Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
In recent years a remarkable catalog of mechanisms has been identified by which eukaryotic cells expand their genome coding capacity beyond the linear blocks of nucleotide sequence originally thought to encode individual proteins. These coding strategies not only create diversity by increasing the number of proteins encoded but also provide a means by which to regulate the expression of these proteins. Nowhere is this more apparent than with eukaryotic viruses. In most cases they have provided the prototypic example of molecular mechanisms that serve to maximize coding potentials. The compactness of the viral genome may be an important factor contributing to the success of the virus as a cellular parasite. The viruses with RNA genomes have used many different forms of gene expression, and the influenza viruses provide some illuminating examples of genome diversity. Although the functions of some of the influenza virus gene products described below are not yet known, the variety of mechanisms used for their synthesis provides a paradigm of successful exploitation of a genome.

**FIG 1**

A schematic diagram of the structure of the influenza A virus particle. Three types of integral membrane protein—hemagglutinin (HA), neuraminidase (NA) and small amounts of M2—are inserted through the lipid bilayer of the viral membrane. The virion membrane protein M1 is thought to underlie the lipid membrane. Within the envelope are the eight segments of single-stranded genome RNA contained in the form of helical ribonucleoproteins (RNP). The nucleocapsid protein (NP) is associated with approximately every 20 nucleotides of the RNA. Associated with the RNPs are small amounts of the transcriptase complex, consisting of proteins PB1, PB2 and PA. The coding assignments of the eight RNA segments are also illustrated. RNA segments 7 and 8 each code for more than one protein (M1 and M2, and NS1 and NS2, respectively). NS1 and NS2 are found only in infected cells and are not thought to be structural components of the virus. Influenza B virus does not encode an M2 integral membrane protein, but the NB glycoprotein encoded by RNA segment 6 (which also encodes NA) is of very similar structure to M2, and it seems possible it will perform a similar function in cells and virions. Influenza C virus contains seven RNA segments, lacking an RNA segment for NA. However, the HA also has a neuraminic-O-acetylemesterase activity (reviewed in Ref. 1). An equivalent of M2 or NB has yet to be identified in influenza C viruses.

**Diversity of coding strategies in influenza viruses**

ROBERT A. LAMB AND CURT M. HORVATH

Influenza viruses have exploited a variety of strategies to increase their genome coding capacities. These include unspliced, spliced, alternatively spliced and bicistronic mRNAs, translation from overlapping reading frames and a coupled stop-start translation of tandem cistrons.

The influenza A, B and C viruses comprise a family of related enveloped viruses with a segmented single-stranded RNA genome that has been called 'negative stranded' because the viral mRNAs are transcribed from the viral RNA segments (vRNAs) by a virus-encoded RNA transcriptase. The complete
nucleotide sequence of the eight RNA segments of influenza A and B viruses and the seven RNA segments of the influenza C viruses has been obtained: the PR/8/34 strain of influenza A virus contains 13,588 nucleotides (reviewed in Refs 1, 2). A schematic diagram of the structure of influenza A virus is shown in Fig. 1. The virion lipid envelope, which is derived from the host cell plasma membrane, contains three integral membrane proteins: hemagglutinin, neuraminidase and small amounts of M2. The viral membrane protein (M1) is thought to underlie the lipid bilayer. The nucleocapsid protein is the major structural protein that interacts with the vRNA segments to form the ribonucleoprotein particles (RNPs). Associated with each RNP core are small amounts of a complex consisting of the three transcriptase-associated polypeptides PB1, PB2 and PA.

