JOURNAL OF VIROLOGY, Feb. 1983, p. 555–564
0022-538X/83/020555-10$02.00/0

Vol. 45, No. 2

Nucleotide Sequence and Organization of the Adeno-Associated Virus 2 Genome

ARUN SRIVASTAVA,† EDWARD W. LUSBY, AND KENNETH I. BERNS*

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

Received 29 July 1982/Accepted 27 October 1982

The complete nucleotide sequence of the adeno-associated virus 2 genome was determined. The single-stranded genome is 4,675 nucleotides in length and contains inverted terminal repeats of 145 nucleotides, the first 125 nucleotides of which form a palindromic sequence. Within the inverted terminal repetitions, there are two distinct sequences representing an inversion of 43 nucleotides that can exist on either terminus. The 5′ and 3′ termini of three major mRNA transcripts, which are present in both spliced and unspliced forms, were also mapped on the viral genome. Potential initiation and termination codons for efficient protein synthesis were identified, and genome segments were assigned that code for three major viral capsid proteins and, possibly, some as-yet-unidentified, nonstructural viral proteins.

Adeno-associated virus (AAV) is a defective parvovirus that requires coinfection with either adenovirus or herpesvirus for its growth and multiplication (1, 2, 9, 25). The AAV genome is a linear, single-stranded DNA with a molecular weight of 1.5×10^6, and DNA strands of both polarities are encapsidated in separate mature virions with equal frequency (3, 8, 21, 33, 36). AAV provides a good model system to study eucaryotic genome organization and gene expression because it is one of the smallest DNA-containing viruses. Extensive studies have been carried out on the physical characterization of AAV DNA (5, 6). The genome is flanked by inverted terminal repeats that are 145 nucleotides in length. Nucleotide sequence analysis of the inverted terminal repetition has been done (31). The first 125 nucleotides of the inverted terminal repeats are palindromic, and the significance of the palindromic termini in DNA replication has been well documented (6, 17, 39, 40).

Although DNA strands of both polarities are encapsidated into separate virions, the transcription of AAV occurs only from the minus strand (14–16, 37). A significant body of data is available concerning several mRNA transcripts of AAV, which are polyadenylated and present in both spliced and unspliced forms (12, 23, 29). The 5′ termini of the three major transcripts have recently been mapped on the AAV genome (22, 29). All of the transcripts appear to have a common 3′ terminus (29). The spliced form of the shortest mRNA, which is also the most abundant species late in infection, is believed to code for all three major viral capsid proteins (26, 29). This transcript is 2.3 kilobases long, and its promoter maps at 0.385 map unit. The transcription is initiated 31 ± 1 bases downstream from the Goldberg-Hogness box, and the RNA is spliced between map units 0.41 and 0.49 (22–24, 29). Data are scant with regard to possible nonstructural viral proteins and polypeptides and their coding regions. The existence of large transcripts which do not code for structural proteins suggests that nonstructural proteins might exist (23, 25, 26). In light of the foregoing, it was of interest to ascertain the nucleotide sequence of the entire AAV genome to gain further insight into the genome organization and gene expression of AAV. Lusby and Berns (30) have recently reported the nucleotide sequence of the left 45% of the genome (nucleotides 1 to 2,116). We report here the complete nucleotide sequence of the AAV genome, possible gene organization, and the amino acid sequences of major viral capsid proteins and of putative nonstructural viral proteins.

MATERIALS AND METHODS

Cells and viruses. AAV 2H (25) was propagated in HeLa cells in suspension culture with adenovirus type 2 (Ad2) helper as described before (36).

Virus and DNA purification. AAV was purified from Ad2 by banding in CsCl after the treatment of infected cell lysates with trypsin and deoxycholate as described before (36). AAV DNA labeled with [^{3}H]thymidine was purified by sedimentation through alkaline su-

† Present address: Department of Medicine, University of Arkansas for Medical Sciences, Little Rock, AR 72205.

crose gradients, and the resulting single strands were annealed to form duplex DNA (7, 8).

