Proc. Natl. Acad. Sci. USA
Vol. 74, No. 3, pp. 827–831, March 1977
Biochemistry

Structure of a large segment of the genome of simian virus 40 that does not encode known proteins

(messenger RNA/DNA sequence/gene structure/transcription)

RAVI DHAR*, K. N. SUBRAMANIAN†, JULIAN PAN, AND SHERMAN M. WEISSMAN‡

Department of Human Genetics, Yale University School of Medicine, New Haven, Connecticut 06510

Communicated by Aaron B. Lerner, November 9, 1976

ABSTRACT **The nucleotide sequence of the region of DNA of simian virus 40 extending from 0.595 to 0.790 map unit has been derived. The sequence includes the DNA complementary to the 5′ end of early mRNA and to the 5′ end of some of the forms of late RNA. Because there are termination codons in all three phases in early and late RNA, there is a sequence of almost 800 nucleotides of simian virus 40 DNA that probably does not code for known viral proteins. The sequence spans the 5′ end of the early mRNA at 0.67 map unit and overlaps a species of late RNA whose 5′ end is at 0.65 map unit and whose 3′ end is at 0.77 map unit. This RNA is retained on oligo(dT)-cellulose columns in high salt concentrations. Analysis of the sequence of late strand RNA suggests that this RNA is not covalently linked to the mRNA that encodes structural proteins. There is another species of late RNA of simian virus 40 whose 5′ end is at 0.775 map unit.**

The nucleotide sequence of this region of simian virus 40 DNA contains several examples of repeated sequences, most of which are located in DNA that does not encode known peptides. These may be analogous to the reiterated sequences that have been described in animal cell DNA.

The genome of simian virus 40 (SV40) is a double-stranded DNA molecule that contains approximately 5200 base pairs. This DNA codes for three structural proteins: VP1, VP2, and VP3. The amino acid sequences of VP1 and VP2 are different, while VP3 contains many peptides in common with VP2 and may consist of a part of the longer protein (1). In addition, the viral DNA codes for the synthesis of an early protein, termed the A protein, which is thought to be closely related to the T antigen found in virus-induced tumors. The estimated total molecular weight of these proteins is in the range of 160,000, and approximately 4800 base pairs of DNA would be needed to code for their amino acid sequences. There is very little additional genetic information in the virus, and most of the steps in transcription, post-transcriptional processing of viral message, and replication of viral DNA must be performed by host proteins. The structure, mechanisms of expression, and replication of the viral DNA therefore should serve as a model for analogous processes and structures that occur in the uninfected host.

We and others have been analyzing in detail the nucleotide sequence of SV40 viral DNA, and the nucleotide sequence of over half the DNA is known. We report here that the distribution of termination codons in transcripts of the viral genome is such that a contiguous stretch of 17% of the viral DNA could not code for the A protein, VP1, or VP2. Transcription of the DNA sequence produces an RNA that, if translated, would direct the synthesis of peptides shorter than any of the known viral proteins. RNA complementary to this region of the DNA was found in cytoplasm of infected cells.

The sequence of this region of DNA is also of interest in that it includes the sequences overlapping the 5′ ends of early and late messenger RNA and the origin of DNA replication.

The DNA in this region contains some remarkable repeated sequences. We suggest that these may be analogous to reiterated sequences in host cell DNA.

MATERIALS AND METHODS

The source and preparation of the materials used for this work have been reviewed elsewhere (2–4).

The analysis of the sequence of SV40 DNA was performed primarily by transcription of restriction endonuclease fragments of the DNA, followed by RNA sequence analysis and by limited exonuclease digestion of terminally labeled DNA fragments (5). To determine the "strandedness" of sequences we prepared early strand cRNA by transcription of SV40 DNA with *Escherichia coli* RNA polymerase, and annealed this RNA to restriction fragments.