The first 12 nucleotides at the 3'-end, and the first 13 nucleotides at the 5'-end of each vRNA segment are conserved in all eight RNA segments (see Fig. 2); recent studies using RNAs of altered sequence, and reconstitution of the transcriptase activity, indicate a promoter-like role for these sequences in transcriptase recognition3. In the nucleus of infected cells, transcription of the vRNAs into mRNAs by the virus-encoded RNA transcriptase requires cooperation with ongoing transcription by cellular RNA polymerase II. The synthesis of influenza virus mRNAs is an unusual process, as its initiation requires m7GpppXm-containing capped primers that are generated from a subset of host cell RNAs by an influenza virus-encoded cap-dependent endonuclease. The primers average 10–13 nucleotides in length and in many cases contain the sequence 5'-GCA-3' immediately adjacent to the cleavage site, suggesting that the RNA endonuclease has some nucleotide sequence specificity (reviewed in Ref. 4). Viral transcription is initiated by the addition of a G residue complementary to the penultimate C residue on the vRNAs, and transcription elongates until it reaches the poly(A) addition site located 15–22 nucleotides short of the 5'-end of the vRNA. Like most cellular polymerase II transcripts, the influenza virus mRNAs are methylated on internal A residues, but the role of methylation is not known4.

In addition to stealing caps, influenza virus mRNAs make use of another aspect of host cell nuclear function, namely the splicing machinery. As detailed below, the splicing of influenza virus mRNA transcripts provides the only known example of splicing of RNA that is not transcribed from DNA by RNA polymerase II. Since mRNAs synthesized by the viral RNA transcriptase are incomplete transcripts of the vRNA and contain host cell sequences at their 5'-end and poly(A) at their 3'-end, replication of influenza virus vRNA requires the synthesis of full-length copies of the vRNA segments, known as cRNA or template RNAs (reviewed in Ref. 4). The template RNAs are faithful copies of the vRNA, they have no 5' cap structure or 3' poly(A) tail, and they serve as the template for genome replication. The differences between the vRNA, mRNA and template RNA species are shown in Fig. 2.

**Spliced mRNAs and overlapping reading frames**

Genetic and biochemical evidence had shown that vRNA segment 8 (890 nucleotides), the smallest of the influenza virus RNA segments, encoded two discrete proteins, NS1 (26 kDa) and NS2 (14 kDa), and that NS2 is translated from a small mRNA (~350 nucleotides)5,6. Nucleotide sequencing indicated that the NS1 mRNA is unspliced, directly encoding the NS1 protein (237 amino acids) whereas the NS2 pre-mRNA contains a 473 nucleotide intron. NS1 and NS2 share the same AUG codon for initiation of protein synthesis and nine subsequent amino acids before the intron, and then translation of the body of the NS2 mRNA continues in the +1 reading frame, which overlaps the NS1 frame by 70 amino acids7. Figure 3A shows the arrangement of the NS1 and NS2 mRNAs and their open reading frames (ORFs). An analogous arrangement of unspliced and spliced mRNA transcripts is found to encode the NS1 and NS2 proteins derived from RNA segment 8 of influenza B virus and RNA segment 7 of influenza C viruses, respectively8,9.
Analysis of the RNA transcripts derived from RNA segment 7 in cells infected with influenza A virus indicated that in addition to the unspliced transcript encoding the M1 protein there are two alternatively spliced RNAs, M2 and mRNA3, that involve the use of different 5' splice sites (Fig. 3B). The M2 mRNA, which encodes the M2 protein, contains a 51 nucleotide virus-specific leader sequence, a 689 nucleotide intron and a 271 nucleotide [excluding poly(A) tail] body region. The leader sequence of the M2 mRNA contains the AUG initiation codon and codons for eight subsequent residues which are shared with the M1 protein; the M2 mRNA body region encodes 88 residues in the +1 reading frame and overlaps the M1 protein by 14 amino acids. The alternatively spliced mRNA3 has a 5' leader sequence of 11 virus-specific nucleotides, and shares the same 3' splice site as the M2 mRNA. If this mRNA is translated it would yield a nine residue peptide identical to the carboxy-terminal region of the M1 protein but to date the peptide has not been identified in infected cells. An analogous situation does not occur with influenza B virus RNA segment 7, but this RNA segment does contain a second ORF and this is described in detail below.

