

Walking along the rabies genome: Is the large G-L intergenic region a remnant gene?

(nucleotide sequence/rabies proteins/transcription signals/unsegmented negative-strand RNA viruses)

NOËL TORDO*, OLIVIER POCH†, ALAIN ERMINE*, GÉRARD KEITH†, AND FRANÇOIS ROUGEON‡

*Unité Rage Recherche and †Unité de Génétique et Biochimie du Développement, Institut Pasteur, 25 rue du Docteur Roux, 75724 Paris, France; and ‡Groupe Biochimie II, Institut de Biologie Moléculaire et Cellulaire, 15 rue Descartes, 67084 Strasbourg, France

Communicated by André Lwoff, January 27, 1986

ABSTRACT Rabies cDNA clones, obtained by "walking along the genome" using two successive DNA primers, have allowed the sequence determination of the genes encoding the N, M1, M2, G, and the beginning of the L protein as well as the rabies intergenic regions. Start and stop transcription signals located at the border of each gene encoding a protein have been identified and are similar to the corresponding signals from vesicular stomatitis virus (VSV) and Sendai virus. Except for limited stretches of the nucleoprotein, there is no homology between corresponding structural proteins of these three viruses. Rabies intergenic regions are variable both in length and sequence. Evidence for the existence of a remnant protein gene in the 423 nucleotide long G-L intergenic region is presented. This finding is discussed in terms of the evolution of unsegmented negative-strand RNA viruses.

Unsegmented negative-strand RNA viruses are grouped into two families, Rhabdoviridae and Paramyxoviridae. Although they differ widely in their host range and their interaction with the host cell, all these viruses share a similar genomic organization and use the same multiplication strategy during their lytic cycle. The RNA genome is first used as template for the synthesis of monocistronic transcripts, then for replication into positive-strand genomes that are then replicated into the negative-strand genomes and these are encapsulated in progeny virion proteins (1-3). This implies that genomic signals must exist to distinguish between transcription and genome replication. In vesicular stomatitis virus (VSV) and Sendai virus two conserved sequences, probably recognized as initiation and termination sites for transcription, have been characterized at every gene junction (4, 5). The deduced mRNA 5'-start and 3'-end consensus sequences are very similar in the two viruses as are the di- or trinucleotide intergenic regions whose complement does not appear in the mRNA.

Rabies virus belongs to the Rhabdoviridae family. Its 12,000-base genome is sequentially transcribed into one leader RNA at the 3' end (6) and five monocistronic mRNAs (7, 8) encoding successively the N, M1, M2, G, and L proteins (9). In the rabies virion two structural and functional units can be distinguished: (i) the lipidic envelope that contains the transmembrane glycoprotein G, the major viral antigen (10), and the M2 protein that is located on the innerside (11) and (ii) the nucleocapsid core that is transcriptionally active (12) and is composed of the RNA genome always associated with the nucleoprotein N, the phosphoprotein M1, and the polymerase L.

We undertook the cloning of the complete rabies genome to define the structure of the viral genes and to characterize transcriptional and replicational regulatory sequences. The strategy we employed was to obtain cDNA clones by walking

along the genome using two DNA primers. From the nucleotide sequence, we have deduced the leader RNA and the N, M1, M2, and G protein sequences as well as the beginning of the L protein sequence. Furthermore, we have characterized consensus sequences, presumed start and stop transcription signals bordering each mRNA. Comparison of the rabies genome with those of other unsegmented negative-strand RNA viruses indicates extensive divergence except for the transcription regulatory signals and limited stretches of the nucleoprotein (30). The presence of a remnant protein gene between the G and L cistrons in rabies virus suggests a rapid evolution of this region among Rhabdoviridae.

MATERIALS AND METHODS

Cloning and Sequencing. Purification of the rabies RNA genome, conditions for cDNA synthesis, and cloning in pBR322 plasmid vectors have been extensively described elsewhere (30). Sequence determination was performed by the chain-terminating inhibitors method (13) after inserting endonuclease restriction fragments of the cDNA inserts into M13 vectors (14).

Computer Analysis of Amino Acid Sequences. Rabies proteins hydrophobicity profile was determined using the program of Kyte and Doolittle (15). Comparison with the Protein Identification Resource of the National Biomedical Research Foundation (NBRF-PIR)[§] and PGtrans protein data banks (16) was carried out using the program of Wilbur and Lipman (17) with the following parameters: K-tuple size = 2, window size = 40, and gap penalty = 2.

RESULTS

cDNA clones of the rabies RNA genome have been obtained using the following two DNA primers (Fig. 1): the first was complementary to the 3' extremity of the RNA genome, and the second was deduced from the end of the nucleotide sequence of the M2 gene. We have determined the nucleotide sequence of 5500 bases covering the 3' moiety of the rabies genome (Fig. 2). The sequence contains four major nonoverlapping open reading frames corresponding to the N, M1, M2, and G proteins, respectively. Other open reading frames either on the plus or the minus strand, never consisted of more than 50 codons, except one that was 102 codons in another phase of the M1 gene (nucleotide position 1896 to 2201).

