Mapping of the 5' Termini of Two Adeno-Associated Virus 2 RNAs in the Left Half of the Genome

E. W. LUSBY AND K. I. BERNIS*

*Department of Immunology and Medical Microbiology, University of Florida College of Medicine,
Gainesville, Florida 32610*

Received 15 June 1981/Accepted 10 September 1981

The left 45% of the adeno-associated virus 2 genome was sequenced. The 5' termini of two adeno-associated virus-specific RNAs were mapped at the nucleotide level within this region of the genome (nucleotides 286 to 288 and 871 to 874). Both of these 5' termini map 31 ± 2 nucleotides downstream from the start of "TATA" boxes.

Adeno-associated virus (AAV) is a defective parvovirus that requires coinfection with adenovirus for multiplication (1, 18). Although the viral genome is a linear, single-stranded DNA with a molecular weight of 1.5×10^6 (13), strands of both polarities are encapsidated with equal frequency into separate virions (4, 26, 28).

Transcription of AAV2 occurs only from the minus strand (11, 12, 29). Early studies identified stable transcripts from both the nucleus and polysomes which mapped from about 0.2 to 0.96 on the genome (10). This RNA was characterized as being capped at the 5' terminus and polyadenylated at the 3' terminus (9). Subsequently, a transcript mapping from 0.04 to 0.96 was found in the precipitate of a Hirt (17) extraction (19). More recently, detailed mapping studies of AAV transcripts have been reported. Laughlin et al. (23) identified transcripts whose 5' termini mapped at 0.05, 0.13, 0.19, and 0.40. The 3' termini all mapped at 0.96. All four species were found in both spliced and unspliced forms, although the unspliced forms of the three largest species were rare. In each spliced species the splice occurred between 0.41 and 0.49. All eight transcripts were isolated from the cytoplasm. Similar results were reported by Green and Roeder (15) except that they could not identify a transcript with a 5' terminus at 0.13 and could only identify one spliced transcript corresponding to the shortest species described by Laughlin et al. (23). Green et al. (16) were able to isolate only the spliced and possibly the unspliced forms of the shortest transcript from polysomes. The spliced form (2.3 kilobases) is the most abundant form in the cell and is believed to code for all three AAV capsid proteins (20).

Synthesis of RNA by RNA polymerase II in eucaryotic cells frequently begins 30 to 32 bases downstream from a Goldberg-Hogness (M. Goldberg, Ph.D. thesis, Stanford University,

Stanford, Calif., 1979) or "TATA" box (7). By sequencing a cDNA copy of the 5' end of the 2.3-kilobase RNA Green and Roeder were able to map this structure at a position 30 to 33 bases downstream from a TATA box at 0.385 on the AAV genome (14). Further they were able to demonstrate specific initiation of transcription *in vitro* at this site on AAV DNA. From these data they concluded that there are at least two initiation sites for AAV transcription (the site they mapped and the site at 0.06). No data with regard to possible initiation of transcription were presented on the species with a 5' terminus at 0.19 or, indeed, on the species reported by Laughlin et al. to have a 5' terminus at 0.13.

Lusby et al. (24) reported the sequence of the first 145 nucleotides comprising the inverted terminal repetition in AAV2 DNA. In this paper we report the DNA sequence from 146 to 2116, constituting the left 45% of the genome and containing the DNA sequences complementary to the 5' termini of all of the RNA species so far described for AAV2 infection. We were able to map AAV-specific RNA species with 5' termini at 0.06 and 0.19 in extracts of total cellular polyadenylated RNA but were unable to detect a species with a 5' terminus at 0.13. The 5' termini at both 0.06 and 0.19 are 31 ± 2 bases downstream from TATA boxes. We were also able to confirm the presence of the TATA box at 0.385, as reported by Green and Roeder (14).

MATERIALS AND METHODS

Cells and virus. AAV-2H (2) was grown on HeLa cells in suspension culture with adenovirus type 2 (Ad2) helper, as described previously (6).

Virus and DNA purification. Virus was purified by banding in CsCl after treatment of infected cell lysates with trypsin and deoxycholate as described (5). DNA labeled with [3 H]thymidine was purified by sedimentation through alkaline sucrose gradients, and the resulting single strands were annealed to form duplex DNA (5, 6).

Enzymes. Restriction endonucleases and exonuclease III were purchased from Bethesda Research Laboratories, Inc. (Gaithersburg, Md.). Bacterial alkaline phosphatase was purchased from Worthington Diagnostics (Freehold, N.J.), and T4 polynucleotide kinase was purchased from Boehringer Mannheim Corp. (New York, N.Y.) or Bethesda Research Laboratories. Reverse transcriptase from BAI strain A avian myeloblastosis virus was obtained from J. Beard (Life Sciences, Inc., Gulfport, Fla.).

