JOURNAL OF VIROLOGY, Apr. 1979, p. 279-296
0022-538X/79/04-0279/18$02.00/0

Vol. 30, No. 1

Simian Virus 40 Early mRNA's

I. Genomic Localization of 3′ and 5′ Termini and Two Major Splices in mRNA from Transformed and Lytically Infected Cells

V. B. REDDY,[1] P. K. GHOSH,[2] P. LEBOWITZ,[2]* M. PIATAK,[1] AND S. M. WEISSMAN[1, 2]

Departments of Human Genetics[1] and Internal Medicine,[2] Yale University School of Medicine, New Haven, Connecticut 06510

Received for publication 6 October 1978

We have studied the structure of polyadenylated virus-specific cytoplasmic mRNA's in mouse and human cells transformed by simian virus 40 and in monkey cells infected with simian virus 40 in the presence of cytosine arabinoside by means of reverse transcriptase-catalyzed complementary DNA synthesis and complementary DNA sequencing. Abundant mRNA species containing splices from residues 4490 to 4557 (0.533 to 0.546 map units [m.u.]) and 4490 to 4837 (0.533 to 0.600 m.u.) were identified in both transformed and infected cells. Two principal reverse transcriptase stops were observed at the 5′ termini of these mRNA's, both occurring with approximately equal frequency. The most distal of these stops was localized at residues 5152 to 5154 (0.660 m.u.), and the second was at residues 5147 to 5148 (0.659 m.u.). Several additional minor stops, between approximately 0.62 and 0.65 m.u., were also found on complementary DNA copied from transformed cell mRNA; in contrast, only one additional stop was present on complementary DNA copied from early lytic mRNA. These data suggest the presence of a principal 5′ terminus of early lytic and transformed cell mRNA's at residues 5152 to 5154 and raise the possibility of additional 5′ termini at one or more locations in the 0.62 to 0.659 m.u. region of these mRNA's. Transformed cell mRNA was also found to contain a single 3′ terminus at positions 2504 and 2505 (0.153 m.u.); termini lying beyond this site were not detected.

The early gene region of simian virus 40 (SV40) consists of approximately 2,650 nucleotides lying on the minus strand of the circular double-stranded SV40 genome. It is transcribed in a counterclockwise or leftward direction from 0.660 to 0.153 map units (m.u.) throughout both lytic infection and in cells transformed by SV40 (16–18, 25–29, 48, 49).

Expression of the early region of the SV40 genome is required for initiation of new rounds of viral DNA synthesis during lytic infection (13, 54), for initiation of transformation (1, 2, 33, 55), and, in certain transformed lines, for maintenance of transformation as well (12, 32, 33, 36, 43, 55). Until recently, it was thought that early gene expression was also required for initiation of late gene transcription. However, a recent report that transcriptional complexes from cells infected with mutants of SV40 blocked in early functions synthesize late RNA (21) suggests that late gene transcription is not totally dependent upon prior early gene expression.

Because of the importance of the viral early gene region for induction and maintenance of cell transformation, the products of the early gene region are currently under intensive investigation. By means of immunoprecipitation with antisera from animals bearing SV40-induced tumors, two major early proteins have been consistently demonstrated in cells lytically infected or transformed by SV40 (41, 51, 53). These two proteins have also been synthesized in vitro in translational systems programmed with early lytic or transformed cell mRNA's (37, 41). The principal protein has been designated large T antigen or A protein because it has a molecular weight in the range of 90,000 to 95,000 and is heat labile in cells infected with temperature-sensitive mutants in SV40 early functions (*tsA* mutants). The second protein with a molecular weight of approximately 17,000 to 20,000 has been designated small t antigen.

Three important observations made in the past 2 years have suggested great complexity in the organization of the genes for these two early proteins. First, sequences specifying translation termination codons are present in all three reading phases on the viral minus strand at 0.535 to 0.55 m.u., only 25% of the distance into the early region (22, 44). Second, large and small T antigens share amino-terminal amino acids (38, 51), starting with a methionine residue copied from an AUG codon at 0.647 m.u. (38), but small t contains additional tryptic peptides not present in large T and large T contains additional peptides not present in small t antigen (38, 41, 51). Third, viral mutants containing deletions in the 0.54 to 0.59 m.u. region specify normal large T antigens but produce either altered or no small t antigens (15, 52; A. E. Smith, personal communication), whereas mutants containing deletions at 0.21 and from 0.325 to 0.43 m.u. synthesize normal small t but altered large T antigens (15, 47).

On the basis of these and other findings, it has been proposed (15) that small t antigen is encoded in a continuous segment of viral DNA extending from approximately 0.65 to 0.55 m.u. and that the template for large T consists of two separate segments of DNA, the 3′ segment containing sequences from 0.65 to 0.59 m.u. and the 5′ segment containing sequences from 0.54 to 0.17 m.u. In support of this model, Berk and Sharp (5) have demonstrated two mRNA's early in SV40 lytic infection by means of S1 nuclease mapping techniques, one with a splice which removes a limited number of nucleotides in the 0.54 m.u. region and the second with a splice which removes approximately 300 nucleotides and fuses sequences from approximately 0.54 and 0.60 m.u. It has been suggested that these mRNA's code for small and large T antigens, respectively.

In the present report, we have analyzed the principal viral mRNA's present in SV40-transformed cells and early lytic infection by means of primer-directed, reverse transcriptase-catalyzed complementary DNA (cDNA) synthesis followed by cDNA sequencing. In both systems, mRNA's presumed to code for small and large T antigens were found to contain splices between residues 4490 to 4557 (0.533 to 0.546 m.u.) and 4490 to 4837 (0.533 to 0.600 m.u.), respectively. (Nucleotide residues are numbered according to the system of Reddy et al. [44]; numbers decrease in a 5′ → 3′ direction on early mRNA's.) cDNA's synthesized on both transformed cell and early lytic mRNA's demonstrated two equally strong reverse transcriptase stops at residues 5152 to 5154 (0.660 m.u.) and 5147 to 5148 (0.659 m.u.), suggesting the presence of a major mRNA 5′ terminus at the former site and possibly a second 5′ terminus at the latter site. Several additional reverse transcriptase stops in the 0.62 to 0.65 m.u. region were observed in cDNA's copied from transformed cell mRNA's, and one additional stop was noted in this region in cDNA copied from early lytic mRNA's, raising the possibility of additional minor 5′ termini in these RNA's. Transformed cell mRNA's were also found to contain a single 3′ terminus at residues 2504 to 2505 (0.153 m.u.). A portion of these results have been noted in a prior report (44).

MATERIALS AND METHODS

The Vero line of African green monkey kidney cells, the SV40-transformed mouse fibroblast lines SVT2, SV101, and SV215 (kindly provided by G. Todaro, R. Pollack, and P. Mora), and the transformed human lung fibroblast SV80 (the generous gift of D. Livingston) were used in the present studies. These cells were grown in roller bottles with a 750-cm^2 surface area in Eagle minimal essential medium containing 2 mM glutamine, 250 U of penicillin/ml, 250 μg of streptomycin/ml, and 10% fetal calf serum. Upon reaching confluence, Vero cells were washed with normal saline and infected with strain 776 SV40 at a multiplicity of infection of 20 PFU/cell in the same medium as above except for replacement of fetal calf serum with 2% agammaglobulinemic calf serum. Five hours postinfection, cytosine arabinoside was added to cultures to a final concentration of 20 μg/ml, and 30 h postinfection cells were harvested by treatment with trypsin-EDTA and washed with normal saline. At confluence, transformed cells were harvested similarly. RNA was then extracted from the cytoplasm of infected and transformed cells by a modification of the method of Penman (40) and was passed through columns of oligodeoxy-thymidylic acid [oligo(dT)]-cellulose as described elsewhere (P. K. Ghosh et al., J. Mol. Biol., in press). SV40 DNA was prepared as previously described (63).

Restriction enzymes were obtained from New England Biolabs or prepared by published procedures; DNA digestions wêre carried out under the conditions specified by this supplier or described in the literature. The *Hinf*-G fragment was prepared by digestion of SV40 DNA with *Hinf* restriction endonuclease followed by electrophoresis of the resultant fragments on a 4% polyacrylamide gel in 0.04 M Tris, 0.02 M sodium acetate, and 0.002 M EDTA, pH 7.8. This fragment extends from residues 4378 to 4486 (0.512 to 0.533 m.u.) on the SV40 minus DNA strand and was used as a primer for sequence analysis across the splices in the early viral mRNA's. Three fragments were used for analysis of sequences at the 5′ termini of the early mRNA's. The fragment extending from residues 5054 to 5089 (0.641 to 0.648 m.u.) on the viral minus strand was obtained by first digesting SV40 DNA with the *Hinf* enzyme, subjecting the digest to electrophoresis on a 4% polyacrylamide gel, and isolating the *Hinf*-A fragment from this gel. This fragment was then digested with *Hin*dIII restriction enzyme, and the smallest of the resultant fragments, extending from residues 5054 to 5089, was isolated from a similar gel. The fragment extending from residues 4629 to 4689 (0.560 to 0.572 m.u.) was obtained by redigesting the *Hinf*-D fragment with *Mbo*I restriction enzyme and isolating the smallest of the resultant three fragments from a polyacrylamide gel. The third fragment, spanning the sequence from residues 4835 to 4883 (0.600 to 0.609 m.u.), was prepared by first digesting SV40 DNA with the *Eco*RII restriction enzyme, isolating the I fragment, with termini at map positions 4810 and 5009, on a 4% gel, redigesting this fragment with the *Mbo*II restriction enzyme, and isolating the second largest of the resultant fragments, extending from residues 4835 to 4883, on a 6% gel. For determination of viral sequences at the 3′ termini of the early mRNA's, the fragment extending from residues 2571 to 2586 (0.166 to 0.169 m.u.) was prepared. This was accomplished by first digesting SV40 DNA with the *Hin*dII and *Hin*dIII restriction enzymes, isolating the G fragment on a 4% polyacrylamide gel, subjecting this fragment to digestion with the *Alu* restriction enzyme, and isolating the smallest of the resultant fragments on a 10% polyacrylamide gel. The localizations of these five fragments within the SV40 early gene region are shown on Fig. 1. Before being used as primers, these fragments were labeled at their 5′ termini with [γ-^{32}P]ATP (New England Nuclear) by use of the enzyme T4 polynucleotide kinase (46).

Hybridization of radiolabeled primers to cellular RNA, synthesis of cDNA's on these primers, fractionation of cDNA's, and determination of the nucleic acid sequences of individual cDNA's were carried out by methods described in detail previously (23; Ghosh et al., in press). Human placental RNase inhibitor prepared by the method of Blackburn et al. (6) was included in certain cDNA syntheses. Certain cDNA's were cleaved with *Hae*III restriction endonuclease prior to gel electrophoretic fractionation. This enzyme is known to cleave single- and double-stranded DNAs at identical sites (7, 24); however, since the efficiency of cleavage of single-stranded DNA is only 5 to 10% that of double-stranded DNA (7), digestions were carried out overnight at 37°C with 20-fold the concentration of enzyme required for complete cleavage of double-stranded DNA in 3 h.