From S1 nuclease protection experiments, we have located the 5' start site of the mRNA for the N protein around position 59 on the genome (30). Consequently, the first open reading frame, encoding the nucleoprotein N, extends between the ATG codon in position 71 and the TAA codon at position 1421. The 450-amino acid N protein (M_r 50,500) and

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "advertisement" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviations: VSV, vesicular stomatitis virus.

[§]National Biomedical Research Foundation (1986) Protein Sequence Database (Natl. Biomed. Res. Found., Washington, DC 20007), Tape Release 7.0.

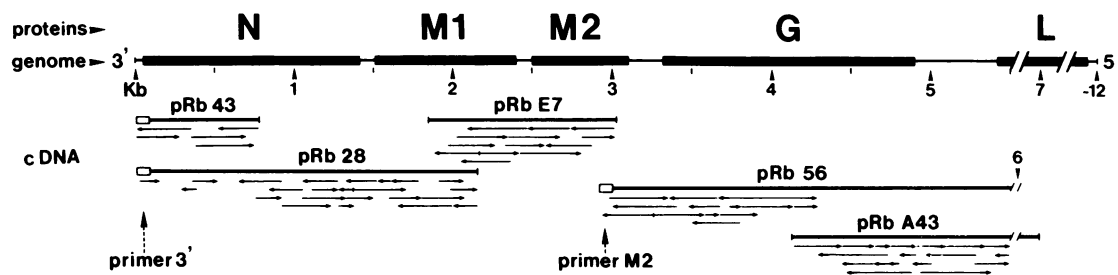


FIG. 1. Cloning and sequencing strategies of the rabies genome.

the leader RNA encoded upstream from the mRNA for nucleoprotein N have been extensively studied elsewhere (30).

The second open reading frame [position 1514 (ATG) to 2405 (TAA)] encodes the phosphoprotein M1. The hydrophobicity profile of the corresponding 297-amino acid sequence (M_r 33,000) reveals this to be the most hydrophilic of the four proteins examined (15). Hydrophilic residues are dominant in the first two-thirds of the amino end of the protein. A particularly hydrophilic region is observed between amino acids 139 and 170 (nucleotide position 1928 to 2023). Interestingly, this region of the M1 protein contains 13 of the 40 putative serine and threonine phosphorylation sites. Since phosphate residues of the corresponding NS protein of VSV have been shown to be mostly anchored in the hydrophilic amino-terminal moiety (18, 19), it is possible that some of these serine and threonine residues could be used as phosphate acceptors in the rabies M1 protein.

The sequence encoding the M2 protein extends from position 2496 to 3101. This corresponds to a 202-amino acid polypeptide with a M_r of 23,000. The sequence exhibits a central segment of 19 amino acids (nucleotide position 2760 to 2816) very rich in hydrophobic residues. The hydrophobicity program (15) predicts that this region has a high probability of being membrane bound. This is of note since it has been shown that in the case of negative-strand RNA viruses, the membrane protein M is located in the inner side of the lipidic envelope (11, 20) and could, therefore, interact with the lipid bilayer and the ribonucleoprotein core (21). For VSV this interaction was shown to be at least partially hydrophobic (21, 22), although the amino acid sequence of the VSV M protein does not contain any long hydrophobic segments (23).

The transmembrane glycoprotein G responsible for the induction and binding of virus neutralizing antibodies (10), consists of 524 amino acids (M_r 58,500) encoded between an ATG in position 3318 and a TGA in position 4890. The amino acid sequence possesses the following two hydrophobic regions previously described in the glycoprotein structure of ERA and CVS strains of rabies virus (24, 25): the first 19 amino acids of the amino extremity that constitute the signal peptide that is absent in the mature protein (26); and the transmembrane segment located in the carboxyl-terminal part of the protein that separates the hydrophilic cytoplasmic domain from the glycosylated spike. Four potential glycosylation sites appear in the spike. Three of these are in the same location as in the ERA glycoprotein (Asn-37, -247, and -319; nucleotide position 3483, 4113, and 4329, respectively) but an additional site appears at Asn-158 (position 3846).

No significant homology was found when comparing M1, M2, and G proteins using data banks. Only the nucleoprotein N has been shown to contain homologous stretches, presumed essential in the nucleocapsid structure, between rabies virus and VSV (30).

Consensus Sequences at Boundaries of Structural Genes. To characterize a rabies mRNA 5'-start consensus sequence, we have compared the nucleotide sequence at the 5' side of each protein initiator codon with the 5'-start sequence of the

mRNA for the N protein (30). The conclusion, summarized in Fig. 3, indicates that a consensus mRNA start sequence of nine nucleotides exists 12–30 residues upstream of the translation initiation codon. The first four and the two last nucleotides of the consensus sequence are invariant; in position 5 and 6 there are two pyrimidine residues but cytosine is much more frequent than thymidine; finally the seventh position is variable. This organization is related to those of Sendai virus and VSV where one or two variable nucleotides, respectively, separate two conserved regions (4, 5). Rabies virus and VSV mRNA start consensus sequences that share five invariant positions appear to be more closely related.