In influenza C virus, another variation on the splicing theme has been found for the M protein-encoding RNA segment 6 (influenza C virus has just seven RNA segments as it lacks a separate RNA segment encoding a neuraminidase). The primary transcript mRNA contains an ORF encoding 374 amino acids and this mRNA is spliced such that a translational termination codon is introduced after 242 residues of the ORF. The available evidence suggests that the spliced mRNA encodes the M protein (see Fig. 4). At present a protein product translated from the unspliced transcript mRNA has not been identified, which is surprising considering the abundance of the primary transcript RNA.

The 5' and 3' splice junctions of the influenza virus RNAs fit reasonably closely with the consensus sequences found at both sides of the exon-intron boundaries of eukaryotic RNAs. Evidence that the cellular splicing machinery is used to generate the influenza virus spliced mRNAs comes from the findings that the
splice junctions are used when the cDNAs are expressed using DNA vectors\textsuperscript{13,15}. In most cases of splicing of eukaryotic mRNAs, only the spliced mRNA products are detected in the cytoplasm. In influenza virus, and also in retroviruses, both spliced and unspliced RNAs code for proteins and thus both spliced and unspliced RNAs are transported from the nucleus to the cytoplasm.

For influenza virus, the extent of splicing is controlled, with a steady-state amount of about 10% spliced mRNA to unspliced mRNA (reviewed in Ref. 3). Some evidence suggests that this regulation is mediated by a \textit{trans} mechanism requiring the action of a viral protein\textsuperscript{16}. Other evidence from studies of the splicing of NS\textsubscript{1} mRNA \textit{in vitro} indicates that the regulation is via a \textit{cis}-acting mechanism, but the nucleotide sequences involved in the control system have yet to be identified\textsuperscript{17,18}. In addition, the extent of splicing of the NS\textsubscript{1} mRNA can be regulated by the efficiency of nucleocytoplasmic transport of unspliced NS\textsubscript{1} mRNA\textsuperscript{19}. In human immunodeficiency virus type 1, the \textit{rev} gene product acts through a structured target RNA sequence located near the \textit{env} gene to activate the transport of unspliced RNA from the nucleus\textsuperscript{20}. It is possible either that the influenza virus NS\textsubscript{1} protein regulates the transport of unspliced NS\textsubscript{1} mRNA from the nucleus, analogous to \textit{rev}, or that NS\textsubscript{1} acts as a \textit{trans}-dominant negative regulator of NS\textsubscript{1} spliced mRNA transport to the cytoplasm (R.M. Krug, pers. commun.).

It is interesting to speculate how the overlapping genes might have evolved. The length of the NS\textsubscript{1} protein varies between 124 and 237 residues in different strains of influenza A virus, and analysis of the nucleotide sequence in the region of overlap between NS\textsubscript{1} and NS\textsubscript{2} indicates that the amino acid sequence of NS\textsubscript{2} is conserved at the expense of NS\textsubscript{1} (reviewed in Ref. 1). These observations support the earlier suggestion, made by Winter and colleagues\textsuperscript{21}, that perhaps NS\textsubscript{1} and NS\textsubscript{2} were originally collinear in the vRNA but not overlapping; readthrough of a terminator at the end of NS\textsubscript{1} then allowed the NS\textsubscript{1} protein to become longer and use more of the reading frame.

**Bicistronic mRNAs**

At the translational level, diversity of polypeptide production can arise by the selection of alternative initiation sites for translation on bicistronic or poly-cistronic mRNAs. The coding potential of the influenza B virus genome is expanded, since its RNA segment 6 gives rise to a functional bicistronic mRNA containing two initiating AUG codons that are separated by four nucleotides, so two discrete proteins are expressed from different reading frames: the 100 amino acid NB glycoprotein and the 466 amino acid neuraminidase glycoprotein\textsuperscript{22-24}. The NB ORF overlaps the neuraminidase reading frame by 292 nucleotides (see Fig. 5).

