JOURNAL OF VIROLOGY, Aug. 1981, p. 519-528
0022-538X/81/080519-10$02.00/0

Vol. 39, No. 2

Nucleotide Sequences of the mRNA's Encoding the Vesicular Stomatitis Virus G and M Proteins Determined from cDNA Clones Containing the Complete Coding Regions

JOHN K. ROSE* AND CAROL J. GALLIONE

Tumor Virology Laboratory, The Salk Institute, San Diego, California 92138

Received 24 March 1981/Accepted 20 April 1981

The complete nucleotide sequences of the vesicular stomatitis virus mRNA's encoding the glycoprotein (G) and the matrix protein (M) have been determined from cDNA clones that contain the complete coding sequences from each mRNA. The G protein mRNA is 1,665 nucleotides long, excluding polyadenylic acid, and encodes a protein of 511 amino acids including a signal peptide of 16 amino acids. G protein contains two large hydrophobic domains, one in the signal peptide and the other in the transmembrane segment near the COOH terminus. Two sites of glycosylation are predicted at amino acid residues 178 and 335. The close correspondence of the positions of these sites with the reported timing of the addition of the two oligosaccharides during synthesis of G suggests that glycosylation occurs as soon as the appropriate asparagine residues traverse the membrane of the rough endoplasmic reticulum. The mRNA encoding the vesicular stomatitis virus M protein is 831 nucleotides long, excluding polyadenylic acid, and encodes a protein of 229 amino acids. The predicted M protein sequence does not contain any long hydrophobic or nonpolar domains that might promote membrane association. The protein is rich in basic amino acids and contains a highly basic amino terminal domain. Details of construction of the nearly full-length cDNA clones are presented.

Vesicular stomatitis virus (VSV) buds from the surface of the host cell and thereby acquires a membrane that contains spikes composed of the single viral glycoprotein (G). Our previous studies have shown that G protein has a COOH-terminal basic domain of 29 amino acids which is internal to the lipid bilayer of the virion and an adjacent hydrophobic segment of 20 amino acids which spans the bilayer (27). The NH_2-terminal 95% of G is external to the lipid bilayer and contains two asparagine-linked complex oligosaccharides (10, 21). G plays an essential role in binding of the virus to the host cell and presumably plays a role in directing budding of virus from the plasma membrane of the host cell (2, 5).

The G protein is synthesized on membrane-bound polyribosomes (18) and inserted into the rough endoplasmic reticulum as a nascent protein chain (29, 34). A short hydrophobic signal sequence of 16 amino acids is cleaved from the NH_2 terminus after insertion into the rough endoplasmic reticulum (12, 15, 22). G assumes a transmembrane configuration in the rough endoplasmic reticulum (6, 13) with only a small COOH-terminal segment exposed on the cytoplasmic face of the rough endoplasmic reticulum. Transport of G to the plasma membrane occurs via the Golgi apparatus (1a), with interorganelle transport probably occurring via coated vesicles (28). At a late stage of transport to the plasma membrane, one or two molecules of fatty acid are esterified to G (31).

The VSV matrix protein (M) is thought to be a peripheral membrane protein that lines the inner surface of the virion envelope, perhaps interacting with the lipid bilayer, the internal portion of G, and the nucleocapsid core (reviewed by Wagner [35]). In addition to a structural function, M plays a role in directing budding of virus from infected cells (14) and may be involved in regulating transcription of mRNA from the single negative strand of genomic RNA (4, 7, 9, 16).

In contrast to G, M is synthesized on free polyribosomes (18) and associates rapidly with the plasma membrane fraction after synthesis (14). The nature of the association of M with the plasma membrane fraction is not clear. It could be by association with the lipid bilayer in regions containing G protein or by association with other proteins such as those in nucleocapsids which are already associated with the plasma membrane (14).

KELONIA EXHIBIT 1028

Knowledge of the complete primary amino acid sequences of G and M proteins is clearly critical to a molecular understanding of their interactions with membranes and other cellular and viral components. Our previous studies have employed VSV cDNA clones that contain only fractions of the coding sequences from each mRNA (24, 27), and only terminal mRNA and protein sequences were reported. We report here the isolation and the complete nucleotide sequences of cDNA clones that contain the entire coding sequences of G and M protein mRNA's. From these sequences we predict the amino acid sequences of the G and M proteins and discuss features of these predicted sequences.

MATERIALS AND METHODS

Materials. Reverse transcriptase was supplied by J. Beard, St. Petersburg, Fla. The Klenow fragment of DNA polymerase I and most restriction endonucleases were purchased from New England Biolabs, Beverly, Mass. Nuclease S1, *Pst*I, and T4 polynucleotide kinase were from Boehringer Mannheim, Indianapolis, Indiana, and terminal deoxynucleotidyl transferase was from Bethesda Research Laboratories, Bethesda, Md. Oligo$(dT)_{12-18}$ and oligo(dT) cellulose (T3) were from Collaborative Research, Waltham, Mass. [α-^{32}P]dCTP and dGTP were from Amersham/Searle, Chicago, Ill., and unlabeled dNTP's were from P-L Biochemicals, Milwaukee, Wis.

RNA synthesis. VSV mRNA was synthesized in vitro in a 10-ml reaction exactly as described previously (26), with the following modification. After 4 h the reaction was stopped by addition of sodium dodecyl sulfate and sodium acetate to final concentrations of 1% and 0.5 M, respectively. The entire mixture was then passed through a column containing 0.5 g of oligo(dT)-cellulose, and the column was washed with 10 ml of 0.4 M sodium acetate. The bound mRNA was eluted with distilled water and precipitated with ethanol. All mRNA was from the San Juan strain of the Indiana serotype of VSV.

First-strand DNA synthesis and purification. First-strand DNA copies of the mRNA's encoding the N, NS, M, and G proteins were synthesized in a 4-ml reaction containing 300 μg of total VSV mRNA, 0.5 mM dATP, dGTP, and dTTP, 0.25 mM [α-^{32}P]-dCTP (10 Ci/mmol), 30 mM β-mercaptoethanol, 120 mM KCl, 10 mM $MgCl_2$, 4,000 U of reverse transcriptase, 300 μg of oligo$(dT)_{12-18}$ and a cytoplasmic extract (1 mg of total protein) from baby hamster kidney (BHK-21) cells prepared as described below. After incubation for 30 min at 42°C, the reaction mixture was extracted with phenol. Unincorporated dNTP's were separated from cDNA by chromatography of the aqueous phase on Sephadex G-50. The excluded fraction was lyophilized in a silicated glass tube. The yield of total cDNA from 300 μg of VSV mRNA was approximately 40 μg as determined from the incorporation of [α-^{32}P]dCTP. The full-length cDNA copies of each mRNA species were purified by electrophoresis on a 1.5% alkaline agarose gel (30 mM NaOH, 5 mM EDTA; 1.5 mm by 18 cm by 20 cm). The cDNA's were located by autoradiography (Fig. 1) or by staining with ethidium bromide and UV illumination. Stained bands were excised from the gel, electroeluted, and precipitated with ethanol. The yield of full-length cDNA's was approximately 0.7 μg of G, 2 μg of N, and 3 μg of both NS and M. Reaction conditions given were optimized with respect to the concentrations of reverse transcriptase, KCl, and cell extracts to give maximal sizes of cDNA's.

