Proc. Natl. Acad. Sci. USA
Vol. 77, No. 7, pp. 3884–3888, July 1980
Biochemistry

Vesicular stomatitis virus glycoprotein is anchored in the viral membrane by a hydrophobic domain near the COOH terminus

(transmembrane protein sequence/signal sequence/cDNA cloning/DNA sequence determination)

J. K. ROSE*, W. J. WELCH*, B. M. SEFTON*, F. S. ESCH†, AND N. C. LING†

*Tumor Virology Laboratory and †Laboratories for Endocrinology, The Salk Institute, Post Office Box 85800, San Diego, California 92138

Communicated by Renato Dulbecco, April 7, 1980

ABSTRACT We have determined the COOH-terminal and NH_2-terminal amino acid sequences of the vesicular stomatitis virus (VSV) glycoprotein (G). A sequence of 122 COOH-terminal amino acids was deduced from the complete sequence of a cloned DNA insert carrying 470 nucleotides derived from the 3′ end of the G mRNA. Evidence presented indicates that this portion of the polypeptide includes the domains of G that reside inside the virion and span the lipid bilayer of the virion. This seems clear because a partial amino acid sequence of a fragment of G that remains associated with the membrane of the virion after exhaustive proteolytic digestions can be located unambiguously in the predicted sequence. This predicted sequence contains an uninterrupted hydrophobic domain beginning 49 amino acids and ending 30 amino acids from the COOH terminus. This region presumably spans the lipid bilayer. The COOH-terminal portion of 29 amino acids contains a high proportion of basic residues and resides inside the virion. The COOH-terminal portion of the VSV G protein therefore resembles in structure that of glycophorin, an erythrocyte membrane protein well characterized previously. The configuration of G in the viral membrane demonstrated here is probably similar for other viral glycoproteins, although this has not been tested as directly in any other case. From the sequence of a DNA primer extended on the RNA genome from the adjacent M protein gene into the G protein gene, we have deduced an NH_2-terminal G protein sequence of 53 amino acids, including the leader sequence of 16 amino acids. Our sequence confirms, extends, and corrects two partial amino acid sequences reported for this region previously.

Enveloped viruses, such as vesicular stomatitis virus (VSV) serve as simple and useful model systems for the study of the structure, biosynthesis, and function of membrane proteins. VSV virions contain a single glycoprotein, G, which forms spikes on the surface of the virion. This 70,000-dalton glycoprotein contains two asparagine-linked complex oligosaccharides (1, 2) and is positioned in the virion such that almost 90% of the polypeptide chain is external to the lipid bilayer (3, 4).

G plays two roles in the life cycle of the virus. First, it is responsible both for the binding of the virus to susceptible host cells and for inducing the uptake of the virus by the cell (5, 6). Second, during virus maturation the interaction between the internal components of the virion and the portion of G exposed on the cytoplasmic face of the plasma membrane probably directs envelopment and virus budding.

The G protein is inserted into the rough endoplasmic reticulum (RER) as a nascent polypeptide chain (7, 8). Both ends of the molecule appear to have critical roles. Like the majority of the secreted and membrane proteins, the nascent G polypeptide has a short hydrophobic NH_2-terminal signal or leader peptide (9–12), which appears to initiate association of the nascent polypeptide–mRNA–ribosome complex with the membrane of the endoplasmic reticulum and which is removed prior to chain termination (13). Unlike secreted proteins, however, membrane proteins such as G are not discharged completely across the microsomal membrane and instead become anchored stably in the membrane. Preliminary studies have suggested that G is bound to microsomal membranes at a site very near to its COOH terminus (12, 14). Hence, this region of G must differ from the COOH termini of secreted proteins in some fundamental but as yet not understood way so as to halt extrusion into the lumen of the RER.

A knowledge of the complete primary amino acid sequence of both ends of this polypeptide is clearly critical to a molecular understanding of how G interacts with cellular membranes. Because only partial amino acid sequences had been determined for the NH_2 terminus of both the leader peptide and of the mature protein (10, 11) and no sequences had been determined for the COOH terminus of the protein, we used two different strategies to obtain the G mRNA sequences encoding these domains. Furthermore, we have been able to demonstrate, by determining the partial amino acid sequence of a fragment of G that is resistant to digestion by protease by virtue of protection by the viral membrane, that G interacts with the membrane of the mature virion near its COOH terminus. We discuss features of this COOH-terminal sequence that anchor G in the membrane and that may be critical in formation of functional virions.

