Volume 16 Number 13 1988 Nucleic Acids Research

Analysis of the complete nucleotide sequence of the group IV RNA coliphage SP

Yoshio Inokuchi*, Ann B.Jacobson[2], Tadaaki Hirose[1], Seiichi Inayama[1] and Akikazu Hirashima+

Department of Molecular Biology and [1]Pharmaceutical Institute, School of Medicine, Keio University, Shinjuku-ku, Tokyo 160, Japan and [2]Department of Microbiology, SUNY at Stony Brook, Stony Brook, NY 11794-8621, USA

Received December 7, 1987; Revised and Accepted May 11, 1988

ABSTRACT

We report the nucleotide sequence of the Group IV RNA bacteriophage SP. The entire sequence is 4276 nucleotides long. Four cistrons have been identified by comparison with the related Group III phage Qβ. The maturation protein contains 449 amino acids, the coat protein contains 131 amino acids, the read-through protein contains 330 amino acids and the replicase β-subunit contains 575 amino acids. SP is 59 nucleotides longer than Qβ. We have analyzed both sequence and structural conservation between SP and Qβ and shown that the sequences for the coat and central region of the replicase are strongly conserved between the two genomes. We also show that the S and M replicase binding sites of Qβ are strongly conserved in SP. Interestingly, the base composition of SP and Qβ differ significantly from one another , and most of the differences can be accounted for by a strong preponderance of U in the third position of each codon of Qβ relative to SP. We also compare conserved hairpins associated with potential coat protein and replicase binding sites.

INTRODUCTION

RNA phages of *Escherichia coli* are single-stranded viruses in which genomic RNA functions as mRNA. For a recent review on these coliphages see Van Duin (1). A large number of RNA coliphage strains have been islated from sewage extracts in Japan, Southeast Asia, and Latin America (2,3). Sixteen of those strains have been studied in detail. By serological criteria, they form four distinct groups. The groups differ from one another in the length of their generation time, and the temperature at which they replicate optimally (2) They also differ in the molecular weight of their genomic RNAs and in the size of their proteins (4). Sequence analysis at the 3' end of these RNAs showed that sequences of strains within the same group differ from one another by approximately ten percent; greater sequence divergence was observed between different groups (5).

Two RNA coliphage strains, MS2 (and its close relatives f2, R17 and M12) and Qβ, have been studied extensively. MS2 belongs to the Group I phage type and Qβ belongs to the Group III phage type. Both the nucleotide sequence and the genetic map of MS2 and Qβ are known (6,7) and the genome of Qβ is

approximately 20% larger than the genome of MS2. Recently we reported the nucleotide sequence of GA, a Group II RNA coliphage that is relatively closely related to MS2 (8). Here we report the sequence of SP, a Group IV coliphage that is more closely related to Qβ. The sequence analysis was undertaken, in part, to obtain more information concerning the evolutionary relationships between the viral groups. In addition, since comparative studies with related sequences can be used to identify conserved regions that are likely to have functional significance, we are using these sequences to look for conserved structure both in the viral RNAs and in the proteins. In this manuscript we compare the structure of Qβ and SP and focus on a comparative analysis of sequences associated with viral replication. We discuss conservation of sequence within the viral replicase gene itself, examine conservation of replicase binding sites between SP and Qβ, and also examine the conservation of structure in the midivariant strain MDV-1. We describe briefly some mutagenic studies in one conserved region have evolved from the above-mentioned studies. A more detailed comparative analysis of the four known RNA coliphage nucleotide sequences will be presented elsewhere.

METHODS

Bacteria, bacteriophages and plasmids

Escherichia coli strain A/λ was used for growing of RNA phage SP. *E. coli* strain HB101 was used as recipient cells for cDNA cloning of SP RNA and strain JM103 was used for subcloning of SP cDNA clones to analyze the nucleotide sequence of the cloned cDNAs. RNA coliphage SP was grown in *E. coli* A/λ in peptone-glucose medium supplemented with 0.25% yeast extract and 10 mM $CaCl_2$, and then purified as described previously (8). SP RNA was extracted from the purified SP phage particles (4). Bacteriophage M13mp8, mp9, mp18 and mp19 RFI DNA were obtained from PL Biochemicals. Plasmid pBR322 DNA was purchased from Boehringer Mannheim.

Chemicals and enzymes

High specific activity grade [α-^{32}P]dATP and [^{3}H]dTTP were purchased from the Radiochemical Center (Amersham). Four kinds of dideoxynucleotides and M13 single-stranded sequencing primer were obtained from PL Biochemicals. Calf intestine alkaline phosphatase was purchased from Boehringer Mannheim. The Klenow fragment of *E. coli* DNA polymerase I, terminal deoxynucleotidyl transferase, T4 DNA polymerase, T4 DNA ligase and nuclease BAL-31 were purchased from Bethesda Research Laboratories. The reverse transcriptase from avian myeloblastosis virus and RNasin were obtained from Seikagaku Kogyo and nuclease S1 from Sankyo. Restriction nucleases were purchased from Bethesda Research Laboratories, Promega Biotec, Boehringer Mannheim, New England Biolabs, Nippon Gene, PL Biochemicals and Toyobo.

Construction of cDNA clones

Polyadenylation of SP RNA and isolation of the polyadenylated RNA were carried out as described previously (5). The reaction conditions for synthesizing the complementary DNA (cDNA) to SP RNA and annealing the cDNA with pBR322 DNA were the same as previously described (8). Transformation was carried out according to Ruther et al (9). The identification of clones carrying chimera plasmid DNA molecules was done by the method of Birnboim and Doly (10).

Subcloning cloned cDNA into phage M13 DNA and DNA sequencing

The cloned cDNA was digested by restriction enzymes and the DNA fragments were purified by gel electrophoresis. The fragments were inserted into M13 RFI DNA digested at the various cloning sites. Identification of the recombinants and preparation of single-stranded DNA for sequencing were done according to Sanger et al., (11). DNA sequencing was done essentially according to Messing (12). Gel electrophoresis was performed as described by Maxam and Gilbert (13) or Sanger et al. (14).

Sequencing of SP RNA by reverse transcription

Thirty nucleotides at the 5' end were determined by reverse transcription of SP RNA as previously described (5) using the oligonucleotide primer (5'-ATTCTCCTCTGTAGTGC-3').