Cell extract preparation. BHK cells (8×10^8 cells) growing in a 1-liter suspension were pelleted by centrifugation and suspended in 1.5 times the packed cell volume of 10 mM HEPES (*N*-2-hydroxyethylpiperazine-*N*′-2-ethanesulfonic acid) (pH 7.5), 10 mM KCl, 1.5 mM magnesium acetate, and 7 mM β-mercaptoethanol. Cells were broken by 50 strokes of a Dounce homogenizer and centrifuged at $15,000 \times g$ for 20 min. The supernatant was dialyzed for 16 h against 1 liter of 10 mM HEPES (pH 7.5), 90 mM KCl, 1.5 mM magnesium acetate, and 7 mM β-mercaptoethanol. The total protein concentration in the extract was 15 to 20 mg/ml. All steps were carried out at 4°C, and samples of the extract were stored at −80°C.

Second-strand cDNA synthesis. Reactions for second-strand DNA synthesis (200 μl) included 100 mM HEPES buffer (pH 6.9), 10 mM $MgCl_2$, 60 mM KCl, 1 mM dATP, dCTP, dGTP, and dTTP, 0.5 μg of a purified cDNA, and 5 U of the Klenow fragment of DNA polymerase I. Reactions were for 15 h at 15°C and were terminated by extraction with phenol followed by precipitation with ethanol. Samples were then resuspended in 75 μl of a solution containing 3 mM $ZnCl_2$, 30 mM sodium acetate (pH. 4.5), 300 mM NaCl, and 75 U of S1 nuclease. Digestion was for 30 min at 37°C. S1-treated DNAs were extracted with phenol and precipitated with ethanol and then purified by electrophoresis on a 6% polyacrylamide gel. The nearly full-length double-stranded DNAs were detected by autoradiography or staining with ethidium bromide. The amounts of double-stranded DNAs recovered were 100 ng (G size class), 180 ng (N size class), and 450 ng (NS and M size class), as calculated from the cpm of ^{32}P in the first-strand DNA. Double-stranded DNAs were electroeluted and precipitated with ethanol.

Homopolymer addition and cloning. Homopolymer addition was by a modification of a published procedure (30). Addition of dCMP residues to double-stranded cDNA was for 10 to 30 min at 37°C in a 25-μl reaction containing 100 to 500 ng of DNA, 0.2 mM dCTP, and an amount of terminal deoxynucleotidyl transferase (Tdt) that was known to add approximately 10 to 20 dC residues to *Hae*III fragments (5 pmol of total ends) of pBR322 DNA. The number of dC residues added was calculated from the incorporation of [α-^{32}P]dCTP (100 Ci/mmol) that was included in test reactions. Units reported by the supplier were generally 10- to 50-fold greater than what we observed. Addition of 10 to 20 dG residues to *Pst*I-cleaved pBR322 DNA was carried out as above (0.2 mM dGTP) using 1 to 2 μg of DNA. The linear form of *Pst*I-cleaved pBR322 was always purified by agarose gel electrophoresis and electroelution before tailing.

Equimolar quantities of dC-tailed insert DNA and dG-tailed plasmid DNA were annealed for 15 min at 4°C in 10 μl of 0.5 M NaCl-10 mM Tris (pH 7.4) and used to transform *Escherichia coli* strain C600 to tetracycline resistance (30). From 10 to 40 colonies were obtained per ng of insert DNA. Approximately 80% of the colonies obtained were ampicillin sensitive and had inserts at the *Pst*I site. Small preparations of plasmid DNA (1 to 2 μg) were analyzed for the size of the insert DNA by *Pst*I digestion. *Hae*III and *Hin*fI digestions gave partial restriction maps for each insert. By comparing these maps with previous data (24, 25, 27) we were able to assign the clones unambiguously to the appropriate mRNA's. Although apparently full-length double-stranded DNAs were used as starting material, a large fraction (80% for G) of the inserts had deletions at one or both ends.

DNA sequence analysis. All sequence analysis was by the Maxam-Gilbert procedure as described previously (17, 25). DNA fragments (25 to 50 pmol) were prepared by restriction enzyme digestion, alkaline phosphatase treatment, polyacrylamide gel electrophoresis, and electroelution (25). End labeling was with 500 μCi of [γ-^{32}P]ATP and 2 U of polynucleotide kinase (24). Labeled DNA strands were separated by electrophoresis on thin 5% polyacrylamide gels (0.35 mm by 16 cm by 40 cm) rather than the thicker gels suggested by Maxam and Gilbert (17). Thin gels gave better resolution of closely spaced DNA strands. Gels were run at 1,200 V with an electric fan for cooling, and strand separation of fragments up to 800 base pairs long was accomplished by using electrophoresis times of less than 8 h. Occasionally fragments for sequencing were obtained by cleavage with a second restriction enzyme followed by gel electrophoresis. Sequencing gels (0.35 mm by 16 cm by 40, 85, or 152 cm) were electrophoresed at 1,500 to 6,000 V, such that the gel remained warm (>40°C) throughout the run. Reaction times were reduced substantially for samples on which the sequence was to be read for more than 300 nucleotides from the labeled end. Generally 100,000 to 500,000 dpm of ^{32}P-labeled fragment were loaded per lane on 152-cm gels so that exposure could be carried out at high resolution without fluorescent screens. Long (85 or 152 cm) gels were cut in sections, transferred to old exposed sheets of X-ray film for backing, covered with plastic wrap, and exposed to Kodak XR-5 film for 1 to 4 days.

RESULTS

The basic strategy we employed to obtain nearly full-length cDNA clones was: (i) to optimize reverse transcription of full-length first-strand DNA copies of each mRNA; (ii) to use isolated, full-length first strands as templates for second-strand DNA synthesis; and (iii) to purify the nearly full-length double-stranded DNAs (obtained after S1 nuclease treatment) by gel electrophoresis before cloning. All cloning was of dC-tailed inserts into the dG-tailed *Pst*I site of pBR322 (3).