We have investigated whether a consensus sequence also exists at the 3' end of each structural gene by analyzing the genomic region downstream of the translation stop codon. One oligo(A) run that constitutes the initiation of the mRNA poly(A) tail has been invariably found at the end of the gene coding for each protein of VSV and Sendai virus (A_7 for VSV, A_5 for Sendai virus). The rabies N-M1 and M1-M2 intergenic regions contain a single run, A-A-A-A-A-A at positions 1476 and 2469, respectively, and we, therefore, postulate that these represent the polyadenylation signals of the mRNA for the N and M1 proteins. Interestingly, the M2-G and G-L intergenic regions have two putative polyadenylation signals at positions 3193 or 3279 and 4957 or 5357, respectively. If we consider the size of the mRNA for the M2 protein (8), the second run A-A-A-A-A-A (position 3279) is more likely to be the polyadenylation site. Nevertheless, we cannot exclude the possibility that the first poly(A) run may also be used or represent a vestigial signal. In the case of the mRNA for the G protein, we can conclude from the nucleotide sequence that the first poly(A) sequence (position 4957) is used as polyadenylation signal (24). The comparison of the rabies mRNA 3' end has enabled us to construct a consensus sequence composed of two invariant nucleotides followed by a polyadenylation signal of seven, and exceptionally eight (mRNA for the G protein), adenosine residues. This consensus sequence is identical to that of VSV that is also comprised of two additional invariant nucleotides (Fig. 3).

Intergenic Regions. These are defined between the 3' end of one mRNA and the 5' start of the following one. The intergenic regions of VSV and Sendai virus are two (GA) or three (GAA) nucleotides long, respectively (4, 5). The rabies intergenic regions appear more variable both in length and nucleotide composition. Even though the N-M1 intergenic region has been shown to be identical to VSV (30), the following M1-M2, M2-G, and G-L intergenic regions are 5, 5, and 423 nucleotides long, respectively. Interestingly, they all start with a guanosine. At either end of the unusually long G-L intergenic region are two striking sequences. The first one, located 10 nucleotides downstream from the stop signal for the mRNA of the G protein (position 4974), resembles the rabies consensus mRNA start signal, while the poly(A) run at position 5357, 25 nucleotides upstream from the L gene, looks like the polyadenylation signal encountered at the end of each mRNA. Nevertheless, this G-L intergenic region cannot encode a peptide larger than 18 amino acids.

DISCUSSION

The results of this study on the primary structure of the rabies genome have enabled us to define sequences involved in the transcription of its structural genes and to outline fundamen-

tal aspects of its evolutionary relationship with other unsegmented negative-strand RNA viruses.

Two consensus sequences located at the boundary of each structural gene were deduced from analyses of nonprotein coding regions and by comparison with the 5'-start sequence

► LEADER RNA

ACGCTTAAACAACAGATCAAAAGAAAAAGACAGCGCTCAATGGCAGAGCAAAAAATGTAACACCTCTACAATGCGTGGCACAAGATTGTATTCAAAGTCAATAATCAGTGGTCTCTTT
 100

LysProGluIleIleValAspGlnTyrGluTyrLysTyrProAlaIleLysAspLeuLysLysProCysIleThrLeuGlyLysAlaProAspLeuAsnLysAlaTyrLysSerValLeu
 GAAGCTGAGATTATCGTGCATCAATAGTACAGTACCCTGCCATCAAAGATTGAAAAAGCCCTGTATAACTTAGGAAAGCTCCCGATTTAAATAAAGCATACAAGTCAGTTTT
 200

SerCysMetSerAlaAlaLysLeuAspProAspAspValCysSerTyrLeuAlaAlaAlaMetGlnPhePheGluGlyThrCysProGluAspTrpThrSerTyrGlyIleValIleAla
 ATCATGATGAGCGCCGCAAACTTGATCTCAGCATGTATGTTCTTATTGGCGCGGCAATGAGTTTTTTAGGGGACATGTCCGGAAGACTGGACCAGTATGGAATCGTGATTGC
 300

ArgLysGlyAspLysIleThrProGlySerLeuValGluIleLysArgThrAspValGluGlyAsnTrpAlaLeuThrGlyGlyMetGluLeuThrArgAspProThrValProGluHis
 ACGAAAAGAGATAAGATCACCACAGTTCTCTGCTGGAGATAAAACGTAAGTGTAGAGGGAAATGGGCTCTGACAGGAGGCACTGGAAGTACAAGACACCCACTGTCTCCAGCA
 400