Since extension of 5′ terminally labeled primers takes place in a 5′ → 3′ direction with respect to DNA but with a 3′ → 5′ polarity with respect to template RNA, sequence analysis of the resultant cDNA's provides sequences of internal and 5′-terminal regions of template RNAs (see Fig. 1). However, cDNA's obtained by this method cannot be used for sequencing the 3′ termini of RNAs, and the following two-step extension procedure was carried out to localize the 3′ terminus of transformed cell early mRNA. In the first step, a cDNA copy was made of polyadenylated cytoplasmic RNA from the SV80-transformed line by use of oligo(dT) as a primer and the enzyme reverse transcriptase. The reaction for this synthesis contained, in a volume of 400 μl, 50 mM Tris-hydrochloride, pH 8.3, 6 mM magnesium acetate, 60 mM NaCl, 10 mM dithiothreitol, 1 mM each of dCTP, dATP, and dGTP, and dTTP (the dTTP containing 10 μCi of ^{3}H label [New England Nuclear]), 500 μg of RNA, 50 μg of oligo(dT), and 10 U of reverse transcriptase (kindly provided by J. Beard, National Cancer Institute Viral Research Program). Actinomycin D was also included in the reaction in a final concentration of 1 μg/ml to prevent the synthesis of double-stranded DNA. Incubation was carried out at 41°C for 3 h, after which RNA templates were degraded by addition of NaOH to a concentration of 0.2 N and incubation at 41°C was continued for 1 h. The reaction mixture was then brought to 0.25% in sodium dodecyl sulfate, extracted twice with water-saturated phenol, and passed through a column of Sephadex G100 in 0.01 M Tris-hydrochloride, pH 7.5, 0.01 M $MgCl_2$, and 1 mM EDTA in order to separate cDNA from degraded RNA fragments and any residual oligo(dT).

In the second phase of this procedure, the SV40 DNA fragment extending from residues 2571 to 2586 (0.166 to 0.169 m.u.) labeled in the 5′-terminal positions with ^{32}P was annealed to the viral cDNA and extended in a 5′ → 3′ direction to yield extended products with sequences identical to those at the 3′ termini of early mRNA's. First, approximately 0.5 to 1 μg of this fragment was taken up in 50 μl of 100% formamide and denatured at 95 to 100°C for 5 min. After addition of approximately 100 μg of cDNA in an equal volume of 4 × SSC (SSC is standard saline citrate: 0.15 M NaCl, 0.015 M sodium citrate, pH 7.0), the mixture was taken up into a capillary tube, kept at 100°C for 10 min, and then incubated at 50°C for 24 h. At the conclusion of this reaction, the mixture was diluted with 10 volumes of water, and nucleic acids were precipitated by addition of 0.1 volume of 3 M sodium acetate and 2 volumes of ethanol. Pelleted nucleic acids were then subjected to reverse transcription under the conditions noted above except for the deletion of oligo(dT) and actinomycin D. At the conclusion of the reaction, phenol extraction was carried out, and nucleic acids were precipitated with ethanol. The pellet, containing double-stranded DNA-DNA hybrids, was then taken up in 20 μl of 0.05 N NaOH and 5 M urea, incubated at 95°C for 5 min to denature the DNA, and subjected to electrophoresis on an 8% polyacrylamide-7 M urea slab gel in the Tris-EDTA-borate system of Peacock and Dingman (39). The procedures for extraction of individual fragments from this gel and the determination of their nucleotide sequences have been described (Ghosh et al., in press; V. B. Reddy et al., Nucleic Acids Res., in press).

RESULTS

Localization of splices. We were guided in our choice of a priming fragment with which to probe for splices in transformed cell and early lytic mRNA's by the presence of termination codons in all three translational reading frames in early mRNA at 0.535 to 0.55 m.u. (22, 44) and the demonstration that deletion mutants of SV40 in the 0.54 to 0.59 m.u. region directed synthesis of wild-type T antigen (15, 52). On these bases, we chose to use the *Hinf*-G fragment, extending from residues 4378 to 4486 (0.512 to 0.533 m.u.), as a primer. Furthermore, to reduce the size of the cDNA products to an analyzable level (less than 300 nucleotides) and to reduce the possible complication of extended products with artifactual 3′ termini (due either to RNA degradation during extraction or to reverse transcription or premature termination of transcription), we chose to digest the single-stranded cDNA's with the *Hae*III restriction enzyme as described in Materials and Methods before fractionating them on 8% polyacrylamide-7 M urea gels. This enzyme cleaves the early region of SV40 DNA at two sites downstream from the *Hinf*-G primer, at residues 4781 and 5110 (0.589 and 0.652 m.u.). Figure 1 indicates the site on the SV40 genome to which *Hinf*-G binds, the direction of reverse transcription, and the locations where *Hae*III cuts the viral DNA.

The electrophoretic pattern of the *Hae*III-treated cDNA's synthesized on polyadenylated cytoplasmic RNA from the SV80 transformed line is shown in Fig. 2. A number of bands were visualized, four of which (no. 3, 4, 6, and 7) contained over 90% of the total radioactivity. Since this pattern was more complex than expected, we proceeded to perform sequence analyses on these four bands as well as the shorter bands 8, 9, and 10 (Fig. 3). Unfortunately, the amounts of radioactivity in bands 1 and 2 were not great enough to permit analyses, and band 5 could not be separated adequately from band 6 and thus could not be analyzed.

Bands 3 and 7 exhibited identical patterns on Maxam-Gilbert DNA sequencing gels over a distance of approximately 250 nucleotides at their 5′ ends. The patterns of bands 4, 6, 8, 9, and 10 were also identical to one another over a similar span, but different from those of bands 3 and 7. Figure 4 presents the sequencing gels for bands 4 and 7. cDNA sequences are read in a 5′ → 3′ direction, starting from the bottom of the gels and proceeding upward. Since reverse transcriptase faithfully copies RNA templates, the complements of the 5′ → 3′ sequences of the respective cDNA's provide the sequences of their RNA templates with a 3′ → 5′ polarity. cDNA's 4 and 7 both started with the sequence ...TTCCATAGGTTGGAATCT, corresponding perfectly with the sequence of the SV40 minus DNA strand from residues 4460 to 4465 through

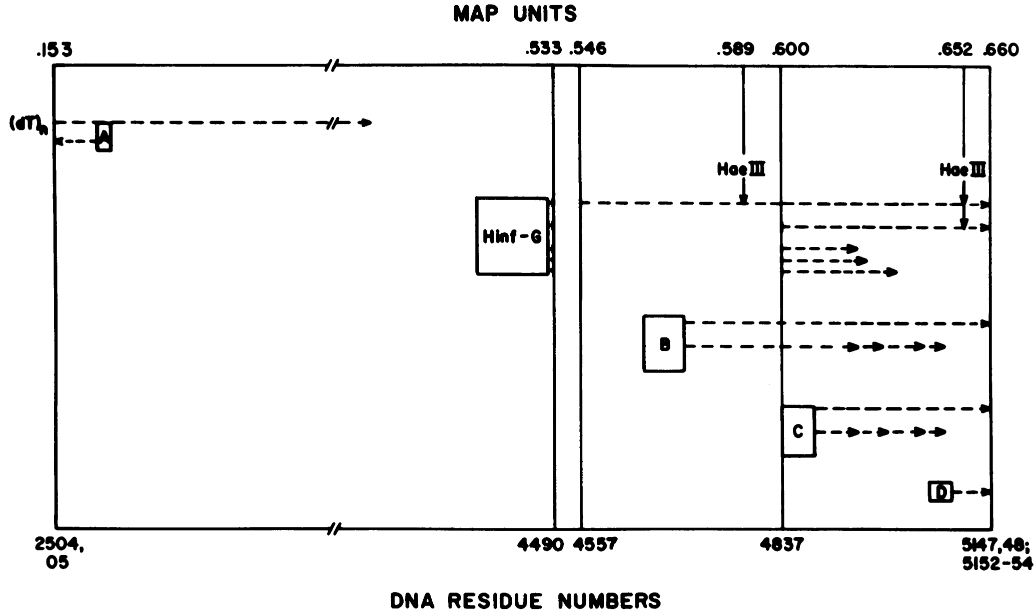


FIG. 1. *Map of the SV40 early gene region showing the localizations of specific restriction endonuclease fragments (in boxes) which were bound to early SV40 mRNA's and used as primers for reverse transcriptase-catalyzed cDNA syntheses and the major cDNA's produced (dashed lines with arrows). On this map, early SV40 mRNA's have a right to left 5′ → 3′ direction; cDNA's are reverse-transcribed on early mRNA's in a left to right direction. Restriction fragments A, B, C, D, and Hinf-G span, respectively, residues 2571 to 2586 (0.166 to 0.169 m.u.), residues 4629 to 4689 (0.650 to 0.572 m.u.), 4835 to 4883 (0.600 to 0.609 m.u.), 5054 to 5089 (0.641 to 0.648 m.u.) and 4378 to 4486 (0.512 to 0.533 m.u.); the methods by which they were prepared are given in Materials and Methods. The dashed lines extending from fragments C and D containing multiple arrow heads represent a number of cDNA's whose precise 3′ termini have not yet been determined. Digestion with HaeIII restriction enzyme at the indicated sites was performed on cDNA's synthesized on the Hinf-G primer prior to their electrophoresis and analysis. Note that determination of 3′-terminal mRNA sequences required a cDNA synthesis using oligo(dT) as a primer and a second reverse transcription in which oligo(dT)-terminal cDNA served as template and the A fragment as primer. Map units are fractional genomic lengths from the unique EcoRI cleavage site, and DNA residue numbers are taken from Reddy et al. (44).*

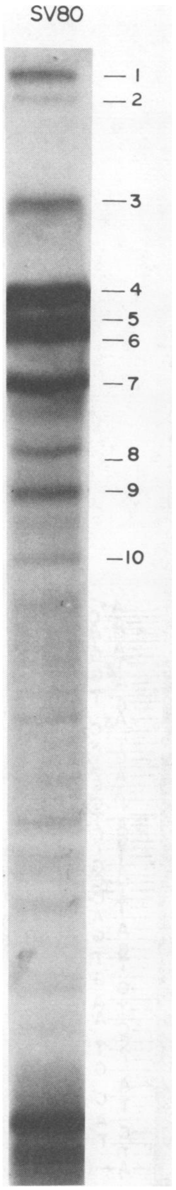


FIG. 2. *Autoradiogram of the 8% polyacrylamide-7 M urea gel electrophoresis of the 5' ^{32}P-labeled products obtained by reverse transcriptase-catalyzed extension of the Hinf-G primer (0.512 to 0.533 m.u.) bound to polyadenylated cytoplasmic RNA isolated from the SV80-transformed line followed by HaeIII digestion of the resultant cDNA. The methods used are presented in Materials and Methods and prior reports (23; Ghosh et al., in press).*

residue 4491 (0.534 m.u.). Beyond this point, however, neither band showed the contiguous sequence. Band 7 continued TATAGCTTT..., corresponding to the sequence from residue 4558 onwards, and band 4 continued CAGTTGCAT..., corresponding to the sequence from residue 4838 onwards. Beyond these points, the sequences of all the bands continued colinearly with SV40 DNA until their gel patterns became unclear (bands 3-8) or until they reached 3' termini (at approximately residues 4865 and 4850 for bands 9 and 10, respectively).