In previous experiments we and others found that addition of a crude cell extract greatly enhanced the synthesis of full-length VSV mRNA's by the VSV virion transcriptase (1, 26). Because our initial reverse transcription experiments showed that the yields of full-length cDNA's of VSV G and N mRNA's were low relative to the yields of cDNA's of the smaller mRNA's (NS and M), we examined the effect of such an extract on reverse transcription. Figure 1 shows the products of reverse transcription of total VSV mRNA synthesized in the absence and presence of the extract. The yields of full-length G and N cDNA's were increased 5- and 2-fold, respectively, whereas the M and NS cDNA's were increased only marginally (1.2-fold). Because of the dramatic enhancement of the yield of large cDNA, the cell extract was included in all syntheses of single-stranded cDNA's for cloning. The mechanism of action of the cell extract is not known, but it presumably acts by inhibiting RNase or by promoting unfolding of mRNA secondary structure.

G and M cDNA clones contain the complete coding sequences. Two recombinant plasmids, pG1 and pM309, were obtained which had insert sizes of ca. 1,700 and 850 nucleotides, respectively, consistent with their having nearly complete copies of the G and M mRNA sequences. To determine the exact 5′ endpoints of each cloned sequence relative to the mRNA, we determined the nucleotide sequence at the end of the insert corresponding to the 5′ end of the RNA. The sequencing gel showing this region from the pG1 insert is shown in Fig. 2. The sequence preceding the homopolymer tails is

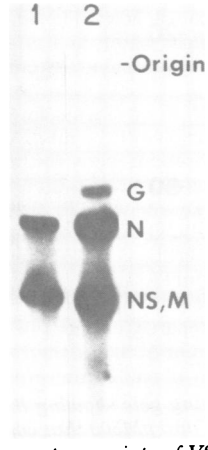

FIG. 1. *Reverse transcripts of VSV mRNA synthesized in the presence (lane 2) or absence (lane 1) of a cytoplasmic cell extract. Reverse transcripts were labeled with [α-^{32}P]dCTP and analyzed by electrophoresis on an alkaline agarose gel. Identities of the cDNA bands in the autoradiogram are indicated.*

complementary to all but nine nucleotides at the 5′ end of the G mRNA sequence (24). Similarly, the sequence of the pM309 insert (Fig. 2) showed that it contained all but 19 nucleotides from the 5′ M mRNA sequence (24). Because the translation initiation codons for the G and M mRNA's are located 30 and 42 nucleotides from the 5′ ends of the mRNA's (23, 24), these clones clearly contain the sequences encoding the NH_2 termini of G and M proteins. Subsequent sequence analysis showed that these clones contained the 3′-terminal mRNA sequences as well.

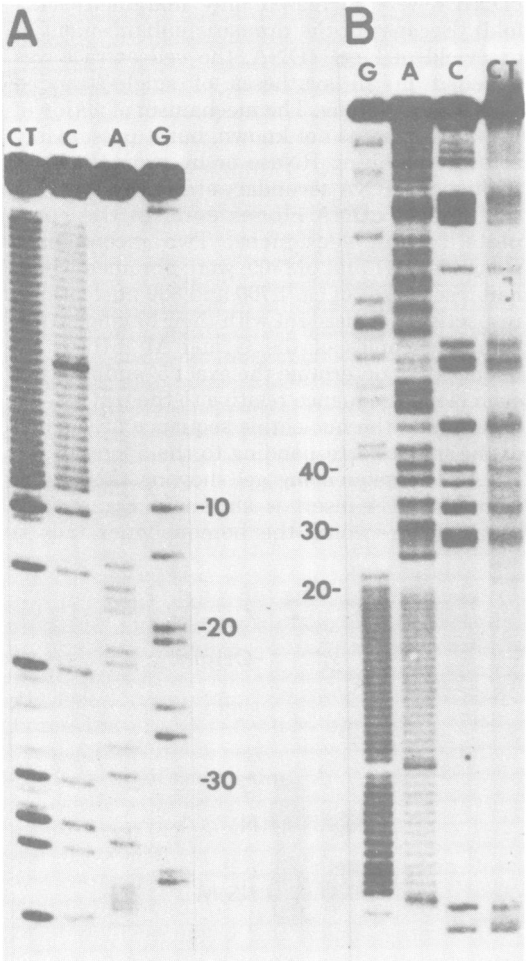

FIG. 2. *Sequencing gels showing the sequences in the regions of pG1 and pM309 that correspond to the 5′-proximal regions of G and M mRNA's. Panels: A, sequence from the 5′-proximal AluI site of the pG1 insert through the C-tails. Panel B, sequence from the pBR322 HpaII site proximal to the 5′ end of pM309 reading into the insert DNA. Numbers indicate the alignment of the sequences with the mRNA sequence (see Fig. 4 and 6).*

The mechanism of formation of pG1 and pM309 can be explained from the 5′-terminal sequences of the mRNAs (Fig. 4 and 6 of reference 24). The priming of second-strand DNA synthesis in pG1 presumably occurred from the 3′ end of the cDNA looped back to base pair perfectly with positions 13 through 18, leaving an unpaired loop of six nucleotides. After second strand synthesis, digestion of the loop with nuclease S1 must have left the unpaired nucleotides 10 through 12 intact because they appear in the clone. In the formation of pM309, the 3′ end of the cDNA presumably looped back to pair with positions 21 through 24 and began priming. Nuclease S1 digestion was complete or nearly complete, generating a clone lacking 19 nucleotides of the 5′ mRNA sequence.

Sequencing strategy for pG1. Initial restriction mapping combined with previous partial sequences (24) allowed us to locate the two *Hae*III sites within the pG1 insert (Fig. 3). Initial sequences were then established from the central *Hae*III site reading 520 nucleotides toward the 5′ end of the mRNA and 350 nucleotides toward the 3′ end. Additional sequence was determined from the 3′ *Hae*III site for 520 nucleotides toward the 5′ end of the mRNA. This approach, with sequencing gels 152 cm long, allowed us to establish a tentative sequence for almost the entire G gene using only two restriction sites. A restriction map was then established from this sequence by computer analysis (32). Subsequently we were able to identify specific restriction fragments required to complete the sequence on both DNA strands and correct a few errors that were generated by reading sequences more than 450 nucleotides from the labeled end. Figure 3 shows the locations of sites for restriction enzymes that were used in the sequence analysis and the specific regions sequenced. Several *Eco*RII sites (CC $\frac{A}{T}$ GG) were encountered in this analysis. These sequences showed gaps in the sequence ladder at the second C residue (20), and the sequences at these sites were always confirmed by sequencing the complementary strand.

