1983) and Compton et al. (1987). Inhibition of in vitro RNA transcription by antibody to the N protein suggests a role in coronavirus replication (Compton et al., 1987); whether this inhibition reflects a direct function of N in RNA synthesis or is the result of rapid RNA degradation is not known. Since the synthesis of mRNA 7 and of N protein is not inhibited in cells infected with defective interfering particles, in contrast to the other mRNAs and their translation products (Makino et al., 1985), the N protein may well have a role in the replication of viral RNA.

The subgenomic mRNAs which form a 3' coterminial nested set are synthesized in non-equimolar, but constant, amounts during the replication cycle (reviewed by Siddell et al., 1983). RNase T1 fingerprinting of MHV genomic and subgenomic mRNAs revealed unique oligonucleotides that do not fit into the nested set structure, suggesting that these oligonucleotides are derived from a leader sequence which all mRNAs might share (Spaan et al., 1982; Lai et al., 1983). Direct evidence for a common leader sequence was obtained by nucleotide sequence analysis of the 5' ends of IBV and MHV mRNAs (Spaan et al., 1983; Lai et al., 1984; Brown et al., 1984). The IBV and MHV leader sequences of about 60 and 72 nucleotides, respectively, are transcribed from the 3' end of the negative-stranded template (Spaan et al., 1983; Lai et al., 1984; Brown et al., 1986; Bredenbeek et al., 1987; Shieh et al., 1987). However, the coronavirus mRNAs are not generated by splicing; replication occurs exclusively in the cytoplasm, and u.v. inactivation studies have shown that the mRNAs are transcribed independently (reviewed by Siddell et al., 1983).
Fig. 2. Consensus sequences (in the sense orientation) of the reinitiation sites involved in the synthesis of subgenomic mRNAs of MHV, BCV, FIPV, TGEV and IBV. For details see text.

To explain the presence of a common leader sequence, a discontinuous transcription of coronavirus mRNAs was proposed (Baric et al., 1983; Lai et al., 1983; Spaan et al., 1983). Several models for the required fusion of non-contiguous sequences have been suggested: jumping of the polymerase caused by the looping out of the negative-stranded template, post- or cotranscriptional ligation (or trans-splicing) and priming of the mRNA body transcription by the leader. Looping out of the template is unlikely to occur since loop structures have not been encountered in the RI isolated from MHV-infected cells (Baric et al., 1983). Also, the leader RNAs can be exchanged between the mRNAs of co-infecting coronaviruses during a mixed infection (Makino et al., 1986a).

Oligonucleotide fingerprints of non-denatured RI contained the oligonucleotides 10 and 19 (markers for the leader sequence), which would exclude the post-transcriptional trans-splicing model (Baric et al., 1983). Both the leader-primed transcription and the cotranscriptional trans-splicing model are compatible with the findings that leader sequences can be freely exchanged (Makino et al., 1986a) and that leader-containing transcripts of various sizes are present both in cells infected with wild-type virus and with an RNA- temperature-sensitive (ts) mutant (Baric et al., 1985, 1987). However, the leader RNA transcripts which function in mRNA body transcription have not been identified. In both models the leader RNA is transcribed independently of the mRNA body and, after termination, is translocated to conserved sequences (reinitiation sites) on the negative-stranded template to serve as a primer for the mRNA body synthesis. Alternatively, it may be spliced cotranscriptionally to mRNA body transcripts.

Several reinitiation sites (i.e. junction or homology sequences) have been identified by comparing the sequences of the 5' ends of the mRNAs and the corresponding regions on the genome (Spaan et al., 1983; Bredenbeek et al., 1987) or by S1 mapping (Brown & Boursnell, 1984; de Groot et al., 1987b). The remaining reinitiation sites have been recognized by homology searches and by their position in the genome (Boursnell & Brown, 1984; Binns et al., 1985; Boursnell et al., 1985b; Bredenbeek et al., 1986, 1987; Budzilowicz et al., 1985; Rasschaert et al., 1987; Jacobs et al., 1987; Skinner & Siddell, 1985; Skinner et al., 1985; Kapke & Brian, 1986; Lapps et al., 1987; Schmidt et al., 1987; Rasschaert & Laude, 1987; Luytjes et al., 1987, 1988). The consensus sequences of the reinitiation sites of IBV, MHV, BCV, TGEV and FIPV are listed in Fig. 2. It is evident that there is also a limited extent of reinitiation site homology between these viruses. Additionally, in IBV, MHV and TGEV there is a conserved sequence of 10 nucleotides about 80 bases from the 3' end of the genomic RNA (Kapke & Brian, 1986). This site may be critical for the synthesis of the negative-stranded template.