Purification of restriction fragments and terminal labeling. AAV DNA (20 to 50 μ g) was digested with a given restriction enzyme according to the conditions specified by the supplier (Bethesda Research Laboratories). The DNA was then labeled at the 5' termini with 32 P according to Maxam and Gilbert (25), and the fragments were resolved on a polyacrylamide gel, usually 6%. Fragments detected by autoradiography were extracted by the method of Maxam and Gilbert (25), except that the extracted DNA was centrifuged for 5 min in a microfuge before ethanol precipitation. The fragments were usually strand separated on polyacrylamide gels (25) or sometimes subjected to digestion with a second restriction enzyme. Ethidium bromide staining of the DNA was avoided because it seemed to interfere with strand separation.

DNA sequencing. DNA sequencing was carried out by the method of Maxam and Gilbert (25) or by the method of Sanger et al. (32). The method of Maxam and Gilbert was carried out exactly as described, except that no effort was made to make the Eppendorf caps airtight during the heating steps.

Template for the Sanger method was prepared by treating AAV DNA with exonuclease III (34). 3 H-labeled AAV DNA (8 pmol) in 30 μ l of 0.01 M Tris (pH 7.5)–3 μ l of 10 $\times$ buffer (0.5 M Tris, 10 mM $MgCl_2$, 0.1 M mercaptoethanol [pH 8.0])–32 U of exonuclease III was incubated at 37°C until 40 to 50% of the DNA was digested (usually 4 h). The extent of digestion was monitored by removing about 0.3 μ l and spotting it on a small strip of polyethyleneimine cellulose. Chromatography was carried out with 1 N HCl as the solvent. The strip was cut in half, with the top containing digested nucleotides and the bottom containing undigested DNA. The sample was diluted with 75 μ l of 0.01 M Tris (pH 7.5)–2 μ l of 0.5 M EDTA–2 μ l of 5 M NaCl. The DNA was extracted twice with phenol and once with ether and then precipitated with ethanol.

The Sanger dideoxytriphosphate sequencing reactions were carried out according to R. Roberts (personal communication), using avian myeloblastosis virus reverse transcriptase. Template (0.5 pmol), primer (1.5 pmol), and 1 μ l of 10 $\times$ Hin buffer (0.5 M NaCl, 66 mM Tris, 66 mM $MgCl_2$, 10 mM dithiothreitol, pH 7.5) in a total volume of 7 μ l were heated at 100°C in a flame-sealed micropipette for 2 min and then incubated at 60°C for 30 to 60 min. [α - 32 P]dCTP (10 μ Ci) was dried down, and the template-primer was added along with 3 μ l of water. A 2- μ l amount of the [α - 32 P]dCTP template-primer mix was added to each of four siliconized Eppendorf tubes. A 1- μ l amount of a dNTP mix (0.5 mM dATP, 0.5 mM dGTP, 0.5 mM dTTP, 10 $\times$ Hin buffer; equal volumes of each) was then added to each tube, and 1 μ l of a different dideoxytriphosphate (obtained from P-L Biochemicals [Milwaukee, Wis.]) was added to each tube. The concentration of each dideoxytriphosphate was determined empirically.

Typical concentrations were 0.5 μ M ddCTP, 0.5 mM ddATP, 0.5 mM ddGTP, and 0.5 mM ddTTP. Reverse transcriptase (1.5 U) was then added to each tube, and incubation was carried out at 42°C for 15 min. An additional 1 μ l of 0.5 mM dCTP was then added to each tube, and the incubation was continued for 15 min. The reactions were stopped with 10 μ l of 10 mM EDTA. After lyophilization to dryness, the samples were redissolved in deionized formamide containing 0.1% xylene cyanol. The samples were electrophoresed on 8% thin gels, as described by Sanger and Coulson (31). The gels were autoradiographed with Du Pont Cronex 4 or Kodak XAR-5 X-ray film.

RNA isolation. HeLa cells were labeled with [3 H]uridine (2 mCi/liter; 50 Ci/mmol; from Schwarz/Mann) at 18 h postinfection. The cells were collected by centrifugation at 24 h postinfection and frozen at –70°C in pellets containing cells from 500 ml of culture medium. Each pellet was thawed and lysed in 40 ml of 0.15 M NaCl–0.05 M sodium acetate–0.005 M EDTA–0.5% sodium dodecyl sulfate (pH 5.2), homogenized with 1 stroke in a Dounce homogenizer, and quickly added to an equal volume of phenol. The solution was incubated in a 70°C water bath with periodic shaking until the temperature of the solution was 65°C for 15 min. The solution was fast cooled on ice and centrifuged to separate phases. The RNA was extracted with chloroform-phenol (1:1) and precipitated with ethanol. Polyadenylated RNA was isolated by chromatography on oligodeoxythymidylic acid cellulose, type 3 (Collaborative Research Inc., Waltham, Mass.; 2).

