STRUCTURE AND REPLICATION OF ALPHAVIRUS RNAs

Thesis by

Jing-hsiung James Ou

In Partial Fulfillment of the Requirements
For the Degree of
Doctor of Philosophy

California Institute of Technology
Pasadena, California

1982

(Submitted April 8, 1982)

Acknowledgements

First of all I would like to thank my thesis advisor, Jim Strauss, for his advice and guidance throughout this work. I am also grateful to Norman Davidson, Tom Maniatis and Ellen Strauss for their past helpful discussions, and to Giuseppe Attardi, Norman Davidson, Elliot Meyerowitz and Herschel Mitchell for being my thesis committee.

Lynn Dalgarno, Steve Monroe, Charlie Rice, Sondra Schlesinger, Ellen Strauss and Dennis Trent have in the past collaborated with me on my research.

Many thanks are also due to these people:

to Charlie Rice, for his solving many of my research problems and reading many of my manuscripts;

to Jeff Mayne, for his teaching me how to harvest viruses and his sometimes being my English teacher;

to John Bell, for his frequent help in fixing instruments so that my experiments could proceed;

to Vann Parker, for his suggestion of doing some of the projects in this thesis;

to Edith Lenches, for preparing numerous chick cell cultures for me;

to Tim Hunkapiller and Michael Douglas, for teaching me how to use computer programs;

to Connie Katz, for help in putting my thesis together;

to the people in the office who in the past have typed manuscripts for me.

Finally, my deepest appreciation goes to my wife, Ker-hwa, whose patience and encouragement (and sometimes pushing and dragging) bring this thesis to completion.

This work has been supported by the Arthur McCallum Fund, Lucy Mason Clark Endowment Fund, Li Ming Endowment Fund, California Foundation for Biomedical Research and Training Grant (GM 00086) from the National Institutes of Health. Preparation of this thesis was supported by the Jean Weigle Memorial Fund.

Abstract

Both ends of the alphavirus genomic RNA are potentially important in its replication. The region preceding and including the 5'-end of the subgenomic 26S RNA in genomic RNA might also be involved in 26S RNA transcription. Sequences of these regions of up to 10 alphaviruses were determined by using strategies including enzymatic, chain-termination and cDNA sequencing methods.

Comparison of the nucleotide sequences reveals three highly conserved sequences. The first conserved sequence is 19 nucleotides in length and is located at the extreme 3'-end next to the poly(A) tail. The second conserved sequence, which is 21 nucleotides in length, precedes the 5'-end of 26S RNA and includes the first two nucleotides of it. The third conserved sequence is 51 nucleotides in length and is located at a position of about 130 to 150 nucleotides from the 5'-end, depending on the virus. The last conserved sequence in all alphaviruses examined is capable of forming two stable hairpin structures and could also base-pair stably with the 3'-terminal sequences to cyclize genomic RNAs. Besides these three conserved sequences, a highly conserved stem and loop structure could also be formed at the extreme 5'-end of genomic RNA.

Defective interfering (DI) RNAs of alphaviruses are mutated genomic RNAs which often contain deleted, repeated and translocated sequences, but yet retain all elements essential for their replication. By studying the sequence organization of alphavirus DI RNAs, and the 3'-terminal sequences of the genomic RNAs of two alphavirus variants and their replication, the importance of these conserved sequences and secondary structures in alphavirus replication are discussed.

Both the 3'- and 5'-terminal sequences of several alphavirus 26S RNAs were also determined. Results show that 26S and genomic RNAs are coterminal. Together with the results previously published, the total length of the 26S RNAs of two alphaviruses, Sindbis virus and Semliki Forest virus, were determined to be 4102 and 4074 nucleotides, respectively.

The NH_2- and COOH-terminal sequences of the precursors of nonstructural proteins (translated from genomic RNA) and structural proteins (translated from 26S RNA) of several alphaviruses were deduced from the nucleotide sequences determined. The initiation codons of most alphavirus genomic and 26S RNAs are preceded by the sequence CANN. To determine the importance of these tetranucleotides, their sequences in 65 eucaryotic mRNAs were surveyed. Results show that the sequence distribution of these tetranucleotides are non-random and they might be involved in initiation of translation.

The 3'-noncoding regions of alphavirus genomic RNAs contain AU rich sequences. Sequence organization in the 3'-noncoding regions is similar to those in alphavirus DI RNAs. Mechanisms for the generation of these sequence rearrangements are discussed.

Table of Contents

Chapter 1

Replication of alphavirus RNAs

Alphaviruses which belong to the family Togaviridae comprise about 20 species of viruses (Table 1). Transmitted by insect vectors many of them can cause severe epidemic diseases in man and equine animals (Porterfield, 1980).

Matured alphavirus particle is enveloped in a lipid bilayer. This lipid envelope which is acquired from the host cell membrane through the budding process is found associated with two to three species of virus-specific glycoproteins. Inside this envelope there is the nucleocapsid which is composed of a single-stranded genomic RNA of plus polarity and about 300 molecules of capsid proteins (for review see Strauss and Strauss, 1977). The genomic RNA, which has a molecular weight of 4.0-4.5 x 10^6 daltons and a sedimentation coefficient of 49S, is infectious as naked RNA (Wecker, 1959). After entering the host cell, the genomic 49S RNA serves as the message for translating a large precursor polypeptide which is subsequently cleaved to several nonstructural proteins required for RNA synthesis. Four stable nonstructural proteins in the order of ns70-ns86-ns78-ns60 (Lachmi and Kaariainen, 1976) and three stable nonstructural proteins in the order of ns60-ns89-ns82 (Brzeski and Kennedy, 1977) have been identified in Semliki Forest virus (SFV) and Sindbis virus (SIN) infected cells, respectively. SFV and SIN are the two best studied alphaviruses. Most of the complementation experiments have been done with SIN and four complementation groups, A, B, F and G, have been assigned to its temperature-sensitive (ts) mutants which show an RNA-negative phenotype at the nonpermissive temperature (Strauss and Strauss, 1980). It has been suggested that ns60 and ns89 found in SIN infected cells, are the gene products of G and A, respectively (Fuller and Marcus, 1980).

The genomic 49S RNA as well as a subgenomic 26S RNA are the two major species of RNAs found in alphavirus-infected cells (Simmons and Strauss, 1974a). The latter, which is coterminal with 49S RNA (Ou et al., 1981; also Chapter 2, this thesis) and contains the 3'-one-third of the sequence information of it (Simmons

Table 1. The Alphaviruses[a]

Subgroup	Virus	Abbreviation used in this thesis
EEE	Eastern equine encephalitis	EEE
WEE	Western equine encephalitis	WEE
	Highlands J[b]	HJ
	Sindbis	SV or SIN
	Middleburg	MBV
	Fort Morgan	
	Whataroa	
	Aura	
	Ndumu	
VEE	Venezuelan equine encephalitis	VEE
	Semliki Forest	SFV
	Ross River	RRV
	Bijou Bridge	
	Everglades	
	Mucambo	
	Pixuna	
	Chikungunya	
	O'nyong-nyong	
	Bebaru	
	Getah	
	Sagiyama	
	Mayaro	
	Una	

a. This Table is modified from Porterfield (1980).

b. Highlands J virus was originally not listed on the Table. Its classification is according to the report of Trent and Grant (1980).

and Strauss, 1972a; Kennedy, 1976), is the message for structural proteins (Simmons and Strauss, 1974b; Garoff et al., 1980a; Garoff et al., 1980b; Rice and Strauss, 1981). As for most eucaryotic mRNAs, both 49S and 26S RNAs are capped at their 5'-termini (Hsu-Chen and Dubin, 1976) and polyadenylated at their 3'-termini (Sawicki and Gomatos, 1976; Frey and Strauss, 1978). The size of the poly(A) ranges from 40 to 200 nucleotides and has a mean length of about 70 nucleotides (Frey and Strauss, 1978).

When alphaviruses are propagated at high multiplicities, defective interfering (DI) particles are generated. DI particles are unable to propagate alone, but in the presence of helper viruses (which are usually their parental or closely-related viruses), they can often replicate and interfere with the replication of those viruses (Stollar, 1980). DI RNAs often contain deleted, repeated and translocated sequences (Soderlund et al., 1981; Monroe et al, 1982) but yet retain all elements essential for their replication. For this reason DI RNAs are useful tools for studying the replication of alphavirus RNAs.

Replication Forms of Alphavirus RNAs

Besides 49S and 26S RNAs, a multi-stranded RNA (replicative intermediate, RI) can also be isolated from alphavirus-infected cells (Simmons and Strauss, 1972b; Segal and Sreevalsan, 1974; Martin and Burke, 1974; Sawicki and Gomatos, 1976). Digesting the single-stranded tails of RI with ribonucleases releases three double-stranded RNA cores which are called RF I, RF II and RF III (Simmons and Strauss, 1972b; Martin et al., 1979). Hybridization-competition experiments showed that RF I and RF III contain the complements of 49S and 26S RNAs, respectively, and RF II contains the complement of the non-26S RNA region of 49S RNA. It was postulated that RF I is derived from an RI engaged in 49S RNA synthesis (RIa) and RF II and RF III are derived from another RI engaged in 26S RNA synthesis (RIb) (Simmons and Strauss, 1972b). This hypothesis predicts that the complement of

49S RNA (minus-stranded RNA) is the same template for transcribing both 49S and 26S RNAs. The fact that RF I can be isolated directly from infected cells, but RF III cannot, greatly supports this hypothesis (Segal and Sreevalsan, 1974; Martin et al., 1979).

Because of the finding of RF II it was proposed that the synthesis of 49S and 26S RNAs could begin simultaneously on the same template and certain control function(s) associated with RIb prevent the transcription of 49S RNA reading into 26S RNA region. This control function(s) is reversible and thus could convert RIb into RIa for 49S RNA transcription (Strauss and Strauss, 1977). Alternatively, it seems possible that 49S RNA transcription is actually controlled at the level of initiation and thus, once a template is used for 26S RNA transcription no 49S RNA transcription could be initiated on this template. This would require that 49S RNA anneal to RIb during the RNA isolation processes to explain why RF II is generated when the RNA isolated is treated with ribonucleases. Such a mechanism could also explain why the time required for labeling RF II is much longer than those of RF I and RF III in vivo (Simmons and Strauss, 1972b), because labels can be incorporated directly into RI which are the precursors of RF I and RF III, but labeling RF II would depend on the amount of 49S RNA labeled in infected cells, which requires a certain length of time to reach its equilibrium state.

Transcription Initiation of Minus-Stranded RNA

Comparative studies of the 3'-terminal sequences of several alphavirus genomic RNAs revealed a highly conserved sequence 19 nucleotides in length which is located at the extreme 3'-end next to the poly(A) tail (Ou et al., 1981; Ou et al., 1982b; also Chapters 2 and 3, this thesis). This sequence is also conserved in DI RNAs (Monroe et al., 1982; Lehtovaara et al., 1981; also Chapter 4, this thesis) and thus might be the replicase recognition site for minus-stranded RNA replication. Evidence supporting the involvement of this conserved sequence in minus-stranded RNA synthesis comes from studying the replication of a Sindbis variant, SIN-16. SIN-16 was isolated from a persistently infected BHK cell line 16 months postinfection with SIN (Weiss and Schlesinger, 1981), The replication of SIN-16 is not interfered with by DI particles derived from wild-type SIN. Cell cultures coinfected with SIN-16 and DI particles show no DI RNA synthesis. Studying the 3'-terminal sequence of SIN-16 genomic RNA revealed a single-base substitution in the 3'-highly conserved region (Monroe et al., 1982; also Chapter 4, this thesis). These results suggest that SIN-16 has an altered replicase which recognizes the new 3'-conserved sequence and thus does not replicate DI RNAs.

A simple recognition signal at the 3'-end of genomic RNA for synthesizing minus-stranded RNA, however, cannot explain two observations: first, 26S RNA, which has the 3'-terminal sequence of the genomic RNA, is apparently not used as the template for minus-stranded RNA synthesis; and second, SIN-1, a variant isolated from the same BHK cell line as that of SIN-16 one month postinfection (Weiss and Schlesinger, 1981), has the parental 3'-conserved sequence in its genome, but is somewhat resistant to the interference of DI particles (Monroe et al., 1981; also Chapter 4, this thesis). The simplest explanation of these two observations is that the replication of the minus-stranded RNA involves other non-26S RNA sequences. The genomic RNAs of SIN and SFV are able to form similar circular structures (Hsu et al., 1973; Frey et al., 1979). Because serological experiments (Porterfield, 1980) and nucleotide sequence studies (Ou et al., 1982b; also Chapter 3, this thesis) suggest that SIN and SFV are distally related alphaviruses, this circular structure is probably conserved in all alphavirus genomic RNAs and might be important in alphavirus replication. It is conceivable that cyclizing the RNA molecule could bring the non-26S RNA sequence close to the 3'-end and enable the replicase to also recognize it and initiate minus-stranded RNA synthesis. Sequences which could be involved in cyclization have been located in the genomic RNAs of several alphaviruses: the 3'-end sequences involved in base-pairing are located at about 240 nucleotides from the poly(A) tails, and the 5'-end sequences involved are either partly or completely located within a highly conserved sequence which is 51 nucleotides in length and is located at about 145 nucleotides from the 5'-end (Chapter 5, this thesis). These conserved 51 nucleotides could be the non-26S RNA sequence involved in minus-stranded RNA synthesis. Note that although the 3'-end sequence potentially involved in forming circles has been deleted in an SFV DI RNA, the conserved 51 nucleotides have been rearranged to a position about 190-250 nucleotides from the poly(A) tail (Lehtovaara et al., 1981). However, if this highly conserved 5'-end sequence is involved in minus-stranded RNA synthesis, its translocation would result in a different configuration of the two sequences which are recognized by the replicase.

Poly(U) tracts with a size range similar to that of poly(A) are found at the 5'-end of minus-stranded RNA (Sawicki and Gomatos, 1976; Frey and Strauss, 1978) and thus it is possible that once the replicase recognizes the 3'- and 5'-conserved sequences, the transcription of the minus-stranded RNA starts from the first nucleotide at the 3'-end of genomic 49S RNA.

Termination of Minus-Stranded RNA Synthesis in Infected Cells

The synthesis of minus-stranded RNA ceases at about 3-4 hours postinfection (Bruton and Kennedy, 1975; Sawicki and Sawicki, 1980). 26S RNA is transcribed at a faster rate than is 49S RNA. The molar ratio of newly synthesized 26S RNA to 49S RNA is about 9 in SIN infected cells at the time when the rate of plus-stranded RNA synthesis is constant (Sawicki et al., 1981b). For this reason a conceivable mechanism for the cessation of minus-strand synthesis is that 26S RNA,which has the same 3'-conserved sequence as that of 49S RNA (the template for minus-stranded RNA transcription), might function as a competitive inhibitor and shut off minus-stranded RNA synthesis.

Functions associated with the A gene are required for termination of minus-stranded RNA syntheis (Sawicki et al., 1981b). Because the A gene is required for 26S RNA transcription (Keranen and Kaariainen, 1979), it is possible that mutations in the A gene reduce the amount of 26S RNA transcribed and enable minus-stranded RNA synthesis to continue. Polymerase activities involved in alphavirus RNA synthesis have been found associated with certain cytopathic vacuoles (Grimley et al., 1972). From pulse-labeling experiments it has been observed that plus-stranded RNAs are first synthesized at these vacuoles and then transported to the rough endoplasmic reticulum for protein translation or encapsidation (Friedman et al., 1972). It is thus likely that only those 26S RNAs which are still associated with these cytopathic vacuoles could be actively involved in terminating minus-stranded RNA synthesis. This could explain why upon shifting a SIN ts mutant belonging to the complementation group A to the nonpermissive temperature at 12 hours postinfection, when large amounts of 26S RNA have accumulated in infected cells, minus-stranded RNA synthesis can still resume (Sawicki et al., 1981b). Most of the plus-stranded RNA synthesized could have been transported away from these cytopathic vacuoles, and termination of 26S RNA synthesis would immediately change the ratio of 26S to 49S RNAs associated with these vacuoles and lead to a resumption of minus-stranded RNA synthesis. A detailed description of the nonstructural genes required for alphavirus RNA synthesis will be discussed later.

Transcription Initiation of Genomic RNA

Comparative studies of the 5'-terminal sequences of several alphavirus genomic RNAs reveals a highly conserved stem and loop structure at the extreme 5'-end, next to the cap structure (Chapter 6, this thesis). Except for an additional G at the 3'-end, the 3'-terminal sequence of SFV minus-stranded RNA has been shown complementary to the 5'-end of its genomic RNA (Wengler et al., 1979). Because of the sequence complementarity similar stem and loop structures would also be present at the 3'-ends of minus-stranded RNAs. The position of this structure suggests that it might be the replicase recognition signal for controlling the genomic 49S RNA transcription.

Because no complete 5'-terminal sequences of alphavirus DI RNAs have yet been published, it is not clear whether this stem and loop structure is conserved in them or not. However, a common, modified 5'-terminal sequence which might affect the configuration of the stem and loop structure has been found at the 5'-ends of a group of SIN DI RNAs (Chapter 6, this thesis). Heterogeneous 5'-terminal sequences are also found at the 5'-ends of a group of SFV DI RNAs (Pettersson, 1981). How these modifications are generated is not clear at this time; however, because similar kinds of sequence modifications are not found in genomic RNAs, there must be certain selection pressures against them. It is conceivable that an unmodified 5'-terminal sequence is required for the translation of nonstructural proteins. Alternatively, as discussed above, there might be certain control functions associated with RIb which prevent the initiation of 49S RNA transcription, and an unmodified 5'-terminal sequence might be involved in functions such as these which are not required for DI RNA replication.

Transcription Initiation of 26S RNA

Comparative sequence studies also revealed a highly conserved sequence which is about 21 nucleotides in length (Ou et al., 1982a; also Chapter 5, this thesis). This sequence precedes the 5'-end of 26S RNA and includes the first two nucleotides of it. (For simplicity this sequence will be referred to as the junction sequence.) The location of this junction sequence again suggests that its complement in the minus-strand might be the initiation site for 26S RNA transcription. In contrast to the conserved sequences required for minus-stranded RNA synthesis described above, this junction sequence has been deleted in an SFV DI RNA (Lehtovaara et al., 1981).

Polyadenylation of Plus-Stranded RNAs

Because poly(U) tracts are present at the 5'-ends of alphavirus minus-stranded RNAs, it is conceivable that poly(A) of both 49S and 26S RNAs are transcribed directly from the template. However, poly(A) tails of the genomic RNAs of SIN (Frey and Strauss, 1978) and poliovirus (Spector and Baltimore, 1974), another plus-stranded RNA virus, if removed can be regenerated during virus replication, and thus polyadenylation could also be a post-transcriptional event. AU-rich sequences are found at the 3'-ends of the genomic RNAs of alphaviruses (Ou et al., 1981; Ou et al., 1982b; also Chapters 2 and 4, this thesis) and poliovirus (Racaniello and Baltimore, 1981). The consensus polyadenylation signal, AAUAAA, of most eucaryotic mRNAs (Proudfoot and Brownlee, 1976) is absent in the genomic RNAs of these viruses, and instead these AU-rich sequences might be their polyadenylation signals. AU-rich sequences serving as polyadenylation signals have also been proposed for two minus-stranded RNA viruses, influenza virus (Robertson et al., 1981) and vesicular stomatitis virus (Schubert and Lazzarini, 1981).

Poly(A) has been thought important in stabilizing mRNAs (Revel and Groner, 1978). The conservation of poly(A) and AU-rich sequences in DI RNAs (Lehtovaara et al., 1981; Monroe et al., 1982; also Chapter 4, this thesis) suggests that they are important in alphavirus replication.

Nonstructural Genes Required for RNA Synthesis

We have proposed that the three distinct recognition signals described above are involved in synthesizing minus-stranded RNA, genomic 49S RNA and 26S RNA. This would require at least three replicases (or replicases with at least three different functional recognition sites). As mentioned above, four genes, A, B, F and G, are required for alphavirus RNA synthesis. It is possible that the gene products of three of them are RNA polymerases (or initiation factors) which recognize the three different recognition signals. Ts-mutants of complementation group A are unable to synthesize 26S RNA at the nonpermissive temperature, but can continue the synthesis of minus-stranded and genomic RNAs (Keranen and Kaariainen, 1979). This suggests that the gene product of A cistron could be the replicase (or initiation factor) which recognizes the complement of the conserved junction sequence required for 26S RNA synthesis. The only ts-mutant, ts-11 of SIN, assigned to the complementation group B is unable to synthesize minus-stranded RNA at the nonpermissive temperature (Sawicki et al., 1981a) and thus the gene product of B cistron might be the replicase (or initiation factor) which recognizes both 3'- and 5'-conserved sequences required for minus-stranded RNA synthesis. Three SIN ts-mutants, ts-7, ts-18 and ts-134 have been identified belonging to the complementation group G (Strauss and Strauss, 1980). The replication of these ts-mutants has not been studied in detail. It was suggested that G cistron is also required for 26S RNA synthesis (Keranen and Kaariainen, 1979). Alternatively, the gene product of G cistron might be the replicase (or the initiation factor) which recognizes the stem and loop structure at the 3'-end of minus-stranded RNA and is essential for genomic RNA synthesis. The mutant ts-6 belongs to the last complementation group F. This mutant is unable to continue the synthesis of all three RNA species when shifted to the nonpermissive temperature and does not accumulate large precursors of nonstructural proteins (Keranen and Kaariainen, 1979; Sawicki et al., 1981a). This suggests that F-peptide is important for synthesizing all three alphavirus RNAs and is most likely an elongation factor of the RNA polymerases.

Generation of DI RNAs

As discussed above DI RNAs of alphaviruses are aberrant genomic RNAs which contain deleted, repeated and translocated sequences. Similar kinds of sequence rearrangements are also found at the 3'-noncoding regions of genomic RNAs where evolutionary constraints are less prominent (Ou et al., 1982b; also Chapter 3, this thesis). It is conceivable that the generation of these sequence rearrangements is due to improper copying of the template RNA by the replicase during RNA replication. A possible mechanism for generating DI RNAs is shown in Figure 1: jumping over a secondary structure would delete a sequence (Figure 1a); if this "jumping" occurs inside a loop structure, repeated sequence would be generated (Figure 1b); several independent events of deletions and duplications could translocate sequences (Figure 1c). A second hypothesis is that of a template-switching mechanism in which the polymerases are able to switch templates during RNA transcription. Such a mechanism has been proposed for the generation of the DI RNAs of minus-stranded RNA viruses (Lazzarini et al., 1981), and it is possible that a similar kind of mechanism could be responsible for generating alphavirus DI RNAs. To investigate the possibilities of these mechanisms, sequences which have been deleted from DI RNAs were examined. Interestingly, sequences at the ends of deletions are often capable of forming stable hairpin structures (Figures 2a and 2b). In addition, a deletion of 21 nucleotides, which code for a string of seven amino acids has been found in the 26S RNA of a mutant of Ross River virus (RRV), another alphavirus (Dalgarno et al., personal communication) and these 21 nucleotides could also form a stable hairpin structure (Figure 2c). These three observations suggest a jump-over mechanism. The structure in Figure 2d, on the contrary, suggests a template-switching mechanism: the 3'-terminal sequence of an SFV DI RNA is identical to that of its genomic RNA for 106 nucleotides (excluding the poly(A) tail) (Lehtovaara et al., 1981). The DI RNA sequence then diverges into non-26S RNA sequence (a deletion of more than 4000 nucleotides). Nucleotides 107-138 of the genomic RNA could form a stable hairpin structure (Figure 2d). This observation suggests that secondary structures might impede the process of RNA transcription and result in the translocation of the replicase to another position of the template. An alternative explanation which involves

Figure 1. Possible Models for Generating Alphavirus DI RNAs

(a) Jump-over a hairpin structure would generate a deletion. (b) Circling around inside a hairpin loop would generate repeated sequences. (c) Several independent events of duplication and deletion would generate translocated sequences. Dash-lined arrows indicate sequences being deleted.

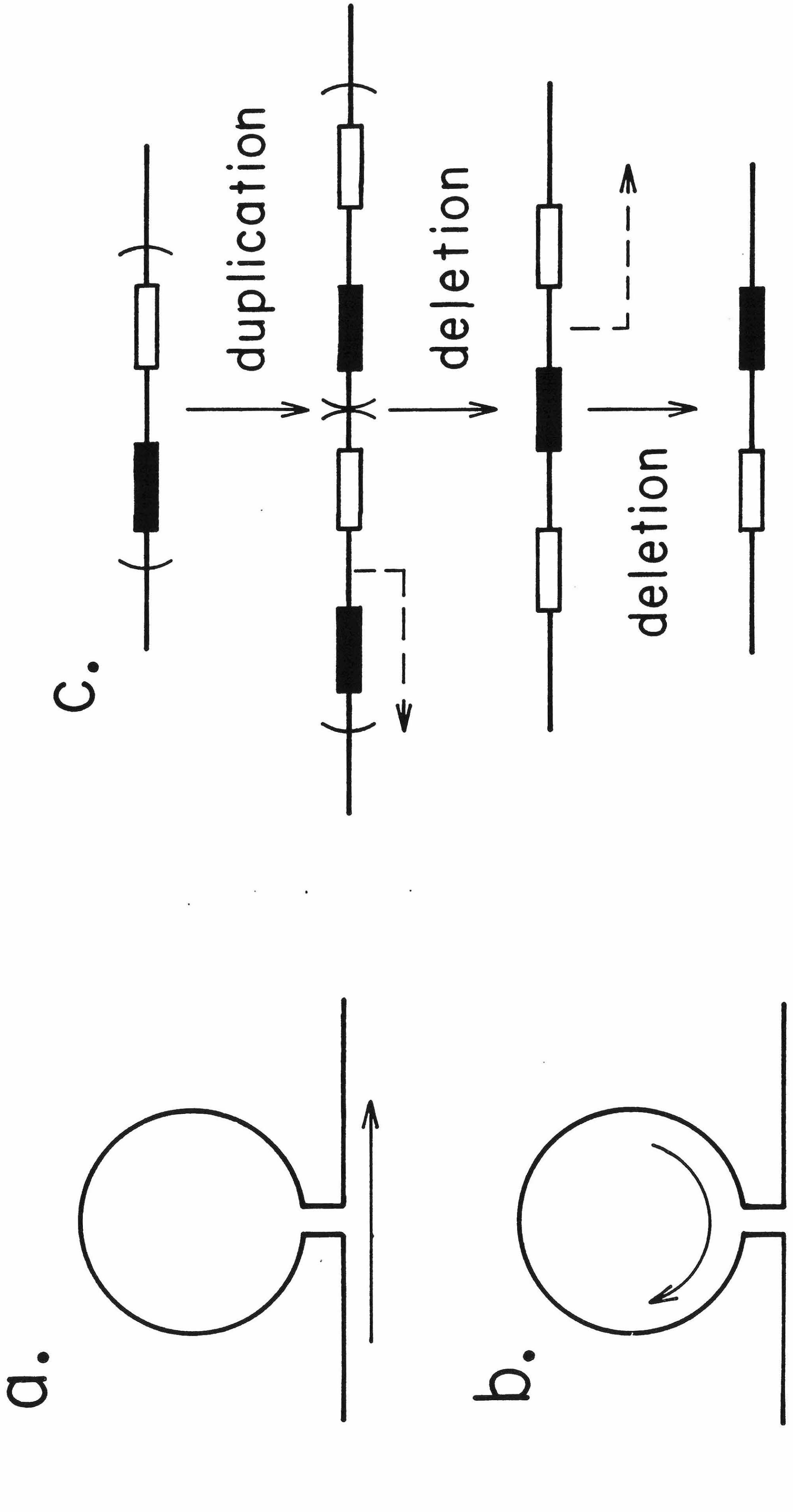
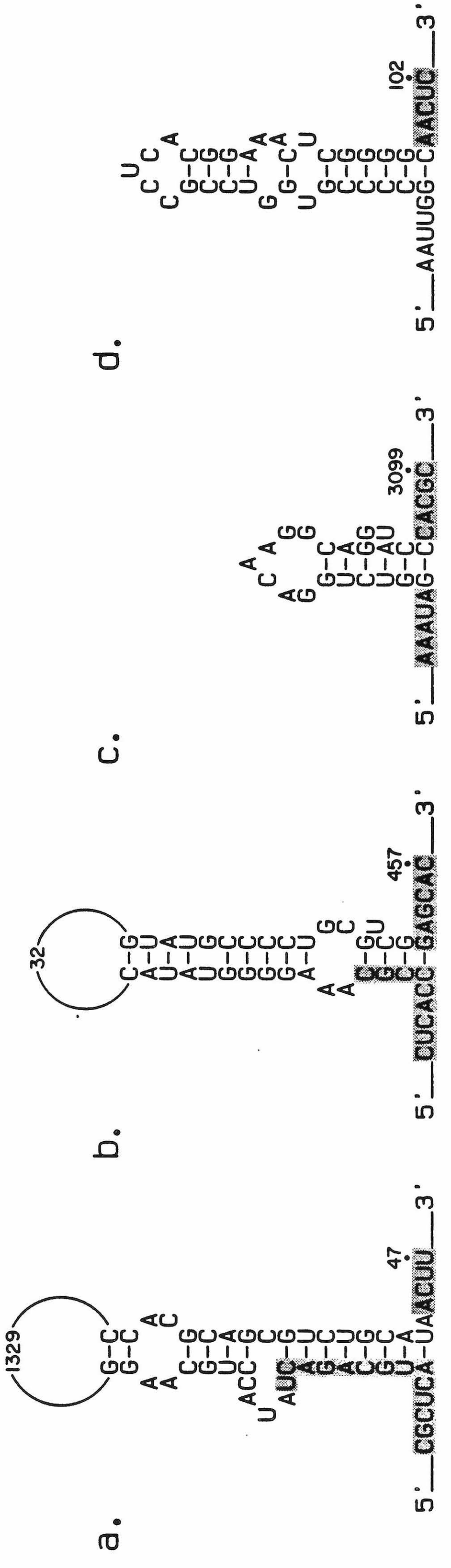

a.
b.
c.
duplication
deletion
deletion

Figure 2. Possible Hairpin Structures Involved in Generating Deletions

Shaded regions indicate the sequences of DI RNAs or the RRV mutant. Numbers give either the size of the hairpin loop or the position of the nucleotide from poly(A) tail. (a) SIN DI RNA (Monroe et al., 1982; also Chapter 4, this thesis). (b) SFV DI RNA (Lehtovaara et al., 1981). Sequence not shaded indicates the 60-nucleotide "insert" of one of the repeated sequences. (c) Sequence of an RRV mutant (see text). (d) SFV DI RNA (Lehtovaara et al., 1981).


a.
1329
47
5'
3'
b.
32
457
5'
3'
c.
3099
5'
3'
d.
102
5'
3'

multiple jump-over events to generate the sequence organization of this DI RNA, however, cannot be ruled out.

Repeated sequences in the 3'-noncoding regions of different genomic RNAs were also examined. Most of the repeats show certain degrees of sequence variation (Ou et al., 1982b; also Chapter 3, this thesis). This suggests that they were generated some time ago and subsequent evolution results in their sequence divergence. The repeats of Middelburg virus (MBV), however, are identical in sequences and the generation of them could be a recent event. Sequence organization of MBV repeats is illustrated in Figure 3A(e): there are two large tandem repeats, and inside each large repeat there is also a set of small tandem repeats. According to the "jumping" model ancestral sequences flanking either small (Figure 3A(a) and (b), circles) or large (Figure 3A(c) and (d), crosses) repeats would be able to base-pair stably. As predicted, stable base-pairings formed by these sequences were found (Figure 3B).

It has been found that sequences preceding repeated sequences in the 3'-noncoding regions of genomic RNAs often resemble the 3'-conserved sequences (Ou et al., 1982b; also Chapter 3, this thesis). It is conceivable that for either jumping or template-switching mechanism, these sequences are the preferential reinitiation sites of replicases.

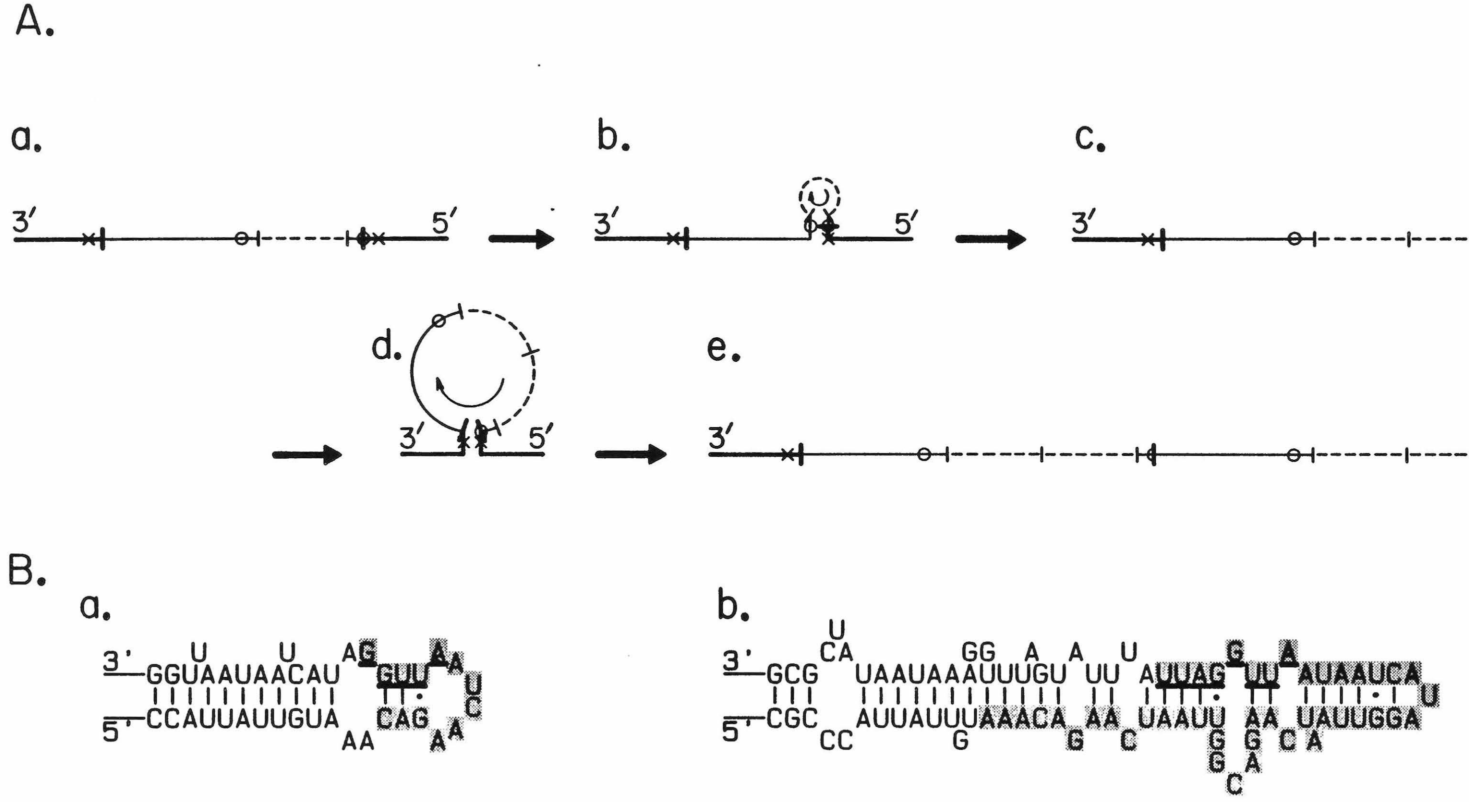
Conclusion

After infecting host cells, alphavirus genomic RNA is released into the cytoplasm and translated into possibly four nonstructural proteins required for RNA synthesis. This genomic 49S RNA then serves as the template for transcribing a minus-stranded RNA which is again the template for transcribing more 49S as well as 26S RNAs. The discovery of three different potential replicase recognition signals gives clues as to how alphaviruses regulate their RNA synthesis.

From complementation experiments, four genes have been found to be required for alphavirus RNA synthesis. The A and B genes are probably initiation factors

Figure 3. The Generation of MBV Repeats

A. Processes for generating MBV repeats. Circles indicate the sequences which could base-pair stably and generate small repeats. Crosses indicate the complementary sequences involved in generating large repeats. B. (a) The hairpin structure of the ancestral sequence as shown in A(b) for generating small repeats. (b) The hairpin structure of the ancestral sequence as shown in A(d) for generating large repeats. Sequences resembling the 3'-highly conserved sequence are underlined.