The sequencing gels thus demonstrate the presence of two different splices in early RNAs isolated from the SV80 transformed line. The first splice removes 66 nucleotides from the RNA(s) serving as template(s) for cDNA's 3 and 7, and the second splice takes out a total of 346 nucleotides from the RNA(s) from which cDNA's 4, 6, and 8-10 are copied. However, ambiguities arise in assigning the precise nucleotides at the 5' and 3' termini of RNA segments contributing to the splices in these RNAs. Similar ambiguities have been described for five splices in the late RNAs of SV40 (Ghosh et al., in press; Reddy et al., in press) and for splices in a mouse immunoglobulin light-chain mRNA (59) and ovalbumin mRNA (10). All come about as a result of short identical sequences which are present at both sites involved in each splice and the fact that splices are constructed so that only single copies of the duplicated sequences are retained in the final RNAs. For the RNAs giving rise to bands 3 and 7 and bands 4, 6, 8, 9, and 10, the duplicated sequence in a 5' → 3' direction is the dipurine AG at residues 4491-4490, 4557-4556, and 4837-4836 (for the late SV40 RNAs, duplicated sequences have been AGGU, GGU, GG, and AG). To facilitate discussion, we have adopted the convention of constructing splices between the central dipurine nucleotides of these di-, tri-, and tetranucleotides. Thus, we

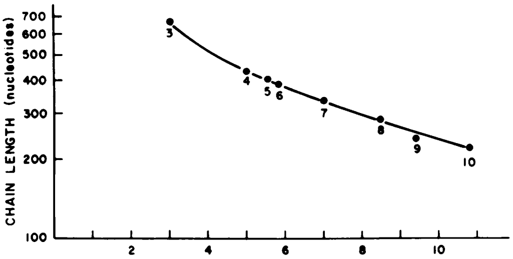


FIG. 3. *Semilogarithmic plot relating the migration of cDNA's in Fig. 2 to their chain lengths. The lengths of cDNA's 9 and 10 were determined by direct nucleotide sequence analyses. Estimates of the lengths of cDNA's 3, 4, 6, and 7 were based in part on sequence analyses and in part on knowledge of the sites at which HaeIII restriction endonuclease cleaves the SV40 genome (see text for further explanation). The chain lengths of cDNA's 5 and 8 were derived from their migrations in Fig. 2 and this plot.*

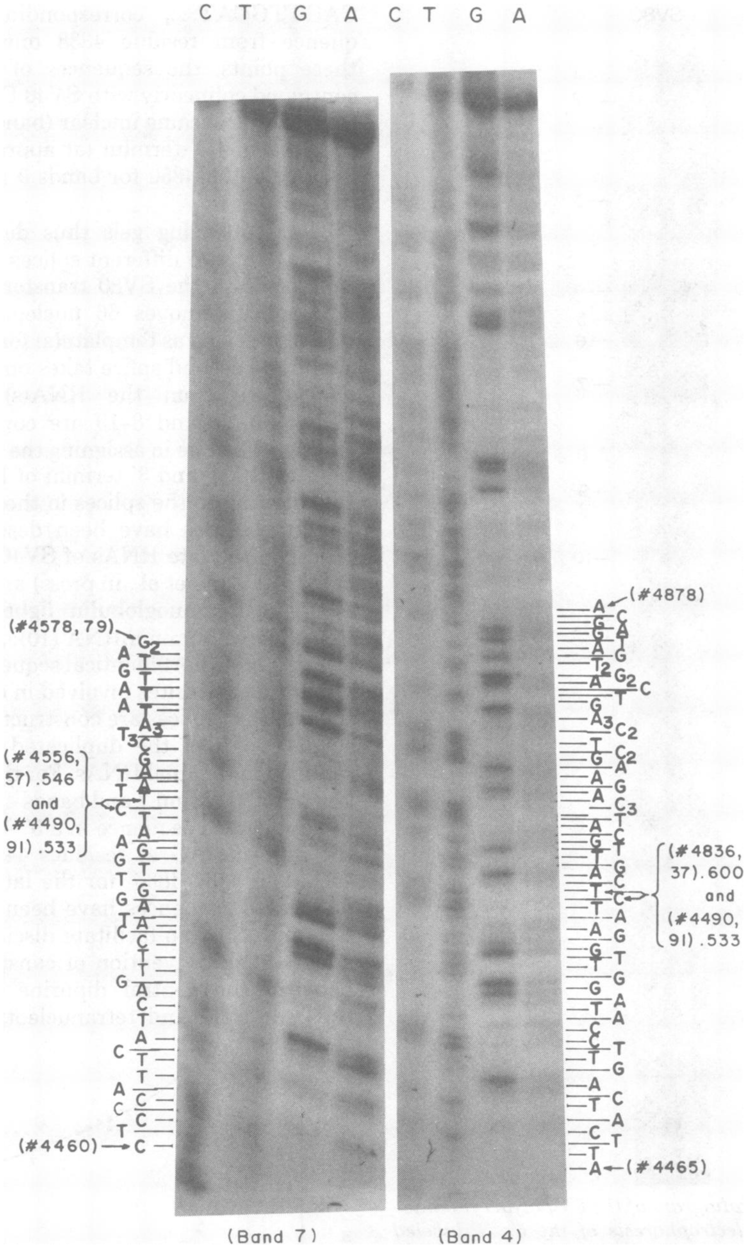

C T G A C T G A
(#4578,79)
G2
T3
A3
(#4556, 57).546 and (#4490, 91).533
(#4460)
(#4878)
(#4836, 37).600 and (#4490, 91).533
(#4465)
(Band 7)
(Band 4)

FIG. 4. *Electrophoretic fractionations obtained when cDNA's 4 and 7 from Fig. 2 were subjected to DNA sequence analysis by the procedure of Maxam and Gilbert (34). C, T, A, and G over gel channels refer to products generated by means which preferentially cleave at cytidylic acid residues, at cytidylic and thymidylic acid residues, at guanylic and to a lesser extent adenylic acid residues, and at adenylic and guanylic acid residues. DNA sequences are shown in the margins of the autoradiograms and are read in an upward (5' → 3') direction. Decimal numbers represent map units on the SV40 genome, and whole numbers represent the positions of specific DNA residues (44).*

have designated residues 4490 and 4557 as the spliced nucleotides in the RNA(s) giving rise to cDNA's 3 and 7 and residues 4490 and 4837 as the spliced nucleotides in the RNA template(s) for cDNA's 4, 6, 8, 9, and 10. We have also found the identical two splices in polyadenylated cytoplasmic RNA isolated from the transformed lines SVT2 and SV101 and in early lytic mRNA.

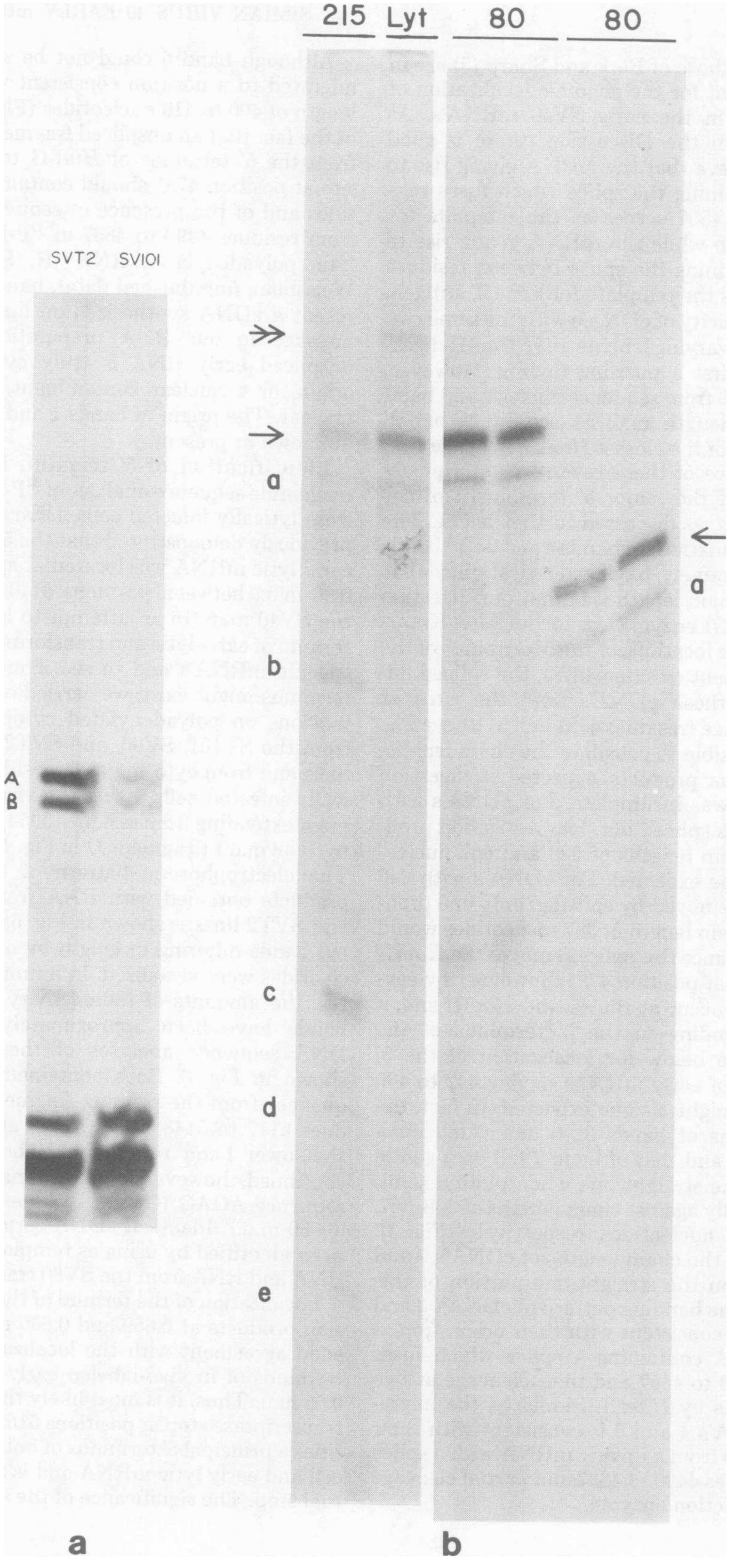
Our data and those of Berk and Sharp (5) are in good agreement for the genomic localization of these splices in the early SV40 mRNA's. As brought out in the Discussion, there is good reason to believe that the mRNA giving rise to cDNA's containing the splice which fuses residues 4490 to 4557 serves as the template for small t antigen while the mRNA giving rise to cDNA's containing the splice between residues 4490 to 4837 is the template for large T antigen.

The multiplicity of cDNA's with the same two splices but of varying lenghts after *Hae*III digestion was at first a puzzling finding. However, bands 9 and 10 from sequence analysis and band 8 from chain length analysis (see Fig. 3) had 3′ termini short of the closest *Hae*III cleavage site; the significance of these reverse transcriptases stops short of the major 5′ terminus(i) of the early mRNA's, as discussed further below. The simplest explanation for bands 3 and 7 and bands 4 and 6, respectively, having identical splices but differing in chain length was that our digestion with the *Hae*III enzyme was incomplete. Knowing the precise locations of the 5′ terminus of the *Hinf*-G fragment (residue 4378), the spliced nucleotides in these cDNA's, and the sites of *Hae*III cleavage (residues 4781 and 5110; see Fig. 1), it was possible to calculate the chain lengths of the various products expected if digestion with *Hae*III was incomplete. For cDNA's with 66 nucleotides spliced out, two restriction products with chain lengths of 338 and 667 nucleotides would be expected. For cDNA's with 346 nucleotides removed by splicing, only one product with a chain length of 387 nucleotides would be expected, since this splice removes the *Hae*III cleavage site at position 4781; however, if cleavage does not occur at the second *Hae*III site, a product extending to the 5′ terminus of the molecule (see below for localization of the 5′ terminus [i] of early mRNA) having 442 to 430 nucleotides might also be expected. In fact, the gel migrations of bands 3, 4, and 6 fell on a straight line, and that of band 7 fell on a gentle upslope of the straight line when plotted semilogarithmically against chain lengths of 338, 387, 430, and 667 nucleotides, respectively (Fig. 3); furthermore, the chain lengths of cDNA's 9 and 10 also fell on the straight-line portion of this plot. Thus, the banding pattern of cDNA's 3 and 7 in Fig. 1 is consistent with their origin from a single mRNA containing a splice which fuses residues 4490 to 4557 and their cleavage at two different sites by *Hae*III. Similarly, the migration of cDNA's 4 and 6 is consistent with their transcription from a unique mRNA with a splice fusing residues 4490 to 4837 and partial cleavage by this restriction enzyme.