The complete sequence of the G mRNA predicted from the sequence of the pG1 clone (and previous data, references 24 and 27) is shown in Fig. 4. The sequence contains a single open reading frame for translation beginning at the 5′-proximal ATG codon (positions 29 through 31) and ending 100 nucleotides from the 3′ end of the mRNA. That this is the reading frame encoding G protein has been confirmed by direct sequencing of G protein at both ends of the molecule (12, 15, 27).

Nucleotide sequencing of the M mRNA. A

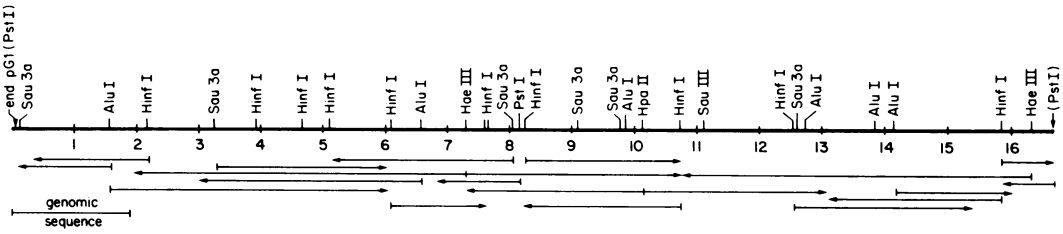

FIG. 3. *Restriction map of the pGF1 insert DNA. Sites for restriction enzymes that were used in sequencing are shown. The left end corresponds to the 5′ end of G mRNA. Arrows indicate the regions sequenced and the direction of sequencing. Numbers are in hundreds of nucleotides starting from the 5′ end of the mRNA. Terminal PstI sites are at the junctions of pBR322 sequences with the dG:dC tails. The 5′-terminal sequence of G mRNA determined previously from the genomic sequence is indicated (27). The sequence of the 3′-terminal 470 nucleotides were reported previously from a smaller clone, but also were sequenced on at least one DNA strand in pG1.*

sequence from the 5′ end of the M protein mRNA had been determined from the sequence of a DNA primer extended from the adjacent NS gene along the genome into the M mRNA-coding region (24). This sequence overlaps and agrees with the sequences determined from the pM309 insert (Fig. 5) and with the sequence of a partial M clone (pM101; L. Iverson, unpublished results; 12a). The remainder of the M mRNA sequence was determined by sequencing from the restriction sites indicated in Fig. 5. All regions were sequenced on both DNA strands or sequenced at least twice (from different sites) to ensure accuracy.

The M mRNA is 831 nucleotides long without polyadenylic acid and contains a single open reading frame for translation extending from the 5′-proximal ATG codon at positions 42 through 44 to a TAG termination codon at positions 729 through 831 (Fig. 6). Other reading frames are blocked extensively throughout the mRNA sequence, leaving little doubt that the predicted M protein sequence is correct. Although no direct protein sequence is available for M, the sequence of the ribosome binding site determined directly from the mRNA (23) is in complete agreement with the sequence predicted here from the cDNA clone. The sequence also matches the partial RNA sequence of the M mRNA determined by priming on genomic RNA from the adjacent NS gene (24).

DISCUSSION

VSV G protein structure and modification. The nucleotide sequence presented here for the VSV glycoprotein mRNA predicts a protein of 511 amino acids (57,416 daltons), including the NH_2-terminal signal sequence of 16 amino acids that is not present on the mature protein (12, 15, 27). The predicted COOH-terminal sequence of G contains a hydrophobic domain of 20 amino acids (residues 463 through 482) followed by a hydrophilic domain (residues 483 through 511). We have presented evidence previously that the hydrophobic domain spans the lipid bilayer of the viral envelope, and that the COOH-terminal charged domain residues inside the viral envelope leaving more than 95% of the protein protruding from the virus (27). The COOH-terminal basic domain probably interacts with internal virion proteins and cellular proteins during budding of the virus from the plasma membrane. This portion of G could also play a role in targeting the protein to the plasma membrane.

The VSV G protein contains two apparently identical Asn-linked complex oligosaccharides (10, 21). The precise locations of the oligosaccharide attachment sites in G are of special interest because the timing of the glycosylation events relative to the timing of G protein synthesis has been examined in detail in vitro. Addition of oligosaccharide chains by transfer from a lipid carrier occurs on the nascent polypeptide chain (29, 34). The first addition occurs when about 38% of the chain has been synthesized, and the second occurs when about 70% has been synthesized (29), assuming that the rate of protein synthesis is uniform in the in vitro system. *N*-glycosidic linkage of oligosaccharides to proteins occurs at Asn-X-Ser or Asn-X-Thr sequences (19). Inspection of the predicted G protein sequence reveals 18 Asn residues, but only two of these (amino acid residues 178 and 335, Fig. 4) occur in canonical glycosylation sequences. We therefore presume that these are the actual glycosylation sites. These sites are at fractional distances of 0.35 and 0.66 from the NH_2 terminus. The nearly exact correspondence of the positions of these sites with fraction of G synthesized when glycosylation occurs suggests that transfer of oligosaccharides to the nascent protein chain occurs when the appropriate Asn residues traverse the rough endoplasmic reticu-

MET LYS CYS LEU LEU TYR LEU ALA PHE LEU PHE
AACAGAGATCGATCTGTTTCCTTGACACTATGAAGTGCCTTTTGTACTTAGCCTTTTTAT
10 20 30 40 50 60

ILE GLY VAL ASN CYS LYS PHE THR ILE VAL PHE PRO HIS ASN GLN LYS GLY ASN TRP LYS
TCATTGGGGGTGAATTGCAAGTTCACCATAGTTTTTCCACACAACCAAAAAGGAAACTGGA
70 80 90 100 110 120

ASN VAL PRO SER ASN TYR HIS TYR CYS PRO SER SER SER ASP LEU ASN TRP HIS ASN ASP
AAAATGTTCCTTCTAATTACCATTATTGCCCGTCAAGCTCAGATTTAAATTGGCATAATG
130 140 150 160 170 180

LEU ILE GLY THR ALA ILE GLN VAL LYS MET PRO LYS SER HIS LYS ALA ILE GLN ALA ASP
ACTTAATAGGCACAGCCATACAAGTCAAAATGCCCAAGAGTCACAAGGCTATTCAAGCAG
190 200 210 220 230 240

GLY TRP MET CYS HIS ALA SER LYS TRP VAL THR THR CYS ASP PHE ARG TRP TYR GLY PRO
ACGGTTGGATGTGTCATGCTTCCAAATGGGTCACTACTTGTGATTTCCGCTGGTATGGAC
250 260 270 280 290 300

