VIRUS GENES 2:3, 223-239, 1988


# Computer Analysis Suggests a Role for Signal Sequences in Processing Polyproteins of Enveloped RNA Viruses and as a Mechanism of Viral Fusion

J.K. FAZAKERLEY[1] AND A.M. ROSS[2]
[1]*Department of Microbiology, School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA*
[2]*Department of Biochemistry, School of Dental Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA*

Received August 2, 1988
Accepted November 21, 1988

Requests for reprints should be addressed to J.K. Fazakerley, Department of Immunology, Research Institute of Scripps Clinic, 10666 N. Torrey Pines Rd., La Jolla, CA 92037, USA

Key words: viral polyprotein, fusion, signal peptidase, signalase, computer analysis

***Abstract***

We have used a computer program to scan the entire sequence of viral polyproteins for eucaryotic signal sequences. The method is based on that of von Heijne (1). The program calculates a score for each residue in a polyprotein. The score indicates the resemblance of each residue to that at the cleavage site of a typical N-terminal eucaryotic signal sequence. The program correctly predicts the known N-terminal signal sequence cleavage sites of several cellular and viral proteins. The analysis demonstrates that the polyproteins of enveloped RNA viruses—including the alphaviruses, flaviviruses, and bunyaviruses—contain several internal signal-sequence-like regions. The predicted cleavage site in these internal sequences are often known cleavage sites for processing of the polyprotein and are amongst the highest scoring residues with this algorithm. These results indicate a role for the cellular enzyme signal peptidase in the processing of several viral polyproteins. Not all high-scoring residues are sites of cleavage, suggesting a difference between N-terminal and internal signal sequences. This may reflect the secondary structure of the latter. Signal sequences were also found at the N-termini of the fusion proteins of the paramyxoviruses and the retroviruses. This suggests a mechanism of viral fusion analogous to that by which proteins are translocated through the membranes of the endoplasmic reticulum at synthesis.

## Introduction

The protein synthesis strategy of the RNA viruses of eucaryotic cells is different from that of DNA viruses. In DNA viruses, separation of information coding for functional viral proteins occurs, as in the host cell, at the level of transcription. A series of monocistronic mRNAs is produced, each of which is translated into a single functional protein. In many of the RNA viruses, separation of functional proteins occurs by segmentation of genomic information or by post-translational cleavage of a polyprotein.

The majority of positive-sense RNA viruses have a single large genome, which functions directly as a message and codes for a single large polyprotein (2-4). Since these viral mRNA molecules code for more than one functional protein, they can be considered to be polycistronic. Such messages are also produced by some genome segments of negative-sense RNA viruses (5-7). The translated product of a viral polycistronic message must be processed by a viral or a cellular protease to yield the individual viral proteins. In the well-defined case of poliovirus, it is a viral protease. The single, large mRNA is translated in the cytoplasm to produce a single, large polyprotein, at least two parts of which have protease activity and cleave the polyprotein to its several constituent parts (8). Similarly, the capsid protein of the alphaviruses has a protease activity and cleaves its own release from the translated polyprotein (9). It is also likely that cellular proteases are involved in the processing of viral polyproteins; one possibility is signal peptidase.

Membrane-associated or secretory proteins are usually preceded by a signal sequence (10) marking them for synthesis at, and translocation through, the endoplasmic reticulum (ER). The signal sequence is removed by signal peptidase. The enzyme is present on the inner surface of the membranes of the ER. There are numerous reviews and hypotheses on the mechanisms involved in membrane-bound protein synthesis, translocation, and signal peptidase cleavage (11-13). N-terminal signal sequences are present on the proteins translated from many monocistronic viral messages, for example, the hemagglutinin (HA) molecule of influenza virus (14) and the G protein of vesicular stomatitis virus (15,16). Translation of a viral polycistronic message coding for an N-terminal signal sequence will also take place on ER-bound ribosomes, with translocation of the extending polypeptide into the lumen of the ER. Just as signal peptidase cleaves the N-terminal signal sequence, so it may recognize and cleave other sites of similar specificity internally within polyprotein sequences. To examine this possibility, we have used a computer algorithm to scan the entire length of many viral polyproteins to determine the presence or absence of signal sequences. Several sequences analogous to N-terminal signal sequences were detected internally in viral polyproteins and also at the N-termini of some viral fusion proteins. We have examined known processing cleavage sites in these polyproteins and determined their similarity to signal-sequence cleavage sites. The results indicate a role for signal peptidase in the processing of many viral polyproteins.

## Methods

No strict consensus sequence can be determined in eucaryotic N-terminal signal sequences, but certain types of amino acids are present at certain positions relative to the site of cleavage (17,18). Cleavage usually follows a small neutral residue, preferably alanine, but permissibly threonine, serine, or cysteine. A similar residue is also usually found at position -3 (three residues from the cleavage site towards the N-terminus). At position -2 most residues are acceptable, except glycine and proline. Glycine and proline are usually found at position -4 or -5, from where a hydrophobic helix then extends to the N-terminus (17).

Based on a sample of 450 known eucaryotic signal sequences, von Heijne (1) devised a method for statistically predicting the most likely cleavage site of the signal peptide. A score is calculated for each amino acid in a putative signal sequence, reflecting the probability of cleavage occurring at each position in the sequence. The score is based on an analysis of 13 amino acids that are N-terminal and two that are C-terminal of the putative cleavage site. Each amino acid located in this 15-residue window is assigned a value that is a function of its relative frequency at that position in the known signal-sequence database (1). The score indicating the probability of cleavage occurring after the residue in question is given by the formula:

$$\text{Score}(i) = \sum_{p=-13}^{p=2} W(a_{i+1+p}, p),$$

where $p$ is the position of the residue in the window, $w$ is the weight matrix score derived from the von Heijne probability table (1), and $i$ is the sequence position (sequence numbering begins with residue 1).

We have developed a computer program, written in C language for the IBM PC, which implements the von Heijne algorithm (1) and scans the entire length of an amino acid sequence. Probability scores for signal peptidase cleavage sites are reported for each residue in a sequence. Viral sequences were obtained from the Genbank database through Bionet™. The results show the known sites at which viral polyproteins are cleaved, the 15-residue window around the site, the residue number at which cleavage occurs, the computer program score, indicating the resemblance of this site to a eucaryotic signal sequence cleavage site, and the rank of the score relative to the scores for all the other residues on the polyprotein. Scores greater than 0 are possible signal peptidase cleavage sites (1). The higher the score, the closer the residue and the surrounding sequence resemble that around a typical, eucaryotic N-terminal signal-sequence cleavage site.

