MOLECULAR BIOLOGY OF CORONAVIRUS 1986

Michael M.C. Lai

Departments of Microbiology and Neurology
University of Southern California, School of Medicine
Los Angeles, CA 90033

Introduction

Molecular biology took the center stage at the Second International Coronavirus Symposium held in 1983. Much of the discussion focused on the structure and synthesis of virus-specific RNAs, and viral structural proteins, particularly envelope glycoproteins. At that time, the application of recombinant DNA technology to coronavirus research was just beginning to change the direction of studies in this field. This trend became even more evident in the Third International Coronavirus Symposium in 1986, as a vast body of sequence data was presented. These data not only provided a deeper understanding of the viral RNAs and their genes, but also revealed many details of viral structural proteins, complementing the studies of protein biochemistry. This sequence information has also contributed significantly to our understanding of coronavirus pathogenesis and facilitated future development of effective coronavirus vaccines.

The structure of coronavirus genomic RNA

The genome of coronaviruses has been shown to be a 60 S single-stranded RNA, with a molecular weight ranging from 5.4×10^6 for murine coronavirus (Lai and Stohlman, 1978; Wege et al, 1978) to 8×10^6 for avian coronavirus (Lomniczi and Kennedy, 1977). The genomic RNA is infectious and contains a cap structure at the 5'-end and poly (A) sequence at the 3'-end. Since the coronavirus genomic RNA is considerably larger than any known stable RNA species, the molecular weight determination represents a rough estimate because of lack of reliable molecular weight markers. Indeed, these molecular weight estimates turned out to be considerable underestimates. In this meeting, Boursnell et al (this volume) presented results of their monumental efforts of sequencing the complete genome of avian infectious bronchitis virus (IBV). This sequencing revealed that the IBV genome is of 27.6 kilobases (kb), making it the largest viral RNA and stable RNA known to exist. Although the complete sequences of other coronavirus genomic RNAs are not yet available, it appears that the size of most of the other coronavirus genomes might be close to that of IBV. This large size might suggest that coronaviruses have the capacity to code for many proteins. However, the 5'-most 20 kb of the IBV RNA appears to belong to a single gene, which most likely encodes the RNA-dependent RNA polymerases. Analysis of these 20 kb sequences revealed two long overlapping open reading frames (ORF), which have the capacity to code for two proteins of 440 Kd and 300 Kd

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

respectively. These proteins have yet to be identified in IBV-infected cells. It is presumed that the genome-sized mRNA in infected cells would serve as the mRNA for the first protein at the 5'-end of the gene. However, no corresponding mRNA for the second potential protein has been detected. It is conceivable that ribosomal frameshifting (Jacks and Varmus, 1985) could result in a single protein product, which would have a molecular weight of more than 700,000 daltons.

Although no biochemical evidence is yet available to support the assignment of these potential protein products as RNA-dependent RNA polymerases, genetic evidence argues for such a likelihood: First, since coronavirus contains a positive-stranded genome and does not carry an RNA polymerase in the virus particles (Brayton et al, 1982; Dennis and Brian, 1982), the polymerase has to be translated from the incoming genomic RNA of the infecting virus. This is only possible if the polymerase gene is localized at the 5'-end of the genomic RNA, inasmuch as coronavirus does not employ post-translational proteolytic cleavage of a precursor protein (Sturman and Holmes, 1983). Second, in vitro translation of the genomic RNA yielded proteins of more than 200,000 daltons (Leibowitz et al. 1982b; Denison and Perlman, 1986), which are largest proteins approaching the size of the potential polymerases. In vitro translation products should reflect only proteins from the 5'-most ORF of mRNAs. Third, RNA recombinants derived from temperature-sensitive mutants of mouse hepatitis virus (MHV) localized the ts lesions of some of the RNA (-) mutants to the 5'-end of the coronavirus genome (Keck et al, this volume).

Although RNA polymerase activities have been detected in several coronavirus-infected cells, including MHV (Brayton et al, 1982; Mahy et al, 1983) and porcine transmissible gastroenteritis virus (TGEV) (Dennis and Brian, 1982), these activities have not been associated with any protein. If the 440 Kd and 300 Kd proteins are indeed viral RNA polymerases, why does coronavirus need such a huge polymerase? It might be due to the fact that the virus utilizes a complex process for synthesizing its mRNAs. The exceptionally large size of the potential proteins suggests multiple functions of the polymerase.