Reverse transcriptase mapping. Polyadenylated RNA was dissolved in HB buffer {0.3 M NaCl, 0.01 M PIPES [piperazine-*N,N'*-bis(2-ethanesulfonic acid)], 0.01 M EDTA, 0.1% sodium dodecyl sulfate [pH 6.7]} at a concentration of 2 to 4 mg/ml. Polyadenylated RNA (50 μ l) was incubated with about 10^6 cpm of 32 P-end-labeled primer for 2 h at 50°C. The nucleic acids were precipitated in ethanol, ethanol rinsed, and redissolved in 400 μ l of 0.04 M Tris–0.025 M NaCl–5 mM $MgCl_2$ –5 mM dithiothreitol, pH 8.0. A solution containing 10 mM of each deoxytriphosphate in a volume of 50 μ l was added along with 5 μ l (80 U) of reverse transcriptase. After incubation at 42°C for 3 h, 10 μ l of 0.5 M EDTA and 50 μ l of 1 N NaOH were added, and the solution was incubated at 37°C for 16 h. DNA was then precipitated with ethanol and redissolved in formamide–0.1% xylene cyanol. The cDNA's were resolved by electrophoresis on 8 or 12% thin gels (31).

RESULTS

The left 45% of the AAV genome was sequenced by the methods of Maxam and Gilbert (25) and Sanger et al. (32; Fig. 1). The various restriction fragments used for the sequencing studies are listed in Table 1. A major reason for deducing this sequence was to be able to map the 5' termini of the several species of AAV RNA (6, 23). From these data it might be possible to gain insight into whether the various AAV RNAs were derived by processing of a larger precursor or were the result of independent sites of initiation of transcription. The Goldberg-Hogness TATA box (Goldberg, Ph.D. thesis) has been associated with the initiation of transcrip-

5'TTGGCACTOCTCTCTCGGGCTCGCTCGCTCACTGAGGCGGGGGAACAAAGTGGCCGACGCGGGGCTTTGGCGGGGGCTCAGTGAGCGAGC 100
 GAGCGCGCAGAGAGGAGTGGCACTOCTCACTAGGGGTCTCTGAGGGGTGGAGTCTGACGTGAATTACGTCATAGGGTTAGGAGGCTCTGTATT
 AGAGGTCAOGTAGTGTTTTGGACATTTTGGACACCATGTGGTCAOCTGGGTTTAAAGCGGAGTGACAGCGAGGCTTCCATTTTGAAGCGGA
 GGTTTGAAGCGCAGCGGCATGCGGGGTTTACGAGATTGTGATTAAGGTGCGGAGCGCTTGAAGCGGACCTGCGCGGCACTTCTGACAGCTTTGT
 GAACTGGTGGCGAGAGGAATGGAGTTCGCGCAGATTCTGACATGGATCTGAATCTGATTGACAGCGCACCGCTCAOCTGGCGCAGAGCTCAG 500
 CGGACTTTCTGACGGAATGGCGCGCTGTGAGTAAGCGCGCGAGCGCGCTTTCTTTGTGCAATTTGAGAGCGCAGAGCTACTTCCACATGCACGTGC
 TCGTGGAAACGCGGGGTGAATCCATGCTTTTGGGAGCTTCTGAGTCAGATTTCGGAAGAACTGATTACAGAAATTTACGCGGATCGAGCGAC
 TTTGCGAAACTGTTGGCGTCACAAAGACGAAATGGCGCGAGCGCGGAAAGGTGGTGGATGAGTGCTACATCCCAATTACTTGTCTCCCAAA
 AOCAGCGTCACTCCAGTGGCGGTGACTAATATGAACAGTATTTAAAGCGGCTTTTGAATCTCAOAGCGTAAAGGTTGGTGGCGCAGCATCTGA
 CGCACTGTGCGCAGCGCAGGAGCAGAAAGAGAAATCAGAAATCCCAATTTCTGATGCGCGGTGATCAGATCAAAACTTCAGCAGGTACATGGAGCT 1000
 GGTGGGCTGGCTGGTGGCAAGGGGATTACCTCGGAGAAGCAGTGGATCGAGGAGCAGCGCTCATATCTCTCTCAATGGCGCTCCAACTGGCGG
 TCCCAATCAAGGCTGCTTGGCAATGGGGGAAAGATTATGAGCTGACTAAACCGCGCGGAGTCACTGGTGGCGCAGCAGCGCGTGGAGCAGATT
 CGAGCAATCGGATTATATAAATTTTGGAACTAAAGGGTACGATCCCAATATCGCGCTTCGCTCTTTCTGGGATGGCGCAGGAAAGTTGGCAAGAG
 GAACACATCTGGCTGTTTGGCGCTGCAACTACCGGGAAGCAACATCGCGAGGCGCATAGCGCAGCTGTGCGCTTCTACGGGTGGTAAACTGGAOC
 AATGAGAACTTTCCCTTCAAGCACTGTGTGACAAGATGGTATCTGGTGGGAGGAGGGAAGATGACCGCAAGGCTGTGGAGTGGCGCAAGGCATTC 1500
 TCGGAGGAAGCAAGGTGCGCGTGGACAGAAATGCAAGTCTCGGCGCAGATAGACCGGCTCGCGTATCTGACCTTCAACACCAACATGCGCGCT
 GATTGACGGGAAGTCAACGACTTTCGAACACGACGACCGGTTGCAAGACCGGATGTTCAAATTTGAAGTCAACCGCGCTCTGGATCATGACTTTGGGAG
 GTCAOAGCAGGAGTCAAGACTTTTTCGGTGGGCAAGGATCAGTGGTTGAGGTGGAGCATGAATTTCTACGTCAAAAGGCTGGAGCGCAAGAAAA
 GACCGCGCGCGAGTCAAGCATATAAGTGAAGCGCAAGCGGTGGCGGAGTCAAGTGGCGCAGCATGACGTCAGACGCGGAGCTTCGATCAACTACGC
 AGACAGGTACCAAAACAAATGTTCTGTCAGCTGGCATGAATCTGATGCTGTTTCCTGCGAGCAATGGCAGAGAAATGAATCAGAAATCAATATCTGC 2000
 TTCACTACCGCAGAGAAAGCTGTTAGAGTCTTTCCGCTGTCAGAACTCAACCGCTTTCTGTCGTCAAAAGCGTATCAGAACTGTGCTCATCTC
 ATCATATCATGGAAA