Sequence data obtained from IBV and MHV genomic RNA demonstrate a complementarity between the transcription reinitiation sites and the 3' end of the leader transcript (Brown et al., 1986; Bredenbeek et al., 1986, 1987; Shieh et al., 1987). This complementarity would allow base-pairing between the leader transcript and the internal transcription initiation sites. For the leader-primed transcription model it has been suggested that base-pairing is a prerequisite for the initiation of body RNA synthesis (Spaan et al., 1983; Brown et al., 1986; Shieh et al., 1987; Bredenbeek et al., 1986, 1987), and that the degree of complementarity between the free leader

<table>
<thead>
<tr>
<th>Virus</th>
<th>Reinitiation Site</th>
</tr>
</thead>
<tbody>
<tr>
<td>MHV-A59</td>
<td>AAUCUAAC</td>
</tr>
<tr>
<td>BCV</td>
<td>AAUCUAAC</td>
</tr>
<tr>
<td>FIPV</td>
<td>AACUAAC</td>
</tr>
<tr>
<td>TGEV</td>
<td>AACUAAC</td>
</tr>
<tr>
<td>IBV</td>
<td>CUUAACAA</td>
</tr>
</tbody>
</table>

(c) (u) (c) (g)
and the different reinitiation sites may regulate the expression of the mRNAs (Budzilowicz et al., 1985; Shieh et al., 1987).

At the time it was formulated, the leader-primed transcription regulation model was based on limited sequence data. Subsequently, computer-assisted analyses of two completely sequenced 'intergenic' regions of IBV strain M42 (Brown et al., 1986) and MHV-A59 (Bredenbeek et al., 1987; Luytjes et al., 1988) have shown that stability differences between intramolecular base-pairings involving the leader and the different reinitiation sites are not sufficient to explain the discrete mRNA moieties in infected cells (D. A. M. Konings, P. J. Bredenbeek, J. F. H. Noten, P. Hogeweg & W. J. M. Spaan, unpublished data).

RNA recombination

The RNA of coronaviruses can undergo recombination (Lai et al., 1985). A high frequency of recombination was detected in cells infected with a ts mutant of MHV strain A59 and wild-type JHM virus at the non-permissive temperature (Makino et al., 1986b). Multiple recombination sites have been detected at the 5' end of MHV genomic RNA (Keck et al., 1987). Although sequence data of the crossover sites are not available, the close relationship between the parental genomic RNAs suggests homologous recombination.

Results from studies of poliovirus recombination strongly suggest a copy-choice (or polymerase jumping) mechanism (Kirkegaard & Baltimore, 1986). The ability of the coronavirus polymerase to switch templates during the discontinuous mRNA transcription supports the copy-choice concept. The pool of leader RNA containing incomplete transcripts that have been detected in MHV-infected cells (Baric et al., 1985, 1987) can be the result of a discontinuous and non-processive replication mechanism. These RNA intermediates may play an important role in homologous recombination but also in the generation of defective interfering RNAs (Keck et al., 1987; Makino et al., 1984, 1985).

Recombination is an important feature of coronavirus evolution. Not only has recombination between MHV strains been shown to occur in vitro (see above) but also in mouse brain (Keck et al., 1988b). Comparison of IBV M protein sequences (Cavanagh & Davis, 1988) with the S1 sequences of the same strains (Niesters, 1987) suggests that some IBV field strains are recombinants. The high sequence divergence at the N terminus between the spike proteins of TGEV and FIPV could be the result of an RNA recombination event (Jacobs et al., 1987).

Sequence analysis of mRNA 2 of MHV-A59 has revealed a high similarity between the predicted amino acid sequence of a second ORF located at its 5' end and the HA1 subunit of the influenza C spike protein. This could be the result of a non-homologous recombination event (Luytjes et al., 1988).

These features and the differences among coronaviruses with respect to the number and order of the genes (de Groot et al., 1987c), the length of the intergenic region between the ORFs encoding D3 and M of IBV (Cavanagh & Davis, 1988) and the presence of a translatable gp65 ORF all testify to the inconstant nature of coronavirus genomes. This is likely to be a major area of future research.

CONCLUDING REMARKS

Progress has been rapid in coronavirology during the last few years, and interesting insights have been gained for virology in general. Thus recombination may be a major mechanism responsible for biological variation within the family, including expansion of the host spectrum. 'New' viruses may arise when members of different families occupying the same ecological niche exchange genetic information; the similarity between a predicted amino acid sequence in mRNA 2 of MHV-A59 and the HA1 subunit of the influenza C spike protein could be the result of such a non-homologous recombination event (Luytjes et al., 1988). Apart from evolutionary implications this insight must have consequences for the use of modified 'live' virus vaccines.

There are some areas to which future research will have to be directed. The role of spike protein cleavage and gp65 in the infection process is unclear. Apart from the molecular mechanisms of coronavirus pathogenesis the non-structural proteins, e.g. the polymerase, await
identification, while the details of mRNA transcription, RNA replication and recombination need to be elucidated.

We would like to thank Jacqueline Dumernet for typing the manuscript, and Drs Peter Bredenbeck, Peter Rottier, Willem Luytjes, Raoul de Groot and Harry Vennema for many helpful discussions. We are grateful to Drs K. V. Holmes, P. Palese, S. G. Siddell, B. A. M. van der Zeyt and S. Inglis for making data available to us prior to their publication and to Dr E. Chirnside for critical reading of the manuscript.

REFERENCES


Review: Coronavirus proteins and genomic RNA