A.
a.
3′
5′
b.
3′
5′
c.
3′
5′
d.
3′
5′
e.
3′
5′
B.
a.
3'
5'
U U AG A
GGUAAUAACAU GUU A
CCAUUAUUGUA CAG U
AA A A C
b.
3'
5'
U
CA GG A A U G A
GCG UAAUAAAUUUGU UU AUUAG UU AUAAUCA
CGC AUUAUUUAAACA AA UAAUU AA UAUUGGA U
CC G G C G G C A
G A
C

for transcribing 26S and minus-stranded RNAs, respectively. The F gene is required for synthesizing all three RNA species and is probably an elongation factor. The role of G gene is unclear; however, it is possible that it is the initiation factor for transcribing 49S RNA.

Mechanism for generating alphavirus DI RNAs could be very complicated and might involve many factors. It has been suggested that host factors might be involved in alphavirus RNA replication (Kowal and Stollar, 1981) and thus, they might also be involved in DI RNA generation. How host factors participate in alphavirus RNA replication is not clear at this moment and more experimental results would be required to answer this question.

References

Bruton, C. J. and Kennedy, S. I. T. (1975). Semliki Forest virus intracellular RNA: Properties of multistranded RNA species and kinetics of positive and negative strand synthesis. J. Gen. Virol. 28, 111-127.

Brzeski, H. and Kennedy, S. I. T. (1977). Synthesis of Sindbis nonstructural polypeptides in chicken embryo fibroblasts. J. Virol. 22, 420-429.

Frey, T. K ., Gard, D. L. and Strauss, J. H. (1979). Biophysical studies on circle formation by Sindbis virus 49S RNA. J. Mol. Biol. 132, 1-18.

Frey, T. K. and Strauss, J. H. (1978). Replication of Sindbis virus. VI. Poly(A) in virus-specific RNA species. Virology 86, 494-506.

Friedman, R. M., Levin, J. G., Grimley, P. M. and Berezesky, I. K. (1972). Membrane-associated replication complex in arbovirus infection. J. Virol. 10, 504-515.

Fuller, F. J. and Marcus, P. I. (1980). Sindbis virus. I. Gene order of translation in vivo. Virology 107, 441-451.

Garoff, H., Frischauf, A.-M., Simons, K., Lehrach, H. and Delius, H. (1980a). The capsid protein of Semliki Forest virus has clusters of basic amino acids and prolines in its amino terminal region. Proc. Nat. Acad. Sci. USA 77, 6376-6380.

Garoff, H., Frischauf, A.-M., Simons, K., Lehrach, H. and Delius, H. (1980b). Nucleotide sequence of the cDNA coding for the Semliki Forest virus membrane glycoproteins. Nature 288, 236-241.

Grimley, P. M., Levin, J. G., Berezeskey, J. K. and Friedman, R. M. (1972). Specific membranous structures associated with the replication of group A arbovirus. J. Virol. 10, 492-503.

Hsu, M. T., Kung, H. J. and Davidson, N. (1973). An electron microscope study of Sindbis virus RNA. Cold Spring Harbor Symp. Quant. Biol. 38, 943-950.

Hsu-Chen, C. C. and Dubin, D. T. (1976). Di- and trimethylated cogeners of 7-methylguanine in Sindbis virus mRNA. Nature 264, 190-191.

Kennedy, S. I. T. (1976). Sequence relationships between the genome and intracellular RNA species of standard and defective interfering Semliki Forest virus. J. Mol. Biol. 108, 491-511.

Keranen, S. and Kaariainen, L. (1979). Functional defects of RNA-negative temperature-sensitive mutants of Sindbis and Semliki Forest viruses. J. Virol. 32, 19-29.

Kowal, K. J. and Stollar, V. (1981). Temperature-sensitive host-dependent mutants of Sindbis virus. Virology 114, 140-148.

Lachmi, B. E. and Kaariainen, L. (1976). Sequential translation of nonstructural proteins in cells infected with a Semliki Forest virus mutant. Proc. Nat. Acad. Sci. USA 73, 1936-1940.

Lehtovaara, P., Soderlund, H., Keranen, S., Pettersson, R. F. and Kaariainen, L. (1981). 18S defective interfering RNA of Semliki Forest virus contains a triplicated linear repeat. Proc. Nat. Acad. Sci. USA 78, 5353-5357.

Martin, B. A. B. and Burke, D. C. (1974). The replication of Semliki Forest virus. J. Gen. Virol. 24, 45-66.

Martin, J. D., Riggsby, W. S. and Beck, R. W. (1979). The effect of ribonuclease on the replicative forms of Sindbis virus RNA. Archives Virol. 60, 131-146.

Monroe, S. S., Ou, J.-H., Rice, C. M., Schlesinger, S., Strauss, E. G. and Strauss, J. H. (1982). Sequence analysis of cDNAs derived from the RNA of Sindbis virions and of defective interfering particles. J. Virol. 41, 153-162.

Ou, J.-H., Rice, C. M., Dalgarno, L., Strauss, E. G. and Strauss, J. H. (1982a). Sequence studies of several alphavirus genomic RNAs in the region containing the start of the subgenomic RNA. Proc. Nat. Acad. Sci. USA (submitted).

Ou, J.-H., Strauss, E. G. and Strauss, J. H. (1981). Comparative studies of the 3'-terminal sequences of several alphavirus RNAs. Virology 109, 281-289.

Ou, J.-H., Trent, D. W. and Strauss, J. H. (1982b). The 3'-noncoding regions of alphavirus RNAs contain repeating sequences. J. Mol. Biol. (in press).

Pettersson, R. F. (1981). 5-Terminal nucleotide sequence of Semliki Forest virus 18S defective interfering RNA is heterogeneous and different from the genomic 42S RNA. Proc. Nat. Acad. Sci. USA 78, 115-119.

Porterfield, J. S. (1980). In "The Togaviruses" (R. W. Schlesinger, ed.). Antigenic characteristics and classification of Togaviridae. pp. 13-46, Academic Press, New York.

Proudfoot, N. J. and Brownlee, G. G. (1976). 3'-Noncoding region sequences in eucaryotic messenger RNA. Nature 263, 211-214.

Racaniello, V. R. and Baltimore, D. (1981). Molecular cloning of poliovirus cDNA and determination of the complete nucleotide sequence of the viral genome. Proc. Nat. Acad. Sci. USA 78, 4887-4891.

Revel, M. and Groner, Y. (1978). Post-transcriptional and translational controls of gene expression in eucaryoties. Ann. Rev. Biochem. 47, 1079-1126.

Rice, C. M. and Strauss, J. H. (1981). Nucleotide sequence of the 26S mRNA of Sindbis virus and deduced sequence of the encoded virus structural proteins. Proc. Nat. Acad. Sci. USA 78, 2062-2066.

Robertson, J. S., Schubert, M. and Lazzarini, R. A. (1981). Polyadenylation sites for influenza virus mRNA. J. Virol. 38, 157-163.

Sawicki, D. L. and Gomatos, P. J. (1976). Replication of Semliki Forest virus: Polyadenylate in plus-strand RNA and polyuridylate in minus-strand RNA. J. Virol. 20, 446-464.

Sawicki, D. L., Sawicki, S. G., Kaariainen, L. and Keranen, S. (1981a). Specific Sindbis virus-coded function for minus-strand RNA synthesis. J. Virol. 39, 348-358.

Sawicki, D. L. and Sawicki, S. G. (1980). Short-lived minus-strand polymerase for Semliki Forest virus. J. Virol. 34, 108-118.

Sawicki, S. G., Sawicki, D. L., Kaariainen, L. and Keranen, S. (1981b). A Sindbis virus mutant temperature-sensitive in the regulation of minus-strand RNA synthesis. Virology 115, 161-172.

Schubert, M. and Lazzarini, R. A. (1981). In vivo transcription of the 5'-terminal extracistronic region of vesicular stomatitis virus RNA. J. Virol. 38, 256-262.

Segal, S. and Sreevalsan, T. (1974). Sindbis virus replicative intermediates: Purification and characterization. Virology 59, 428-442.

Simmons, D. T. and Strauss, J. H. (1972a). Replication of Sindbis virus. I. Relative size and genetic content of 26S and 49S RNA. J. Mol. Biol. 71, 599-613.

Simmons, D. T. and Strauss, J. H. (1972b). Replication of Sindbis virus. II. Multiple forms of double-stranded RNA isolated from infected cells. J. Mol. Biol. 71, 615-631.

Simmons, D. T. and Strauss, J. H. (1974a). Replication of Sindbis virus. V. Polyribosomes and mRNA in infected cells. J. Virol. 14, 552-559.

Simmons, D. T. and Strauss, J. H. (1974b). Translation of Sindbis virus 26S and 49S RNA in lysates of rabbit reticulocytes. J. Mol. Biol. 86, 397-409.

Soderlund, H., Keranen, S., Lehtovaara, P., Palva, I., Pettersson, R. and Kaariainen, L. (1981). Structural complexity of defective-interfering RNAs of Semliki Forest virus are revealed by analysis of complementary DNA. Nucleic Acids Res. 9, 3403-3417.

Spector, D. H. and Baltimore, D. (1974). Requirement of 3'-terminal poly(adenylic acid) for the infectivity of poliovirus RNA. Proc. Nat. Acad. Sci. USA 71, 2983-2987.

Stollar, V. (1980). In "The Togaviruses" (R. W. Schlesinger, ed.) Defective interfering alphaviruses, pp. 427-457. Academic Press, New York.

Strauss, E. G. and Strauss, J. H. (1980). In "The Togaviruses" (R. W. Schlesinger, ed.). Mutants of alphaviruses: Genetics and physiology, pp. 393-426, Academic Press, New York.

Strauss, J. H. and Strauss, E. G. (1977). In "The Molecular Biology of Animal Viruses" (D. P. Nayak, ed.) Vol. 1, pp. 111-166, Marcel Dekker, New York.

Trent, D. W. and Grant, J. A. (1980). A comparison of new world alphaviruses in the western equine encephalomyelitis complex by immunochemical and oligonucleotide fingerprint techniques. J. Gen. Virol. 47, 261-282.

Wecker, E. (1959). The extraction of infectious virus nucleic acid with hot phenol. Virology 7, 241-243.

Weiss, B. and Schlesinger, S. (1981). Defective interfering particles of Sindbis virus do not interfere with the homologous virus obtained from persistently infected BHK cells but do interfere with Semliki Forest virus. J. Virol. 37, 840-844.

Wengler, G., Wengler, G. and Gross, H. J. (1979). Replicative form of Semliki Forest virus contains an unpaired guanosine. Nature 282, 754-756.

Chapter 2

This chapter was published in Virology.

VIROLOGY 109, 281–289 (1981)

Comparative Studies of the 3′-Terminal Sequences of Several Alphavirus RNAs

JING-HSIUNG OU, ELLEN G. STRAUSS, AND JAMES H. STRAUSS[1]

Division of Biology, California Institute of Technology, Pasadena, California 91125

Accepted September 15, 1980.

The 3′-termini of 26 S RNA of Sindbis virus and of the virion 49 S RNAs of Sindbis virus, Semliki Forest virus, and Middelburg virus have been sequenced by using a chain-terminating sequencing method with reverse transcriptase. Both 49 S and 26 S RNAs of Sindbis virus have identical 3′-terminal sequence for at least 146 nucleotides from the poly(A) tracts. The 26 S RNA is known to be a subgenomic RNA located at or near the 3′-end of 49 S RNA; our present results indicate that 26 S and 49 S RNA are coterminal. The 49 S RNAs of all three viruses contain a region of 20 nucleotides adjacent to the poly(A) which is highly conserved. This conserved sequence could be the replicase recognition site and/or a recognition site for encapsidation. After this conserved region the sequences of Sindbis and of Semliki Forest virus diverge significantly but there does appear to be some residual sequence homology in these regions. The 3′-termini of Sindbis virus and Semliki Forest virus RNAs contain a high content of A and U; of the first 50 nucleotides adjacent to the poly(A) tail over half are U and over one quarter are A. In addition repeating sequences are present in the 3′-termini.

INTRODUCTION

Sindbis virus, Semliki Forest virus (SFV), and Middelburg virus all belong to the Togavirus family (for review see Strauss and Strauss, 1977). They cross-react serologically in complementation-fixation assays (Casals and Clarke, 1965) and their structural proteins have similar amino acid compositions (Bell *et al.*, 1979). Two major species of messenger RNAs, 49 S RNA and 26 S RNA, are found in Sindbis- or SFV-infected cells. The 49 S RNA (4.5×10^6 daltons) is the genomic RNA and serves as mRNA for viral nonstructural proteins. The 26 S RNA (1.6×10^6 daltons) has about one-third the molecular weight of 49 S RNA and encodes the three structural proteins of the virus (Simmons and Strauss, 1972, 1974). Both 49 S and 26 S RNA contain a stretch of polyadenylic acid [poly(A)] at their 3′-end (Frey and Strauss, 1978).

The results of hybridization-competition experiments with Sindbis RNAs showed that 26 S RNA is a subgenomic RNA containing one-third of the base sequences of 49 S RNA (Simmons and Strauss, 1972). Results with SFV RNAs have shown that 26 S RNA is located at or near the 3′-end of 49 S RNA (Kennedy, 1976; Wengler and Wengler, 1976).

In this paper we describe the use of the chain-terminating sequencing method developed by Sanger *et al.* (1977) to sequence the 3′-termini of Sindbis, Semliki Forest, and Middelburg virion RNAs and of Sindbis 26 S RNA.

MATERIALS AND METHODS

Materials and enzymes. Triphosphates, primers, and T4 polynucleotide kinase were purchased from PL Biochemicals, Inc. Bacterial alkaline phosphatase was purchased from Bethesda Research Labs, Inc., and oligo(dT)-cellulose from Collaborative Research, Inc. [^{3}H]Uridine was from New England Nuclear, [α-^{32}P]deoxyribonucleoside triphosphates (400 Ci/mmol) from Amersham Searle, and [γ-^{32}P]adenosine triphosphate (>5000 Ci/mmol) was from ICN. Reverse transcriptase from avian myelo-

[1] To whom reprint requests should be addressed.

0042-6822/81/040281-09$02.00/0

blastosis virus was obtained from the Division of Cancer Cause and Prevention, National Cancer Institute.

Growth of viruses and preparation of virion RNAs. Sindbis virus was grown in primary cultures of chick embryo fibroblasts at 37° in medium containing low salt followed by harvest in high salt, and purified by velocity and isopycnic centrifugation, as previously described (Frey *et al.*, 1979). SFV was grown and harvested in regular salt Eagle's medium (Eagle, 1959) at 37°, precipitated in 8% polyethylene glycol, resuspended in 0.2 M NaCl, 0.001 M EDTA, 0.05 M Tris, pH 7.4 (Pierce *et al.*, 1974), and purified by velocity and isopycnic sedimentation as above. Middelburg virus was grown in regular salt Eagle's medium at 30°, harvested at 24 hr postinfection, and then precipitated and purified as for SFV.

The purified virus was pelleted by centrifugation at 50 K rpm for 3 hr in an SW 50.1 rotor and resuspended in 0.01 M Tris, pH 7.4, containing 1 mM EDTA, 60 mM NaCl, and 1% SDS. Virion RNAs were extracted with phenol/chloroform using the method of Hsu *et al.* (1973) and precipitated with 2.5 vol of ethanol in 0.2 M sodium acetate at −70°.

Purification of Sindbis virus 26 S RNA. Chick monolayers were infected with Sindbis virus at 37° and 5 to 7 hr postinfection the cells were lysed with hypertonic buffer (0.5 M NaCl, 15 mM $MgCl_2$, 5 mM EGTA, 50 mM PIPES, pH 6.5) containing 1% Nonidet P40. The nuclei were removed by centrifugation at 660 *g* for 5 min. SDS was added to the supernatant to a final concentration of 1% and the cytoplasmic lysate extracted immediately with phenol/chloroform. The RNA was precipitated with ethanol, resuspended in 0.2% SDS, heated to 56° for 10 min, and ribosomal RNA removed by two cycles of oligo(dT)-cellulose chromatography. RNA was loaded on an oligo(dT)-cellulose column in 0.2 M NaCl, 0.01 M EDTA, 0.01 M Tris, pH 7.4, 0.2% SDS, and the poly(A)-containing RNA eluted with 0.2% SDS, essentially as previously described (Frey and Strauss, 1978). The RNA preparation was resuspended in 0.01 M NaCl, 1 mM EDTA, 0.01 M Tris, pH 7.4, 0.2% SDS, and sedimented through a 15–30% sucrose gradient in the same buffer at 23° and 40 K rpm in the SW40 rotor for 6.5 hr. The peak of 26 S RNA was pooled, ethanol precipitated, resuspended in sterile H_2O, and kept at −70°.

Purification of primers. As reported earlier (Zimmern and Kaesberg, 1978), the commercially available primers were usually contaminated with other oligonucleotides and it was necessary to purify the primer before use. Primers $p(dT)_7rG$ and $p(dT)_7rA$ were first treated with alkali (0.1 N NaOH) at 56° for 1 hr to ensure that only one ribonucleotide was present on the 3′-end. One microgram of primer thus treated was dephosphorylated with 5 units of alkaline phosphatase in 0.05 M Tris (pH 8.3) for 30 min at 37°, phenol/chloroform extracted, and lyophilized, followed by end-labeling with 100 pmol of [γ-^{32}P]ATP and 5 units of T4 kinase at 37° for 30 min in 0.1 M Tris (pH 7.4) containing 16 mM 2-mercaptoethanol and 10 mM $MgCl_2$. The resulting products were separated on a 0.4-mm-thick, 15% acrylamide gel (acrylamide:bisacrylamide, ratio 20:1). The gel was radioautographed and the major band cut out and eluted with distilled water at 37° with constant shaking. Alkali-treated, nonlabeled primers were purified by electrophoresis on another gel using 1000 cpm of the purified labeled primer as a marker. The radioactive band on the gel, located by autoradiography using preflashed film and an intensifier screen at −70°, was cut out and eluted with H_2O as described above. The labeling and purification of the oligonucleotide $p(dT)_{10}$ was essentially the same, except that the alkali treatment was omitted.

Preparation of the template/primer mixtures. Eight micrograms of virion RNA or 3 μg of Sindbis 26 S RNA were mixed with 0.5 μg of purified primer in a total of 9 μl H_2O and the mixture heated to 56° for 5 min in order to disrupt the secondary structures of the RNAs and to allow better association of primers and templates. Template/primer mixtures were then chilled to 0° and 1 μl of 0.8 M KCl was added to each mixture. Mixtures were stored at −70° for subsequent use.

Determination of the first few nucleotides adjacent to the poly(A). In our preliminary experiments it was found very difficult to determine the first few nucleotides with the chain-terminating sequencing method. This seems to be a common phenomenon (Both and Air, 1979; McGeoch and Turnbull, 1978; Symons, 1979). Therefore, an alternative approach for sequencing the first few nucleotides was necessary. Details of these reactions are described under Results and in the legends of Figs. 1 and 2.

Dideoxy sequencing of the 3′-termini of virus RNAs. The 3′-termini of the viral RNAs were sequenced using the dideoxy technique developed by Sanger *et al.* (1977). Transcription with reverse transcriptase was initiated specifically at the 3′-end by using the primer $p(dT)_7rG$. Specific termination of the cDNA chain was obtained by including one of the four dideoxynucleoside triphosphates in each of four parallel reactions. The products of these reactions were separated by size on acrylamide gels, basically as described by Sanger and Coulson (1978). Details are given in the figure legends.

For the determination of the sequence of the first 15 nucleotides, a deoxy to dideoxy ratio of 1:5 was used. Higher concentrations of deoxynucleotides usually caused nonspecific bands to appear in all four lanes and therefore gave ambiguous results. For the determination of the sequences from 15 to 100 nucleotides, a ratio of 5:1 or 10:1 was used. For sequences more than 100 nucleotides in length, a 10:1 ratio was found to be best. The concentrations of dGTP, dCTP, dTTP were 10 μM and the concentration of [^{32}P]dATP was 1 μM (specific activity = 400 Ci/mmol) in all reactions except those with a ratio of 1:5, where the concentration of all four deoxyribonucleoside triphosphates was 1 μM (and the concentration of the corresponding dideoxyribonucleoside triphosphates was 5 μM).

All reaction mixtures were in 50 mM Tris, pH 8.3, 8 mM $MgCl_2$, 0.4 mM dithiothreitol, 50 mM KCl, in a final volume of 10 μl. Reactions were started by the addition of 1 μl template/primer mix and 2.1 units of reverse transcriptase. All reactions were carried out at 0° for 20 min followed by 1 hr at 37°. One microliter of a solution which was 0.5 mM in each deoxynucleoside triphosphate was then added to each reaction and they were incubated for an additional 20 min at 37° in order to chase nascent chains. Reactions were stopped by adding 5 μl of 10 M urea/dye mix followed by boiling at 100° for 2 min and loaded onto gels. All the sequencing gels used in our experiments were 0.4 mm in thickness (Sanger and Coulson, 1978).

RESULTS

The First Nucleotide Adjacent to the Poly(A) of Sindbis RNAs

The most convenient way to determine the first nucleotide adjacent to the poly(A) at the end of the RNAs is to use [α-^{32}P]-dATP, [α-^{32}P]dGTP, or [α-^{32}P]dCTP individually, as the only triphosphate in the reverse transcriptase assay, using an oligo-(dT) primer. One of these reaction mixtures, that containing the deoxynucleoside triphosphate complementary to the first nucleotide in the RNA, will produce a new electrophoretic band consisting of the primer which has been elongated by one (or more) nucleotides.

It had been previously reported that after complete digestion with RNase T1 which cleaves specifically after guanosine, a stretch of pyrimidines is still associated with the poly(A) of either 49 S RNA or 26 S RNA of Sindbis virus (Frey and Strauss, 1978). Thus, the first non-A nucleotide on the 3′-end of these RNAs is a pyrimidine. With this in mind, we used [α-^{32}P]dATP or [α-^{32}P]dGTP as the only nucleoside triphosphate in the reaction using 49 S RNA as template and found that deoxyguanosine was incorporated into the primer $p(dT)_{10}$ (Fig. 1, lanes 2 and 3), indicating that the first nucleoside in the template is cytosine. This result was confirmed by the finding that $p(dT)_7rG$ was a good primer (see below), whereas $p(dT)_7rA$ was a poor primer (data not shown). Lanes 4 and 5 of Fig. 1 are controls containing both dATP and dGTP; one deoxynucleotide is labeled (as in lanes 2 and 3) but the reaction also contains a 10-fold excess of the other (unlabeled) deoxynucleotide.

The Sequence of the First Few Nucleotides Adjacent to Poly(A) at the 3′-Termini of Sindbis Virus RNAs

As shown in lane 2 of Fig 2 for 26 S RNA and in lane 6 of Fig 1 for 49 S RNA, when [α-^{32}P]dATP was used as the only deoxynucleoside triphosphate in the reverse

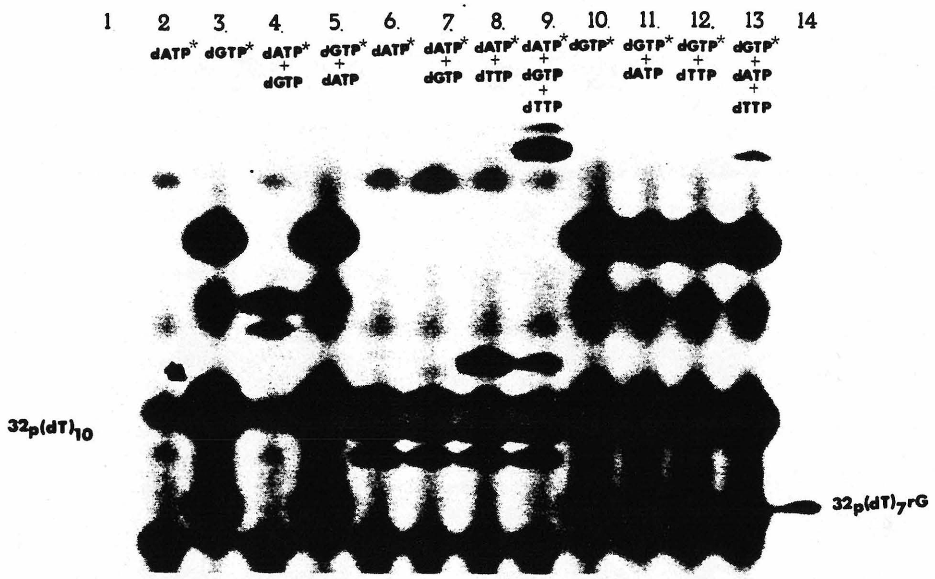


FIG. 1. The determination of the first few nucleotides adjacent to the poly(A) of Sindbis 49 S RNA. All reactions were in a final volume of 10 μl of 50 mM Tris, pH 8.3, 8 mM $MgCl_2$, 0.4 mM dithiothreitol, 50 mM KCl, and contained 1 μl of template/primer mix (prepared as described under Materials and Methods). After adding 2.1 units of reverse transcriptase the mixture was incubated at 0° for 20 min followed by incubation at 37° for another 20 min. Reactions were stopped by adding urea/dye mix followed by boiling for 2 min. Samples were chilled on ice and then loaded on a 0.4-mm-thick 15% acrylamide gel (acrylamide/bisacrylamide = 20/1). Primers used in lanes 2–5 were $p(dT)_{10}$ and in lanes 6–13 $p(dT)_7rG$. Lane 1: $^{32}P(dT)_{10}$. Lane 2: 10 pmol of [α-^{32}P]dATP was used as the only triphosphate in the reaction. Lane 3: 10 pmol of [α-^{32}P]dGTP was used as the only triphosphate. Lane 4: 10 pmol [α-^{32}P]-dATP plus 10 μM dGTP. Lane 5: 10 pmol [α-^{32}P]dGTP plus 10 μM dATP. Lane 6: 10 pmol of [α-^{32}P]-dATP only. Lane 7: 10 pmol [α-^{32}P]dATP plus 10 μM dGTP. Lane 8: 10 pmol [α-^{32}P]dATP plus 10 μM dTTP. Lane 9: 10 pmol [α-^{32}P]dATP plus 10 μM each of dGTP and dTTP. Lane 10: 10 pmol [α-^{32}P]-dGTP only. Lane 11: 10 pmol [α-^{32}P]dGTP plus 10 μM dATP. Lane 12: 10 pmol [α-^{32}P]dGTP plus 10 μM dTTP. Lane 13: 10 pmol [α-^{32}P]dGTP plus 10 μM each of dATP and dTTP. Lane 14: ^{32}P-$(dT)_7rG$. Arrow of lane 3 denotes the specific oligonucleotide synthesized in this reaction.

transcriptase reaction with $p(dT)_7rG$ as the primer, bands one, two, and three nucleotides longer than the primer were observed. This indicates that the first four nucleotides adjacent to the poly(A) in the templates are 3′-C-U-U-U-5′. If the [α-^{32}P]dATP was supplemented with dGTP in the reactions, no other bands appeared (Fig. 1, lane 7, and Fig. 2, lane 3). However, when dTTP was used with [α-^{32}P]dATP an extra band could be seen (Fig. 1, lane 8 and Fig. 2, lane 4). These results indicated that the fifth nucleotide in the template is A. Since the longest oligonucleotide synthesized under these conditions is only four nucleotides longer than the primer, the sixth nucleotide thus must either be C or G in the template. We found it to be C by including dGTP together with [α-^{32}P]dATP and dTTP as triphosphates in the reaction (Fig. 1, lane 9, and Fig. 2, lane 5). A band 12 nucleotides longer than the primer was observed, as determined by comparing its position with those of the bands in lane 9 of Fig. 2. Thus, the sequence of the first six nucleotides adjacent to the poly(A) is poly(A)-C-U-U-U-A-C and the 14th nucleotide is G. This is true for both 49 S and 26 S RNAs.

3′-Terminal Sequences of Sindbis RNAs, SFV 49 S RNA, and Middelburg RNA

The remaining sequence data was obtained using the dideoxy chain termination method of Sanger *et al.* (1977). Figure 3 shows representative gels from the sequencing experiments with Sindbis RNAs, and Fig. 4 gives the sequence derived from

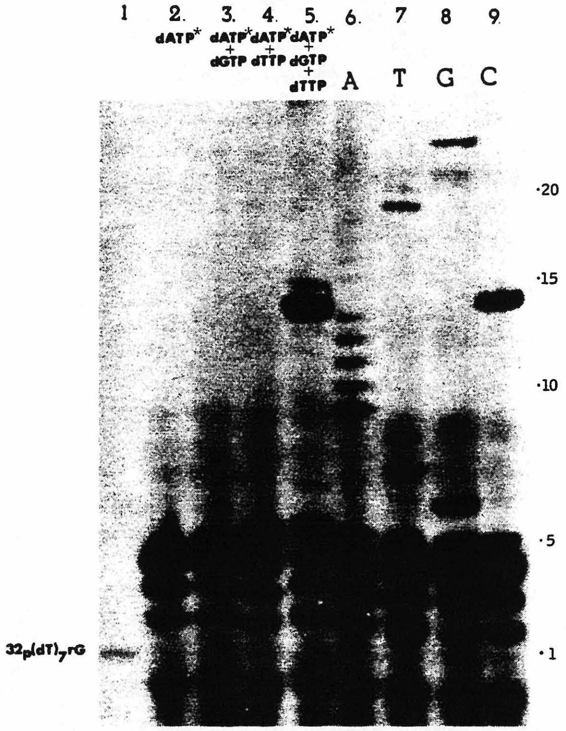


FIG. 2. The determination of the first few nucleotides adjacent to the poly(A) of Sindbis 26 S RNA. Experiments shown in lanes 2–5 were similar in design to those of Fig. 1 (lanes 6–9), with p(dT)$_7$rG as the primer. Lane 1: ^{32}P-(dT)$_7$rG standard. Lane 2: 10 pmol of [α-^{32}P]dATP only. Lane 3: 10 pmol [α-^{32}P]-dATP plus 10 μM dGTP. Lane 4: 10 pmol of [α-^{32}P]dATP plus 10 μM dTTP. Lane 5: 10 pmol of [α-^{32}P]-dATP plus 10 μM each of dGTP and dTTP. The experiments shown in lanes 6–9 are dideoxy sequencing of the 3′-terminus of Sindbis 26 S RNA with a deoxy:dideoxy ration of 1:5, and p(dT)$_7$rG as the primer. The dideoxynucleotide used in the reaction in lane 6 was A, in lane 7 dideoxy T was used, in lane 8 dideoxy G and dideoxy C was used in lane 9. Acrylamide concentrations were the same as for Fig. 1.

it. Only a single sequence is shown since we found the 3′-terminal sequences from 26 S RNA and 49 S RNA to be identical.

In Fig. 5 representative data for SFV 49 S RNA is shown, and in Fig. 6 the sequences of the 3′-termini of SFV, Middelburg and Sindbis are presented together, and have been aligned to maximize possible sequence homology.

Several features of these sequence data should be noted. The first is the presence of anomalies in the polymerase reaction itself. The sequence for the 49 S SFV RNA cannot be determined beyond nucleotide 104 because at that point there is a very strong stop and the reverse transcriptase reaction terminates (see Fig. 5). There is little or no readthrough under our conditions. A similar stop, but not nearly as complete, occurs in the Sindbis 49 S or 26 S RNA near nucleotide 145. There are similar ambiguities at positions 62, 79, 82 of Sindbis RNA, where bands appear in all four lanes of the electrophereogram. We believe these ambiguities are caused by regions of tight secondary structure through which the reverse transcriptase has difficulties reading. Increasing the triphosphate concentration during the transcriptase reaction and increasing the running temperature of the gel help in resolving some, but not all, of these ambiguities. Following the ambiguity at position 62, we note that a hairpin structure can be formed, since nucleotides 63 to 68 are complementary to nucleotides 78 to 83, and of the six pairs of hydrogen bonds which can be formed, five are GC.

Another type of anomaly is presented by Middelburg virus RNA after nucleotide 21. Following this, bands appear in all four lanes

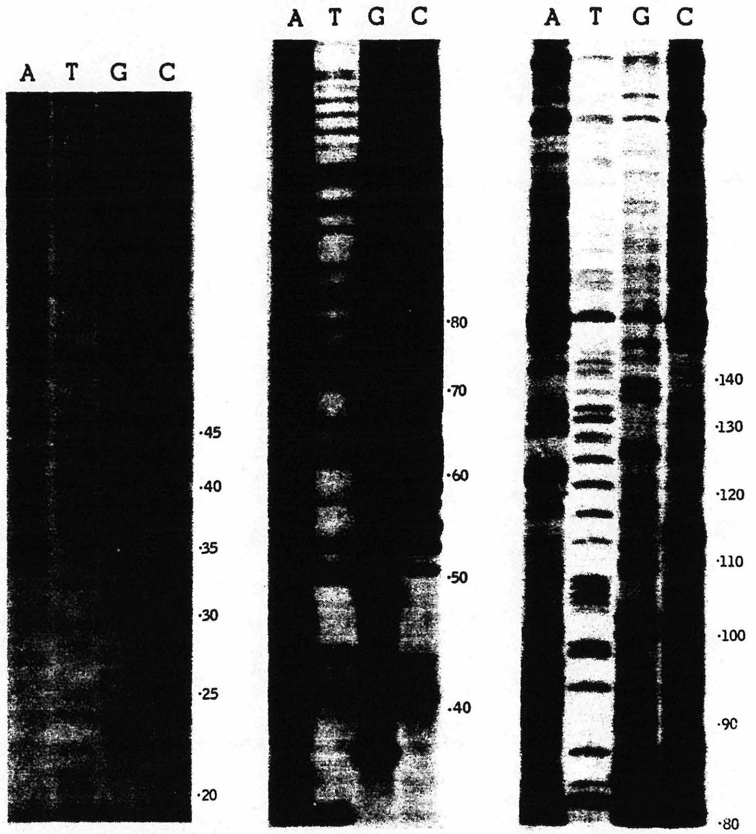


FIG. 3. Dideoxy sequencing of Sindbis virus RNAs. The concentrations of dGTP, dCTP, and dTTP were 10 μM in all reactions, the concentration of [α-^{32}P]dATP was 1 μM (specific activity = 400 Ci/mmol), and the deoxy to dideoxy ratio was 10:1. The primer used was p(dT)$_7$rG. The acrylamide gels used were 8% acrylamide (acrylamide/bisacrylamide = 20/1) and 0.4 mm in thickness. Both 49 S RNA and 26 S RNA showed the same results and only part of the data is shown. The left set of results was obtained with 26 S RNA and the two sets on the right with 49 S RNA. The dideoxy nucleotide in each reaction is identified above the appropriate lane.

of the gel for about 30 nucleotides. The reasons for this are unclear, but it could be also related to the secondary structure of the RNA. Attempts to modify the reaction conditions so as to obtain the sequence of this region have so far been unsuccesful. We are continuing to study this problem.

There are three sets of repeating sequences at the 3′-end of Sindbis virus mRNAs, located at positions 29–37 and 41–49, 36–42 and 90–96, and positions 50–56 and 71–77 (Fig. 4). Repeating sequences are also present in SFV RNA: U-A-A-U-U-G-G and U-U-U-U-A repeat three times and A-A-U-U-G-G repeats four times. Repeating sequences also have been observed in the 3′-termini of other mRNAs, such as chicken ovalbumin mRNA (McReynolds *et al.*, 1978) and N protein mRNA of vesicular stomatitis virus (McGeoch and Turnbull, 1978). The significance of these repeating sequences in alphavirus RNAs is not known.

Seven nonsense codons are present, representing all three reading frames in the sequenced regions of both Sindbis and SFV RNA. These are indicated with numbered brackets in Fig. 4 for Sindbis and are listed in the figure legend to Fig. 6, for SFV.