FIG. 1. DNA sequence (1 to 2116) of the left half of the AAV genome. The underlined bases represent TATA boxes, and the arrows indicate the approximate (± 1 base) 5' end of AAV RNAs.

tion of many, but not all, eucaryotic cellular and viral DNAs. The DNA sequence shown in Fig. 1 contains potential TATA boxes starting at 255 (0.06), 677 (0.144), 830 (0.177), 843 (0.179), 1214 (0.258), 1661 (0.353), and 1824 (0.388). Those at 255, 677, 830/843, and 1824 correspond closely to the reported positions of the 5' termini of different AAV RNAs. These positions have been characterized by heteroduplexing and endo- and exonuclease studies (15, 23). Additionally, Green and Roeder (14) determined the map position of the most abundant AAV RNA species whose 5' terminus starts near 0.385 at the level of the nucleotide sequence. Therefore, we concentrated our mapping efforts on locating the 5' termini of the three other less abundant RNAs. To do this we used specific restriction

fragments of AAV DNA which were annealed to total infected cell polyadenylated RNA. Hybrids so formed were used to prime the synthesis of runoff cDNA's, which, after separation by gel electrophoresis, were then sequenced by the method of Maxam and Gilbert (25) in a manner analogous to that first employed to deduce the 5'-terminal sequences of simian virus 40 RNAs (27). Because it seemed likely that the TATA boxes described above might be near the 5' termini of the RNA, restriction fragments starting approximately 45 to 60 nucleotides downstream from these sequences were selected for testing as potential primers.

The 5' end of the RNA mapping near 0.06 was determined by using the separated strands of *HpaII*-Q (325 to 379) as the primer. The resulting

TABLE 1. Restriction fragments used to sequence the left half of AAV DNA

Bases sequenced	Method ^a	Fragment	Primer
122 260	MG	<i>Hae</i> III-E	
400 305	S		<i>Hae</i> III-M
411 185	MG	<i>Hae</i> III-E	
325 389	MG	<i>Hpa</i> II-Q	
390 509	MG	<i>Hpa</i> II-M	
535 374	S		<i>Hpa</i> II-P
540 497	MG	<i>Hpa</i> II-M	
606 474	S		<i>Hpa</i> II-N
613 547	MG	<i>Hpa</i> II-P	
614 724	MG	<i>Hpa</i> II-N	
666 600	MG	<i>Taq</i> I-Y ₁	
721 782	MG	<i>Taq</i> I-Y ₂	
743 660	MG	<i>Hpa</i> II-N	
744 822	MG	<i>Hpa</i> II-J	
868 780	S		<i>Hpa</i> II-J/ <i>Hha</i> I-M ^b
961 889	MG	<i>Hpa</i> II-J	
957 811	S		<i>Hpa</i> II-F/ <i>Bam</i> -B ^c
962 1043	MG	<i>Hpa</i> II-F	
1019 945	MG	<i>Bam</i> HI-B	
1046 1141	MG	<i>Bam</i> -A	
1170 1018	S		<i>Hae</i> III-K
1332 1153	MG	<i>Hpa</i> II-F	
1430 1281	MG	<i>Hind</i> II-B	
1431 1568	MG	<i>Hind</i> II-C	
1644 1383	S		<i>Hpa</i> II-O
1649 1535	MG	RI-B	
1767 1625	MG	RI-C	
1861 1725	MG	<i>Taq</i> I-F	
1768 1938	MG	RI-C	
1887 2025	MG	<i>Taq</i> I-E	
1984 1869	MG	<i>Hpa</i> II-G	
1984 2116	MG	RI-A	

^a MG, Maxam-Gilbert method; S, Sanger dideoxy-triphosphate method.