## Results

The program was tested by scanning the sequences of several preproteins with N-terminal signal sequences in which the site of cleavage was known. These included the cellular proteins: insulin, growth hormone, alpha-amylase, relaxin, melittin, interferon, and ovalbumin (19-28), the viral influenza HA, and the vesicular stomatitis virus (VSV) G proteins (14,16, 29, 30). In most cases (Table 1), the highest scoring site on the molecule was the known site of signal sequence cleavage. Scores at cleavage sites ranged from +3.6 to +15.2. The highest scoring site was not always the cleavage site. The known cleavage site of interferon (26) gave the second highest score.

The mature proteins, after the removal of the signal peptides, demonstrated a few residues with a score above 3.6 (the lowest score found at any of the known signal-sequence cleavage sites). Mouse alpha-amylase (23) has one such residue located in a hydrophobic C-terminal domain of the molecule, a position that may not be translocated across the membrane. Ovalbumin (Table 1) has no N-terminal signal sequence but does have an uncleaved internal signal sequence (31), which is functional in the insertion of the extending molecule into the membrane (32). The site of cleavage of the signal peptide of influenza HA is the highest score on the

*Table 1.* Known N-terminal signal sequences

| Protein[a] | Sequence at[b] cleavage site | | Residue[c] | Signal[d] Score | Rank |
|---|---|---|---|---|---|
| Insulin | LVLLVVSWPGSQA | VA | 24 | 11.8 | 1 |
| Growth hormone | SLLCLLWPQEAGA | LP | 26 | 11.3 | 1 |
| Ovalbumin | MSMLVLLPDEVSG:LE | | 252 | 5.3 | 1 |
| Rat relaxin | LGFWLFLSQPCRA | RV | 22 | 10.7 | 1 |
| Mouse alpha-amylase | FVLLLSLIGFCWA | QY | 15 | 15.2 | 1 |
| Honeybee prepromelittin | LVFMVVYISYIYA | AP | 21 | 6.6 | 1 |
| Human interferon Alpha 16. | VLVLSYKSICSLG | CD | 23 | 3.9 | 2 |
| VSV G | LLYLAFLFIGVNC | KF | 16 | 10.5 | 1 |
| Influenza $HA_0$ | IIALSYIFCLALG | QD | 16 | 6.8 | 1 |
| $HA_1$-$HA_2$ | ATGMRNVPEKQT(R) | GL | 344 | −7.7 | 139 |

[a]Cellular and viral proteins with N-terminal, signal-sequence cleavage sites: insulin (19), growth hormone (21), ovalbumin (28), vesicular stomatitis virus G protein (29), and influenza virus HA (30). The influenza $HA_0$ molecule is cleaved into $HA_1$ and $HA_2$ at a trypsin-like cleavage site (33).

[b]The 13 residues N-terminal and two residues C-terminal of the cleavage site (the window scanned by the algorithm) are shown. : denotes the highest scoring site in an uncleaved signal sequence. Basic residue cleavage sites are underlined. ( ) denotes a residue that is lost at the cleavage site.

[c]Residue after which cleavage occurs.

[d]Computer calculated score for the residue after which cleavage occurs. The score indicates the similarity of the residue to that at the cleavage site of a typical N-terminal signal-sequence cleavage site. The rank position of the score relative to the scores for all other residues on the polyprotein is shown. (Most polyproteins have very few positive scoring residues, see Figs. 1 and 2).

molecule. The HA molecule is post-translationally cleaved at a trypsin-like site (33) to form two polypeptides, $HA_1$ and $HA_2$. The negative score at this site (Table 1) indicates that it has no similarity to a eucaryotic signal sequence.

The alphaviruses, bunyaviruses, arenaviruses, and picornaviruses produce polycistronic messages (2-5,34). With the exception of the picornaviruses, these are all enveloped viruses with a requirement both for processing of a polyprotein and membrane insertion of the viral envelope proteins. The coronaviruses also produce a series of messages containing information for several proteins (35), but in this case only the first protein on each message is translated. Each message is thus functionally monocistronic, and there is no requirement for the cleavage of a polyprotein.

*Alphaviruses*

The best-understood alphaviruses are Sindbis virus (SV) and Semliki Forest virus (SFV). The alphaviruses produce two mRNA species. The smaller (26S) codes for the viral structural proteins: the capsid and envelope proteins (36). The larger (42S) contains, in addition, the nonstructural proteins. The organization, cleavage, and calculated cleavage scores of the translated product of the 26S RNA of Sindbis virus are shown in Fig. 1. Following translation of the 26S RNA, the capsid protein catalyzes its own release from the following envelope proteins (9). As expected, this conserved site of autocatalytic cleavage has a negative score (Table 2) and thus has no similarity to a signal sequence.

The mechanism of cleavage that separates the envelope proteins is unclear, although the involvement of signal peptidase has been suggested (37,38). The N-terminus of the p62 (E3/E2) precursor protein contains a known signal sequence, which initiates ER-associated protein synthesis and translocation (41). This signal sequence is uncleaved (42), but for each virus the highest scoring site is shown in Table 2. The p62 precursor protein is cleaved at a conserved series of negative-scoring basic residues (Table 2). Like the cleavage of influenza $HA_0$ to $HA_1$ and $HA_2$, this is a late event, occurring distant from the site of protein synthesis (43).

The sites of cleavage between the C-terminal of p62 and the N-terminal of the following nonstructural 6 kD peptide, and between this peptide and E1, rank amongst the highest scoring residues on the polyproteins (Table 2, Fig. 1). Of the six sites in these three viruses, five are in the top five scoring residues for their polyprotein and score higher than several of the known N-terminal eucaryotic signal-sequence cleavage sites examined (Table 1). The high-scoring site preceding E1 in the C-terminus of the 6 kD peptide not only provides a cleavage site, but is also an internal signal sequence functional in the initiation of membrane-bound protein synthesis and translocation (44).