MATERIALS AND METHODS

The Indiana serotype of VSV (San Juan strain) was used in all studies reported here. Isolation of cDNA clones of VSV mRNAs, DNA sequence determination, and DNA primer extension on VSV RNA were as described (15–17).

Labeling and Purification of Virus. BHK (baby hamster kidney) cell monolayers were infected with VSV (multiplicity of infection 50–100) and labeled with amino acids between 3 and 11 hr after infection. The labeled amino acids were L-[4,5-^{3}H(N)]leucine (New England Nuclear, 58 Ci/mmol), L-[2-^{3}H]glycine (ICN, 5–15 Ci/mmol), L-[^{35}S]methionine (Amersham/Searle, 500 Ci/mmol), and L-[4,5-^{3}H(N)]isoleucine (New England Nuclear, 100 Ci/mol) in medium lacking the appropriate amino acid (1 Ci = 3.7×10^{10} becquerels). Virus was purified by sedimentation through 20% sucrose onto a cushion of 60% sucrose, followed by centrifugation to equilibrium on a 20–50% sucrose gradient.

Protease Treatment of VSV Virus. Labeled virus (final protein concentration of 10–20 mg/ml) was digested with trypsin (Worthington, final concentration 2 mg/ml), or thermolysin (Calbiochem, final concentration 1 mg/ml) in 50 mM Tris·HCl, pH 8.0/100 mM NaCl/5 mM $CaCl_2$. After digestion

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; RER, rough endoplasmic reticulum.

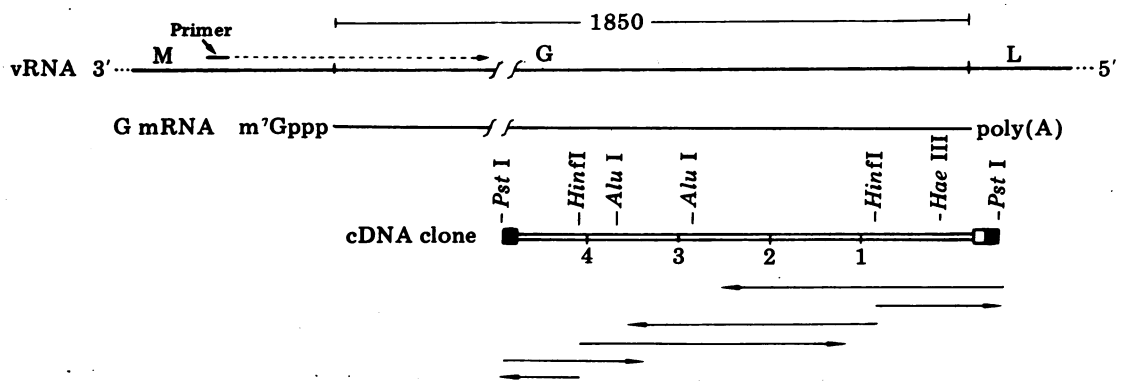


FIG. 1. Diagram showing the methods used to obtain the 5'-terminal and 3'-terminal G mRNA sequences. The region of the VSV genome containing the G gene and the flanking M and L genes is shown. A single-stranded DNA primer derived from a cDNA clone of the VSV M mRNA (pM 32) was extended from the indicated position in the M gene into the G gene and then analyzed to obtain the 5'-terminal G mRNA and protein sequences. Isolation and extension of this primer was as described (16). The G mRNA and the region of the mRNA sequence containing the pG65 cDNA clone (double lines) are indicated. Arrows represent the number of nucleotides sequenced from each ^{32}P-labeled 5' end. Restriction endonuclease sites used in sequence determination are indicated. Lengths are in hundreds of nucleotides.

for 1 hr at 37°C, 2 mM phenylmethylsulfonyl fluoride and tosyllysylchloromethyl ketone at 5 μg/ml were added, followed by incubation for 10 min on ice. Virus was pelleted at 200,000 × *g* for 3 hr through 6 ml of a solution containing 20% sucrose, 400 mM Tris·HCl, pH 7.4/100 mM NaCl/10 mM dithiothreitol/10 mM ethylenediaminetetraacetic acid.