Late cytoplasmic RNA was prepared by infecting monolayers of VERO lines of continuous cultures of African green monkey cells with 10–20 infectious units per cell. The cells were exposed to [^{32}P]phosphate from 12 to 36 hr after infection. RNA was prepared and fractionated on oligo(dT)-cellulose (6) as described (2). The retained RNA was annealed to appropriate restriction endonuclease fragments of SV40 DNA immobilized on nitrocellulose filters. The protected RNA was analyzed by T1 RNase digestion and oligonucleotide analysis.

RESULTS

Analysis of the nucleotide sequence of this region of SV40 DNA was expedited by the number of restriction endonuclease cleavage sites it contains (7, 8) (Figs. 1 and 2). Analysis of the sequences from nucleotides 229 to 537 and from 538 to 592 is presented in detail elsewhere (ref. 9; K. N. Subramanian and S. M. Weissman, *Cell*, in press). The largest single uncleaved stretch was 157 nucleotides, from nucleotide 386 to nucleotide 537. The fragment from nucleotides 28 to 233 was cleaved at several points by the restriction endonuclease from *Moroxella bovis, Mbo* II, an enzyme that cleaves eight nucleotides from the recognition site G-A-A-G-A. This enzyme tended to produce partial digestion products, which presented some difficulties. Cytoplasmic ^{32}P-labeled RNA was prepared from cells late after infection with SV40. The fraction retained on oligo(dT)-cellulose at high salt concentration was then annealed to various restriction fragments spanning this region of SV40 DNA. The complementary RNA was digested with T1 or pancreatic ribonuclease and the resulting products were compared with those obtained from transcripts of the fragment prepared *in*

Abbreviation: SV40, simian virus 40.

* Present address: Laboratory of the Biology of Viruses, National Cancer Institute, National Institutes of Health, Bethesda, Md. 20014.
† Present address: Department of Microbiology, University of Illinois Medical Center, Chicago, Ill. 60680.
‡ To whom requests for reprints should be sent.

