JOURNAL OF VIROLOGY, Aug. 1981, p. 529-535
0022-538X/81/080529-07$02.00/0

Vol. 39, No. 2

Nucleotide Sequences of the mRNA's Encoding the Vesicular Stomatitis Virus N and NS Proteins

CAROL J. GALLIONE,[1] JONATHAN R. GREENE,[2] LINDA E. IVERSON,[1,2] AND JOHN K. ROSE[1]*

Tumor Virology Laboratory, The Salk Institute, San Diego, California 92138,[1] and Department of Biology, University of California, San Diego, La Jolla, California 92093[2]

Received 24 March 1981/Accepted 20 April 1981

The complete nucleotide sequences of the vesicular stomatitis virus (VSV) mRNA's encoding the N and NS proteins have been determined from the sequences of cDNA clones. The mRNA encoding the N protein is 1,326 nucleotides long, excluding polyadenylic acid. It contains an open reading frame for translation which extends from the 5'-proximal AUG codon to encode a protein of 422 amino acids. The N mRNA is known to contain a major ribosome binding site at the 5'-proximal AUG codon and two other minor ribosome binding sites. These secondary sites have been located unambiguously at the second and third AUG codons in the N mRNA sequence. Translational initiation at these sites, if it in fact occurs, would result in synthesis of two small proteins in a second reading frame. The VSV mRNA encoding the NS protein is 815 nucleotides long, excluding polyadenylic acid, and encodes a protein of 222 amino acids. The predicted molecular weight of the NS protein (25,110) is approximately one-half of that predicted from the mobility of NS protein on sodium dodecyl sulfate-polyacrylamide gels. Deficiency of sodium dodecyl sulfate binding to a large negatively charged domain in the NS protein could explain this anomalous electrophoretic mobility.

The genome of vesicular stomatitis virus (VSV) is a single negative strand of RNA about 11,000 nucleotides long (5, 15). The viral N, NS, and L proteins associate with the genomic RNA to form the helical nucleocapsid structure within the enveloped virion (reviewed by Wagner [22]). The transmembrane glycoprotein (G) and the peripheral membrane protein (M) are separable from the nucleocapsid core after disruption of the viral membrane with detergent (22). The L and NS proteins can be dissociated from the nucleocapsid and have been shown to be required for mRNA synthesis from a template consisting of N protein and genomic RNA (3). The synthesis, capping, methylation, and polyadenylation of the five VSV mRNA's are probably carried out by the L and NS proteins during transcription from the N protein-genome RNA complex (reviewed by Banerjee et al. [1]). A knowledge of the amino acid sequences of the N and NS proteins should facilitate further structural and functional analyses of these proteins and will be critical to a molecular understanding of their interactions with genomic RNA and other virion components. As part of an analysis of VSV mRNA and protein structure we have determined the nucleotide sequences of the VSV mRNA's encoding the N and NS proteins from cDNA clones. The amino acid sequences of the proteins are predicted from these sequences. Features of the mRNA and protein sequences are discussed.

MATERIALS AND METHODS

Details of mRNA synthesis, cDNA cloning, and DNA sequencing are provided in the accompanying paper (18).

Primer isolation. Isolation of the primer illustrated in Fig. 3 from pNS173 DNA was as follows. pNS173 DNA (50 μg) was digested with *Alu*I and treated with alkaline phosphatase, and the 198-base-pair blunt-ended *Alu*I fragment (56 base pairs of pBR322 DNA linked to 142 base pairs of insert DNA) was purified by gel electrophoresis followed by electroelution. The 5'-ends of this fragment were labeled with ^{32}P by using polynucleotide kinase and [γ-^{32}P]ATP. The fragment was digested with *Hin*fI, which cuts once and generates staggered ends. The products were then separated by electrophoresis on a sequencing gel (40 cm by 16 cm by 0.35 mm). The 53-nucleotide, single-stranded primer fragment was located by autoradiography and eluted by soaking the gel slice in 0.5 ml of 0.4 M sodium acetate for 5 h. The labeled primer fragment separates from the unlabeled complementary strand because it is three nucleotides shorter.