Computer analysis

A variety of different computer programs were used in the analysis of the sequences. Protein sequence and nucleic acid comparisons were done using a 'Needleman-Wunsch' type of algorithm that was developed by Michael Zuker, National Research Council of Canada, Ottawa, Canada. For studies of protein similarity, weights were assigned to matched amino acids using the PAM 250 mutation matrix of Dayhoff (15). Gap penalties were treated as described in Fitch et al. (16). The programs used were all written in Fortran or Pascal. Sequence alignments were performed on an IBM mainframe computer. Computations which generated graphic output were obtained with a Scientific Microsystem DS11X microcomputer with an LSI 11/23 processor. Dot plot analysis was done according to White et al. (17). Plotting was performed with a DEC VT100 monitor equipped with Retrographics Terminal Enhancement, and hard copy was obtained with a Watanabe WX4633 XY plotter.

RESULTS

Determination of the complete nucleotide sequence of the SP genome

To determine the nucleotide sequence of the SP genome, we first synthesized DNAs complementary to SP RNA, and cloned them into pBR322 as described in the Methods section. Two clones (pSP131 and pSP219) were chosen for sequence analysis. The clones covered more than 99% of the SP

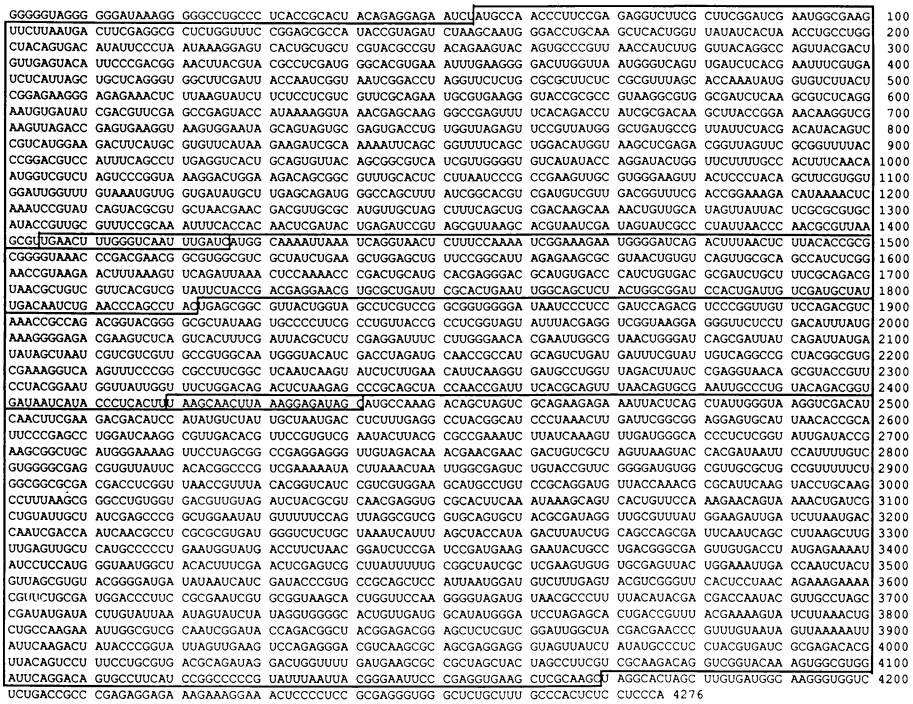


Figure 1. The complete nucleotide sequence of SP RNA. Numbering starts from the 5' end of the RNA. The solid lines indicate the regions of the four viral genes; maturation (A2), coat (C), readthrough (A1), and replicase β-subunit (R) Hyphens have been omitted from the sequence for clarity. There are differences at eleven positions in the sequence between the present data and the sequence at the 3' end of SP RNA that was published previously (5). The reason for the discrepancies is unknown. The original autoradiograms have been reexamined and the interpretation of both the current and the original sequence has been confirmed. It should be noted that the original determination was made by reverse transcription of polyadenylated SP RNA, rather than a cloned DNA derivative.

genome, but lacked 30 bases at the 5' terminal region of the sequence. These bases were sequenced directly by reverse transcription of SP RNA (5). For sequencing we digested the cloned DNA with appropriate restriction enzymes, and subcloned the DNA fragments into phage M13 RFI DNA. Thirty clones were used and 4246 bases were determined by the dideoxynucleotide sequencing method (14). All of the genome was sequenced at least twice and in both directions. The sequence of SP is presented in Figure 1. It contains 4276 nucleotides and is 59 nucleotides longer than the sequence of Qβ RNA (7). The identification of the viral genes in the SP sequence is based on known

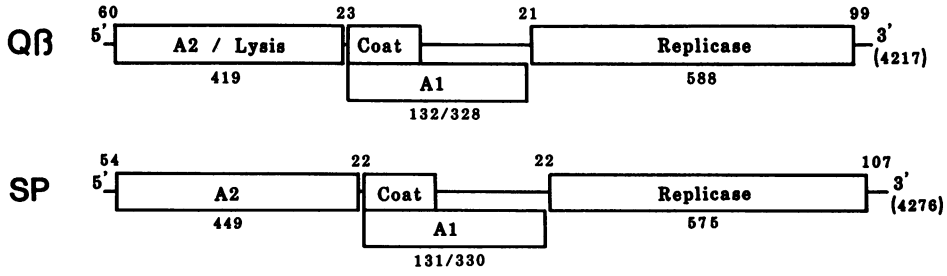


Figure 2. Genomic organization of the RNA phages Qβ and SP. The data for Qβ are from Mekler (7) and have been updated according to Billeter (personal communication). The length of the nontranslated regions are indicated in nucleotides above each map. The length of the viral proteins are given in amino acids beneath each map. These values do not include the initiating formylmethionine.

properties of some of the viral proteins, and in part by analogy with the Qβ sequence. Four protein products of the SP genes have been characterized by gel electrophoresis and shown to be similar to those of Qβ (4). These are A2 (maturation protein), coat, A1 (the coat readthrough protein) and replicase (the b-subunit of the viral replicase). In Qβ three of the proteins (coat, maturation, and readthrough) are found in the virion. The coat protein is the major constituent. The maturation protein is found at low concentration and is essential for binding of the virus to host F pili. A1 is also found at low concentration in the virion. Its function is less clear, although it has been shown to be essential for the formation of infectious particles (18). The coat protein gene in SP was identified on the basis of known carboxy and aminoterminal sequences (19). The A1 protein was identified as being the next large open reading frame downstream from the coat protein. The maturation and replicase gene were identified from the large open reading frames that are flanked by Shine-Dalgarno sequences that are consistent with the size of the viral proteins as determined by gel electrophoresis.