Enzymes and reagents. Restriction endonucleases and T4 polynucleotide kinase were purchased from Bethesda Research Laboratories, Inc. (Gaithersburg, Md.). Bacterial alkaline phosphatase was purchased from Worthington Diagnostics (Freehold, N.J.). Reverse transcriptase from BAI strain A avian myeloblastosis virus was obtained from J. Beard, Life Sciences, Inc., Gulfport, Fla. Chemicals for DNA sequencing were dimethylsulfate (Aldrich Chemical Co., Milwaukee, Wis.), hydrazine (Kodak, Rochester, N.Y.), and piperidine (Sigma Chemical Co., St. Louis, Mo.). [γ-^{32}P]ATP (specific activity, >2,000 Ci/mmol) was obtained from Amersham Corp. (Arlington Heights, Ill.). Oligodeoxythymidylic acid [oligo(dT)$_{\sim 12-18}$] and deoxynucleoside triphosphates were purchased from P. L. Biochemicals, Inc. (Milwaukee, Wis.).

Purification of restriction fragments. Usually 50 μg of AAV DNA was digested with a given restriction endonuclease under the conditions specified by the supplier. Restriction fragments were then subjected to bacterial alkaline phosphatase treatment at 65°C and labeled at their 5′ termini with 0.2 mCi of [γ-^{32}P]ATP and bacteriophage T4 polynucleotide kinase at 37°C. Labeled fragments were then resolved on and extracted from polyacrylamide gels (acrylamide-bisacrylamide, 40:1) as described by Maxam and Gilbert (32), except that the extracted DNA was sedimented at top speed in a microcentrifuge for 5 min to remove small acrylamide pieces before precipitation with at least 3 volumes of cold ethanol.

DNA sequencing. All nucleotide sequences were determined by the partial chemical degradation method essentially as described by Maxam and Gilbert (32) with slight modifications. Piperidine reactions with 100 μl of 1 M piperidine were carried out at 90°C for 30 min without making any efforts to make the Eppendorf tubes airtight during the heating step. The gels were autoradiographed at −70°C with Cronex-4 (Du Pont Co., Wilmington, Del.) or XAR-5 (Kodak) X-ray films.

Mapping of 3′ termini of RNA. Polyadenylated RNA was isolated from Ad2-AAV-infected HeLa cells as described before (30). A 200-μg amount of RNA dissolved in 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) was annealed to 20 μg of oligo(dT)$_{\sim 12-18}$ dissolved in 0.1 M NaCl at 45°C for 60 min, slowly cooled, and ethanol precipitated. The precipitate was washed once with cold ethanol, dried, and suspended in sterile 0.05 M Tris (pH 8.0) containing 0.025 M NaCl, 5 mM $MgCl_2$, and 5 mM dithiothreitol. A 1 mM concentration of each deoxynucleoside triphosphate, 1 μg of actinomycin D per ml, and 80 U of reverse transcriptase were added, and the reaction mixture was incubated at 41°C for 3 h. RNA was hydrolyzed with 0.2 N NaOH at 37°C for 16 h and neutralized, and single-stranded cDNA was ethanol precipitated. cDNA was washed once with cold ethanol, dried, and dissolved in 0.05 M NaCl–6 mM Tris–6 mM $MgCl_2$–1 mM dithiothreitol.

A specific primer, 16 nucleotides in length, was obtained by digesting a 5′-terminally labeled *Bgl*I fragment (4,375 to 4,632) with *Hin*fI. The fragments were strand separated on a 20% polyacrylamide gel, and the labeled 16-nucleotide-long primer was extracted from the gel, annealed to cDNA at 60°C for 60 min, cooled slowly, ethanol precipitated, washed, and suspended in 0.04 M Tris–5 mM dithiothreitol–5 mM $MgCl_2$, pH 8.0. Each deoxynucleoside triphosphate (to 1 mM) and 80 U of reverse transcriptase were added, and the reaction mixture was incubated at 41°C for 3 h. Sodium acetate was added to a 0.3 M final concentration, and the extended primer was ethanol precipitated, washed, dried, dissolved in deionized water, and sequenced as described before (35).