Primer extension and sequencing. Primer derived from approximately 25 μg of pNS173 DNA (10^6 dpm) was combined with 100 μg of total VSV mRNA

prepared as described previously (18). The mixture of primer DNA and mRNA was then precipitated with ethanol. The pellet was washed with 80% ethanol, dried, and suspended in 10 μl of 4× reverse transcription buffer containing 200 mM Tris-hydrochloride (pH 8.4), 40 mM $MgCl_2$, 120 mM β-mercaptoethanol, 440 mM KCl, and each of the four dNTP's at 2 mM. This mixture was incubated at 40°C for 10 min followed by the addition of 20 μl of water and 200 U of reverse transcriptase in 10 μl. Incubation was for 1 h at 42°C followed by phenol extraction and ethanol precipitation. The pellet was resuspended in 100 μl of 10 mM Tris (pH 7.4) containing 10 U of RNase A (Worthington Diagnostics) and heated to 90°C for 1 min. The extended primer was then purified by electrophoresis on a 20% polyacrylamide gel. The gel was subjected to autoradiography, and the largest band (250 nucleotides) was excised, electroeluted, and sequenced.

RESULTS

Determination of the N mRNA sequence. The procedures used for generation of nearly full-length cDNA clones from the VSV N and NS mRNA's are described in the accompanying paper (18). Two cDNA clones designated pN3 and pN4 were used for the N mRNA sequence analysis reported here. The inserts in both of these clones were about 1,350 nucleotides long, including 30 to 40 dG:dC tails at each end, and were flanked by *Pst*I sites. Determination of the nucleotide sequences at the ends of the pN4 insert DNA showed that it contained all but 61 nucleotides from the 5′ end of the mRNA and the complete 3′ mRNA sequence. This was clear from comparison of the insert DNA sequence with published sequences from both the 5′ and 3′ ends of the N mRNA (11, 13, 19).

The basic strategy for sequencing the N gene was first to isolate the two large *Hae*III fragments (Fig. 1) from the insert DNA and then to determine the sequences of the separated strands of each fragment by the Maxam-Gilbert procedure (9). Combined with data published previously (11, 13, 16, 19), these sequences enabled us to establish a tentative sequence for the entire gene. A restriction map was generated from this sequence, and the fragments needed to verify the sequence on both DNA strands were isolated and sequenced (Fig. 1). We also sequenced through all sites used as sequencing endpoints and thus discovered the short sequence between the two central *Hae*III sites.

The predicted sequence of the N mRNA is shown in Fig. 2. This 1,326-nucleotide sequence contains a single open reading frame for translation beginning at the AUG in positions 14 through 16 (within the previously identified ribosome binding site [16]) and ending at positions 1,280 through 1,282. The two other translational reading frames are blocked extensively throughout the sequence. Comparison of the 3′-terminal N mRNA sequences from the VSV Indiana and New Jersey serotypes suggested that the reading frame shown here was correct (12). The predicted N protein sequence is 422 amino acids long with a calculated molecular weight of 47,355.

Determination of the NS mRNA sequence. A cDNA clone of the VSV NS mRNA (pNS173, reference 9) was used for the majority of the sequence analyses reported here. A restriction map of the insert DNA illustrating the sites for restriction enzymes used in the sequence analysis is shown in Fig. 3. To obtain the 5′-terminal NS mRNA sequence that was not present in the cDNA clone, we isolated a 5′ ^{32}P-labeled DNA primer from the position indicated in Fig. 3 (see above). This primer was then extended to the 5′ end of the NS mRNA by using reverse transcriptase. The sequence obtained from this elongated primer extended into the 5′-terminal sequence of the NS mRNA that was known previously from the sequence of the ribosome binding site (16). This analysis showed that pNS173 lacks 126 nucleotides of the 5′ NS mRNA sequence. An additional clone, pNS319,