The sequences around the cleavage sites on either side of the 6 kD nonstructural peptide of the alphaviruses are indicative of cleavage by signal peptidase. In sup-

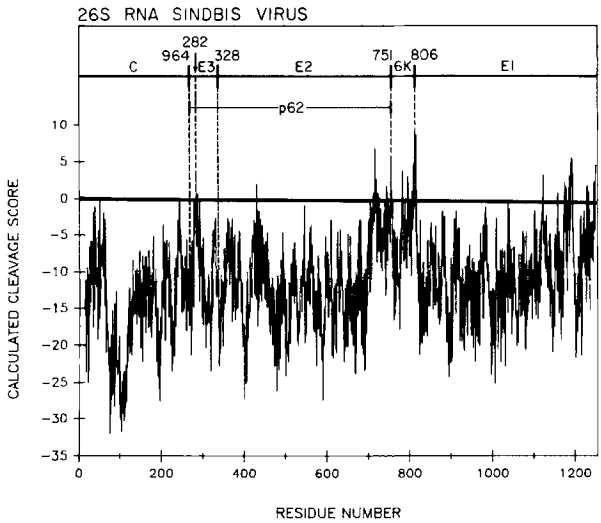


*Fig. 1.* Processing sites on the polyprotein translated from the 26S RNA of Sindbis virus. The order of the functional proteins and the residues at which cleavage occurs are shown. Very few residues within the polyprotein have a positive score (indicative of a signal-sequence cleavage site). The polyprotein is cleaved by protease activity of the capsid protein (C) at residue 964. An uncleaved signal sequence is present at the N-terminus of the precursor p62 protein, with a high score at residue 282. P62 is cleaved to E3 and E2 at residue 328, a basic residue cleavage site with a negative score. Cleavages either side of the 6 kD nonstructural protein after residues 751 and 806 give high scores, and are likely to be produced by signal peptidase.

port of this, translation studies in cell-free systems demonstrate that cleavage at these sites is associated with the activity of a membrane-bound protease. If no membranes are present in the translation system, a large polyprotein containing p62-6kD-E1 is produced (45). Addition of membranes results in the production of separate viral proteins (41,43,46).

High-scoring residues in these alphavirus polyproteins fall into groups in which cleavage often, but not necessarily, occurs after the highest scoring residue. This is also apparent in the N-terminal signal sequences of cellular proteins and in the analogous internal sequences of other viral polyproteins. The cleavage, p62-6kD at residue 751 of Sindbis virus (Table 2) is the highest score in a group of three high-scoring residues (748, 751, 753). Cleavage after residue 815 in the SFV polyprotein occurs after the second highest score in a group of four high-scoring residues (810, 812, 813, 815). The p62-6kD cleavage at residue 756 of Ross River virus (Table 2) does not have adjacent high-scoring residues.

### *Flaviviruses*

The viral message (10 kb for yellow fever virus) is the same length as the genomic RNA, there being no subgenomic message as seen in alphaviruses. The message

*Table 2*[a]. Alphaviruses

| Cleavage[b] | Sequence at cleavage site | | Residue | Signal Score | Signal Rank |
|---|---|---|---|---|---|
| Sindbis virus (SV) | | | | | |
| C-p62 (E3) | KTIKTTPEGTEEW | SA | 264 | −20.6 | 1156 |
| p62 sig. seq. | VTAMCLLGNVSFP:CD | | 282 | 3.5 | 9 |
| E3-E2 | AILRCGSSGRSRR | SV | 328 | −3.7 | 149 |
| p62 (E2)-6kD | SLALLCCVRSANA | ET | 751 | 10.9 | 1 |
| 6kD-E1 | LVVAGAYLAKVDA | YE | 806 | 2.8 | 16 |
| Semliki Forest virus (SFV) | | | | | |
| C-p62 (E3) | MVTRVTPEGSEEW | SA | 267 | −17.4 | 1009 |
| p62 sig. seq. | ITAMCVLANATFP:CF | | 284 | 3.8 | 11 |
| E3-E2 | AALTCRNGTRHRR | SV | 333 | −13.4 | 755 |
| p62 (E2)-6kD | TLGILCCAPRAHA | AS | 755 | 5.9 | 5 |
| 6kD-E1 | FLVLLSLGATARA | YE | 815 | 8.6 | 2 |
| Ross River virus (RRV) | | | | | |
| C-p62 (E3) | MVTRVTPEGTEEW | SA | 270 | −17.9 | 1024 |
| p62 sig. seq. | ALMMCILANTSFP:CS | | 285 | 5.0 | 7 |
| E3-E2 | ASMTCRNRSRHRR | SV | 334 | −15.1 | 859 |
| p62 (E2)-6kD | TLGLLCCAPRANA | AS | 756 | 7.7 | 3 |
| 6kD-E1 | FLVLLSLGASAKA | YE | 816 | 6.6 | 4 |

[a]For explanation, see legend to Table 1.
[b]The cleavage sites in the polyprotein of the translated 26S RNA of SV(2), SFV(39), and RRV(40) are known. C is the capsid protein. P62 is the precursor protein of E3 and E2. E1 and E2 are viral envelope proteins and 6kD is a 6 kD nonstructural protein. Sig. seq. is the uncleaved signal sequence at the N-terminal of p62.

contains one long open reading frame (47-50) and is the longest known mRNA that is completely translated in eucaryotic cells. Many putative viral proteins have been observed in infected cells, and it has proved difficult to determine how many proteins are coded for on the message and how they are processed. N-terminal amino-acid sequencing of some of the proteins has determined their position, but, with the exception of Kunjin virus (51), this has not been possible for all the nonstructural proteins. As with the alphaviruses, it has been speculated that cellular proteases, including signal peptidase, are involved in the processing (47,48).

***Kunjin virus.*** The processing of the polyprotein of Kunjin virus is the best understood of the flaviviruses. N-terminal sequencing has clearly established the positions of the nonstructural proteins (51). Fig. 2 shows the organization of the polyprotein, the cleavage sites, and their scores. Of the 10 cleavage sites, four give a high score with the algorithm and rank among the highest scoring residues on the polyprotein (Table 3). A fifth-site NS1-NS2A has a signal-sequence-like configuration at the five residues N-terminal of the cleavage site. However, this is not pre-