DISCUSSION

Two major species of virus specific mRNA are produced in cells infected with

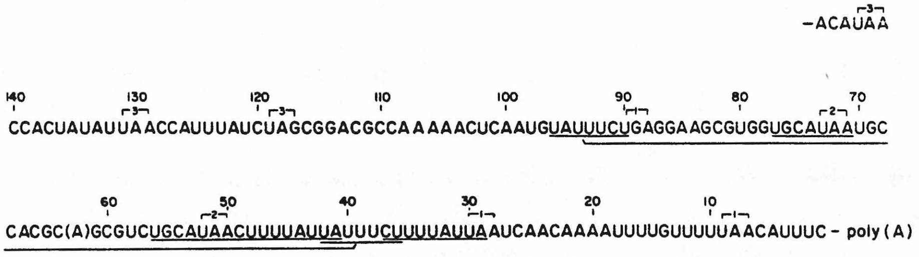

-ACAUAA
3
140
130
120
110
100
90
80
70
CCACUAUAUUAACCAUUUAUCUAGCGGACGCCAAAACUCAAUGUAUUUCUGAGGAAGCGUGGUGCAUAAUGC
60
50
40
30
20
10
CACGC(A)GCGUCUGCAUAACUUUUAUUAUUUCUUUUAUUAAUCAACAAAAUUUGUUUUUAACAUUUC - poly(A)
1
2

FIG. 4. 3'-Terminal sequences of Sindbis virus RNAs. Both 49 S RNA and 26 S RNA gave the same results and the sequence therefore is for both RNAs. Repeating sequences are underlined. Numbers above termination codons denote their reading frames.

the alphaviruses: 49 S RNA, which is identical to the virion RNA, and 26 S RNA (for review see Strauss and Strauss, 1977). The former serves as messenger for the nonstructural proteins of the virus, the latter for the structural proteins. The 26 S RNA is known to be a subgenomic RNA which contains base sequences of the 3' one-third of 49 S RNA. The results shown here, that the 3'-terminal sequences of 26 S and 49 S RNA are identical, clearly demonstrate that the two RNAs are coterminal.

The RNAs of the three viruses examined contain a highly conserved sequence of about 20 nucleotides adjacent to the poly(A) (Fig. 6). There are but three base changes between positions 1 and 20 of Sindbis and positions 2 to 20 of SFV (counting a deletion as a single change). Positions 1 to 19 of SFV are identical to positions 1 to 17 of Middelburg with the deletion of two U's in the latter. This conservation implies an important function for this region. Two obvious possibilities are that it contains sequences involved in RNA replication or in packaging of the RNA into nucleocapsids.

After this conserved region, the sequences of Sindbis and of SFV diverge significantly, although there does appear to be significant homology. The evolutionary constraints on this region, which appears to be noncoding (see below), are thus not as marked. An attempt has been made in Fig. 6 to align the sequences on the basis of apparent homology and aligning nonsense codons and stops. Although the alignment is somewhat arbitrary, there appear to be numerous deletions in the SFV sequence when compared to the Sindbis sequence.

The 3'-terminal sequences of the three alphavirus RNAs examined show a striking feature in that of the first nucleotides after the poly(A) 56% are U and 30% are A in the case of Sindbis (A + U = 86%), while 54% are U and 26% are A in SFV RNA (A + U = 80%). A + U rich regions were also observed at the 3'-termini of the N protein mRNA of vesicular stomatitis virus (McGeoch *et al.*, 1980), cow pea mosaic virus RNAs (Davies *et al.*, 1979), picornavirus

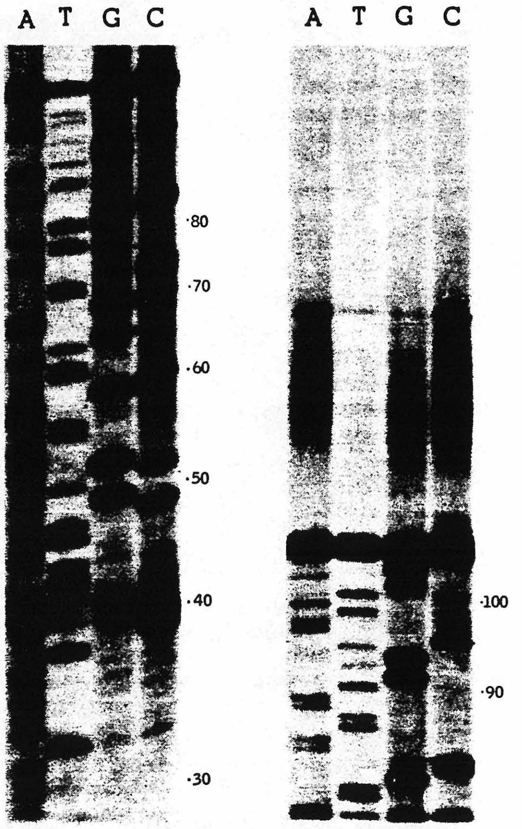


FIG. 5. Dideoxy-sequencing of Semliki Forest virus 49 S RNA. Methods are as in Fig. 3.

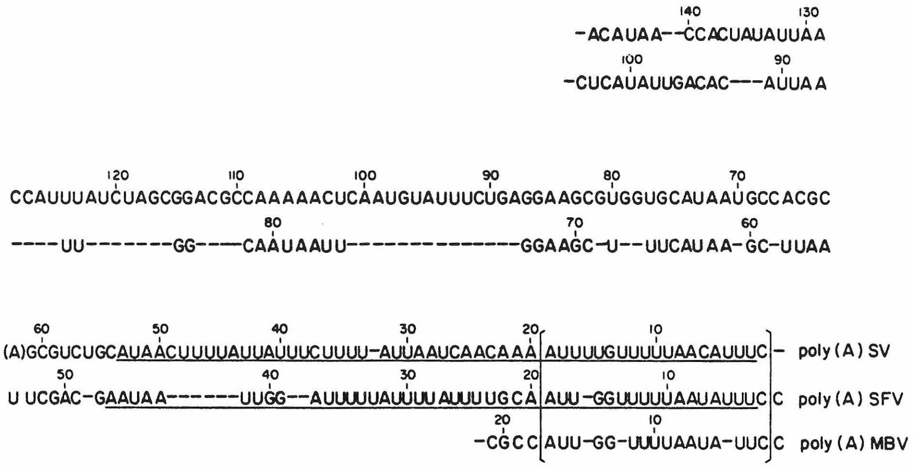

140 130
-ACAUAA--CCACUAUAUUAA
100 90
-CUCAUAUUGACAC---AUUAA
120 110 100 90 80 70
CCAUUUAUCUAGCGGACGCCAAAAACUCAAUGUAUUUCUGAGGAAGCGUGGUGCAUAAUGCCACGC
80 70 60
----UU-------GG----CAAUAAUU---------------GGAAGC-U--UUCAUAA-GC-UUAA
60 50 40 30 20 10
(A)GCGUCUGCAUAACUUUUAUUAUUUCUUUU-AUUAAUCAACAAA AUUUUGUUUUUAACAUUUC - poly (A) SV
50 40 30 20 10
U UCGAC-GAAUAA-------UUGG--AUUUUUAUUUUAUUUUGCA AUU-GGUUUUUAAUAUUUC C poly (A) SFV
20 10
-CGCC AUU-GG-UUUUAAUA-UUC C poly (A) MBV

FIG. 6. Comparison of the 3′-terminal sequences of alphavirus RNAs. RNA sequences are aligned to maximize sequence homology. The most highly conserved sequences at the 3′ end of the RNA are bracketed. The very high A + U regions of Sindbis and SFV RNAs are underlined. Note that termination codons are present in SFV RNA at the following locations: 43–45, 55–57, and 61–63 (frame 1), 87–89 (frame 2), and 8–10, 77–79, and 95–97 (frame 3). Termination codons in Sindbis RNA are noted in Fig. 4.

(Porter and Fellner, 1978), and chicken ovalbumin mRNA (McReynolds *et al.*, 1978), although their A + U contents are not as high as in alphavirus RNAs. In the N protein mRNA of vesicular stomatitis virus the A + U rich region was thought to contain signals for termination of transcription and polyadenylation of the mRNA (McGeoch and Turnbull, 1978). In the case of alphaviruses the poly(A) tract appears to be transcribed from poly(U) in the minus strand (Frey and Strauss, 1978; Sawicki and Gomatos, 1976). However, the poly(A) of poliovirus (Spector and Baltimore, 1974), and Sindbis virus (Frey and Strauss, 1978), if removed is regenerated and thus signals for poly(A) addition may be present. The sequence A–A–U–A–A–A, which was thought to be common to the 3′-termini of eucaryotic mRNAs (Proudfoot and Brownlee, 1976), is not present in the mRNAs of Sindbis or SFV. It is also noteworthy that nucleotides 14 through 54 of Sindbis RNA comprise a single T1 oligonucleotide of 41 residues.

The region from the poly(A) to nucleotides 146 (Sindbis) or 104 (SFV) contains numerous nonsense codons present in all three reading frames, implying that this region is not translated. In addition, C. Rice in our laboratory has sequenced most of the 26 S RNA region encoding E1, the most distal protein known to be encoded in this RNA. His data indicate that the termination codon for E1 is about 320 nucleotides removed from the poly(A). Thus, this entire region appears to be devoted to control functions.

ACKNOWLEDGMENTS

We are grateful to E. M. Lenches for expert technical assistance, to T. E. Crowley and C. M. Rice for help in the preparation of Sindbis 26 S RNA. This work was supported by Grant PCM 77-26728 from the National Science Foundation, by Grants GM 06965 and AI-10793 from NIH, and by Biomedical Research Support Grant 507 RR 07003 from NIH. J.O. was supported by fellowships from the California Foundation for Biochemical Research and from the Li Ming Scholarship Endowment Fund.

REFERENCES

BELL, J. R., STRAUSS, E. G., and STRAUSS, J. H. (1979). Purification and amino acid composition of the structural proteins of Sindbis virus. *Virology* **97**, 287–294.

BOTH, G. W., and AIR, G. M. (1979). Nucleotide sequence coding for the N-terminal region of the matrix protein of Influenza virus. *Eur. J. Biochem.* **96**, 363–372.

CASALS, J., and CLARKE, D. H. (1965). Arboviruses; Group A. *In* "Viral and Rickettsial Infections of Man" (F. L. Horsfall and I. Tamm, ed.), 4th ed., Chap. 26, pp. 583–605. Lippincott, Philadelphia.

DAVIES, J. W., STANLEY, J., and VAN KAMMEN, A. (1979). Sequence homology adjacent to the 3' terminal poly(A) of cow pea mosaic virus RNAs. *Nucleic Acids Res.* **7**, 493–500.

EAGLE, H. (1959). Amino acid metabolism in mammalian cell cultures. *Science* **130**, 432–437.

FREY, T. K., GARD, D. L., and STRAUSS, J. H. (1979). Biophysical studies on circle formation by Sindbis virus 49 S RNA. *J. Mol. Biol.* **132**, 1–18.

FREY, T. K., and STRAUSS, J. H. (1978). Replication of Sindbis virus IV. Poly(A) and poly(U) in virus-specific RNA species. *Virology* **86**, 494–506.

HSU, M. T., KUNG, H. J., and DAVIDSON, N. (1973). An electron microscope study of Sindbis virus RNA. *Cold Spring Harbor Symp. Quant. Biol.* **38**, 943–950.

KENNEDY, S. I. T. (1976). Sequence relationships between the genome and the intracellular RNA species of standard and defective-interfering Semliki Forest virus. *J. Mol. Biol.* **108**, 491–511.

MCGEOCH, D. J., and TURNBULL, N. T. (1978). Analysis of the 3'-terminal nucleotide sequence of vesicular stomatitis virus N protein mRNA. *Nucleic Acids Res.* **5**, 4007–4024.

MCGEOCH, D. J., DOLAN, A., and PRINGLE, C. R. (1980). Comparison of nucleotide sequence in the genomes of the New Jersey and Indiana serotypes of vesicular stomatitis virus. *J. Virol.* **33**, 69–77.

MCREYNOLDS, L., O'MALLEY, B. W., NISBET, A. D., FOTHERGILL, J. E., GIVOL, D., FIELDS, S., ROBERTSON, M., and BROWNLEE, G. G. (1978). Sequence of chicken ovalbumin mRNA. *Nature (London)* **273**, 723–728.

PIERCE, J. S., STRAUSS, E. G., and STRAUSS, J. H. (1974). Effect of ionic strength on the binding of Sindbis virus to chick cells. *J. Virol.* **13**, 1030–1036.

PORTER, A. G., and FELLNER, P. (1978). 3'-terminal nucleotide sequences in the genome RNA of picornavirus. *Nature (London)* **276**, 298–300.

PROUDFOOT, N. J., and BROWNLEE, G. G. (1976). 3'-noncoding region sequences in eukaryotic messenger RNA. *Nature (London)* **263**, 211–214.

SANGER, F., and COULSON, A. R. (1978). The use of thin acrylamide gels for DNA sequencing. *FEBS Lett.* **87**, 107–110.

SANGER, F., NICKLEN, S., and COULSON, R. (1977). DNA sequencing with chain-terminating inhibitors. *Proc. Nat. Acad. Sci. USA* **74**, 5463–5467.

SAWICKI, D. L., and GOMATOS, P. J. (1976). Replication of Semliki Forest virus: Polyadenylate in plus-strand RNA and polyuridylate in minus-strand RNA. *J. Virol.* **20**, 446–464.

SIMMONS, D. T., and STRAUSS, J. H. (1972). Replication of Sindbis virus. I. Relative size and genetic content of 26 S and 49 S RNA. *J. Mol. Biol.* **71**, 599–613.

SIMMONS, D. T., and STRAUSS, J. H. (1974). Translation of Sindbis virus 26 S RNA and 49 S RNA in lysates of rabbit reticulocytes. *J. Mol. Biol.* **86**, 397–409.

SPECTOR, D. H., and BALTIMORE, D. (1974). Requirement of 3'-terminal poly(adenylic acid) for the infectivity of poliovirus RNA. *Proc. Nat. Acad. Sci. USA* **71**, 2983–2987.

STRAUSS, J. H., and STRAUSS, E. G. (1977). *In* "The Molecular Biology of Animal Viruses" (D. P. Nayek, ed.), Vol. 1, pp. 111–166. Marcel Dekker, New York.

SYMONS, R. H. (1979). Extensive sequence homology at the 3'-termini of the four RNAs of cucumber mosaic virus. *Nucleic Acids Res.* **7**, 825–837.

WENGLER, G., and WENGLER, A. (1976). Localization of the 26 S RNA sequence in the viral genome type 42 S RNA isolated from SFV-infected cell. *Virology* **73**, 190–199.

ZIMMERN, D., and KAESBERG, P. (1978). 3'-terminal nucleotide sequence of encephalomyocarditis virus RNA determined by reverse transcriptase and chain-terminating inhibitors. *Proc. Nat. Acad. Sci. USA* **75**, 4257–4261.

Chapter 3

This chapter is in press in the _Journal of Molecular Biology_.

The 3'-Noncoding Regions of Alphavirus RNAs Contain Repeating Sequences

Jing-hsiung Ou,[1] Dennis W. Trent[2] and James H. Strauss[1]

[1]Division of Biology, California Institute of Technology
Pasadena, CA 91125 USA

[2]Centers for Disease Control, Center for Infectious Diseases, Vector-Borne Diseases Division
Fort Collins, CO 80522

Running title: 3'-Noncoding Sequence of Alphavirus RNA

Abstract

We have compared the 3' terminal noncoding sequences of the RNAs from 10 alphaviruses and found this region to be composed of distinct domains in terms of base composition, degree of sequence conservation, and sequence organization. The first 50-60 nucleotides adjacent to the 3'-terminal poly(A) tract are extremely AU rich (up to 90% A+U). Of these, the first 19 nucleotides are highly conserved, and we postulate that this conserved sequence serves as a replicase recognition signal. For strains of Venezuelan, Western, and Eastern equine encephalitis viruses, Highlands J virus and Sindbis virus, only the 6th nucleotide of this sequence shows any variation. This conserved region is slightly more variable for Semliki Forest virus and Middelburg virus. The remainder of the AU rich region shows only limited homology among viruses and may contain signals for polyadenylation. Upstream from the AU rich domain, between 60 and 300 nucleotides from the poly(A) tract, there are repeated sequences in each viral RNA. These repeats are up to 60 nucleotides in length and can be either tandemly or nontandemly arranged. The repeated sequences show considerable conservation among closely related viruses, in contrast to the nonrepeated sequences in this region which contain little homology.

1. Introduction

The alphavirus group of the family Togaviridae is comprised of viruses which infect both invertebrates and vertebrates, and many of them are pathogenic for man (Strauss & Strauss, 1977). The genome of the alphaviruses consists of a single-stranded RNA which has a molecular weight of about $4.2\text{x}10^6$ daltons. After infection of the host cell, this RNA serves as the messenger for nonstructural proteins involved in viral RNA replication (Kennedy, 1980). A minus-stranded RNA is then transcribed which serves as the template for production of genomic RNA as well as a subgenomic RNA called 26S RNA. The latter is the messenger for the virus structural proteins.

We have reported previously the sequence of the 150 nucleotides at the 3' terminus of both the genomic and subgenomic RNAs of Sindbis virus (HR strain) and showed that they are coterminal (Ou et al., 1981). We have also reported the sequence of the 3'-terminal 100 nucleotides of Semliki Forest virus (SFV) and the 3'-terminal 20 nucleotides of Middelburg virus genomic RNAs (Ou et al., 1981). In this paper we extend these observations to the genomic RNAs of several additional alphaviruses, including two strains of Eastern equine encephalitis virus (EEE), Western equine encephalitis virus (WEE), Highlands J virus, and two strains of Venezuelan equine encephalitis virus (VEE). Improved techniques have allowed us to determine up to 350 nucleotides of 3'-terminal sequence for WEE and more than 250 nucleotides for EEE and Highlands J viruses.

2. Materials and Methods

The procedures for determining the 3'-termini of alphavirus RNAs were the same as described (Ou et al., 1981) with several modifications. First, concentrations of nucleoside triphosphates were raised five-fold. These higher concentrations greatly reduced the number of non-specific bands and enabled us to determine longer sequences. Second, 80 cm long gels with 6% acrylamide were used to determine sequences longer than 30 nucleotides (Smith & Calvo, 1980). (Sequences shorter than 30 nucleotides were determined on a 40 cm, 15% gel as before.) Third, the concentration of both the electrode and gel buffers were doubled. This enabled us to run gels at higher temperature and reduce the compression of bands (Maxam & Gilbert, 1980).

Semliki Forest virus (SFV) and Middelburg virus were grown and purified as previously described (Ou et al., 1981). Sindbis virus wild-type strain was isolated in the same way as Semliki Forest Virus (Ou et al., 1981). Preparation of the other viruses was according to the methods of Trent (Trent et al., 1979; Trent & Grant, 1980). Isolation of the virion RNAs was as previously described (Ou et al., 1981).

3. Results

The 3'-terminal sequences of the genomic RNAs of the various alphaviruses were obtained by a chain terminating method (Sanger et al., 1977), using reverse transcriptase (provided by the Division of Cancer Cause and Prevention, National Cancer Institute) with $(dT)_7rG$ as a primer (Ou et al., 1981). This primer is inefficient in comparison to poly(dT), but gives a specific start point, and sufficient radioactivity can be incorporated using α-^{32}P-ATP to develop the sequence ladder in a reasonable time. The sequences obtained are shown in Fig. 1, and have been confirmed at least once in independent sequencing experiments. All sequences in this and other figures are shown as the RNA sequence and given from 3' to 5'. In all cases numbering of the nucleotides is also 3' to 5' and begins with the first nucleotide adjacent to the poly(A) tract. From 130 to 350 nucleotides of sequence have been obtained depending upon the virus. Reverse transcription is template dependent in our experiments, and a good template like WEE RNA gave readable sequence data for more than 350 nucleotides. For VEE RNA the sequence ladder became faint after 130 nucleotides. We presume the differences have to do with secondary structure in the template. As an example of this, nucleotides 105-141 of SFV RNA are able to form a stable

Figure 1. The 3'-Terminal Sequences of Alphaviruses Genomic RNAs. The sequences are read from left to right with a polarity of 3' to 5'. The first nucleotide shown is immediately adjacent to the poly(A) tract. The Sindbis HR and SFV sequences are from Ou *et al.* (1981), Rice & Strauss (1981a)and Garoff *et al.* (1980). The letter N represents any of the four possible nucleotides, P represents either purine, and S represents either G or C. EEE NA: Eastern equine encephalitis virus, North America strain; EEE SA: Eastern equine encephalitis virus, South America strain; SIN HR: Sindbis virus, HR strain; SIN WT: Sindbis virus wild-type (AR339 strain); WEE: Western equine encephalitis virus; HJ: Highlands J virus; SFV: Semliki Forest virus; MBV: Middelburg virus; VEE TC: Venezuelan equine encephalitis virus, TC-83 strain; VEE TD: Venezuelan equine encephalitis virus, Trinidad donkey strain.

```
                                  |                        |                        |                        |
EEE NA    CUUUAUAAUUUUUGUUUUAUUUUUGUAUUUUCUUUUUAAUUAUACUAAAACAUUUAACUAUAUUUUCCGUCGUAAUACGCGACGGACAUCACAAAUAUUC   100
          UCCGCCAUAUAAUGGACAUCACAUUAACAUAUACGCCGUUAAUACGGACUAAUAUUUCACACAUAUUUUAGUGAPUUCUCGUCGUAAUACGUGACGGGAA   200
          UCACAAUAUUACACCAUUAUAUCGUCGGAAUACGUGACGUAGCCGCCGCAAUACGUGGUGGCNGUGGUAUCGAGCCUUCAUGUAGCUUGGUCCCACUUAA   300
                                  |                        |                        |                        |
EEE SA    CUUUAUAAUUUUUGUUUUAUUUUUGUAUUUUCUUUUUGACAUAUACUAAAAUUUUCUAAUAUUUUCCGUCAUAAUACGUGACGGUCAUCACAAUAUUUUC   100
          CGUCGUAAUCGUGACGGUCAUCACAAACAUUUUCCGUCAUAAUACGUGACGGUCAUCACAAUAAUCACCAUCAUUCGAUGGACAUCACACUAAUAUAUCC   200
          GGUUAUCCGAUUGACAAGCGUCGUAAUACGUGACGCGUCCACCGUAAUACGUGGUGGAASUGGCGACGAGCCUUCAUGUAGCUUGUCGCUAUUAUGUUAU   300
          UACAACCAUCGUAAAUGAAAUCACAAACACCC
                                  |                        |                        |                        |
SIN HR    CUUUACAAUUUUUGUUUUAAAACAACUAAUUAUUUUCUUUAUUAUUUUCAAUACGUCUGCGACGCACCGUAAUACGUGGUGCGAAGGAGUCUUUAUGUAA   100
          CUCAAAAACCGCAGGCGAUCUAUUUACCAAUUAUAUCACCAAUACACCGUUGUGNCGCGUNUAAUACGUGACGCGAAGGAGCCUUCAUGUAGCUCAAAAA   200
          UCACAACGAUAUAACGGGCGCCAUUCGCUAGAUUACAUGGUCGGACUACGUAAUACGUGUAGUCAAGGAGCCUUCAUGUAGCUCAAAACGACCAGCCUAG   300
          UAACCCCGCAUCGCCAGUAGAAGCACACGAUCAGUCGUAGUACGACGUUC
                                  |                        |                        |                        |
SIN WT    CUUUACAAUUUUUGUUUUAAAACAACUAAUUAUUUUCUUUAUUAUUUUCAAUACGUCUGCGACGCACCGUAAUACGUGGUGCGAAGGAGUCUUUAUGUAA   100
          CUCAAAAACCGCAGGCGAUCUAUUUACCAAUUAUAUCACCAAUACACCGUUGUGACGCGUCNAAUACGUGACGCGAAGGAGCCUUCAUGUAGCUCAAAAA   200
          UCACAACGAUAUAACGGGCGCCAUUC
                                  |                        |                        |                        |
WEE       CUUUAAAAUUUUUGUUUUAUUUUUGUUUUCUUUUUAAUAUUUUCUUAAAUAUUUCACACAUAUCACAAAAAUUCAUUUUAUACACAUCACAUAUAAUCUC   100
          UGGGUAUACUCAGAGUAUUUUCACUAUAUCACAACAAUUAUACAUCACACUAUAUGCUAAUUACACUGUGACGUGUCGUAAUACGUGACACGAAGGAGCC   200
          UUUAUGUAGCUCAAAAUUACAGUUCGCCGCGCACCGUAAUACGUGGUGCGAAGGAGCCUUCAUGUAGCUCAAAAUGGCGAUACAGUCACCGGCGCGAGUC   300
          AGUAGAUGCACACAAAUAUUCGUAUCUCGACGUCUGGUUGUGAUAUUCAGG
                                  |                        |                        |                        |
HJ        CUUUAAAAUUUUUGUUUUAUUUUCUUUUUAUAUCAUUAGACUAGUUUUAAAAUUUUCAAAAAUAUUCACACAUCACACAAACUUAUUUUAAGUACAUCAC   100
          AACUAAUAUCACCUUGGGUGAUCUAUGCCAUUGCAUAUGAUCAUUUUAAGGACAUCACAAUGUACAAAUCUCAAUAUUACACAUCACUAUUGUAUUGUAU   200
          ACCUAUUAUUAUUAACCAAUUUACGCUGUGACGUGUCGUAAUACGUGACACGAAGGAGCCUUCAUGUAGCUCAAAGCC
                                  |                        |                        |                        |
SFV       CCUUUAUAAUUUUUGGUUAACGUUUUAUUUUAUUUUUAGGUUAAUAAGCAGCUUAAUUCGAAUACAUUCGAAGGUUAAUAACGGUUAAUUACACAGUUAU   100
          ACUCAACGGGGCUCAAAGGCACUCCGCCUGGUGCCCCGGUUAACGCGUCCAGAACAACGCGAAACAAUGUUCAAUAUGUCAAUACCAAUAUACGGUAACG   200
          AAUGGGAUUGAAAAAGACAAAAGUUAAAAGAACGAUAUAGUUACGGUAACGGAUGGGAUUGAAUAGACGCCUCGGGUUACGUUCACUGGUGUUGGUCGUG   300
          GUCCUAUCGCGGCUAACGCUUCCGGGGGUCUGGUGGGCUCUAAAAGACGU
                                  |                        |                        |                        |
MBV       CCUUAUAAUUUUGGUUACCGCUGAUCGUUAGCAGCGCUAUAAUAAAGGUUUAGUAUUUAUUAGGUUAAUAAUCAUAGGUUAAUCAAGACGGUUAAUCAAG   100
          ACAAAUUAGGUUANUAAUCAUAGGUUANUCAANACGGUUANUCAAGACAAAUGUUAUUACCCGCCG
                                  |                        |                        |                        |
VEE TC    CUUUAUAAUUUUUGUUUUAGGCUAAGCCUUUUCUUUUCUUUUUAUUUUAUUUUUAAAAUUCCGCCGUACGGUUAGCGGCGCUCAAGAUACAUUCGUCGAA   100
          CGGUUAACGACGACAUAAGUUANUACAAAGACC
                                  |                        |                        |                        |
VEE TD    CUUUAUAAUUUUUGUUUUAGGCUAAGCCUUUUCUUUUCUUUUUUAUUUUAUUUUUAAAAUUCCGCCGUACGGUUAGCGGCGCUCAAGAUACAUUCGUCGA   100
          ACGGUUAACGACGACAUAAGUUANUACAAAGACC
```

hairpin structure and reverse transcription through this region is very inefficient. Only sequences that could be read with confidence are reported, and any ambiguities in the sequence ladder are given in Fig. 1 using the appropriate ambiguous symbols.

The North American and South American isolates of EEE, which are very closely related, are shown first, in Fig. 1, and these RNAs begin to diverge at nucleotide 38. Thereafter there is considerable divergence in sequence with homologous regions interspersed (also see below). The next two viruses are the HR and wild-type strains of Sindbis, which were separated in the laboratory some 15 years ago (Burge & Pfefferkorn, 1966). These RNAs are identical over the region sequenced (226 nucleotides for wild-type). The next two viruses are the closely related WEE and Highlands J virus (Trent & Grant, 1980). These viruses diverge at nucleotide 24.

The data for SFV consist of 104 nucleotides obtained by dideoxy sequencing together with the data of Garoff et al. (1980) obtained by chemical sequencing of cloned cDNA. Our data agree with those of Garoff et al. (1980) except for a single nucleotide change at position 66. The SFV sequence and that of Middelburg virus RNA, which has been extended to 167 nucleotides, are quite distinct from the other viruses except for the conserved 3'-terminal sequence described below.

Finally, 133 nucleotides have been sequenced in the virulent Trinidad donkey strain of VEE and the avirulent vaccine strain TC83 derived from it (Trent et al., 1979). The two sequences are identical with the single exception that a U at position 55 of the Trinidad donkey strain has been deleted in the vaccine strain.

The sequences of the first 40 nucleotides adjacent to the poly(A) tract of these viruses are compared to the EEE sequence in Fig. 2. There is a striking conservation of the first 19 nucleotides. For all of the viruses except SFV and Middelburg, only position 6 shows any variability. SFV contains 2 changes in this region, and Middelburg has 5 changes (counting deletions as single changes); these last two viruses show a strong resemblance to one another in this region (also see below). The striking conservation of this region indicates that the precise sequence is important for virus replication.

Figure 2. Comparison of the First 40 Nucleotides at the 3'-End of Alphavirus RNAs. Sequences are shown from 3' to 5' reading from left to right and begin with the first nucleotide adjacent to the poly(A) tract. Abbreviations are defined in the legend to Figure 1. Since the first 40 nucleotides of the two VEE strains are identical, they are presented together. This is also true for the two strains of Sindbis virus. The highly conserved sequences are boxed. A space represents a deletion.

```
EEE NA       CUUUAUAAUUUUUGUUUUAUUUUUGUAUUUUCUUUUUAAU
EEE SA       -------------------------------------G-C
WEE          -----A-------------------  --- ---------AU
HJ           -----A-----------------CU-U--A-A-CA---GA
SIN          -----C-------------AAACAACUAA--A----CUU-
VEE          -------------------GGC-AAGCC---UC----CU-
SFV         C--------------G--A-CG--- ------A------GG
MBV         C--- -------- -G--ACCGC-GA-CG--AGCAGCGCUAU
```

Following this conserved region there is an area of the genome where all of the viruses have sequences rich in A and U but not otherwise conserved. The fraction of A and U in the first 50 nucleotides 5' to the poly(A) tract for these alphaviruses and for two unrelated viruses is shown in Table 1; the AU content is quite striking. Such AU rich regions may function as signals for polyadenylation (McGeoch & Turnbull, 1978).

Following the conserved 3' terminus, the various sequences appear quite divergent, even among closely related viruses (Fig. 1). The region shown in Figure 1 is almost entirely noncoding and the evolutionary constraints on sequence divergence appear less than in the coding region (Rice & Strauss, 1981b; also see below). For Sindbis, the coding region begins at nucleotide 319 (Rice & Strauss, 1981a), whereas for SFV it begins at nucleotide 265 (Garoff et al., 1980). We have examined the translated sequence of all the other viruses in all 3 reading frames, looking for homology to the carboxy terminus of the E1 glycoprotein of Sindbis and SFV (the last protein encoded in the RNA). From such homology we have determined that the coding region of WEE RNA begins at nucleotide 301, and that E1 of WEE terminates with two arginine residues as do the E1s of both Sindbis and SFV. From this analysis we believe that for the other viruses shown in Figure 1 the coding region has not been entered. Note in this regard that the EEE (South American strain) sequence extends for 332 nucleotides and its 3'-untranslated region is apparently longer than that of Sindbis, WEE, or SFV.

In order to visualize the relationships of the sequences to one another, a computer matrix was generated and is shown in Figure 3. Each sequence was searched for homology against itself (where homologies indicate repeated sequences) and against all of the other viral 3'-terminal sequences from Figure 1; a dot was put in the matrix whenever 6 out of the 7 nucleotides being compared were identical. The solid diagonal line results from the file being compared with itself (100% homology). Diagonal

Table 1: Percentages of Adenosine and Uridine in the 3'-Terminal 50 Nucleotides

VIRUS[a]	A_{50}[b]	U_{50}[b]	$(A+U)_{50}$[b]
WEE	22%	68%	90%
EEE NA	26%	64%	90%
EEE SA	24%	62%	86%
HJ	26%	60%	86%
SIN	56%	30%	86%
SFV	26%	54%	80%
VEE TC	16%	64%	80%
VEE TD	16%	64%	80%
MBV	26%	36%	62%
CMV[c]	32%	34%	66%
VSV[d]	36%	36%	72%

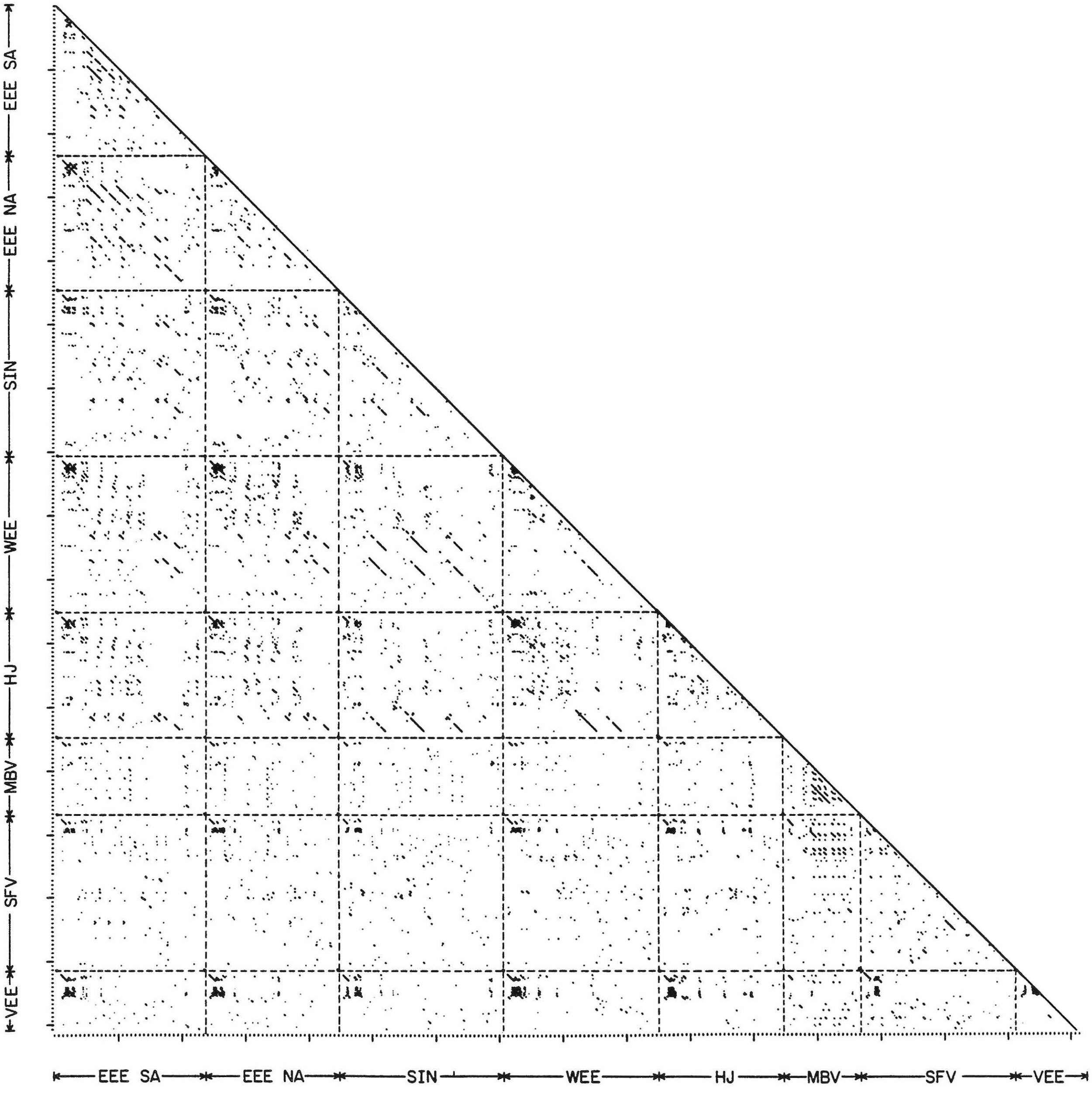
[a] Abbreviations as in the legend to Figure 1 except as otherwise noted.

[b] Percent of A, of U, or of A+U in the 50 nucleotides immediately adjacent to the poly(A) tract.

[c] Cowpea mosaic virus (Davies et al., 1979). Percentages shown are the average of M and B RNAs.

[d] Vesicular stomatitis virus (McGeoch & Turnbull, 1978). Percentages are calculated from N-protein mRNA of Indiana serotype strain.