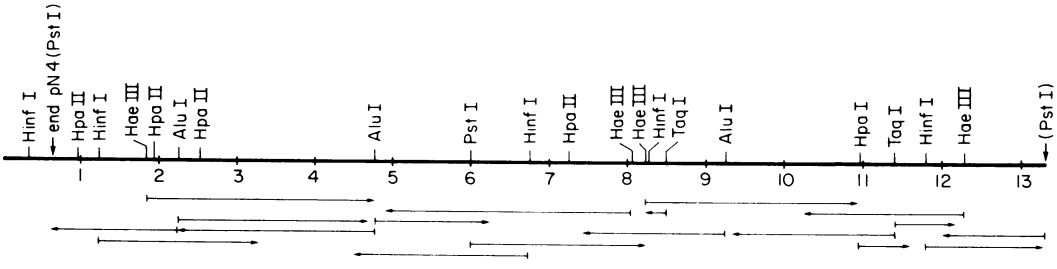

FIG. 1. *Restriction map of the pNF4 insert DNA. Sites for restriction enzymes that were used in sequencing are shown. The left end corresponds to the 5′ end of the N mRNA. Arrows indicate the regions sequenced and the direction of sequencing. Terminal PstI sites are at the junctions of pBR322 sequences with the dG:dC tails. The lengths of the dG:dC tails are not indicated in this figure or in Fig. 3.*