The initiation of protein synthesis from eukaryotic mRNAs usually occurs at the 5' proximal AUG codon, although other AUG codons can be used for translational initiation in some cellular and many viral mRNAs. Most examples of translational initiation can be accommodated by the 'modified scanning hypothesis' of eukaryotic translation\textsuperscript{25} in which 40S ribosomal subunits enter the 5'-end of the mRNA and 'scan' the nucleotide sequence, examining each AUG codon for its potential use as an initiator. From examination of the NB/neuraminidase mRNA nucleotide sequence in conjunction with the rules of the modified scanning hypothesis\textsuperscript{25}, it would be expected that only NB and not neuraminidase should be synthesized in infected cells. However, NB and neuraminidase accumulate in a 0.6:1 ratio, suggesting that approximately 60% of ribosome preinitiation complexes scanning the NB/neuraminidase mRNA do not initiate protein synthesis at the first AUG codon but continue scanning until reaching the second AUG codon four nucleotides downstream\textsuperscript{26}. The importance of the nucleotide sequences surrounding the two AUG codons was investigated by altering many of the residues flanking the initiation codons. Most of the changes had very little effect on the ratio of NB/neuraminidase that accumulated. However, deletion of 40 of the 46 influenza virus-specific 5'-untranslated region nucleotides decreased neuraminidase synthesis tenfold, suggesting that an unrecognized feature of the entire region, such as a specific secondary structure, also has a major role in the initiation of protein synthesis at the second AUG codon\textsuperscript{27}.

**Coupled translation of tandem cistrons**

The nucleotide sequence of influenza B virus RNA segment 7 indicated that, in addition to the gene encoding the 248 amino acid M\textsubscript{1} protein, there is a second ORF in the +2 frame that has a coding capacity of 195 amino acids and is designated BM2ORF (reviewed in Ref. 1) (see Fig. 6). To search for a polypeptide product derived from BM2ORF, an antiserum to
an engineered β-galactosidase-BM2ORF fusion protein was produced and used to identify a polypeptide, designated BM2 protein (12 kDa), that is synthesized in cells infected with influenza B virus. A mutational analysis of the cloned DNA was performed and the altered DNAs expressed in eukaryotic cells, in an attempt to reveal the mechanism by which the BM2 protein is synthesized. The data indicate that the BM2 protein initiation codon overlaps with the termination codon of the M1 protein in a translational stop-start pentanucleotide UAAUG, and that the expression of the BM2 protein requires termination of M1 synthesis adjacent to the 5'-end of the BM2 coding region (Fig. 6).

The simple notion that initiation of BM2 synthesis is due solely to migration of the ribosome scanning from the 5'-end of the mRNA past 21 AUG codons in all three reading frames is unlikely, given that BM2 synthesis did not occur when the M1 and BM2 reading frames were fused. The results for translation of this construct also make it unlikely that the ribosome enters internally in the mRNA. Internal entry of ribosomes has been found to occur with picornaviruses, where initiation of protein synthesis occurs at the seventh AUG codon from the 5'-end of the viral RNA. A large body of data indicates that cis-acting RNA sequences adjacent to the initiator are involved in the ‘ribosome landing pad’ or ‘internal ribosome entry site’ and that specific cellular factors associate with these sequences (reviewed in Ref. 27). A related but seemingly different form of internal initiation of protein synthesis also appears to occur with the polymerase (reverse transcriptase) gene of the hepatitis B virus, since it is located 3' to an overlapping gene (C) yet mutations affecting translation of C have only minimal effects on expression of polymerase.

Reinitiation of translation at downstream AUG codons has been found to occur with some other naturally occurring and artificially constructed bicistronic mRNAs, the latter mostly using preproinsulin as a model system. In these cases, as with the BM2 protein, initiation of protein synthesis at the downstream AUG codon depends on translational scanning of the polypeptide encoded by the upstream ORF. In almost all of these natural or artificially created mRNAs, the upstream ORF is small and has been characterized as a ‘mini-cistron’. In a study on the effect of intercistronic length on the efficiency of reinitiation at downstream AUG initiation codons, using altered preproinsulin mRNAs, it was found that when the terminator/initiator was in the nucleotide sequence UAAUG, reinitiation was inefficient, but when the intercistronic distance was lengthened the efficiency of reinitiation was greatly increased. It has been suggested that initiation factors may be released stochastically during chain elongation, and if these factors are needed for reinitiation, then the ability to reinitiate would be inversely proportional to the size of the upstream ORF. The comparatively large size of the influenza B virus M1 ORF (248 amino acids) may in part explain the relative inefficiency of reinitiation (~25 mol % of M1). In addition, the influenza B virus BM2 AUG initiation codon is not in the most favored nucleotide context for initiation of protein synthesis, and this may affect the efficiency of reinitiation. The major difference between the analysis of the naturally occurring influenza B virus M1/BM2 bicistronic mRNA and the data obtained using artificial RNAs is that when the intercistronic distance was increased, the synthesis of BM2 could not be detected, whereas in the artificial constructions, translation from the second ORF was found to be increased. These data suggest that for BM2 synthesis to occur, the termination of translation and the reinitiation event have to be closely coupled.