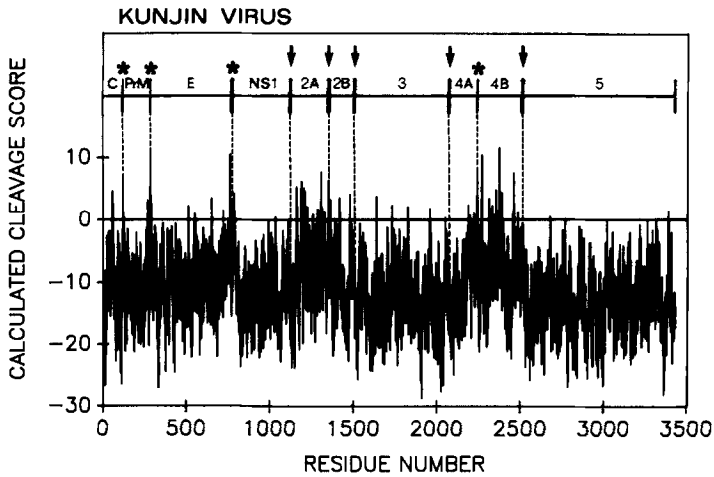


*Fig. 2.* Processing sites on the polyprotein of Kunjin virus (50,51). Processing occurs at a mixture of signal peptidase and basic residue sites. The signal peptidase sites are indicated by a star, and the basic residue sites are indicated by an arrow. The constituent proteins are shown. C is the capsid protein; M and E are viral structural proteins. NS1, 2A, 2B, 3, 4A, 4B, and 5 are nonstructural proteins.

ceded by the usual hydrophobic stretch of amino acids, and this site does not give a high score. The remaining five cleavage sites occur after pairs of basic residues, have no similarity to a signal sequence, give a negative score (Table 3), and are probably cleaved by a virally coded protease (48,51).

***Yellow fever virus.*** N-terminal sequencing has established the positions of the C, M, E, NS1, NS3, and NS5 proteins (48,52). On the basis of molecular weights, space available in the open reading frame, and likely cleavage sites, the nonstructural proteins NS2a/b and NS4a/b have been located between NS1 and NS3, and NS3 and NS5, respectively (48).

The first protein is C, the C-terminal of which is unknown but has been suggested (48) to be at residue 101, the last of four basic amino acids. Similarly, the known cleavage sites at the N-termini of M, NS3, and NS5 occur after pairs of basic residues. As expected, these basic residue sites all give a negative score (Table 3).

Analysis of possible signal-sequence cleavage sites (Table 3) suggests a signal-sequence-like region prior to residue 123, which probably functions (48,53) in the initiation of membrane-bound protein synthesis and translocation of Pre-M through the ER. In addition, if yellow fever virus follows the strategy of Kunjin virus, then residue 123 is probably a signal peptidase cleavage site yielding the N-terminus of Pre-M (Table 3). The C-terminal region of M is also signal-sequence-like, and the known site of cleavage (52) between M and E has the third highest score on the polyprotein. The known cleavage (52) between E and NS1, after

*Table 3*[a]. Flaviviridae

| Cleavage | Sequence at cleavage site | | Residue | Signal Score | Signal Rank |
|---|---|---|---|---|---|
| Kunjin virus[b] | | | | | |
| Known cleavage sites | | | | | |
| C-Pre M | IAFMIGLIAGVGA | VT | 123 | 7.4 | 9 |
| Pre M-M | RCTKTRHSRRSRR | SL | 215 | −13.0 | 2049 |
| M-E | FAVLLLLVAPAYS | FN | 290 | 11.5 | 2 |
| E-NS1 | GGVLLFLSVNVHA | DT | 791 | 4.3 | 25 |
| NS1-NS2a | HDEKTLVQSQVNA | YN | 1143 | −7.0 | 835 |
| NS2a-NS2b | AAGLVACDPNRKR | GW | 1374 | −13.6 | 2185 |
| NS2b-NS3 | VVGFWITLQYTKR | GG | 1505 | −6.8 | 813 |
| NS3-NS4a | ALKSFKDFASGKR | SQ | 2124 | −11.4 | 1687 |
| NS4a-NS4b | LICVLTLVGAVAA | NE | 2273 | 10.5 | 4 |
| NS4b-NS5 | TLVKNMEKPGLKR | GG | 2528 | −13.8 | 2225 |
| Other high-scoring sites | | | | | |
| | GVSALLLAAGCWG:QV | | 2377 | 11.7 | 1 |
| | LLLSLLLMIVLIP:EP | | 2244 | 8.5 | 5 |
| | LILVSLAALVVNP:SV | | 2463 | 7.6 | 8 |
| Yellow fever virus[c] | | | | | |
| Known cleavage sites | | | | | |
| Pre M-M | KCDSAGRSRRSRR | AI | 210 | −10.5 | 1465 |
| M-E | IALLVLAVGPAYS | AH | 285 | 9.2 | 3 |
| E-NS1 | GVIMMFLSLGVGA | DQ | 778 | 5.1 | 15 |
| ORF-NS3 | LAGWLFHVRGARR | SG | 1484 | −1.1 | 159 |
| ORF-NS5 | GMYNLWKMKTGRR | GS | 2506 | −10.9 | 1544 |
| Suggested cleavage sites | | | | | |
| ORF(C)-Pre M | ASLMRGLSSRKRR | SH | 101 | −8.6 | 1044 |
| NS1-NS2a | QVTLLDLLKLTVA | VG | 1187 | 1.2 | 70 |
| NS2a-NS2b | LCAFLATRIFGRR | SI | 1354 | −3.1 | 300 |
| NS3-NS4a | ALSEFIKFAEGRR | GA | 2107 | −8.9 | 1124 |
| NS4a-NS4b | GIKAQQSKLAQRR | VF | 2394 | −17.1 | 2800 |
| Possible alternative signal sequence cleavage sites | | | | | |
| ORF-Pre M | LILGMLLMTGGVT | LV | 123 | 3.8 | 28 |
| NS1-NS2a | VLLGAMLVGQVTL | LD | 1178 | 5.9 | 12 |
| NS2a-NS2b | LGLCAFLATRIFG | RR | 1352 | 7.8 | 6 |
| ORF-NS3 | LAGWLFHVRG (ARR) | SG | 1481 | 10.4 | 2 |
| NS4a-NS4b | MAGCGYLVSAVAA | NE | 2256 | 7.0 | 8 |

[a]For explanation, see legend to Table 1.

[b]The known cleavage sites of Kunjin virus (50,51) are shown. Four are indicative of signal peptidase cleavage. In addition, three other high-scoring sites in the polyprotein are shown. No cleavage occurs at these sites, which would in each case follow a proline or glycine residue (see discussion).

[c]The known (48) and suggested (48,53) cleavage sites in the yellow fever virus polyprotein are shown. Five alternative, high-scoring signal-sequence cleavage sites are also shown. If the processing strategy of the yellow fever virus polyprotein follows that of Kunjin virus (51), the sites predicted by the algorithm to yield the N-terminus of Pre-M, NS2a, and NS2b are likely to be used. The sites at residues 1352 and 1481 are probably not used, as each would follow a glycine residue (see discussion).

residue 778 is at another high-scoring site. Cleavage between NS1 and NS2a has been suggested (48) to occur after residue 1187. This site has a positive but low score (Table 3). Residue 1178 has a higher score (Table 3), is nine residues N-terminal of the suggested cleavage site and is a more likely signal-peptidase cleavage site in this region.