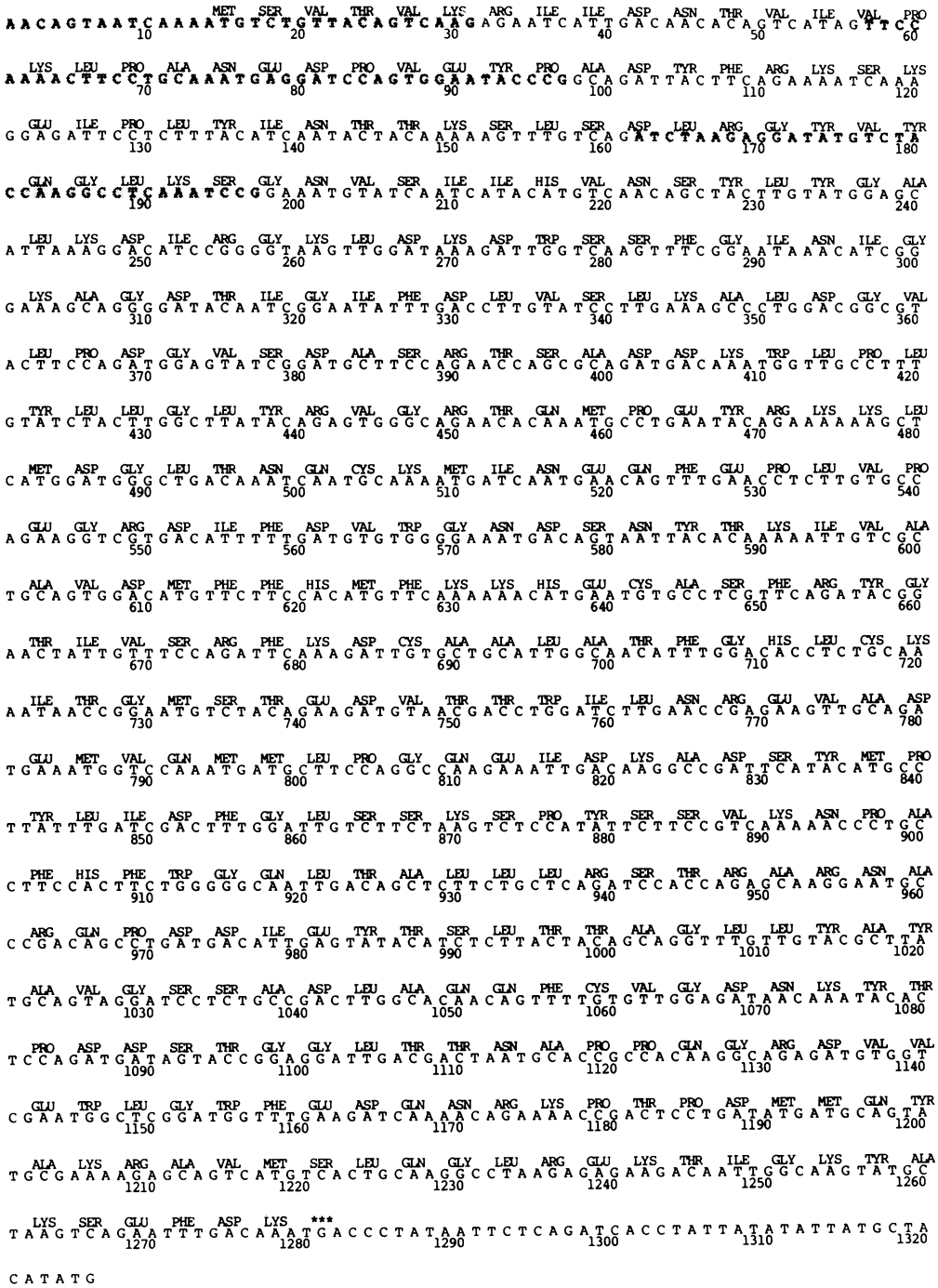

FIG. 2. *Nucleotide sequence of the VSV mRNA encoding the N protein and the predicted N protein sequence. The shaded nucleotide sequences represent the three ribosome binding sites in the N mRNA.*

was isolated which has the complete NS mRNA sequence except for 32 nucleotides from the 5′ end. The 5′ sequence determined from this clone is indicated in Fig. 3.

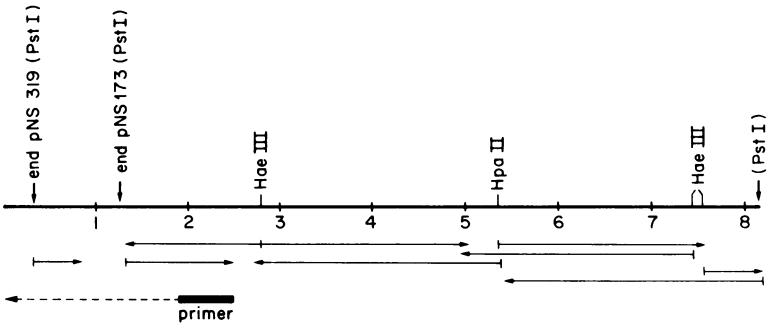

FIG. 3. *Restriction map of the pNS173 insert DNA. Sites for restriction enzymes that were used in sequencing are shown. The heavy line indicates the DNA segment used as a primer on the NS mRNA to obtain the 5'-terminal mRNA sequence.*

DISCUSSION

Translation of the N mRNA. The nucleotide sequence presented here for the VSV mRNA encoding the N protein predicts a protein of 422 amino acids initiated at the first AUG codon in the mRNA (Fig. 2). Our previous analysis of ribosome-protected initiation sites in N mRNA showed that three separate sites were protected within the sequence (16). The major site occurred at the 5'-proximal AUG codon and contained about 75% of the protected material. The minor sites were each found to contain 10 to 15% of the protected material. Partial sequence analysis showed that both minor sites contained single AUG codons. The complete sequence of the N mRNA contains 45 AUG codons in all three frames. We compared the partial sequences of the T1 oligonucleotides from the minor ribosome binding sites with the sequences around each AUG codon. These partial sequences can be aligned perfectly at the second and third AUG codons in the mRNA (Fig. 4), but not at any of the other AUG codons. An assignment of one of these sites to the second AUG codon was made previously on the basis of partial sequence information for the N mRNA (17).

The generally accepted view of eucaryotic protein synthesis initiation (reviewed in reference 8) is that eucaryotic ribosomes initiate only at single sites on mRNA's. These sites are usually, but not always, the first AUG in the mRNA (8). The fact that ribosomes bind only at the three 5'-proximal AUG codons in the N mRNA rather than at any of the 42 other AUG codons suggests that this binding results from "entry" of ribosomes at the 5' end of the mRNA. Presumably, features of the mRNA structure (such as distance from the 5' end) determine the extent of binding at the three sites. These multiple binding sites could be artifacts of the in vitro binding

A

```
  60          70          80          90
TTCCAAAACTTCCTGCAAATGAGGATCCAGTGGAATACCCGG
------------YUG       AG(G)     UG(G)
                YAAAUG    ---YAG    ------CG(G)
```

B

```
      170         180         190
ATCTAAGAGGATATGTCTACCAAGGCCTCAAATCCGG
---YAAG    -YAUG          ---------CG(G)
       AG(G)    -----YAAG(G)
```

FIG. 4. *Alignment of T1 oligonucleotide sequences with N mRNA sequences. Partial sequences of two secondary ribosome binding sites from the N mRNA (16) are shown aligned with the appropriate portions of the N mRNA sequence. Y indicates a pyrimidine, and parentheses indicate nearest neighbors. Sizes of the T1 oligonucleotides predicted are consistent with the migration of each oligonucleotide on the two-dimensional homochromatography system (16). Numbers are relative to the 5' end of the N mRNA.*

conditions, although we did not observe secondary binding sites in other VSV mRNA's under identical conditions. If ribosomes initiated translation at these sites, only small proteins of 30 and 23 amino acids would be produced. Such small proteins would have gone undetected previously. Because the amino acid sequences of these putative proteins can be predicted from the N mRNA sequence, it may be possible to detect them by labeling VSV-infected cells with the appropriate amino acids.