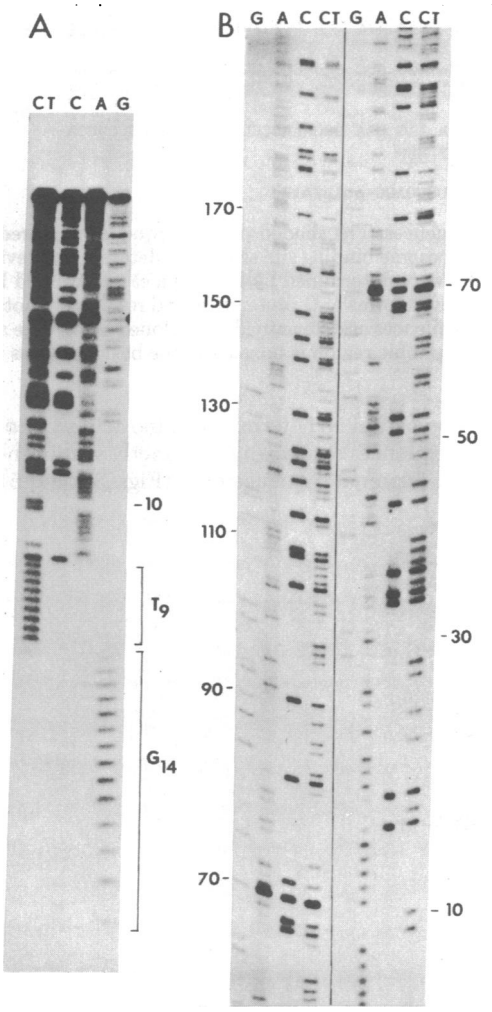


FIG. 2. Gel autoradiogram showing sequences derived from pG65 insert DNA. (*A*) Sequence of G_{14} and T_9 leading into the complement of the 3' G mRNA sequence. (*B*) Sequence of 170 nucleotides complementary to the 3' end of the G mRNA. Thin gels [40 cm by 16 cm by 0.35 mm (20)] of either 20% acrylamide (*A*) or 6% acrylamide (*B*) (85 cm by 16 cm by 0.35 mm) were used. The lower half of the 6% gel was subjected to autoradiography. Numbering is from the poly(A)-proximal nucleotide.

Purification and Sequence Determination of the Protease-Resistant Fragments. Purification of fragments was by polyacrylamide gel electrophoresis as described (18). The fragments were excised from the gel, electroeluted, filtered through cotton to remove particulate matter, dialyzed extensively against three changes of 0.03% sodium dodecyl sulfate, and lyophilized. These fragments were subjected to automatic sequential degradation on a Beckman 890C sequencer. The nonprotein carrier Polybrene and a 0.33 M Quadrol program similar to that described by Hunkapiller and Hood (19) were employed. The residues from the sequential analyses were transferred into scintillation vials with two 150-μl washes of methanol and their radioactivities were measured in 10 ml of Budget Solve (Research Products International, Elk Grove, IL).

RESULTS

Determination of the COOH-Terminal G Protein Sequence. Katz *et al.* (14) and Chatis and Morrison (12) observed that newly synthesized G protein was oriented in vesicles derived from the RER such that proteolysis degraded only about 3000 daltons of the protein. The tryptic peptides of G that were degraded appeared to map to the COOH terminus of the protein. Assuming that the orientation of G in the mature virion was similar to that in vesicles derived from the RER, we expected the COOH-terminal portion of G to contain both the part of the polypeptide that spans the lipid bilayer and the part that is inside of the lipid bilayer, presumably interacting with the internal components.