^b *Hpa*II-J was the primary restriction fragment; *Hha*I was used for secondary cleavage. That portion of *Hpa*II-J overlapping with *Hha*I-M was used as the primer.

^c *Hpa*II-F was the primary restriction fragment; *Bam*HI was used for secondary cleavage. That portion of *Hpa*II-F overlapping with *Bam*-B was used as the primer.

cDNA was electrophoresed on a 12% acrylamide-8 M urea gel. In parallel channels the sequencing reactions for *Hpa*II-G (1651 to 1336) were run as size markers (Fig. 2). From this gel it was estimated that approximately 38 bases had been added to the 3' end of the primer to form the final runoff product. This would place the 5' terminus of this RNA (presumably the initiation site) near base 287, or 32 bases downstream from the start of the TATA box at 255. To ensure that the cDNA produced actually corresponded to an AAV RNA, the cDNA was isolated from the gel and sequenced. The sequencing gel (Fig. 3) demonstrates that the cDNA directly

corresponds to an extension of the *Hpa*II-Q primer along the AAV DNA to base 286 to 288.

A similar approach was used to map the 5' terminus of another RNA near 0.19. *Hpa*II-J (744 to 960) was labeled at the 5' termini with ³²P by means of polynucleotide kinase and then digested with *Hha*I. *Hha*I cuts *Hpa*II-J at nucleotides 852 and 890. The digestion products were strand separated on a 10% polyacrylamide gel, and the shorter labeled strand (891 to 960) was used as the primer for the reverse transcriptase reaction after hybridization to total infected cell polyadenylated RNA. Three cDNA's were observed after electrophoresis in an 8% polyacrylamide gel (Fig. 4). The most rapidly migrating cDNA (cDNA 3) was 16 bases longer than the primer, using the *Hpa*II-G sequencing reactions as a source of size markers (data not shown). This would place the 5' end of the RNA at base

FIG. 2. Mapping the 5' ends of AAV RNA with *Hpa*II-Q primer. *Hpa*II-Q was terminally labeled with polynucleotide kinase and strand separated on an 8% acrylamide gel. Both strands were annealed to polyadenylated RNA from AAV-infected cells and extended with reverse transcriptase, as described in the text. *Hpa*II-G was terminally labeled with polynucleotide kinase and cleaved with *Hind*II, and the resulting large fragment was sequenced by the Maxam and Gilbert (25) procedure. Lane 1, slow strand of *Hpa*II-Q as primer; lane 2, fast strand of *Hpa*II-Q as primer; lane 3, fast strand of *Hpa*II-Q without annealing to RNA and reverse transcriptase extension.

FIG. 3. Sequencing gel of the cDNA synthesized with *HpaII*-Q primer. The cDNA was extracted from the gel (24) and sequenced by the Maxam and Gilbert procedure (25). The three T's in parentheses were readable on the original autoradiogram. A's were determined from the AC reaction.

873, 30 bases downstream from the TATA box at base 843. When this cDNA was isolated from a preparative gel and sequenced, its sequence matched perfectly the AAV DNA sequence to at least base 871-4 (Fig. 5). Our conclusion is that this cDNA represents the 5' terminus of the AAV RNA reported to begin near 0.19 on the genome.

A second cDNA approximately 20 bases longer than the first was also noted on the analytical gel. Because there is a possible TATA box starting at 830, this cDNA was also sequenced to see whether it corresponded to another AAV RNA with a 5' terminus nearer the latter TATA box. The sequence obtained, however, did not correspond to an extension of the primer along an AAV RNA template; rather, the sequence was compatible with the primer having hair-pinned on itself to serve as a template-primer to copy its own 5' end. We suspect that the primer

doublets in Fig. 2 may be caused by the same phenomenon although we do not have sequence data in this case.

We next considered whether the longest cDNA seen in Fig. 4 could correspond to the RNA reported by Laughlin et al. (23) to map at 0.13. Sequencing of this cDNA by the Maxam-Gilbert technique demonstrated that it represented an RNA transcribed from the AAV genome, but the cDNA was too long to sequence all the way to the end. We therefore moved to using a primer further to the left on AAV DNA. Because the only potential TATA box between those at 255 (0.06) and 830 (0.177) was at 677 (0.144), we chose to use as a primer the *HpaII*-

FIG. 4. Synthesis of cDNA with the *HpaII*-J/*Hha* primer and polyadenylated RNA from AAV-Ad-infected cells.

FIG. 5. DNA sequencing gel of cDNA 3 synthesized with the *HpaII-J/Hha* primer. The primer was prepared as described in the text.