RESULTS

Nucleotide sequence of the AAV 2 genome. We determined the complete nucleotide sequence of AAV DNA. The sequencing strategy for nucleotides 1 through 2,116 has been presented before (30). An outline of the sequencing strategy for the remainder of the genome is shown in Fig. 1. This part of the sequence was determined exclusively by the method of Maxam and Gilbert (32), using virion DNA. For the majority of the genome, strands of both polarities were independently sequenced. All of the restriction sites used as starting points were also sequenced as internal points in overlapping fragments and could be accounted for from the known restriction maps of AAV DNA. The complete nucleotide sequence of AAV DNA is shown in Fig. 2. The sequence is 4,675 nucleotides in length and is shown with the same polarity as AAV mRNA.

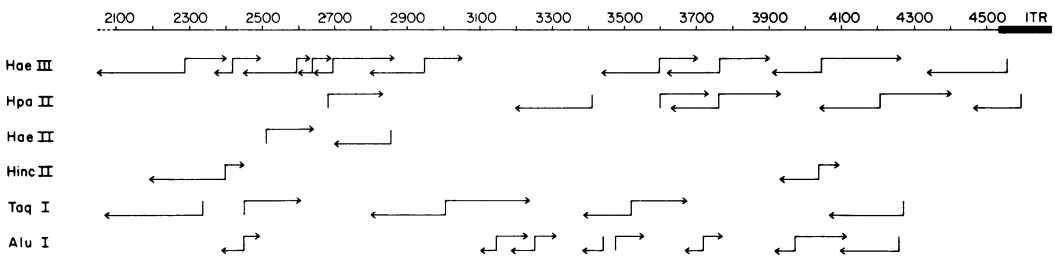

FIG. 1. Strategy for sequencing the right 55% of the AAV 2 genome. All restriction fragments were labeled at the 5′ terminus. The vertical bars represent the restriction sites, and the arrows indicate the direction and the extent of the sequence obtained. The inverted terminal repeat (ITR) is shown as a closed box.

FIG. 2. Complete nucleotide sequence of the AAV 2 genome. The nucleotide sequence of the plus strand is shown. Inverted terminal repeats, sequence heterogeneity (flip and flop orientations), palindromic stretches, TATA boxes, splice region, polyadenylation signals, 5' and 3' termini of major mRNA transcripts, and major coding regions are highlighted and discussed in the text. Figure 2 is continued on the next two pages.

A 1- to 2-base heterogeneity at either end of the DNA has been described previously. At the 5' termini, 35, 50, and 15% of the molecules contain TTGG, TGG, and GG, respectively (19). Another feature of the genome is the presence of inverted terminal repeats that are 145 nucleotides long. The first 125 nucleotides are palindromic (30) and have been shown to be implicated in AAV DNA replication in vivo (6, 17, 39, 40). In detail, there are two internal palindromes contained within a larger overall palindrome in the terminal 125 nucleotides. As a consequence of DNA replication, there is an inversion of the terminal 125 nucleotides, leading to two possible sequences at either end of the genome. The heterogeneity is only in the 43 nucleotides contained within the two small internal palindromes (31). In Fig. 2, the flip and flop orientations are highlighted in the left and right terminal repeats, respectively. There are no sites for restriction enzymes *Bgl*II, *Hpa*I, *Pvu*I, *Xba*I, and *Xor*II and only one site each for restriction enzymes

GTCACCAAGC AGGAAGTCAA AGACTTTTTC CGGTGGGCAA AGGATCACGT GGTTGAGGTG GAGCATGAAT TCTACGTCAA AAAGGGTGGA GCCAAGAAAA

Promoter 3 5' terminus of RNA 3
GACCCGCCCC CAGTGACGCA GATATAAGTG AGCCCAAACG GGTGCGCGAG TCAGTTGCGC AGCCATCGAC GTCAGACGCG GAAGCTTCGA TCAACTACGC