To obtain the 3'-terminal G mRNA sequence and the corresponding COOH-terminal protein sequence, we have determined the sequence an insert from a cDNA clone that contains 470 nucleotides derived from the 3' end of the G mRNA (16). Fig. 1 illustrates the region of G mRNA sequence included in this cDNA clone and the restriction endonuclease sites used to generate fragments for complete sequence determination of both DNA strands by the Maxam–Gilbert procedure (17). Fig. 2 contains examples of sequencing gels that show the sequence extending from one end of the insert through the "G-tails" and "T-primer" used in cDNA cloning (15) into the complement of the mRNA sequence. The mRNA sequence derived from the complete sequence of the cDNA clone is shown in Fig. 3*B*.

The reading frame for the G protein was deduced as follows.

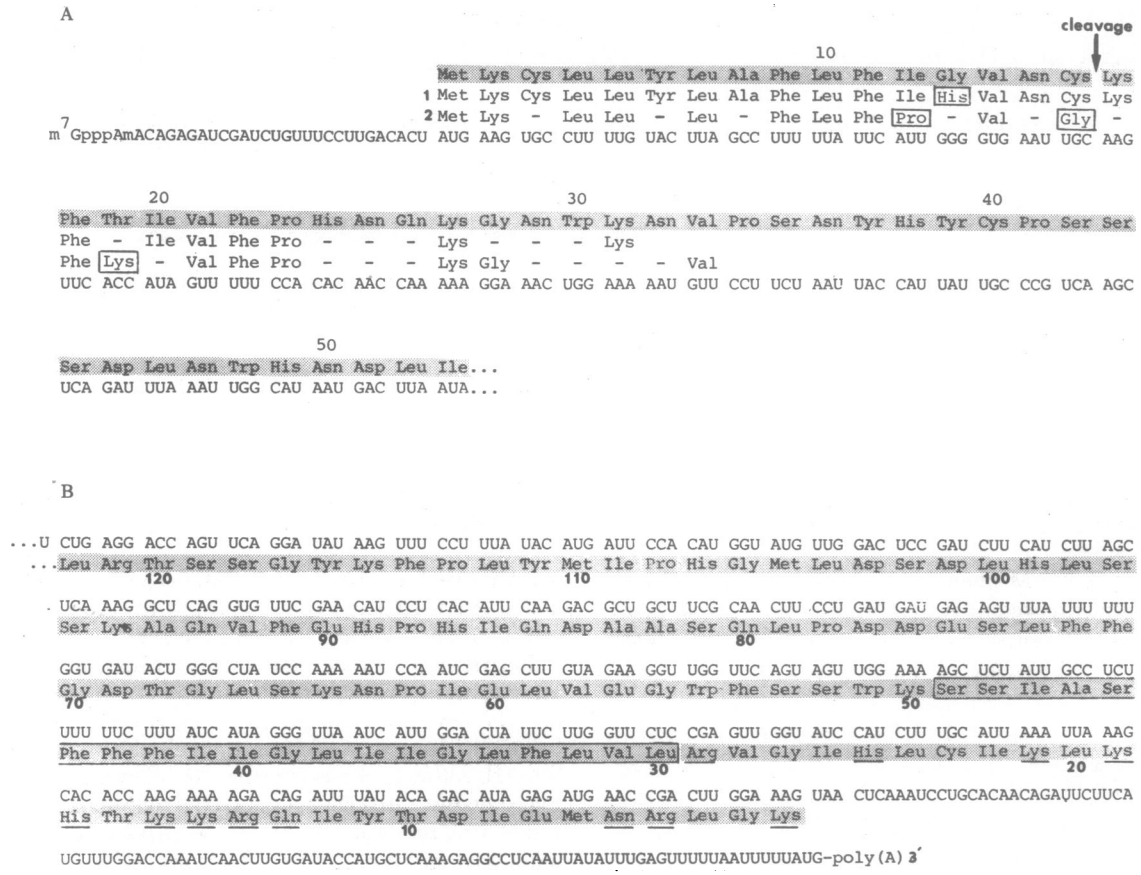


FIG. 3. G mRNA and protein sequences. (*A*) 5′-Terminal G mRNA and protein sequences. The shaded protein sequence was predicted from the cDNA (mRNA) sequence beginning from the single initiator AUG codon in the ribosome binding site sequences determined previously from ribosome-protected G mRNA (9). Partial amino acid sequences numbered 1 and 2 were determined by Lingappa *et al.* (10) and Irving *et al.* (11), respectively. These sequences were determined directly from cleaved and uncleaved forms of G proteins. Boxed residues do not agree with those predicted from the cDNA sequences. (*B*) 3′-Terminal G mRNA and protein sequences as determined from cloned DNA. The amino acid sequence shown was predicted from the mRNA sequence. The uninterrupted hydrophobic region is boxed and the basic residues in the COOH-terminal 29 amino acids are underlined.