N/FnuDII fragment (718 to 742). If there were an AAV RNA with a 5' terminus downstream from the TATA box at 677, we should have seen two cDNA's. The first would have extended the primer by about 10 bases and would have corresponded to the TATA box at 677 (+31 bases), and the second would have extended the primer by about 430 bases and would correspond to the RNA with a 5' terminus near 0.06 on the genome. Only the latter species could be detected on an 8% gel (data not shown). To make certain that we were correct in our size assignment, the products of this reaction and suitable markers were glyoxal treated and run on a 6% acrylamide-8 M urea gel so that we could get a more accurate sizing of the cDNA produced. The mobility of the cDNA did indeed correspond to a runoff product of the size expected for the RNA mapping at 0.06 (data not shown). Thus, we were unable to detect an RNA corresponding to that mapping at 0.13 reported by Laughlin et al. (23). Although several faint bands could be seen between cDNA 1 and cDNA 2 in

Fig. 4, the process by which they were formed is uncertain. They probably represent premature termination of cDNA synthesis by reverse transcriptase. However, it is possible that a band(s) could represent complete cDNA copies of AAV RNA present in very low concentration, such as the one mapping at 0.13 (23).

Because of the dependence of AAV on Ad helper function for any AAV macromolecular synthesis, we were interested in any potential sequence homology between the genomes of the two viruses. Two regions of potentially significant homology were noted. The first involves the 22 nucleotides immediately to the right of the left inverted terminal repetition in AAV DNA (nucleotides 146 to 167). This sequence has a 19/22 homology with a sequence within the Ad5 inverted terminal repetition (36) starting at nucleotide 48. There are data that suggest that this region of the Ad genome may have a long range or secondary effect upon transcription (P. Sharp, personal communication; T. Shenk, personal communication). The second homologous sequence noted involved the TATA box at nucleotide 255. Starting one nucleotide upstream (254), there is a 17/19 nucleotide homology with the sequence of the Ad5 early region 1a promoter (36). This is of particular interest because transcription from this promoter seems to control subsequent Ad transcription (32). The AAV sequence around 0.06 (nucleotides 130 to 330) is shown in Fig. 6. In this region, which extends from just within the inverted terminal repetition to just downstream from the 5' terminus of the largest AAV RNA species, there are a large number of direct repeats and reverse repeats (those sequences appearing in reverse in the same strand) which are illustrated in Fig. 6. The longest reverse repeat (10 nucleotides) occurs from 142 to 151 and 186 to 195. The longest direct repeat is also 10 nucleotides long and occurs in tandem between 217 and 236. The extent of these repetitions might suggest that they play some role in the initiation of transcription in this region, but there is no direct evidence of this.

DISCUSSION

The AAV genome is a small linear DNA with a coding capacity which would be very limited if transcriptional units were arranged in the non-overlapping fashion found in cellular DNA. Three structural (21, 30) and two nonstructural (9) AAV proteins have been described. The total estimated molecular weight for these five proteins is 263,000. If all of these proteins were to be encoded in AAV DNA in a linear non-overlapping way, DNA with a molecular weight of approximately 2.4×10^6 (single stranded) would be required, i.e., almost twice the appar-

FIG. 6. AAV DNA sequence from 130 to 330. The numbered arrows indicate natural or reverse repetitions. The TATA box and RNA initiation site are also indicated.

ent coding capacity known to exist in the AAV genome. Thus, it would seem clear that, in common with other small DNA viruses, AAV has had to evolve additional mechanisms of encoding information to account for the number of known proteins. Two general models are possible to account for the known data. The first would involve the synthesis of a common precursor polypeptide which could be cleaved at various specific points to produce mature proteins. Because of the sizes of the mature proteins, the cleavage points would have to be such that the mature proteins would contain overlapping amino acid sequences. Indeed, this has been reported to be the case for the three AAV structural proteins (22), as well as the structural proteins of the autonomous parvovirus minute virus of mice (35). The two nonstructural proteins could then be envisioned as the remnants of the precursor polypeptide after cleavage to produce the two smaller structural proteins. The second model would predict individual RNA transcripts for some or all of the AAV proteins. The likelihood of this model is enhanced by the discovery of the several AAV RNA species with different 5' termini that map at specific points on the genome. For the species starting at 0.06 there is an open reading frame from the first AUG extending into the splice region, whereas for the species starting at 0.18, the open reading frame after the first AUG extends for only three amino acids; the second AUG, however, is in phase with the first AUG in the 0.06 species. It is quite possible that the five AAV proteins result from a combination of these two models. Indeed, it has recently been reported that the three capsid proteins can be translated *in vitro* from the shortest RNA species starting at 0.385 (20).