Splice
AGACAGGTAC CAAAACAAAT GTTCTCGTCA CGTGGGCATG AATCTGATGC TGTTTCCCTG CAGACAATGC GAGAGAATGA ATCAGAATTC AAATATCTGC 2000

TTCACTCACG GACAGAAAGA CTGTTTAGAG TGCTTTCCCG TGTCAGAATC TCAACCCGTT TCTGTCGTCA AAAAGGCGTA TCAGAAACTG TGCTACATTC

Polyadenylation signal
ATCATATCAT GGGAAAGGTG CCAGACGCTT GCACTGCCTG CGATCTGGTC AATGTGGATT TGGATGACTG CATCTTTGAA CAATAAATGA TTTAAATCAG

Splice Terminator
GTATGGCTGC CGATGGTTAT CTTCCAGATT GGCTCGAGGA CACTCTCTCT GAAGGAATAA GACAGTGGTG GAAGCTCAAA CCTGGCCCAC CACCACCAAA

GCCCGCAGAG CGGCATAAGG ACGACAGCAG GGGTCTTGTG CTTCCTGGGT ACAAGTACCT CGGACCCTTC AACGGACTCG ACAAGGGAGA GCCGGTCAAC

GAGGCAGACG CCGCGGCCCT CGAGCACGTA CAAAGCCTAC GACCGGCAGC TCGACAGCGG AGACAACCCG TACCTCAAGT ACAACCACGC CGACGCGGAG 2500

TTTCAGGAGC GCCTTAAAGA AGATACGTCT TTTGGGGGCA ACCTCGGACG AGCAGTCTTC CAGGCGAAAA AGAGGGTTCT TGAACCTCTG GGCCTGGTTG

AGGAACCTGT TAAGACGGCT CCGGGAAAAA AGAGGCCGGT AGAGCACTCT CCTGTGGAGC CAGACTCCTC CTCGGGAACC GGAAAGGCGG GCCAGCAGCC

TGCAAGAAAA AGATTGAATT TTGGTCAGAC TGGAGACGCA GACTCAGTAC CTGACCCCCA GCCTCTCGGA CAGCCACCAG CAGCCCCCTC TGGTCTGGGA

Coding region 3
ACTAATACGA TGGCTACAGG CAGTGGCGCA CCAATGGCAG ACAATAACGA GGGCGCCGAC GGAGTGGGTA ATTCCTCCGG AAATTGGCAT TGCGATTCCA

CATGGATGGG CGACAGAGTC ATCACCACCA GCACCCGAAC CTGGGCCCTG CCCACCTACA ACAACCACCT CTACAAACAA ATTTCCAGCC AATCAGGAGC 3000

CTCGAACGAC AATCACTACT TTGGCTACAG CACCCCTTGG GGGTATTTTG ACTTCAACAG ATTCCACTGC CACTTTTCAC CACGTGACTG GCAAAGACTC

ATCAACAACA ACTGGGGATT CCGACCCAAG AGACTCAACT TCAAGCTCTT TAACATTCAA GTCAAAGAGG TCACGCAGAA TGACGGTACG ACGACGATTG

FIG. 2. *Continued.*

*Acc*I, *Bam*HI, *Bcl*I, *Bst*EII, *Eca*I, *Hin*dIII, *Mst*I, *Sac*I, *Sal*I, *Sst*I, and *Xho*II.