Although band 5 could not be sequenced, it migrated to a position consistent with a chain length of 400 to 410 nucleotides (Fig. 3). In view of the fact that an unspliced fragment extending from the 5′ terminus of *Hinf*-G to the *Hae*III site at position 4781 should contain 404 nucleotides and of the presence of sequences derived from residues 4490 to 4837 in ^{32}P-labeled early lytic polyadenylated RNA (R. Dhar and S. Weissman, unpublished data), band 5 may represent a cDNA synthesized on unspliced early message in our RNA preparation. Whether unspliced early RNA is truly cytoplasmic in origin, or a nuclear contaminant, is presently unclear. The origin of bands 1 and 2 in Fig. 1 is unknown at present.

Identification of 5′ termini. By means of nucleotide sequence analysis of ^{32}P-labeled RNA from lytically infected cells, Dhar et al. (16–18) previously demonstrated that the 5′ terminus of early lytic mRNA was located at approximately 0.66 m.u., between positions 5140 and 5150 on the SV40 map. In an attempt to localize the 5′ termini of early lytic and transformed cells virus-specific mRNA's and to see if more than one terminus might exist, we carried out primer extensions on polyadenylated cytoplasmic RNA from the SV101, SV80, and SVT2 transformed lines and from cytosine arabinoside-treated lytically infected cells, using the viral DNA fragment extending from residues 5054 to 5089 (0.641 to 0.648 m.u.) (fragment D in Fig. 1) as a primer. The electrophoretic pattern of the extended products obtained with RNA from the SV101 and SVT2 lines is shown in Fig. 5a. Two principal bands differing in length by only a few nucleotides were visualized. In a number of analyses, the amounts of radioactivity in these two bands have been approximately equal. The DNA sequence analyses of these bands are shown in Fig. 6. Both contained identical sequences from the priming fragment up to residues 5147 to 5148 (0.659 m.u.), at which point the lower band terminated. The upper band continued, however, and terminated with the sequence AGAG (GC) at residues 5152 to 5154 (0.660 m.u.). Identical cDNA termini have also been identified by using as templates early lytic RNA and RNA from the SV80 transformed line.

Localization of the termini of these two extension products at 0.659 and 0.660 m.u. is in very good agreement with the localization of the 5′ terminus of in vivo-labeled early lytic RNA at 0.66 m.u. Thus, it is most likely that the reverse transcriptase stop at positions 5152 to 5154 identifies a principal 5′ terminus of both transformed cell and early lytic mRNA and is not an artifactual stop. The significance of the second reverse


SVT2
SVI01
A
B
215
Lyt
80
80
a
b
c
d
e
a
a
b

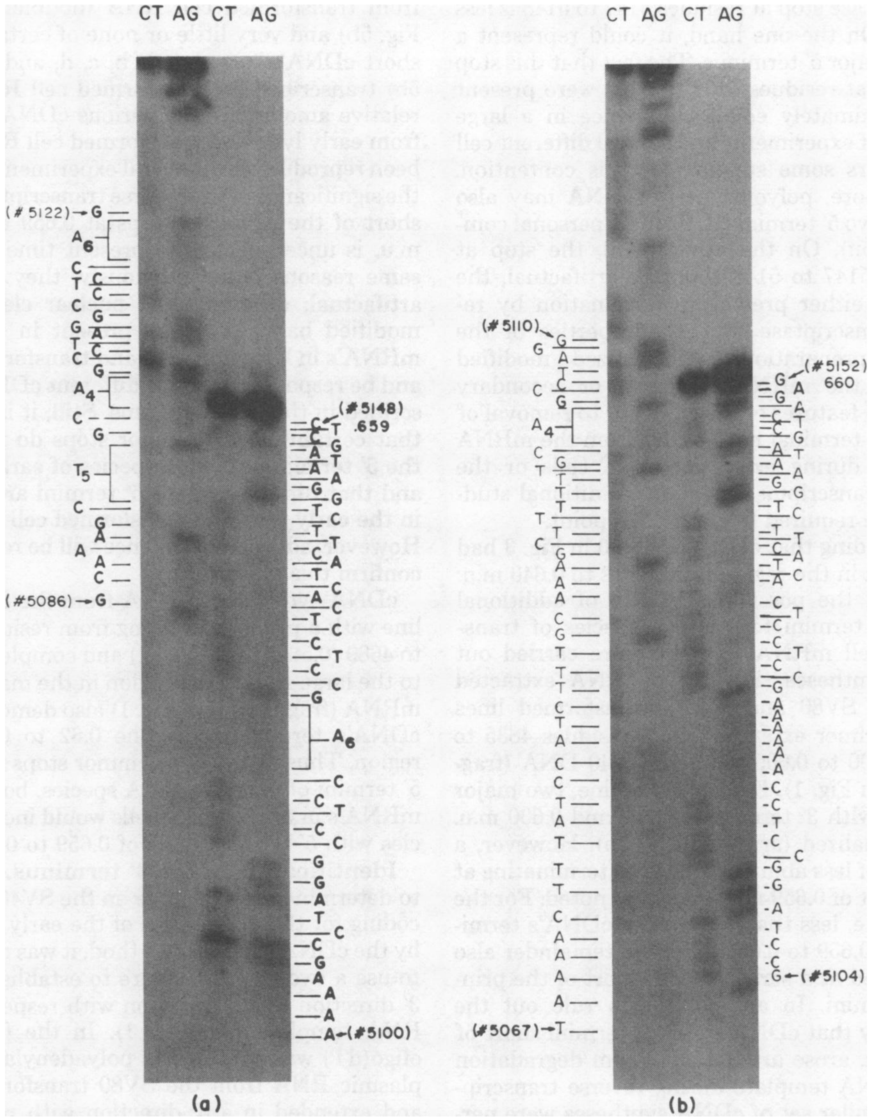


FIG. 6. *Electrophoretic fractionations of the cleavage products obtained when the two major cDNA's of the 0.641 to 0.649 m.u. primer (Fig. 5) were subjected to DNA sequencing by the method of Maxam and Gilbert (34). DNA sequences are read as described in the legend to Fig. 4.*

FIG. 5. *Autoradiograms of the 8% polyacrylamide-7 M urea gel electrophoreses of the 5′ ^{32}P-labeled products obtained by reverse transcriptase-catalyzed extension of (a) the 0.641 to 0.649 m.u. primer bound to RNA isolated from the SVT2 and SV101 transformed lines and (b) the 0.600 to 0.609 m.u. primer bound to RNA isolated from the SV215 and SV80 transformed lines and early lytic (Lyt) RNA. In (b), all six channels came from the same electrophoretic gel; however, migration of bands in the four channels on the right was slightly more rapid as a result of a slight inequality in electrical current running through opposite sides of the gel. The cDNA's synthesized on SV80 RNA were subjected to electrophoresis for two different lengths of time; the two center channels in (b), for the same period of time as the cDNA's derived from SV215 and early lytic RNA; and the two rightmost channels, for a longer period of time. The longer SV80 cDNA electrophoresis demonstrates that the principal cDNA band is a doublet. The methods used for primer extension and electrophoreses are provided in Materials and Methods and prior reports (23; Ghosh et al., in press).*

transcriptase stop at residues 5147 to 5148 is less certain. On the one hand, it could represent a second major 5′ terminus. The fact that this stop and that at residues 5152 to 5154 were present in approximately equal abundance in a large number of experiments and several different cell lines offers some support for this contention. Furthermore, polyoma early mRNA may also contain two 5′ termini (R. Kamen, personal communication). On the other hand, the stop at residues 5147 to 5148 could be artifactual, the result of either premature termination by reverse transcriptase (due to properties of the enzyme preparation, unrecognized modified bases in the mRNA template, or secondary structural features of the RNA) or to removal of several 5′-terminal nucleotides from the mRNA template during extraction from cells or the reverse transcriptase reaction. Additional studies will be required to clarify this point.

The finding that cDNA's 8 to 10 in Fig. 3 had 3′ termini in the region from 0.622 to 0.640 m.u. suggested the possible existence of additional minor 5′ termini for various species of transformed cell mRNA. We therefore carried out cDNA syntheses (Fig. 5b) on RNA extracted from the SV80 and SV215 transformed lines with a primer extending from residues 4835 to 4883 (0.600 to 0.609 m.u.) on SV40 DNA (fragment C in Fig. 1). For the SV80 line, two major cDNA's with 3′ termini at 0.659 and 0.600 m.u. were visualized (arrow in Fig. 5b). However, a number of less abundant cDNA's terminating at sites short of 0.659 m.u. were also noted. For the SV215 line, less than half of the cDNA's terminated at 0.659 to 0.660 m.u.; the remainder also terminated at a series of sites short of the principal termini. In an attempt to rule out the possibility that cDNA's with 3′ termini short of 0.659 m.u. arose artifactually from degradation of the RNA template during reverse transcription, a similar set of cDNA syntheses were performed on the SV215 RNA template with addition of various amounts (0, 13, 26, or 130 U) of placental RNase inhibitor to the transcription reaction. The relative amounts of cDNA terminating short of 0.659 m.u. were unchanged by the addition of this inhibitor (results not presented). A similar cDNA synthesis was also carried out with early lytic RNA and the primer extending from 0.600 to 0.609 m.u. As expected, the two principal cDNA's terminated at 0.659 and 0.660 m.u. In addition, one shorter cDNA terminating just short of 0.659 m.u. comigrated with similar cDNA's derived from the transformed lines (band a in Fig. 5b). However, cDNA copied from early lytic RNA included one species longer than that seen among cDNA's copied from transformed cell RNA (double arrow in Fig. 5b) and very little or none of certain of the short cDNA's (e.g., bands b, c, d, and e in Fig. 5b) transcribed on transformed cell RNA. The relative amounts of the various cDNA's copied from early lytic and transformed cell RNA have been reproducible in several experiments. Again, the significance of the reverse transcriptase stops short of the prinicpal stops at 0.659 and 0.660 m.u. is uncertain at the present time. For the same reasons noted previously, they could be artifactual; differences in nuclear cleavage or modified bases could be present in the early mRNA's in lytic infection and transformed cells and be responsible for the different cDNAs transcribed in these two systems. Still, it is possible that certain of these minor stops do represent the 5′ termini of in vivo species of early mRNA and that different minor 5′ termini are present in the early lytic and transformed cell mRNA's. However, additional evidence will be required to confirm this point.

cDNA synthesis on RNA from the SV80 cell line with a primer extending from residues 4629 to 4689 (0.560 to 0.572 m.u.) and complementary to the large spliced out region in the major early mRNA (fragment B in Fig. 1) also demonstrated cDNA's terminating in the 0.62 to 0.65 m.u. region. Thus, if any of the minor stops mark the 5′ termini of in vivo mRNA species, both major mRNA's in transformed cells would include species with 5′ termini short of 0.659 to 0.660 m.u.