Genetic maps of SP and Qβ are shown in Figure 2. In SP, the maturation protein (A2) contains 449 amino acid residues, the coat protein has 131 amino acids, and the replicase protein has 575 amino acids. The maturation protein of SP has 30 amino acids more than the maturation protein of Qβ, and the replicase gene of SP has 13 amino acids less than the replicase gene of Qβ. Small differences can also be observed in the size of the non-translated regions.

Comparison of the sequence of SP with the sequence of Qβ

Protein coding regions

In Figure 3 we show a dot plot comparing the nucleotide sequence of SP and Qβ. The strongest sequence conservation between these two phages occurs in

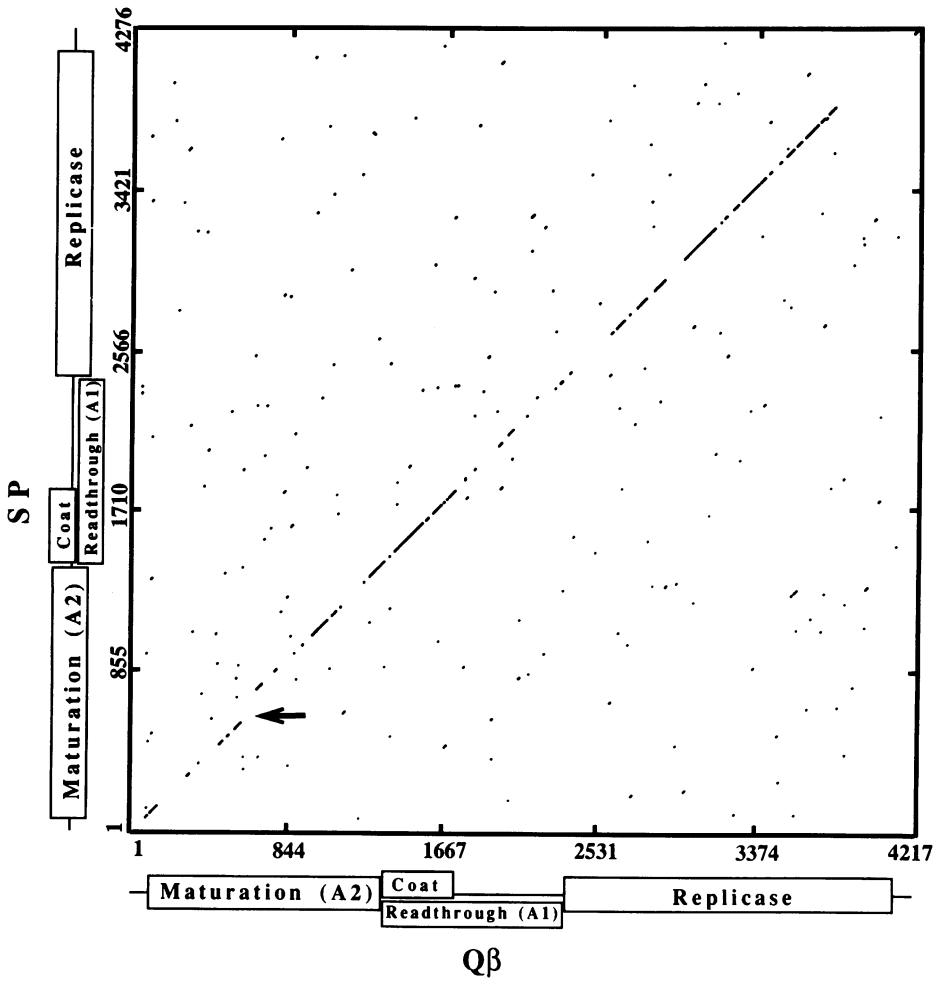


Figure 3. Dot plot showing regions of homology between SP and Qβ RNA. A single dot indicates the center of each window. A continuous line represents the overlap of several windows. The stringency was set at 14 matches in a window of 20 nucleotides. Numbers start from the 5' end of each RNA. The capital letters A2, C, A1 and R represent maturation, coat, readthrough, and replicase subunit, respectively. The arrow indicates a deletion in Qβ RNA.

the coat and in the center of the replicase genes. The strongest sequence divergence is seen in the region of the maturation gene and in the carboxy terminus of the readthrough protein.

In Figure 4 we show the alignment of viral proteins of Qβ and SP. The alignments show both exact matches and conservative amino acid changes. Consis-

tent with the results shown in Figure 3, the greatest similarity can be observed in the coat protein genes (Fig. 4B), where 80% of the amino acids are identical between the two sequence. The center of the replicase gene is also well conserved. A region with 29 perfect matches begins with amino acid 206 of the SP sequence and a region with 24 perfect matches begins at amino acid 313. Sequence divergence is observed in SP and Qβ replicase genes beginning at amino acid 513 of the SP sequence. A series of short deletions totaling 13 amino acids are observed in this region of the RNA. Sequence divergence is also seen in maturation protein (Fig. 4A) where only 48% of the amino acids are identical between the two sequences and two adjacent insertions of 18 and 13 amino acids are found in the SP sequence beginning at position 185. Sequence divergence also occurs in the readthrough protein (Fig. 5B) in the region that lies downstream of the coat coding region. Only 43% of the 200 amino acids that occur in this region of the sequence are identical, compared with 80% in the coat coding region. Although the coat protein genes of SP and Qβ are well conserved at the amino acid level, 58% of the conserved codons show third position base changes at the nucleotide sequence level. While the frequency of third position changes is not surprising, almost half of the differences are changes to U in Qβ. This pattern is observed in the other viral genes as well.