LYS TYR ILE THR [GLN] SER ILE ARG SER PHE THR PRO SER VAL GLU GLN CYS LYS GLU SER
CGAAGTATATAACACA[G]TCCATCCGATCCTTCACTCCATCTGTAGAACAATGCAAGGAAA
310 320 330 340 350 360

ILE GLU GLN THR LYS GLN GLY THR TRP LEU ASN PRO GLY PHE PRO PRO GLN SER CYS GLY
GCATTGAACAAACGAAACAAGGAACTTGGCTGAATCCAGGCTTCCCTCCTCAAAGTTGTG
370 380 390 400 410 420

TYR ALA THR VAL THR ASP ALA GLU ALA VAL ILE VAL GLN VAL THR PRO HIS HIS VAL LEU
GATATGCAACTGTGACGGATGCCGAAGCAGTGATTGTCCAGGTGACTCCTCACCATGTGC
430 440 450 460 470 480

VAL ASP GLU TYR THR GLY GLU TRP VAL ASP SER GLN PHE ILE ASN GLY LYS CYS SER ASN
TGGTTGATGAATACACAGGAGAATGGGTTGATTCACAGTTCATCAACGGAAAATGCAGCA
490 500 510 520 530 540

ASN SER THR
TYR ILE CYS PRO THR VAL HIS THR TRP HIS SER ASP TYR LYS VAL LYS GLY
ATTACATATGCCCCACTGTCCATAACTCTACAACCTGGCATTCTGACTATAAGGTCAAAG
550 560 570 580 590 600

LEU CYS ASP SER ASN LEU ILE SER MET ASP ILE THR PHE PHE SER GLU ASP GLY GLU LEU
GGCTATGTGATTCTAACCTCATTTCCATGGACATCACCTTTCTTTCTCAGAGGACGGAGAGC
610 620 630 640 650 660

SER SER LEU GLY LYS GLU GLY THR GLY PHE ARG SER ASN TYR PHE ALA TYR GLU THR GLY
TATCATCCCTGGGAAAGGAGGGCACAGGGTTCAGAAGTAACTACTTTGCTTATGAAACTG
670 680 690 700 710 720

GLY LYS ALA CYS LYS MET GLN TYR CYS LYS HIS TRP GLY VAL ARG LEU PRO SER GLY VAL
GAGGCAAGGCCTGCAAAATGCAATACTGCAAGCATTGGGGAGTCAGACTCCCATCAGGTG
730 740 750 760 770 780

TRP PHE GLU MET ALA ASP LYS ASP LEU PHE ALA ALA ALA ARG PHE PRO GLU CYS PRO GLU
TCTGGTTCGAGATGGCTGATAAGGATCTCTTTGCTGCAGCCAGATTCCCTGAATGCCCAG
790 800 810 820 830 840

GLY SER SER ILE SER ALA PRO SER GLN THR SER VAL ASP VAL SER LEU ILE GLN ASP VAL
AAGGGTCAAGTATCTCTGCTCCATCTCAGACCTCAGTGGATGTAAGTCTAATTCAGGACG
850 860 870 880 890 900

GLU ARG ILE LEU ASP TYR SER LEU CYS GLN GLU THR TRP SER LYS ILE ARG ALA GLY LEU
TTGAGAGGATCTTGGATTATTCCCTCTGCCAAGAAACCTGGAGCAAAATCAGAGCGGGTC
910 920 930 940 950 960

PRO ILE SER PRO VAL ASP LEU SER TYR LEU ALA PRO LYS ASN PRO GLY THR GLY PRO ALA
TTCCAATCTCTCCAGTGGATCTCAGCTATCTTGCTCCTAAAAACCCAGGAACCGGTCCTG
970 980 990 1000 1010 1020

ASN GLY THR
PHE THR ILE ILE ASN GLY THR LEU LYS TYR PHE GLU THR ARG TYR ILE ARG VAL ASP ILE
CTTTCACCATAATCAATGGTACCCTAAAATACTTTGAGACCAGATACATCAGAGTCGATA
1030 1040 1050 1060 1070 1080

ALA ALA PRO ILE LEU SER ARG MET VAL GLY MET ILE SER GLY THR THR THR GLU ARG GLU
TTGCTGCTCCAATCCTCTCAGAATGGTCGGAATGATCAGTGGAACTACCACAGAAAGGG
1090 1100 1110 1120 1130 1140

LEU TRP ASP ASP TRP ALA PRO TYR GLU ASP VAL GLU ILE GLY PRO ASN GLY VAL LEU ARG
AACTGTGGGATGACTGGGCACCATATGAAGACGTGGAAATTGGACCCAATGGAGTTCTGA
1150 1160 1170 1180 1190 1200

THR SER SER GLY TYR LYS PHE PRO LEU TYR MET ILE GLY HIS GLY MET LEU ASP SER ASP
GGACCAGTTCAGGATATAAGTTTCCTTTATACATGATTGGACATGGTATGTTGGACTCCG
1210 1220 1230 1240 1250 1260

LEU HIS LEU SER SER LYS ALA GLN VAL PHE GLU HIS PRO HIS ILE GLN ASP ALA ALA SER
ATCTTCATCTTAGCTCAAAGGCTCAGGTGTTCGAACATCCTCACATTCAAGACGCTGCTT
1270 1280 1290 1300 1310 1320

GLN LEU PRO ASP ASP GLU SER LEU PHE PHE GLY ASP THR GLY LEU SER LYS ASN PRO ILE
CGCAACTTCCTGATGATGAGAGTTTATTTTTTGGTGATACTGGGCTATCCAAAAATCCAA
1330 1340 1350 1360 1370 1380

SER SER SER ILE ALA SER PHE PHE PHE ILE
GLU LEU VAL GLU GLY TRP PHE SER SER TRP LYS SER
TCGAGCTTGTAGAAGGTTGGTTCAGTAGTTGGAAAAGCTCTATTGCCTCTTTTTTCTTTA
1390 1400 1410 1420 1430 1440

ILE GLY LEU ILE ILE GLY LEU PHE LEU VAL LEU ARG VAL GLY ILE HIS LEU CYS ILE LYS
TCATAGGGTTAATCATTGGACTATTCTTGGTTCTCCGAGTTGGTATCCATCTTTGCATTA
1450 1460 1470 1480 1490 1500

LEU LYS HIS THR LYS LYS ARG GLN ILE TYR THR ASP ILE GLU MET ASN ARG LEU GLY LYS
AATTAAAGCACACCAAGAAAAGACAGATTTATACAGACATAGAGATGAACCGACTTGGAA
1510 1520 1530 1540 1550 1560