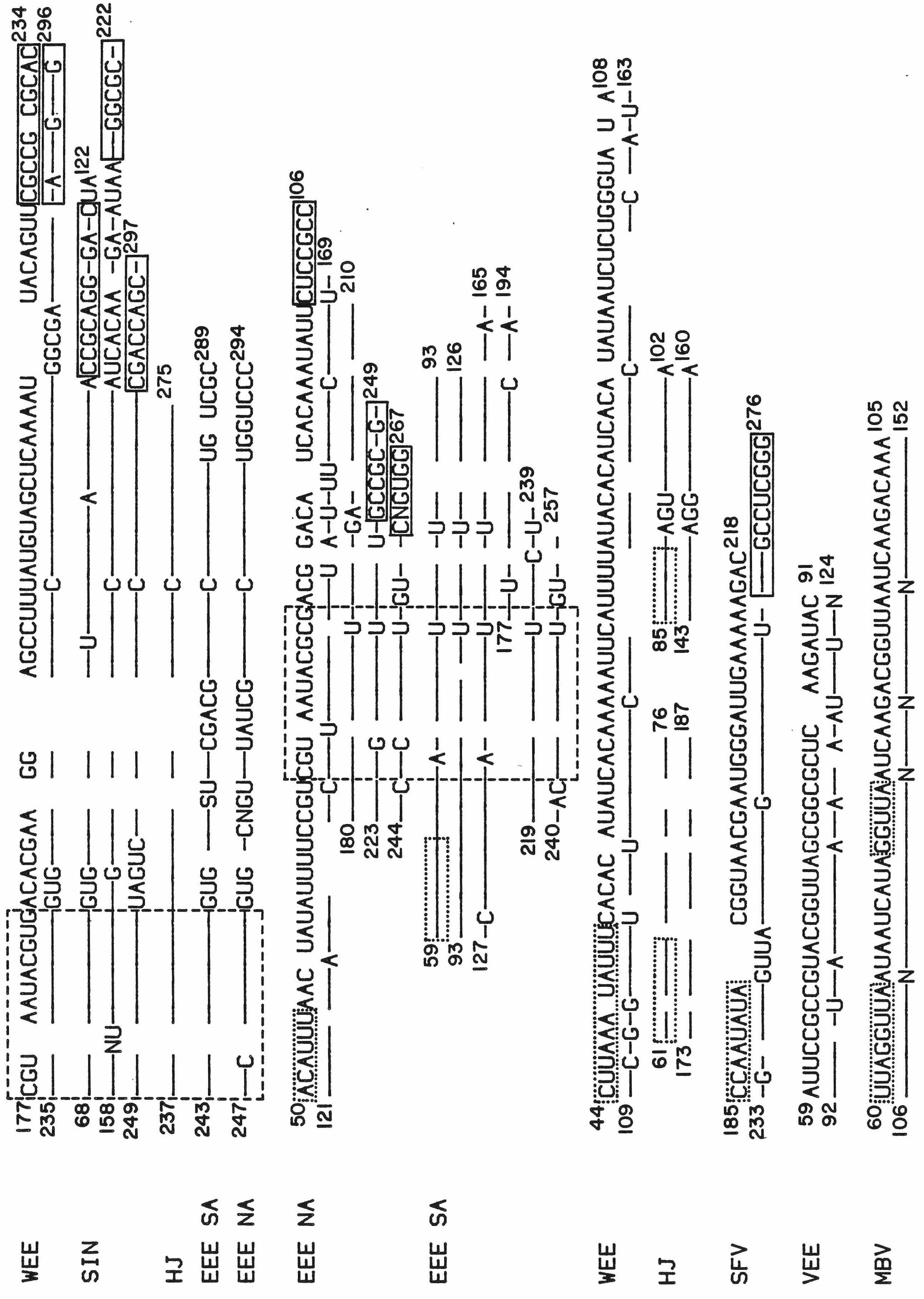
Figure 3. Computer-Plotted Dot Matrix. A file was generated which contained all of the sequences of Figure 1 (but containing only one strain of VEE and one of Sindbis since the two strains in the two cases have identical or virtually identical sequences) separated by short spacers. The file was then searched against itself with a program in which the search file was progressively shifted by one nucleotide relative to the test file; a dot was placed in the matrix whenever 6 of a string of 7 nucleotides being examined were identical. In this way each virus RNA sequence was tested against itself (triangular areas along the diagonal) and against every other viral RNA sequence (retangular areas where two RNA sequences intersect). Diagonal lines represent identical or homologous sequences. Note that sequences which are shared by two or more viruses are usually repeated in one or more viruses. The conserved 20 nucleotides at the 3' terminus are visible in each pattern; the cluster of dots in the shape of a square just after this is due to the AU rich region generating many matches. Nucleotides progress from 3' to 5' reading from left to right or from up to down. Tick marks are present every 10 nucleotides. Only the lower left half of the complete search is shown since the upper right half is a mirror image.


EEE SA
EEE NA
SIN
WEE
HJ
MBV
SFV
VEE
EEE SA
EEE NA
SIN
WEE
HJ
MBV
SFV
VEE

lines below and to the left represent repeated sequences within an RNA (if the RNA is being tested against itself) or a sequence shared between RNAs (when an RNA is tested against another RNA). Broken lines indicate the sequence is highly homologous but not identical. Several points emerge from an examination of this figure. First, repeated sequences are found in every virus RNA, but the number of repeats and length of the repeats varies with each virus. Second, most sequences that are shared by two or more viruses (excluding the 20 nucleotide conserved sequence at the 3'-terminus) are also repeated within one or more viruses. Third, by comparing the length and number of the diagonal lines which indicate homology within the various boxes the relatedness of the viruses is apparent (based on this noncoding region). Thus the two strains of EEE show long broken diagonal stretches when compared with one another, indicating they are closely related. These two viruses also show a few scattered sequences in common with Sindbis, WEE, and Highlands J viruses. These last 3 viruses show homologies in several repeated segments, indicating they are related. SFV shows little homology to any of the other viruses. The sequence data for Middelburg and VEE are limited, but little apparent homology to any of the other viruses is evident at this level of analysis.

The repeated sequences are examined in more detail in Figure 4. A second computer search program was used to examine the RNA sequences of Figure 1 for homologous or repeated sequences, and all such sequences of 15 nucleotides or greater are shown aligned to maximize homology. The top line shows a sequence of about 60 nucleotides which is present as a tandem repeat in WEE RNA (from nucleotides 177 to 234 and from 235 to 296) with only a few differences in the two copies of this sequence. The same sequence, again with minor changes, is present three times in Sindbis RNA, in this case as nontandem repeats (Fig. 4). One copy of a somewhat shortened version of this sequence is also present in Highlands J RNA and in the two strains of EEE RNA. In all cases the single copy of this sequence or one of

Figure 4. Repeating Sequences of Alphavirus RNAs. Repeating sequences were found with a computer search program and aligned to show maximal homology. In each of the 6 series of sequences the sequence in the first line is given in full, and each of the other sequences of that series is compared with it. If the nucleotide is the same as on the first line, only a dash is shown; otherwise the different nucleotide is given. Spaces in a sequence represent deletions (or additions in the corresponding regions of other sequences). Sequences at the beginning of a repeat which are homologous to parts of the conserved 20-nucleotide-sequence at the 3'-terminus are boxed with dotted or dashed lines. GC rich sequences following a repeat are boxed with solid lines. Sequences read 3' to 5' from left to right. The numbers indicate the first and last nucleotide shown in a sequence, numbering from 3' to 5' beginning with the first nucleotide adjacent to the poly(A) tract.


WEE 177 CGU AAUACGUGACACGAA GG AGCCUUUUAUGUAGCUCAAAAU UACAGUUCGCCG CGCAC 234
235 GUG C GGCGA -A- G- G 296
SIN 68 GUG U A ACCGCAGG-GA-CUA 122
158 NU G C AUCACAA -GA-AUAA-GGCGC- 222
249 UAGUC C CGACCAGC- 297
HJ 237 C 275
EEE SA 243 GUG SU CGACG C UG UCGC 289
EEE NA 247 C GUG CNGU UAUCG C UGGUCCC 294
EEE NA 50 ACAUUUAAC UAUAUUUUCCGUCGU AAUACGCGACG GACA UCACAAAAUAUUCUCCGCC 106
121 A C U U A-U-UU C U- 169
180 G U -GA- 210
223 G U U-GCCGC-G- 249
244 C C U-GU- CNGUGG 267
EEE SA 59 A- U U 93
93 U U 126
127 C A- U U A- 165
177 U C A- 194
219 U C-U- 239
240-AC U-GU- 257
WEE 44 CUUAAA UAUUUCACAC AUAUCACAAAAAUUCAUUUUAUACACAUCACA UAUAAUCUCUGGGUA U A 108
109 C-G-G U U C C C A-U- 163
HJ 61 76 85 AGU A 102
173 187 143 AGG A 160
SFV 185 CCAAUAUA CGGUAACGAAUGGGAUUGAAAAAGAC 218
233 G GUUA G U GCCUCGGG 276
VEE 59 AUUCCGCCGUACGGUUAGCGGCGCUC AAGAUAC 91
92 U A A A A-AU U-N 124
MBV 60 UUAGGUUAAUAAUCAUAGGUUAAUCAAGACGGUUAAUCAAGACAAA 105
106 N N N N N 152

the multiple copies starts about 240 nucleotides from the poly(A) tail. The sequences of Middelburg RNA and VEE RNA do not extend far enough to determine whether this sequence is also present in these viruses. It is clearly not present in SFV RNA, but there is a repeated sequence in this RNA, one copy of which begins at nucleotide 233 (Figure 4).

Next shown in Fig. 4 is a sequence of about 50 nucleotides in EEE (NA) RNA which is repeated once in almost its entirety and also repeated in part three more times in this RNA. Parts of this sequence are found six times in EEE (SA). Note that the first three times this sequence occurs in EEE (SA) it forms a set of tandem repeats which are highly conserved among themselves. The central region of this highly repeated sequence is homologous to parts of the WEE-repeated sequence discussed above (Fig. 4; also see below).

There is a second sequence in WEE RNA, about 60 nucleotides long, which is found twice as a tandem repeat (Fig. 4). Large portions of this repeated sequence are also found twice in Highlands J RNA (Fig. 4) but not in other virus RNAs. Note that in Highlands J RNA parts of this repeat have been translocated relative to the WEE sequence.

Next shown in Fig. 4 is a repeated sequence of about 30 nucleotides in SFV RNA, one copy of which begins at nucleotide 233 as discussed above. Below this is shown a sequence in VEE RNA which is tandemly repeated (Fig. 4). Finally, there is a long (45 nucleotide) tandem repeat in Middelburg RNA. The Middelburg repeat appears to be identical, suggesting that the generation of this repeat is a recent event. Note also that within this long tandem repeat are shorter tandem repeats (positions 77-89 and 90-102 are repeats, as are 124-136 and 137-149).

4. Discussion

We undertook a comparative study of the 3'-terminal sequences of alphaviruses for two reasons. First, conservation of sequences is an important clue in determining the function of a region and its importance in the virus life cycle. Second, a comparative study of these viruses is important in our understanding of the evolutionary pressures operating on the alphaviruses and of their relatedness to one another.

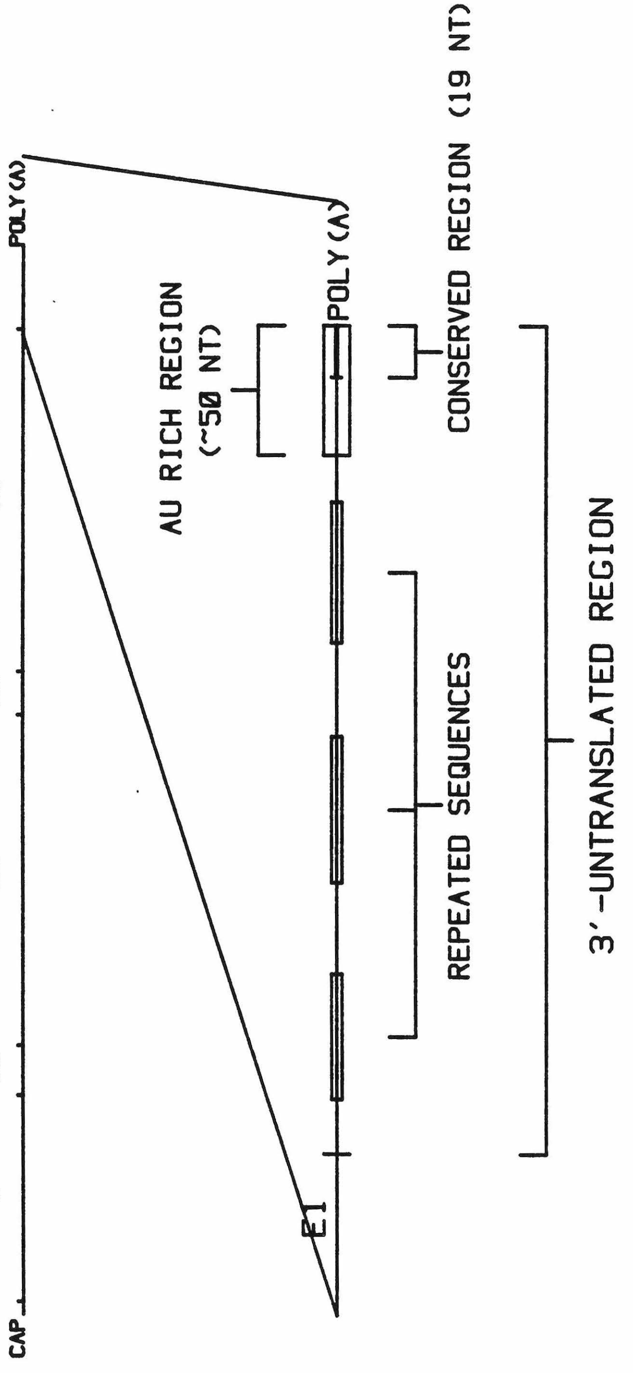
The overall organization of the 3'-terminal noncoding regions of alphavirus RNAs is schematically represented in Figure 5. The major features include the conserved 19 nucleotide sequence next to the poly(A) tract, the A + U rich segment which comprises the 50-60 nucleotides immediately adjacent to the poly(A) tract, and the repeating sequence region which extends from nucleotide 50 or 60, numbering from the poly(A), to the coding region.

We have examined representatives of several serologic subgroups of the alphaviruses. These are the EEE group, the VEE group, the WEE group, which also contain Highlands J virus and Sindbis virus, and SFV which is sometimes grouped with VEE (Porterfield, 1980) and sometimes with Chickungunya as a separate group (Theiler & Downs, 1973). Very little has been done with Middelburg. It has been reported to be more closely related to SFV than to Sindbis (Kokernot et al., 1957) but is also placed at times in the WEE group (Porterfield, 1980). This grouping is based on serologic cross reaction which measures primarily the relatedness of the virion glycoproteins. The results from the 3'-terminal sequences are in general agreement with this grouping. We note that based on the conserved 19 nucleotide sequence SFV and Middelburg appear related to one another and different from the other viruses.

The degree of sequence divergence in the 3'-terminal noncoding regions of the various viruses is perhaps surprising. Even closely related strains show considerable divergence, and at the extreme Sindbis and SFV show little sequence homology in the 3'-noncoding region, even though they are 45% homologous at the amino acid level in the three structural proteins (Rice & Strauss, 1981a). Thus the conserved sequences found in the 3'-terminal noncoding region take on added significance.

Figure 5. Schematic diagram of the organization of the 3'-terminal noncoding regions of alphavirus RNAs. The organization of 26S RNA is shown for scale in the upper part of the figure. The 3'-untranslated region is expanded in the lower part of the figure. The figure is drawn to scale for Sindbis virus. Other alphavirus RNAs differ in detail but contain the same general features.

SINDBIS VIRUS 26S RNA (41Ø2 NT)



We have found a stretch of 19 nucleotides immediately adjacent to the poly(A) tract which is very highly conserved. We hypothesize that this sequence forms a replicase recognition site and the precise sequence of this region is required for binding. We have recently observed that a variant of Sindbis virus which is no longer interfered with by the parental defective interfering particles (Weiss & Schlesinger, 1981) has a single change within this conserved region (Monroe et al., 1981). One interpretation of this result is that the variant has evolved a replicase whose recognition sequence is slightly altered (by one nucleotide) and which is now able to recognize the variant RNA for replication but not the defective interfering RNA.

It is of interest to note that even if this conserved 3'-terminal sequence is a replicase binding site for production of minus strands, the replicase must also interact with other sequences in the RNA, presumably found near the 5'-terminus of 49S RNA. The rationale for this conclusion is that 26S RNA is not replicated although it is the 3'-terminal one third of 49S RNA. The cyclization of 49S RNA (Hsu et al., 1973; Frey et al., 1979) may be involved in this dual recognition phenomenon, as it brings sequences at the 5' end of the RNA near to those at the 3' end, and could allow the replicase to interact simultaneously with sequences at both ends. The study of defective interfering particles should be useful in this regard, as such recognition signals would be preserved. The recent findings that alphavirus DI RNAs do not contain simple deletions but instead contain numerous sequence rearrangements (Monroe et al., 1981; Lehtovaara et al., 1981) suggests a possible mechanism for the enhanced replicative ability of these RNAs. If the 5' sequences recognized by the replicase are relocated near the 3' end, the resulting molecule would always be in a configuration to be replicated. On the other hand, initiation of minus strand synthesis on 49S RNA would require that the molecule be cyclized.

We have also found a sequence which is conserved in large part among all the members of the EEE and WEE groups, and in some cases is repeated in a viral RNA

(Fig. 4). One copy (or the only copy) of this sequence begins about 240 nucleotides from the 3'-terminal poly(A). This conservation of position as well as of sequence implies an important function in the virus life cycle. This function is unclear at present, although it could be involved in cyclization of the virus RNA (unpublished observations). In any event, this sequence appears to be absent from the 3'-terminal regions of RNAs from defective interfering particles of Sindbis virus (Monroe et al., 1981), and thus RNA replication and packaging can apparently proceed without this sequence.

The finding of long repeated sequences in the 3'-terminal noncoding region was not expected. Moreover, except for the 19-nucleotide-conserved sequence at the end, these repeated sequences represent the areas of greatest conservation between the different alphaviruses. In an attempt to probe the origins of these repeated sequences we have examined them in some detail (Fig. 4). They are often preceded by short sequences which are homologous to a part of the conserved sequence at the 3' end (dashed or dotted boxes in Fig. 4) and are often followed by GC rich sequences (solid boxes in Fig. 4). In Table 2 we compare in more detail these boxed sequences preceding the repeats with the corresponding homologous sequences found in the 19-nucleotide-conserved 3'-terminus. It has been proposed that template-switching during replication is one of the mechanisms for generating the defective interferring RNAs of minus-stranded RNA viruses (Lazzarini et al., 1981). We suggest that the repeating sequences in the alphavirus RNAs could also be generated by a template-switching mechanism, and that the short preceding sequences could encourage template switching because of their homology to the normal replicase binding site at the 3' end of the RNA. The presence of GC rich regions following a number of the repeated sequences could also be of importance in their generation, by creating areas of secondary structure difficult to read through by the replicase and encouraging template switching.

Table 2: Comparison of the Sequences Preceding Repeated Sequences with Homologous Regions of the 20 Nucleotide Conserved Sequence at 3'-terminus

Virus[a]	Position of Preceding Sequence[b]	Preceding Sequence Conserved Sequence[c]
WEE	44-54	CUUAAAUAUUU CUUUAAAAUUU
	177-183	CGUAAUA CUUUAAA
SIN	68-74	CGUAAUA CUUUACA
EEE NA	50-55	ACA-UUU AUAAUUU
EEE SA	59-65	AUA-UUUU AUAAUUUU
HJ	61-66	AAUAUU AAAAUU
	85-89	AUUUU AUUUU
SFV	185-192	CCAAUAUA CCUUUAUA
MBV	60-67	UUAGGUUA UUUGGUUA
	77-81	GGUUA GGUUA

[a] Abbreviations as in the legend to Figure 1.

[b] Nucleotides are numbered from 3' to 5' beginning with the first nucleotide adjacent to the poly(A) tract.

[c] The preceeding sequence boxed in Figure 4 is compared with a segment of the 20 nucleotide conserved sequence at the 3'-terminus of the same virus.

The lengths of the 3'-terminal noncoding regions (exclusive of the poly A tract) are known for Sindbis virus (318 nucleotides), SFV (264 nucleotides), and WEE (300 nucleotides) and are all approximately the same size. It is possible that this region undergoes frequent (in evolutionary terms) rearrangements through a series of sequence duplication and deletions, resulting in the observed divergence in sequence. The variability in size and sequence of the noncoding region would be limited by packaging constraints and by the requirement that certain sequences be retained because they have a function in replication.

Acknowledgements

We thank E. M. Lenches for preparing chick embryo cell cultures and E. G. Strauss for help in preparing the manuscript. We are grateful to Michael Douglas and Tim Hunkapiller for assistance with the computer work and to Lee Hood for use of his computer facilities. This work was supported by grants PCM 8022830 from the National Science Foundation and grant GM06965 from the National Institutes of Health. J.-H. Ou was supported by fellowships from the California Foundation for Biomedical Research and from the Li Ming Scholarship Endowment Fund.

References

Burge, B. W. & Pfefferkorn, E. R. (1966). Virology **30**, 204-213.

Davies, J. W., Stanley, J. & Kammen, A. V. (1979). Nucleic Acids Res. 7(2), 493-500.

Frey, T. L., Gard, D. & Strauss, J. H. (1979). J. Mol. Biol. **132**, 1-18.

Garoff, H., Frischauf, A.-M., Simons, K., Lehrach, H. & Delius, H. (1980). Nature **288**, 236-241.

Haaster, C. C.-V. & Bishop, D. (1980). Virology **105**, 564-574.

Hsu, M. T., Kung, H. J. & Davidson, N. (1973). Cold Spring Harbor Symp. Quant. Biol. **38**, 943-950.

Kennedy, S. I. T. (1980). In The Togaviruses (Schlesinger, R. W., ed.), pp. 351-369, Academic Press, New York.

Kokernot, R. H., Demeillon, B., Paterson, H. E., Heymann, C. S. & Smithburn, K. C. (1957). S. Afr. J. Med. Sci. **22**, 145-153.

Lazzarini, R. A., Keene, J. D. and Schubert, M. (1981). Cell **28**, 145-154.

Lehtovaara, P., Soderlund, H., Keranen, S., Pettersson, R. F. & Kaariainen, L. (1981). Proc. Natl. Acad. Sci. USA, in press.

Maxam, A. M. & Gilbert, W. (1980). Methods in Enz., Part 1, **65**, 499-560.

McGeoch, D. J. & Turnbull, N. T. (1978). Nucleic Acids Res. **5**(11), 4007-4024.

Monroe, S., Ou, J.-H., Rice, C. M., Schlesinger, S., Strauss, E. G. & Strauss, J. H. (1981). J. Virol., in press.

Ou, J.-H., Strauss, E. G. & Strauss, J. H. (1981). Virology **109**, 281-289.

Porterfield, J. S. (1980). In The Togaviruses (Schlesinger, R. W., ed.), pp. 13-46, Academic Press, New York.

Rice, C. M. & Strauss, J. H. (1981a). Proc. Natl. Acad. Sci. USA **78**, 2062-2066.

Rice, C. M. & Strauss, J. H. (1981b). J. Mol. Biol. **150**, 315-340.

Sanger, F., Nicklen, S. & Coulson, R. (1977). Proc. Natl. Acad. Sci. USA **76**, 5463-5467.

Smith, D. R. & Calvo, J. M. (1980). Nucleic Acids Res. **8**, 2255-2274.

Strauss, J. H. & Strauss, E. G. (1977). In The Molecular Biology of Animal Viruses (Nayek, D. P., ed.), vol. I, pp. 111-166, Marcel Dekker, Inc., New York.

Theiler, M. & Downs, W. G. (1973). In The Arthropod Borne Viruses of Vertebrates Yale University Press, New Haven, Conn., Chapter 6, pp. 113-146.

Trent, D. W., Clewley, J. P., France, J. K. & Bishop, D. H. L. (1979). J. Gen. Virol. **43**, 365-381.

Trent, D. W. & Grant, J. A. (1980). J. Gen. Virol. **47**, 261-282.

Weiss, B. & Schlesinger, S. (1981). J. Virol. **37**(2), 840-844.

Chapter 4

This chapter was published in the Journal of Virology.

Journal of Virology, Jan. 1982, p. 153–162
0022-538X/82/010153-10$02.00/0

Vol. 41, No. 1

Sequence Analysis of cDNA's Derived from the RNA of Sindbis Virions and of Defective Interfering Particles

STEPHAN S. MONROE,[1] JING-HSIUNG OU,[2] CHARLES M. RICE,[2] SONDRA SCHLESINGER,[1]* ELLEN G. STRAUSS,[2] AND JAMES H. STRAUSS[2]

Department of Microbiology and Immunology, Washington University School of Medicine, St. Louis, Missouri 63110,[1] and Division of Biology, California Institute of Technology, Pasadena, California 91125[2]

Received 15 June 1981/Accepted 28 August 1981

Sindbis virus generates defective interfering (DI) particles during serial high-multiplicity passage in cultured cells. These DI particles inhibit the replication of infectious virus and can be an important factor in the establishment and maintenance of persistent infection in BHK cells. In an effort to understand how these DI particles are generated and how they interfere with the replication of standard virus, we performed a partial sequence analysis of the RNA obtained from two independently isolated populations of DI particles and from two Sindbis virus variants and compared these with the RNA of the parental wild-type virus. The 3'-terminal regions of the RNAs were sequenced by the dideoxy chain terminating method. Internal regions of the RNA were examined by restriction endonuclease digestion of cDNA's made to the various RNAs and by direct chemical sequencing of 5' end-labeled restriction fragments from cDNA made to the DI RNAs. One of the variant viruses examined was originally derived from cells persistently infected with Sindbis virus for 16 months and is resistant to interference by the DI strains used. In the 3'-terminal region of the RNA from this variant, only two base changes were found; one of these occurs in the 20-nucleotide 3'-terminal sequence which is highly conserved among alphaviruses. The DI RNA sequences were found to have been produced not by a single deletional event, but by multiple deletion steps combined with sequence rearrangements; all sequences examined are derived from the plus strand of Sindbis virion RNA. Both DI RNAs had at least 50 nucleotides of wild-type sequence conserved at the 3' terminus; in addition, they both contained conserved and perhaps amplified sequences derived from the non-26S region of the genome which may be of importance in their replication and interference ability.

Defective interfering (DI) particles are nonviable deletion mutants produced by serial high-multiplicity passage of most viruses in cultured cells. These particles are characterized by their ability to interfere effectively with the replication of the homologous standard virus. The mechanism by which deletions are generated may vary for different viruses, but those sequences in the parental virus genome which are essential for replication and packaging should be conserved in the defective nucleic acids. To develop an understanding of how DI particles are generated and how they interfere with the growth of infectious virus, it is necessary to determine the sequence relationships between the standard and DI genomes. Our laboratories have had a long-standing interest in alphavirus replication. There is now considerable information available about the primary structure of the RNAs from two alphaviruses, Sindbis virus (12, 14; J.-h. Ou and E. Strauss, unpublished data) and Semliki Forest virus (3, 4), making it feasible to begin to probe the sequences of the nucleic acids from their DI particles.

The alphavirus genome is a single-strand 49S RNA which is infectious and can be translated to produce the nonstructural proteins involved in RNA replication (17, 20). During the infection cycle, intracellular full-length minus-strand RNA is made and serves as a template for the synthesis of progeny genome RNA and a 26S mRNA. Both of these RNA products are capped at the 5' end and polyadenylated at the 3' end. The 26S RNA is the 3' third of the genome and codes for the virion structural proteins. Passaging either Sindbis or Semliki Forest virus at high multiplicity leads to the formation of DI particles in which the RNA appears to undergo successive deletions until a limit size of 18 to 20S RNA is reached (5, 19). On the basis of oligonucleotide fingerprinting and nucleic acid hybridizations, a model for the structure of these DI RNAs has been proposed in which 3'- and 5'-terminal sequences of virion RNA are conserved and internal sequences of the genome are progressively deleted (6, 8). Recently, however Pettersson reported that the 5′-terminal sequence of a limit size population of Semliki Forest virus DI RNAs is heterogeneous and different from the standard sequence (13). These results suggest that the structure of alphavirus DI RNAs may be more complex than had originally been thought.

Because of their ability to limit productive infection, DI particles can be involved in the establishment and maintenance of persistent viral infections in cultured cells (7). Weiss et al. established a line of BHK cells persistently infected with Sindbis virus by infecting the cells with a preparation of this virus containing a high ratio of DI particles to standard virus (21). One month after the persistent infection was established the infectious virus [Sin-1(1)] released from the infected cells was characterized as a small-plaque, temperature-sensitive mutant which was slightly less sensitive than the wild-type virus to interference by the original DI particles. Sixteen months after the persistently infected cells were in culture, the infectious virus [Sin-1(16)] cloned from these cells had become resistant to interference by the original DI particles (22). This phenotypic alteration must reflect a specific change or changes in the genome of the virion which prevent interference from occurring.

In this report we describe the results obtained by examining sequences of the RNA from two independently isolated DI particle populations of Sindbis virus and from the two variants cloned from persistently infected cells.

MATERIALS AND METHODS

Materials. Deoxy- and dideoxyribonucleoside triphosphates, primers, and T_4 polynucleotide kinase were purchased from P-L Biochemicals, Inc. The p(dT)$_7$rG primer was purified as described previously (12). Bacterial alkaline phosphatase was purchased from Bethesda Research Laboratories, Inc. Oligo(dT)-cellulose was purchased from Collaborative Research, Inc. [^{3}H]uridine was purchased from New England Nuclear, [α-^{32}P]deoxynucleoside triphosphates, from Amersham Searle, and [γ-^{32}P]ATP from ICN. Restriction enzymes were purchased from New England Biolabs. Avian myeloblastosis virus reverse transcriptase was a generous gift from J. W. Beard.

Isolation and purification of Sindbis virus RNAs. The derivation of virus strains and DI particles is summarized in Table 1. Virus preparations to be used for RNA isolation were obtained by one-cycle infections of secondary chicken embryo fibroblasts grown in glass roller bottles. Virus particles released into the medium were concentrated by centrifugation and puri fied by sucrose gradient centrifugation. The RNA was isolated by extraction with phenol and chloroform and purified by oligo(dT)-cellulose chromatography (9). The RNA from Sin-1(1) was also purified by velocity gradient centrifugation. All procedures were carried out with aseptic techniques. Samples of the purified RNA were analyzed by agarose gel electrophoresis after denaturation with glyoxal (1). The RNAs isolated from the two DI preparations were predominantly mixtures of DI RNAs (Fig. 1) and are referred to as DI RNA populations to reflect their heterogeneity. The RNAs isolated from Sin-1(1) and Sin-1(16) comigrated with wild-type Sindbis 49S RNA (Fig. 1).

Sequencing by dideoxy chain termination. The 3′-terminal sequences of the viral RNAs were determined by the dideoxy chain termination technique (16) with the use of reverse transcriptase and the p(dT)$_7$rG primer as described previously (12). The reactions contained 50 mM Tris-hydrochloride, pH 8.3, 50 mM KCl, 8 mM $MgCl_2$, 0.4 mM dithiothreitol, 0.4 pmol of RNA, and 0.1 μg of primer in 20 μl. cDNA synthesis was initiated by the addition of 4 U of avian myeloblastosis virus reverse transcriptase. The chain termination reactions contained 5 μM [α-^{32}P]dATP (400 Ci/mmol), 50 μM of the other three unlabeled deoxyribonucleoside triphosphates, and 0.5 to 50 μM dideoxyribonucleoside triphosphates. All reactions were incubated for 20 min on ice followed by 1 h at 37°C and were terminated by the addition of 10 μl of 10 M urea containing 0.1 mM dATP and dyes. Samples were heated for 2 min at 90°C, quickly cooled on ice, and loaded onto thin sequencing gels (15).

Synthesis of cDNA. We followed the methods described by Rice and Strauss (14a). Briefly, cDNA was synthesized in reactions containing: Tris-hydrochloride, pH 8.3, 50 mM; KCl, 50 mM; $MgCl_2$, 8 mM; dithiothreitol, 2 mM; sodium pyrophosphate, 4 mM; oligo(dT)$_{12-18}$, 10 to 20 μg/ml; digested calf thymus DNA, 300 μg/ml; template RNA, 30 to 60 μg/ml; and avian myeloblastosis virus reverse transcriptase, 500 U/ml. Analytical reactions contained ~1 μM α-^{32}P-labeled and 0.2 mM of each unlabeled deoxyribonucleoside triphosphate in 20 μl. Preparative reactions contained 1 mM unlabeled triphosphates and 50 μCi of [^{3}H]dTTP in 200 μl. Reactions were incubated at 42.5°C for 1 h, stopped by adding EDTA to 25 mM, phenol extracted, treated with 0.1 M NaOH at 60°C for 1 h to remove template RNA, and neutralized. Preparative cDNA's were purified further on a 5-ml Bio-Gel A5M column. The excluded fractions were pooled, lyophilized, resuspended in water, and ethanol precipitated two times.

Restriction enzyme digestion of cDNA. Single-stranded cDNA's were cleaved by digestion with restriction enzymes *Hae*III, *Hha*I, *Rsa*I, and *Taq*I (14a). To prepare fragments for sequencing, the digests were treated with alkaline phosphatase, phenol extracted, and labeled at the 5′ ends with 1 mCi of [γ-^{32}P]ATP and polynucleotide kinase. The products were resolved on 5% or 6% sequencing gels and localized by autoradiography. Fragments to be sequenced were excised from the gel, eluted in 0.6 M sodium acetate, 0.1 M Tris-hydrochloride, pH 8, 2 mM EDTA, and 25 μg of tRNA per ml for 40 to 65 h at room temperature. The eluted DNA was ethanol precipitated twice to remove salt and was lyophilized before sequencing.

DNA sequencing. The procedures for base-specific chemical cleavages were based on those of Maxam and Gilbert (11) as modified by Smith and Calvo (18). The four base-specific reactions used were C, C + T, G + A and G. Samples were denatured for 2 min at

TABLE 1. Source of Sindbis RNAs used for sequencing studies

Virus	Description	Reference
DI-1	Obtained after 16 undiluted passages of Sindbis virus in BHK cells	9
DI-2	Obtained after 18 undiluted passages of Sindbis virus in chicken embryo fibroblasts	2
Sin-1(1)	Cloned from BHK cells 1 month after the establishment of persistent infection	21
Sin-1(16)	Cloned from the BHK cells described above but 16 months after persistent infection was established	22
Wild type	This is the parent of the HR strain and was obtained originally from B. Burge	

90°C in 80% formamide, quickly cooled, and analyzed on thin sequencing gels.

Thin sequencing gels. The gels were essentially the same as those described originally by Sanger and Coulson (15). The gels were either 30 by 40 by 0.04 cm or 30 by 80 by 0.04 cm. The voltage (1,800 to 2,400 V for 80-cm gels) was adjusted during the run to maintain the temperature of the gels at ~50°C. After electrophoresis, gels were transferred to Whatman 3 MM paper and exposed to presensitized film with intensifying screens at −70°C (10).

RESULTS

Sequencing of the 3′-terminal regions of Sindbis virus RNAs. On the basis of T_1-resistant oligonucleotide fingerprinting of Sindbis DI RNAs, Dohner et al. identified a characteristic 3′ oligonucleotide conserved in all of the DI RNAs (2). They were not able to ascertain, however, to what extent these RNAs were identical to the standard virion RNA at the actual 3′ terminus. Furthermore, all of the DI RNAs examined had a deletion near the 3′ terminus, raising the possibility that the deletions always occurred at the same site. We sequenced the 3′-terminal regions of two populations of DI RNAs to determine which sequences are conserved and where the deletions begin. We also sequenced the 3′ termini of the RNA from the two Sindbis virus variants, Sin-1(1) and Sin-1(16). The latter virus is no longer sensitive to interference by the DI particles used to initiate the persistent infection (22). We wished to know whether there was any correlation between resistance to interference by DI particles and alterations at the 3′ terminus of the RNA.

Chain termination methods were used to determine the nucleotide sequence at the 3′ terminus of these RNAs. Since the first nucleotide adjacent to polyadenylate [poly(A)] is C for many alphaviruses (12), we carried out avian myeloblastosis virus reverse transcriptase reactions using $p(dT)_7rG$ as the specific primer. The first few nucleotides adjacent to the priming site were determined with only one or two deoxyribonucleoside triphosphates in the reverse transcriptase reactions. The sequence of the first five nucleotides adjacent to the 3′ poly(A) of all the viral RNAs analyzed is 5′ A-U-U-U-C-poly(A) 3′ and is identical to that found for Sindbis 49S RNA (12).