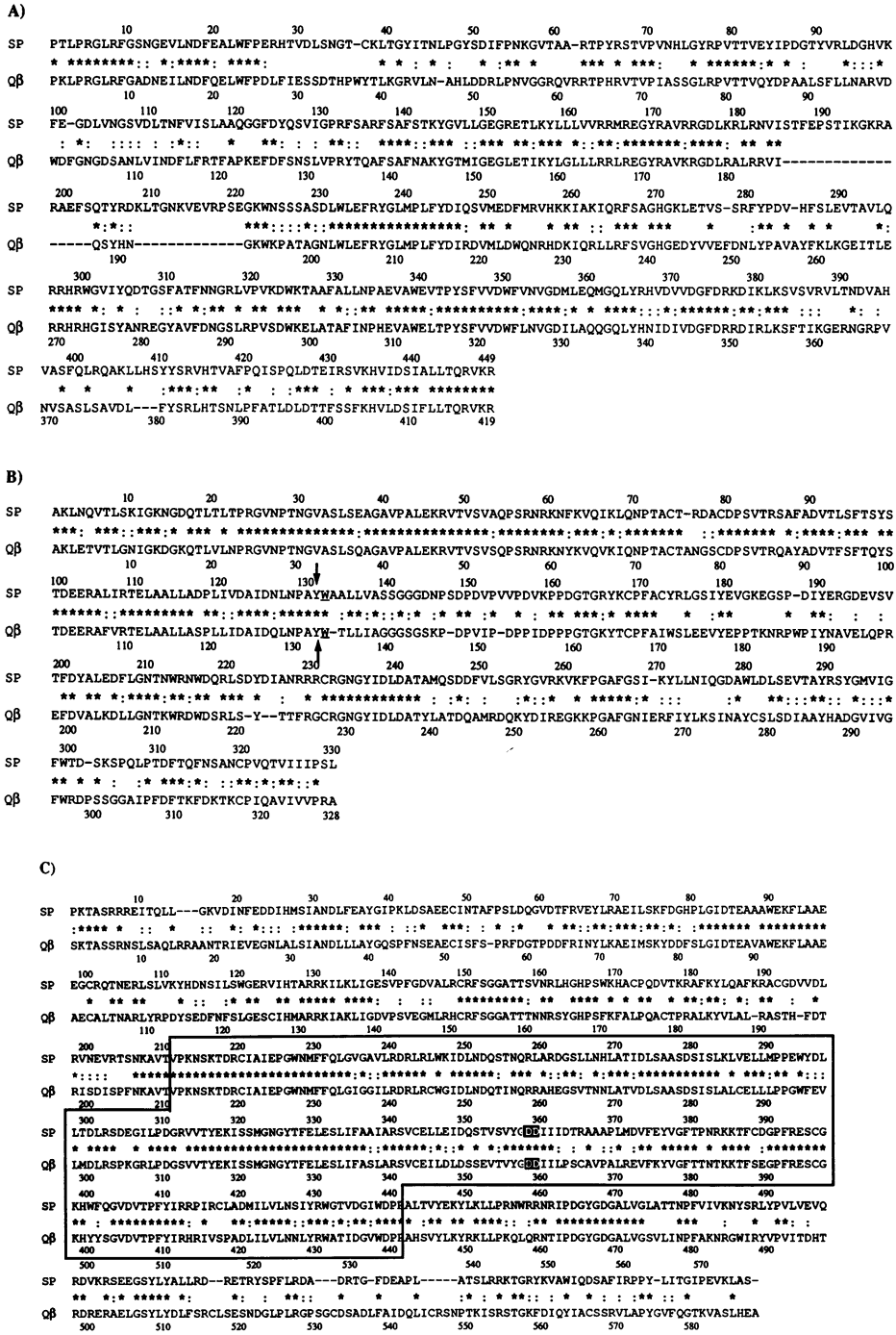
It has been known for some time that the base distribution of Qβ is asymmetric and is unusually high in U (20). This is not observed in SP. In Figure 5 we show the relative frequency of A and U along the sequences of SP and Qβ RNAs. The differences between the two sequences are striking. In Qβ, large regions of the sequence are very rich in U , whereas in SP both nucleotides are found in roughly equal frequency. Both the asymmetric distribution of bases as well as the rather high frequency of U suggest that the global secondary structure of Qβ RNA will be less stable than that of SP (see discussion below). In Table 1 we compare the relative distribution of U between SP and Qβ in the first, second, and third position of each codon of the individual viral genes. All of the differences in U between SP and Qβ are localized to the third position of each codon.

Non-translated regions

In Figure 6 we show the alignment of the nucleotide sequences of the 5' and 3' nontranslated regions of SP and Qβ. As described previously, the last 35 nucleotides of the 3' end of the sequences are identical (5). The 5'end of the sequences are not well conserved and will be discussed below.

Hairpins in the replicase initiation region

The coat proteins of the RNA bacteriophages are known to serve as translational repressors of the viral polymerase gene. The binding site or translational operator contains a small hairpin of 21 (MS2, R17) or 22 (Qβ) nucleotides that


A)
SP
Qβ
B)
SP
Qβ
C)
SP
Qβ

includes the initiation codon for the viral replicase gene (21,22). The structural features of the hairpin that are required by R17 coat for binding have been studied in detail by Uhlenbeck and collaborators (23). Essential features include an upper stem containing 2 base pairs and a lower stem containing 5 base pairs that is separated from the upper stem by a bulge loop consisting of a single adenine. The hairpin loop at the top of the upper helix contains 4 unpaired nucleotides. The base composition of the single-stranded regions of the hairpin are important and changes in several nucleotides alter the binding constant of coat protein to the hairpin. A uridine in position -5 of this loop is particularly important and is believed to interact with cysteine in the viral coat (24). The comparable hairpins in SP and Qβ are shown in Figure 7. The structures have been drawn in a manner that maximizes their similarity with the R17 hairpin. As seen in the figure, these hairpins differ from the one characterized in R17 in several ways. The lower helix in each stem has 4 rather than 5 base pairs . The hairpin loops are larger and do not contain uridine.

Studies to date have not explored the effect of loop or stem size on the interaction of the coat protein with the hairpin, nor is it known yet whether the binding properties of Qβ and SP coat differ from the binding properties of R17 coat. It should be noted that the upper helix of the hairpins could be drawn in an alternate form with four base pairs. They are not shown here because the calculated free energy of the hairpins is lower without them (see figure legend).