NS mRNA and protein structure. The nucleotide sequence predicted here for the VSV mRNA encoding the NS protein is 815 nucleotides long, excluding polyadenylic acid. It encodes a protein of 222 amino acids (25,110 daltons) initiated from the 5'-proximal AUG codon in the mRNA (Fig. 4). Previous estimates of the molecular weight of NS from its migration on sodium dodecyl sulfate-polyacrylamide gels

range from 39,000 to 50,000 (see reference 7). This discrepancy suggests that NS has an unusual monomeric structure or that it perhaps migrates as a dimer. Phosphorylation of NS cannot explain its anomalous migration because removal of phosphate residues from NS decreases rather than increases its mobility on sodium dodecyl sulfate-polyacrylamide gels (4). Inspection of the distribution of charged residues within the predicted NS protein sequence does reveal the unusual feature of a large domain containing 18 negatively charged residues and no positively charged residues (Fig. 5). A deficiency of sodium dodecyl sulfate binding to this negatively charged domain might explain the anomalous electrophoretic mobility.

Codon usage. As found for the G and M mRNA's (18), there is also a deficiency in the CG dinucleotide in the N and NS mRNA's. One would predict that 58 CGs would occur at random in the N mRNA coding region and 33 would occur in the NS coding region, whereas only 30 and 16, respectively, are found. The codon usage for both mRNA's is shown in Fig. 6. In the N mRNA there is a clear bias against arginine codons of the form CGN and against NCG codons (N = A, G, C, or T). A similar bias is also seen between adjacent codons where 31 junctions of the type NNC-GNN would be expected at random, whereas only 18 are found. In the NS mRNA there is no bias against the CGN codons, but the NCG codons are deficient. Codon junctions NNC-GNN are strongly deficient in NS, as only 3 are found where 14 would be expected at random. Deficiency in the CG dinucleotide may prove to be a general feature in the sequences of RNA genomes of eucaryotic viruses as it has been observed in the mRNA encoding an influenza virus hemagglutinin (14).

Remainder of the genome. The sequences presented here, combined with the previously determined leader RNA (2), 3'-terminal genomic sequences (6, 11, 20), and intergenic sequences (10, 17), provide a continuous sequence of 4,723