***
AGTAACTCAAATCCTGCACAACAGATTCTTCATGTTTGGACCAAATCAACTTGTGATACC
1570 1580 1590 1600 1610 1620

ATGCTCAAAGAGGCCTCAATTATATTTGAGTTTTTAATTTTTATG
1630 1640 1650 1660

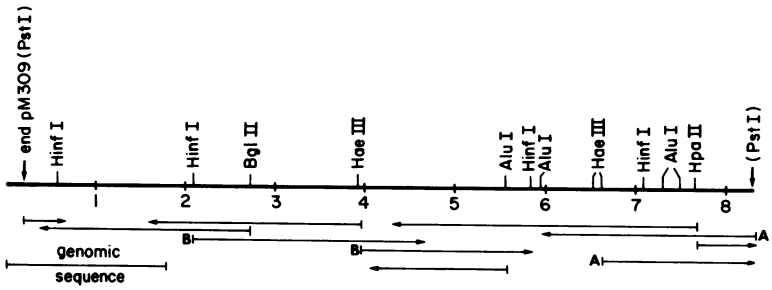

FIG. 5. *Restriction map of the pM309 insert DNA. Sites for restriction enzymes used in sequencing are shown. The left end corresponds to the 5' end of M mRNA. Arrows indicate the regions and directions sequenced; numbers represent hundreds of nucleotides. Arrows labeled A and B represent sequences determined from cDNA clones pM32 and pM37 which contain smaller portions of the M sequence than pM309. Terminal PstI sites are at the junctions of pBR322 sequences with the dG:dC tails. The 5'-terminal mRNA sequence (genomic sequence) reported previously is indicated (24).*

lum. Extensive folding of the polypeptide chain on the COOH-terminal site of the glycosylation sites is apparently not required to specify glycosylation.

One to two molecules of fatty acid are esterified to G protein at a late stage in passage of G from the rough endoplasmic reticulum to the plasma membrane (31). The evidence suggests that the fatty acid is esterified to serine residues (31). Our preliminary evidence (W. J. Welch, B. M. Sefton, and J. K. Rose, unpublished data) indicates that [^{3}H]palmitate label can be found associated specifically with the 64-amino acid, membrane-protected tailpiece of G isolated after proteolysis of intact virions (27). There are only five serine residues in this portion of G, and they are clustered around the NH_2-terminal side of the hydrophobic domain (Fig. 4). Presumably, one or more of these residues are esterified to fatty acid. Fatty acid esterified in this region could serve the obvious function of promoting stable association of the hydrophobic protein domain with the membrane.

Structure and function of VSV M. The nucleotide sequence of the VSV M mRNA presented here predicts a protein sequence of 229 amino acids (26,064 daltons). The predicted amino-terminal sequence of M is highly basic. Eight lysine residues occur in the first 19 positions, and these are the only charged residues in this region (Fig. 6). A triple proline sequence separates this domain from the remainder of the molecule. There is evidence from analysis of M protein mutants and from in vitro studies that M may regulate VSV transcription (4, 7, 9, 16). This basic domain might play a role in such regulation, perhaps by interacting with the genomic RNA. The predicted sequence indicates that M is the most basic of the VSV N, NS, M, and G proteins, having 21 Lys, 10 Arg, and 8 His residues and only 13 Asp and 13 Glu residues. Although there is no direct amino acid sequence analysis available for M protein, this basic character is consistent with its isoelectric point determined by gel electrophoresis (4). Our attempts to obtain an NH_2-terminal sequence from M have been unsuccessful, suggesting that the NH_2 terminus is blocked.

Evidence indicates that M protein is released into the soluble fraction of infected cells after synthesis. It then associates rapidly with the plasma membrane fraction (14). It is not clear, however, whether there is any direct association of M with the lipid bilayer of the plasma membrane. There is evidence that M will associate with membrane fractions from uninfected cells, although it is not clear that this interaction is biologically significant (8). Inspection of the predicted M protein sequence (Fig. 6) does not reveal any long hydrophobic or nonpolar domains that might be inserted into the membrane. This situation contrasts with that of the influenza virus matrix protein which has a central hydrophobic domain that may be membrane

FIG. 4. *Nucleotide sequence of the VSV G mRNA and the predicted protein sequence. The shaded nucleotide sequence is the ribosome binding site. NH_2-terminal and COOH-terminal hydrophobic domains are shaded, as are the two potential glycosylation sites and basic residues in the COOH-terminal hydrophilic domain. The bracketed G residue (317) appeared clearly as a G on two sequencing gels of the DNA strand shown. The sequence of the complementary DNA strand using the same DNA preparation showed a clear A residue instead of the expected C in this position in two separate experiments. This residue and the amino acid encoded should therefore be considered tentative. Nucleotide number one is linked to the 5' cap in the mRNA sequence.*