The sequence information obtained from reactions containing dideoxynucleoside triphosphates is shown in Fig. 2. The 20-nucleotide stretch at the 3′ end of the RNA adjacent to the poly(A) tract, which has been shown to be highly conserved in the alphaviruses (12; J.-h. Ou, unpublished data), is identical in the wild-type parental strain, the two DI populations, and Sin-1(1). Sin-1(16) shows a single base change in this region, a T to C change at position 7 of the cDNA. Sin-1(1) is identical to the wild type for at least the first 150 nucleotides, whereas Sin-

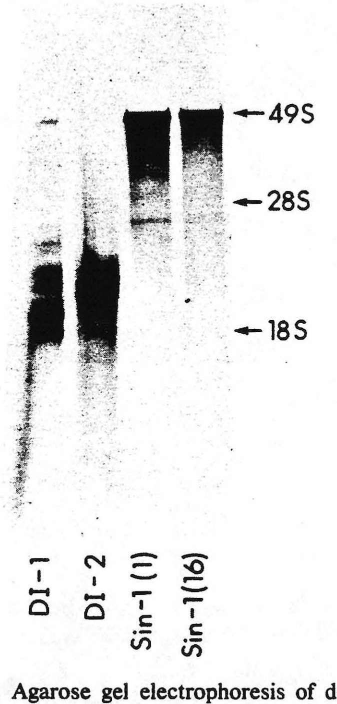


FIG. 1. Agarose gel electrophoresis of denatured Sindbis virus RNAs. Samples of [^{3}H]uridine-labeled RNAs were denatured in 10 mM sodium phosphate, pH 7, 1 M glyoxal, and 50% dimethyl sulfoxide for 1 h at 50°C. They were loaded on a horizontal 1% agarose gel (16 by 20 by 0.3 cm) made in 10 mM phosphate and electrophoresed for 3 h at 4 V/cm with constant buffer recirculation. The gel was impregnated with 2,5-diphenyloxazole (PPO), dried, and exposed to presensitized film at −70°C (10).

```
                                                   SEQUENCE
  cDNA

26S  (HR)     1    GAAATGTTAA AAACAAAATT TTGTTGATTA ATAAAAGAAA TAATAAAAGT TATGCAGACG CTGCGTGGCA TTATGCACCA     80
Wild Type     1    .......... .......... .......... .......... .......... .......... .......... ..........     80
Sin-1(1)      1    .......... .......... .......... .......... .......... .......... .......... ..........     80
Sin-1(16)     1    ......C... .......... .......... .......... .......... .......... .......... ........N.     80
DI-1          1    .......... .......... .......... .......... N......... N......... .N...N.... ...            73
DI-2          1    .......... .......... .......... .......... .........N AGTCTGSATG AGCGGCCGGC TGACATNCCA     80

26S  (HR)    81    CGCTTCCTCA GAAATACATT GAGTTTTTGG CGTCCGCTAG ATAAATGGTT AATATAGTGG TTATGTGGCA               150
Wild Type    81    .......... .......... .......... .......... .......... .......... ..........               150
Sin-1(1)     81    .......... .......... .......... .......... .......... .......... ..........               150
Sin-1(16)    81    .......... .......... .......... .......... .......... .....G.... .....N....               150
DI-2         81    AGGAGCCGNA GCATTTGPTT TTTGCTAGGA NCGGCAGAAG GTPATTTTGC ATGGTGTTGG NGTTCCT                  147
```

FIG. 2. The 5′-terminal sequences of cDNA's transcribed from the 3′ end of Sindbis virus RNAs. The sequence of 26S cDNA was determined previously (14, 14a). In each case the 5′ terminal G is derived from the p(dT)$_7$rG primer. The sequences were compared for homology by using a computer program. A dot indicates that the base is identical to the base listed for 26S cDNA. N, any nucleotide; P, purine; Y, pyrimidine; R, A or T; S, C or G.

1(16) shows, in addition to the change at position 7, a second change at position 136. DI-2 is identical to the wild type for the first 50 nucleotides, after which the sequences diverge. The following 54 nucleotides of DI-2 RNA (positions 51 to 104) are identical to positions 1410 to 1463 of Sindbis 26S RNA (HR strain) (14) and presumably to the parental wild-type 49S RNA. Positions 105 to 147 of DI-2 RNA are derived from the non-26S region of 49S RNA (E. Strauss, unpublished data). DI-1 RNA is identical to wild-type RNA for the first 73 nucleotides. After this, the DI-1 sequence becomes heterogeneous and could not be unambiguously determined. This indicates that at least some of the RNAs in the DI population diverge from the wild-type sequence at this point.

Restriction endonuclease digestion of viral cDNA's. An examination of the pattern of fragments obtained by digestion of DNA with restriction enzymes provides a means of comparing related nucleic acids. Several of the type II restriction enzymes cleave single-strand DNA, making it possible to apply this mapping procedure to cDNA copies of RNA. Accordingly, cDNA's were synthesized with reverse transcriptase by using the genomic RNAs as templates and a mixture of oligo(dT)$_{12-18}$ and digested calf thymus DNA as primers. After removal of the template, the cDNA's were digested with three different restriction enzymes, and the products were resolved on 5% sequencing gels (Fig. 3). Included in Fig. 3 is the pattern obtained from cDNA to 26S RNA to identify those fragments from this region of the genome.

The restriction patterns of the çDNA's from the two Sindbis variants [Sin-1(1) and Sin-1(16)] resemble quite closely those of Sindbis wild-type virion cDNA; the few differences are marked by arrows in Fig. 3. Sin-1(1) has one missing *Taq*I fragment derived from the non-26S region of the genome. Sin-1(16) shows alterations in two *Rsa*I fragments derived from 26S RNA, as well as one *Rsa*I fragment and four *Taq*I fragments derived from the non-26S RNA region of the RNA. We note, however, that many of the fragments observed result from partial (incomplete) cleavage (14a), and it is possible that a single change in the RNA could result in an observable change in more than one band.

The restriction patterns from the two DI cDNA's are quite different from those of the wild-type virus. Most of the wild-type restriction fragments and especially those from the 26S region are missing in the DI patterns. In addition, the two DI cDNA's have very different restriction patterns, although there are several bands in common (*Hae*III fragments C and D, for example). We note also that the patterns from the DI cDNA's are relatively simple, with few prominent bands. This limited complexity is similar to what was observed in the T$_1$-resistant oligonucleotide maps (2, 9).

Sequencing of *HaeIII* fragments of DI-1 and DI-2. Preparations of cDNA to DI-1 and DI-2 were digested with *Hae*III, and the resulting fragments were labeled at the 5′ ends with T$_4$ polynucleotide kinase and [γ-^{32}P]ATP and were separated on a preparative gel. Fragments corresponding to those labeled in Fig. 3 were excised, eluted from the gel, and sequenced by the chemical methods of Maxam and Gilbert (11). The fragments varied in length from DI-1 fragment A, which is >850 nucleotides long, to fragment K, only 30 nucleotides long. (Fragment K is not shown in Fig. 3.) Some of the shorter fragments were sequenced in their entirety. For other fragments, either the first 35 nucleotides or the sequence beyond ~400 nucleotides from the 5′ end or both were not determined. The sequences obtained are shown in Fig. 4 and were aligned to maximize homology between fragments.

Figure 5A illustrates the location of 26S sequences in the DI cDNA's. This scheme incorporates data from both the dideoxy sequencing of the RNA from the 3′ terminus and from the chemical sequence of the internal restriction fragments. Three fragments from DI-1 contain 26S sequences. Fragment F is identical (except for a single base change at position 141) to the *Hae*III fragment of 26S RNA, 334 nucleotides long, located in the 26S sequence between nucleotide 1090 and nucleotide 1424 (14a). DI-1 fragment E and DI-1 fragment I, on the other hand, are composite fragments of 26S and non-26S sequence. Fragment E contains at its 5′ end part of the non-26S fragment 410 discussed in detail below (see also Fig. 5B), linked to 57 nucleotides found between nucleotides 1032 and 1089 of the 26S sequence, which is located directly adjacent to fragment F in 26S RNA. Fragment I, on the other hand, begins at its 5′ end with the 39 nucleotides of 26S sequence (nucleotides 1425 to 1463) adjacent to the other end of fragment F and terminates with 120 nucleotides of non-26S sequence which are common to fragment C1 from DI-1 (see Fig. 5B). There is no direct evidence that the three DI-1 fragments (E, F, and I) actually are joined together and are found in the same molecule, but the direct alignment with known 26S sequence suggests that they could be contiguous segments in the DI sequence.

Most of the remaining *Hae*III fragments which were sequenced contain some sequences in common, and the organization of these is shown in Fig. 5B. The top line shows two adjacent *Hae*III fragments from non-26S regions of the 49S virion RNA (called 410 and 450 here)

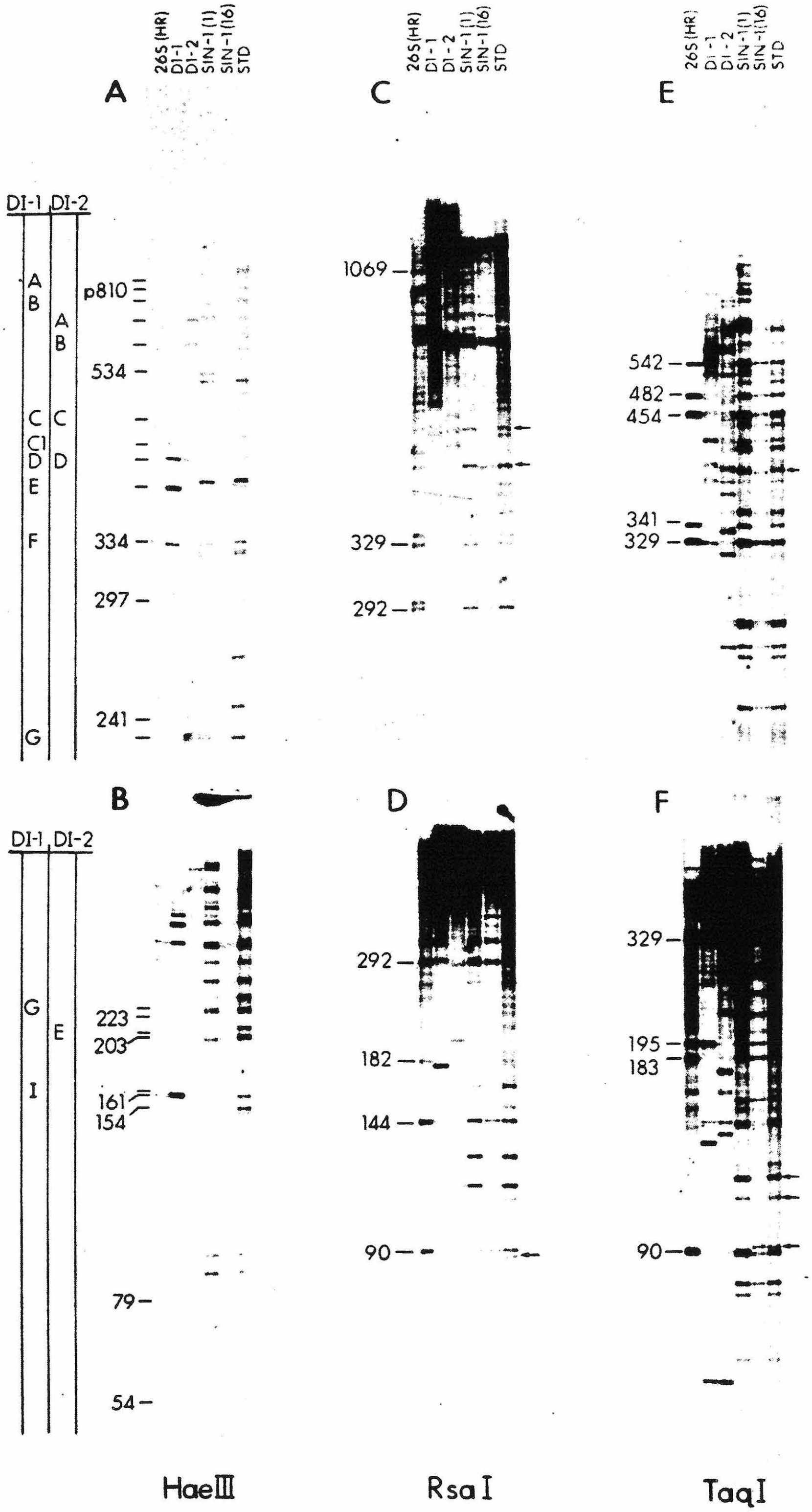


FIG. 3. Acrylamide gel analysis of restriction fragments of cDNA's transcribed from Sindbis virus RNAs. The conditions for synthesis of cDNA by use of random primers and subsequent cleavage with restriction enzymes are described in the text. Samples of the fragments were resolved on 40-cm 5% sequencing gels made in 100 mM Tris-borate, 2 mM EDTA, and 8.3 M urea. The gels were run at 40 W until the xylene cyanol dye had migrated 60 cm (upper panels) or 20 cm (lower panels). The lengths of fragments from 26S cDNA which were sequenced previously are indicated. *Hae*III fragments from the DI cDNA's which were subjected to direct sequencing are identified with letters. Fragment K ran off this gel and was identified from other gels. Arrows in each panel indicate fragments from wild-type Sindbis cDNA missing from variant cDNA.

```
A
FRAGMENT                                  SEQUENCE

DI-1  A                                              CCTTT TACTGTGTCA GTAACTTTGC ATAGCAAGAA GCCCTCGCTA TTGTGTGTAA CCGCGTATCC
      B                                              ..... .......... .......... .......... .......... .......... ..........
      C1                                             ..... .......... .......... .......... .......... .......... ..........
      D                                              ..... .......... .......... .......... .......... .......... ..........
      E                                              ..... .......... .......... .......... .......... .......... ..........
      G       GGATGTACGT GCACACAGGG AACGATACCC GTTCT ..... .......... .......... .......... .......... .......... ..........
DI-2  A                                              ..... .......... .......... .......... .......... .......... ..........
      B                                              ..... .......... .......... .......... .......... .......... ..........
      D                                              ..... .......... .......... .......... .......... .......... ..........

DI-1  A       CACGGTTTCT CCCGTGATCC CGGGACTGAT GGTGATTTTC  TTCACTACGT AGCCTTCGCA ACTCACCACT GTATCACAGC GGCAAGTGTA CGACTGCTTT
      B       .......... .......... .......... ..........  .......... .......... .......... .......... .......... ..........
      C1      .......... .......... .......... ..........  .......... .......... .......... .......... .......... ..........
      D       .......... .......... .......... ..........  .......... .......... .......... .......... .......... ..........
      E       .......... .......... .......... ..........  .......... .......... .......... .......... .......... ..........
      G       .......... .......... .......... ..........  .......... .......... .......... .......... .......... ...
DI-2  A       .......... .......... .......... ..........  .......... .......... .......... .......... .......... ..........
      B       .......... .......... .......... ..........  .......... .......... .......... .......... .......... ..........
      D       .......... .......... .......... ..........  .......... .......... .......... .......... .......... ..........

DI-1  A       CCATTCAAGT GGAACACCGA TGGAAGATGC CAGCTCTGCA  AGCTGGCCTG TGTTCTGGAT AAAGTGTCGA TCCTACGGAG AAATAAACCC GCGACCCGGG
      B       .......... .......... .......... ..........  .......... .......... .......... .......... .......... ..........
      C1      .......... .......... .......... ..........  .......... .......... .......... .....GCCCA TTTGCTGAAC CCTTGTGCTA
      D       .......... .......... .......... ..........  .......... .......... .......... .......... .......... ..........
      E       .......... .......... .......... ..........  .......... .......... .......... .......... .......... ..........
DI-2  A       .......... .......... .......... ..........  .......... .......... .......... .......... .......... ..........
      B       .......... .......... .......... ..........  .......... .......... .......... .......... .......... ..........
      D       .......... .......... .......... ..........  .......... .......... .......... .......... .......... ..........

DI-1  A       CTTCAACTCC TTCTTCCTCA TTATCGACAA TTTTCCTGTC  CTACCTTCAC TCAGCTTTGT GCTGCAAAGT CCGATGTTAC GCGCTTCAAG GACTTTCTCG
      B       .......... .......... .......... ..........  .......... .......... .......... .......... .......... ..........
      C1      TGATCGGCAG AAGGTAATTT TGCATGGTGT TGGTGTTCCT  GTTAGTCCTA CCGTTAATGA CAATTCGCTG GTTGAGCCCA ACCAGAAGTT TTTGTGCATC
      D       .......... .......... .......... ..........  .......... .......... .......... .......... .......... ..........
      E       .......... .........C AAACGCTCCT GGTTTCATCG  CTCCATATTC CGGGAAGTCA TAGTTGTAC
DI-2  A       .......... .......... TTCGGCACGC GTGTTGCAGG  TAACATCGTT GTGAAAGCAG AGCGATGGTG TTTCAGCATC CGGCGTATCA AGTACGGTCC
      B       .......... .......... .......... ..........  .......... .......... .......TCC GGCGTATCAA GTACGGTC
      D       .......... .......... .......... ..........  .......... .......... .......... .......... .......... ..........

DI-1  A       TCGGCCYAGY YGGYGYYGYA
      C1      GTCAGGTGAT ATATCC
      D       ....
DI-2  A       GGAGATCCTT AATCTTCTCA TGCAAGTTCT TGTTTGTAAT  CTTGCACGCT TTTTCCGCCA GTTTACTGGC TATTTCATCA TGCGGTCCGG GTCTTCTGGA
      D       ....

B
FRAGMENT                                  SEQUENCE

DI-1  C       AAAAACATGA ACTGGGTGGT GTCGAAGCCA ATCCAGTACA  GGGTCCGCAC ACCTTCCATA GCCTGATGAT AGATAGTTCC GGGAGCGTTG YATATACGTC
DI-2  C       .......... .......... .......... ..........  .......... G....T.... .......... .......... .......... ..........

DI-1  C       CTGCATGACG GAATATTCGG CACGCATGTT GCAGGTAACA  TCGTTGTGAA AGCAGAGCGA TGGTGTTTCA GCATCCGGCC TATCAAGTAC GGTCCGGAGA
DI-2  C       .......... .......... .......... ..........  .......... .......... .......... .......... .......... ..........

DI-1  C       TCCTTAATCT TCTCATGCAA GTTCTTGTTT GTAATCTTGC  ACGCTTTTTC CGCCAGTTTA CTGGG
DI-2  C       .......... .......... .......... ..........  .......... .......... .....

DI-1  F       GATGCTGAAA TTGGTCCAGC TATGACTTTC AAGTCTTTAG  ACGTTCCTGG TGTGACTCCG TTCACGTACA CATCTAGGAA ACTGGTAGTG TTCCCGTACA
              CTATACGCAG TCCTACTTTC ATCGCGGCAG TGTGCACCTT  AATCGCCTGC GCGTGGTCAG ACGCGCATAC TGCTGACAAT TCGACGTACG CCTCACTCAT
              CTGGCTGTTC TCACTGTCGC AAAAACATTG CGCTCCTCCC  CACATAAAGG GGTAGACCCC TCCGAAGACC TTGCAGGTAT AGTCTGCATG

DI-1  I       GCTGACATTC CAAGGAGCCG CAGCATTTGA TTTTTGCTAT  GATCGGCAGA AGGTAATTTT GCATGGTGTT GGTGTTCCTG TTAGTCCTAC CGTTAATGAC
              AATTCGCTGG TTGAGCCCAA CCAGAAGTTT TTGTGCATCG  TCAGGTGATA TATCCG

DI-1  K       TTATACCAGT CATCTGATCG CATATGGTGA
DI-2  K       .......... .......... ..........

DI-2  E       GATGCGAAA  ATGCTCTGGC ATTAGCATGG TCATTTGGAG  TGACCTGCTG TGCTACTACC TCAAATTGCG GGAAGCTTTT TTGCAGTTGC ACGACAAACG
              AACTCTGGGG GTCTACGTCT ACGTTTACTA CTGGCTTCTC  CATTGTGATG GTAGTGCAAT TGGTCGAC
```

FIG. 4. 5′-Terminal sequences of restriction fragments generated by *Hae*III cleavage of DI cDNA's. The fragments are lettered as indicated in Fig. 3. The sequences were determined by chemical cleavage and analysis on acrylamide gels (11, 18). (A) The sequences of fragments with shared terminal regions are aligned to show the extent of homology. A dot indicates that the base is identical to the base listed for DI-1 fragment A. The ambiguous letter code is described in the legend for Fig. 2. Sequences of DI-1 fragments A, B, C1, D, and E and DI-2 fragments A, B, and D were determined beginning approximately 35 bases from the 5′ end. (B) The sequences of the remaining fragments analyzed from the DI cDNA's. The single base change between 26S cDNA and DI-1 fragment F is underlined.

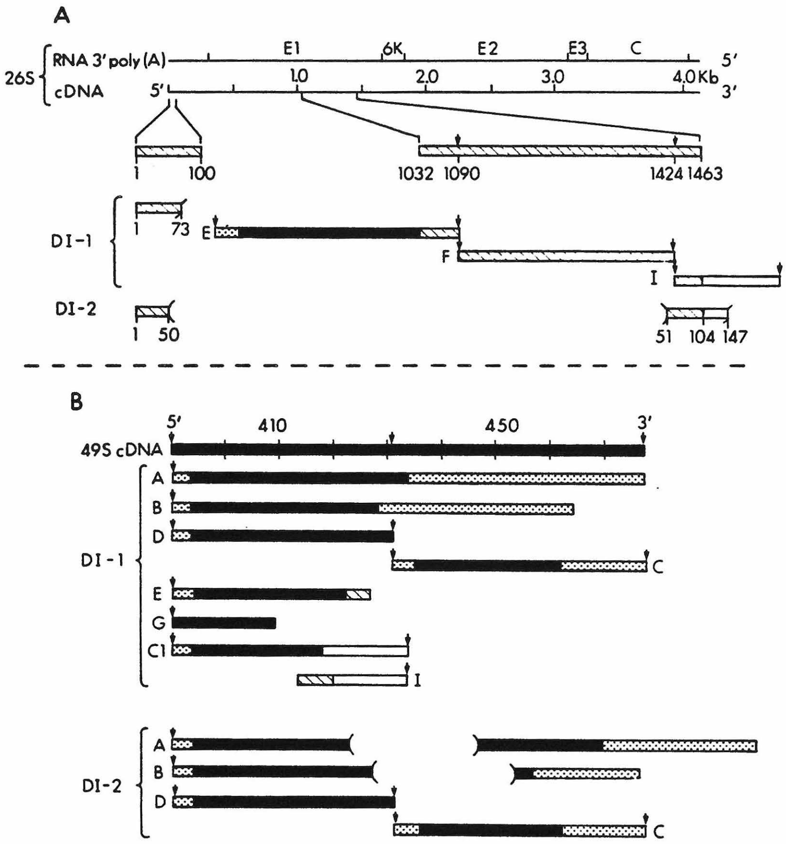


FIG. 5. Model for the sequence organization of Sindbis DI cDNA's. (A) Organization of the 5' region of the DI cDNA's based on chain termination sequencing and analysis of *Hae*III fragments. The 26S RNA and cDNA are shown at the top of the figure. Sequences are numbered from the 5'-terminal G residue of the cDNA. *Hae*III fragments from DI-1 cDNA are identified by letters. The order of the structural protein genes in 26S RNA is as follows: C, capsid; E3, E2, and E1, viral glycoproteins (17). (B) Organization of several internal *Hae*III fragments of DI cDNA which contain sequences in common. At the top are two contiguous *Hae*III fragments, 410 and 450 nucleotides in length, respectively, present in the non-26S region of HR 49S cDNA. Several *Hae*III fragments of cDNA to DI-1 and DI-2 RNAs are diagrammed below. The following designations are used for both A and B: cross-hatched bars, sequences found in 26S cDNA; solid bars, sequences in the nonstructural (non-26S) part of 49S cDNA homologous with fragments 410 and 450; open bars, other non-26S sequences; stippled bars, regions of DI fragments for which sequence data were not obtained. The vertical arrows indicate *Hae*III cleavage sites in the cDNA.

whose sequence is known (E. Strauss, unpublished data) but whose location in the genome is unknown. The region defined by these two fragments is 860 nucleotides long or approximately 10% of the non-26S portion of the Sindbis genome. Fragments C and D (of both DI-1 and DI-2) are the same as 450 and 410, respectively, beginning and ending at legitimate *Hae*III sites, and they appear to represent unaltered sequences in this region. Fragment A of DI-1 is a partial digestion product containing both fragments (probably in their entirety). Four other of the most prominent *Hae*III fragments of DI-1 (fragments B, C1, E, and G) as well as two fragments from DI-2 (fragments A and B) begin with the same sequence (the 5' end of fragment 410). Fragment B of DI-1 is shorter than A and probably contains a deletion whose end points are unknown. Fragment C1 of DI-1 contains the common 5' sequence plus the same sequence found attached to 26S sequence in fragment I. This sequence is present at an unknown location in 49S cDNA (E. Strauss, unpublished data). Fragment E of DI-1 is a rearrangement of non-26S and 26S sequences as described above. DI-2 fragments A and B are simple deletions of 241 and 245 nucleotides, respectively.

The other fragments which were sequenced, E from DI-2 and K from both DI populations, contain sequences found in the non-26S regions of the 49S sequence (E. Strauss, unpublished data). Thus, all of the DI sequence obtained so far can be identified in the 49S RNA of Sindbis virus, and no host or nonviral sequences have been found. This is consistent with earlier studies with nucleic acid hybridization which showed that Sindbis DI RNA can be completely protected from RNase by hybridization to RNA complementary to virion 49S RNA (6).

DISCUSSION

The 3'-terminal sequences of the RNAs of several different alphaviruses show strong conservation of the first 20 nucleotides, suggesting that there are evolutionary pressures resulting in conservation of this region (12; J.-h. Ou, unpublished data). This 3'-terminal conserved sequence was also found in two DI RNA populations of Sindbis virus and two variants of this virus obtained from persistently infected BHK cells, supporting the conclusion that these sequences play an essential role in virus replication.

The variant [Sin-1(16)] from persistently infected cells which is resistant to interference by DI particles (derived from wild-type virus) had two base changes in the 3'-terminal sequence. One is at position 7 within the sequence which is most highly conserved among related alphaviruses, and the other is at position 136. If the first 20 nucleotides form a replicase binding site and if interference involves competition for replicase molecules, then the change from A to G at position 7 may be important for the phenotype of this variant. However, resistance to interference must be more complex, since variant virus cloned from cells after 1 month of persistent infection had 3' sequences identical to those of the wild-type parental strain, yet showed some resistance to interference.

The RNA obtained from the two DI strains had sequences identical to Sindbis HR or wild-type virus for at least the first 50 nucleotides adjacent to the 3' poly(A). Both DI RNA populations were heterogeneous, but since only one terminal sequence was obtained all of these species must contain the standard 3' sequence (within the limits of detection). The two DI RNA populations diverge from the wild-type sequences at different locations (Fig. 2), demonstrating that deletions can occur at more than one site during the generation and subsequent evolution of DI RNAs. Since DI-1 was generated by passing Sindbis virus in BHK cells and DI-2 was generated by passing the virus in chicken cells, it is possible that the host cell might influence the exact site or extent of deletions.

In the case of DI-2 the first (3'-terminal) deletion comprises 1,359 nucleotides and extends from position 51 through 1409 of the HR 26S RNA sequence. Following a stretch of 54 nucleotides (positions 51 to 104 of the DI RNA), there is a second deletion which jumps into non-26S sequence (Fig. 5A). In the case of DI-1, the 3'-terminal sequence was identical to the HR sequence until nucleotide 73 from the 3' poly(A), where it became heterogeneous. It will be necessary to sequence the individual DI RNA species to determine whether the heterogeneity is due to only part of the DI population diverging at position 73 and others diverging later or whether all diverge at position 73 but restart at different places in the wild-type genome. In this regard we note that both DI RNA populations lack a specific 3' oligonucleotide which we believe extends from nucleotides 117 to 148 of the wild-type sequence (2, 14). Thus, the entire DI-1 population must diverge by nucleotide 148, and in all cases the 3'-terminal deletion starts within the relatively narrow region between nucleotides 50 and 148 of the 26S RNA sequence.

It was surprising that so many of the *Hae*III fragments from both DI cDNA's contain different lengths of identical sequences. These sequences correspond to sequences identified in the nonstructural region of 49S RNA. They are not at the 5' end of the genome, but their exact location has not yet been determined (J.-h. Ou and E. Strauss, unpublished data). Some of these fragments must have been transcribed from different RNA molecules or from different regions of the same molecule. For example, fragments E and C1 from DI-1 cannot be derived from any of the larger fragments and fragments A, B, and D from DI-2 are distinct (Fig. 5B). The DI RNAs are heterogeneous, and it is possible that the fragments are derived from different DI RNA molecules. An alternative explanation—that they represent repeated sequences in the same molecule—is supported by the observation that repeated sequences are present in at least one Semliki Forest DI RNA (10a).

A comparison of sequences present in the two different DI cDNA's indicates that there may be some other regions of special significance. The DI-2 sequences from 66 to 147 and DI-1 fragment I are identical. They both leave the 26S sequence at the same nucleotide and enter the same non-26S sequence at the same nucleotide. This suggests that the two DI populations may follow a similar pathway of evolution. The same non-26S sequence (beginning, however, at a different point) is also found in DI-1 fragment C1 attached to the parental sequence of fragment 410 (Fig. 5B). Furthermore, two fragments diverge near nucleotide 321 of the wild-type sequence of fragment 410 (Fig. 5B). DI-1 fragment E diverges at this nucleotide to 26S sequence, indicating a rearrangement, whereas DI-2 fragment A diverges closely nearby in a simple deletion. It is possible that some regions of the RNA may have a secondary structure which facilitates deletions.

In conclusion, the structure of the DI RNA and the relation of the DI sequence to the virion genome are more complex than was first envisaged. It is clear the DI RNA cannot arise solely from large deletions in the middle of the genome, retaining only the 3′ and 5′ ends. This result is in agreement with T_1 oligonucleotide studies which showed that Sindbis DI RNAs contain multiple deletions (2). However, the 3′ termini of both DI RNAs are conserved, and the entire DI RNA sequence appears to be derived from the plus strand of the viral 49S RNA. Whether the conserved termini represent replicase binding sites or whether they are important for a three-dimensional configuration essential for replication or encapsidation, or both, cannot be determined at this time. Similarly, there is conservation (and perhaps amplification) of one region of non-26S sequence and conservation of a short stretch of 26S sequence. It is possible that these conserved sequences could also be important for DI function. One possible hypothesis is that they could serve as binding sites for capsid protein during encapsidation. Further sequence studies on these and on related DI RNAs will be necessary to resolve these questions.

ACKNOWLEDGMENTS

We thank T. Hunkapiller for the computer program which made it possible to determine the sequence homologies reported here and Barbara Weiss for encouragement, critical comments, and help in preparing the RNA.

This research was supported by Public Health Service grants AI 11377 (to S.S.) and AI 10793 (to J.H.S.) from the National Institute of Allergy and Infectious Diseases, GM 06965 (to J.H.S.) from the National Institute of General Medical Science, and P30CA16217 (to S.S.) from the National Cancer Institute, and by grant PCM-8022830 (to J.H.S.) from the National Science Foundation.

LITERATURE CITED

1. **Carmichael, G. G., and G. K. McMaster.** 1980. The analysis of nucleic acids in gels using glyoxal and acridine orange. Methods Enzymol. **65:**380–391.
2. **Dohner, D., S. Monroe, B. Weiss, and S. Schlesinger.** 1979. Oligonucleotide mapping studies of standard and defective Sindbis virus RNA. J. Virol. **29:**794–798.
3. **Garoff, H., A.-M. Erischauf, K. Simons, H. Lehrach, and H. Delius.** 1980. Nucleotide sequence of cDNA coding for Semliki Forest virus membrane glycoproteins. Nature (London) **288:**236–241.
4. **Garoff, H., A.-M. Frischauf, K. Simons, H. Lehrach, and H. Delius.** 1980. The capsid protein of Semliki Forest virus has clusters of basic amino acids and prolines in its amino terminal region. Proc. Natl. Acad. Sci. U.S.A. **77:**6376–6380.
5. **Guild, G. M., and V. Stollar.** 1975. Defective interfering particles of Sindbis virus. 3. Intracellular viral RNA species in chick embryo cell cultures. Virology **67:**24–41.
6. **Guild, G. M., and V. Stollar.** 1977. Defective interfering particles of Sindbis virus. 5. Sequence relationships between SV_{STD} 42S RNA and intracellular defective viral RNAs. Virology **77:**175–188.
7. **Holland, J. J., S. I. T. Kennedy, B. L. Semler, C. L. Jones, L. Roux, and E. A. Grabau.** 1980. Defective interfering RNA viruses and the host-cell response, p. 137–192. *In* H. Fraenkel-Conrat and R. R. Wagner (ed.), Comprehensive virology, vol. 16. Plenum Press, New York.
8. **Kennedy, S. I. T.** 1976. Sequence relationships between the genome and the intracellular RNA species of standard and defective-interfering Semliki Forest virus. J. Mol. Biol. **108:**491–511.
9. **Kennedy, S. I. T., C. J. Bruton, B. Weiss, and S. Schlesinger.** 1976. Defective interfering passages of Sindbis virus: nature of the defective virion RNA. J. Virol. **19:**1034–1043.
10. **Laskey, R. A.** 1980. The use of intensifying screens or organic scintillators for visualizing radioactive molecules resolved by gel electrophoresis. Methods Enzymol. **65:**363–371.

10a. **Lehtovaara, P., H. Söderlund, S. Keränen, R. F. Pettersson, and L. Kääriäinen.** 1981. 18S defective interfering RNA of Semliki Forest virus contains a triplicated linear repeat. Proc. Natl. Acad. Sci. U.S.A. **78:**5353–5357.

11. **Maxam, A. M., and W. Gilbert.** 1980. Sequencing end-labeled DNA with base-specific chemical cleavages. Methods Enzymol. **65:**499–560.
12. **Ou, J.-h., E. G. Strauss, and J. H. Strauss.** 1981. Comparative studies of the 3′-terminal sequences of several alphavirus RNAs. Virology **109:**281–289.
13. **Pettersson, R. F.** 1981. 5′-Terminal nucleotide sequence of Semliki Forest virus 18S defective interfering RNA is heterogenous and different from the genomic 42S RNA. Proc. Natl. Acad. Sci. U.S.A. **78:**115–119.
14. **Rice, C. M., and J. H. Strauss.** 1981. Nucleotide sequence of the 26S mRNA of Sindbis virus; deduced sequence of the encoded viral structural proteins. Proc. Natl. Acad. Sci. U.S.A. **78:**2062–2066.

14a. **Rice, C. M., and J. H. Strauss.** 1981. Synthesis, cleavage and sequence analysis of DNA complementary to the 26S messenger RNA of Sindbis Virus. J. Mol. Biol. **150:**315–340.

15. **Sanger, F., and A. R. Coulson.** 1978. The use of thin acrylamide gels for DNA sequencing. FEBS Lett. **87:**107–110.
16. **Sanger, F., S. Nicklen, and A. R. Coulson.** 1977. DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. U.S.A. **74:**5463–5467.
17. **Schlesinger, M. J., and L. Kääriäinen.** 1980. Translation and processing of alphavirus proteins, p. 371–392. *In* R. W. Schlesinger (ed.), The togaviruses. Academic Press, Inc., New York.
18. **Smith, D. R., and J. M. Calvo.** 1980. Nucleotide sequence of the *E. coli* gene coding for dihydrofolate reductase. Nucleic Acids Res. **8:**2255–2274.
19. **Stark, C., and S. I. T. Kennedy.** 1978. The generation and propagation of defective-interfering particles of Semliki Forest virus in different cell types. Virology **89:**285–299.
20. **Strauss, J. H., and E. G. Strauss.** 1977. Togaviruses, p. 111–166. *In* D. P. Nayak (ed.), The molecular biology of animal viruses, vol. 1, Marcel Dekker, Inc., New York.
21. **Weiss, B., R. Rosenthal, and S. Schlesinger.** 1980. Establishment and maintenance of persistent infection by Sindbis virus in BHK cells. J. Virol. **33:**463–474.
22. **Weiss, B., and S. Schlesinger.** 1981. Defective interfering particles of Sindbis virus do not interfere with the homologous virus obtained from persistently infected BHK cells but do interfere with Semliki Forest virus. J. Virol. **37:**840–844.

Chapter 5

This chapter has been prepared for submission to the Proc. Natl. Acad. Sci. USA.

Sequence studies of several alphavirus genomic RNAs in the region containing the start of the subgenomic RNA

(Sindbis virus/RNA sequencing/sequence homology/nonstructural proteins/transcription initiation)

Jing-hsiung Ou, Charles M. Rice, Lynn Dalgarno,[1] Ellen G. Strauss and James H. Strauss[2]

Division of Biology
California Institute of Technology
Pasadena, California 91125

[1] Present address: Biochemistry Department
The Australian National University
P.O. Box 4
Canberra, Australia 2600

[2] To whom reprint requests should be addressed.