Mechanism of coronavirus mRNA synthesis

All of the coronaviruses synthesize six to seven mRNA species in infected cells. These mRNAs have nested-set 3'-coterminal structure, i.e. they represent sequences from the 3'-end of the genome and extend for various distances toward the 5'-end (Stern and Kennedy, 1980; Lai et al 1981). Only the unique sequence at the 5'-end of each mRNA, which does not overlap with the next smaller mRNAs is used for translation. Thus, these mRNAs are physically polycistronic but functionally monocistronic. In view of the monocistronic nature of the mRNAs, the coronavirus genomes would encode only six to seven proteins, despite the large size of the virion RNA genome. Several different strains of coronavirus synthesize additional mRNA species, although it is not clear whether they have any functional roles. However, one coronavirus, bovine coronavirus (BCV), does synthesize an additional mRNA species which might encode a functional protein, hemagglutinin, unique to BCV (Keck, this volume).

Coronavirus mRNAs have another unique feature, namely, the presence of a 5'-leader sequence of 50-70 nucleotides. This finding was first suggested by the analysis of T1-oligonucleotides in MHV (Lai et al, 1982). In the last coronavirus symposium, additional evidence for the presence of leader RNA was presented by several groups (Lai et al, 1983, 1984, Spaan, et al, 1984). Subsequent sequence studies have firmly established the presence of leader sequences in at least MHV and IBV. Furthermore, a large body of data has been obtained that suggests a novel mechanism for MHV mRNA transcription, termed "leader RNA-primed transcription" in which a leader RNA is synthesized from the 3'-end of the negative-strand RNA template, dissociates from this template and then reassociates ("primes") at downstream sites where transcription of each of the

mRNAs is initiated (Baric et al, 1983). This mechanism has been supported by the detection of free leader RNA (Baric et al, 1985) and by the observation that the leader sequences can be exchanged with very high frequency between co-infecting viruses, suggesting that the leader RNA serves as a separate transcriptional unit (Makino et al, 1986b). At this meeting, the detailed mechanism of this transcription model was revealed from the 5'-end genomic sequence of the JHM strain of MHV. It appears that the 3'-end of the leader RNA shares sequence homology of 7-18 nucleotides with the sequence at the initiation sites of mRNAs (Soe et al, this volume). Thus, the leader RNA is complementary to sequences in the intergenic regions on the template RNA. Furthermore, the sequence homology extends beyond the putative leader-body junction points of mRNAs. Thus, it was proposed that the RNA polymerase contains an endonucleolytic activity which cleaves the leader RNA before it is used as the primer. The extent of sequence homology appears to correlate well with the molar amounts of particular mRNA species. This transcription mechanism reveals some similarity to the CAP-snatching transcription of influenza viruses (Plotch et al, 1981). This finding has now been confirmed in the A59 strain of MHV (Bredenbeek, et al. this volume) and also agrees well with the sequence data of IBV (Brown et al, 1986).

Thus, the mechanism of coronavirus mRNA transcription provides an exciting and novel area of molecular biology, and represents an alternative mechanism to conventional RNA splicing. The enzymology of RNA synthesis is still not clear. It is known, however, that the putative RNA polymerases are extremely large and that at least six complementation groups are involved in RNA synthesis (Leibowitz et al, 1982a). Whether any of the detected or presumed nonstructural proteins are required for transcription has still not been established. Obviously, this area will be a subject of intensive future studies.

RNA recombination and defective-interfering (DI) particles--a model of discontinuous jumping transcription?

Two recent observations further contributed to our understanding of coronavirus RNA synthesis. The first observation is that coronaviruses could undergo RNA-RNA recombination at a very high frequency (Makino et al, 1986a). Keck et al (this volume) expanded on this observation and presented evidence that multiple recombination events could take place between two strains of MHVs. Furthermore, by using appropriate selection pressure, it is possible to obtain recombinants with cross-over sites at practically any part of the genome. A series of recombinants between MHV-2 and A59 of MHV is particularly revealing. These recombinants were selected by their ability to cause cell fusion and by use of a ts marker, both of which are probably localized in the gene encoding the peplomer protein. Most of the recombinants had additional cross-overs in other parts of the genome in which no selection pressure was applied. All of these data further suggest that the recombination frequency is extremely high, approaching the frequency of RNA reassortment of segmented RNA viruses. Hence, what distinguishes coronaviruses from other RNA viruses which do not recombine or recombine at very low frequency? It was speculated that coronavirus RNA replication proceeds by a discontinuous and non-processive "stop-and-go" mechanism, thus yielding free RNA intermediates which become precursors for RNA recombination by a copy-choice mechanism. Indeed, such free RNA intermediates have been detected in MHV-infected cells (Baric et al, this volume). Sequence analysis of a limited number of recombinants suggests that the recombination sites correspond to the regions of secondary structure on the RNA template as well as the sizes of free RNA intermediates. Thus, RNA recombination could be generated by a copy-choice mechanism involving these free RNA intermediates.