Transcription of the viral genome. Although a significant amount of data on AAV-specific RNA transcripts is available, a knowledge of the complete nucleotide sequence of the genome has enabled us to gain further insight into AAV transcription. Details of the RNA transcripts have been reported earlier (14, 22–24, 29, 30). Three major viral transcripts have been detected in infected cells in both spliced and unspliced forms, and all are polyadenylated (12, 23, 29). Initiation sites have previously been mapped for all three transcripts at nucleotides 287 and 873 by Lusby and Berns (30) and at 1,853 by Green and Roeder (22), and it was noted that the sequence around the leftward-most promoter (175 to 320) contained a large number of direct and reverse repeats. In this same region, a 17/19 base homology was noted with the Ad5 E1a promoter (30). Green and Roeder (22) have mapped the splice region which spans nucleotides 1,907 to 2,227, a total of 320 nucleotides. It is interesting to note that the genome contains two potential polyadenylation signals, AATAAA, occurring at nucleotide positions 2,182 and 4,420. The internal polyadenylation signal, however, lies within the splice region and may not be functional. It has been reported that all major AAV transcripts share common 3′ termini mapping at position 0.96 on the genome (29). We carried out the precise mapping of the 3′ termini of the transcripts as described above (35). A 12% sequencing gel depicting the nucleotide sequence ladder corresponding to the sequence between nucleotides 4,377 and 4,447 is shown in Fig. 3. This sequence places the common 3′ termini of the AAV transcripts at nucleo-