Cleavage at the N-terminus of the putative nonstructural 2b, 4a, and 4b proteins is suggested (48,53) to follow pairs of arginine residues (RR), as found at the cleavage site at the known N-terminus of NS3. With the exception of the cleavage at the N-terminus of 4b, this would be consistent with the situation in Kunjin virus (51). An alternative signal-sequence cleavage site after residue 2256 or 2385 could generate the N-terminus of NS4b (Table 3). Residue 2256 has a higher score and is the more likely cleavage site. Residue 2385 is nine residues N-terminal from the proposed (48,53) basic residue site. It is also worth noting (Table 3) two other high-scoring residues: Residue 1352 is two residues away from the suggested basic residue cleavage site between NS2a and NS2b, and residue 1481 if four residues N-terminal from the known N-terminal of NS3.

The results indicate that signal peptidase may be involved in the processing of flavivirus polyproteins. In this case, processing of the flavivirus polyprotein would occur at a combination of signal-peptidase and basic-residue cleavage sites.

*Bunyaviruses*

Several bunyaviruses, representing four of the five genera of the family have been sequenced. The viruses have a tripartate, predominantly negative-sense, RNA genome (54). Of the three genome segments, the smallest, S RNA, has two reading frames coding for the capsid protein and a small nonstructural protein (55). The S segment has no requirement for the processing of a polyprotein. The large L RNA segment codes for the viral polymerase (56), which is as yet unsequenced. The middle-sized M RNA codes for a polyprotein containing two envelope glycoproteins and a nonstructural protein (57-60). There is a requirement for processing of this M polyprotein. It is possible that any of the nonstructural proteins from the S, M, or L segments could encode a protease responsible for the cleavage of the M polyprotein. However, translation of M RNA in cell-free systems demonstrates only a requirement for microsomal membranes (61,62).

Table 4 shows the known cleavage sites of the M polyproteins of viruses representative of three different genera. In each case, the known cleavage sites are among the highest scoring sites on the polyprotein. In the case of Punta Toro and Rift Valley fever viruses, the two documented cleavage sites are the first and second highest scoring sites on the whole 1440-residue polyprotein. It is likely that signal peptidase is responsible for cleavage at all the sites indicated in Table 4, except for the cleavage site at the C-terminal position of the first envelope glycoprotein of snowshoe hare virus. This follows an arginine residue, has a negative score, and is probably mediated by a cellular trypsin-like enzyme (65). The internal

*Table 4*[a]. Bunyaviridae

| Cleavage[b] | Sequence at cleavage site | | Residue | Signal Score | Signal Rank |
|---|---|---|---|---|---|
| Punto Toro virus | | | | | |
| NS-G1 | VALLSSSVAPIIA | AP | 270 | 9.4 | 1 |
| G1-G2 | LLLTLIMMTGGNA | CS | 809 | 8.1 | 2 |
| Rift Valley fever virus | | | | | |
| NS-G2 | ALAVFALAPVVFA | ED | 153 | 9.5 | 2 |
| G2-G1 | YLMLLLIVSYASA | CS | 690 | 9.8 | 1 |
| Hantaan virus | | | | | |
| NH2-G1 | WLVMASLVWPVLT | LR | 18 | 6.9 | 4 |
| ORF-G2 | LLVLESILWAASA | SE | 648 | 7.4 | 3 |
| Snowshoe hare virus | | | | | |
| $NH_2$-G2 | MICILILFAVTAA | SP | 13 | 7.6 | 3 |
| G2-NS | GLCPGYKSLRAAR | VM | 299 | −12.1 | 745 |
| NS-G1 | LTLIKDSAIVVQA | AG | 473 | 2.4 | 15 |

[a]For explanation, see Table 1.

[b]The known cleavage sites of Punta Toro virus (4), Rift Valley fever virus (63), Hantaan virus (64) and snowshoe hare virus (5) are shown. NS denotes a nonstructural protein. G1 and G2 are viral envelope proteins. NH2 indicates the amino terminus of the polyprotein, and ORF the open reading frame, where the preceding protein is unknown.

signal-sequence-like region preceding the G1 protein of Rift Valley fever virus, as with those preceding the p62 and E1 proteins of the alphaviruses, not only provides a cleavage site, but is also functional in membrane insertion and translocation (62).

*Picornaviruses*

As expected, none of the cleavage sites of poliovirus (4) gives a positive score with this algorithm. Polio virions are assembled in the cytoplasm and processing occurs by way of virally coded proteases, as the polyprotein elongates on the free polysomes in the cytoplasm. However, 27 of the 2209 amino acid residues of the polyprotein give positive scores, four of which are greater than 4.0. The highest scoring residue is alanine at position 1105, with a score of 8.1. It is not clear why these sites in the poliovirus polyprotein are not recognized by the signal recognition particle and why translocation through the ER is not initiated.

*Viral fusion proteins*

The mechanism of viral fusion is not yet fully understood for any virus. In the paramyxoviruses, orthomyxoviruses, and retroviruses, fusion has been associated with an N-terminal hydrophobic domain of a transmembrane viral-envelope glycoprotein (76). The active glycoprotein is generated by cleavage of a larger precursor molecule at a basic residue site. Gallaher (77) has pointed out N-terminal sequence homology between the paramyxoviruses and human immunodeficiency virus (HIV) and the occurrence of the tripeptide Phe-X-Gly (where X represents any amino-acid). This peptide is also seen in reverse orientation Gly-X-Phe at the N-terminus of the influenza $HA_2$ molecule.