Another interesting observation concerns DI particles. Unlike the DI particles of other viruses, the DI particles of cornaviruses synthesize distinct

subgenomic polyadenylated RNAs in infected cells. Furthermore, these intracellular subgenomic DI-specific RNAs contain sequences derived from several discontiguous parts of the DI genome (Makino et al, 1985). These defective intracellular RNAs are not appreciably detected in virions and thus are probably transcribed de novo by a discontinuous mechanism from the DI particle genome. However, it has not been ruled out that a small amount of subgenomic DI RNA is incorporated in virions and thus serves as the template for its transcription. Evidence was presented that the transcription of these subgenomic DI RNAs requires a helper function, while the replication of the DI genomic RNA does not. These complex processes of DI RNA transcription and RNA recombination are probably related to "leader-primed transcription" which involves a discontinuous transcriptional process, and may provide valuable insights into the normal mechanism of RNA transcription.

Structure of the structural proteins

Cloning and sequencing of coronaviral RNAs has increased our understanding of viral structural and nonstructural proteins. Characterization of viral structural proteins was one of the earliest advances in the study of the biochemistry of coronaviruses (Sturman, 1977). Three structural proteins, gp180 peplomer protein (E2), gp25 matrix protein (E1) and pp60 nucleocapsid protein (N), have been detected in all of the coronaviruses studied. In addition, some coronaviruses, such as bovine coronavirus (BCV), contain an additional protein gp65 hemagglutinin protein (Hogue and Brian, this volume). It is unclear if the latter protein plays an important biological function.

The E2 protein has been the subject of intensive investigation. This protein is responsible for interaction with cellular receptors, induction of cell fusion, and elicitation of neutralizing antibody and cell-mediated cytotoxicity. It is usually cleaved, probably by cellular proteases into two different 90 Kd subunits in virions (Sturman et al, 1985). In most instances, the cleavage is required for virus infectivity and cell fusion. However, the cleavage of E2 is not observed in feline infectious peritonitis virus (FIPV). The complete sequence of the gene encoding the E2 protein has been obtained for IBV (Binns et al, 1985), MHV (de Groot et al, this volume; Schmidt et al, 1987), TGEV (Laude, this volume) and FIPV (de Groot et al, this volume). The sequences showed that the protein contains approximately 1400 to 1800 amino acids. It has features typical of membrane proteins, such as the presence of a signal peptide at the N-terminus and a hydrophobic membrane-anchoring domain at the carboxyl terminus of the protein. The N-terminal half exhibits greater divergence in contrast to the C-terminal half. There is a cleavage site in the middle of the protein for some of the viruses. The sequence suggests that the peplomer contains long alpha-helix chains, which may interact with each other to form a coiled coil structure, which may be the basis of the peplomer stalks.

RNA recombination studies (Keck et al, this volume) suggest that the C-terminal half of the E2 protein actually contains the neutralization epitopes, neuropathogenic determinants and the determinants of cell fusion-inducing activity of the MHV peplomer. This result would suggest that the carboxyl-terminal half of the peplomer protein is more exposed. How this information reconciles with the proposed peplomer structure is not clear. Preparation and characterization of a large number of monoclonal antibodies specific for E2 have also been reported in the literature and at this meeting. These monoclonal antibodies will be useful for understanding the structural domains and structure-function relationship of the E2 protein. The sequence data and structural studies of E2 would obviously facilitate future development of effective coronavirus vaccines.

One of the functions of E2 is the binding to receptors on the cell membrane of target cells. Holmes et al (this volume) reported the detection of coronavirus

receptors on the cell surface of enterocytes and hepatocytes of the genetically susceptible Balb/c mice. Absence of the receptors in SJL/J strain corresponds to its resistance to coronavirus infection. The study of receptors will contribute significantly to our understanding of the biological roles of E2 and molecular mechanism of initial stages of virus replication.