The several species of AAV RNA with different 5' termini could be either the products of the processing of a large precursor, presumably initiating at 0.06 on the genome, or the products of initiation at different sites along the genome. The results of Green and Roeder (14) strongly implied that the latter was the case for the shortest species (0.385) because of the proximity of the 5' terminus to a TATA box in the DNA and the fact that RNA polymerase II would initiate at the site in an *in vitro* system. The significance of the latter evidence might be questioned though, in light of the recent report by Benoist and Chambon (3) that a TATA box in the early region of simian virus 40 served as a site for the initiation of transcription *in vitro* but not *in vivo*. In any event the 5' termini of the species at 0.06 and 0.19 that we have mapped in this paper both are 31 ± 1 nucleotides from the starts of TATA boxes in the DNA. As was the case with the shortest transcript characterized by Green and Roeder (14), these are all *in vivo* products. Therefore, either these three transcripts are all initiated in exactly the positions predicted by the TATA boxes or cleavage of a precursor transcript fortuitously occurs at at least two of the sites. The former certainly seems more likely. Of interest are the observations that the TATA boxes at 0.06 and 0.19 are identical and that the general sequence around the TATA box at 0.06 is very similar to that observed for the TATA box preceding the initiation of the Ad early 1a gene (17/19 bases are identical; 35).

We were unable to detect a 5' terminus corresponding to the species reported by Laughlin et al. (23) to have its 5' end map at 0.13. We do not know the reason for this discrepancy. The near-

est TATA box is near 0.15, but use of a suitable primer mapping to the right of this position yielded only a cDNA corresponding to the 0.06 species. Several potential TATA boxes were observed within this region of the DNA (i.e., at positions 677, 830, 1214, and 1661) that did not correspond to observed RNA species. This result tends to support conclusions derived from studies of simian virus 40 that there are signals in addition to a TATA box required for initiation of transcription *in vivo* (3).

Finally, there were two other interesting features of the AAV DNA sequence whose significance is not yet understood. The first is that 20/22 bases immediately after the inverted terminal repetition (146 to 167) match a sequence in the middle of the Ad5 inverted terminal repetition. The second is the large number of short repeats and reverse repeats observed near the TATA box at 0.06.

In summary, we have sequenced the left 45% of the AAV genome and mapped the 5' ends of two RNA species in this region. The data are compatible with both representing sites of initiation for transcription. No evidence was found for splicing near the 5' termini of the species. We are currently completing the sequence of the AAV genome to learn the positions of potential terminator codons and signals for polyadenylation. However, it is fair to note that independent data derived from the primary sequence of the AAV proteins will also be required to complement the nucleotide sequencing data to gain a more complete understanding of the regulation of AAV gene expression.

ACKNOWLEDGMENTS

We thank W. Hauswirth and P. Laipis for useful criticism. We also thank the reviewer who offered a helpful suggestion with regard to cDNA 2 in Fig. 4.

These investigations were supported by Public Health Service research grant R01 AI16326.