ABSTRACT The sequence of the region of the 49S genomic RNA which contains the 5'-end of the subgenomic 26S RNA and the 5'-flanking sequences in 49S RNA were determined for several alphaviruses. A highly conserved sequence of 21 nucleotides was found which includes the first two nucleotides of 26S RNA and the 19 nucleotides preceding it. We propose that the complement of this sequence in the minus strand is the recognition site used by the viral transcriptase for initiation of transcription of 26S RNA. The COOH-terminal sequence of the nonstructural polyprotein precursor, which is translated from 49S RNA, has been deduced for each virus. These protein sequences are highly homologous, but the particular triplets used for a given amino acid have diverged markedly between viruses, indicating that virus evolution is quite rapid at the nucleotide level. Clusters of in phase stop codons for the nonstructural polyprotein were found in the nontranslated region of 26S RNA in each case. The length of untranslated sequence at the 5'-end of 26S RNA is between 48 and 51 nucleotides, depending on the virus.

Alphaviruses compose a group of about 20 enveloped viruses which are able to replicate in both invertebrates and vertebrates. The alphavirus genome is a single-stranded RNA molecule of plus polarity which has a sedimentation coefficient of 49S and a molecular weight of about 4.2×10^6 (reviewed in 1). After infecting host cells, the genomic RNA is released from the virion and serves as messenger for the nonstructural proteins required to replicate the viral RNAs. It then serves as a template for producing a minus-stranded, full-length RNA. This minus-stranded RNA in turn serves as the template for producing more plus-stranded 49S RNA and, in addition, as a template for transcription of a subgenomic 26S RNA which is identical to the 3'-terminus of 49S RNA (2, 3). This 26S RNA is the message for structural proteins (4, 5). We are interested in the mechanisms involved in transcription of this subgenomic RNA, and have found that a comparative approach to the study of alphavirus RNA sequences is useful in elucidating important control regions in the RNA (3, 6, 7). In this paper, we extend this comparative approach to the region of the genomic RNA which contains both the 5'-end of 26S RNA and the 5'-flanking sequences in 49S RNA, which we will refer to as the junction region, in order to study the transcriptional control of 26S RNA.

MATERIALS AND METHODS

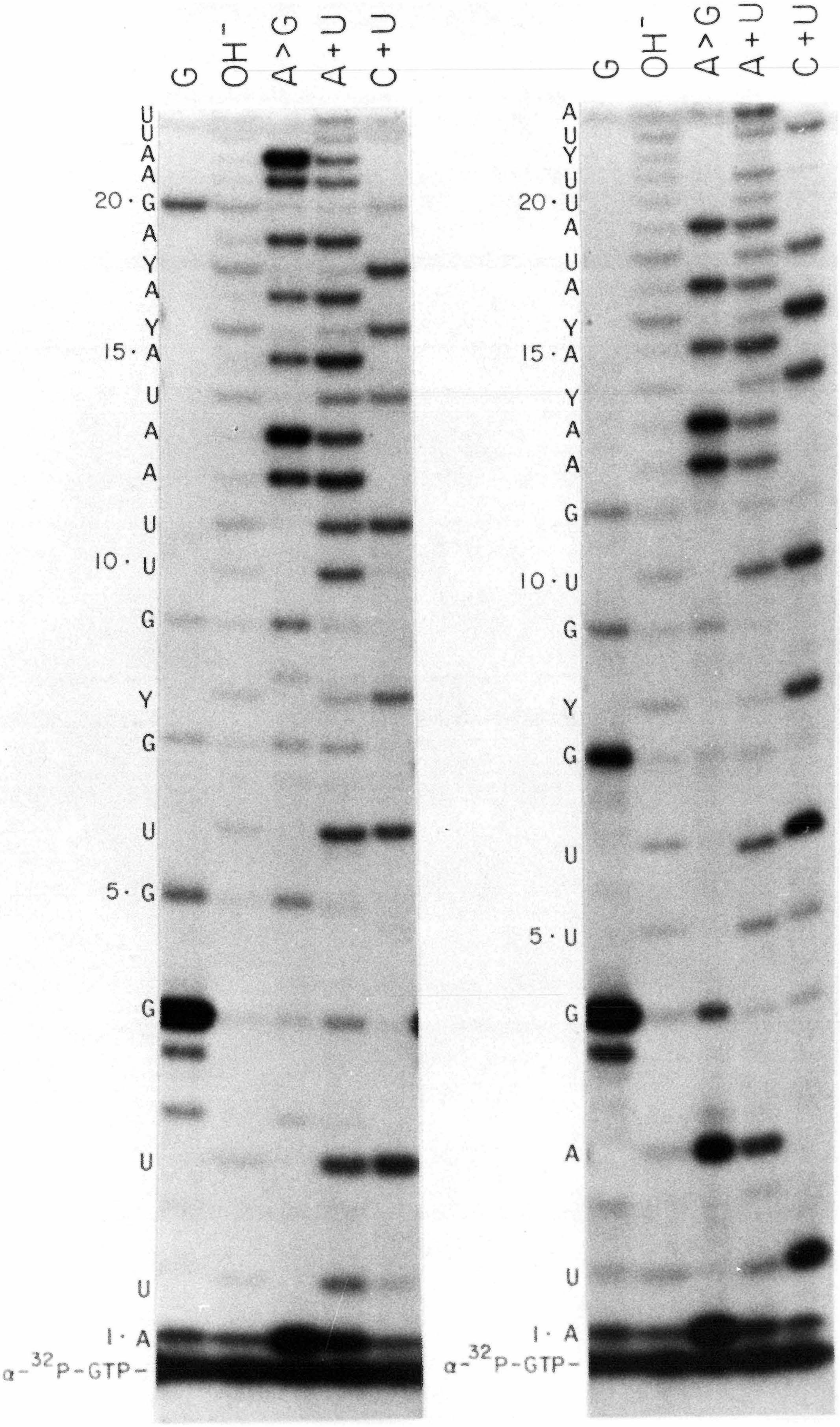
Preparation of vaccinia guanylyl transferase. Guanylyl transferase was prepared from vaccinia virus (WR strain, isolate 11; a gift from Dr. W. K. Joklik) by simplifying the method of Paoletti et al. (8). 2 mg of vaccinia virus in 640 μl 1 mM Tris-HCl, pH 9.0, were used for each preparation of the enzyme. After disruption of the virus cores, the preparation was briefly sonicated to shear viral DNA, followed by centrifugation at 136,000 xg for 60 min. The guanylyl transferase activity present in the supernatant was stable for 2 weeks at 4°C.

5'-end labeling of alphavirus 26S RNA. The 26S RNA of Sindbis virus (HR strain), Semliki Forest virus (SFV), and Middelburg virus were prepared as previously described (3). 30 µg of 26S RNA were decapped by β-elimination using the method of Rose and Lodish (9). Decapped RNA was then end-labeled using conditions modified from Ahlquist et al. (10): the reaction (100 µl) contained 0.16 mM ATP, 1 mM $MgCl_2$, 1 mM DTT, 2.5 µM α-^{32}P-GTP (410 Ci/mmol, Amersham), 50 mM Tris-HCl, pH 7.8 and 15 µl of guanylyl transferase extract. After incubating at 37°C for 15 min, the reaction was terminated by phenol/chloroform extraction followed by ethanol precipitation as previously described (3). RNA pellets were resuspended in 50 µl 10 mM Tris-HCl, pH 8.2, containing 10 mM NaCl, 1 mM EDTA and 0.1% SDS and purified from the unincorporated label by gel filtration using a Biogel A-5M column. The excluded volume fractions containing the RNA were pooled and the RNA was ethanol precipitated twice with carrier RNA and resuspended in water to a final concentration of about 2,000 cpm/µl. More than $1x10^5$ cpm were incorporated per 30 µg of RNA which was sufficient for several experiments.

Direct enzymatic sequencing of the RNA from the 5'-end. Partial ribonuclease digestion of the 5'-end labeled RNA was performed by using the method recommended by P. L. Biochemicals, Inc. (method E998) and the products were separated on sequencing gels. A brief description of the reaction conditions is contained in the legend of Fig. 1.

Single-stranded cDNA synthesis and chemical sequencing. The 49S RNA of Sindbis, Middelburg, and SFV were prepared as previously described (3). Ross River virus (RRV) (T48 strain from R. Shope) 26S and 49S RNAs were prepared from infected BHK cells. Single-stranded cDNA to these RNAs were synthesized with reverse transcriptase using calf thymus DNA or oligo (dT) primers as previously described (11). Preparation, isolation, and chemical sequencing of 5'-end-labeled restriction fragments from these cDNAs has been previously described (11).

FIG. 1. Direct 5'-end sequencing of the 26S RNAs of SFV and Middelburg virus. About 2,000 cpm of end-labeled RNA were enzymatically digested at 56°C for from 3 min to 10 min depending on the degree of digestion preferred. Alkaline hydrolysis was done at 90°C for 6 min. G reaction: 1 µl RNA was added to 3 µl buffer I containing 1 unit RNase T_1; A>G reaction: 1 µl RNA was added to 3 µl buffer II containing 1 unit RNase U_2; A+U reaction: 1 µl RNA was added to 3 µl buffer I containing 1 unit of Physarum M RNase; U+C reaction: 1 µl RNA was added to 3 µl buffer III containing 1 unit of Bacillus cereus RNase; alkaline hydrolysis: 1 µl RNA was added to 2 µl of 50 mM $NaHCO_3$, pH 9.0, containing 1.6 mM EDTA and 1 µg carrier tRNA. All reactions were terminated by chilling on ice. For the U+C reaction and the alkaline hydrolysis 3 µl of loading buffer (0.01% xylene cyanol FF, 0.01% bromophenol blue and 10 M urea in 1/10 x electrode buffer) were added before loading onto the gel. Buffer I: 33 mM Na citrate, pH 5.0, 1.7 mM EDTA, 0.04% xylene cyanole FF, 0.08% bromophenol blue, 1 mg/ml carrier tRNA and 7 M urea; Buffer-II: the same as Buffer-I, except that the pH of Na citrate is 3.5; Buffer III: the same as Buffer I except that it contains no urea or dyes. Acrylamide gel electrophoresis was previously described (3) and the gels were autoradiographed using preflashed film and exposure at -70°C for one week with enhancing screens. Figure shown is a 20% sequencing gel. Y is used for ambiguous pyrimidines. Left: SFV 26S RNA; right: Middelburg 26S RNA.


G
OH$^-$
A>G
A+U
C+U
U
U
A
A
20 · G
A
Y
A
Y
15 · A
U
A
A
U
10 · U
G
Y
G
U
5 · G
G
U
U
1 · A
α-^{32}P-GTP-
G
OH$^-$
A>G
A+U
C+U
A
U
Y
U
20 · U
A
U
A
Y
15 · A
Y
A
A
G
10 · U
G
Y
G
U
5 · U
G
A
U
1 · A
α-^{32}P-GTP-

RESULTS

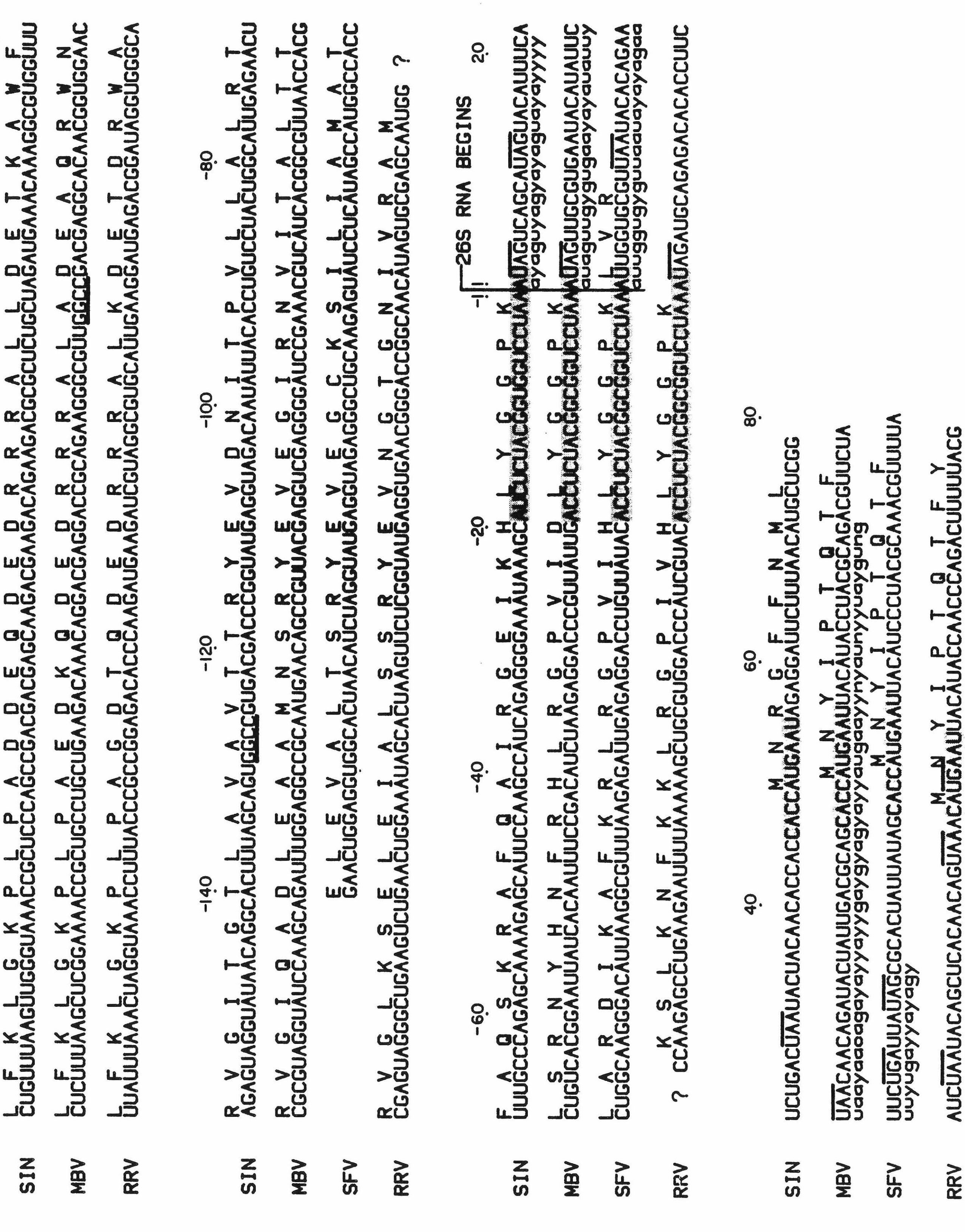
The 5' terminal sequences of the 26S RNAs of Sindbis virus, Middelburg virus, and SFV were obtained as follows: the RNAs were decapped by β-elimination, end-labeled with guanylyl transferase, and then partially digested using endonucleases with 4 different base specificities. The resulting products were separated on acrylamide gels and the sequence determined from the ladder produced. Fig. 1 presents part of the data in order to illustrate the technique. Each RNA was sequenced at least twice, and the sequences obtained are shown in lower case letters in Fig. 2. In general this enzymatic method gives a clean signal for the purines, but the pyrimidines are often ambiguous (Fig 1).

Also shown in Fig. 2 (in upper case letters) are the nucleotide sequences of the RNAs in the junction region deduced from chemical sequencing of single-strand cDNA made to 49S RNA of Sindbis, Middelburg, SFV, and RRV. The method used involved sequencing HaeIII restriction fragments of the cDNA (11) by the methods of Maxam and Gilbert (12). The start of 26S RNA has been identified by alignment with the 5'-RNA sequence data, and the RNA sequence and the cDNA sequence are in agreement. Taken together with the previously published sequences of 26S RNA (4, 13, 14), the complete sequence of both Sindbis and SFV 26S RNA is now known. Sindbis 26S RNA is 4102 nucleotides in length and that of SFV is 4074 nucleotides in length.

Finally, Fig. 2 also contains the deduced amino acid sequences of the COOH-terminal region of the nonstructural polyprotein (translated from 49S RNA) and the NH_2-terminal region of the capsid protein (translated from 26S RNA) (4, 13) (discussed in more detail below).

For Sindbis and Middelburg viruses, restriction fragments produced by HaeIII digestion of cDNA to 49S RNA were randomly selected for sequencing. Identification and alignment of the sequences in the junction region was done by computer. The

FIG. 2. The nucleotide sequence of the junction region. The conserved sequence near the start of 26S RNA and that around the initiation codon of the capsid proteins are shaded. The deduced protein sequences are shown above the nucleotide sequences. Nucleotide sequences of the 5'-end of 26S RNA determined from the enzymatic method are shown in lower case letters. In phase termination codons for nonstructural proteins are overlined. Hae III sites not directly sequenced are underlined. Numbering of the nucleotides begins with the start of 26S RNA; positive numbers are used for nucleotides in the 26S region and negative numbers for the 5'-flanking sequence in 49S RNA. Abbreviations: y, pyrimidines; n, any nucleotide; A, Ala; C, Cys; D, Asp; E, Glu; F, Phe; G, Gly; H, His; I, Ile; K, Lys; L, Leu; M, Met; N, Asp; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp; Y, Tyr; SIN, Sindbis virus; MBV, Middelburg virus.


SIN
MBV
RRV
SFV
-240
-220
-200
-180
-160
-140
-120
-100
-80
-60
-40
-20
-1
1
20
40
60
80
26S RNA BEGINS
?

cDNA sequences were stored in a computer, converted to RNA sense, and translated in all three reading frames. The RNA sequences and the possible protein sequences were searched using homology routines and the fragments which contained the start of 26S RNA and the start of the capsid protein were identified. These fragments end at nucleotide -126 for Sindbis and -180 for Middelburg (numbering begins at the start of 26S RNA) (Fig. 2). Furthermore, because the HaeIII sites are found at different positions in the two viruses, the adjacent HaeIII fragments were identified by homology. This is possible because of the pronounced homology at the protein level among the various alphaviruses, and use of homology routines often allow corresponding fragments in any two viruses to be identified and aligned. Thus the HaeIII sites at the junction of the two Sindbis fragments which join at -126 and that of the two Middelburg fragments which join at -180 were not directly sequenced but were inferred from the activity of HaeIII and the homology between the two RNAs and their translated proteins. (The HaeIII site at Middelburg position -129 was obtained by sequencing a partial fragment which had not been cleaved at this site.)

For SFV, a simplified procedure was used to identify the junction fragment. cDNA to both 26S and 49S RNA were made using an oligo (dT) primer, digested with HaeIII, and the resulting fragments compared on gels. Because of premature termination by the reverse transcriptase the complexity of the 49S pattern was only slightly greater than that of the 26S RNA pattern. From our direct 5'-end sequencing and the previously published 26S RNA sequence (13) it was known that the first HaeIII site in 26S RNA was 83 nucleotides from the cap. Only 12 49S RNA specific fragments larger than 83 nucleotides were found, and these were excised and scanned by chemical sequencing using the G + A reaction to find the junction fragment. A fragment of 222 nucleotides in length was found which contained the right purine ladder for the 5'-end sequence of 26S RNA and its sequence was then determined (Fig. 2).

A third procedure was used for RRV. RNA was prepared from infected cells which sedimented at about 26S and which contained poly(A). This preparation contained not only 26S RNA but also fragments of degraded 49S RNA and thus the sequences in the junction region were present as well as 26S RNA sequences. cDNA was made with calf thymus primer and the HaeIII fragments randomly selected for sequencing. The correct fragments were identified by homology. Six nucleotides of sequence are apparently missing which we believe to be present in a small unsequenced HaeIII fragment. Nonetheless, corresponding sequences can be readily aligned.

Inspection of the junction sequences (Fig. 2) reveals a highly conserved sequence which extends from nucleotides -19 to 2 which we postulate is the recognition signal for the viral transcriptase. Translation of the RNA sequence reveals stop codons for the nonstructural proteins at nucleotides 2-4 for 3 of the viruses and nucleotides 11-13 for SFV. A diagramatic summary of the genomic organization in the junction region is shown in Fig. 3.

DISCUSSION

We have used a comparative approach to study the structure of the junction region of alphavirus 49S RNA. The pronounced homology among the various alphaviruses allows corresponding sequences between any two viruses to be aligned without obtaining the complete RNA sequence of each virus. The homology is much more striking at the protein level than at the RNA level, and the homology in the COOH-terminal sequence of the nonstructural proteins is illustrated in Fig. 4. The homology is striking but nonuniform. Some regions are highly conserved, while other regions are less well conserved. The use of a computer search program to detect such homology is illustrated in Fig. 5. In this figure a dot matrix has been constructed in which the deduced COOH-terminal portions of the nonstructual proteins of Sindbis and Middelburg are compared. A dot is placed in the matrix whenever at least 4 out

FIG. 3. Schematic of the genome of alphaviruses with the junction region expanded. The boxes represent conserved nucleotide sequences found near the start of 26S RNA and around the start codon of the capsid protein. The start of 26S RNA, the stop point of the nonstructural proteins, and the start of the capsid protein are indicated. In phase termination codons are shown by solid circles. The figure is drawn to scale for Sindbis virus.

SINDBIS VIRUS 49S RNA (13000 NT)

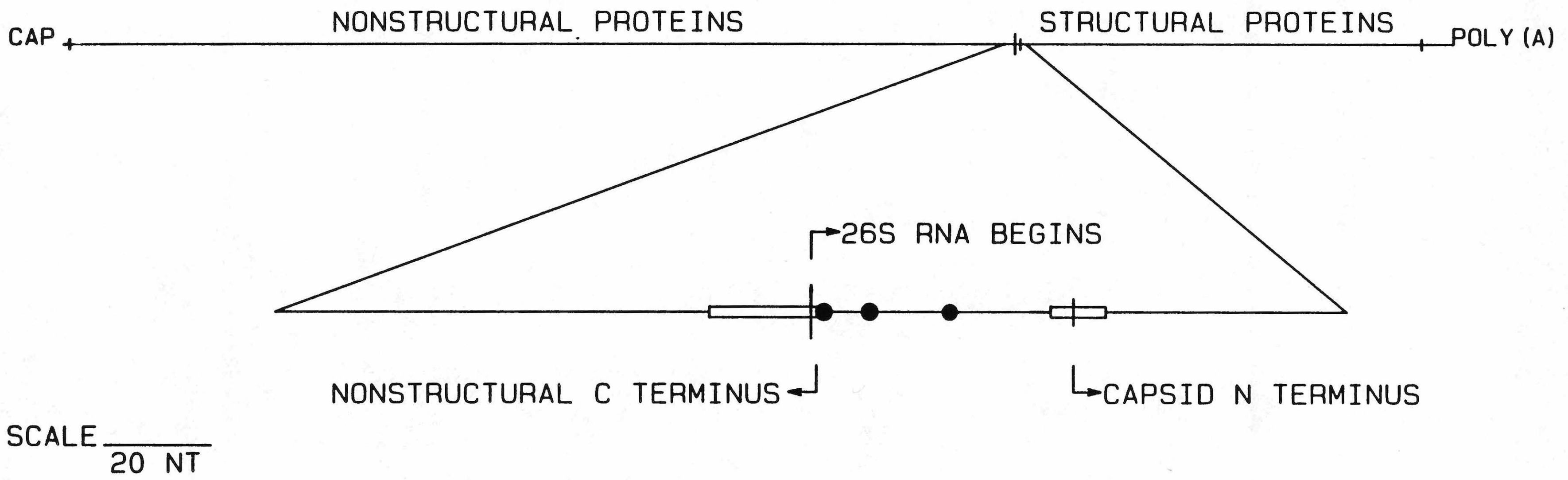
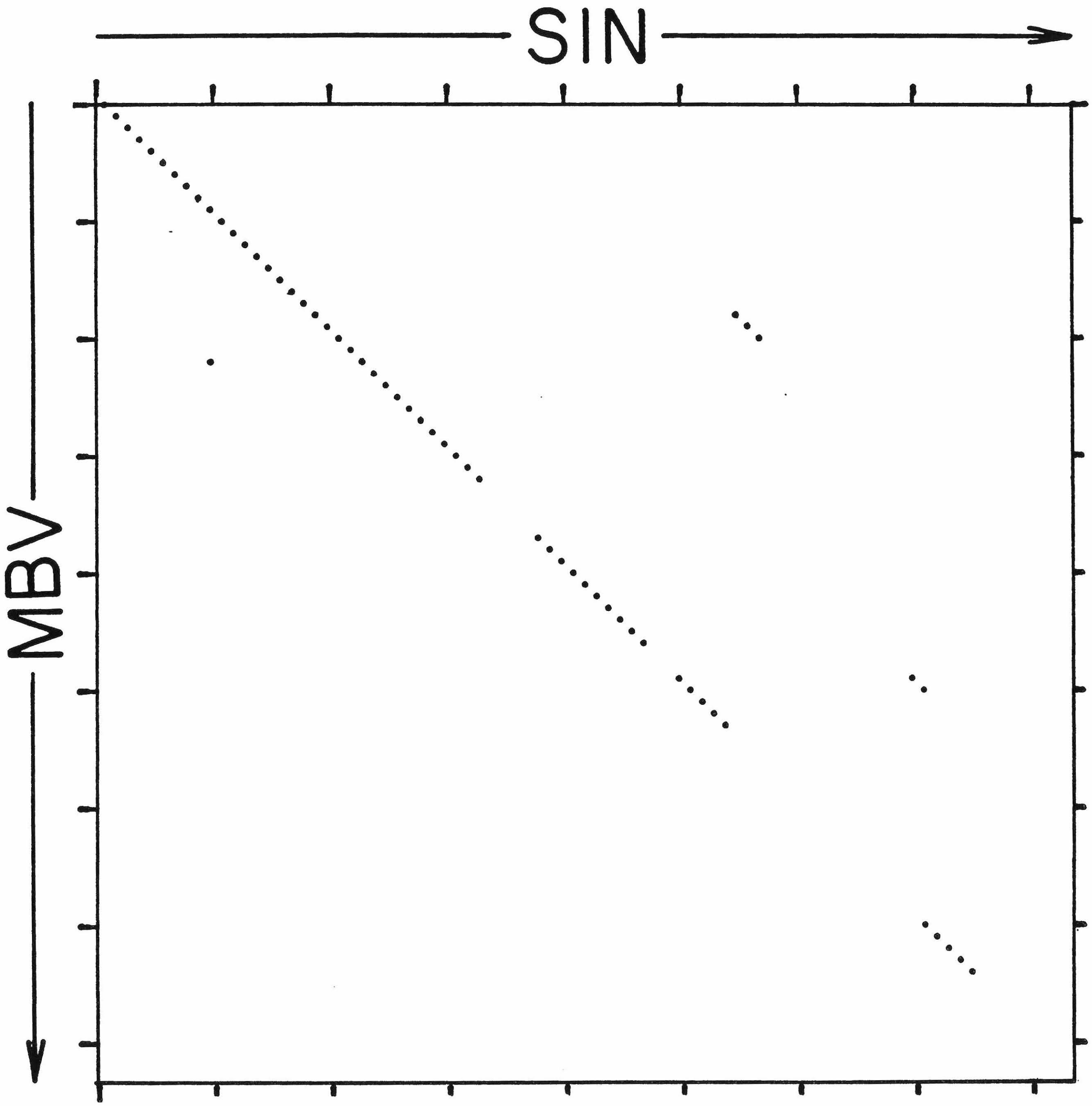


FIG. 4. Comparison of the carboxyl-terminal sequences of the nonstructural proteins of alphaviruses. The single letter amino acid code is used and is defined in the legend to Fig. 2. A dash means the amino acid is the same as is found in Sindbis virus (SIN). MBV is used for Middelburg virus.

```
SIN   LFKLGKPLPADDEQDEDRRRALLDETKAWFRVGITGTLAVAVTTRYEVDNITPVLLALRTFAQSKRAFQAIRGEIKHLYGGPK
MBV   ----------E-K---------A--AQR-N----QAD-EA-MNS----EG-RN-IT--T-LSRNYHN-RHL--PVID------
RRV   ----------G-T---------K---DR-A---LKSE-EI-LSS----NGTGNIVR-M ? -K-LKN-KKL--P-V-------
SFV                                       E-E--L-S----EGCKSI-I-MA-L-RDIK--KRL--PVI-------LVR
```

FIG. 5. Dot matrix comparing the amino acid sequence in the COOH-terminal region of the Sindbis nonstructural polyprotein to that of the corresponding sequence in Middelburg. The amino acid sequences are compared 9 residues at a time and a dot placed in the matrix whenever at least 4 amino acids in the string are identical. The ends of the arrows mark the COOH-termini for the nonstructural polyprotein of Sindbis virus (horizontal axis) and Middelburg virus (vertical axis).


SIN
MBV

of 9 amino acids being compared are identical. The strong diagonal line shows the homology between the two sequences, while the low background in other areas of the matrix illustrates that homologies of even less than 50% can be readily detected. Furthermore the comparative approach yields information on the translation phase of the RNA sequences. Inspection of Fig. 2 reveals that even when the amino acid sequence is conserved, the codons used are not conserved. The evolution of these viruses is so rapid that the third codon position in degenerate codon families is essentially randomized. As examples, note that 3 different codons are used for the Ala at position -78 to -80 and for the Leu at position -135 to -137. A second way to illustrate this point is to note that outside of the conserved sequence near the start of 26S RNA (see below) the Sindbis and Middelburg nonstructural polypeptides shown in Fig. 2 contain the same amino acid at 40 positions, and different codons are used in 25 of these instances. Because of this, translation of corresponding RNA sequences will reveal extensive protein homology only when the proper translation phase is chosen. Thus we feel confident that the deduced amino acid sequences shown are in fact correct and that termination of the nonstructural proteins occurs where shown.

It is of interest that the termination codons of the nonstructural proteins lie in the 26S region of the genomic RNA. Thus the untranslated region between the nonstructural and structural polypeptides is less than 50 nucleotides as the initiation codon of the capsid protein begins at nucleotide 50, 51, 52, and 49 in Sindbis, Middelburg, SFV, and RRV 26S RNAs, respectively. In each case the untranslated region contains a cluster of stop codons which are in phase but slightly separated (3 for Sindbis and SFV, 2 for Middelburg, 4 for RRV (Fig. 2)), implying that multiple stop codons are important. It is also intriguing that Sindbis, Middelburg, and RRV terminate at the same position, whereas the corresponding SFV codon is a sense codon and a later stop codon (corresponding to the second in phase Sindbis stop codon) is used. We note that in vitro translation of 49S RNA does not lead to production of structural proteins (reviewed in 1), as would be expected from these results.

There is conserved stretch of 21 nucleotides near the beginning of 26S RNA, from nucleotide -19 to 2. It is unlikely that conservation of this sequence is caused by a need to conserve the amino acid sequence. As discussed above conservation of protein sequence does not necessitate such high conservation of the nucleotide sequence. We note that in contrast to the situation outside this region cited above in which the same codon is used for only 15 of the 40 conserved amino acids between Sindbis and Middelburg, in this conserved region 5 of 6 amino acids are encoded by the same codon. As another example, we note that the leucine in this region is encoded by CUC in all four viruses, whereas outside this region multiple codons are used for conserved leucine residues (Fig. 2, and see above).

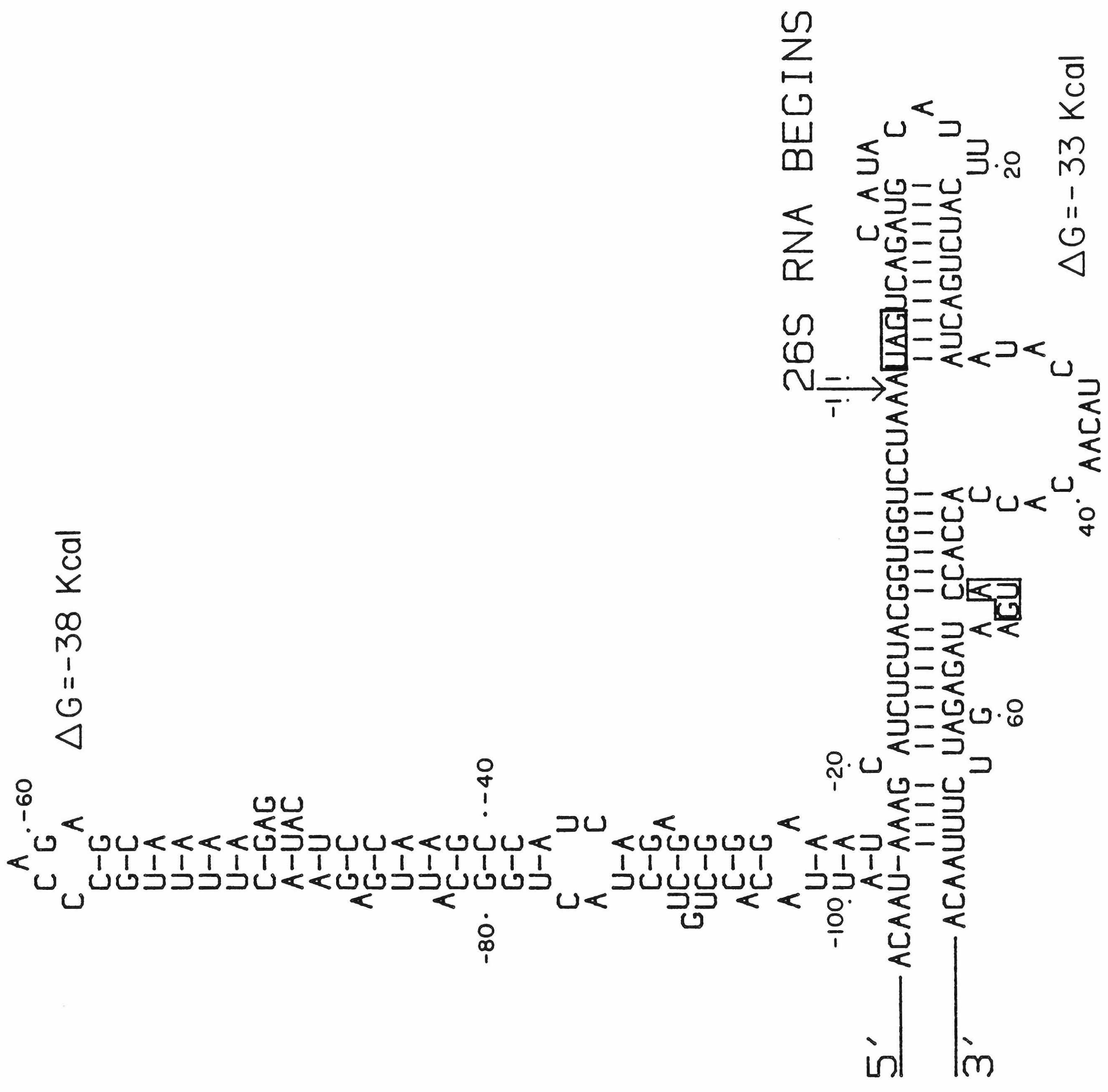
We have previously shown that there is a conserved sequence of 19 nucleotides at the 3'-end of the alphavirus genome and have postulated that this forms a replicase recognition site on this plus strand for production of minus strands (3, 6). Similarly we postulate that the complement of the conserved sequence around the start of 26S RNA forms a recognition site for the viral transcriptase which transcribes 26S RNA from the genomic length minus strand. Because these two recognition sequences are quite distinct and possess no sequence homology, differential control of 26S RNA and minus strand production is possible. We postulate that different virus polypeptides are required for initiation of minus strand synthesis and for 26S RNA synthesis. These two initiation proteins are probably encoded in different complementation groups and could be components of a viral replicase/transcriptase complex. The nonstructural proteins of Sindbis virus have been grouped into four different complementation groups (15), all of which affect RNA synthesis, and 3 or 4 nonstructural polypeptides have been described in infected cells (16). This subject will be discussed in more detail elsewhere (J.-H. Ou, Ph.D. thesis, California Institute of Technology).

A very stable hairpin structure can be formed from nucleotides -99 to 67 in the Sindbis junction region (Fig. 6). It is conceivable that this structure is involved in initiation of 26S RNA synthesis. However, we could not find comparable structures in the other 3 viruses and this structure, despite its stability, may not possess any particular function in the virus life cycle. This result emphasizes the importance of a comparative approach for identification of sequences and structures of importance in the control of virus replication.

Stable hairpin structures ($\Delta G \backsim 12$ Kcal) can also be formed in the 5' untranslated regions of 26S RNA in each of the 4 viruses, beginning just after the cap, and these may be of importance in translation of the RNA. The size and nucleotide sequence of these structures are not conserved, however.