AlaSerLeuValGlyLeuLeuLeuSerLeuTyrArgLeuSerLysIleSerGlyGlnSerThrGlyAsnTyrLysThrAsnIleAlaAspArgIleGluGlnIlePheGluThrAlaPro
 TCGCTCTAGTCCGTCTCTCTGAGTCTGTATAGTTGAGCAAAATATCCGGGCAAGCACTGTAACATAAGACAACATGACAGACAGATAGAGCAGATTTTGGAGACAGCC
 500 600

PheValLysIleValGluHisHisThrLeuMetThrThrHisLysMetCysAlaAsnTrpSerThrIleProAsnPheArgPheLeuAlaGlyThrTyrAspMetPhePheSerArgIle
 TTTTGTAAAATCGGAAACCATACTCTAATGACAACCTCACAATATGTCGTCTAATGGAGTACTATACCAACTTCAGATTTTGGCGGAACTTACATGTTTTTCTCCCGAT
 700

GluHisLeuTyrSerAlaIleArgValGlyThrValValThrAlaTyrGluAspCysSerGlyLeuValSerPheThrGlyPheIleLysGlnIleAsnLeuThrAlaArgGluAlaIle
 TGAGCATCTATATCAGCAATCAGATGCGGACAGTTGCTCACTGCTTATGAAGACTGTCAGACTGTGCTGTTACTGGGTTCAAAAACCAATCAATCTCACCGCTAGAGAGCAAT
 800

LeuTyrPhePheHisLysAsnPheGluGluGluIleArgArgMetPheGluProGlyGlnGluThrAlaValProHisSerTyrPheIleHisPheArgSerLeuGlyLeuSerGlyLys
 ACTATATTTCTCCACAAGAATTGAGGAAGATAAAGAAGATCTTTGAGCCAGGGCAGGACAGCTTCTCTCACTCTTATTTCATCCACTTCGTTTCTGAGTGGAA
 900

SerProTyrSerSerAsnAlaValGlyHisValPheAsnLeuIleHisPheValGlyCysTyrMetGlyGlnValArgSerLeuAsnAlaThrValIleAlaAlaCysAlaProHisGlu
 ATCTCTTATTCATCAAAATGCTGTTGGTCAAGTGTTCATCTCATTCACTTTGAGGATGCTATATGGTCAAGTCAGATCCCTAAATGCAACGGTTATGCTGCATGCTCTCATGA
 1000

MetSerValLeuGlyGlyTyrLeuGlyGluGluPhePheGlyLysGlyThrPheGluArgArgPhePheArgAspGluLysGluLeuGlnGluTyrGluAlaAlaGluLeuThrLysThr
 AATGCTGTTCTAGGGGCTATCTGGAGAGGAATTTCTCGGAAAGGCACATTTGAAGAAGATTCTTCAGAGATGAGAAAGAACTTCAAGAATACGAGCGGGCTGAAGTCAAAAGAC
 1100 1200

AspValAlaLeuAlaAspAspGlyThrValAsnSerAspAspGluAspTyrPheSerGlyGluThrArgSerProGluAlaValTyrThrArgIleIleMetAsnGlyGlyArgLeuLys
 TGACGTAGCACTGGCAGATGATGAACTCTCACTCTGACGACGAGGACTACTTCTCAGGTAACCAAGTCCGGAAGCTGTTTACTCGAATCAATAATGAATGGAGTCCGACTGAA
 1300

ArgSerHisIleArgArgTyrValSerValSerSerAsnHisGlnAlaArgProAsnSerPheAlaGluPheLeuAsnLysThrTyrSerSerAspSer***
 GAGATCCGCACATACGGAGATATGCTCAGTCACTCAATCACTCAAGCTCGTCAAACTCATGCGCGAGTTTCTAAACAAGACATATTCGAGTACTCATAAGAACTGAATAACAAAA
 1400

TGCCGAAATCTACGGATTGTATATCCATCATGAAAAAACTAACACCCCTCTTTTCGAAACCCCAACATGAGCAAGATCTTGTCAATCTAGTCTATTAGAGCGGGCTGCGC
 1500

M1► MetSerLysIlePheValAsnProSerAlaIleArgAlaGlyLeuAla

AspLeuGluMetAlaGluGluThrValAspLeuIleAsnArgAsnIleGluAspAsnGlnAlaHisLeuGlnGlyGluProIleGluValAspAsnLeuProGluAspMetGlyArgLeu
 CGATCTTGAGTGGCTGAAGAACTGTGATCTGATCAATAGAAATATCGAAGACAATCAGGCTCATCTCCAAGGGAAACCCATAGAAGTGGCAATCTCCCTGAGGATATGGGGGACT
 1600

HisLeuAspAspGlyLysSerProAsnProGlyGluMetAlaLysValGlyGluGlyLysTyrArgGluAspPheGlnMetAspGluGlyGluAspProSerLeuLeuPheGlnSerTyr
 TCACCTGGATGATGAAAAATCGCCCAACCTGGTGGAGTGGCAAGTGGGAGAAGGAAGTATCGAGAGGACTTTCAGATGGATGAGGAGAGGATCCTAGCTCTCTGCTCAGTATA
 1700 1800