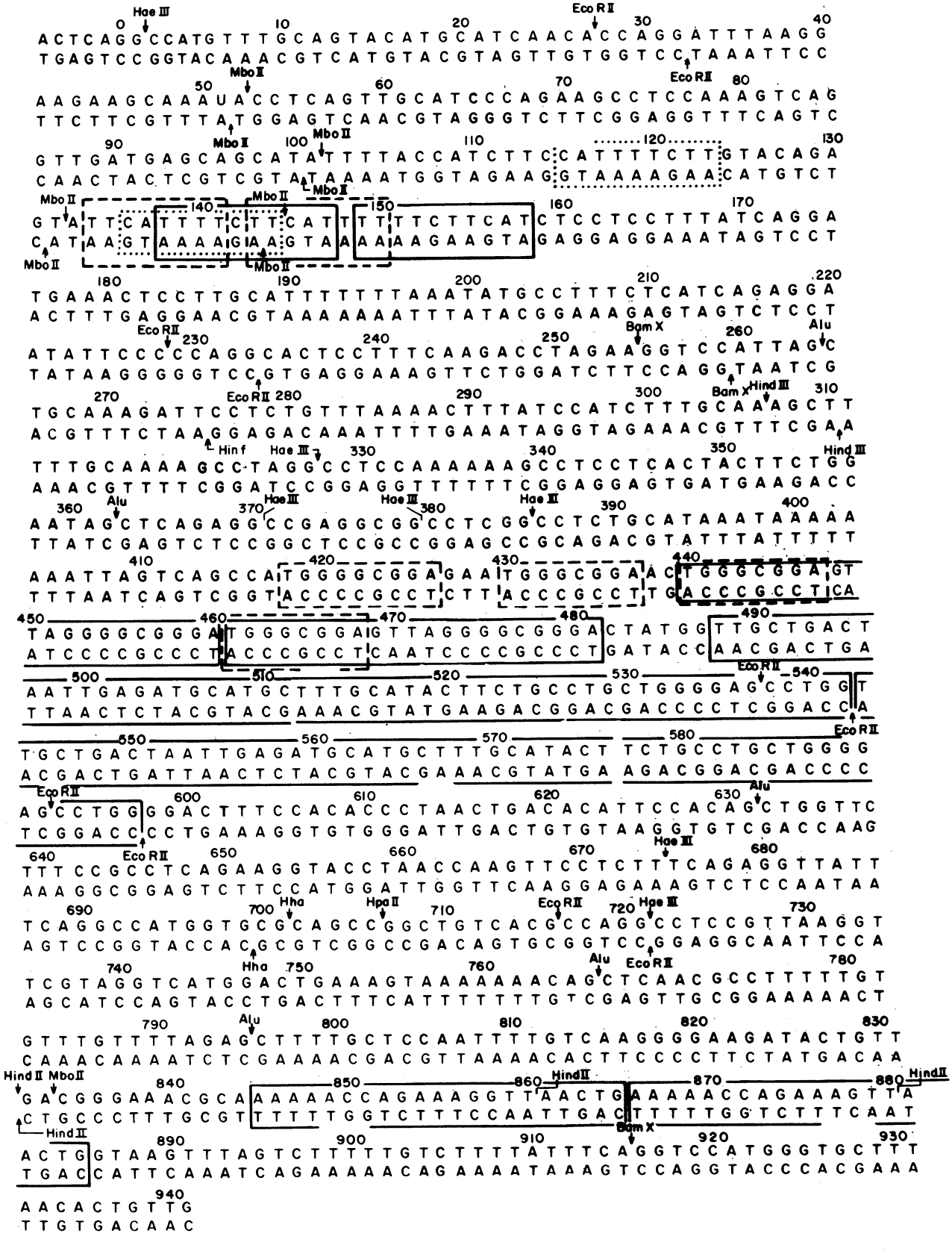

FIG. 1. Nucleotide sequence of a part of SV40 DNA. Arrows indicate restriction endonuclease cleavage sites. Abbreviations for restriction endonucleases are: *Hin*dII, *Hemophilus influenzae* serotype d; *Hin*dIII, *H. influenzae* serotype d; *Hin*f, *H. influenzae* serotype f; *Eco*RII, *Escherichia coli* containing class II plasmid; *Bam* X, *Bacillus* species purified by Dr. Frank Young; *Mbo*I, *Moraxella bovis;* *Mbo*II, *M. bovis;* *Hae* II, *H. aegyptius;* *Hae* III, *H. aegyptius;* *Hpa* II, *H. parainfluenzae;* *Hha*, *H. hemolyticus;* *Alu*, *Arthrobacter luteus.* Reiterated sequences are enclosed in solid boxes, dashed boxes, and dotted boxes.