Identification of the 3′ terminus. In order to determine the exact site on the SV40 genome coding for the 3′ terminus of the early mRNA's by the cDNA synthetic method, it was necessary to use a two-step procedure to establish a 5′ → 3′ direction of transcription with respect to the RNA template (see Fig. 1). In the first step, oligo(dT) was annealed to polyadenylated cytoplasmic RNA from the SV80 transformed line and extended in a 3′ direction with respect to DNA (5′ with respect to the RNA template) with reverse transcriptase in the presence of actinomycin. In a second step, isolated viral cDNA was duplexed with the residue 2571 to 2586 (0.166 to 0.169 m.u.) SV40 restriction fragment labeled in the 5′ terminal position with ^{32}P, and extension was carried out again with reverse transcriptase (in a 3′ direction with respect to both the DNA primer and the original RNA template). After deproteinization and alkaline denaturation, the single-stranded cDNA products were subjected to electrophoresis on an 8% polyacrylamide–7 M urea gel (Fig. 7a). Six major bands were visualized, all of which on DNA sequence analysis (Fig. 7b) showed the sequence AGCTTATAA . . . TCACTGC, extending from

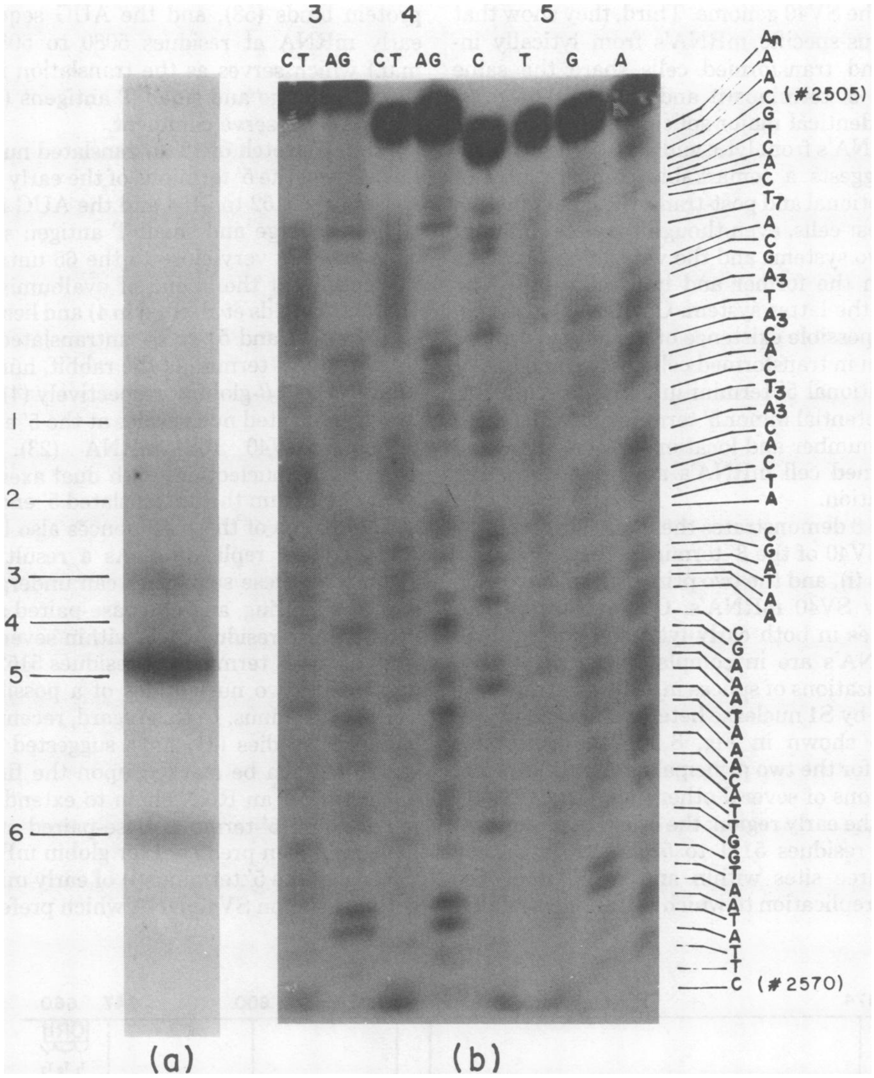


FIG. 7. *(a) Autoradiogram of the 8% polyacrylamide-7 M urea gel electrophoresis of the 5′ ^{32}P-labeled products obtained by reverse transcriptase-catalyzed extension of the 0.166 to 0.169 m.u. primer bound to cDNA complementary to the 3′ terminus of polyadenylated cytoplasmic RNA isolated from the SV80-transformed line. The methods used for primer extension and electrophoresis are provided in Materials and Methods and prior reports (23; Ghosh et al., in press). (b) Electrophoretic fractionations of the cleavage products obtained when bands 3 to 5 from Fig. 7a were subjected to DNA sequence analysis. DNA sequences are read as described in the legend of Fig. 4.*

residues 2572 to 2505 (0.166 to 0.153 m.u.) on the plus strand of viral DNA. Following this sequence, all the bands contained stretches of polydeoxyadenylic acid of varying length. These results clearly indicate that the 3′ terminus of transformed cell early mRNA's is copied from the C residue at position 2505 or the A residue at position 2504 (0.153 m.u.) on the minus strand of viral DNA. It is of interest that the sequence purine-C precedes the polyadenylic acid tract at the 3′ termini of early and late SV40 mRNA's (45) and human α- and β-globin (42, 61, 62) and rabbit β-globin (20, 42) mRNA's.

DISCUSSION

Our present results are significant in several major respects. First, they demonstrate two principal splices, a single major 3′ terminus, a principal 5′ terminus, and possibly a second principal 5′ terminus in virus-specific mRNA's from SV40-transformed mouse and human cells. Second, they permit precise localization of these sites on the SV40 genome. Third, they show that early virus-specific mRNA's from lytically infected and transformed cells share the same principal 5′ terminus(i) and splices. The presence of identical major splices and 5′ termini in viral mRNA's from lytic and transforming infections suggests a remarkable comparability of transcriptional and post-transcriptional function in the host cells, even though they are different in the two systems and the viral DNA template is free in the former and integrated into host DNA in the latter system. Our results also suggest the possible existence of a number of minor 5′ termini in transformed cell mRNA and one or two additional 5′ termini in early lytic mRNA. These potential minor 5′ termini and differences in their number and locations in early lytic and transformed cell mRNA's are presently under investigation.

Figure 8 demonstrates the localizations on the map of SV40 of the 3′ terminus, the principal 5′ terminus (i), and the two prinicpal splices within the early SV40 mRNA's. Our localizations of splice sites in both early lytic and transformed cell mRNA's are in complete agreement with the localizations of splices in early lytic mRNA's obtained by S1 nuclease heteroduplex digestions (5). Also shown in Fig. 8 are the structures deduced for the two principal early mRNA's and localizations of several other important sites related to the early region: the origin of replication between residues 5111 to 5195 (0.652 to 0.668 m.u.), three sites within and overlapping the origin of replication to which a T antigen-related protein binds (58), and the AUG sequence on early mRNA at residues 5080 to 5082 (0.646 m.u.) which serves as the translation initiation codon for large and small T antigens (38). Several points deserve comment.

First, a stretch of 72 untranslated nucleotides lie between the 5′ terminus of the early mRNA's at residues 5152 to 5154 and the AUG initiation codon for large and small T antigen synthesis. This figure is very close to the 65 untranslated nucleotides at the 5′ end of ovalbumin mRNA (L. McReynolds et al. cited in 4) and lies between the 33 to 37 and 51 to 54 untranslated nucleotides at the 5′ termini of the rabbit, human, and mouse α- and β-globins, respectively (4), and the 243 untranslated nucleotides at the 5′ end of the principal SV40 16S mRNA (23). Several stretches of nucleotides with dual axes of symmetry lie within the untranslated 5′ end of early mRNA; some of these sequences also lie within the origin of replication. As a result of their symmetry, these sequences can undergo extensive base-pairing, and one base-paired stem can extend up to residue 5145, within seven nucleotides of the 5′ terminus at residues 5152 to 5154 and within two nucleotides of a possible additional 5′ terminus. In this regard, recent crystallographic studies (31) have suggested that cap structures can be stacked upon the first oligonucleotide of an RNA chain to extend a 5′-terminal stem; 5′-terminal base-paired structures have also been predicted for globin mRNA's.

Second, the 5′ terminus(i) of early mRNA lies within a site on SV40 DNA which preferentially

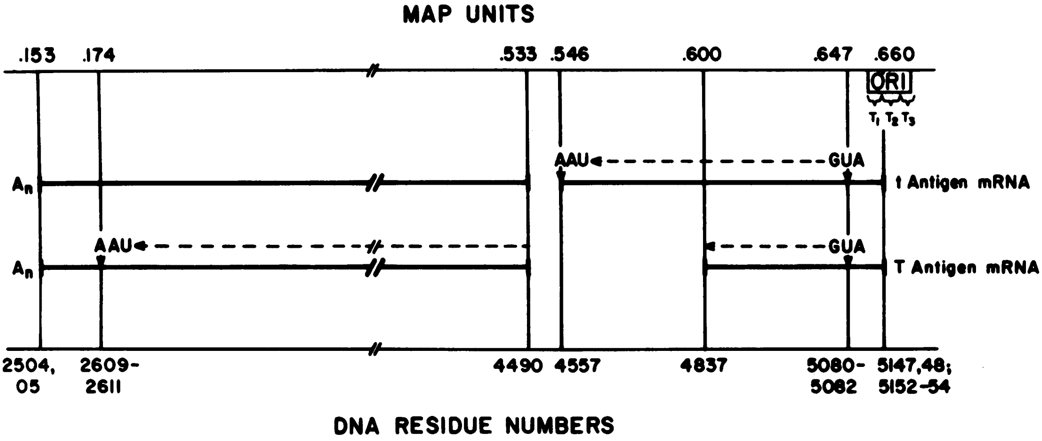


FIG. 8. *Map of the SV40 early gene region showing the expanses of the two principal early mRNA's and the regions of these mRNA's (dashed lines) which appear to code for the small t and large T antigens (see text). Transcription of these mRNA's and translation of the proteins is from right to left.* A_n *indicates 3′ polyadenylic acid mRNA termini. AUG and UAA indicate initiation and termination codons for translation of the two T antigens. The localization of the origin of replication (ORI) as well as three sites overlapping ORI which bind a T antigen-related protein are also shown. Map units are fractional genomic lengths from the unique EcoRI cleavage site, and DNA residue numbers are taken from Reddy et al. (44).*

binds a T antigen-related protein isolated from an SV40 adenovirus hybrid (58; Fig. 8). Furthermore, the following lines of evidence suggest that at least one promoter for early gene transcription lies at or very close to the template for the 5′ end of early mRNA, within the region of DNA binding the T antigen-related protein: (i) a number of deletion mutants of SV40 which lack sequences extending from about 0.67 m.u. within the late gene region clockwise through the remainder of the late region express early mRNA (14, 35, 50; K. N. Subramanian and T. Shenk, in press and personal communication); (ii) early mRNA is expressed in SV40-transformed cells in which the late gene region is interrupted (7); and (iii) mutants of SV40 with an extra copy of the origin of DNA replication inserted at various loci on the viral genome synthesize polyadenylic acid terminal RNA with 5′ termini transcribed from the inserted DNA (T. Shenk, personal communication). The proximity of the template for the 5′ terminus(i) of early mRNA, T antigen-binding sites, and an early gene promoter is significant since transcription of the early gene region is "autoregulated" by its end product, T antigen — low levels of the antigen permitting and high levels inhibiting early gene transcription (3, 30, 45, 56). The mechanism by which T antigen inhibits early gene transcription is unknown. However, in a relevant procaryotic system, transcription of the *lac* operon, *lac* repressor binds to an operator site which codes for the 5′ terminus of *lac* mRNA (19) in a manner formally similar to binding of T antigen at sites coding for the 5′ terminus(i) of SV40 early mRNA. Based on this model and the proximity of T antigen binding sites, an early gene promoter, and the 5′ terminus(i) of early mRNA, it seems reasonable to suggest that T antigen acts as a repressor of initiation of early gene transcription by binding to SV40 DNA at regions overlapping the early gene promoter and transcription initiation site(s). If the reverse transcriptase stops on early mRNA at 0.659 m.u. and from 0.62 to 0.65 m.u. correspond to the 5′ termini of mRNA's, it would be interesting to know whether these sites represent transcription initiation sites and, if so, whether synthesis of these mRNA's is coordinately autoregulated with that of the major mRNA with a 5′ terminus at 0.660 m.u.