LeuAspAsnValGlyValGlnIleValArgGlnIleArgSerGlyGluArgPheLeuLysIleTrpSerGlnThrValGluGluIleIleSerTyrValAlaValAsnPheProAsnPro
 CCTGGACAATGTTGGAGTCAAAATAGTCAGACAAATAAGTTCAGGAGAGAGATTTCTCAAGATATGGTCACAGCCGTAGAAGAGATTATATCTATGTCGGGTCAACTTTCCCAACCC
 1900

ProGlyLysSerSerGluAspLysSerThrGlnThrThrGlyArgGluLeuLysLysGluThrThrProThrProSerGlnArgGluSerGlnSerSerLysAlaArgMetAlaAlaGln
 TCCAGGAAAGTCTTACAGCAATAAATCAACCCAGACTACCGCCGAGAGCTCAAGAAAGGAGACAACCCCACTCTCTTCAGAGAGAAAGCCAAATCTCCGAAAGCCAGATGCGCGCTCA
 2000

ThrAlaSerGlyProProAlaLeuGluTrpSerAlaThrAsnGluGluAspAspLeuSerValGluAlaGluIleAlaHisGlnIleAlaGluSerPheSerLysLysTyrLysPhePro
 AACTGCTTCTGCGCTCCAGCCCTGAAATGCTGCGCCCAATGAAGAGATGATCTACGTGGGCTGAGATCGCTCACCAGATGCGAAAGTTTCTCAAAAAATATAAGTTTCC
 2100

SerArgSerSerGlyIleLeuLeuTyrAsnPheGluGlnLeuLysMetAsnLeuAspAspIleValLysGluAlaLysAsnValProGlyValThrArgLeuAlaArgAspGlySerLys
 CTCTCGATCTCAGGATACTCTTGTATAATTTGAGCAATTGAAATGAACCTTGATGATATAGTTAAAGAGGCAAAAAATGTACCAGGTGACCCGTTTACCCCTGACGGGTCCAA
 2200

LeuProLeuArgCysValLeuGlyTrpValAlaLeuAlaAsnSerLysLysPheGlnLeuLeuValGluSerAsnLysLeuSerLysIleMetGlnAspAspLeuAsnArgTyrThrSer
 ACTCCCCTAAGATGCTACTGGATGGTGCCTGGCCAACTTAAGAAATTCAGTGTGTTAGTCCGAATCCAACAAGCTGACTAAAATCATGCAAGATGACTTGAATCCGTATACAT
 2300 2400

Cys***

TGCTAACCGAACCTCTCCACTCAGTCCCTCTAGACAATAAGTCCGAGATGTCTAAAGTCAACAATGAAAAAAGCGCAACACCACTGATAAAATGAACCTTCTACGTAAGATAGTGA
 2500

M2► MetAsnPheLeuArgLysIleValLys

AsnCysArgAspGluAspThrGlnLysProSerProValSerAlaProLeuAspAspAspLeuTrpLeuProProGluTyrValProLeuLysGluLeuThrSerLysLysAsn
 AAAATTGCAGGGACGAGGACACTCAAAAACCTCTCCCGTCTCAGCCCTCTGGATGACGATGACTTGTGGCTCCACCCTGAACTACGCTCCCGCTAAAAGAAGTACAAGCAAGAGA
 2600

ArgArgAsnPheCysIleAsnGlyValLysValCysSerProAsnGlyTyrSerPheGlyIleLeuArgHisIleLeuArgSerPheAspGluIleTyrSerGlyAsnHisArgMet
 ACAGGAGAACTTTGTATCAACGGAGGGTTAAAGTGTAGCCCAATGGTTACTCGTTCGGATCTCGCCGACATCTGAGATCATTCGAGAGATATATCTGGGAATCATAGGA
 2700

ValGlyLeuValLysValIleGlyLeuAlaLeuSerGlyAlaProValProGluGlyMetAsnTrpValTyrLysLeuArgArgThrLeuIlePheGlnTrpAlaAspSerArgGly
 TGGTGGGTTAGTCAAGTACTTATGGACTGGCTTTGTCAGGAGCTCCAGTCCCTGAGGGCAATGAACTGGGTATACAAGTTGAGGAGAACCCCTATCTCCAGTGGGTGATTCAGGG
 2800

ProLeuGluGlyGluGluLeuGluTyrSerGlnGluIleThrTrpAspAspAsnThrGluPheValGlyLeuGlnIleArgValSerAlaLysGlnCysHisIleArgGlyArgIleTrp
 GCCCTCTGAAGGGGAGGACTGGAATACTCTCAGGAGATCACTTGGGATGATAATCTGACTTCTCGGATTCGAAATAAGAGTGAAGTCAAAAACAGTCTCATATCCGGGCGAGAACT
 2900 3000

CysIleAsnMetAsnSerArgAlaGlyGlnLeuTrpSerAspMetSerLeuGlnThrGlnArgSerGluGluAspLysAspSerSerLeuLeuLeuGlu***
 GGTGATCAACATGAACCTCAGAGCAGGTCAACTATGGTCTGACATGCTCTTCAGACACAAAGTCCGAAGAGGACAAAGATTCCTCTGCTTCTAGAATAATCAGATTATATCCCGC
 3100

AAATTTACTCTGTTTACTCTGAGGAGAGAACAATATGGGCTCAACTCAACCCCTGGGGCAATATAACAAAAAATGTTATGGTCCATAAACCCGCTGCATTTTCATCAAAGT
 3200

FIG. 2. (Figure continues on the opposite page.)