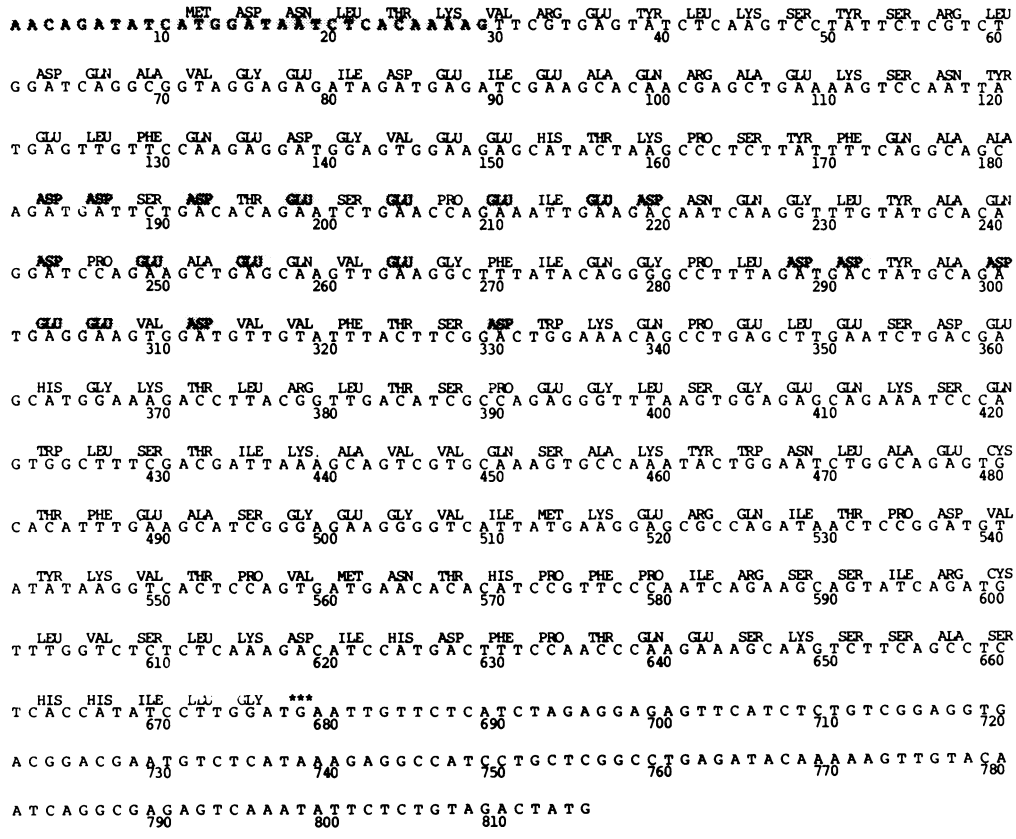

FIG. 5. *Nucleotide sequence of the VSV mRNA encoding the NS protein, and the predicted NS protein sequence. The shaded nucleotide sequence indicates the ribosome binding site. Negatively charged amino acids in the negatively charged domain are shaded.*

N Protein

	U	C	A	G	
U	Phe 8	Ser 8	Tyr 9	Cys 3	U
	9	7	12	2	C
	Leu 2	7	Term 0	Term 1	A
	15	2	0	Trp 7	G
C	Leu 9	Pro 9	His 2	Arg 1	U
	5	0	3	0	C
	3	8	Gln 10	2	A
	4	3	4	1	G
A	Ile 8	Thr 6	Asn 9	Ser 4	U
	9	5	7	2	C
	5	13	Lys 20	Arg 15	A
	Met 14	2	10	1	G
G	Val 5	Ala 6	Asp 22	Gly 3	U
	9	6	13	8	C
	6	15	Glu 13	16	A
	6	1	4	4	G

NS Protein

	U	C	A	G	
U	Phe 5	Ser 8	Tyr 7	Cys 1	U
	2	3	1	1	C
	Leu 3	1	Term 0	Term 1	A
	4	4	0	Trp 3	G
C	Leu 3	Pro 2	His 5	Arg 2	U
	3	1	1	1	C
	0	6	Gln 6	1	A
	2	2	8	1	G
A	Ile 3	Thr 4	Asn 4	Ser 3	U
	5	2	1	2	C
	3	5	Lys 6	Arg 2	A
	Met 3	1	7	0	G
G	Val 3	Ala 2	Asp 11	Gly 2	U
	4	2	7	1	C
	3	8	Glu 14	6	A
	4	1	14	2	G

FIG. 6. *Codon usage in the N and NS mRNA's. First positions are on the left, second positions are across the top, and third positions are on the right.*

nucleotides from the 3′ half of the VSV genome. Using the value of 11,278 nucleotides calculated by Repik and Bishop (15) for the size of the genome, this leaves 6,496 nucleotides to encode the L mRNA which terminates 59 nucleotides from the 5′ end of the genome (21). This sequence is more than sufficient to encode the L protein, which is estimated to have a molecular weight of about 190,000 (22).

ACKNOWLEDGMENTS

We thank Jim Ostlund, Joe Zwass, and Jeff Velten for help with computer analysis of sequence data and the members of the Tumor Virology Laboratory for their helpful suggestions. We are grateful to Carolyn Goller for typing this and the accompanying manuscript.

This work was supported by Public Health Service grants AI-15481 from the National Institute of Allergy and Infectious Diseases and CA-14195 from the National Cancer Institute.