The length of G mRNA is about 1850 nucleotides without poly(A), and it encodes a polypeptide of 62,500 daltons (21) or about 570 amino acids. In the mRNA there are 30 noncoding nucleotides at the 5′ end (16), leaving the capacity to encode 606 amino acids. Thus we would expect no more than 108 noncoding nucleotides at the 3′ end of the mRNA. The mRNA sequence has an uninterrupted reading frame extending toward the 5′ end from the UAA terminator located 100 nucleotides from the poly(A). This must be the proper reading frame because the two other frames are blocked many times, including termination codons at positions 449 and 468. The amino acid sequence predicted for this open reading frame is shown in Fig 3*B*. Note that the predicted sequence contains an uninterrupted hydrophobic domain (boxed) near the COOH terminus.

The previous data on the orientation of G in microsomal membranes (12, 14) led us to suspect that the amino acid sequence deduced above contained the domain of G associated with the membrane of the virion. However, to test this directly we digested VSV virions with proteases and determined partial amino acid sequences of those fragments of G that were protected from protease by the lipid bilayer. As shown by Mudd (3) and Schloemer and Wagner (4), digestion of VSV with protease results in complete disappearance of the intact G protein and the appearance of a small fragment (ca. 7000 daltons) of G that was presumably protected from digestion by the lipid bilayer. We found that virions purified after proteolysis contained peptides of ca. 6700 daltons (trypsin digestion) or ca. 7000 daltons (thermolysin digestion) that were not present prior to digestion (Fig. 4). We isolated both of these fragments and determined the partial NH_2-terminal amino acid sequences of each. These sequences can be aligned exactly within the predicted COOH-terminal sequence of G (Figs. 3*B* and 5). The

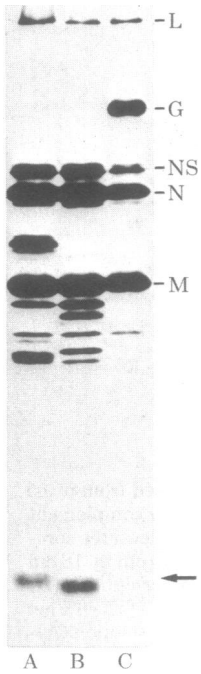
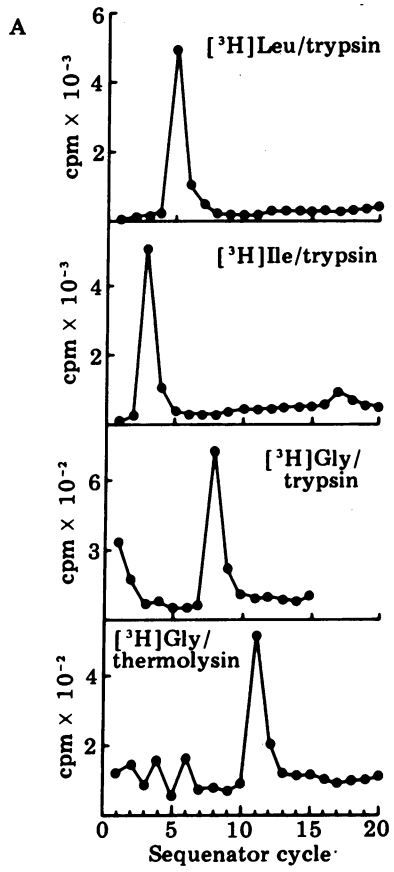


FIG. 4. Polyacrylamide gel electrophoresis of VSV proteins before and after proteolytic digestion. Samples of VSV labeled with [^{3}H]leucine were subjected to electrophoresis on a 20% polyacrylamide gel and detected by fluorography (18). Slot A, VSV digested with thermolysin; slot B, VSV digested with trypsin; and slot C, undigested VSV.