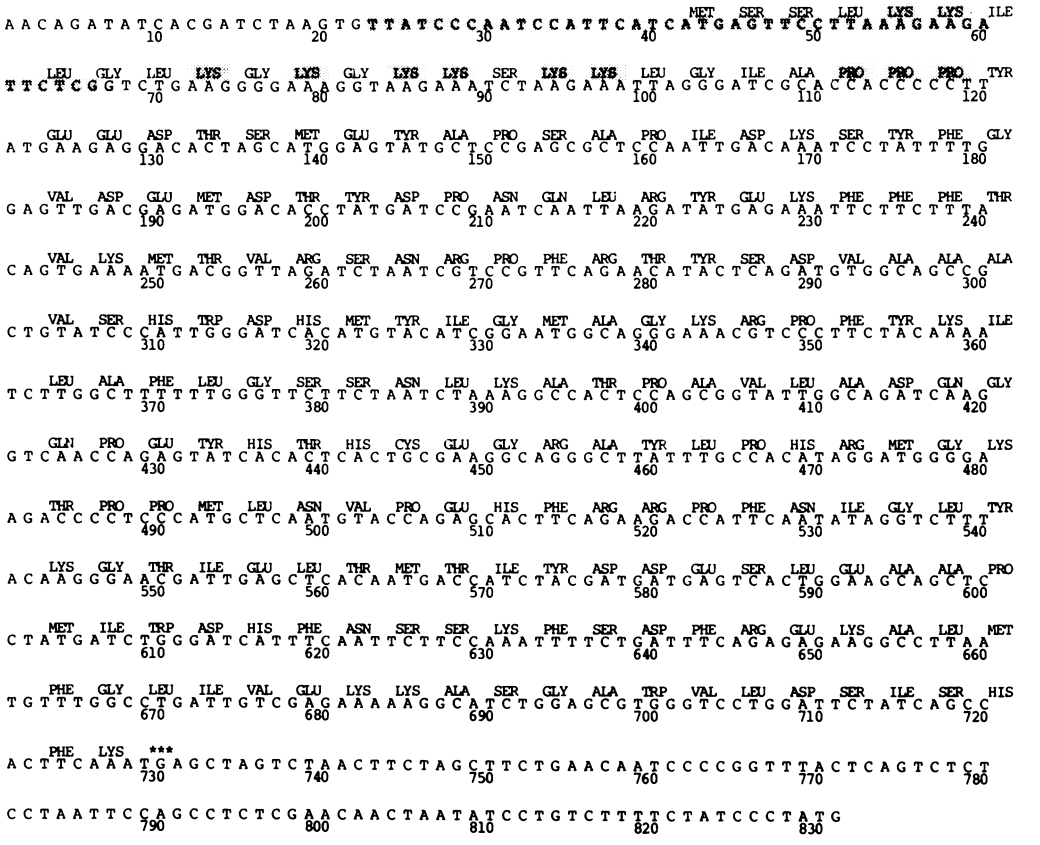

FIG. 6. *Nucleotide sequence of the VSV M mRNA and the predicted M protein sequence. The shaded nucleotide sequence is the ribosome binding site (23). Basic residues near the NH_2 terminus and the triple Pro sequence are shaded.*

associated (37). The nature of VSV M protein association with the viral envelope clearly requires further analysis.

CG dinucleotide deficiency and codon usage. There is an overall deficiency of the dinucleotide CG in the G and M mRNA's. This deficiency is also seen in the N and NS mRNA's (11). On a random basis considering base composition, one would predict 74 CGs in the G mRNA-coding region and 34 CGs in the M mRNA-coding region, whereas only 24 and 18, respectively, are observed. This deficiency is clear from inspection of the codon usage in both mRNA's (Fig. 7). Arginine codons AGA or AGG are preferred over CGN arginine codons, and there is infrequent use of codons having CG in the second and third positions. The CG deficiency is also observed between adjacent codons. That is, if one considers the frequency of codons having C in the third position and G in the first position, random assortment would generate 37 CGs in G mRNA and 20 CGs in M mRNA between adjacent codons, whereas only 14 and 9, respectively, occur. Because the CG deficiency is observed between adjacent codons as well as within codons, the deficiency is probably not an adaptation of the virus to a relative shortage of cellular tRNA's recognizing codons containing CG. Consistent with this interpretation, we find that CG is also deficient in the 479 noncoding nucleotide of the G, M, N, and NS mRNA's where 18 CGs are expected and only 9 are found.

Vertebrate DNA is known to be deficient in the CG dinucleotide (33), although the reason for this is not known. If the mechanism of selection against CG is the same for cellular DNA and VSV RNA the selection must be able to operate at the RNA level because VSV has only RNA intermediates in transcription and replication.

Both the G and M mRNA's are moderately A rich (32 and 30% A, respectively). The preference for codons having A in the third position presumably reflects this base composition (Fig. 7).

G protein

	U		C		A		G	
U	Phe	10	Ser	10	Tyr	11	Cys	4	U
		16		8		9		11	C
	Leu	8		11	Term	1	Term	0	A
		4		1		0	Trp	15	G
C	Leu	9	Pro	10	His	11	Arg	0	U
		7		3		5		1	C
		6		12	Gln	11		3	A
		5		2		7		0	G
A	Ile	14	Thr	10	Asn	10	Ser	10	U
		15		12		8		7	A
		8		7	Lys	15	Arg	8	C
	Met	11		2		16		3	G
G	Val	8	Ala	13	Asp	17	Gly	9	U
		9		6		11		4	C
		3		5	Glu	16		19	A
		9		1		9		6	G

M protein

	U		C		A		G	
U	Phe	5	Ser	8	Tyr	7	Cys	0	U
		9		4		5		1	C
	Leu	4		2	Term	0	Term	1	A
		4		0		0	Trp	3	G
C	Leu	1	Pro	3	His	3	Arg	2	U
		3		3		5		0	C
		1		7	Gln	3		0	A
		4		3		0		0	G
A	Ile	4	Thr	3	Asn	6	Ser	1	U
		6		3		0		3	C
		1		3	Lys	11	Arg	6	A
	Met	11		2		10		2	G
G	Val	2	Ala	6	Asp	9	Gly	5	U
		2		3		4		2	C
		3		6	Glu	3		4	A
		2		2		10		4	G

FIG. 7. *Codon usage in the G and M mRNA's. First positions of codons are indicated on the left, second positions are indicated across the top, and third positions are indicated on the right.*

ACKNOWLEDGMENTS

We thank Bart Sefton and Linda Iverson for critical comments on the manuscript and other members of the Tumor Virology Laboratory for their helpful suggestions. We are grateful to Linda Iverson for providing sequence data from pM101 (Iverson and Rose, in press) which overlaps and confirms our 5'-terminal sequence from pM309.

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. **Ball, L. A., and C. N. White.** 1976. Order of transcription of genes of vesicular stomatitis virus. Proc. Natl. Acad. Sci. U.S.A. **73**:442–446.

1a. **Bergman, J. E., K. T. Tokuyasu, and S. J. Singer.** 1981. Passage of an integral membrane protein, the vesicular stomatitis virus glycoprotein, through the Golgi apparatus en route to the plasma membrane. Proc. Natl. Acad. Sci. U.S.A. **78**:1746–1750.

2. **Bishop, D. H. L., P. Repik, J. F. Obijeski, N. F. Moore, and R. R. Wagner.** 1975. Restoration of infectivity to spikeless vesicular stomatitis virus by solubilized viral components. J. Virol. **16**:75–84.

3. **Bolivar, F., R. L. Rodriguez, M. C. Betlach, and H. W. Boyer.** 1977. Construction and characterization of new cloning vehicles. Gene **2**:95–112.

4. **Carroll, A. R., and R. R. Wagner.** 1979. Role of the membrane (M) protein in endogenous inhibition of in vitro transcription by vesicular stomatitis virus. J. Virol. **29**:134–142.

5. **Cartwright, B., C. J. Smale, and F. Brown.** 1969. Surface structure of vesicular stomatitis virus. J. Gen. Virol. **5**:1–10.

6. **Chatis, P. A., and T. G. Morrison.** 1979. Vesicular stomatitis virus glycoprotein is anchored to intracellular membranes near its carboxyl end and is proteolytically cleaved at its amino terminus. J. Virol. **29**:957–963.

7. **Clinton, G. M., S. P. Little, F. S. Hagen, and A. S. Huang.** 1978. The matrix (M) protein of vesicular stomatitis virus regulates transcription. Cell **15**:1455–1462.

8. **Cohen, G., P. Atkinson, and D. Summers.** 1971. Interaction of vesicular stomatitis virus structural proteins with HeLa plasma membranes. Nature (London) **231**: 121–123.