LITERATURE CITED

1. **Banerjee, A. K., G. Abraham, and R. J. Colonno.** 1977. Vesicular stomatitis virus: mode of transcription. J. Gen. Virol. **34**:1-8.
2. **Colonno, R. J., and A. K. Banerjee.** 1978. Complete nucleotide sequence of the leader RNA synthesized in vitro by vesicular stomatitis virus. Cell **15**:93-101.
3. **Emerson, S. U., and Y. H. Yu.** 1975. Both NS and L proteins are required for in vitro RNA synthesis by vesicular stomatitis virus. J. Virol. **15**:1348-1356.
4. **Hsu, C. H., and D. Kingsbury.** 1980. Vesicular stomatitis virus morphogenesis is accompanied by covalent protein modifications, p. 613-622. *In* B. N. Fields, R. Jaenisch, and C. F. Fox (ed.), Animal virus genetics. Academic Press, Inc., New York.
5. **Huang, A. S., and R. R. Wagner.** 1966. Comparative sedimentation coefficients of RNA extracted from plaque-forming and defective particles of vesicular stomatitis virus. J. Mol. Biol. **22**:381-384.
6. **Keene, J. D., M. Schubert, and R. A. Lazzarini.** 1980. Intervening sequence between the leader region and the nucleocapsid gene of vesicular stomatitis virus RNA. J. Virol. **33**:789-794.
7. **Knipe, D., J. K. Rose, and H. F. Lodish.** 1975. Translation of individual species of vesicular stomatitis viral mRNA. J. Virol. **15**:1004-1011.
8. **Kozak, M.** 1978. How do ribosomes select initiation regions in mRNA? Cell **15**:1109-1123.
9. **Maxam, A. M., and W. Gilbert.** 1980. Chemical sequencing of DNA. Methods Enzymol. **65**:501-561.
10. **McGeoch, D. J.** 1979. Structure of the gene N:gene NS intercistronic junction in the genome of vesicular stomatitis virus. Cell **17**:673-681.
11. **McGeoch, D. J., and A. Dolan.** 1979. Sequence of 200 nucleotides at the 3′-terminus of the genome RNA of vesicular stomatitis virus. Nucleic Acids Res. **6**:3199-3211.
12. **McGeoch, D. J., A. Dolan, and C. R. Pringle.** 1980. Comparisons of nucleotide sequences in the genomes of the New Jersey and Indiana serotypes of vesicular stomatitis virus. J. Virol. **33**:69-77.
13. **McGeoch, D. J., and N. Turnbull.** 1978. Analysis of the 3′-terminal sequence of vesicular stomatitis virus N protein mRNA. Nucleic Acids Res. **5**:4007-4024.
14. **Porter, A. G., C. Barber, N. H. Carey, R. A. Hallewell, G. Threlfall, and J. S. Emtage.** 1979. Complete nucleotide sequences of an influenza virus haemagglutinin gene from cloned DNA. Nature (London) **282**:471-477.
15. **Repik, P., and D. H. L. Bishop.** 1973. Determination of the molecular weight of animal RNA viral genomes by nuclease digestions. J. Virol. **12**:969-983.
16. **Rose, J. K.** 1978. Complete sequences of ribosome recognition sites in VSV mRNAs. Cell **14**:345-353.
17. **Rose, J. K.** 1980. Complete intergenic and flanking gene sequences from the genome of vesicular stomatitis virus. Cell **19**:415-421.
18. **Rose, J. K., and C. J. Gallione.** 1981. Nucleotide sequences of the mRNA's encoding the vesicular stomatitis virus G and M proteins determined from cDNA clones containing the complete coding regions. J. Virol. **39**:519-528.

19. **Rose, J. K., and L. E. Iverson.** 1979. Sequences from the 3′ ends of vesicular stomatitis virus mRNA's as determined from cloned DNA. J. Virol. **32**:404-411.
20. **Rowlands, D. J.** 1977. Sequences of vesicular stomatitis virus RNA, N protein mRNA, and their junction. Proc. Natl. Acad. Sci. U.S.A. **76**:4793-4797.
21. **Schubert, M., J. D. Keene, R. C. Herman, and R. A. Lazzarini.** 1980. The site on the VSV genome specifying polyadenylation and the end of the L gene mRNA. J. Virol. **34**:550-559.
22. **Wagner, R. R.** 1975. Reproduction of rhabdoviruses, p. 1-93. *In* H. Fraenkel-Conrat and R. R. Wagner (ed.), Comprehensive virology. Plenum Publishing Corp., New York.