B

THERMOLYSIN TRYPSIN

65 60 55 50

Leu Ser Lys Asn Pro Ile Glu Leu Val Glu Gly Trp Phe Ser Ser Trp Lys Ser Ser Ile COOH

– – Ile – Leu – – (Gly) – – (Ser Ser) – – – – Ile

– – – – – (Ile) – – – – Gly – – – –

FIG. 5. (*A*) Partial NH_2-terminal sequence analyses of the protease-resistant fragments of the VSV G protein. Protease-resistant fragments of the VSV G protein labeled with individual tritiated amino acids were purified by gel electrophoresis and subjected to automated sequential Edman degradation. Release of radioactive amino acids at each cycle of degradation is plotted. (*B*) Partial amino acid sequences aligned with the amino acid sequence predicted by the mRNA sequence. Tentative assignments based on amino acid sequence data which gave a high background in the initial residue (Gly/trypsin, for example) are indicated by parentheses. Data for the tentative assignments of serine residues in the trypsin-generated fragment and isoleucine in the thermolysin fragment are not shown. Positions of cleavage by thermolysin and trypsin deduced by alignment of the partial amino acid sequences with the G protein sequence predicted from the mRNA are indicated by the arrows and are consistent with the enzyme specificities.

new NH_2 terminus generated by trypsin results from cleavage after the lysine located 64 amino acids from the COOH terminus of the G protein. Similarly, the slightly larger fragment generated by thermolysin must result from cleavage on the NH_2-terminal side of the leucine located 66 residues from the COOH terminus. The bands between N and M and just below M (Fig. 4, slots A and B) were reproducible and were seen only after protease digestion. They may be partial cleavage products of G.

Determination of the NH_2-Terminal Protein Sequences. To obtain the 5′-terminal sequence of the G protein mRNA and the corresponding NH_2-terminal protein sequence, we used the approach illustrated in Fig. 1. A 52-nucleotide DNA primer isolated from a cDNA clone of the VSV M protein mRNA was labeled at its 5′ end by using [γ-^{32}P]ATP and polynucleotide kinase. After hybridization to the single negative strand of VSV genome RNA, the primer was extended into the adjacent G gene by using reverse transcriptase and unlabeled deoxynucleoside triphosphates. Primer extended more than 350 nucleotides was purified by gel electrophoresis and its sequence was determined by using the Maxam–Gilbert procedure (17). This method previously revealed an intercistronic dinucleotide separating the M and G genes, the 5′-noncoding G mRNA sequence, and a sequence corresponding to the G mRNA ribosome binding site (9, 16). To obtain further G mRNA and protein sequence from the extended primer, we employed longer electrophoresis times on 6% polyacrylamide sequencing gels (20) and were able to read an unambiguous sequence extending 190 nucleotides into the G mRNA. This sequence is shown in Fig. 3*A* with a predicted G protein sequence of 53 amino acids beginning at the single AUG codon in the ribosome binding site. Two partial sequences of the NH_2 terminus of G protein synthesized *in vitro* (10, 11) are largely in agreement with the sequence predicted from the cDNA (Fig. 3*A*) but differ from it in a few positions. The single difference from the sequence of Lingappa *et al.* (10) is in a region of protein sequence that was considered tentative. Because three base changes would be required to convert the glycine codon (GGG, position 12) to specify the histidine reported (10), we believe that this assignment was probably incorrect. The three differences with the sequence of Irving *et al.* (11) might be due to VSV strain differences. However, this possibility seems unlikely at positions 12 and 19 because at least two base changes would be required to convert each codon to specify the amino acids suggested. From the previous protein sequence determination data on G synthesized *in vitro* in the presence or absence of cellular membranes (10, 11), it is clear that the first 16 amino acids constitute the leader sequence, which is cleaved during or shortly after membrane insertion.