The second structural protein of coronaviruses, E1, constitutes the matrix protein. This protein has two interesting properties: First, glycosylation of the E1 of murine coronavirus occurs through O-glycosidic bond instead of more common N-glycosidic bond (Niemann and Klenk, 1981), and thus, is not inhibited by tunicamycin (Holmes, et al, 1981). Second, the coronavirus matures into endoplasmic reticulum instead of budding through plasma membrane (Sturman and Holmes, 1983). The maturation is probably mediated through the E1 protein. These two properties of E1 were examined by expression of the E1 gene in mammalian cells (Niemann et al, this volume). It was shown that the O-glycosylation is not essential for the function of the E1 protein. Furthermore, the E1 sequence itself is responsible for the transport of the E1 into endoplasmic reticulum in the perinuclear region. The sequence determining such a property is mapped within the transmembrane domain of the E1 protein. These properties make E1 an interesting protein for the study of the transport of membrane proteins.

The third protein is the nucleocapsid protein, N, which is a phosphorylated protein interacting with virion genomic RNA (Robbins et al, 1986). It probably plays a role in RNA transcription and viral morphogenesis. However, its precise biological function, in addition to its structural roles, is not clear. The relative paucity of knowledge on the functional roles of viral structural proteins is partly due to lack of suitable genetic mutants. Initial attempts at isolating temperature-sensitive mutants affecting the viral structural proteins have already been made (Sawicki; Sturman et al, this volume). The characterization of ts mutants should help us better understand the functions of these proteins.

Expression of non-structural proteins

Based on the number of mRNAs, coronaviruses have six to eight genes but encode at most three or four structural proteins. Therefore, the viruses have capacity to encode at least three or four nonstructural proteins, one of which must be the RNA-dependent RNA polymerase. Some of the nonstructural proteins have been detected. For instances, p30 from gene B (mRNA 2) and p15 from gene D (mRNA 4) of MHV were detected by polyacrylamide gel electrophoresis of proteins from infected cells. These proteins were reported prior to the second coronavirus symposium; however, their function still remains elusive. Recent sequencing data revealed additional ORFs in several genes encoding nonstructural proteins, such as mRNA D of IBV (Smith et al, this volume) and mRNA 5 of MHV (Skinner et al, 1985). An ORF has also been detected at the 3'-end noncoding regions of the genomes of TGEV (Kapke and Brian, 1986). It is not known whether all of these ORFs are utilized. The identification of the bona fide protein-encoding ORFs would require further characterization of favorable translation initiation sequences and favorable codon usage patterns in these ORFs.

The availability of cDNA clones and sequences of these genes should enable the generation of specific antibodies, which could be utilized to assess the functions of the potential nonstructural proteins. In addition, isolation and characterization of genetic mutants would particularly facilitate the progress in this area.

Future

In the last 5-6 years, the molecular biology of coronaviruses has progressed at an extremely rapid pace. The basic description of the structure of the viral RNA and proteins and the major events of viral replication cycle has been completed. Now this field is entering another phase, that is, the utilization of recombinant DNA technology to provide more detailed knowledge, as evidenced by the presentations in this symposium. Several major issues remain to be studied:

(1) The mechanism of RNA synthesis: The data obtained so far indicated that coronaviruses use a novel mechanism of leader RNA-primed transcription. Many details of this transcription model have been obtained from the sequences of mRNAs and genomic RNA, and also from the molecular studies on infected cells. More information is forthcoming from the studies of defective-interfering RNAs and RNA recombination. However, the precise mechanism of RNA transcription will require studies using an in vitro transcription system. Several in vitro systems have been described (Brayton et al, 1982; Dennis and Brian, 1982; Mahy et al, 1983) but none are very efficient, nor able to utilize exogenous RNA as a transcription template or primer. Although in vitro transcription systems were not presented in this meeting, this approach should eventually become the focus of future efforts to understand the mechanism of RNA synthesis. These studies are particularly important in establishing the priming activity of the leader RNA.

(2) The nature of the RNA polymerase: RNA polymerase is among the last remaining nonstructural proteins to be identified in coronavirus-infected cells. Since this enzyme is involved in leader-primed transcription, high-frequency RNA-RNA recombination, generation and transcription of DI RNA, and is of extremely large size, the RNA polymerase of coronaviruses would be of extreme interest. A major advance has been made by the completion of sequencing of the gene encoding the RNA polymerase of IBV (Boursnell et al, this volume). This sequence provides a glimpse of the possible structure of the RNA polymerase. It would now be important to identify these proteins in the infected cells and study the processing and functions of these polymerases. These studies would have to be complemented by genetic studies. Although many ts mutants affecting RNA synthesis have been obtained, which fall into six complementation groups, surprisingly few studies have been performed to identify the defects of these ts mutants. These genetic studies are needed to complement the biochemical characterization of the protein.