Replicase binding sites

We have examined the conservation of several regions of the sequence that have been implicated as replicase binding sites. Two internal replicase binding sites, the S and M sites have been studied extensively in Qβ (25,26). The S site precedes the coat initiation region and is believed to function by blocking ribosome binding at the coat initiation region. It is located between nucleotides 1247 and 1360, extending from the end of the maturation gene through the first 15 nucleotides of the coat gene. The M site is essential for the replication of the RNA. It contains three replicase binding fragments. The first is known as

Figure 4. Amino acid sequence similarities in the maturation (A2), coat, readthrough (A1), and replicase β-subunit proteins between SP and Qβ . (A) maturation, (B) coat and readthrough, (C) replicase β-subunit protein. Numbers start at the amino terminus of the protein. Each amino acid is represented as a single letter. Asterisks indicate identical amino acids. Colons indicates conservative amino acid changes that have a positive score in the Dayhoff PAM 250 mutation data matrix In each case the initiation formylmethionine is omitted. The arrow in Figure 4B indicates the end of the coat protein. The suppressed codon (UGA) in the A1 protein is underlined. The Asp-Asp segment conserved in RNA dependent polymerases from many viruses is shown in reverse contrast (31). Bars are used to represent probable insertion/deletion sites. The enclosed area indicates the regions of the two sequences that are conserved when they are aligned with GA and MS2. The analysis of homology was done using a 'Needleman-Wunsch' type of algorithm as described in the Methods section.

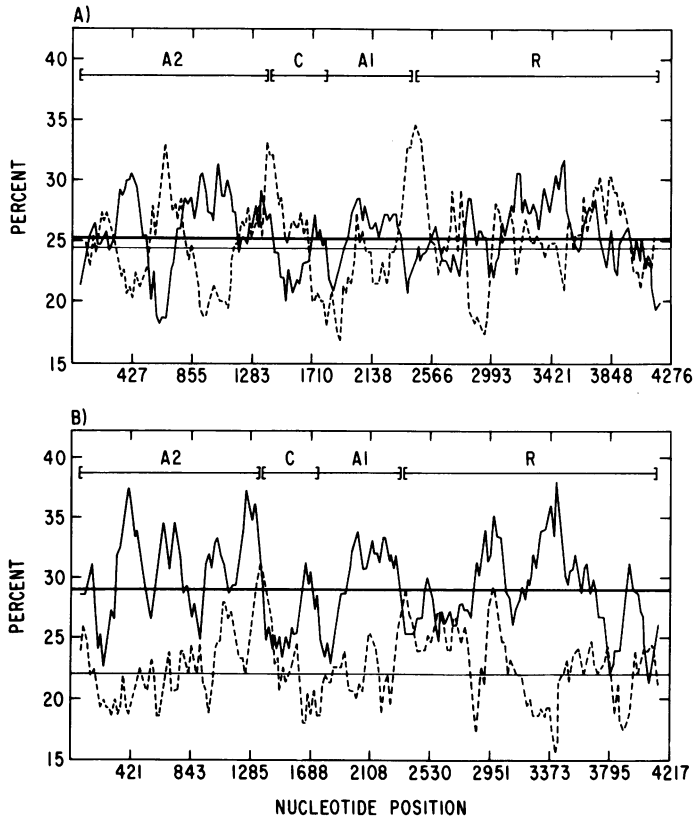


Figure 5. Relative distribution of adenine and uracil in the sequences of SP (A) and Qβ (B). The horizontal lines indicate the average composition of U (upper line) and A (lower line). The graphs were obtained by sliding a window of length 150 nucleotides along the sequence. The percent U (——) and the percent A (- - -) were calculated in each window. Points were taken at 20 nucleotide intervals. The relative position of the individual viral genes A2 (maturation), C (coat), A1 (readthrough) and R (replicase) are indicated at the top of each figure.

M5 and extends from nucleotides 2545 to 2613. The second region is called M2b and extends from nucleotides 2637 to 2811. The third region is M11 and extends from nucleotides 2844 to 2872. Studies by electron microscopy (27 and our own unpublished results) have shown that the S and M sites probably lie in close juxtaposition to one another in the folded RNA and may lie in the stem of a large open loop.

The alignment of the S and M regions of Qβ and SP are shown in Figure 8. Based on this alignment the S region of SP appears to extend from nucleotides

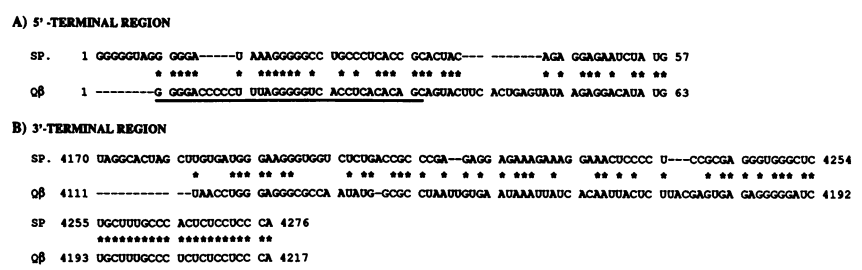


Figure 6. Comparison of the non-translated 5' and 3' regions of Qβ and SP RNA. The alignments were performed with a 'Needleman-Wunsch' algorithm. Match value = 1, mismatch value = 0, gap value = 2 + 0.05 x gap length. The underscored sequence in Figure 6A corresponds to a conserved region between Qβ and midivariant MDV-1 RNA. The asterisks indicate identical bases. Bars are used to represent probable insertion/deletion sites.

1331 to 1443. Approximately 80% of the nucleotides are conserved between the sequences The homologous M region in SP extends from nucleotides 2629 to 2956. The overall similarity between the two sequences in the M region is approximately 66%. A 42 nucleotide stretch beginning at nucleotide 2601 of the Qβ sequence is conserved almost exactly . This region overlaps the end of the M5 fragment and the start of the M2b region. Although the M region is located within the replicase gene, it precedes the central region of the gene where the sequence is strongly conserved at the amino acid level (see above). It includes one of three regions in the replicase gene where the nucleotide sequence is conserved exactly for more than 15 nucleotides. It should be noted that replicase binding has not been studied directly with SP RNA.