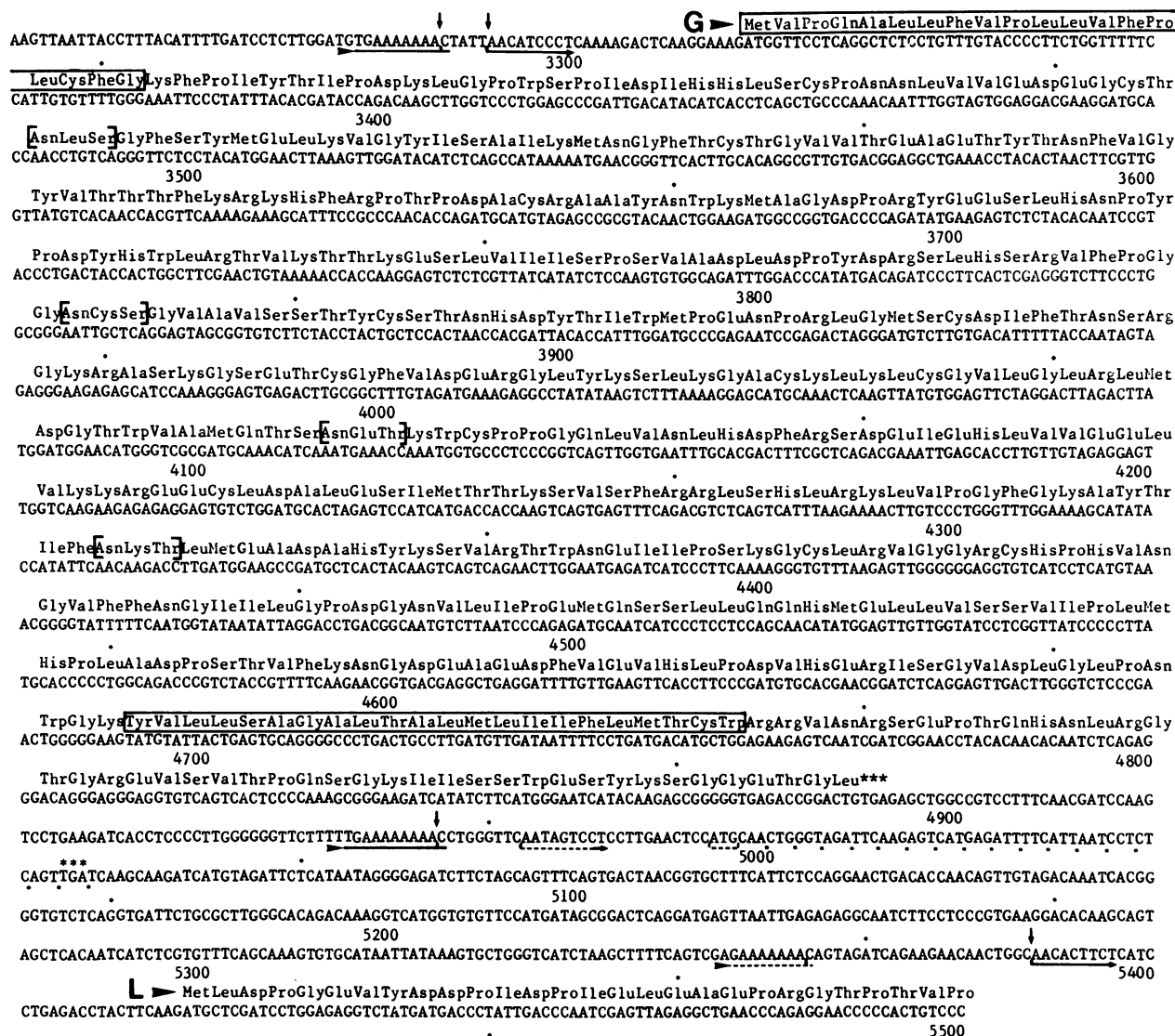


FIG. 2. Sequence of the first 5500 nucleotides of the rabies genome presented as DNA, positive (+) sense. Vertical arrows indicate probable mRNA extremities. mRNA 5'-start and 3'-end consensus sequences are underlined with a solid line. Alternative 3' end of the mRNA for the M2 protein as well as sequences similar to consensus sequences in the G-L intergenic region are underlined with a dotted line. Deduced amino acid sequences of N, M1, M2, G, and start of L proteins are shown. Solid lines box the hydrophobic segments in M2 and G proteins. The hydrophilic region of M1 protein is boxed with dotted lines and putative phosphorylation sites are indicated by crosses. Potential glycosylation sites of the G protein are between brackets.