Scanning of several viral fusion proteins with the computer algorithm demonstrates high-scoring residues in the N-terminal fusion sequences of the paramyxoviral and retroviral fusion proteins (Table 5). The N-terminal sequences of these fusion proteins give scores as high as those of many N-terminal cellular signal sequences (Table 1), indicating their similarity to N-terminal signal sequences of cellular preproteins. The N-terminal signal sequences of some other viral fusion proteins do not resemble signal sequences. The only sites scoring greater than 2.0

*Table 5*[a]. Viral fusion proteins

| Protein[b] | Sequence at cleavage site | Residue | Signal Score | Signal Rank |
|---|---|---|---|---|
| Paramyxoviruses | | | | |
| SV5 | *FAGVVIGLAALGVATA AQ | 16 | 8.4 | 1 |
| Measles | *FAGVVLAGAALGVATA AQ | 16 | 7.6 | 1 |
| Newcastle DV | *FIGAIIGGVALGVATA AQ | 16 | 4.6 | 3 |
| Sendai | IGTIALGVATSAQITA GI | 21 | 4.4 | 1 |
| ReSV | *FLGFLLGVGSAIA SG | 13 | 7.2 | 1 |
| Retroviruses | | | | |
| Visna | *GIGLVIVLAIMIIAAAG AG | 18 | 5.6 | 1 |
| AKVp15E | *EPVSLTLALLLGGLTMG GI | 17 | 8.2 | 2 |
| MoMULVp15E | *EPVSLTLALLLGGLTMG GI | 17 | 8.2 | 1 |
| FrMULVp15E | *EPVSLTLALLLGGLTMG GI | 17 | 8.2 | 2 |
| HIVgp41 | *FVGIGALFLGFLGAAGS TM | 17 | 5.9 | 1 |

[a]For explanation, see legend to Table 1.

[b]N-terminal sequences of viral fusion proteins: SV5 (66), measles (67), Newcastle disease virus (NDV) (68), respiratory syncytial virus (ReSV) (69), Sendai virus (70), Visna virus (71), AKV murine leukemia virus (72), Moloney murine leukemia virus (MoMULV) (73), Friend murine leukemia virus (FrMULV) (74), human immunodeficiency virus (HIV-I) (75).

In order to show the complete N-terminal sequence of these fusion proteins, the sequence given is longer than the 15-residue window considered by the algorithm. * indicates the N-terminal residue. The Phe-X-Gly (FXG) tripeptide sequence noted by Gallaher (76) is underlined.

in the mature G protein of New Jersey strain vesicular stomatitis virus are residues 496 (score 6.4) and 208 (score 3.2). The highest scoring residue in the E1 (putative fusion) protein of the alphavirus Sindbis is residue 419 (score 6.7). Residues 420, 431, and 437 also score greater than 2.0.

## Discussion

N-terminal signal sequences of viral proteins are known to function in the initiation of membrane-bound protein synthesis and translocation (41,44,62). N-terminal signal sequences are cleaved by the cellular enzyme signal peptidase, the activity of which is associated with the membranes of the ER. Similarly, in-vitro translation studies have demonstrated requirement of microsomal membranes for the correct processing of some viral polyproteins (41,43,46,61,62). Indeed, it has previously been suggested that internal signal sequences function as cleavage sites in the processing of viral polyproteins (37,38,47,48). However, there has been no previous attempt to determine the number of signal-sequence-like regions on a viral polyprotein, nor to determine how closely known processing cleavage sites on a viral polyprotein resemble those in the N-terminal signal sequences of cellular proteins. The results from our computer study, based on the method of von Heijne (1), demonstrate that regions of primary sequence characteristic of the N-terminal signal sequences of eucaryotic preproteins are present internally within the polyproteins of enveloped RNA viruses. In many cases, the internal processing cleavage sites of viral polyproteins closely resemble cleavage sites in the N-terminal signal sequences of cellular preproteins. These findings support the suggestion that the cellular enzyme, signal peptidase, is involved in the processing of viral polyproteins. Our results also demonstrate the presence of signal-sequence-like sequences at the N-termini of certain known viral fusion proteins. This suggests a possible mechanism of viral fusion analogous to that whereby newly synthesized proteins are inserted into and translocated through the internal cell membranes.

This computer program can be used to locate signal-sequence-like regions within polyproteins. However, not all high-scoring residues within a viral polyprotein are sites of cleavage, and in its present form the algorithm cannot be used to definitively predict a processing cleavage site. However, the program may be of use in determining possible cleavage sites, particularly when the sizes of component proteins, and therefore the likely regions of cleavage, are known. If, for example, the program had been used to predict the cleavage sites in the polyproteins translated from the M RNAs of the bunyaviruses or the 26S RNAs of the alphaviruses, it would have been quite successful. Of the seven viruses shown in Tables 2 and 4, 12/14 of the signal peptidase cleavage sites are in the top four scoring groups for their polyprotein.

The eucaryotic signal sequence has at least three functions: a binding site for the signal recognition particle (SRP), a general hydrophobic nature of sufficient

length to allow membrane insertion, and a recognition site for signal peptidase (11-13). In addition, a binding site for a membrane translocation system may also be present (78). It is not known whether these different recognition functions are mediated by the same or different areas on the signal sequence. If separate, it is possible that some signal sequences give a high score but lack one or more of these functions. In this way, a signal sequence could be functional for recognition by SRP and for translocation, but not for cleavage. This may explain why some high-scoring sites are cleaved but others are not. This may be the case for the signal-sequence-like regions at the N-termini of the viral fusion proteins in Table 5. Analysis of many such internal sequences may determine a difference in the primary sequence between sites that are cleaved and those that are not. But absolute prediction of utilized signal peptidase cleavage sites from primary sequence analysis may not be possible. Cleavage may depend upon the secondary and tertiary structure of the molecule at the site and time of cleavage, parameters that probably cannot be determined from primary sequence nor included in a modified algorithm.

There is evidence to suggest (78) that extending polypeptides pass through a protein translocation system of the ER as a series of units with secondary structure rather than as a continuous thread. Secondary structure in translocating proteins may explain the failure to cleave certain sequences. The most favorable cleavage sites in the internal uncleaved signal sequences of ovalbumin (Table 1); the N-terminus of the p62 proteins of the alphaviruses SV, SFV, and RRV (Table 2); the high-scoring sites 2377, 2244, and 2463 in Kunjin virus, and 1352 and 1481 in yellow fever virus (Table 3); and in four of the five retroviral fusion proteins in Table 5 would in each case follow a proline or glycine residue. The presence of proline or glycine at this position in an N-terminal signal sequence does not inhibit cleavage. However, cleavage after proline or glycine residues was not found at any of the internal signal-sequence-like cleavage sites examined. It is possible that the presence of these residues in an internal sequence may introduce β-turns into the secondary structure affecting cleavage at such sites.

The sequences at the N-termini of viral fusion proteins originally pass through the ER as internal sequences within the viral fusion protein precursor, where their activity could be inhibited by their secondary structure at translocation. Conversion to an N-terminal position by cleavage of the precursor molecule occurs distant from the ER, late in viral replication, and could be a mechanism of activation just prior to release from the cell. On infection of a new cell, likely candidates for a receptor recognizing an N-terminal signal-sequence-like fusion region would be that component of the SRP responsible for signal sequence recognition, a translocation channel similar to that through which an extending polypeptide is translocated into the ER, the receptor component of such a channel, or even the signal peptidase enzyme. Such a receptor would have to be present on the outer cell membrane or the luminal side of the endosomal membrane. However, to date none of these components have been observed in either location.