Although replicase binds primarily to the internal S and M regions of Qβ sequence, additional fragments at the 3' end of the RNA are protected from nuclease T1 digestion under initiation conditions (26). They are located at posi-

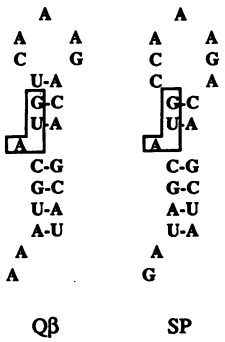


Figure 7. Proposed secondary structure models for the translational operator of Qβ and SP RNA. Boxes indicate the initiation triplets of the replicase β-subunit gene. The hairpin for Qβ is from (21). The calculated free energies for the hairpins are: Qβ -4.0 Kcal/mole, SP -2.3 Kcal/mole. Unpaired terminal nucleotides were not included in the calculation.

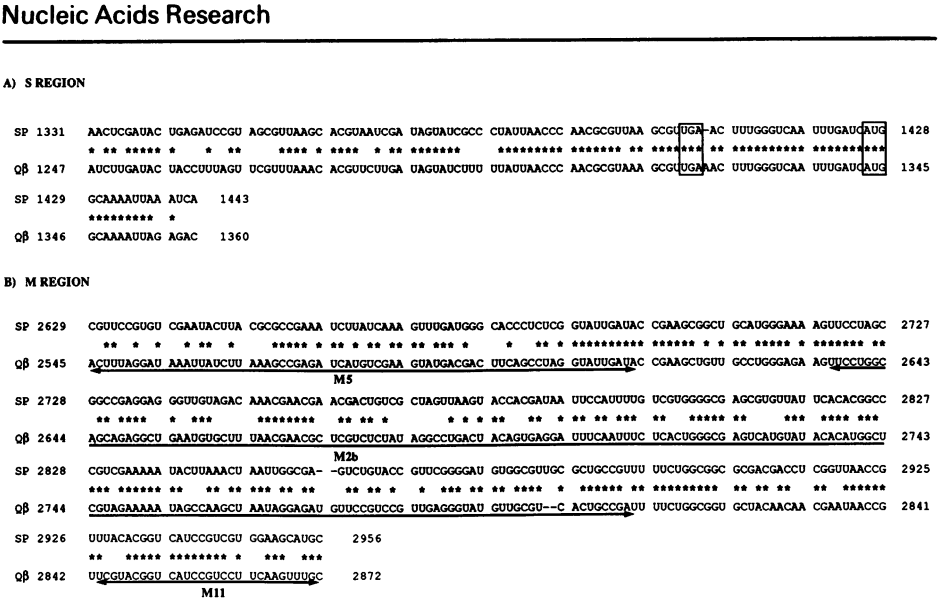


Figure 8. Alignments showing the homology between the S (A) and M (B) replicase binding sites of Qβ and SP. The positions of the replicase binding fragments M5, M2b, and M11 are indicated. Boxes indicate the termination (A2) and initiation (coat) triplets.

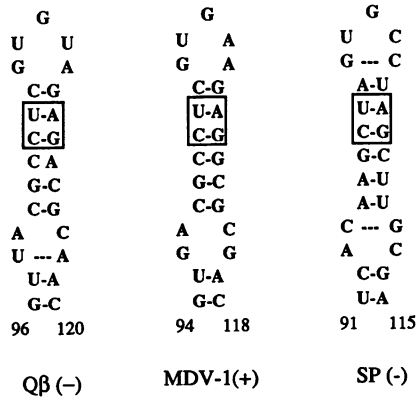


Figure 9. Proposed secondary structure for a hairpin that is conserved between Qβ, midivariant (MDV-1), and SP RNA. Dotted lines indicate base pairs that are not conserved between the three structures. The boxed areas indicate a stacked region that is conserved in all three hairpins. The calculated free energies for the hairpins are: Qβ -3.0 Kcal/mole, MDV -8.3 Kcal/mole, SP -7.7 Kcal/mole.

tions 3778-3795, 4155-4180, 4203-4217 (the 3' end of the Qβ sequence). We have examined sequence similarity in the corresponding regions of SP RNA. As discussed above, 35 nucleotides at the 3' terminus are conserved almost exactly between SP and Qβ, including the fragment at 4203-4217. No conservation of sequence is observed for the other two fragments.

In addition to the replicase binding regions of Qβ we also examined homology in regions that are conserved between MDV-1 (+) RNA and Qβ(-) RNA. MDV-1 RNA is a small naturally occurring template for Qβ replicase. Several years ago, Nishihara et al. (28) showed that a central hairpin within the sequence of MDV-1 RNA was required for replicase binding. The sequence of the hairpin loop is almost identical with nucleotides 84 to 127 of Qβ minus strand RNA. As shown in Figure 9, this hairpin loop is conserved in SP RNA as well. It is located between nucleotides 95 and 119 of the SP minus strand. Only 7 nucleotides are conserved between Qβ and SP in this region of the sequence. Four of them are paired and lie in the center of the upper stem of the hairpin. They are conserved in MDV-1 as well. The calculated free energy of the hairpins (see figure legend) shows that the SP and MDV-1 hairpins are more stable than the Qβ hairpin due to the presence of some additional base pairs (29). The conservation of the detailed structure of these three hairpins is intriguing and suggests that they may be associated with replicase binding in SP and Qβ. Similar but not identical hairpins have been identified in CT and microvariant RNAs (30). These small RNAs can also be replicated with Qβ replicase.

Another conserved region between MDV-1 and Qβ minus strand is located within the last 35 nucleotides at the 3' end of both RNAs. A short stable hairpin consisting of five Gs paired with five Cs can be identified in both sequences. The corresponding region in SP sequences is not well conserved (Figure 6A). Nine additional nucleotides appear to have been added to the 5' end of the SP sequence and a run of 5 C's that correspond to the stem of the small hairpin have been deleted. These differences could indicate differences in the specificity of the SP and Qβ replicases during the initiation of minus strand replication; alternatively, the hairpin may not have functional significance.