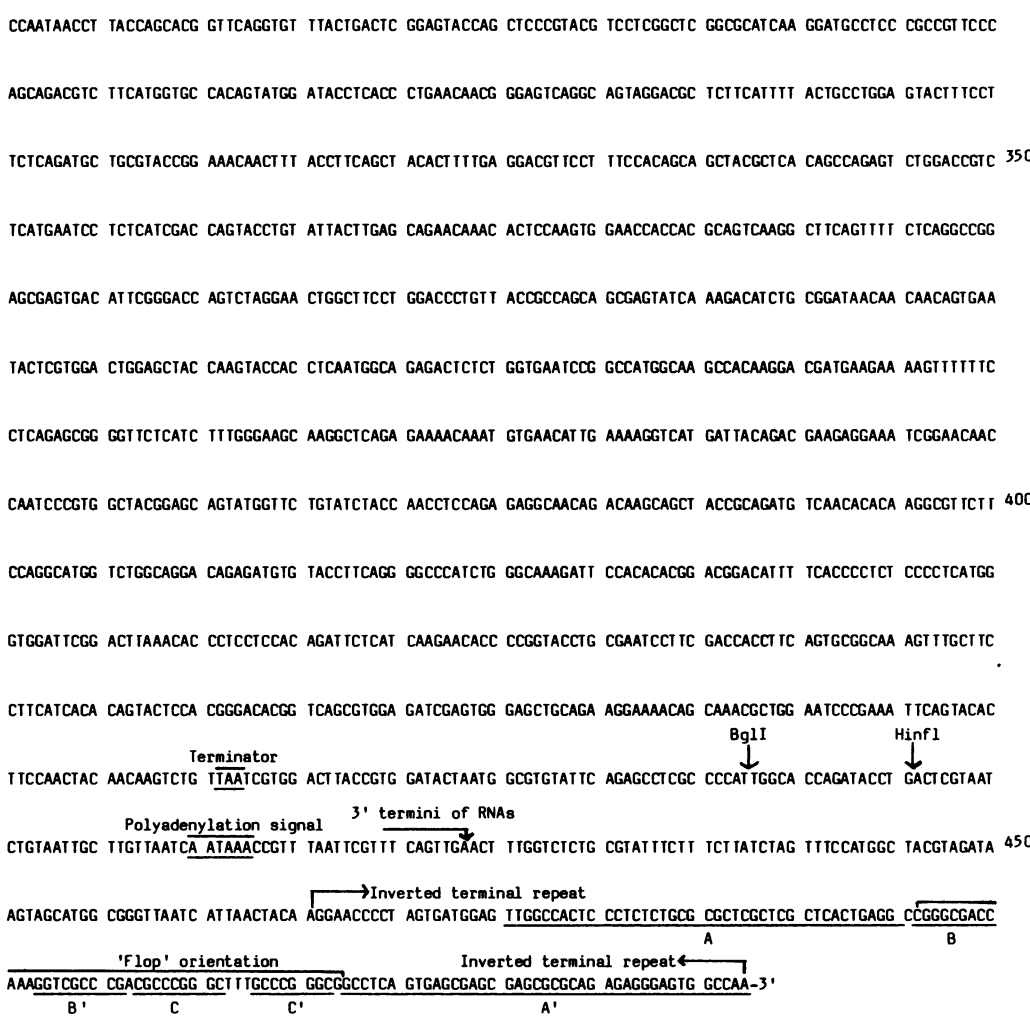

FIG. 2. *Continued.*

tide 4,447 on the genome, 22 nucleotides downstream from the polyadenylation signal, consistent with other eucaryotic transcriptional termini (20). It is also interesting to note that the first TATA box and the functional polyadenylation signal AATAAA are both approximately 250 nucleotides into the genome from either end, which is in agreement with the data of others that virtually the entire AAV genome is transcriptionally active.

Open reading frames and viral gene products. Open reading frames in the DNA sequence were determined by computer analysis. Figure 4 represents such an analysis and a schematic representation of the distribution of initiator triplets (ATG) and terminator triplets (TGA, TAA) across the AAV genome, as well as the spliced and unspliced forms of the three transcripts. As is evident from Fig. 4, both spliced and unspliced transcripts contain at least three major open reading frames, two of which lie in the left half of the genome and share a common reading frame. The third, in the right half of the genome, overlaps with a fourth smaller one, but different reading frames are used.

Although no viral proteins have yet been identified that are coded for by the transcripts originating at 0.06 and 0.19 map unit positions, the spliced forms of these transcripts can code for two proteins with approximate molecular weights of 67,000 and 40,000. Their unspliced counterparts, however, can code for two larger proteins with approximate molecular weights of 77,000 and 50,000, respectively. These two open reading frames are in phase and can code for proteins with a common carboxy terminus. Numerous shorter open reading frames exist, but none would code for a polypeptide of molecular