Lastly, it may now be possible to predict the sequences of the T antigens coded for by the two major early mRNA's on the basis of the sequence of SV40 DNA (22, 44), localization of their translation initiation codons to residues 5082 to 5080 (0.646 m.u.) (38), and demonstration of the precise sequences across their splices in the 0.533 to 0.600 m.u. region. The only requirement that must be met before a prediction can be made with full confidence is that additional splices be ruled out in these mRNA's. Two lines of evidence suggest fulfillment of this requirement: (i) the results presented in these experiments showing colinearity of SV40 DNA and early mRNA sequences from 0.646 to 0.600 m.u. and additional results of our cDNA sequencing studies which show that at least a portion of early mRNA molecules contain sequences colinear with SV40 DNA from 0.533 to 0.153 m.u., and (ii) the demonstration by Berk and Sharp (5) that early SV40 mRNA protects a large fraction of, if not all, radiolabeled SV40 DNA from digestion with S1 nuclease in approximately these same genomic regions in a system in which S1 nuclease can recognize and cleave through single mismatched bases of a heteroduplex (60).

With regard to small t antigen, the available genetic and biochemical data strongly suggest that the mRNA containing the splice which fuses residues 4490 to 4557 (0.533 to 0.547 m.u.) serves as the template for this protein. One may thus predict that the template for small t antigen consists of a continuous stretch of 174 sense codons extending from the known AUG initiator at residues 5082 to 5080 to a UAA termination codon at residues 4559 to 4557 (0.546 m.u.) (Fig. 8). The sequence of these 174 codons and the amino acids they specify are shown in Fig. 9. The molecular weight of the unmodified small t antigen composed of these amino acids is 20,500, in good agreement with estimates based on electrophoretic migration. It is noteworthy that the splice in the mRNA for small t antigen occurs immediately after the UAA termination codon.

The nucleotide sequence of the mRNA containing the splice between residues 4490 and 4837 (0.533 to 0.600 m.u.) is compatible with its coding for a T antigen of high molecular weight. Starting with the AUG initiation triplet at residues 5082 to 5080, this mRNA contains codons for 82 amino acids before reaching the splice point at residue 4837. From the expanses of the two early mRNA's, it is seen that these 82 amino acids are shared by small and large T antigens. The 3′ segment of the mRNA for large T antigen starting from residue 4490 then contains a total of 626 consecutive sense triplets before reaching a UAA termination triplet at 0.174 m.u. (residues 2609 to 2611) provided no additional small splices are present in this region of the mRNA. The predicted chain length of large T antigen from this analysis is 708 amino acids, coding for an unmodified protein of about 85,000 daltons.

```
          Met Asp Lys Val Leu Asn Arg Glu Glu Ser Leu Gln Leu Met Asp Leu Leu Gly Leu Glu  20
t and T - AUG GAU AAA GUU UUA AAC AGA GAG GAA UCU UUG CAG CUA AUG GAC CUU CUA GGU CUU GAA

          Arg Ser Ala Trp Gly Asn Ile Pro Leu Met Arg Lys Ala Tyr Leu Lys Lys Cys Lys Glu  40
          AGG AGU GCC UGG GGG AAU AUU CCU CUG AUG AGA AAG GCA UAU UUA AAA AAA UGC AAG GAG

          Phe His Pro Asp Lys Gly Gly Asp Glu Glu Lys Met Lys Lys Met Asn Thr Leu Tyr Lys  60
          UUU CAU CCU GAU AAA GGA GGA GAU GAA GAA AAA AUG AAG AAA AUG AAU ACU CUG UAC AAG

          Lys Met Glu Asp Gly Val Lys Tyr Ala His Gln Pro Asp Phe Gly Gly Phe Trp Asp Ala  80
          AAA AUG GAA GAU GGA GUA AAA UAU GCU CAU CAA CCU GAC UUU GGA GGC UUC UGG GAU GCA

          Thr Glu                                                                          82
          ACU GAG

          Val Phe Ala Ser Ser Leu Asn Pro Gly Val Asp Ala Met Tyr Cys Lys Gln Trp Pro Glu 102
      t - GUA UUU GCU UCU UCC UUA AAU CCU GGU GUU GAU GCA AUG UAC UGC AAA CAA UGG CCU GAG

          Cys Ala Lys Lys Met Ser Ala Asn Cys Ile Cys Leu Leu Cys Leu Leu Arg Met Lys His 122
          UGU GCA AAG AAA AUG UCU GCU AAC UGC AUA UGC UUG CUG UGC UUA CUG AGG AUG AAG CAU

          Glu Asn Arg Lys Leu Tyr Arg Lys Asp Pro Leu Val Trp Val Asp Cys Tyr Cys Phe Asp 142
          GAA AAU AGA AAA UUA UAC AGG AAA GAU CCA CUU GUG UGG GUU GAU UGC UAC UGC UUC GAU

          Cys Phe Arg Met Trp Phe Gly Leu Asp Leu Cys Glu Gly Thr Leu Leu Leu Trp Cys Asp 162
          UGC UUU AGA AUG UGG UUU GGA CUU GAU CUU UGU GAA GGA ACC UUA CUU CUG UGG UGU GAC

          Ile Ile Gly Gln Thr Thr Tyr Arg Asp Leu Lys Leu                                 174
          AUA AUU GGA CAA ACU ACC UAC AGA GAU UUA AAG CUA UAA

          Ile Pro Thr Tyr Gly Thr Asp Glu Trp Glu Gln Trp Trp Asn Ala Phe Asn Glu Glu Asn 102
      T - AUU CCA ACC UAU GGA ACU GAU GAA UGG GAG CAG UGG UGG AAU GCC UUU AAU GAG GAA AAC

          Leu Phe Cys Ser Glu Glu Met Pro Ser Ser Asp Asp Glu Ala Thr Ala Asp Ser Gln His 122
          CUG UUU UGC UCA GAA GAA AUG CCA UCU AGU GAU GAU GAG GCU ACU GCU GAC UCU CAA CAU

          Ser Thr Pro Pro Lys Lys Lys Arg Lys Val Glu Asp Pro Lys Asp Phe Pro Ser Glu Leu 142
          UCU ACU CCU CCA AAA AAG AAG AGA AAG GUA GAA GAC CCC AAG GAC UUU CCU UCA GAA UUG

          Leu Ser Phe Leu Ser His Ala Val Phe Ser Asn Arg Thr Leu Ala Cys Phe Ala Ile Tyr 162
          CUA AGU UUU UUG AGU CAU GCU GUG UUU AGU AAU AGA ACU CUU GCU UGC UUU GCU AUU UAC

      Thr Thr Lys Glu Lys Ala Ala Leu Leu Tyr Lys Lys Ile Met Glu Lys Tyr Ser Val Thr  182
      ACC ACA AAG GAA AAA GCU GCA CUG CUA UAC AAG AAA AUU AUG GAA AAA UAU UCU GUA ACC

      Phe Ile Ser Arg His Asn Ser Tyr Asn His Asn Ile Leu Phe Phe Leu Thr Pro His Arg  202
      UUU AUA AGU AGG CAU AAC AGU UAU AAU CAU AAC AUA CUG UUU UUU CUU ACU CCA CAC AGG

      His Arg Val Ser Ala Ile Asn Asn Tyr Ala Gln Lys Leu Cys Thr Phe Ser Phe Leu Ile  222
      CAU AGA GUG UCU GCU AUU AAU AAC UAU GCU CAA AAA UUG UGU ACC UUU AGC UUU UUA AUU

      Cys Lys Gly Val Asn Lys Glu Tyr Leu Met Tyr Ser Ala Leu Thr Arg Asp Pro Phe Ser  242
      UGU AAA GGG GUU AAU AAG GAA UAU UUG AUG UAU AGU GCC UUG ACU AGA GAU CCA UUU UCU

      Val Ile Glu Glu Ser Leu Pro Gly Gly Leu Lys Glu His Asp Phe Asn Pro Glu Glu Ala  262
      GUU AUU GAG GAA AGU UUG CCA GGU GGG UUA AAG GAG CAU GAU UUU AAU CCA GAA GAA GCA

      Glu Glu Thr Lys Gln Val Ser Trp Lys Leu Val Thr Glu Tyr Ala Met Glu Thr Lys Cys  282
      GAG GAA ACU AAA CAA GUG UCC UGG AAG CUU GUA ACA GAG UAU GCA AUG GAA ACA AAA UGU

      Asp Asp Val Leu Leu Leu Leu Gly Met Tyr Leu Glu Phe Gln Tyr Ser Phe Glu Met Cys  302
      GAU GAU GUG UUG UUA UUG CUU GGG AUG UAC UUG GAA UUU CAG UAC AGU UUU GAA AUG UGU

      Leu Lys Cys Ile Lys Lys Glu Glu Pro Ser His Tyr Lys Tyr His Glu Lys His Tyr Ala  322
      UUA AAA UGU AUU AAA AAA GAA CAG CCC AGC CAC UAU AAG UAC CAU GAA AAG CAU UAU GCA
```

```
Asn Ala Ala Ile Phe Ala Asp Ser Lys Asn Gln Lys Thr Ile Cys Gln Gln Ala Val Asp 342
AAU GCU GCU AUA UUU GCU GAC AGC AAA AAC CAA AAA ACC AUA UGC CAA CAG GCU GUU GAU

Thr Val Leu Ala Lys Lys Arg Val Asp Ser Leu Gln Leu Thr Arg Glu Gln Met Leu Thr 362
ACU GUU UUA GCU AAA AAG CGG GUU GAU AGC CUA CAA UUA ACU AGA GAA CAA AUG UUA ACA

Asn Arg Phe Asn Asp Leu Leu Asp Arg Met Asp Ile Met Phe Gly Ser Thr Gly Ser Ala 382
AAC AGA UUU AAU GAU CUU UUG GAU AGG AUG GAU AUA AUG UUU GGU UCU ACA GGC UCU GCU

Asp Ile Glu Glu Trp Met Ala Gly Val Ala Trp Leu His Cys Leu Leu Pro Lys Met Asp 402
GAC AUA GAA GAA UGG AUG GCU GGA GUU GCU UGG CUA CAC UGU UUG UUG CCC AAA AUG GAU

Ser Val Val Tyr Asp Phe Leu Lys Cys Met Val Tyr Asn Ile Pro Lys Lys Arg Tyr Trp 422
UCA GUG GUG UAU GAC UUU UUA AAA UGC AUG GUG UAC AAC AUU CCU AAA AAA AGA UAC UGG

Leu Phe Lys Gly Pro Ile Asp Ser Gly Lys Thr Thr Leu Ala Ala Ala Leu Leu Glu Leu 442
CUG UUU AAA GGA CCA AUU GAU AGU GGU AAA ACU ACA UUA GCA GCU GCU UUG CUU GAA UUA

Cys Gly Gly Lys Ala Leu Asn Val Asn Leu Pro Leu Asp Arg Leu Asn Phe Glu Leu Gly 462
UGU GGG GGG AAA GCU UUA AAU GUU AAU UUG CCC UUG GAC AGG CUG AAC UUU GAG CUA GGA

Val Ala Ile Asp Gln Phe Leu Val Val Phe Glu Asp Val Lys Gly Thr Gly Gly Glu Ser 482
GUA GCU AUU GAC CAG UUU UUA GUA GUU UUU GAG GAU GUA AAG GGC ACU GGA GGG GAG UCC

Arg Asp Leu Pro Ser Gly Gln Gly Ile Asn Asn Leu Asp Asn Leu Arg Asp Tyr Leu Asp 502
AGA GAU UUG CCU UCA GGU CAG GGA AUU AAU AAC CUG GAC AAU UUA AGG GAU UAU UUG GAU

Gly Ser Val Lys Val Asn Leu Glu Lys Lys His Leu Asn Lys Arg Thr Gln Ile Phe Thr 522
GGC AGU GUU AAG GUA AAC UUA GAA AAG AAA CAC CUA AAU AAA AGA ACU CAA AUA UUU ACC

Pro Gly Ile Val Thr Met Asn Glu Phe Ser Val Pro Lys Thr Leu Gln Ala Arg Phe Thr 542
CCU GGA AUA GUC ACC AUG AAU GAG UUC AGU GUG CCU AAA ACA CUG CAG GCC AGA UUU GUA

Lys Gln Ile Asp Phe Arg Ala Lys Asp Tyr Leu Lys His Cys Leu Glu Arg Ser Glu Phe 562
AAA CAA AUA GAU UUU AGG GCC AAG GAU UAU UUA AAG CAU UGC CUG GAA CGC AGU GAG UUU

Leu Leu Glu Lys Arg Ile Ile Gln Ser Gly Ile Ala Leu Leu Leu Met Leu Ile Trp Tyr 582
UUG UUA GAA AAG AGA AUA AUU CAA AGU GGC AUU GCU UUG CUU CUU AUG UUA AUU UGG UAC

Arg Pro Val Ala Glu Phe Ala Gln Ser Ile Gln Ser Arg Ile Val Glu Trp Lys Glu Arg 602
AGA CCU GUG GCU GAG UUU GCU CAA AGU AUU CAG AGC AGA AUU GUG GAG UGG AAA GAG AGA

Leu Asp Lys Glu Phe Ser Leu Ser Val Tyr Gln Lys Met Lys Phe Asn Val Ala Met Gly 622
UUG GAC AAA GAG UUU AGU UUG UCA GUG UAU CAA AAA AUG AAG UUU AAU GUG GCU AUG GGA

Ile Gly Val Leu Asp Trp Leu Arg Asn Ser Asp Asp Asp Asp Glu Asp Ser Gln Glu Asn 642
AUU GGA GUU UUA GAU UGG CUA AGA AAC AGU GAU GAU GAU GAU GAA GAC AGC CAG GAA AAU

Ala Asp Lys Asn Glu Asp Gly Gly Glu Lys Asn Met Glu Asp Ser Gly His Glu Thr Gly 662
GCU GAU AAA AAU GAA GAU GGU GGG GAG AAG AAC AUG GAA GAC UCA GGG CAU GAA ACA GGC

Ile Asp Ser Gln Ser Gln Gly Ser Phe Gln Ala Pro Gln Ser Ser Gln Ser Val His Asp 682
AUU GAU UCA CAG UCC CAA GGC UCA UUU CAG GCC CCU CAG UCC UCA CAG UCU GUU CAU GAU

His Asn Gln Pro Tyr His Ile Cys Arg Gly Phe Thr Cys Phe Lys Lys Pro Pro Thr Pro 702
CAU AAU CAG CCA UAC CAC AUU UGU AGA GGU UUU ACU UGC UUU AAA AAA CCU CCC ACA CCU

Pro Pro Glu Pro Glu Thr                                                          708
CCC CCU GAA CCU GAA ACA UAA
```