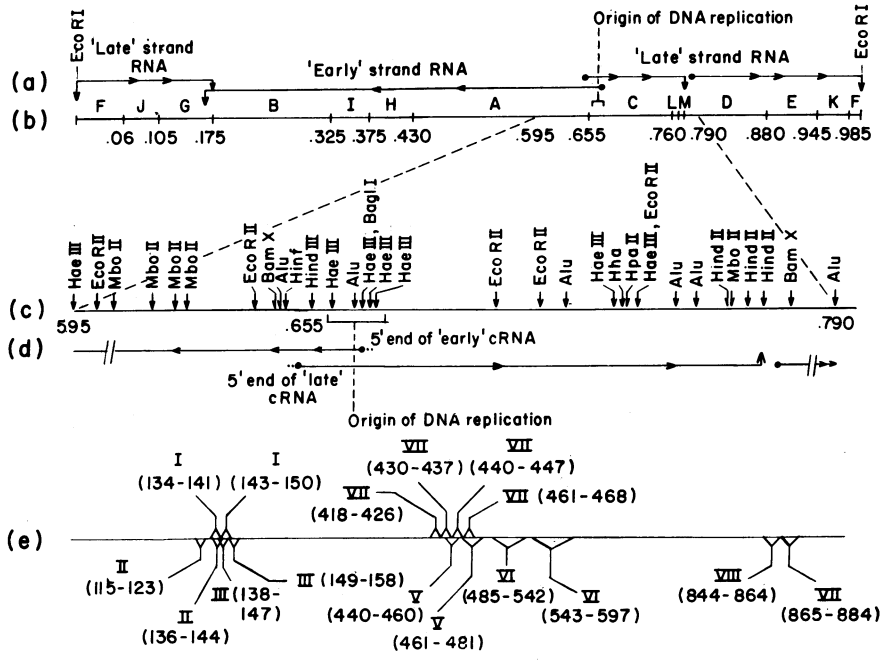

FIG. 2. Schematic sketch of the SV40 genome. (*a*) Regions of the genome where early-strand or late-strand transcripts are found in the cytoplasmic RNA retained on oligo(dT)-cellulose columns. (•) Approximate location of 5′ end of the cytoplasmic RNA. (–|) Approximate location of the last T1 oligonucleotide found toward the 3′ end of the cytoplasmic RNA. (*b*) *Hin*dII+III restriction endonuclease cleavage map of SV40 DNA. Numbers represent the fractional length of the genome. (*c*) Expanded sketch of the segment of the genome corresponding to the nucleotide sequence shown in Fig. 1. The cleavage sites by various restriction endonucleases are shown (see legend to Fig. 1). Numbers refer to the fractional length of the genome from the *Eco*RI cleavage site. (*d*) Portions of the SV40 segment shown above whose early- and/or late-strand transcripts are detectable in cytoplasmic RNA that has been retained by oligo(dT)-cellulose. (*e*) Location of repeating sequences with the segment of SV40 DNA represented in (*c*) above. A roman numeral indicates a particular sequence that occurs more than once. The numbers following the roman numerals indicate the residues in Fig. 1 that constitute that sequence. Repeated segments V and VI are identical tandem repeats (ref 9; K. N. Subramanian and S. M. Weissman, *Cell*, in press). The remaining repeated segments differ by one or more nucleotides from one another and/or are not precisely tandem.

vitro with *E. coli* RNA polymerase. Throughout the region 0.70–0.84 map unit[§], only products from the transcript of the late strand were detected. The results for the region from nucleotides 228 to 537 have been presented in detail elsewhere (9, 10). When the cytoplasmic RNA was annealed to fragments overlapping the region from 0.76 to 0.84 on the genome, more SV40 specific RNA was detected than had been found from the preceding region. The products from the region 0.70 to 0.77 were prominent. Contrary to expectations, these oligonucleotides were more abundant than those from the region beyond nucleotide 916. From nucleotides 886 to 915, four consecutive T1 RNase oligonucleotides were missing from the maps of cytoplasmic RNA (U-A-A-A-G, U-U-U-A-G, U-C-U-U-U-U-U-G, and U-C-U-U-U-U-A-U-U-U-C-A-G). Two of these were very rich in uridylic acid and might have been selectively lost, but the other two, U-A-A-G and U-U-U-A-G, should have been readily apparent. Also, the oligonucleotides immediately preceding these products were darker than those following. Unless there was some unanticipated artifact, such as annealing of adenylic acid-rich sequences to the uridylic acid stretches or selective retention of fragments of RNA transcribed from the DNA immediately preceding 0.77 map unit on the oligo(dT)-cellulose column because of the internal stretches of 5, 6, and 7 adenylic acids, the results indicate that there are two separate late RNA species. One species would start from a location preceding 0.66 and extend to about 0.77, and the second would begin just after 0.775 and extend at least to 0.84, and perhaps through the rest of the late region of the genome.

[§] Map units refer to the fractional genome length from the *Eco*RI cleavage site on SV40 DNA in the direction towards the *Bam* 1 cleavage site.

DISCUSSION

Location of Genes for Proteins Encoded in SV40 DNA. The major structural protein of SV40 is termed VP1. The molecular weight of this protein is estimated to be 42,000–44,000. The NH_2-terminal sequence of VP1 is known and corresponds to a series of codons in mRNA transcribed from DNA at 0.945 map unit on the genome (11). The 3′ end of abundant late mRNA, which codes for VP1, is located precisely at 0.175 map unit on the genome (12–14), and there are termination codons in all three phases about 100 nucleotides before the 3′ end of late mRNA. Therefore, the mRNA coding for VP1 contains a total of approximately 1100 nucleotides. This is slightly less than that anticipated from the estimated molecular weight of VP1.