The tetranucleotide CACC preceeds the AUG initiation codon of the capsid protein in Sindbis, Middelburg, and SFV, whereas AAAC proceeds the initiation codon in RRV. An examination of the sequences of 65 eukaryotic mRNAs showed that 58% of initiation codons were preceded by a tetranucleotide beginning with CA and 14% were preceded by one beginning AA. In total the second position of the tetranucleotide was A in 83% of the examples studied. The remaining two positions of the tetranucleotide were more variable, but G is seldom found. We believe that the sequence of this tetranucleotide in general and of the first two nucleotides of it in particular are of importance in the selection of the proper start codon and initiation of protein synthesis. A more complete discussion of the distribution of nucleotides preceding initiation codons in eukaryotic mRNAs will be presented elsewhere (J. H. Ou, Ph.D. Thesis, California Institute of Technology, Pasadena, California).

FIG. 6. Hairpin structures in the junction region of Sindbis RNA. Free energies were calculated by using the method of Tinoco et al. (17). The start of 26S RNA is indicated and the stop codon of the nonstructural polyprotein and the initiation codon of the capsid protein are boxed.


ΔG=-38 Kcal
26S RNA BEGINS
ΔG=-33 Kcal
5′
3′
-60
-40
-20
-80
-100
-1
20
40
60

ACKNOWLEDGEMENTS

We are grateful to Mrs. E. M. Lenches for preparing chick embryo cell cultures and to T. Hunkapiller for assistance with the computer work. This work was supported by Grants PCM 8022830 from the National Science Foundation, GM 06965 and AI 10793 from the National Institutes of Health, and Biomedical Research Support Grant 507RR 07003 from the National Institutes of Health. J.O. was supported in part by Training Grant GM 00086 from the National Institutes of Health and CMR by a fellowship from the Gosney Foundation.

REFERENCES

1. Strauss, J. H. & Strauss, E. G. (1977) in The Molecular Biology of Animal Viruses, ed., Nayek, D. P. (Marcel Dekker, New York), pp. 111-166.
2. Simmons, D. T. & Strauss, J. H. (1972) J. Mol. Biol. **71**, 599-613.
3. Ou, J.-H., Strauss, E. G. & Strauss, J. H. (1981) Virology **109**, 281-289.
4. Rice, C. M. & Strauss, J. H. (1981) Proc. Natl. Acad. Sci. U.S.A. **78**, 2062-2066.
5. Simmons, D. T. & Strauss, J. H. (1974) J. Mol. Biol. **86**, 397-409.
6. Ou, J.-H., Trent, D. & Strauss, J. H. (1982) J. Mol. Biol. (in press).
7. Monroe, S. S., Ou, J.-H., Rice, C. M., Schlesinger, S., Strauss, E. G. & Strauss, J. H. (1981) J. Virol. **41**, 153-162.
8. Paoletti, E., Rosemond-Hornbeak, H. & Moss, B. (1974) J. Biol. Chem. **249**, 3273-3280.
9. Rose, J. K. & Lodish, H. F. (1976) Nature **262**, 32-37.
10. Ahlquist, P., Dasgupta, R., Shih, D. S., Zimmern, D. & Kaesberg, P. (1979) Nature **281**, 277-282.
11. Rice, C. M. & Strauss, J. H. (1981) J. Mol. Biol. **150**, 315-340.
12. Maxam, A. M. & Gilbert, W. (1980) Methods in Enzymology **65**, 499-560.
13. Garoff, H., Frischauf, A.-M., Simons, K., Lehrach, H. & Delius, H. (1980) Proc. Natl. Acad. Sci. U.S.A. **77**, 6376-6380.
14. Garoff, H., Frischauf, A.-M., Simons, K., Lehrach, H. & Delius, H. (1980) Nature **288**, 236-241.
15. Strauss, E. G. & Strauss, J. H. (1980) in The Togaviruses, ed. Schlesinger, R. W. (Academic Press, New York), pp. 393-426.
16. Schlesinger, M. J. & Kaariainen, L. (1980) in The Togaviruses, ed. Schlesinger, R. W. (Academic Press, New York), pp. 371-392.
17. Tinoco, I., Borer, P. N., Dengler, B., Levine, M. D., Uhlenbeck, O. C., Crothers, D. M. and Gralla, J. (1973) Nature **246**, 40-41.

Chapter 6

Comparative studies of the 5'-terminal sequences of several alphavirus genomic RNAs and a family of their defective interfering RNAs

Jing-hsiung Ou, Ellen G. Strauss & James H. Strauss

Division of Biology, California Institute of Technology, Pasadena, California
91125, USA

Abstract

By developing sequencing strategies we have determined the 5'-terminal sequences of several alphavirus genomic RNAs and a family of their defective interfering (DI) RNAs. A highly conserved sequence and several secondary structures formed by the 5'-terminal sequences which might be important in alphavirus replication were discovered. Comparative studies also enabled us to deduce the NH_2-terminal sequences of their nonstructural polyproteins.

Alphaviruses which include Sindbis virus (SIN) and Semliki Forest virus (SFV) are a group of viruses that infect both vertebrates and invertebrates[1]. The alphavirus genome is a single-stranded RNA molecule which has a sedimentation coefficient of 49S and a molecular weight of about 4.2 x 10^6 daltons. After infecting host cells, this genomic RNA is released from the virus into the cytoplasm and serves as the messenger for translating a large precursor polypeptide which is subsequently cleaved to several nonstructural proteins required for RNA synthesis[2]. The genomic RNA then becomes the template for transcribing a minus-stranded, full-lengthed RNA which again serves as the template for producing more genomic RNA as well as a subgenomic 26S RNA (for review, see ref. 3). 26S RNA which is the message for structural proteins[4,5,6] comprises the 3'-one-third of the 49S RNA sequence and is coterminal with it[7,8,9].

From comparative studies we have previously located two highly conserved sequences in 49S RNA[9,10]. The first conserved sequence is located at the extreme 3'-end next to the poly(A) tail and the second conserved sequence precedes the 5'-end of 26S RNA and includes the first two nucleotides of it. We have proposed that the 3'-end conserved sequence is the replicase recognition site for minus-stranded RNA transcription[10,11] and the complement of the second conserved sequence in the minus-stranded RNA is required for 26S RNA transcription[9].

In this communication, we have developed strategies for determining the 5'-terminal sequences of several alphavirus 49S RNAs. Several interesting secondary structures and a highly conserved sequence which might be important in alphavirus replication have been found. Comparative studies also enabled us to deduce the NH_2-terminal sequences of nonstructural polyproteins. DI RNAs of alphaviruses are aberrant genomic RNAs which often contain deleted, repeated and translocated genomic sequences[11,12], but yet retain all elements essential for their replication[13]. Because they are useful tools for studying alphavirus RNA replication, partial 5'-terminal sequences of a family of SIN DI RNAs were also determined.

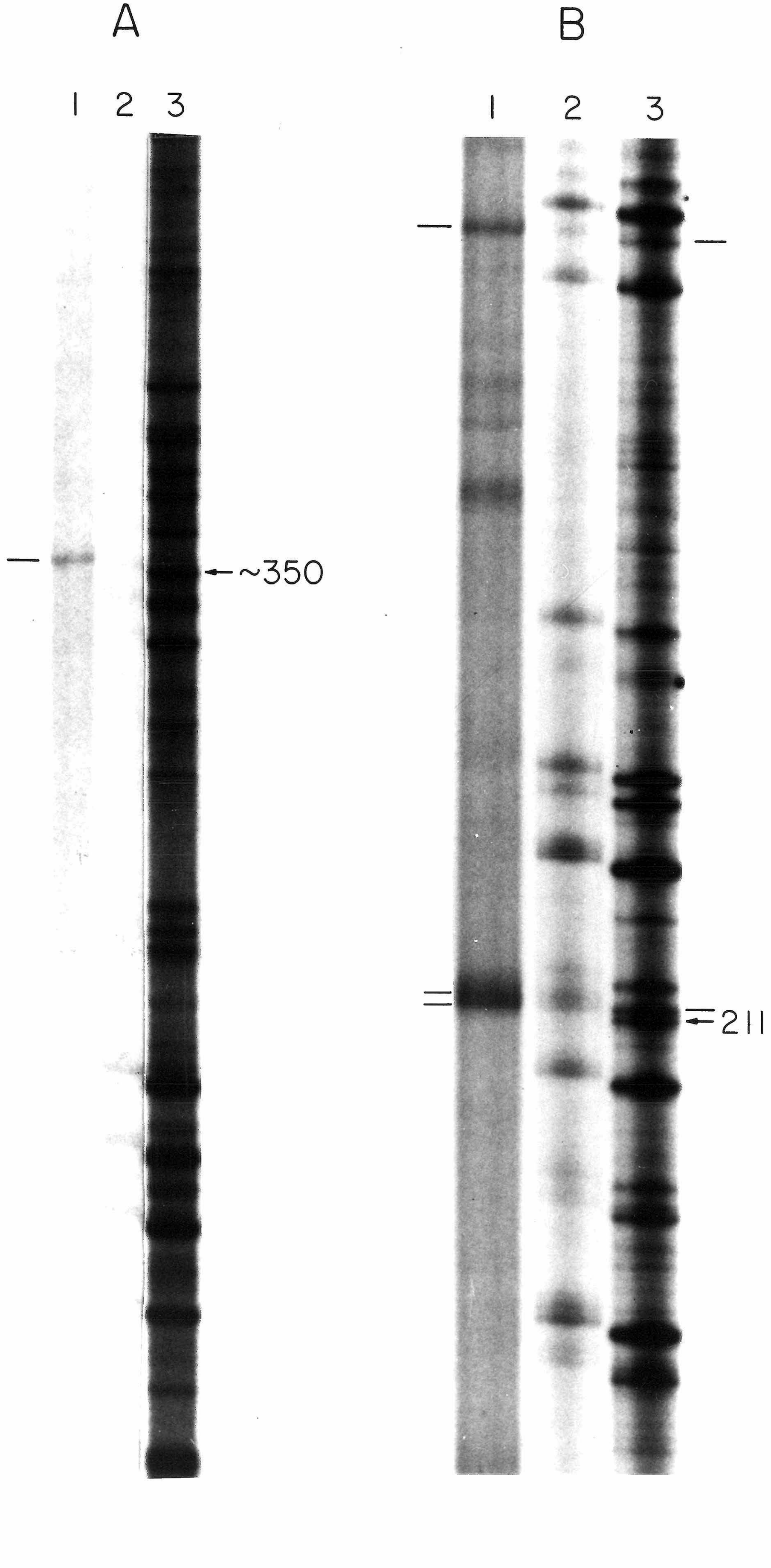
Sequence determination of the 5'-terminus of 49S RNA

Two approaches were used to determine the 5'-terminal sequence of 49S RNA. The first approach used is the direct RNA sequencing method. Details have been described elsewhere[9]. Briefly, RNAs were decapped by β-elimination and recapped with α-^{32}P-GTP and vaccinia guanylyl transferase. The end-labeled RNAs were then partially digested with alkali or base-specific ribonucleases and sequence determined on acrylamide gels. Because no ribonucleases have yet been found which discriminate U and C very well, this approach although it sometimes allowed us to determine sequences up to 160 nucleotides in length, often gave ambiguous results for the pyrimidines.

The second approach used is a cDNA method. The rationale of this approach is that if an RNA template is randomly primed for synthesizing cDNA, a large proportion of cDNA could be transcribed to the 5'-end of the template and have a homogeneous 3'-terminus. When cDNA synthesized this way are 3'-end-labeled and digested with a restriction endonuclease, a distinct 3'-end-labeled fragment containing the complement of the 5'-end of the template would be generated. Most of the other end-labeled fragments, because of their heterogeneous size and 3'-termini, when analyzed on a gel would become the background. Based on this rationale, we were able to identify those HaeIII fragments which are complementary to the 5'-ends of the 49S RNAs of SIN-HR (heat-resistant strain), SIN-WT (wild-type strain), SFV, Middelburg virus (MBV), Highlands J virus (HJ) and eastern equine encephalitis virus-South America strain (EEE-SA). Because not enough labels were incorporated, these fragments were used as markers to identify their relative 5'-end-labeled HaeIII fragments which were subsequently isolated from a preparative gel and sequenced by the chemical method[14].

Figure 1 shows the experimental results of HJ and EEE-SA. cDNA synthesized which were 3'-end-labeled followed by the HaeIII digestion are shown in lane 1. Lane 2 and lane 3 are uniformly 3'- and 5'-end-labeled HaeIII fragments, respectively.

Fig. 1 Identification of the HaeIII cDNA fragments which contain sequences complementary to the 5'-ends of HJ and EEE-SA 49S RNAs. Methods for preparing 49S RNAs and for synthesizing cDNAs using calf thymus random primers were the same as before[9,15]. Lane 1: 0.2 μg of the cDNA synthesized were denatured at 56°C for 5 min and then 3'-end labeled with α-^{32}P-CTP (410 Ci/mmole, Amersham) and terminal transferase (Bethesda Research Laboratory) as described[14]. After two ethanol precipitations with 1 μg tRNA carrier cDNA samples were phenol/chloroform extracted[8] and ethanol reprecipitated. The resulting cDNA were then digested with eight units of HaeIII (New England Biolab) at 37°C for 3 hours in a condition recommended by the manufacturer and were again ethanol precipitated twice with 25 μg tRNA carriers. Finally, the samples were treated with 1 M piperidine and lyophilized as described[14]. Radioactivities incorporated into the samples after these reactions were enough for two loadings on a gel to give reasonable darkness of bands in 48 hours with the intensifying screen and the preflashed film. Lane 2: 0.2 μg of cDNA were treated with HaeIII as described above, followed by phenol/chloroform extractions and ethanol precipitations. The HaeIII digests were then 3'-end labeled, ethanol reprecipitated with 25 μg tRNA carriers, treated with piperidine and finally lyophilized as described above. Samples after these reactions were split for four loadings. Lane 3: 5'-end labeled HaeIII cDNA fragments. Conditions for 5'-end labeling using less than 1 μg of cDNA have been described before[15]. 1% of the labeled samples were used for one loading on an analytical gel. A: experiments with HJ 49S RNA. B: experiments with EEE-SA 49S RNA. Ticks mark the fragments identified. Arrows denote the fragments which contain the complement of the 5'-end of 49S RNA. Numbers indicate the sizes of the fragments which were calculated from either molecular weight markers (HJ) or sequencing results (EEE-SA).


A
1
2
3
~350
B
1
2
3
211

Details of these experiments are described in the figure legend. As expected, only one fragment with a size of about 350 nucleotides was identified in HJ experiments (Fig. 1A). In EEE-SA experiments (Fig. 1B) three fragments were identified, however. When the lower two fragments were isolated and sequenced, the fragment denoted by an arrow with a size of 211 nucleotides was identified to be the fragment of interest. The EEE-SA fragment close to the top of the gel was not sequenced because of its size.

The reason why multiple bands were identified in EEE-SA experiments is probably because of the presence of preferential termination sites for reverse transcriptase in the EEE-SA 49S RNA template. These sites are most likely caused by secondary structures[8,9,15]. It is also possible that the largest fragment of EEE-SA identified is a partially digested HaeIII fragment which also contains the sequence complementary to the 5'-end of EEE-SA 49S RNA.

Those HaeIII fragments specific for the 5'-terminal sequences of 49S RNAs of Venezuelan equine encephalitis viruses (VEE) and EEE-North America strain (EEE-NA) were not found. This could be caused by the presence of a HaeIII site near the 5'-end, or caused by the absence of a HaeIII site within a reasonable length of distance from the 5'-end. For SFV, the 5'-end-labeled HaeIII fragment identified when isolated was contaminated by another fragment of similar size and was unable to be sequenced. Thus, 5'-terminal sequences of the 49S RNAs of VEEs, EEE-NA and SFV were determined only by the direct RNA sequencing method.

5'-terminal sequences of alphavirus 49S RNAs

All the 5'-terminal sequences determined are presented in Fig. 2. Sequences determined only by the direct RNA method are underlined. Except for a single base substitution at nucleotide 35 (Fig. 3), sequences of SIN-HR and SIN-WT are identical for at least 90 nucleotides from the 5'-end and thus only the sequence of SIN-HR is presented. Most of the sequences of the HaeIII and TaqI cDNA fragments of SIN-HR 49S RNA

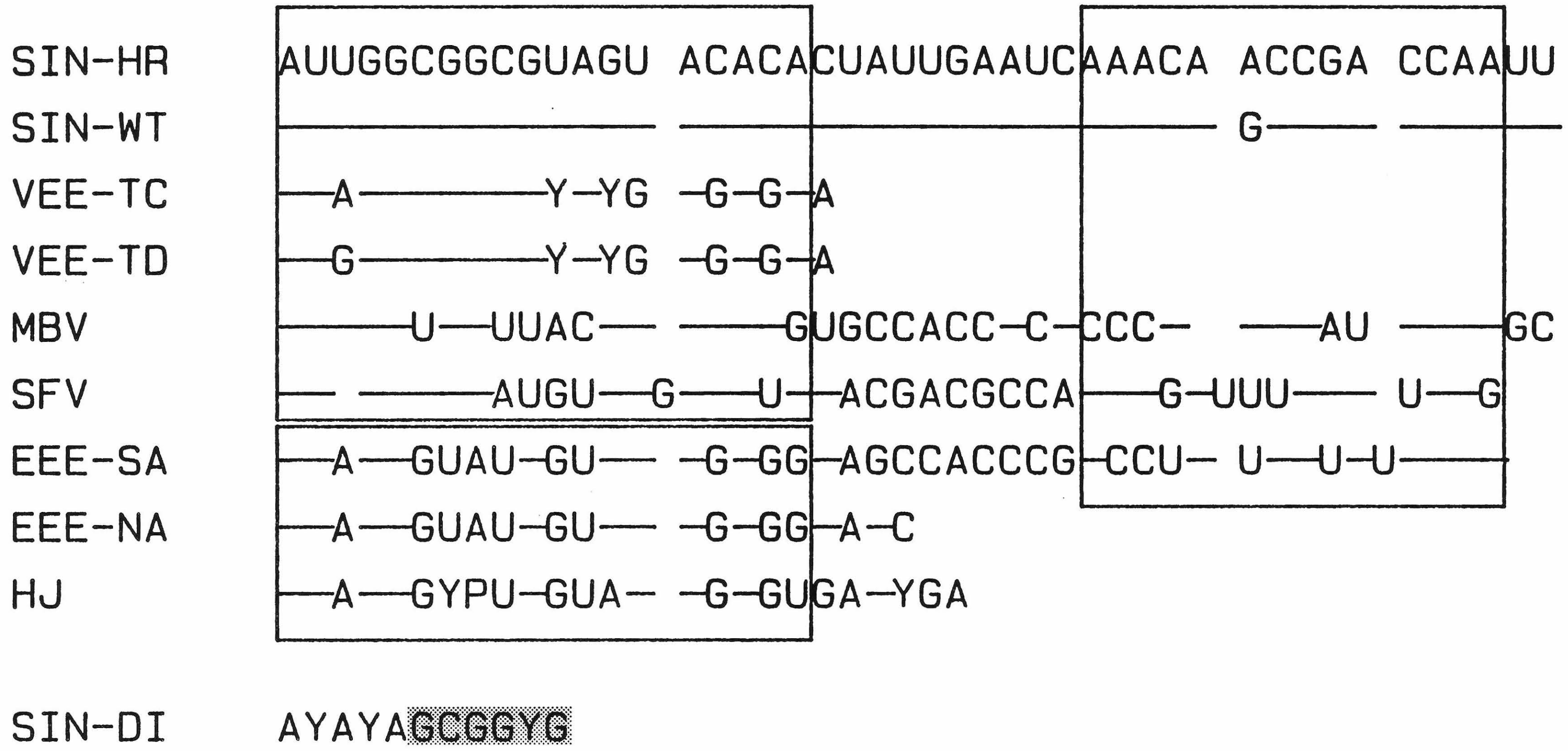
Fig. 2 5'-terminal sequences of alphavirus 49S RNAs. Sequences relying only on the results of the direct RNA method are underlined. For SIN, MBV, EEE-SA and HJ, 5'-end labeled *HaeIII* fragments identified and prepared by the methods described in the legend of Fig. 1 were isolated from preparative gels and sequenced by the chemical method[12]. SFV sequence came from our sequencing results and those previously published[17,18] (see text). Highly conserved sequences are shaded. Amino acid sequences deduced are placed on top of the nucleotide sequences. N's indicate sequences not determined. P and Y are purine and pyrimidine, respectively. Amino acid abbreviations: A, Ala; C, Cys; D, Asp; E, Glu; F, Phe; G, Gly; H, His; I, Ile; K, Lys; L, Leu; M, Met; N, Asn; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp; Y, Tyr.

```
                                                                           M  E  K  P  V  V  N  V
SIN-HR    m7GpppAUUGGCGGCGUAGUACACACUAUUGAAUCAAACAACCGACCAAUUGCACUACCAUCACAAUGGAGAAGCCAGUAGUAAACGUA
                                                                           M  A  R  P  V  V  K  I
MBV                         m7GpppAUUGGUGGUUACGUACACGUGCCACCACCCCCACCAUCCAAGCGAUCCAAAAUGGCGCGCCCUGUUGUGAAGAUA
                                                                           M  A  A  K     V  H  V
SFV                       m7GpppAUGGCGGAUGUGUGACAUACACGACGCCAAAAGAUUUCGAUCAG (N)?CACGAUGGCCGCCAAA   GUGCAUGUU
                                                                           M  E  K        V  H  V
EEE-SA                            m7GpppAUAGGGUAUGGUGUAGAGGCAGCCACCCGACCUAUCCUAUCCAAAAUGGAGAAA      GUACACGUA
                                                                           M  E  K        V  H
EEE-NA                            m7GpppAUAGGGUAUGGUGUAGAGGCAACNNNNNNNNNNNNNNNNAGCAAAAUGGAGAAA      GUACAYG
                                                                           M  E  K
VEE-TC                                        m7GpppAUAGGCGGCGYAYGAGAGAA (N)15AGYCAAAAUGGAGAAAG
                                                                           M  E  K
VEE-TD                                        m7GpppAUGGGCGGCGYAYGAGAGAA (N)18AGYCAAAAUGGAGAAAG


          D  V  D  P  Q  S  P  F  V  V  Q  L  Q  K  S  F  P  Q  F  E  V  V  A  Q  Q  V  T  P  N  D
SIN-HR    GACGUAGACCCCCAGAGUCCGUUUGUCGUGCAACUGCAAAAAGCUUCCCGCAAUUUGAGGUAGUAGCACAGCAGGUCACUCCAAAUGAC
          D  V  E  A  E  S     F  V  K  S  L  Q  K  V  F  P  Q  F  E  I  E  A  E  Q  V  T  P  N
MBV       GACGUUGAGGCYGAAAGYYCAUUUGUCAAGUCUCUACAGAAGGUGUUUCCACAAUUUGAGAUCGAAGCAGAGCAGGUCACUCCGAAUGA
          D  I  E  A  D  S  P  F  I  K  S  L  Q  K  A  F  P  S  F  E  V  E  S  L  Q  V  T  P  N  D
SFV       GAUAUUGAGGCUGACAGCCCAUUCAUCAAGUCUUUGCAGAAGGCAUUUCCGUCAUUCGAGGUGGAGUCAUUGCAGGUCACACCAAAUGAC
          D  L  D  A  D  S  P  Y  V  K  S  L  Q  K  C  F  P  H  F  E  I  E  A  T  Q  V  T  D  N  D
EEE-SA    GACUUAGACGCUGACAGCCCAUACGUCAAGUCGCUGCAAAAGUGCUUUCCGCAUUUUGAGAUAGAAGCCACGCAGGUCACCGACAAUGAC
                                                                                  Q  V  T  A  N  D
HJ                                        m7GpppAUAGGGYPUGGUAUAGAGUGAAYGAY (N)~105 GCAGGUCACAGCCAAUGAC


          H  A  N  A  R  A  F  S  H  L  A  S  K  L  I  E
SIN-HR    CAUGCUAAUGCCAGAGCAUUUUCGCAUCUGGCCAGUAAACUAAUCGAG
          H  A  N  A  R  A  F  S  H  L  A  T  K  L  I  E  Q  E  T  D
SFV       CAUGCAAAUGCCAGAGCAUUUUCGCACCUGGCUACCAAAUUGAUCGAGCAGGAGACUGAC
          H  A  N  A  R  A  F  S  H  L  A  T  K  L  I  E  S  E  V
EEE-SA    CAUGCUAAUGCUAGAGCGUUUUCGCAUCUGGCUACAAAACUCAUCGAAAGCGAAGUGG
          H  A  N  A  R  A  F  S  H  V  A  T  K  L  I  E  S  E  V  D  R  E  Q  I  I
HJ        CAUGCUAAUGCCAGAGCGUUUUCGCAUGUGGCUACAAAGCUCAUAGAGAGCGAAGUCGACCGGGAGCAGAUUAUACU
```

Fig. 3 Comparison of the 5'-terminal sequences of alphavirus 49S RNAs. Sequences that are conserved among viruses are boxed. The sequence of SIN-DI is the common 5'-terminal sequence of a family of DI RNAs as described in the text. These DI RNAs were isolated from BHK cells after 16 undiluted passages of SIN. Details for the preparation of these DI RNAs have been described elsewhere[11]. Shaded region indicates the homologous sequence of SIN-DI to the genomic sequence. Y, pyrimidines.


SIN-HR AUUGGCGGCGUAGU ACACACUAUUGAAUCAAACA ACCGA CCAAUU
SIN-WT G
VEE-TC A Y YG G G A
VEE-TD G Y YG G G A
MBV U UUAC GUGCCACC C CCC AU GC
SFV AUGU G U ACGACGCCA G UUU U G
EEE-SA A GUAU GU G GG AGCCACCCG CCU U U U
EEE-NA A GUAU GU G GG A C
HJ A GYPU GUA G GUGA YGA
SIN-DI AYAYAGCGGYG

have been determined (E. G. Strauss, unpublished results), nucleotides 205 to 221 of SIN-HR presented were derived from overlapping the 5'-terminal sequence determined with the sequence of a TaqI fragment which was again overlapped by another HaeIII fragment.

The HaeIII fragment of MBV which contains the complement of the 5'-end of its 49S RNA is 86 nucleotides in length and thus the sequence from nucleotide 87 to 165 was determined only by the RNA method. The HaeIII fragment identified for HJ is about 350 nucleotides in length; because of the size only part of its sequence was determined. The rest of the HJ sequence was either not determined or determined by the direct RNA method. Partial 5'-terminal sequences of EEE-SA and VEE 49S RNAs are also presented in Fig. 2. Due to band compressions on sequencing gels which were further complicated by ambiguous pyrimidines, sequences represented by N's were not determined. VEE-TC (TC-83 strain) is the avirulent vaccine strain of VEE-TD (Trinidad donkey strain)[16]. The 5'-terminal sequences of these two viruses determined are different at at least four positions: besides having a different nucleotide at position 3 (Fig. 3), VEE-TC might also have three other nucleotide deletions in the region represented by N's (Fig. 2).

The first 12 nucleotides of our SFV sequence determined by the RNA method are the same as those previously published[17]. Nucleotides 11 to 38 also match the 5'-terminal sequence of an SFV DI RNA[18]. From comparative studies of the amino acid sequences deduced (see below), the coding sequence of SFV and the tetranucleotide preceding it were taken out from this SFV DI RNA sequence. Because DI RNAs undergo frequent sequence rearrangements[11,12], we cannot rule out the possibility of any deletion or insertion of the sequences.

Heterogeneous DI RNAs of SIN were also sequenced by the direct RNA method. These DI RNAs have the same 5'-terminal sequence at least up to the region we determined (Fig. 3). The sequence of the first four nucleotides of the genomic RNA

(AUUG) has been mutated to AYAYA in these DI RNAs. From nucleotide 6 the sequence of SIN DI RNAs becomes the same as that of genomic RNA (Fig. 3, shaded region).

In Fig. 3 we also compared the sequences of the first 45 nucleotides or so of the alphavirus RNAs determined. Sequences of the first 20 nucleotides of EEEs and HJ are extremely similar (about 80% homology). Except for these viruses, 5'-terminal sequences in this region in general are not as conserved as those at the extreme 3'-ends next to the poly(A) tails.

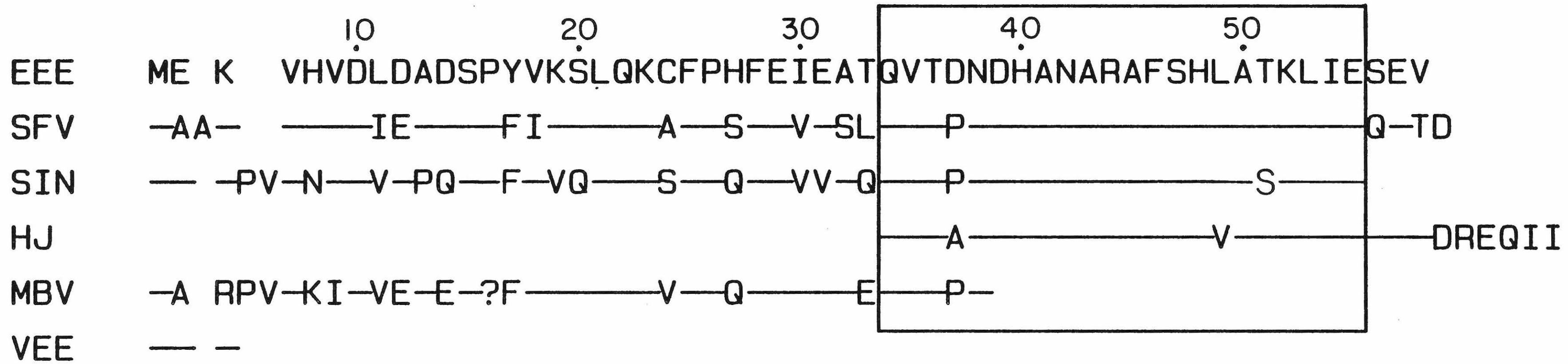
NH_2-terminal sequences of nonstructural polyproteins

It has been noticed that the initiation codons of about 60% of eucaryotic mRNAs are preceded by these tetranucleotides: CANA, CAAN, CANC and CACN (J.-H. Ou, Ph.D. Thesis, California Institute of Technology). Potential initiation codons identified in Fig. 2 are all preceded by these tetranucleotides. In the nucleotide sequences presented, no termination codons are found in the reading frames generated by these initiation codons. In vitro translation of SFV 49S RNA showed that the first two amino acids of SFV nonstructural polyprotein are Met and Ala[19] which are the same as those deduced in Fig. 2. For these various reasons, we are confident that the initiation codons identified are indeed used for translating alphavirus nonstructural proteins.

The 5'-untranslated regions of 49S RNAs range from 42 to 59 nucleotides, depending on the viruses (Fig. 2). It is noteworthy that the SFV HaeIII cDNA fragment we identified which contains the complement of the 5'-end of its 49S RNA is about 88 nucleotides in length. In the SFV sequence presented in Fig. 2, only one HaeIII site is present. This HaeIII site is between nucleotides 4 and 5 from the 5'-end of the coding sequence. Therefore, the 5'-untranslated region of SFV could be as long as 84 nucleotides.

The N-terminal sequences of nonstructural proteins are compared in Fig. 4.

Fig. 4. The comparison of the NH_2-terminal sequences of the nonstructural polyproteins of alphaviruses. Highly conserved sequences are boxed. The numbering of the positions of the amino acids are arbitrary. All the deletions (or insertions) are counted.


10
20
30
40
50
EEE ME K VHVDLDADSPYVKSLQKCFPHFEIEATQVTDNDHANARAFSHLATKLIESEV
SFV —AA— ———IE———FI————A——S——V—SL——P——————————————Q—TD
SIN —— —PV—N——V—PQ——F—VQ———S——Q——VV—Q——P———————————S———
HJ ——A——————————V——————DREQII
MBV —A RPV—KI—VE—E—?F————V——Q—————E——P—
VEE —— —

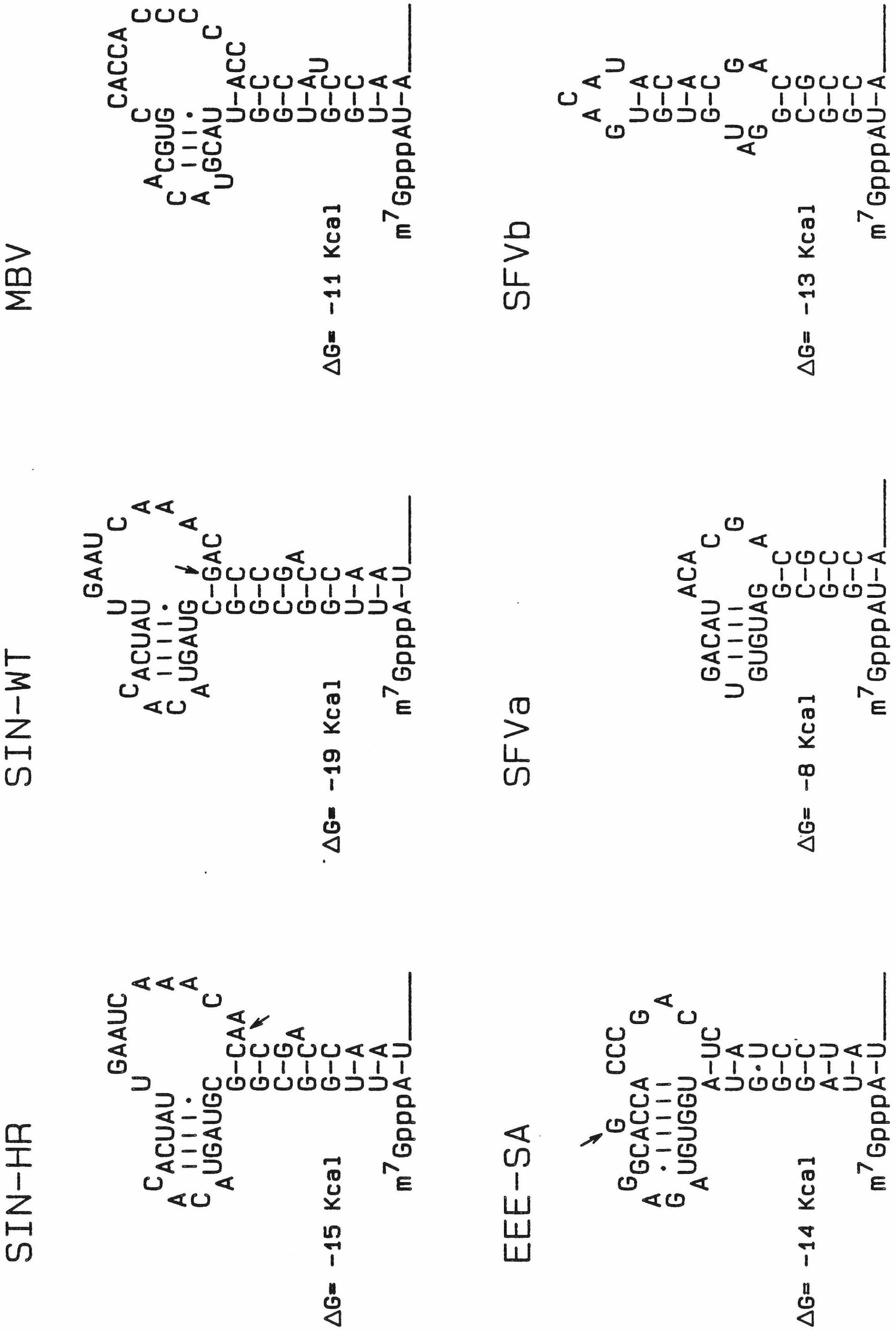
Sequences from positions 34 to 55 are highly conserved. However, most of this region (positions 34 to 50) are encoded by a highly conserved nucleotide sequence (see below). As shown in Fig. 4, at position 3 SFV is distinct from other viruses by having an insertion of Ala, and at positions 5 and 6 SIN and MBV have two insertions of Pro and Val. Except in SIN, nonconserved substitutions of the amino acids always occur at the same positions among viruses, these include amino acids at positions 2, 27, 33 and 37. Other variations of the amino acid sequences are all conserved in the respects of charges (positions 9, 18, 30 and 49), properties—nonpolar amino acid substitution (positions 9, 18, 30 and 49) and R-group structures (positions 17, 24, 32 and 51). The sequence of SIN is different from other viruses by having its unique amino acids at positions 8, 13, 14, 19, 20 and 31.