FIG. 9. *Predicted amino acid sequences for small t and large T antigens based on the nucleotide sequences of the two prinicpal early mRNA's of SV40. The figure is in three sections: the first shows sequences of amino acids 1 to 82 common to small t and large T antigens, the second shows amino acids 83 to 174 of small t antigen, and the third shows amino acids 83 to 708 of large T antigen. The genomic localizations of the translation initiation and termination codons for these proteins are shown in Fig. 8.*

Figure 9 provides the full sequence of the 708 codons present in the presumptive T antigen mRNA and the predicted amino acid sequence for this protein.

With description of two splices in the SV40 early mRNA's, a total of eight splices have now been demonstrated in the RNAs of SV40. The function served by splicing is clear only for the large T antigen mRNA, where it is essential to remove in phase translation termination codons and reconstitute the coding sequence for this antigen. Splices in the mRNA's for globin (57), ovalbumin (11), and mouse immunoglobulin (9) also serve to establish continuous coding sequences for these proteins. We have thus far only been able to speculate on possible functions of splices in the SV40 late RNAs. We have suggested that the major splice in 16S mRNA, which removes 938 nucleotides, may serve to bring the initiation codon for VP1 translation nearer to the 5′ end of this molecule, possibly making it more accessible to ribosomes (23). In the case of 19S late RNA, the principal splice removes a segment of untranslated RNA which can base-pair with the initiation codon for VP2, and we have proposed that this may make the VP2 initiation codon available for initiation of translation (Ghosh et al., in press). We have also suggested that an intraleader splice present in certain species of 16S RNA may play a role in directing the subsequent processing of these RNAs (Reddy et al., in press). In the case of the early mRNA which may code for small t antigen, the function of the splice which removes untranslated codons downstream from the coding region is totally obscure. On the one hand, this small splice may be an intermediate in formation of the larger splice in large T mRNA; alternatively, it may prevent formation of the larger splice; lastly, several groups have suggested that some splicing step may be required for transport of mRNA's from nucleus to cytoplasm.

We have previously pointed out certain structural features common to the precursors of all spliced late SV40 RNAs which may play roles in the splicing reaction (Ghosh et al., in press; Reddy et al., in press). These include (i) the uniform presence of reiterated dipurines or tri- or tetranucleotides containing dipurines (AGGU, GGU, AG, GG) at pairs of sites which undergo splicing, (ii) U-rich stretches at the 3′ ends of the segments of RNA which are removed by splicing, (iii) the sequences CCA or CCU to the 3′ side of the duplicated sequences which in turn lie to the 3′ side of the RNA segments that are removed by splicing, and (iv) the ability of uncleaved precursors to undergo base pairings which result in approximation of sites which undergo cleavage. Of interest, identical sequences (CAGG, AGGU, UCAG, GCAG, GG, and G) have also been found at sites undergoing splicing in a mouse immunoglobulin light-chain mRNA (59) and ovalbumin mRNA (10). On the basis of these features, we have suggested that the splicing reaction may involve approximation of sites which will undergo splicing, recognition of sites which include the reiterated sequences within a base-paired three dimensional structure by a processing enzyme, cleavage at these sites, and fusion of the resultant 5′ and 3′ termini of RNA segments. The continuous precursor of the spliced early mRNA's that we have described in this report also possesses these structural features: the dipurine AG is duplicated at sites which undergo splicing; 16 of 29 nucleotides at the 3′ end of the RNA segments which are removed by splicing are U residues; the sequence CCA lies to the 3′ side of the AG sequence at the 3′ end of the spliced out segments; and the early RNA precursor can undergo base pairing which results in approximation of sites which undergo splicing (Fig. 10).

We have noted above that at least a portion of early mRNA molecules are continuous tran-

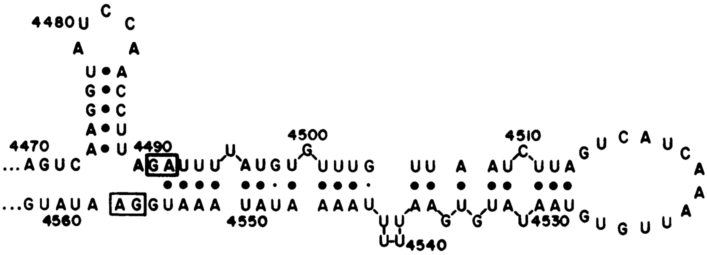

4480 U C C
A A
U • A
G • C
G • C
A • U
4470 A • U 4490 U 4500 G 4510 C G U C A U C
....AGUC AGAUUU AUGU UUUG UU A AU UUA A4520
....GUAUA AG GUAAA UAUA AAAU AA U UA AAU A A
4560 4550 UU G G U4530 G U G U U
U-U4540

FIG. 10. *Base-pairing model of a segment of continuous early RNA which undergoes splicing to yield the mRNA with a gap from residues 4490 to 4557. The model has been constructed in a way which results in approximation of sequences which undergo splicing and maximal base-pairing in this region. GC and AU base pairs are indicated by heavy dots and GU pairs, by lighter ones. Duplicated AG sequences which lie at splice points are enclosed in boxes. Figures represent DNA residue numbers (44).*

scripts of SV40 DNA from 0.533 to 0.153 m.u. However, experiments currently in progress suggest interruptions of RNA in a portion of transformed cell transcripts encoding the carboxy-terminal portion of T antigen, and further studies will be required to determine whether there are additional primary translation products encoded for partly or fully by the early SV40 gene region.

ACKNOWLEDGMENTS

We are grateful to G. Todaro, P. Mora, D. Livingston, and R. Pollack for providing us with transformed cell lines and to J. Beard, Research Resources, Viral Oncology Program, National Cancer Institute, for supplying preparations of avian myeloblastosis virus reverse transcriptase. We also thank P. Chambon for permission to refer to results in press, Regina Ezekial, Vinnie Itri, Alan McCluskey, Nick Davis, and Richard Wang for excellent technical assistance, and Betty Hackett for typing this manuscript.

The research was supported by grants from the National Cancer Institute (CA-16038), American Cancer Society, and the Leonard Eckstein Living Fund for Leukemia Research.