Our computer program allows for change of the algorithm to consider any win-

dow of 15 residues or less. One interesting finding on analysis with a smaller window was that the N-terminus of the $HA_2$ fusion protein of influenza virus scored high for cleavage after residue 7. Interestingly, the N-terminus of $HA_2$ (GLFGAIAGF) has the configuration of a signal-sequence cleavage site without the preceding long hydrophobic sequence.

Evidence for a specific fusion receptor for the paramyxoviruses comes from fusion inhibition studies with peptides resembling the N-terminus of the Sendai virus fusion protein (79). The further the synthetic peptide is extended from the N-terminus (i.e., the more of the signal sequence region that is represented), the greater is its inhibitory effect (79), suggesting that more than the N-terminal Phe-Phe-Gly tripeptide may be important. No N-terminal signal-sequence-like regions are present on the alphavirus E1 protein or on the vesicular stomatitis virus G protein. Fusion mediated by an N-terminal signal sequence is not a possibility for these viruses, although the involvement of an internal signal sequence cannot be ruled out.

The presence of signal-sequence-like regions at the N-termini of some viral fusion proteins is highly suggestive of a mechanism of fusion similar to that by which newly synthesized proteins are inserted and translocated through the ER. Insertion and subsequent translocation of proteins across membranes may be a primitive cellular mechanism present on several cellular membranes. As with other cellular systems, viruses may have evolved to utilize this mechanism to their own advantage.

## References


1. von Heijne G., Nucleic Acids Res *14,* 4683-4690, 1986.
2. Strauss J.H. and Strauss E.G. in Nayak D.P. (ed.) *The Molecular Biology of Animal Viruses.* Marcel Dekker, New York, 1977, pp. 111-166.
3. Rice C.M. and Strauss J.H., Proc Natl Acad Sci USA *78,* 2062-2066, 1981.
4. Kitamura N., Semler B.L., Rothberg P.G., Larsen G.R.,Adler C.J., Dorner A.J., Emini E.A., Hanecak R., Lee J.J., van der Werf S., Anderson C.W. and Wimmer E., Nature *291,* 547-553, 1981.
5. Ihara T., Smith J.,Dalrymple J.M. and Bishop D.H.L., Virology *144,* 246-259, 1985.
6. Eshita Y. and Bishop D.H.L., Virology *137,* 227-240, 1984.
7. Buchmeier M.J. and Oldstone M.B.A., Virology *99,* 111-201, 1979.
8. Pallansch M.A., Kew O.M., Semler B.L., Omilianowski D.R., Anderson C.W., Wimmer E. and Rueckert R.R., J virol *49,* 873-880, 1984.
9. Aliperti, G. and Schlesinger M.J., Virology *90,* 366-369, 1978.
10. Blobel G. and Dobberstein B., J Cell Biol *67,* 835-51, 1975.
11. Rapoport T.A. and Wiedmann M., Curr Top in Membr and Trans *24,* 1-63, 1985.
12. Duffaud G.D., Lehnhardt S.K., March P.E. and Inouye M., Curr Top in Membr and Trans *24,* 65-104, 1985.
13. Walter P. and Lingappa V.R., Ann Rev Cell Biol *2,* 499-516, 1986.
14. Waterfield M.D., Espelie K., Elder K. and Skehel J.J., Br Med Bull *35,* 57-63, 1979.
15. Lingappa V.R., Katz F.N., Lodish H.F. and Blobel G., J Biol Chem *253,* 8667-8670, 1978.
16. Irving R.A., Toneguzzo R., Rhee S.H., Hofmann T. and Ghosh H.P., Proc Natl Acad Sci USA *76,* 570-574, 1979.