The second structural protein of SV40 is termed VP2, and is also coded for by the late region of SV40. The estimated molecular weight of VP2 is well over 35,000. It would require over 1050 base pairs to code for VP2. Therefore a substantial part of SV40 DNA codes for both VP1 and VP2, and the coding region for VP2 must begin within the late region of SV40 DNA, no farther from the origin of DNA replication than about 0.77 map unit. The base triplet ATG at position 920 in the DNA is not followed by any in-phase termination codons in the present sequence. It is therefore a potential candidate for the initiation codon for VP2.

The DNA sequence from 850 to 890 in Fig. 1 predicts termination codons in all three phases of late-strand transcripts, so that it is unlikely that the initiator codon for VP2 lies closer to the origin of DNA replication than position 920.

The protein coded for by the early region of SV40 DNA has an estimated molecular weight in excess of 80,000. It requires

approximately 2500 base pairs of DNA to code for a protein of this size. The sequence of the DNA complementary 5′-terminal region of early mRNA indicates that there are termination codons in all phases of this portion of the RNA. The 3′ terminus of the early RNA is located at approximately 0.16 map unit on the SV40 genome. The total region available for coding for T antigen would therefore be less than 2700 base pairs. An additional difficulty in assigning the coding region of the T antigen is created by recent reports that deletions within the region of 0.54 to 0.59 map unit on SV40 DNA produce virus that is viable and produces a T antigen whose electrophoretic mobility on sodium dodecyl sulfate gels is indistinguishable from that of wild-type T antigen (15).

Part of the Late Region of SV40 DNA Does Not Code for Known Proteins. Between positions 547 and 557 of the sequence shown in Fig. 1 there are termination codons in all three phases of early-strand transcripts. Again, between bases 859 and 881, the DNA sequences show that there would be termination codons at all three phases in late message RNA. Unless one of the codons UAA, UGA, or UAG does not function in cells as a termination of signal, the region of SV40 DNA from positions 557 to 860 could not code for a protein as large as any of the known structural proteins. The transcript of the late strand of SV40 DNA from this region is abundantly represented in the cytoplasmic fraction and is retained on oligo(dT)-cellulose. Although we cannot rigorously exclude leakage of the RNA from nuclei during cell fractionation, this would have to be selective because late-strand nucleotides from the early region of SV40 were not detected in cytoplasmic RNA.

On the other hand, examination of the cytoplasmic RNA complementary to this region of SV40 DNA does not show a detectable amount of oligonucleotides from the sequence between bases 886 and 916. Beyond position 916 the expected oligonucleotides are present. This suggests that the cytoplasmic late-strand RNA from DNA preceding nucleotide 885 may not be covalently linked to the cytoplasmic RNA complementary to nucleotides 916 and onwards. The late-strand transcript contains an AUG corresponding to positions 694 to 696 of the sequence of Fig. 1. This AUG is followed by a series of 61 sense triplets in phase, followed by the triplet UAA. SV40 mutants with deletions of this region grow, albeit slowly, in the absence of helper virus (15). Therefore, if this region did code for a peptide, it probably would not be a structural protein but could function catalytically in virus assembly, in late message or protein production, or by affecting host cell gene expression. There are also sequences of continuous triplets within the 5′ end of early mRNA that are preceded by an AUG and followed by a terminator codon. We have not investigated whether this early RNA exists in cytoplasm as a separate molecular species or as part of the 19S early mRNA.