(3) Other nonstructural proteins: At least three other genes encode nonstructural proteins. Sequences have been obtained on some of these genes, and the possible gene products have been speculated. The functions of these nonstructural proteins have yet to be identified. Like the RNA polymerase, this area requires genetic studies using ts mutants.

Acknowledgment

I thank Carol Flores for excellent typing of this manuscript.

References

Baric, R.S., Stohlman, S.A. and Lai, M.M.C. (1983): J. Virol. 48:633-640

Baric, R.S., Stohlman, S.A., Razavi, M.K. and Lai, M.M.C. (1985): Virus Res. 3:19-33

Binns, M.M., Boursnell, M.E.G., Cavanagh, D., Pappin, D.J.C. and Brown, T.D.K. (1985): J. Gen. Virol. 667:719-726

Brayton, P.R., Lai, M.M.C. and Stohlman, S.A. (1982): J. Virol. 42:847-853

Brown, T.D.K., Boursnell, M.E.G., Binns, M.M. and Tomley, F.M. (1986): J. Gen. Virol. 67:221-228

Denison, M.R., and Perlman, S. (1986): J. Virol. 60:12-18

Dennis, D.E. and Brian, D.A. (1982): J. Virol. 42:153-160

Holmes, K.V., Doller, E.W. and Sturman, L.S. (1981): Virology 115:334-344

Jacks, T. and Varmus, H.E. (1985): Science 230:1237-1242

Kapke, P.A. and Brian, D.A. (1986): Virology 151:41-49

Lai, M.M.C., Baric, R.S., Brayton, P.R. and Stohlman, S.A. (1984): Proc. Nat. Acad. Sci. USA 81:3626-3630

Lai, M.M.C., Brayton, P.R., Armen, R.C., Patton, C.D., Pugh, C. and Stohlman, S.A. (1981): J. Virol. 39:823-834

Lai, M.M.C., Patton, C.D., Baric, R.S. and Stohlman, S.A. (1983): J. Virol. 46:1027-1033

Lai, M.M.C., Patton, C.D., Stohlman, S.A. (1982): J. Virol. 42:1080-1087

Lai, M.M.C. and Stohlman, S.A. (1978): J. Virol. 26:236-242

Leibowitz, J.L., DeVries, J.R. and Haspel, M.V. (1982a): J. Virol. 42:1080-1087

Leibowitz, J.L., Weiss, S.R., Paavola, E. and Bond, C.W. (1982b): J. Virol. 43:905-913

Lomniczi, B and Kennedy, I. (1977): J. Virol. 24:99-107

Mahy, B.W.J., Siddell, S., Wege, H. and ter Meulen, V. (1983): J. Gen. Virol. 64:103-110

Makino, S., Fujioka, N. and Fujiwara, K. (1985): J. Virol. 54:329-336

Makino, S., Keck, J.G., Stohlman, S.A. and Lai, M.M.C. (1986a): J. Virol 57:729-737

Makino, S., Stohlman, S.A. and Lai, M.M.C. (1986b): Proc. Nat. Acad. Sci. USA 83:4204-4208

Niemann, H. and Klenk, H.-D. (1981): J. Mol. Biol. 153:381-392

Plotch, S.J., Bouloy, M., Ulmanen, I., and Krug, R.M. (1981): Cell 23:847-858

Robbins, S.G., Frana, M.F., McGowan, J.J., Boyle, J.F. and Holmes, K.V. (1986): Virology 150:402-410

Schmidt, I., Skinner, M. and Siddell, S. (1987): J. Gen. Virol. 68:47-56

Skinner, M.A., Ebner, D. and Siddell, S.G. (1985): J. Gen. Virol. 66:581-592

Spaan, W., Delius, H., Skinner, M., Armstrong, J., Rottier, P., Smeekens, S., van der Zeijst, B.A.M. and Siddell, S.G. (1983): EMBO J. 2:1939-1944

Stern, D.F. and Kennedy, S.I.T. (1980): J. Virol. 36:440-449

Sturman, L.S. (1977): Virology 77:637-649

Sturman, L.S. and Holmes, K.V. (1983): Adv. Virus Res. 28:35-112

Sturman, L.S., Ricard, C.S. and Holmes, K.V. (1985): J. Virol. 56:904-911

Wege, H., Muller, A. and ter Meulen, V. (1978): J. Gen. Virol. 41:217-227