17. von Heijne G., Eur J Biochem *133,* 17-21, 1983.
18. Perlman D. and Halvorson H.O., J Mol Biol *167,* 391-409, 1983.
19. Hobart P.M., Shen L., Crawford R., Pictet R.L. and Rutter W.J., Science *210,* 1360-1363, 1980.
20. Chan S.J., Edmin S.O., Kwok S.C.M., Kramer J.M., Falkmer S. and Steiner D.F., J Biol Chem *256,* 7595-7602, 1981.
21. Seeburg P.H., Shine J., Martial J.A., Baxter J.D., and Goodman H.M., Nature *270,* 486-494, 1977.
22. Sekine S., Mizukami T., Nishi T., Kuwana Y., Saito A., Sato M., Itoh S. and Kawauchi H., Proc Natl Acad Sci USA *82,* 4306-4310, 1985.
23. Carne T. and Scheele G., J Biol Chem *247,* 4133-4140, 1982.
24. Hudson P., Haley J., Cronk M., Shine J. and Niall H., Nature *291,* 127-131, 1981.
25. Vlasak R., Unger-Ullmann C., Kreil G. and Frischauf A.M., Eur J Biochem *135,* 123-126, 1983.
26. Henco K., Brosius J., Fujisawa A., Fujisawa J.I., Haynes J.R., Hochstadt J., Kovacic T., Pasek M. Schambock A., Schmid J., Todokoro K., Walchli M., Nagata S. and Weissmann C., J Mol Biol *185,* 227-260, 1985.
27. Catterall J.F., Stein J.P., Kristo P., Means A.R. and O'Malley B.W., J Cell Biol *87,* 480-487, 1980.
28. McReynolds L., O'Malley B.W., Nisbet A.D., Fothergill J.E., Givol D., Fields S., Robertson M. and Brownlee G.G., Nature *273,* 723-728, 1978.
29. Gallione C.J. and Rose J.K., J Virol *54,* 374-382, 1985.
30. Verhoeyen M., Fang R., Min Jou W., Devos R., Huylebroeck D., Saman E. and Fiers W., Nature *286,* 771-776, 1980.
31. Lingappa V.R., Lingappa J.R. and Blobel G., Nature *281,* 117-121, 1979.
32. Baty D., Mercereau-Puijalon O., Perrin D., Kourilsky P. and Lazdunski C., Gene *16,* 79-87, 1981.
33. Rott R., Klenk H.D. and Scholtissek C. in Laver W.G., Bachmayer H. and Weil R. (eds.) *The Influenza Virus Hemagglutinin.* Springer, New York, 1978, pp. 83-99.
34. Auperin D.D., Romanowski V., Galinski M. and Bishop D.H.L., J Virol *52,* 897-904, 1984.
35. Lai M.M.C., Brayton P.R., Armen R.C., Patten C.D., Pugh C. and Stohlman S.A., J Virol *39,* 823-834, 1981.
36. Strauss E.G. and Strauss J.H. in Schlesinger S. and Schlesinger M.J. (eds.) *The Togaviridae and Flaviviridae.* Plenum Press, New York, 1986, pp. 35-90.
37. Strauss E.G. and Strauss J.H., Curr Top Microbiol Immunol *105,* 1-98, 1983.
38. Schlesinger M.J. and Schlesinger S. in Schlesinger S. and Schlesinger M.J. (eds.) *The Togarviridae and Flaviviridae.* Plenum Press, New York, 1986, pp. 121-148.
39. Garoff H., Frischauf A.M., Simons K., Lehrach H. and Delius H., Nature *288,* 236-241, 1980.
40. Dalgarno L., Rice C.M. and Strauss J.H., Virology *129,* 170-187, 1983.
41. Garoff H., Simons K. and Dobberstein B., J Mol Biol *124,* 587, 1978.
42. Bonatti S. and Blobel G., J Biol Chem *254,* 12261-12264, 1979.
43. Bonatti S., Cancedda R. and Blobel G., J Cell Biol *80,* 219-224, 1979.
44. Hashimoto K., Erdel S., Keranen S., Saraste J. and Kaariainen L., J Virol *38,* 34-40, 1981.
45. Schlesinger M.J. and Schlesinger S., J Virol *11,* 1013-1016, 1973.
46. Wirth D.F., Lodish H.F. and Robbins P.W., J Cell Biol *81,* 154-162, 1979.
47. Westaway E.G., Adv Virus Res *33,* 45-90, 1987.
48. Rice C.M., Lenches E.M., Eddy S.R., Shin S.J., Sheets R.L. and Strauss J.H., Science *229,* 726-733, 1985.
49. Castle E., Leidner U., Nowak T., Wengler G. and Wengler G., Virology *149,* 10-26, 1986.
50. Coia G., Parker M.D., Speight G., Byrne M.E. and Westaway E.G., J Gen Virol *69,* 1-21, 1988.
51. Speight G., Coia G., Parker M.D. and Westaway E.G., J Gen Virol *69,* 23-34, 1988.
52. Bell J.R., Kinney R.M., Trent D.W., Lenches E.M., Dalgarno L. and Strauss J.H., Virology *143,* 224-229, 1985.
53. Rice C.M., Strauss E.G. and Strauss J.H., in Schlesinger S. and Schlesinger M.J. (eds.) *The Togaviridae and the Flaviviridae.* Plenum Press, New York, 1986, pp. 279-326.
54. Bishop D.H.L., Beaty B.J. and Shope R.E., Ann NY Acad Sci *354,* 84-106, 1980.

55. Fuller F., Bhown A.S. and Bishop D.H.L., J Gen Virol *64,* 1705-1714, 1983.
56. Endres M.J., Jansen R.S., Gonzalez-Scarano F. and Nathanson N., J Gen Virol, *70,* 223-228, 1989.
57. Gentsch J.R. and Bishop D.H.L., J Virol *30,* 767-776, 1979.
58. Fuller F. and Bishop D.H.L., J Virol *41,* 643-648, 1982.
59. Ihara T., Smith J., Dalrymple J.M. and Bishop D.H.L., Virology *144,* 246-259, 1985.
60. Schmaljohn C.S., Schmaljohn A.L. and Dalrymple J.M., Virology *157,* 31-39, 1987.
61. Ulmanen I., Seppala P. and Pettersson R.F., J Virol *37,* 72-79, 1981.
62. Kakach L.T., Wasmoen T.L. and Collett M.S., J Virol *62,* 826-833, 1988.
63. Collett M.S., Purchio A.F., Keegan K., Frazier S., Hayes W., Anderson D.K., Parker M.D., Schmaljohn C. and Dalrymple J.M., Virology *144,* 228-245, 1985.
64. Schmaljohn C.S., Schmaljohn A.L. and Dalrymple J.D., Virology *157,* 31-39, 1987.
65. Fazakerley J.K., Gonzalez-Scarano F., Strickler J., Dietzschold B., Karush F. and Nathanson N., Virology, *167,* 422-432, 1988.
66. Paterson R.G., Harris T.J.R. and Lamb R.A., Proc Natl Acad Sci USA *81,* 6706-6710, 1984.
67. Richardson C., Hull D., Greer P., Hasel K., Berkovich A., Englund G., Bellini W., Rima B. and Lazzarini R., Virology *155,* 508-523, 1986.
68. Chambers P., Millar N.S. and Emmerson P.T., J Gen Virol *67,* 2685-2694, 1986.
69. Collins P.L., Huang Y.T. and Wertz G.W., Proc Natl Acad Sci USA *81,* 7683-7687, 1984.
70. Hsu M.C. and Choppin P.W., Proc Natl Acad Sci USA *81,* 7732-7736.
71. Sonigo P., Alizon M., Staskus K., Klatzmann D., Cole S., Danos O., Retzel E., Tiollais P., Haase A. and Wain-Hobson S., Cell *42,* 369-382, 1985.
72. Lenz J. Crowther R., Straceski A. and Haseltine W., J Virol *42,* 519-529, 1982.
73. Shinnick T.M., Lerner R.A. and Sutcliffe J.G., Nature *293,* 543-548. 1981.
74. Koch W., Huntsmann G. and Friedrich R., J Virol *45,* 1-9, 1983.
75. Ratner L., Haseltine W., Patarca R., Livak K.J., Starcich B., Josephs S.F., Doran E.R., Rafalski J.A., Whitehorn E.A., Baumeister K., Ivanoff L., Petteway S.R. Jr., Pearson M.L., Lautenberger J.A., Papas T.S., Ghrayeb J., Chang N.T., Gallo R.C. and Wong-Staal F., Nature *313,* 277-284, 1985.
76. White J., Kielian M. and Helenius A., Quart Rev Biophys *16,* 151-195, 1983.
77. Gallaher W.R., Cell *50,* 327-328, 1987.
78. Singer S.J., Maher P.A. and Yaffe M.P., Proc Natl Acad Sci USA *84,* 1015-1019, 1987.
79. Richardson C.D., Scheid A. and Choppin P.W., Virology *105,* 205-222, 1980.