In some respects, the finding that the BM2 protein initiation codon overlaps the M1 protein termination codon is similar to the situation found for some coordinately regulated bacterial genes (e.g. the trp operon), where the termination codon of one gene overlaps the initiation codon for a downstream gene (reviewed in Ref. 29). When the intercistronic distance is expanded experimentally in prokaryotic genes, the efficiency of translation of the second ORF is greatly diminished, except in cases where an internal Shine-Dalgarno sequence precedes the second ORF. In these latter cases, it has been suggested that the ribosome translating the first ORF could expose a masked ribosome entry site by melting RNA secondary structure, thereby allowing recognition of the downstream initiation site by ribosomes. However, for the vast majority of eukaryotic mRNAs, ribosomes enter only at the 5'-end of the mRNA. The frequent presence of internal ribosome entry sequences in prokaryotic but not eukaryotic mRNAs may account for one major difference between the M1/BM2 situation and prokaryotic coupled stop–start systems.
in that prokaryotic reinitiation events are often highly efficient.

Although they make use of a wide range of coding strategies, influenza viruses have not been demonstrated to use some other known mechanisms to access 'hidden' reading frames. Among those employed by other viruses is the use of non-AUG initiation codons (ACG or AUA) for protein synthesis by hepatitis B virus, adeno-associated virus and Sendai virus, a paramyxovirus (reviewed in Ref. 31). In addition, frameshifting in the −1 frame occurs with most retroviruses to switch from the gag to pol reading frames, and also in coronaviruses to express their RNA replicase protein. Suppression of termination codons occurs with Sindbis virus to synthesize its ns4 protein. Finally, a form of RNA editing is used by paramyxoviruses (another group of negative strand RNA viruses) to add nontemplated G nucleotides at a specific site in an mRNA. The mechanism for this process is probably similar to the addition of poly(A), involving 'stuttering' of the RNA transcriptase. The consequence of the addition of extra nucleotides to some of the mRNA transcripts is to generate two (and sometimes three) different mRNAs encoding fusion proteins with different carboxy-terminal domains (Refs 35, 36 and references therein; reviewed in Ref. 31). Whether more proteins will be found to be encoded by the influenza virus genome by mechanisms such as non-AUG initiators, ribosomal frameshifting, suppression of termination or RNA editing is of considerable interest and remains to be determined.

Acknowledgements

Research in our laboratory is supported by research grants AI-20201 and AI-23173 from the National Institute of Allergy and Infectious Diseases. C.M.H. was supported by an NIH Training Program in Cell and Molecular Biology GM-08061.