DISCUSSION

In the results section we presented the nucleotide sequence of SP (a Group IV RNA coliphage) and compared the sequence with the known sequence of Qβ (a member of Group III). The genome of SP is 59 nucleotides longer than Qβ, consistent with earlier physical studies that show that all Group IV phage are larger than the Group III RNA coliphages. In general, the difference in nucleotide size is not localized in one large insertion/deletion region, but is dis-

persed throughout the sequences of both RNAs. However, two major insertion/deletion regions have been identified. One of these occurs in the terminal region of the replicase gene of SP relative to Qβ. The other is located in the center of the maturation protein.

Studies with Qβ by other investigators have focused on the details of the mechanism of viral replication. Our comparative analysis has therefore focused on conserved features of the viral replicase gene as well as on the analysis of the replicase binding sites that are known from *in vitro* studies with Qβ replicase. An analysis of this type can provide direct insight into the functional significance of individual regions and help to generate models that can be tested further by genetic engineering. Purified viral replicase has been isolated both from Qβ and SP and the relative specificity of both proteins has been determined with several different viral RNA (31). Qβ replicase replicates its own template RNA approximately 2 to 3 times better than it replicates SP RNA. It shows no template activity with GA or MS2 RNA. Similarly, SP replicase shows higher activity with its own RNA than with Qβ RNA, and no replicase activity with Group I and Group II RNA.

Conserved regions in the coliphage replicases are likely to indicate common structural features like those required for protein subunit interaction, enzyme binding to the RNA and/or the functional site for chain elongation. As indicated in Figure 4C, the region beginning at amino acid 211 and extending through amino acid 441 of the SP sequence is well conserved in MS2 and GA as well as Qβ. The conserved region includes two aspartic acid residues (see Figure 4) that have been shown by Kamer and Argos (32) to be conserved in a large number of viral replicases. Recently we have begun to examine the sequence specificity of this region by site-directed mutagenesis in cloned Qβ DNA. Changes in the glycine residue preceding the aspartic acid residues completely abolish replicase activity. Interestingly, the mutations do not appear to affect replicase binding to Qβ RNA (33) and it is likely that this site is involved in chain elongation. Recent mutagenic studies with the conserved region in the AIDS viral polymerase in this same region support this interpretation and show that similar amino acid substitutions led to loss of reverse transcriptase activity of the protein *in vitro*. (34).

We can also look at sequence divergence between the replicase genes and between replicase binding sites to gain insight into the problem of template specificity. None of the carboxy-termini of the viral repllicases are conserved. The replicase proteins of GA and MS2 are approximately 10% shorter than the replicases of SP and Qβ. Again, most of the extra amino acids are located at the carboxy-termini. These differences suggest that the carboxy-terminus of the replicase could play a role in determining template specificity and might compensate for differences in the replicase binding sites that were described in

the Results section. It will be of interest to see whether this region of the protein can be deleted in cloned SP and Qβ replicase genes without effecting replicase specificity.

The sequence of the maturation or A2 protein shows the greatest divergence between SP and Qβ. In Group I phage the maturation protein has been shown to attach to F pili during viral infection, and to accompany the viral RNA into the cell. In Qβ the maturation protein has been shown to code for the lysis function (35,36). Although the SP maturation gene has been cloned, lysis activity could not be demonstrated. This result may indicate that the lysis activity of the maturation protein is relatively weak and would be consistent with the observation that SP cannot lyse the bacterial host Q13 that is normally used for growing Qβ. The sequence divergence between the maturation proteins of SP and Qβ, as well as the large insertion that is observed in the center of the SP gene, could account for the reduction or loss of lysis activity in the SP maturation protein. Alternatively, genetic information coding for lysis activity may reside in another region of the SP genome.

The A1 (readthrough) protein is believed to be a viral coat protein, since it is found in low concentrations in intact virions. Its function is not known, but it has been shown to be essential for the reconstitution of infectious viral particles. Mutants that are defective in the synthesis of the protein have never been isolated. As discussed in the results section the first part of the A1 gene also codes for viral coat and is well conserved between SP and Qβ, but little conservation is seen in the readthrough region. Not only does the sequences diverge, but codon usage is strikingly different between the coat and readthrough regions. Rare codons are used infrequently in the coat region, and relatively frequently in the readthrough region (data not shown). Evolutionary constraints thus appear to act differently on the individual regions of the protein.

The evolutionary relationships between the viral subgroups have been discussed recently by Furuse (2). He has proposed that Group IV viruses are progenitor viruses from which other groups have evolved. If Qβ has indeed evolved from an SP like progenitor, then the third position preference for U is intriguing and suggests that mutational bias may have occurred during the evolution of the phage. The divergence of the two viruses certainly occurred over time and the third position changes to U may be relatively recent within that time frame. The mechanism by which mutational bias could occur can only be speculated on. Conceivably alterations in the viral replicase have occurred so that uridine is incorporated preferentially when errors in replication occur. Alternatively, the virus may have replicated in an environment that was relatively rich in U and this base was therefore incorporated preferentially. An important consequence of this divergence, if it occurred, might be an overall weakening in the stability of the secondary structure of the RNA. Recent studies by Priano et al. (30)

have shown that the overall secondary forming capacity of midivariant RNAs are important in determining the rate at which the RNAs can replicate. Since SP has both a longer generation time and somewhat different temperature range than Qβ (37), it would be interesting to determine whether these differences can be correlated with the global folding potential of the individual RNAs.

Although the analysis we have presented here has focused primarily on large patterns of conservation between the sequences of SP and Qβ, it is also useful to examine some of the detailed differences that can be observed between residues that are not conserved. As seen in Figure 4, several amino acid changes within the conserved regions of both the viral coat and replicase proteins involve major changes both in charge and size of the side chains. These changes may provide important clues regarding the three dimensional folding of these molecules as well as the potential active sites within them.

ACKNOWLEDGMENTS

We thank Dr. M.A. Billeter (University of Zurich) for providing us the sequence data of Qβ phage. We thank Dr. I. Watanabe for critical discussions, Ms. H. Harigai for her help in determining the terminal amino acids of the coat protein, Mr. Y. Ishii for the help in synthesizing the deoxynucleotide primer, Dr. M. Zuker for making the program for Needleman and Wunsch alignment available to us, and Lawrence Chan for programming assistance. One of us (A.B.J) wishes to acknowledge several thoughtful discussions with Dr. Don Mills concerning the analyses reported here. This work was supported in part by grants from the Ministry, Education, Science and Culture of Japan, from the Ito Science Foundation to A.H., and NIH grants AI 15723 and GM 3842509 to A.B.J.