LITERATURE CITED

- Atchison, R. W., B. C. Casto, and W. McD. Hammon. 1965. Adenovirus-associated defective virus particles. *Science* 194:754-756.
- Aviv, H., and P. Leder. 1972. Purification of biologically active globin messenger RNA by chromatography on oligothymidylic-acid cellulose. *Proc. Natl. Acad. Sci. U.S.A.* 69:1408-1412.
- Benoist, C., and P. Chambon. 1981. *In vivo* sequence requirements of the SV40 early promoter region. *Nature (London)* 290:304-310.
- Berns, K. I., and S. Adler. 1972. Separation of two types of adeno-associated virus particles containing complementary polynucleotide chains. *Virology* 9:394-396.
- Berns, K. I., and J. A. Rose. 1970. Evidence for a single-stranded adenovirus-associated virus genome: isolation and separation of complementary single strands. *J. Virol.* 5:693-699.
- Berns, K. I., J. Kort, K. H. Fife, E. W. Grogan, and I. Spear. 1975. Study of the fine structure of adeno-associated virus DNA with bacterial restriction endonucleases. *J. Virol.* 16:712-719.
- Breathnach, R., and P. Chambon. 1981. Organization and expression of eukaryotic protein coding nuclear split genes. *Annu. Rev. Biochem.* 50:349-384.
- Buller, R. M. L., and J. A. Rose. 1978. Characterization of adenovirus-associated virus-induced polypeptides in KB cells. *J. Virol.* 25:331-338.
- Carter, B. J. 1976. Intracellular distribution and polyadenylate content of adeno-associated virus RNA sequences. *Virology* 73:273-285.
- Carter, B. J., K. H. Fife, L. M. de la Maza, and K. I. Berns. 1976. Genome localization of adeno-associated virus RNA. *J. Virol.* 19:1044-1053.
- Carter, B. J., G. Khoury, and J. A. Rose. 1972. Adenovirus-associated virus multiplication. IX. Extent of transcription of the viral genome *in vivo*. *J. Virol.* 10:1118-1125.
- Carter, B. J., and J. A. Rose. 1972. Adenovirus-associated virus multiplication. VIII. Analysis of *in vivo* transcription induced by complete or partial helper viruses. *J. Virol.* 10:9-16.
- Gerry, H. W., T. J. Kelly, and K. I. Berns. 1973. Arrangement of nucleotide sequences in adeno-associated virus DNA. *J. Mol. Biol.* 79:207-224.
- Green, M. R., and R. G. Roeder. 1980. Definition of a novel promoter for the major adeno-associated virus mRNA. *Cell* 22:231-242.
- Green, M. R., and R. G. Roeder. 1980. Transcripts of the adeno-associated virus genome: mapping of the major RNAs. *J. Virol.* 36:79-92.
- Green, M. R., S. E. Straus, and R. G. Roeder. 1980. Transcripts of the adenovirus-associated virus genome: multiple polyadenylated RNAs including a potential primary transcript. *J. Virol.* 35:560-565.
- Hirt, B. 1967. Selective extraction of polyoma DNA from infected mouse cell cultures. *J. Mol. Biol.* 26:365-369.
- Hoggan, M. D., N. R. Blacklow, and W. P. Rowe. 1966. Studies of small DNA viruses found in various adenovirus preparations: physical, biological, and immunological characteristics. *Proc. Natl. Acad. Sci. U.S.A.* 55:1457-1471.
- Jay, F. T., L. de la Maza, and B. J. Carter. 1979. Parvovirus RNA transcripts containing sequences not present in mature mRNA: a method for isolation of putative mRNA precursor sequences. *Proc. Natl. Acad. Sci. U.S.A.* 76:625-629.
- Jay, F. T., C. A. Laughlin, and B. J. Carter. 1981. Eukaryotic translational control: adeno-associated virus protein synthesis is affected by a mutation in the adenovirus DNA binding protein. *Proc. Natl. Acad. Sci. U.S.A.* 78:2927-2931.
- Johnson, F. B., H. L. Ozer, and M. D. Hoggan. 1971. Structural proteins of adenovirus-associated virus type 3. *J. Virol.* 8:860-863.
- Johnson, F. B., D. A. Vlazny, T. A. Thomason, P. A. Taylor, and M. D. Lubeck. 1978. Adeno-associated virus polypeptides—molecular similarities, p. 411. *In* D. C. Ward and P. Tattersall (ed.), *Replication of mammalian parvoviruses*. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.
- Laughlin, C. A., H. Westphal, and B. J. Carter. 1979. Spliced adeno-associated virus RNA. *Proc. Natl. Acad. Sci. U.S.A.* 76:5567-5571.
- Lusby, E. W., K. H. Fife, and K. I. Berns. 1980. Nucleotide sequence of the inverted terminal repetition in adeno-associated virus DNA. *J. Virol.* 34:402-409.
- Maxam, A. M., and W. Gilbert. 1977. A new method of sequencing DNA. *Proc. Natl. Acad. Sci. U.S.A.* 74:560-563.
- Mayor, H. D., K. Torikaj, J. L. Meilnick, and M. Mandel. 1969. Plus and minus single stranded DNA separately encapsidated in adeno-associated satellite virions. *Science* 166:1280-1282.
- Reddy, V. B., P. K. Ghosh, P. Lebowitz, M. Platak, and S. M. Weissman. 1979. 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. *J. Virol.* **30**:279-296.
28. Rose, J. A., K. I. Berns, M. D. Hoggan, and F. J. Koczot. 1969. Evidence for a single stranded adenovirus-associated virus genome: formation of a DNA density hybrid on release of viral DNA. *Proc. Natl. Acad. Sci. U.S.A.* **64**:863-869.
29. Rose, J. A., and F. J. Koczot. 1971. Adenovirus-associated virus multiplication. VI. Base composition of the deoxyribonucleic acid strand species and strand-specific *in vivo* transcription. *J. Virol.* **8**:771-777.
30. Rose, J. A., J. V. Malzel Jr., J. K. Inman, and A. J. Shatkin. 1971. Structural proteins of adenovirus-associated viruses. *J. Virol.* **8**:766-770.
31. Sanger, F., and A. R. Coulson. 1978. The use of thin acrylamide gels for DNA sequencing. *FEBS Lett.* **87**:107-110.
32. Sanger, F., S. Nicklen, and A. R. Coulson. 1977. DNA sequencing with chain terminating inhibitors. *Proc. Natl. Acad. Sci. U.S.A.* **74**:5463-5467.
33. Shenk, T., N. Jones, W. Colby, and D. Fowlkes. 1980. Functional analysis of adenovirus-5 host-range deletion mutants defective for transformation of rat embryo cells. *Cold Spring Harbor Symp. Quant. Biol.* **44**:367-375.
34. Smith, A. J. H. 1979. The use of exonuclease III for preparing single stranded DNA for use as a template in the chain terminal or sequencing method. *Nucleic Acids Res.* **6**:831-848.
35. Tattersall, P., A. J. Shatkin, and D. C. Ward. 1977. Sequence homology between the structural polypeptides of minute virus of mice. *J. Mol. Biol.* **111**:375-394.
36. Van Ormondt, H., J. Moat, A. de Waard, and A. J. Van der Eb. 1978. The nucleotide sequence of the transforming Hpa I-E fragment of adenovirus type 5 DNA. *Gene* **4**:309-328.