LITERATURE CITED

1. **Abrahams, P. J., C. Mulder, A. Van de Voorde, S. O. Warnaar, and A. J. van der Eb.** 1975. Transformation of primary rat kidney cells by fragments of simian virus 40 DNA. J. Virol. **16**:818–823.
2. **Abrahams, P. J., and A. J. van der Eb.** 1975. In vitro transformation of rat and mouse cells by DNA from simian virus 40. J. Virol. **16**:206–209.
3. **Alwine, J. C., S. I. Reed, J. Ferguson, and G. R. Stark.** 1975. Properties of T antigens induced by wild-type SV40 and tsA mutants in lytic infection. Cell **6**: 529–533.
4. **Baralle, F., and G. G. Brownlee.** 1978. AUG is the only recognizable signal sequence in the 5′ non-coding regions of eukaryotic mRNA. Nature (London) **274**:84–87.
5. **Berk, A. J., and P. A. Sharp.** 1978. Spliced early mRNAs of simian virus 40. Proc. Natl. Acad. Sci. U.S.A. **74**: 1274–1278.
6. **Blackburn, P., G. Wilson, and S. Moore.** 1977. Ribonuclease inhibitor from human placenta. J. Biol. Chem. **252**:5904–5910.
7. **Blakesley, R. W., and R. D. Wells.** 1975. Single-stranded DNA from ϕX174 and M13 is cleaved by certain restriction endonucleases. Nature (London) **257**:421–422.
8. **Botchan, M., W. Topp, and J. Sambrook.** 1976. The arrangement of simian virus 40 sequences in the DNA of transformed cells. Cell **9**:269–287.
9. **Brack, C., and S. Tonegawa.** 1977. Variable and constant parts of the immunoglobulin light chain gene of a mouse myeloma cell are 1250 non-translated bases apart. Proc. Natl. Acad. Sci. U.S.A. **74**:5652–5656.
10. **Breathnach, R., C. Benoist, K. O'Hare, F. Gannon, and P. Chambon.** 1978. Ovalbumin gene: evidence for a leader sequence in mRNA and DNA sequences at exon-intion boundaries. Proc. Natl. Acad. Sci. U.S.A. **75**:4853–4857.
11. **Breathnach, R., J. L. Mandel, and P. Chambon.** 1977. Ovalbumin gene is split in chicken DNA. Nature (London) **270**:314–319.
12. **Brugge, J. S., and J. S. Butel.** 1975. Role of simian virus 40 gene *A* function in maintenance of transformation. J. Virol. **15**:619–635.
13. **Chou, J. Y., and R. G. Martin.** 1975. Products of complementation between temperature-sensitive mutants of simian virus 40. J. Virol. **15**:127–136.
14. **Cole, C., T. Landers, S. Goff, S. Manteuil-Brutlag, and P. Berg.** 1977. Physical and genetic characterization of deletion mutants of simian virus 40 constructed in vitro. J. Virol. **24**:277–294.
15. **Crawford, L. V., C. N. Cole, A. E. Smith, E. Paucha, P. Tegtmeyer, R. Rundell, and P. Berg.** 1978. Organization and expression of early genes of simian virus 40. Proc. Natl. Acad. Sci. U.S.A. **75**:117–121.
16. **Dhar, R., K. N. Subramanian, J. Pan, and S. Weissman.** 1977. Nucleotide sequence of a fragment of SV40 DNA that contains the origin of replication and specifies the 5′ ends of "early" and "late" viral RNA. J. Biol. Chem. **252**:368–376.
17. **Dhar, R., K. N. Subramanian, B. S. Zain, A. Levine, C. Patch, and S. Weissman.** 1975. Sequences in SV40 DNA corresponding to the 'ends' of cytoplasmic mRNA. INSERM **47**:25–32.
18. **Dhar, R., K. Subramanian, B. S. Zain, J. Pan, and S. Weissman.** 1974. Nucleotide sequences about the 3′ terminus of SV40 DNA transcripts and the region where DNA synthesis is initiated. Cold Spring Harbor Symp. Quant. Biol. **39**:153–160.
19. **Dickson, R., J. Abelson, W. Burnes, and W. Reznikoff.** 1975. Genetic regulation and the lac control region. Science **187**:27–35.
20. **Efstratiadis, A., and F. C. Kafatos.** 1977. The primary structure of rabbit B-globin mRNA as determined from cloned DNA. Cell **10**:571–585.
21. **Ferdinand, F., M. Brown, and G. Khoury.** 1977. Characterization of early simian virus 40 transcriptional complexes: late transcription in the absence of detectable DNA replication. Proc. Natl. Acad. Sci. U.S.A. **74**:5443–5447.
22. **Fiers, W., R. Contreras, G. Haegeman, R. Rogiers, A. Van de Voorde, H. Van Hewverswyn, J. Van Herreweghe, G. Volchaert, and M. Ysebaert.** 1978. Complete nucleotide sequences of SV40 DNA. Nature (London) **273**:113–120.
23. **Ghosh, P. K., V. B. Reddy, J. Swinscoe, P. Lebowitz, and S. Weissman.** 1978. The 5′ terminal leader sequence of late mRNA from cells infected with simian virus 40. J. Biol. Chem. **253**:3643–3647.
24. **Horiuchi, F., and N. Zinder.** 1975. Site specific cleavage of single stranded DNA by a hemophilus restriction endonuclease. Proc. Natl. Acad. Sci. U.S.A. **72**:2555–2558.
25. **Khoury, G., J. C. Byrne, and M. Martin.** 1972. Patterns of simian virus 40 DNA transcription after acute infection of permissive and nonpermissive cells. Proc. Natl. Acad. Sci. U.S.A. **60**:1925–1928.
26. **Khoury, G., P. Howley, D. Nathans, and M. Martin.** 1975. Post-transcriptional selection of simian virus 40-specific RNA. J. Virol. **15**:433–437.
27. **Khoury, G., and M. A. Martin.** 1972. Comparison of SV40 DNA transcription *in vivo* and *in vitro*. Nature (London) New Biol. **238**:4–6.
28. **Khoury, G., M. A. Martin, I. N. H. Lee, K. J. Danna, and D. Nathans.** 1973. A map of simian virus 40 transcription sites expressed in productively infected cells. J. Mol. Biol. **78**:377–389.
29. **Khoury, G., M. A. Martin, T. N. H. Lee, and D. Nathans.** 1975. A transcriptional map of the SV40 genome in transformed cell lines. Virology **63**:263–272.
30. **Khoury, G., and E. May.** 1977. Regulation of early and late simian virus 40 transcription: overproduction of early viral RNA in the absence of functional T antigen. J. Virol. **23**:167–176.
31. **Kim, C. H., and R. H. Sarma.** 1978. Spatial configuration of an RNA 5′ terminus Nature (London) **270**:223–227.
32. **Kimura, G., and A. Itagaki.** 1975. Initiation and maintenance of cell transformation by simian virus 40: a viral

genetic property. Proc. Natl. Acad. Sci. U.S.A. 72:673-677.

33. **Martin, R. G., and J. Y. Chou.** 1975. Simian virus 40 functions required for the establishment and maintenance of malignant transformation. J. Virol. 15:599-612.

34. **Maxam, A., and W. Gilbert.** 1977. A new method for sequencing DNA. Proc. Natl. Acad. Sci. U.S.A. 74:560-564.

35. **Mertz, J. E., and P. Berg.** 1974. Viable deletion mutants of simian virus 40: selective isolation by means of a restriction endonuclease from hemophilus parainfluenzae. Proc. Natl. Acad. Sci. U.S.A. 71:4879-4883.

36. **Osborn, M., and K. Weber.** 1975. Simian virus 40 gene *A* function and maintenance of transformation. J. Virol. 15:636-644.

37. **Paucha, E., R. Harvey, and A. E. Smith.** 1978. Cell-free synthesis of simian virus 40 T-antigens. J. Virol. 28: 154-170.

38. **Paucha, E., A. Mellor, R. Harvey, A. E. Smith, R. M. Hewick, and M. D. Waterfield.** 1978. Large and small tumor antigens from simian virus 40 have identical amino termini mapping at 0.65 map units. Proc. Natl. Acad. Sci. U.S.A. 75:2165-2169.

39. **Peacock, A. C., and C. W. Dingman.** 1968. Molecular weight estimation and separation of ribonucleic acid by electrophoresis in agarose-acrylamide composite gels. Biochemistry 7:668-674.

40. **Penman, S.** 1966. RNA metabolism in the HeLa cell nucleus. J. Mol. Biol. 17:117-130.

41. **Prives, C., E. Gilboa, M. Revel, and E. Winocour.** 1977. Cell-free translation of simian virus 40 early messenger RNA coding for viral T-antigen. Proc. Natl. Acad. Sci. U.S.A. 74:457-461.

42. **Proudfoot, N. J.** 1977. Complete 3′ noncoding region sequences of rabbit and human B-globin messenger RNAs. Cell 10:559-570.

43. **Rassoulzadegan, M., B. Perbal, and F. Cuzin.** 1978. Growth control in simian virus 40-transformed rat cells: temperature-sensitive expression of the transformed phenotype in *tsA* transformants derived by agar selection. J. Virol. 28:1-5.

44. **Reddy, V. B., B. Thimmappaya, R. Dhar, K. N. Subramanian, B. S. Zain, J. Pan, M. L. Celma, P. K. Ghosh, and S. M. Weissman.** 1978. The genome of simian virus 40. Science 200:494-502.

45. **Reed, S. I., G. R. Stark, and J. C. Alwine.** 1976. Autoregulation of simian virus 40 gene A by T antigen. Proc. Natl. Acad. Sci. U.S.A. 73:3038-3087.

46. **Richardson, C.** 1971. Polynucleotide kinase from *E. coli* infected with bacteriophage T4. Prog. Nucleic Acid Res. 2:815-828.

47. **Rundell, K., T. K. Collins, P. Tegtmeyer, H. L. Ozer, C. J. Lai, and D. Nathans.** 1977. Identification of simian virus 40 protein A. J. Virol. 21:636-646.

48. **Sambrook, J., P. A. Sharp, and W. Keller.** 1972. Transcription of simian virus 40. I. Separation of the strands of SV40 DNA and hybridization of the separated strands to RNA extracted from lytically infected and transformed cells. J. Mol. Biol. 70:57-71.

49. **Sambrook, J., B. Sugden, W. Keller, and P. A. Sharp.** 1973. Transcription of simian virus 40. III. Orientation of RNA synthesis and mapping of 'early' and 'late' species of viral RNA extracted from lytically infected cells. Proc. Natl. Acad. Sci. U.S.A. 70:3711-3715.

50. **Shenk, T. E., J. Carbon, and P. Berg.** 1976. Construction and analysis of viable deletion mutants of simian virus 40. J. Virol. 18:664-671.

51. **Simmons, D. T., and M. A. Martin.** 1978. Common methionine-tryptic peptides near the amino-terminal end of primate papovavirus tumor antigens. Proc. Natl. Acad. Sci. U.S.A. 75:1131-1135.

52. **Sleigh, M. J., W. C. Topp, R. Hanich, and J. F. Sambrook.** 1978. Mutants of SV40 with an altered small t protein are reduced in their ability to transform cells. Cell 14:79-88.

53. **Smith, A. E., R. Smith, and E. Paucha.** 1978. Extraction and fingerprint analysis of simian virus 40 large and small T antigens. J. Virol. 28:140-153.

54. **Tegtmeyer, P.** 1972. Simian virus 40 deoxyribonucleic acid synthesis: the viral replicon. J. Virol. 10:591-598.

55. **Tegtmeyer, P.** 1975. Function of simian virus 40 gene *A* in transformation and infection. J. Virol. 15:613-618.

56. **Tegtmeyer, P., M. Schwartz, J. K. Collins, and K. Rundell.** 1975. Regulation of tumor antigen synthesis by simian virus 40 gene *A*. J. Virol. 16:168-178.

57. **Tilghman, S. M., D. C. Tiemeier, J. G. Seidman, B. M. Peterlin, M. Sullivan, J. V. Maizel, and P. Leder.** 1978. Intervening sequence of DNA identified in the structural portion of a mouse B globin gene. Proc. Natl. Acad. Sci. U.S.A. 75:725-729.

58. **Tjian, R.** 1978. The binding site on SV40 DNA for a T antigen related protein. Cell 13:165-180.

59. **Tonegawa, S., A. M. Maxam, R. Tizard, O. Bernard, and W. Gilbert.** 1978. Sequence of a mouse germ-line gene for a variable region of an immunoglobulin light chain. Proc. Natl. Acad. Sci. U.S.A. 75:1485-1489.

60. **Weigand, R., G. N. Godson, and C. M. Radding.** 1975. Specificity of the S1 nuclease from Aspergillus oryzae. J. Biol. Chem. 250:8848-8855.

61. **Wilson, J. T., J. K. deRiel, B. G. Forget, C. A. Marotta, and S. M. Weissman.** 1977. Nucleotide sequence of 3′ untranslated portion of human alpha globin mRNA. Nucleic Acids Res. 4:2353-2368.

62. **Wilson, J. T., L. B. Wilson, J. K. deRiel, L. Villa-Komaroff, A. Efstratiadis, B. G. Forget, and S. M. Weissman.** 1978. Insertion of synthetic copies of human globin genes into bacterial plasmids. Nucleic Acids Res. 5:563-581.

63. **Zain, B. S., R. Dhar, S. M. Weissman, P. Lebowitz, and A. M. Lewis, Jr.** 1973. A preferred site for initiation of ribonucleic acid transcription by *Escherichia coli* RNA polymerase within the simian virus 40 deoxyribonucleic acid segment of adenovirus $2^{+}ND_{1}$ and adenovirus $2^{+}ND_{3}$. J. Virol. 11:682-693.