of the mRNA for the nucleoprotein N (30) and the 3' end of the mRNA sequence of the glycoprotein G (24). These sequences most probably represent transcriptional initiation and termination signals. They are clearly similar to those of another Rhabdoviridae, VSV (4), and appear to retain a fair degree of homology with Sendai virus, a Paramyxoviridae (5). This homology between transcription signals strongly contrasts with the divergence found in other parts of the genome, in particular most of the structural genes, with the notable exception of the nucleoprotein N (30). This shows that unsegmented negative-strand RNA viruses, in spite of their high rate of evolution, have conserved the same basal genomic organization and only those signal sequences essential to their characteristic sequential transcription.

A remarkable finding arising from this study is the presence of a 423-nucleotide intergenic region located between the G and L cistrons. The existence of two sequences related to the rabies mRNA start and stop consensus sequences near its extremities suggests the possibility that this intergenic region is transcribed. Even though the corresponding transcript has never been reported, the presence of unresolved minor

mRNA species cannot be excluded (8). Assuming that this region is transcribed, the sizes of the rabies intergenic regions become 2, 5, 5, 9, and 24 residues, respectively. Even though a 19-nucleotide intergenic sequence has been reported in a bicistronic mRNA of respiratory syncytial virus (27), to our knowledge the rabies genome is the first example where the sequences separating the transcriptional units differ so widely both in length and nucleotide composition. The role of this variation, which contrasts with the stability of VSV and Sendai virus intergenic regions (4, 5), is unclear.

As to the functional significance of the putative additional gene between the G and L cistrons, it is doubtful that it encodes a regulatory RNA such as the leader RNA since sequences related to the rabies mRNA transcription signals are found at its boundaries. On the other hand, a structural role is unlikely since transcription of this intergenic region could not lead to a large polypeptide. If we consider that the flanking sequences result from the degeneration of consensus transcription signals and that this G-L intergenic region is the sole large genomic region extensively blocked in all reading frames, we must assume that it represents a remnant protein

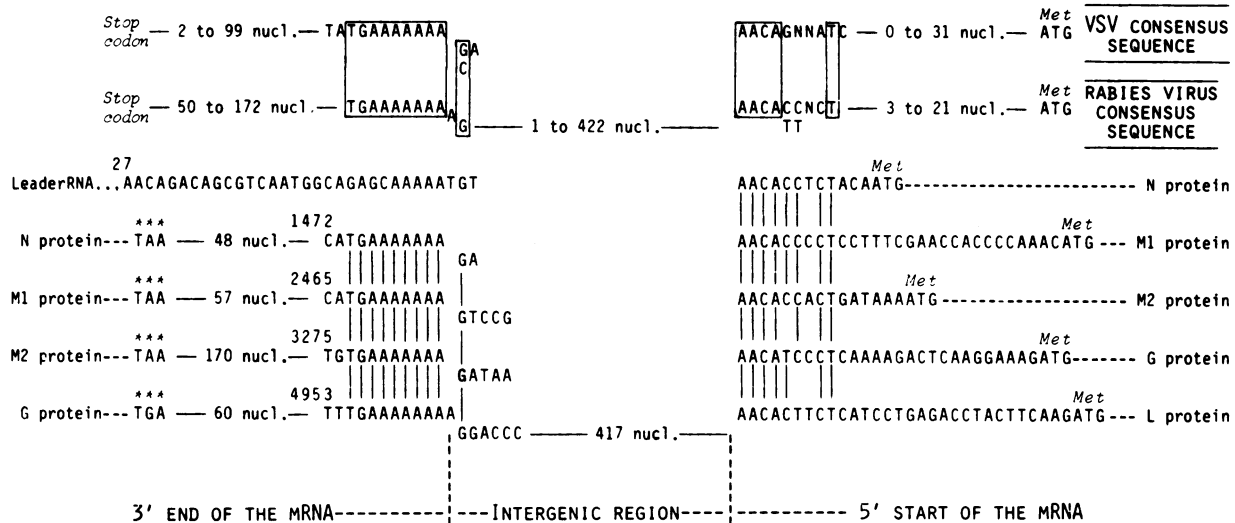


FIG. 3. Deduced rabies mRNA 5'-start and 3'-end consensus sequences. Intergenic regions are indicated in DNA negative (–) sense.

gene. This hypothesis is strongly supported by the identification of a sixth gene, similar in length to the rabies G-L intergenic region, between the *G* and *L* cistrons of a fish Rhabdoviridae, infectious hematopoietic necrosis virus (28, 29). This gene encodes the M_r 12,000 nonvirion structural protein NV.