9. **Combard, A., and C. Printz-Ane.** 1979. Inhibition of vesicular stomatitis virus transcriptase complex by the virion envelope M protein. Biochem. Biophys. Res. Commun. **88**:117–123.

10. **Etchison, J. R., J. S. Robertson and D. F. Summers.** 1977. Partial structural analysis of the oligosaccharide moieties of the vesicular stomatitis virus glycoprotein by sequential chemical and enzymatic degradation. Virology **78**:375–392.

11. **Gallione, C. J., J. R. Greene, L. E. Iverson, and J. K. Rose.** 1981. Nucleotide sequences of the mRNA's encoding the vesicular stomatitis virus N and NS proteins. J. Virol. **39**:529–535.

12. **Irving, R. A., F. Toneguzzo, S. H. Rhee, T. Hofmann, and H. P. Ghosh.** 1979. Synthesis and assembly of membrane glycoproteins: presence of leader peptide in nonglycosylated precursor of membrane glycoprotein of vesicular stomatitis virus. Proc. Natl. Acad. Sci. U.S.A. **76**:570–574.

12a. **Iverson, L., and J. K. Rose.** 1981. Localized attenuation and discontinuous synthesis during VSV transcription. Cell **23**:477–484.

13. **Katz, F. N., and H. F. Lodish.** 1979. Transmembrane biogenesis of the vesicular stomatitis virus glycoprotein. J. Cell. Biol. **80**:416–426.

14. **Knipe, D., D. Baltimore, and H. Lodish.** 1977. Separate pathways of maturation of the major structural proteins of vesicular stomatitis virus. J. Virol. **21**:1128–1139.

15. **Lingappa, V. R., F. N. Katz, H. F. Lodish, and G. Blobel.** 1978. A signal sequence for the insertion of a transmembrane glycoprotein. J. Biol. Chem. **253**:8667–8670.

16. **Martinet, C., A. Combard, C. Printz-Ane, and P. Printz.** 1979. Envelope proteins and replication of vesicular stomatitis virus: in vivo effects of RNA$^+$ temperature-sensitive mutations on viral RNA synthesis. J. Virol. **29**:123–133.

17. **Maxam, A. M., and W. Gilbert.** 1980. Chemical sequencing of DNA. Methods Enzymol. **65**:501–561.

18. **Morrison, T. G., and H. F. Lodish.** 1975. Site of synthesis of membrane and nonmembrane proteins of vesicular stomatitis virus. J. Biol. Chem. **250**:6955–6962.

19. **Neuberger, A., A. Gottschalk, R. D. Marshall, and R. G. Spiro.** 1972. Carbohydrate-peptide linkages in glycoproteins and methods for their elucidation, p. 450-490. *In* A. Gottschalk (ed.) The glycoproteins: their composition, structure and function. Elsevier, Amsterdam.
20. **Ohmori, H., J. I. Tomizawa, and A. Maxam.** 1978. Detection of 5-methylcytosine in DNA sequences. Nucleic Acids Res. **5**:1479-1485.
21. **Reading, C. L., E. E. Penhoet, and C. E. Ballou.** 1978. Carbohydrate structure of vesicular stomatitis virus glycoprotein. J. Biol. Chem. **253**:5600-5612.
22. **Rose, J. K.** 1977. Nucleotide sequences of ribosome recognition sites on messenger RNAs of vesicular stomatitis virus. Proc. Natl. Acad. Sci. U.S.A. **74**:3672-3676.
23. **Rose, J. K.** 1978. Complete sequence of ribosome recognition sites in VSV mRNAs. Cell **14**:345-353.
24. **Rose, J. K.** 1980. Complete intergenic and flanking gene sequences from the genome of vesicular stomatitis virus. Cell **19**:415-421.
25. **Rose, J. K., and L. E. Iverson.** 1979. Sequences from the 3′ ends of VSV mRNA's as determined from cloned DNA. J. Virol. **32**:404-411.
26. **Rose, J. K., H. F. Lodish, and M. L. Brock.** 1977. Giant heterogeneous polyadenylic acid on vesicular stomatitis virus mRNA synthesized in vitro in the presence of S-adenosyl hemocysteine. J. Virol. **21**:683-694.
27. **Rose, J. K., W. J. Welch, B. M. Sefton, F. S. Esch, and N. C. Ling.** 1980. Vesicular stomatitis glycoprotein is anchored in the viral membrane by a hydrophobic domain near the COOH-terminus. Proc. Natl. Acad. Sci. U.S.A. **77**:3884-3888.
28. **Rothman, J. E., and R. E. Fine.** 1980. Coated vesicles transport newly synthesized membrane glycoproteins from endoplasmic reticulum to plasma membrane in two successive stages. Proc. Natl. Acad. Sci. U.S.A. **77**: 780-784.
29. **Rothman, J. E., and H. F. Lodish.** 1977. Synchronized transmembrane insertion and glycosylation of a mascent membrane protein. Nature (London) **269**:775-780.
30. **Roychoudhury, R., E. Jay, and R. Wu.** 1976. Terminal labeling and addition of homopolymer tracts to duplex DNA fragments by terminal deoxynucleotidyl transferase. Nucleic Acids Res. **3**:101-116.
31. **Schmidt, M. F., and M. J. Schlesinger.** 1979. Fatty acid binding to vesicular stomatitis virus glycoprotein: a new type of post-translational modification of the viral glycoprotein. Cell **17**:813-819.
32. **Staden, R.** 1977. Sequence data handling by computer. Nucleic Acids Res. **4**:4037-4051.
33. **Swartz, M. N., T. A. Trautner, and A. Kornberg.** 1962. Enzymatic synthesis of DNA: further studies on nearest neighbor base sequences in deoxyribonucleic acids. J. Biol. Chem. **237**:1961-1967.
34. **Toneguzzo, F., and H. P. Ghosh.** 1977. Synthesis and glycosylation in vitro of glycoprotein of vesicular stomatitis virus. Proc. Natl. Acad. Sci. U.S.A. **74**:1516-1520.
35. **Wagner, R. R.** 1975. Reproduction of rhabdovirus, p. 1-93. *In* H. Fraenkel-Conrat and R. R. Wagner (ed.), Comprehensive virology. Plenum Publishing Corp., New York.
36. **Wensink, P. C., D. J. Finnegan, J. E. Donelson, and D. S. Hogness.** 1974. A system for mapping DNA sequences in the chromosomes of Drosophila melanogaster. Cell **3**:315-325.
37. **Winter, G., and S. Fields.** 1980. Cloning of influenza cDNA into M13: the sequence of the RNA segment encoding the A/PR/8/34 matrix protein. Nucleic Acids Res. **8**:1965-1974.