Certain prokaryotic transcripts, such as those for the late genes of lambda phage and for the tryptophan operon of *E. coli*, have relatively long sequences at their 5′ ends preceding the portions of the messages that are translated to produce known proteins. In these cases, transcription *in vivo* and *in vitro* may terminate in uridylic acid-rich sequences that are located between the 5′ portions of the message and the portions of the sequence known to code for proteins. This termination signal early in the message RNA may be overcome by specific gene products *in vivo* to turn on expression of the downstream genes (16–18). Large parts of low-molecular-weight RNA extending from transcription start to the transcription termination signal may be found *in vivo* or produced by transcription with purified RNA polymerase. These prokaryotic leader sequences are a possible model for the role of the sequences in early and late RNA transcripts of SV40 that do not code for a known SV40 protein. Cells infected with adenovirus also synthesize large amounts of low-molecular-weight RNA (19).

Repeated Sequences in SV40 DNA. The dodecanucleotide A-C-A-A-A-U-A-A-A-G-C-A occurs twice in the early-strand transcript of SV40 DNA near the 3′ end of early RNA (3, 14). The two occurrences of the dodecanucleotide are separated by about 30 bases. Although the function of this repeated sequence is not known, it does contain the heptanucleotide A-A-U-A-A-A, which has been found in the 3′ end of various animal cell mRNAs (20, 21).

Large tandem repeating sequences occurring near the origin of DNA replication between 0.68 and 0.71 map unit have been found (K. N. Subramanian and S. M. Weissman, *Cell*, in press). One of these is a tandem repeat of 21 nucleotides (9); another is a tandem repeat of 55 nucleotides (K. N. Subramanian and S. M. Weissman, unpublished data). The hexanucleotide G-G-G-C-G-G repeats six times within the sequence located between nucleotides 419 and 481 (9).

Another sequence located between nucleotides 844 and 884 (Fig. 1) occurs as a nearly perfect tandem repeat (with 19 of 20 nucleotides being identical). This repeat occurs at a location immediately preceding the DNA encoding the amino terminal sequences of VP2.

These reiterated sequences are shown within boxes in Fig. 1 and are also indicated in Fig. 2.

Viable deletion mutants of SV40 have been generated in which single copies of either the repeating segment containing 9 nucleotides or that containing 55 nucleotides near 0.70 map unit have been moved (ref. 15; K. N. Subramanian, T. Shenk, and S. M. Weissman, unpublished data). There has been no success in obtaining viable virus with deletions of the repeating sequence at 0.76 map unit. Therefore, the former sequence may play at least a quantitatively different role from the latter during SV40 infection.

The short, multiple, imperfectly repeated sequences and the large tandem repeated sequences are reminiscent of the highly and moderately reiterated sequences in most eukaryotic cell DNA (22), and knowledge of the function of the viral sequence could give insight into the role of the noncoding reiterated sequence of the host cell.

Sequences Preceding the 5′ Ends of Cytoplasmic SV40 RNA. The DNA preceding the sequences complementary to the 5′ end of late cytoplasmic RNA, roughly residues 138 to 210, contains a sequence of 14 consecutive AT pairs. The DNA preceding the 5′ end of early RNA contains a sequence of 17 consecutive AT pair (9) and the DNA preceding the 5′ end of the "restart" of late RNA near residue 910 contains a sequence with 19 of 23 AT base pairs (Fig. 2). It is not known whether the 5′ ends of cytoplasmic RNA represent the 5′ ends of initial transcripts or are generated from larger transcripts by endo- or exonucleolytic cleavage. There are several examples of prokaryotic RNAs where a larger precursor is cleaved to produce the more stable RNA species. Extensive base pairing has been noted in T7 mRNA near the sites of RNase III processing, and near the sites where ribosomal 5S RNA precursor of *Bacillus subtilis* is cleaved to produce mature 5S RNA. The secondary structure in tRNA precursors may also be important for their processing.

The sequence near the 5′ end of SV40 early RNA would permit extensive and accurate base pairing of the RNA transcript. There are adenylic acid-rich sequences at positions 843 to 870 in the SV40 sequence that could pair with the transcript of the thymidylic acid-rich sequence near the "restart" of late mRNA. Nevertheless, the analogies to the sequences at sites of

processing of bacterial RNA precursors are weak. The AT-rich sequences of SV40 could represent a somewhat different type of recognition site for RNA processing, but could also be part of promoters for initiation of transcription.