Although all Rhabdoviridae share the same genomic organization, the G-L intergenic region appears to vary extensively in this family. In infectious hematopoietic necrosis virus this region encodes the NV protein, whose role remains unknown, while in VSV it is next to the consensus intergenic dinucleotide G-A. Between these two extreme situations, the rabies G-L intergenic region, similar in length to those of infectious hematopoietic necrosis virus but with multiple stop codons, probably represents an intermediate stage of Rhabdoviridae evolution. In this context, it is interesting to note that, unlike Rhabdoviridae, most Paramyxoviridae have two glycoprotein genes. The additional one is located in a region corresponding to the rabies G-L intergenic region. This emphasizes the plasticity of this region in unsegmented negative-strand RNA viruses. The rabies G-L intergenic region is, to our knowledge, the first example of a pseudogene so far characterized in this group of viruses.

We are deeply indebted to Prof. P. Sureau in whose laboratory this work was carried out. We are grateful to M. A. Akimenko for helpful discussions throughout this work and to M. Goodhardt, G. Langsley, and C. Kean for critical reading of the manuscript. This work was supported by grants from the Centre National de la Recherche Scientifique (ATP 3682 and 955308), from the Institut National de la Santé et de la Recherche Médicale (CRL 811007), and from the Comité Consultatif des Applications de la Recherche de l'Institut Pasteur.

- Banerjee, A. K., Abraham, G. & Colonno, R. J. (1977) *J. Gen. Virol.* **34**, 1–8.
- Ball, L. A. & Wertz, G. W. (1981) *Cell* **26**, 143–144.
- Lazzarini, R. A., Keene, J. D. & Schubert, M. (1981) *Cell* **26**, 145–154.
- Rose, J. K. (1980) *Cell* **19**, 415–421.

- Gupta, K. C. & Kingsbury, D. W. (1984) *Nucleic Acids Res.* **12**, 3829–3841.
- Kurilla, M. G., Cabradilla, C. D., Holloway, B. P. & Keene, J. D. (1984) *J. Virol.* **50**, 773–778.
- Coslett, G. D., Holloway, B. P. & Obijeski, J. F. (1980) *J. Gen. Virol.* **49**, 161–180.
- Holloway, B. P. & Obijeski, J. F. (1980) *J. Gen. Virol.* **49**, 181–195.
- Flamand, A. & Delagneau, J. F. (1978) *J. Virol.* **28**, 518–523.
- Wiktor, T. J., Gyorgy, E., Schlumberger, H. D., Sokol, F. & Koprowski, H. (1973) *J. Immunol.* **110**, 269–276.
- Cox, J. F., Weiland, F., Dietzschold, B. & Schneider, L. G. (1981) in *The Replication of Negative Strand Viruses*, eds. Bishop, D. H. L. & Compans, R. W. (Elsevier/North-Holland, New York), pp. 639–645.
- Kawai, A. (1977) *J. Virol.* **24**, 826–835.
- Sanger, F., Nicklen, S. & Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. USA* **74**, 5463–5467.
- Messing, J., Crea, R. & Seeberg, P. H. (1981) *Nucleic Acids Res.* **9**, 309–321.
- Kyte, J. & Doolittle, R. F. (1982) *J. Mol. Biol.* **157**, 105–132.
- Claverie, J. M., Sauvaget, I. & Bougueleret, L. (1985) *Biochimie* **67**, 437–443.
- Wilbur, W. J. & Lipman, D. J. (1983) *Proc. Natl. Acad. Sci. USA* **80**, 726–730.
- Gill, D. S. & Banerjee, A. K. (1985) *J. Virol.* **55**, 60–66.
- Bell, J. C. & Prevec, L. (1985) *J. Virol.* **54**, 697–702.
- Lenard, J. & Compans, R. W. (1974) *Biochim. Biophys. Acta* **344**, 51–94.
- Wilson, T. & Lenard, J. (1981) *Biochemistry* **20**, 1349–1354.
- Capone, J. & Ghosh, H. P. (1984) *Can. J. Biochem. Cell Biol.* **62**, 1174–1180.
- Rose, J. K. & Gallione, C. J. (1981) *J. Virol.* **39**, 519–528.
- Anilionis, A., Wunner, W. H. & Curtis, P. J. (1981) *Nature (London)* **294**, 275–278.
- Yelverton, E., Norton, S., Obijeski, J. F. & Goeddel, D. V. (1983) *Science* **219**, 614–620.
- Lai, C. Y. & Dietzschold, B. (1981) *Biochem. Biophys. Res. Commun.* **103**, 536–542.
- Elango, N., Satake, M. & Venkatesan, S. (1985) *J. Virol.* **55**, 101–110.
- Kurath, G. & Leong, J. C. (1985) *J. Virol.* **53**, 462–468.
- Kurath, G., Ahern, K. G., Pearson, G. D. & Leong, J. C. (1985) *J. Virol.* **53**, 469–476.
- Tordo, N., Poch, O., Ermine, A. & Keith, G. (1986) *Nucleic Acids Res.* **14**, 2671–2683.