*To whom correspondence should be sent

+Present address: Genetic Engineering Laboratory, Yakult Central Institute for Microbiological Research, Kunitachi-shi, Tokyo 186, Japan

REFERENCES

1. Van Duin, J. (1988) In Frankel-Conrat, H. and Wagner, R. (eds.), The Viruses, Plenum Publishing, New York, in press.
2. Furuse, K. (1987) In Goyal, S.M., Gerba, C.P., and Bitton, G. (eds.), Phage Ecology, John Willey and Sons, Inc. Publishers, New York, p. 87-124.
3. Furuse, K., Sakurai, T., Hirashima, A., Katsuki, M., Ando, A. and Watanabe, I. (1978) Appl. Microbiol. 35 , 995-1002.
4. Furuse, K., Hirashima, A., Harigai, H., Ando, A., Watanabe, K., Kurosawa, Y., Inokuchi, Y., and Watanabe, I. (1979) Virology 97 , 328-341.
5. Inokuchi, Y., Hirashima, A. and Watanabe, I. (1982) J. Mol. Biol. 158: 711-730.
6. Fiers, W., Contreras, R., Duerinck, F., Haegeman, C., Iserentant, D., Merregaert, J., Min

Jou, W., Molemans, F., Raeymaekers, A., Vandenberghe, A., Volckaert, G. and Ysebaert, M. (1976) Nature 260, 500-507.
7. Mekler, P. (1981) Ph.D. thesis, University of Zürich, amended in 9 single nucleotide positions, private communication by Billeter, M.A.
8. Inokuchi, Y., Takahashi, R., Hirose, T., Inayama, S., Jacobson, A.B. and Hirashima, A. (1986) J. Biochem. 99 , 1169-1180.
9. Rüther, U., Koenen, M., Otto, K. and Müller-Hill, B. (1981) Nucl. Acids Res. 9, 4087-4098.
10. Birnboim, H.C. and Doly, J. (1979) Nucl. Acids Res. 7 , 1513-1523.
11. Sanger, F., Coulson, A.R., Barrell, B.G., Smith, A.J.H. and Roe, B.A. (1980) J. Mol. Biol. 143, 161-178.
12. Messing, J. (1983) In Methods in Enzymology, Wu, R., Grossman, L. and Moldave, K. (eds), Vol. 101, pp. 20-78. Academic Press, New York.
13. Maxam, A. and Gilbert, W. (1977) Proc. Natl. Acad. Sci. USA 74 , 560-564.
14. Sanger, F., Nicklen, S. and Coulson, A.R. (1977) Proc. Natl. Acad. Sci. USA 74 , 5463-5467.
15. Schwartz, R.M. and Dayhoff, M.O., In Dayhoff, M.O. (ed.), Atlas of Protein Sequence and Structure, Vol. 5, supp. 3, pp. 353-358, National Biomedical Research Foundation, Washington, D .C.
16. Fitch, W.M. and Smith, T.F. (1983) Proc. Natl. Acad. Sci. USA 80 , 1382-1386.
17. White, C.T., Hardies, S.C., Hutchinson, C.A. III and Edgell, M.H. (1984) Nucl. Acids Res. 12 , 751-766.
18. Hofstetter, H., Monstein, H.J. and Weissmann, C. (1974) Biochem. Biophys. Acta 374 , 238-251.
19. Hirashima, A., Harigai, H. and Watanabe, I. (1982) Microbiol. Immunol 26, 1089-1093.
20. Overby, L.R., Barlow, G.H., Doi, R.H., Jacob, M. and Spiegelman, S. (1966) J. Bacteriol. 92, 739-745.
21. Steitz, J.A. (1969) Nature 224, 957-964.
22. Weber, H. (1976) Biochim. Biophys. Acta 418 , 175-183.
23. Uhlenbeck, O.C., Carey, J., Romaniuk, P.J., Lowary, P.T. and Beckett, D. (1983) J. Biomolecular Structure and Dynamics 1, 539-552.
24. Romaniuk, P., Lowary, P., Wu, H.N., Stormo, G., and Uhlenbeck, O.C., (1987) Biochemistry 26, 1563- 1568.
25. Weber, H., Billeter, M.A., Kahane, S., Weissmann, C., Hindley, J. and Porter, A. (1972) Nature New Biol. 237, 166-170.
26. Meyer, F. (1978) Ph.D. thesis, University of Zürich.
27. Vollenweider, H.J., Koller, Th., Weber, H. and Weissmann, C. (1976) J. Mol. Biol. 101 , 367-377.
28. Nishihara, T., Mills, D.R. and Kramer, F.R. (1983) J. Biochem. 93 , 669-674.
29. Freier, S.M., Kierzek, R., Jaeger, J.A., Sugimoto, N., Caruthers, M.H., Nielson, T., and Turner, D.H. (1986) Proc. Natl.Acad. Sci 83, 9373-9377.
30. Priano, C., Kramer, F.R. and Mills, D.R. (1987) Cold Spring Harbor Symposium Quant. Biol. Cold Spring Harbor Laboratory, New York, in press.
31. Miyake, T., Haruna, I. Shiba, T., Itoh, Y.H. Yamane, K., and Watanabe, I. (1971) Proc. Nat. Acad Soi. USA 68, 2022-2024.
32. Kamer, G. and Argos, P. (1984) Nucl. Acids Res. 12, 7269-7282.
33. Inokuchi,Y. and Hirashima, A. (1987) J. Virol., 61, 3946-3949.
34. Larder, B.A., Purifoy, D.J.M., Powell, K.L. and Darby, G. (1987) Nature 327, 716-717.
35. Karnik, S. and Billeter, M.A. (1983) EMBO J. 2, 1521-1526.
36. Winter, R.B. and Gold, L. (1983) Cell 33, 877-885.
37. Furuse, K., (1982) J. Keio Med. Soc. 59, 265-274.