DISCUSSION

Features of the COOH-Terminal Domain. Previous work has shown that insertion of the VSV G protein into the RER occurs when the protein is a nascent chain (7, 8). Insertion stops when the majority of the protein is inside the lumen of the RER. A region near the COOH terminus spans the membrane, and a small portion is exposed on the cytoplasmic face (12, 14). The COOH-terminal G protein sequence predicted here from the nucleotide sequence of a cDNA clone shows an uninterrupted hydrophobic domain beginning 49 amino acids from the COOH terminus and ending 30 amino acids from the COOH terminus. Presumably this hydrophobic region and the flanking charged amino acids act as a signal to stop the transfer of the protein across the membrane. This region might disrupt a hydrophilic channel (13) through the membrane and then interact strongly with the lipophilic core of the membrane. The lipophilic core of biological membrane is about 3 nm thick (22). The 20-amino acid hydrophobic sequence of G is sufficient to span this region as an α-helix with each residue advancing the helix 0.15 nm (23). The lipophilic region of G is bounded at both ends by basic residues, which might serve to position the protein precisely in the membrane by interacting with phosphates on both membrane surfaces.

Partial amino acid sequences of two small, overlapping, protease-resistant fragments of G obtained from virions could be aligned exactly with the predicted COOH-terminal G protein sequence. Thus, the COOH terminus of G is protected by the viral membrane. The protein therefore appears to be anchored in the virion membrane by the same hydrophobic domain that initially stopped transfer of the protein through the

RER. While it is clear that G spans the membrane of the RER, there is no proof that this is also true of G in virions. It is formally possible that in the virion both the NH_2 and COOH termini of G are external to the bilayer; however, the sizes of the fragments of G that survive proteolysis make this possibility unlikely. The apparent molecular weights of the protected fragments (6700 and 7000), estimated by extrapolating from gel mobility using appropriate standards (18), agree well with the molecular weights predicted for fragments containing the complete COOH-terminus (7311 and 7639; Fig. 2). If the COOH terminus of G were exposed and susceptible to protease, the sizes of the protected fragments would be closer to 4000 daltons. Thus, the hydrophilic portion of 29 COOH-terminal amino acids apparently resides within the virion. Of these internal residues 11 are basic and 2 are acidic, suggesting the possibility of ionic interactions with internal virion components. In addition, there is a single cysteine residue in this region, suggesting a possible disulfide bond with an internal component.

The complete sequence (131 amino acids) of the human erythrocyte membrane protein glycophorin has been determined (24). This protein contains an uninterrupted hydrophobic domain of 23 amino acids that spans the lipid bilayer, leaving 36 COOH-terminal residues inside and 72 NH_2-terminal residues outside the plasma membrane. Thus the location relative to the COOH terminus and the sizes of the hydrophobic portions in VSV G and glycophorin are similar, although there is no exact sequence homology. The internal COOH-terminal portion of glycophorin is highly charged, but does not show the strongly basic character of the internal portion of VSV G. This basic character may be specific for viral structure. The complete sequence of fowl plague virus hemagglutinin (a glycoprotein) has been deduced recently from the sequence of a cDNA clone (25). The sequence predicts a hydrophobic region near the COOH terminus of the protein. Although there is no direct evidence that this region of hemagglutinin spans the viral membrane, it seems likely, on the basis of the results reported here for VSV G. The membrane proteins E1 and E2 of Semliki Forest virus are also presumably anchored in the membrane in a similar manner (26).

Features of the NH_2-Terminal Domain. We have deduced an NH_2-terminal sequence of 53 amino acids for the VSV G protein, including the complete sequence of the leader peptide, from the sequence of a DNA primer extended on the VSV genome from the adjacent M gene into the G gene. The amino acid sequence predicted for the G NH_2 terminus from this DNA copy agrees in most positions with partial amino acid sequences reported previously (10, 11; Fig. 2) for the cleaved and uncleaved NH_2 termini of G. The structure of the 16-amino acid leader peptide is typical of such sequences on other secreted and membrane proteins (27) in that it has a central core of hydrophobic or nonpolar residues (position 4–14) and hydrophilic residues at both ends (positions 2, 3, 15, and 16). Most leader sequences also contain at least one proline or glycine (27), and the sequence predicted from the DNA has a glycine at position 13 (Fig. 2).