References

1. Lamb, R.A. (1989) in The Influenza Viruses (Krug, R.M., ed.), pp. 1-87, Plenum Press.
2. Yamashita, M., Krystal, M. and Palese, P. (1989) Virology 171, 458-466.
3. Parvin, J.D. et al. (1989) J. Virol. 63, 5142-5152.
4. Krug, R.M. (1989) in The Influenza Viruses (Krug, R.M., ed.), pp. 89-152, Plenum Press.
5. Inglis, S.C., Barrett, T., Brown, C.M. and Almond, J.W. (1979) Proc. Natl Acad. Sci. USA 76, 3790-3794.
6. Lamb, R.A. and Choppin, P.W. (1979) Proc. Natl Acad. Sci. USA 76, 4908-4912.
7. Lamb, R.A. and Lai, C.-J. (1980) Cell 21, 475-485.
8. Briedis, D.J. and Lamb, R.A. (1982) J. Virol. 42, 186-193.
9. Nakada, S., Graves, P.N. and Palese, P. (1986) Virus Res. 4, 263-273.
10. Lamb, R.A., Lai, C.-J. and Choppin, P.W. (1981) Proc. Natl Acad. Sci. USA 78, 4170-4174.
11. Inglis, S.C. and Brown, C.M. (1981) Nucleic Acids Res. 9, 2727-2740.
12. Lamb, R.A., Zebedee, S.L. and Richardson, C.D. (1985) Cell 40, 627-633.
13. Yamashita, M., Krystal, M. and Palese, P. (1988) J. Virol. 62, 3349-3355.
14. Lamb, R.A. and Lai, C.-J. (1982) Virology 123, 237-256.
15. Lamb, R.A. and Lai, C.-J. (1984) Virology 135, 139-147.
16. Smith, D.B. and Inglis, S.C. (1985) EMBO J. 4, 2313-2319.
17. Plotch, S.J. and Krug, R.M. (1986) Proc. Natl Acad. Sci. USA 83, 5444-5448.
18. Agnis, C.H., Nemeroff, M.E. and Krug, R.M. (1989) Mol. Cell. Biol. 9, 259-267.
19. Alonso-Caplen, F.V. and Krug, R.M. (1991) Mol. Cell. Biol. 11, 1092-1098.
20. Collen, B.R. and Greene, W.C. (1990) Virology 178, 1-5.
21. Winter, G., Fields, S., Gait, M.J. and Brownlee, G.G. (1981) Nucleic Acids. Res. 9, 237-245.
22. Shaw, M.W., Choppin, P.W. and Lamb, R.A. (1983) Proc. Natl Acad. Sci. USA 80, 4879-4883.
23. Williams, M.A. and Lamb, R.A. (1986) Mol. Cell. Biol. 6, 4317-4328.
24. Williams, M.A. and Lamb, R.A. (1989) J. Virol. 63, 28-35.
25. Kozak, M. (1988) J. Cell. Biol. 107, 1-7.
26. Horvath, C.M., Williams, M.A. and Lamb, R.A. (1990) EMBO J. 9, 2659-2667.
27. Jackson, R.J., Howell, M.T. and Kaminski, A. (1990) Trends Biochem. Sci. 15, 477-483.
28. Chang, L.-J., Ganem, D. and Varmus, H.E. (1990) Proc. Natl Acad. Sci. USA 87, 5158-5162.
29. Normark, S. et al. (1983) Annu. Rev. Genet. 17, 499-525.
30. Das, A. and Yanovsky, C. (1984) Nucleic Acids Res. 12, 4757-4768.
31. Cattaneo, R. (1989) Trends Biochem. Sci. 14, 165-167.
32. Jackson, R., Madhani, H.D., Masiarz, F.R. and Varmus, H.E. (1988) Cell 55, 437-458.
33. Breier, I.P., Digard, P. and Inglis, S.C. (1989) Cell 57, 537-547.
34. Strauss, E.G., Rice, C.M. and Strauss, J.H. (1983) Proc. Natl Acad. Sci. USA 80, 5271-5275.
35. Thomas, S.M., Lamb, R.A. and Paterson, R.G. (1988) Cell 54, 891-902.
36. Paterson, R.G. and Lamb, R.A. (1990) J. Virol. 64, 4137-4145.

R.A. LAMB IS IN THE HOWARD HUGHES MEDICAL INSTITUTE AND DEPARTMENT OF BIOCHEMISTRY, MOLECULAR BIOLOGY AND CELL BIOLOGY, AND M. C. HORBATH IS IN THE DEPARTMENT OF BIOCHEMISTRY, MOLECULAR BIOLOGY AND CELL BIOLOGY, NORTHWESTERN UNIVERSITY, EVANSTON, IL 60208-3500, USA.