Conserved secondary structures and sequence

Stable stem and loop structures could be formed at the extreme 5'-ends of the 49S RNAs of SIN, MBV, EEE-SA and SFV (Fig. 5). Despite the divergence of the nucleotide sequences, the shape of the stem and loop structures of these viruses remains very similar, except the b type of SFV. We do not have enough sequence data for VEEs and EEE-NA and thus do not know whether this stem and loop structure is also present in their 49S RNAs or not. However, it is interesting to note that the only nucleotide that is different between the two EEE sequences determined is the nucleotide 22 (Figs. 2 and 3). This nucleotide is not bonded in the stem and loop structure of EEE-SA (Fig. 5, denoted by an arrow).

Because not enough 5'-terminal sequences of DI RNAs are yet available, it is not clear whether this stem and loop structure is conserved in them or not. However, because the first four nucleotides of SIN which might be involved in forming the stem and loop structure have been mutated to another sequence in the DI RNAs sequenced, it is likely that if stem and loop structures are conserved in DI RNAs they might have different conformations.

Fig. 5. Possible stem and loop structures at the extreme 5'-ends of 49S RNAs. Arrows indicate the positions of the nucleotides which are different between SIN-HR and SIN-WT, and between EEE-SA and EEE-NA. Free energies were calculated by using the method of Tinoco, *et al.*[20].


SIN-HR
ΔG= -15 Kcal
SIN-WT
ΔG= -19 Kcal
MBV
ΔG= -11 Kcal
EEE-SA
ΔG= -14 Kcal
SFVa
ΔG= -8 Kcal
SFVb
ΔG= -13 Kcal
m^7GpppA-U
m^7GpppAU-A

A highly conserved sequence which is 51 nucleotides in length is located at about 130 to 150 nucleotides from the 5'-end of 49S RNA, depending on the viruses (Fig. 2, shaded regions). Although this sequence is in the coding region of nonstructural proteins, we do not believe that its conservation is only because it codes for a highly conserved amino acid sequence. The major reason is that only 5 out of 14 conserved amino acids encoded by this sequence are coded by more than one codon (about 35% silent mutation rate), whereas about 95% of the conserved amino acids encoded by other sequences, such as amino acids at positions 21-23 and 52-55 (Fig. 4), are coded by two or more than two codons.

This conserved sequence could form two stable hairpin structures. Using SIN-HR as an example, the entire 51 nucleotides and the hairpin structures formed by them are shown in Fig. 6. Hairpin structures of other viruses (SFV, EEE-SA and HJ) have about the same stability as those of SIN-HR. In Fig. 6, variable nucleotides are circled.

49S RNAs of SIN and SFV are able to cyclize and form panhandle structures[21,22]. Because serological experiments[23] and nucleotide sequence studies[10] suggest that SIN and SFV are distally related alphaviruses, these circular structures are probably conserved in all alphavirus 49S RNAs. The average length of the panhandle structures is about 0.083 μm which is equivalent to 250 ± 50 nucleotides. For this reason, it has been proposed that there are complementary sequences at both ends of 49S RNA which are responsible for forming the panhandle structure and cyclizing the RNA molecule[21,22]. Free energies of these circular structures have been estimated to be about -13 kcal in 0.1 M salt solution[27]. According to these results, sequences which might be involved in forming circular structures have been searched and the potential 5'- and 3'-terminal sequences have been located (Fig. 7). The 5'-terminal sequences found are either partly or completely located within the conserved 51 nucleotides described above, and the 3'-terminal sequences found are at about 240

Fig. 6 The conserved 51 nucleotides of SIN-HR and the possible hairpin structures formed by them. Variable nucleotides are circles. Because EEE-SA, SFV, HJ and SIN-HR all have similar structures and free energies, only the structures of SIN-HR are presented. Free energies are calculated as described in the legend of Fig. 4. Nucleotide sequence is read from 5'- to 3'-end.

SIN

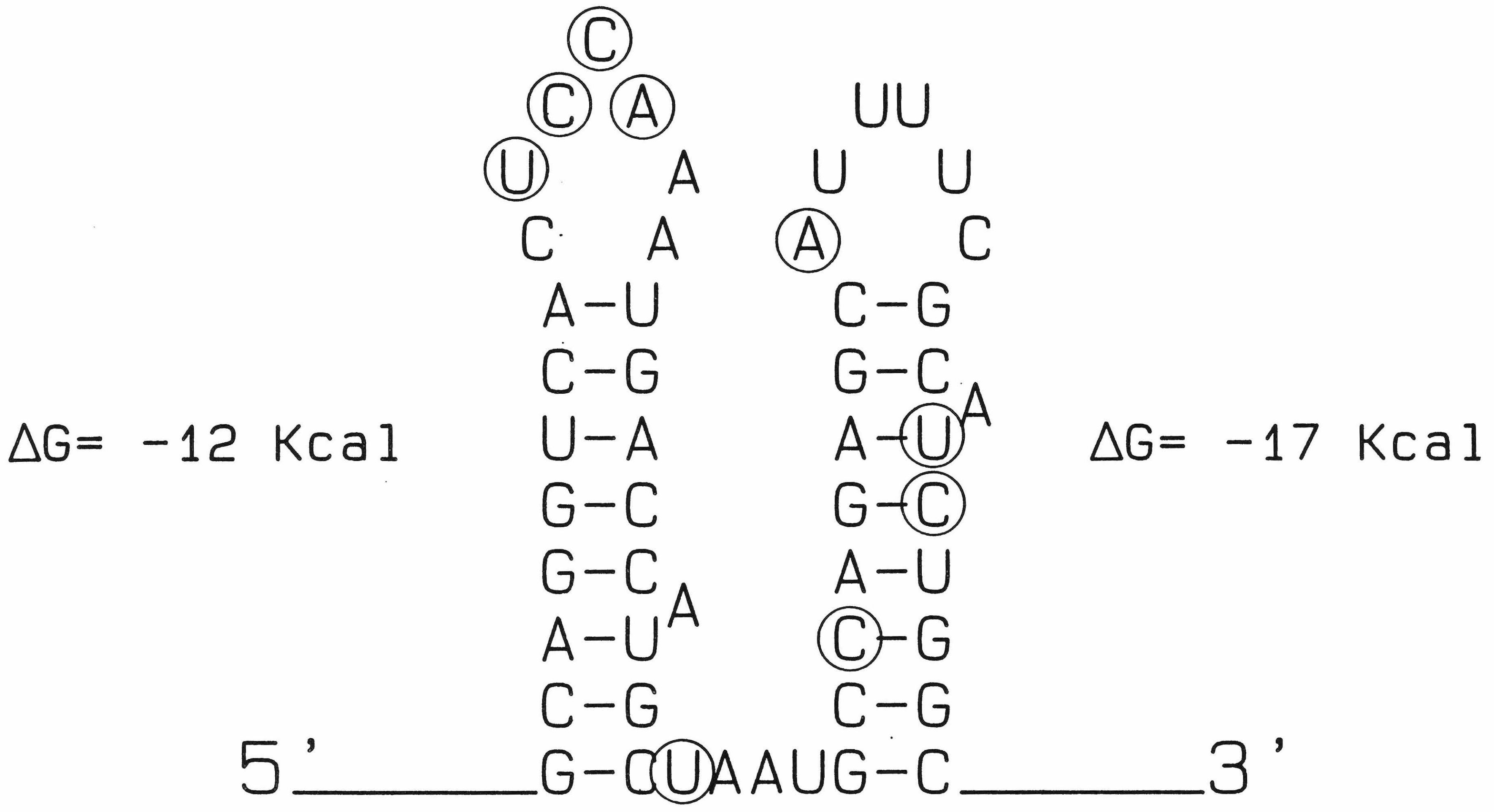
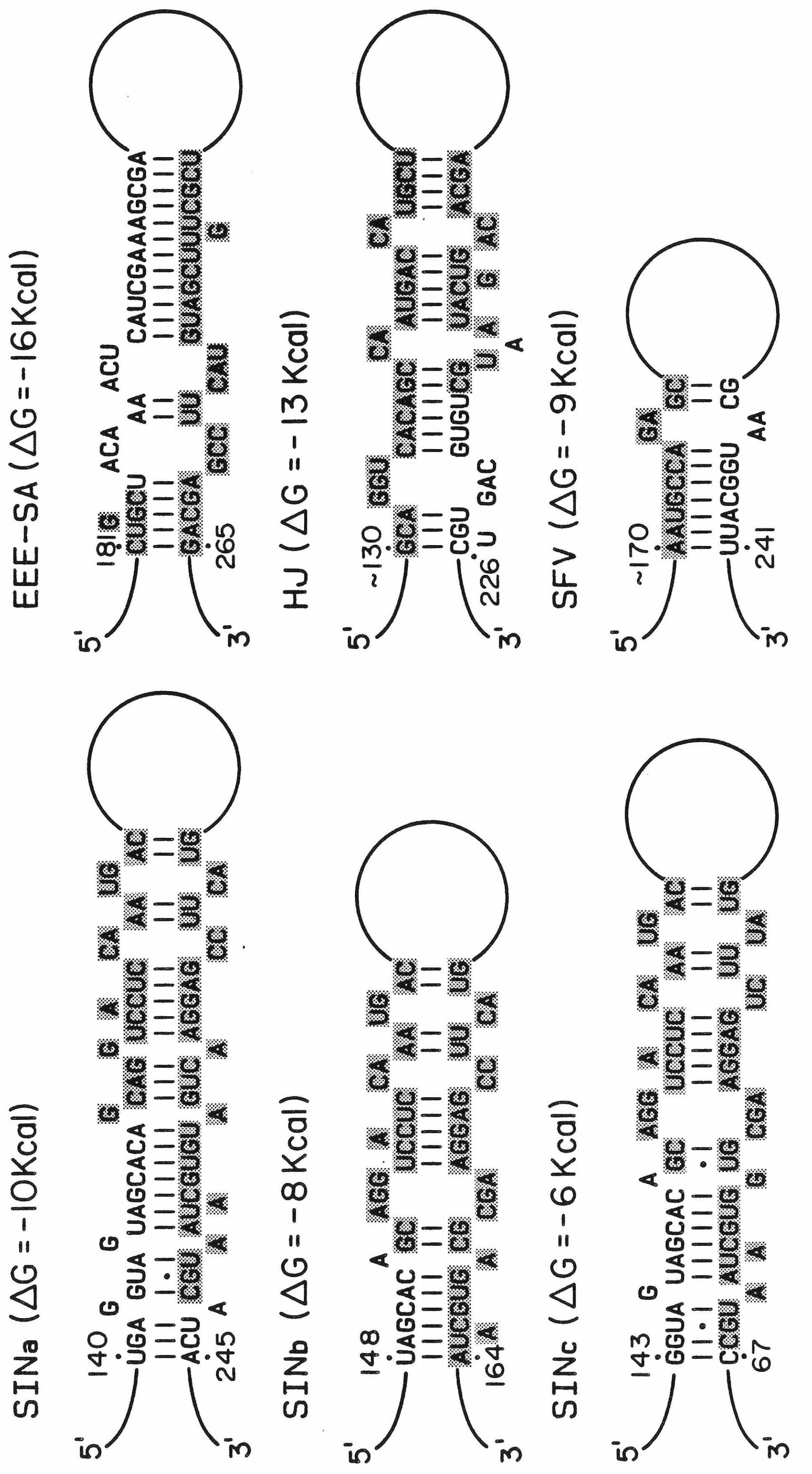


Fig. 7 Secondary structures which would cyclize 49S RNAs. Sequences shaded are in the 5'- or 3'-conserved regions (see text). Numbers indicate the positions of the sequences from either 5'- or 3'-ends (excluding poly(A) tails). Free energies were calculated the same way as described in the legend of Fig. 4. 5'-end sequences are placed on the top of 3'-end sequences.


EEE-SA (ΔG=-16Kcal)
181 G
CUGCU ACA AA ACU CAUCGAAAGCGA
GACGA GCC UU CAU GUACGUUUCGCU
G
265
HJ (ΔG=-13Kcal)
~130
GCA GGU CACAGC CA AUGAC CA UGCU
CGU GUGUCG U A UACUG G AC ACGA
226 U GAC
A
SFV (ΔG=-9Kcal)
~170
AAUGCCA GA GC
UUACGGU AA CG
241
SINa (ΔG=-10Kcal)
140 G
UGA GUA UAGCACA CAG UCCUC CA AA UG AC
ACU CGU AUCGUGU GUC AGGAG UU CA UG
G G G A
245 A A A A A CC
SINb (ΔG=-8Kcal)
148
UAGCAC A GC AGG A UCCUC CA AA UG AC
AUCGUG CG CGA AGGAG CC UU CA UG
164 A A
SINc (ΔG=-6Kcal)
143 G
GGUA UAGCAC A GC AGG A UCCUC CA AA UG AC
CCGU AUCGUG UG G CGA AGGAG UC UU UA UG
67 A A
5'
3'

nucleotides from poly(A) tails. We have noticed before that a sequence which is about 40 nucleotides in length is conserved in the 3'-noncoding regions of SIN, EEE and HJ 49S RNAs[10]. For these viruses, the 3'-terminal sequences which might be involved in 49S RNA cyclization are also partly or completely located within this sequence. Because this sequence has been repeated three times in SIN[10], three kinds of secondary structures are possible for this virus (Fig. 7). For SFV, this sequence is absent; however, the potential 3'-terminal sequence involved in forming the circular structure of its 49S RNA is partly repeated twice[10]. The average length of the poly(A) tails of either SIN or SFV 49S RNA is about 70 nucleotides[24]. Including poly(A) tails and those nucleotides involved in base-pairings, panhandles formed by the three structures of SIN shown in Fig. 7 would have the size of 350, 260 and 170 nucleotides, respectively, and the panhandle formed by the SFV structure would have a size of about 320 nucleotides. Free energies calculated by the method of Tinoco et al.[20] are also shown in Fig. 7. (They are -10, -8, and -6 kcals for the three structures of SIN, respectively, and -9 kcal for the structure of SFV.) Because efforts to search other more stable secondary structures formed by the known sequences of both 3'- and 5'-ends of 49S RNAs failed, it is likely that structures presented in Fig. 7 are indeed involved in 49S RNA cyclization.

Discussion

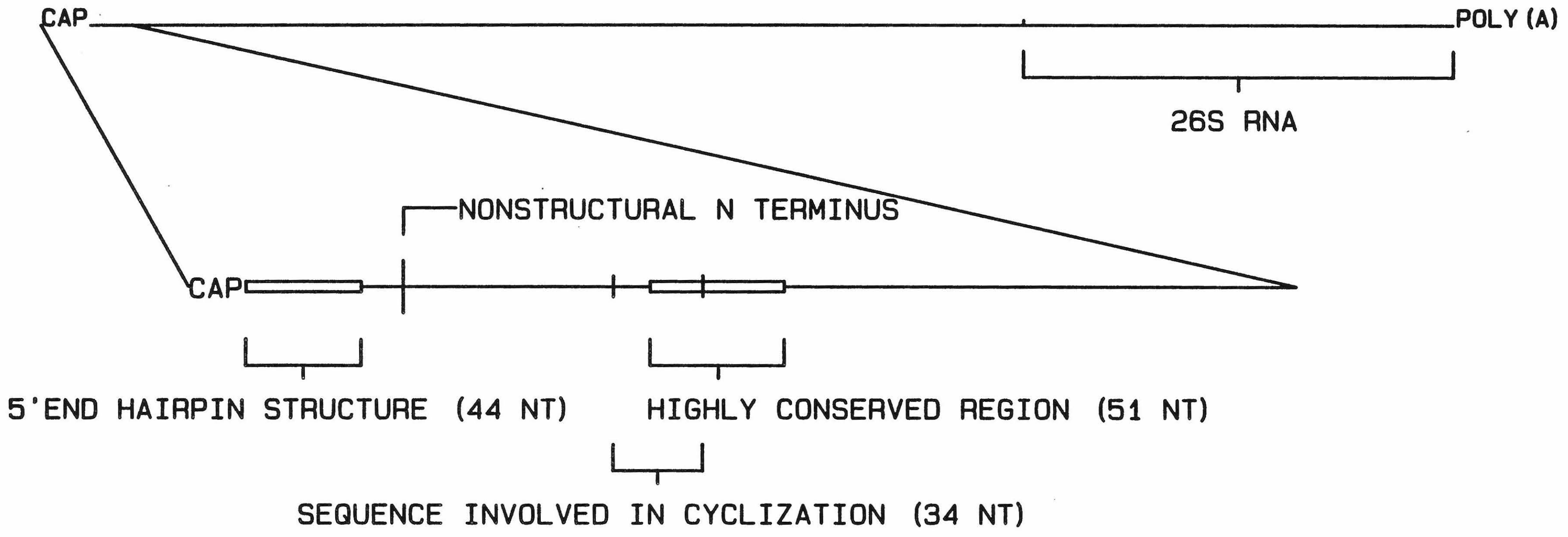
Determining the 5'-terminal sequences of RNAs has always been a problem because of the difficulties to specifically end-label an RNA at its 5'-terminus and the lack of a ribonuclease which can unambiguously distinguish pyrimidines. Two-dimensional fingerprints[25,26] for determining RNA sequences are too time consuming and can determine only limited sequence data. In this paper we described sequencing strategies which are especially useful for determining the 5'-terminal sequences of long RNA transcripts. These strategies could also be used to determine 5'-terminal sequences of a heterogeneous population of RNAs, such as DI RNAs, for which the separation of the RNA species is rather difficult[11]: fragments identified on an analytical gel could all be isolated and sequenced. This would yield sequence data of several different RNA species. However, the strategies described above are still improvable: first, raising the efficiency of 3'-end labeling would enable us to isolate fragments directly from lane 1 of Fig. 1 and determine their sequences. This would greatly simplify the procedures. Second, other restriction endonucleases such as TaqI, HhaI and RsaI which cut single-stranded DNA could also be used in the experiments[15]. This would allow people to select fragments of different length.

Using SIN as an example, the 5'-end region of alphavirus 49S RNA is illustrated in Fig. 8. The conservation of the hairpin structures at the extreme 5'-end of 49S RNAs suggests that they might be important in alphavirus replications. Except for an additional G at the 3'-terminus, the 3'-end of minus-stranded RNA of SFV has been shown complementary to its 49S RNA[17]. Because of this sequence complementarity, similar structures would also be present at the 3'-end of the minus-stranded RNA. It is possible that these stem and loop structures in the minus-stranded RNAs are replicase binding sites which control the replication of 49S RNAs. Conserved secondary structures at the 3'-ends of RNAs which might serve as the initiation sites of replciation have also been observed in other RNA viruses[27]. The possibility that this structure in 49S RNA is required for encapsidation, however, cannot be ruled out. Because of the conservation of the structure the less stable a structure of SFV in Fig. 5 would be more favorable than the b structure for serving these functions. The difference of the free energies could be compensated by the binding of replicases or capsid proteins.

As mentioned above, sequences of SIN-HR and SIN-WT have a single base replacement at nucleotide 35 (Fig. 3). This point mutation would remove one GC base pair from SIN-WT structure and slightly alter the conformation of the 5'-end stem and loop structure of SIN-HR (Fig. 5, denoted by arrows). We have also noticed that the

Fig. 8 Illustration of the 5'-end region of SIN 49S RNA.

SINDBIS VIRUS 49S RNA (13000 NT)


CAP
POLY (A)
26S RNA
NONSTRUCTURAL N TERMINUS
CAP
5'END HAIRPIN STRUCTURE (44 NT)
HIGHLY CONSERVED REGION (51 NT)
SEQUENCE INVOLVED IN CYCLIZATION (34 NT)

nucleotide sequences of two VEEs are possibly different at four positions (Figs. 2 and 3). These changes would also result in different conformations of their stem and loop structures. If the stem and loop structure is indeed involved in 49S RNA replication, changes in its conformation would be expected to affect the replicational efficiencies of viruses.

No complete 5'-terminal sequences of alphavirus DI RNAs are available at present; however, because the sequence of an SFV DI RNA diverged from the genomic sequence at about 10 nucleotides after the 3'-end of the hairpin structure[18], there is the possibility that the hairpin structure is also conserved in DI RNAs. Besides the SIN DI RNAs we studied, a group of SFV DI RNAs were also found to have modified 5'-terminal sequences[28], although these sequences are different from those of SIN DI RNAs and are heterogeneous. What the significance of these modified sequences is and how they were generated are not clear at this moment. However, because these kinds of sequence modifications are not found in genomic RNAs, there must be certain selection pressures against them. One possibility is that an unmodified 5'-terminal sequence is required for translating nonstructural proteins; alternatively, it might be involved in 26S RNA transcription which is not required in DI RNA replication[12].

We have suggested before that the transcription of the minus-stranded RNA, besides involving the highly conserved 19 nucleotides adjacent to the poly(A) tail, must also involve other non-26S RNA sequences[10]. Because 26S RNA, although it has the 3'-terminal sequence of 49S RNA[8], is apparently not used as the template for minus-stranded RNA synthesis[29]. Because of the conservation of the circular structures of 49S RNAs, we proposed that cyclization of 49S RNA is important in bringing the non-26S RNA sequences to the 3'-end for replicases to recognize[10]. If this is true, because of the possible involvement in 49S RNA cyclization, the highly conserved 51 nucleotides found at the 5'-end of the RNA could be the non-26S RNA sequence involved in minus-stranded RNA synthesis. In this regard, we have noticed that although the potential 3'-terminal sequence involved in 49S RNA cyclization has been deleted in an SFV DI RNA, this conserved 5'-terminal sequence has been rearranged to a position about 190-250 nucleotides from the poly(A) tail[18]. It is noteworthy, however, that if this conserved 51 nucleotides is involved in minus-stranded RNA synthesis, its translocation might result in a different configuration of the sequence to be recognized by the replicases.

The hairpin structures of this highly conserved sequence (Fig. 5) might also be important in other functions, such as encapsidation. Hairpin structure serving as the initiation site of encapsidation has been observed in the RNA of tobacco mosaic virus (TMV)[30]. Because this highly conserved sequence has been repeated three times in the SFV DI RNA mentioned above[18], if this sequence is involved in encapsidation, the process of it would be very different from that of TMV.

Acknowledgements

We are deeply grateful to Dr. D. W. Trent for supplying us 49S RNAs of EEE, VEE and HJ, and to Dr. S. Schlesinger and Steve Monroe for allowing us to use their DI RNAs. We thank Mrs. Edith Lenches for preparing chick embryo cell cultures for us.

References

1. Shope, R. E. in The Togaviruses (ed Schlesinger, R. W.) 47-82 (Academic, New York, 1980).
2. Schlesinger, M. J. & Kaariainen, L. in The Togaviruses (ed Schlesinger, R. W.) 371-392 (Academic, New York, 1980).
3. Strauss, J. H. & Strauss, E. G. in The Molecular Biology of Animal Viruses (ed Nayak, D.) Vol. 1, 111-166 (Dekker, New York, 1977).
4. Rice, C. M. & Strauss, J. H. Proc. natn. Acad. Sci. U.S.A. **78**, 2062-2066 (1981).
5. Garoff, H., Frischauf, A.-M., Simons, K., Lehrach, H., & Delius, H. Proc. natn. Acad. Sci. U.S.A. **77**, 6376-6380 (1980).
6. Garoff, H., Frischauf, A.-M., Simons, K., Lehrach, H., & Delius, H. Nature **288**, 236-241 (1980).
7. Simmons, D. T. & Strauss, J. H. J. mol. Biol. **71**, 615-631 (1972).
8. Ou, J.-H., Strauss, E. G., & Strauss, J. H. Virology **109**, 281-289 (1981).
9. Ou, J.-H., Rice, C. M., Dalgarno, L., Strauss, E. G., & Strauss, J. H. Proc. natn. Acad. Sci. (1982, submitted).
10. Ou, J.-H., Trent, D. W., & Strauss, J. H. J. mol. Biol. (1982, in the press).
11. Monroe, S. S., Ou, J.-H., Rice, C. M., Schlesinger, S., Strauss, E. G., & Strauss, J. H. J. Virol. **41**, 153-162 (1982).
12. Soderlund, H., Keranen, S., Lehtovaara, P., Palva, I., Pettersson, R. & Kaariainen, L. Nucleic Acids Res. **9**, 3403-3417 (1981).
13. Stollar, V. in The Togaviruses (ed Schlesinger, R. W.) 427-457 (Academic, New York, 1980).
14. Maxam, A. M. & Gilbert, W. Methods in Enzymology 65, 499-560 (1980).
15. Rice, C. M. & Strauss, H. J. J. mol. Biol. **150**, 315-340 (1981).
16. Trent, D. W., Clewley, J. P., France, J. K., & Bishop, D. H. L. J. gen. Virol. **43**, 365-381 (1979).

17. Wengler, G., & Gross, H. J. *Nature* **282**, 754-756 (1979).

18. Lehtovaara, P., Soderlund, H., Keranen, S., Pettersson, R. F., & Kaariainen, L. *Proc. natn. Acad. Sci. U.S.A.* **78**, 5353-5357 (1981).

19. Glanville, N., Ranki, M., Morser, J., Kaariainen, L., & Smith, A. E. *Proc. natn. Acad. Sci. U.S.A.* **73**, 3059-3063 (1976).

20. Tinoco, I., Borer, P. N., Dengler, B., Levine, M. D., Uhlenbeck, O. C., Crothers, D. M., & Gralla, J. *Nature* **246**, 40-41 (1973).

21. Hsu, M. T., Kung, H. J., & Davidson, N. *Cold Spring Harb. Symp. quant. Biol.* **38**, 943-950 (1973).

22. Frey, T. K., Gard, D. L., & Strauss, J. H. *J. mol. Biol.* **132**, 1-18 (1979).

23. Porterfield, J. S. in *The Togaviruses* (ed Schlesinger, R. W.) 13-46 (Academic, New York, 1980).

24. Frey, T. K., & Strauss, J. H. *Virology* **86**, 494-506 (1978).

25. Sanger, F., Brownlee, G. G., & Barrell, B. G. *J. mol. Biol.* **13**, 373-398 (1965).

26. De Wachter, R. & Fiers, W. *Analyt. Biochem.* **49**, 183-197 (1972).

27. Ahlquist, P., Dasgupta, R., & Kaesberg, P. *Cell* **23**, 183-189 (1981).

28. Pettersson, R. F. *Proc. natn. Acad. Sci. U.S.A.* **78**, 115-119 (1981).

29. Martin, J. D., Riggsby, W. S., & Beck, R. W. *Archives Virol.* **60**, 131-146 (1979).

30. Zimmem, D. *Cell* **11**, 463-482 (1977).

Appendix

Comparison of the tetranucleotides preceding the initiation codons of eucaryotic mRNAs

Table 1. The tetranucleotides preceding the initiation codons of eucaryotic mRNAs[a]

mRNA	Sequence	Reference
CACC group		
1. SIN 26S RNA	CACC	chapter 5
2. SFV 26S RNA	CACC	chapter 5
3. MBV 26S RNA	CACC	chapter 5
4. Chicken ovalbumin	CACC	1
5. Human α-globin	CACC	2
6. Human β-globin	CACC	2
7. Human δ-globin	CACC	3
8. Rabbit α-globin	CACC	2
9. Rat preprocarboxypeptidase A	CACC	4
10. Mouse β-globin	CAUC	2
11. Chicken brain β-tubulin	CAUC	5
12. VSV-M protein	CAUC	6
13. AMV-RNA 4	CAUC	6
14. VSV-G protein	CACU	6
15. Rainbow trout protamin	CACU	7
16. Reovirus s45	CACU	2
17. SFV 49S RNA	CACG	chapter 6
18. Angler fish insulin	CAGC	8
CAAA group		
19. MBV 49S RNA	CAAA	chapter 6
20. EEE 49S RNA	CAAA	chapter 6
21. VEE 49S RNA	CAAA	chapter 6
22. VSV N-protein	CAAA	6

Table 1 (cont.)

23. Mouse liver amylase	CAAA	9
24. Mouse salivary amylase	CAAA	9
25. Influenza hemagglutinin	CAAA	10
26. FPV-NA	CAAA	11
27. FPV-HA	CAAA	11
28. FPV-P_3	CAAA	11
29. Rat insulin[b]	CAAC	12
30. Human fibroblast interferon[b]	CAAC	13
31. TYMV coat protein[b]	CAAC	2
32. Drosophila heat shock protein[b]	CACA	14
33. Human serum albumin[b]	CACA	15
34. SIN 49S RNA[b]	CACA	chapter 6
35. Poliovirus	CAUA	16
36. FPV-$NS_{1,2}$	CAUA	11
37. FPV-P_2	CAAU	11
38. Rabbit β-globin	CAGA	2
<u>AA group</u>		
39. RRV 26S RNA	AAAC	chapter 5
40. Drosophila 68C group IV	AAAC	17
41. Drosophila 68C group II	AACC	17
42. Drosophila 68C group III	AACC	17
43. Mouse α-globin	AACC	2
44. Reovirus m52	AAAG	2
45. FPV-M protein	AAAG	11
46. FPV-L protein	AAUC	11
47. BMV-RNA 4	AAUA	18

Table 1 (cont.)

UA group		
48. VSV-NS protein	UAUC	6
49. FPV-NP	UAUC	11
50. Mouse germ-line V_H gene	UACC	19
51. TMV	UACA	2
52. STNV	UAAC	2
53. Human leucocyte interferon	UACG	20
GA group		
54. Bovine corticotropin-β-lipotropin	GAAG	21
Others		
55. Human chorionic gonadotropin	CGCC	22
56. Human insulin	UGCC	8
57. Human growth hormone	UGCA	23
58. Reovirus s54	CGCA	2
59. FPV-P_1	UUGA	11
60. SV40-VP1	GCUU	2
61. Reovirus s46	AGUU	2
62. Reovirus m30	GGUC	2
63. Reovirus m44	GGUC	2
64. BMV-RNA3	CCCG	24
65. Bovin adrenal preproenkephalin	CCCC	25

a. Abbreviations: SIN, Sindbis virus; SFV, Semliki Forest virus; VSV, vesicular stomatitis virus; AMV, alfalfa mosaic virus; MBV, Middelburg virus; EEE, eastern equine encephalitis virus; VEE, Venezuelan equine encephalitis virus; FPV, fowl plaque virus; TYMV, turnip yellow mosaic virus; RRV,

Table 1 (cont.)

Ross River virus; BMV, Brome mosaic virus; TMV, tobacco mosaic virus; STNV, satellite tobacco necrosis virus; SV-40, Simian virus-40.

b. These mRNAs can also be put into CACC group.

Tetranucleotides preceding the initiation codons of 65 eucaryotic mRNAs are listed in Table 1. About 83% of these tetranucleotides have an A at the second position and about 58% of them begin with the sequence CA. These tetranucleotides were separated into six groups: CACC, CAAA, AA, UA, GA groups and the group whose second position is not A. Tetranucleotides of CACC and CAAA groups have at least three nucleotides identical to their representative sequences. G is found rarely in the last two positions of these two groups.

Statistically, if the sequences of these tetranucleotides are randomly arranged, only 25% of them would be expected to have an A at the second position and slightly less than 7% ($\frac{1}{4}x\frac{1}{4}$) of them would begin with CA. The 26S RNAs of Sindbis virus (SIN) and Semliki Forest virus (SFV) have 81 and 56 AUGs, respectively (26, 27, 28). Excluding the initiation codons, in SIN only 26% of the tetranucleotides preceding AUGs have an A at the second position and only 13% of them begin with CA; in SFV 22% of them have an A at the second position, but only 5% of them begin with CA. Because these figures are close to those predicted, the results shown in Table 1 suggest that tetranucleotides preceding initiation codons might be important in the initiation of translation.

To test this hypothesis, tetranucleotides preceding those AUGs in the 5'-untranslated regions of several eucaryotic mRNAs were investigated. The 5'-untranslated region of poliovirus mRNA has seven unused AUGs. None of the tetranucleotides preceding them have an A at the second position. The initiation codon of poliovirus is preceded by CAUA. Similar observations are also found in amylase mRNAs (9) and the 49S RNAs of eastern equine encephalitis viruses and SFV (chapter 6), in which, tetranucleotides preceding 5' unused AUGs do not have an A at the second position, but those preceding initiation codons begin with CA. This is also true for SV40-VP1 mRNA (2), except that the initiation codon, in this case, is not preceded by NANN.

Consensus sequences in procaryotic mRNAs which could base-pair with the 3'-end of 16S rRNA and help to position ribosomes on the mRNAs have been reported (29). Similar kinds of sequences which could base-pair with the 3'-end of 18S rRNA are not found in eucaryotic mRNAs. From the observations discussed above it seems likely that for at least some eucaryotic mRNAs, especially those in the CACC and CAAA groups, the tetranucleotides preceding the initiation codon are somehow important in the selection of AUG codon which initiates the protein synthesis.

References

1. McReynolds, L., O'Malley, B. W., Nisbet, A. D., Fothergill, J. E., Givol, D., Fields, S., Robertson, M. & Brownlee, G. G. (1978) Nature **273**, 723-728.

2. Kozak, M. (1978) Cell **15**, 1109-1123.

3. Spritz, R. A., DeRiel, J. K., Forget, B. G. & Weissman, S. M. (1980) Cell **21**, 639-649.

4. Quinto, C., Quiroga, M., Swain, W. F., Nikovits, W. C., Standring, D. N., Pictet, R. L., Valenzuela, P. & Rutter, W. J. (1982) Proc. Natl. Acad. Sci. U.S.A. **79**, 31-35.

5. Valenzuela, P., Quiroga, M., Zaldivar, J., Rutter, W. J., Kirschner, M. W. & Cleveland, D. W. (1981) Nature **289**, 650-655.

6. Rose, J. K. (1978) Cell **14**, 345-353.

7. Jenkins, J. R. (1979) Nature **279**, 809-811.

8. Hobart, P. M. Shen, L. P., Crawford, R., Pictet, R. L. & Rutter, W. J. (1980) Science **210**, 1360-1363.

9. Hagenbuchle, O., Tsoi, M., Schibler, U., Bovey, R., Wellauer, P. K. & Young, R. A. (1981) Nature **289**, 643-646.

10. Porter, A. G., Barber, C., Carrey, N. H., Hallewell, R. A., Threlfall, G. & Emtage, J. S. (1979) Nature **282**, 471-477.

11. Robertson, J. S. (1979) Nucleic Acids Res. **6**, 3745-3757.

12. Bell, G. I. Swain, W. F., Pictet, R., Cordell, B., Goodman, H. M. & Rutter, W. J. (1979) Nature **282**, 525-527.

13. Derynck, R., Content, J., DeClercq, E., Volckaert, G., Tavernier, J., Devos, R. & Fiers, W. (1980) Nature **285**, 542-547.

14. Ingolia, T. D., Craig, E. A. & MaCarthy, B. J. (1980) Cell **21**, 669-679.

15. Dugaiczyk, A., Law, S. W. & Dennison, O. E. (1982) Proc. Natl. Acad. Sci. U.S.A. **79**, 71-75.

16. Racaniello, V. R. & Baltimore, D. (1981) Proc. Natl. Acad. Sci. U.S.A. **78**, 4887-4891.

17. Garfinkel, M. (personal communication).

18. Efstratiadis, A., Kafatos, F. C. & Maniatis, T. (1977) Cell **10**, 571-585.

19. Ollo, R., Auffray, C., Sikorav, J. L. & Rougeon, F. (1981) Nucleic Acids Res. **9**, 4099-4109.

20. Taniguchi, T., Mantei, N., Schwarzstein, M., Nagata, S., Muramatsu, M. & Weissmann, C. (1980) Nature **285**, 547-549.

21. Nakanishi, S., Inoue, A., Kita, T., Nakamura, M., Chang, A. C. Y., Cohen, S. N. & Numa S. (1979) Nature **278**, 423-427.

22. Fiddes, J. C. & Goodman, H. W. (1979) Nature **281**, 351-356.

23. DeNoto, F. M. Moore, D. D. & Goodman, H. W. (1981) Nucleic Acids Res. **9**, 3719-3730.

24. Ahlquist, P., Dasgupta, R., Shih, D. S., Zimmern, D. & Kaesberg, P. (1979) Nature **281**, 277-282.

25. Noda, M., Furutami, Y., Takahashi, H., Toyosato, M., Hirose, T., Inayama, S., Nakamishi, S. & Numa, S. (1982) Nature **295**, 202-206.

26. Rice, C. M. & Strauss, J. H. (1981) Proc. Natl. Acad. Sci. U.S.A. **78**, 2062-2066.

27. Garoff, H., Frischauf, A. -M., Simons, K., Lehrach, H. & Delius, H. (1980) Proc. Natl. Acad. Sci. U.S.A. **77**, 6376-6380.

28. Garoff, H., Frischauf, A. -M., Simons, K., Lehrach, H. & Delius, H. (1980) Nature **288**, 236-241.

29. Shine, J. & Dalgarno, L. (1975) Nature **254**, 34-37.