Irving *et al.* (11) have suggested that the NH_2 terminus of the mature G protein is protected by the lipid bilayer from exopeptidase digestion. However, the sequence predicted for this region of the protein reveals few features that might promote association with the membrane. Except for 5 contiguous non-polar residues following the NH_2-terminal lysine, the NH_2 terminus of the mature protein is relatively rich in basic residues (10 out of the first 25). These residues might conceivably interact with phosphate residues on the surface of the lipid bilayer.

We thank Patricia Kelley for excellent technical assistance and Caroly Goller for typing the manuscript. This work was supported by U.S. Public Health Service Grant AI 15481 from National Institute of Allergy and Infectious Diseases, National Science Foundation Grant PCM 77-25974, National Cancer Institute Grant CA 14195, National Institute of Child Health & Human Development Grant HD-09690-05, National Institute of Arthritis, Metabolism and Digestive Diseases Grant AM-18811-05, and a grant from the Sarah Andrews Foundation.

1. Etchison, J. R., Robertson, J. S. & Summers, D. F. (1977) *Virology* **78**, 375–392.
2. Reading, C. L., Penhoet, E. E. & Ballou, C. E. (1978) *J. Biol. Chem.* **253**, 5600–5612.
3. Mudd, J. A. (1974) *Virology* **62**, 573–577.
4. Schloemer, R. H. & Wagner, R. R. (1975) *J. Virol.* **16**, 237–249.
5. Cartwright, B., Smale, C. J. & Brown, F. (1969) *J. Gen. Virol.* **5**, 1–10.
6. Bishop, D. H. L., Repik, P., Obijeski, J. F., Moore, N. F. & Wagner, R. R. (1975) *J. Virol.* **16**, 75–84.
7. Rothman, J. E. & Lodish, H. F. (1977) *Nature (London)* **269**, 775–780.
8. Toneguzzo, F. & Ghosh, H. P. (1977) *Proc. Natl. Acad. Sci. USA* **74**, 1516–1520.
9. Rose, J. K. (1977) *Proc. Natl. Acad. Sci. USA* **74**, 3672–3676.
10. Lingappa, V. R., Katz, F. N., Lodish, H. F. & Blobel, G. (1978) *J. Biol. Chem.* **253**, 8667–8670.
11. Irving, R. A., Toneguzzo, F., Rhee, S. H., Hofmann, T. & Ghosh, H. P. (1979) *Proc. Natl. Acad. Sci. USA* **76**, 570–574.
12. Chatis, P. A. & Morrison, T. G. (1979) *J. Virol.* **29**, 957–963.
13. Blobel, G. & Dobberstein, B. (1975) *J. Cell Biol.* **67**, 835–851.
14. Katz, F. N. & Lodish, H. F. (1979) *J. Cell Biol.* **80**, 416–426.
15. Rose, J. K. & Iverson, L. (1979) *J. Virol.* **32**, 404–411.
16. Rose, J. K. (1980) *Cell* **19**, 415–421.
17. Maxam, A. M. & Gilbert, W. (1977) *Proc. Natl. Acad. Sci. USA* **74**, 560–564.
18. Welch, W. J. & Sefton, B. M. (1979) *J. Virol.* **29**, 1186–1195.
19. Hunkapiller, M. W. & Hood, L. E. (1978) *Biochemistry* **17**, 2124–2133.
20. Sanger, F. & Coulson, A. R. (1978) *FEBS Lett.* **87**, 107–110.
21. Knipe, D., Rose, J. K. & Lodish, H. F. (1975) *J. Virol.* **15**, 1004–1011.
22. Tanford, C. (1978) *Science* **200**, 1012–1018.
23. Dickerson, R. E. & Geis, I. (1969) *The Structure and Action of Proteins* (Harper & Row, London), p. 28.
24. Tomita, M., Furthmayr, H. & Marchesi, V. T. (1978) *Biochemistry* **17**, 4756–4770.
25. Porter, A. G., Barber, C., Carey, N. H., Hallewell, R., Threlfall, G. & Emtage, J. S. (1979) *Nature (London)* **282**, 471–477.
26. Garoff, H. & Söderlund, H. (1978) *J. Mol. Biol.* **124**, 535–549.
27. Inouye, M. & Halegoua, S. (1980) *Crit. Rev. Biochem.* **7**, 339–371.