We are grateful to Dr. Richard Roberts of Cold Spring Harbor Laboratory who furnished us with samples of numerous restriction endonucleases and information about new enzymes prior to publication, and to Dr. Frank Young and his colleagues at the University of Rochester for a gift of the *Bacillus* species enzyme we term *Bam* X. We would also like to thank Christine Gerhardt, Sharlene Ivory, Adeline Tucker, Cheryl Christner, and Anne Goss for their excellent technical assistance. This research was supported by grant 5-PO1-CA-16038 from the National Cancer Institute and by grant VC-1H from the American Cancer Society.

1. Crawford, C. V. (1974) *Br. Med. Bull.* **30,** 253–258.
2. Marotta, C. A., Lebowitz, P., Dhar, R., Zain, B. S. & Weissman, S. M. (1974) in *Methods in Enzymology,* eds. Grossman, L. & Moldave, K. (Academic Press, New York and London), Vol. 29, Part E, pp. 254–272.
3. Zain, B. S., Dhar, R., Weissman, S. M., Lebowitz, P. & Lewis, A. M., Jr. (1973) *J. Virol.* **11,** 682–693.
4. Brownlee, G. G. (1972) *Determination of Sequences in RNA* (North Holland/American Elsevier).
5. Maniatis, T., Jeffrey, A. & Kleid, D. (1975) *Proc. Natl. Acad. Sci. USA* **72,** 1184–1188.
6. Venetianer, P. & Leder, P. (1974) *Proc. Natl. Acad. Sci. USA* **71,** 3892–3895.
7. Subramanian, K. N., Pan, J., Zain, B. S. & Weissman, S. M. (1974) *Nucleic Acids Res.* **1,** 727–752.
8. Yang, R. C. A., Van de Voorde, A. & Fiers, W. (1976) *Eur. J. Biochem.* **61,** 119–138.
9. Subramanian, K. N., Dhar, R. & Weissman, S. M. (1976) *J. Biol. Chem.,* in press.
10. Dhar, R., Subramanian, K. N., Pan, J. & Weissman, S. M. (1976) *J. Biol. Chem.,* in press.
11. Van de Voorde, A., Contreras, R., Rogiers, R. & Fiers, W. (1976) *Cell,* **9,** 117–120.
12. Dhar, R., Zain, B. S., Weissman, S. M., Pan, J. & Subramanian, K. N. (1974) *Proc. Natl. Acad. Sci. USA* **71,** 371–375.
13. Dhar, R., Weissman, S. M., Zain, B. S., Pan, J. & Lewis, A. M., Jr. (1974) *Nucleic Acids Res.* **1,** 595–613.
14. Dhar, R., Subramanian, K. N., Zain, B. S., Pan, J. & Weissman, S. M. (1974) *Cold Spring Harbor Symp. Quant. Biol.* **39,** 153–160.
15. Shenk, T. E., Carbon, J. & Berg, P. (1976) *J. Virol.* **18,** 664–671.
16. Sklar, J., Yot, P. & Weissman, S. M. (1975) *Proc. Natl. Acad. Sci. USA* **72,** 1817–1821.
17. Sklar, J. & Weissman, S. M. (1976) *Fed. Proc.* **35,** 1537.
18. Bertrand, K., Squires, C. & Yanofsky, C. (1976) *J. Mol. Biol.* **103,** 319–337.
19. Reich, P., Rose, J., Forget, B. & Weissman, S. M. (1966) *J. Mol. Biol.* **17,** 428–439.
20. Proudfoot, N. & Brownlee, G. G. (1974) *Nature* **252,** 359–362.
21. Milstein, C., Brownlee, G. G., Cartwright, E. M., Jarvis, M. & Proudfoot, N. J. (1974) *Nature* **252,** 354–367.
22. Lewin, B. (1975) *Cell* **4,** 77–93.