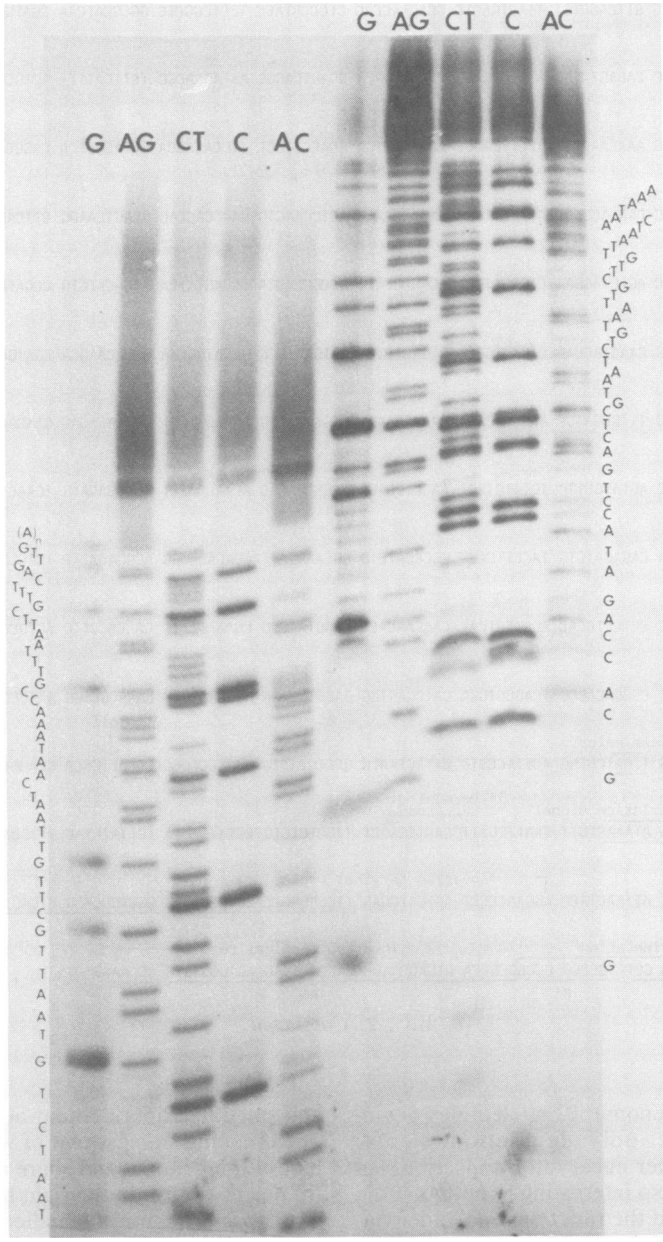

FIG. 3. Mapping of the 3′ termini of AAV 2 transcripts. An autoradiograph of a 12% sequencing gel, depicting the common 3′ termini of the AAV mRNA transcripts, is shown. The indicated sequence is identical to that presented in Fig. 2 between nucleotides 4,377 and 4,447.

weight >7,000. It has been reported that the spliced form of the shortest transcript is the most abundant species late in infection and the only one associated with polysomes at that time (24). Independent data suggest that the shortest transcript codes for all three coat proteins (molecular weights, 85,000, 72,000, and 61,000) (26). In the spliced form of this transcript, there is a leader sequence of approximately 640 nucleotides before the first AUG triplet. The succeeding open reading frame codes for a protein with an approximate molecular weight of 63,000 only. The unspliced form of the same transcript, however, contains an AUG triplet at a position 69 nucleotides downstream from the 5′ terminus of the mRNA, but the reading frame starting at this point could potentially code for a protein of approximately 10,000 daltons only. A larger,

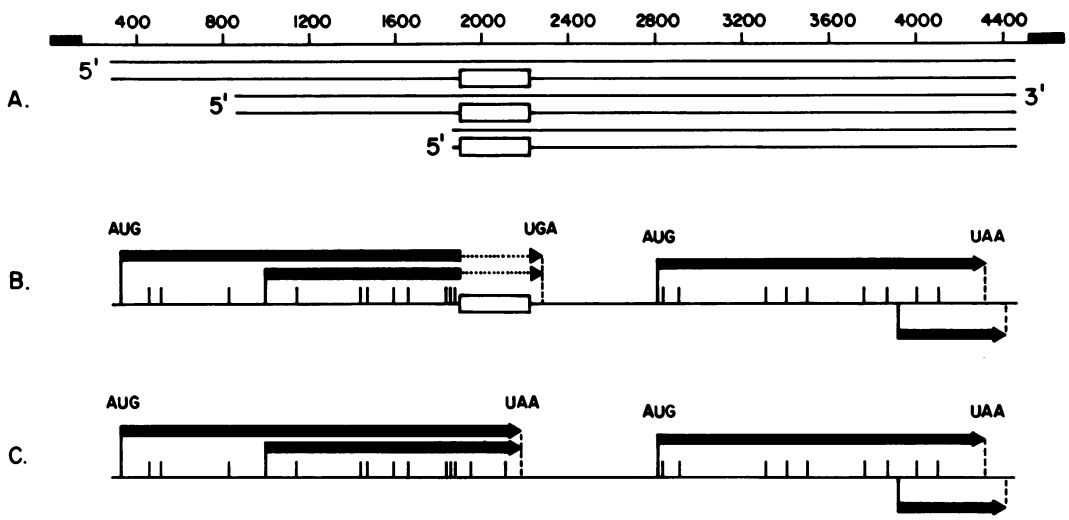

FIG. 4. Organization of the genome for major viral structural and nonstructural proteins of AAV. (A) Transcriptional map of the AAV 2 genome. Spliced and unspliced forms of all three major transcripts are shown. The splice region is indicated as an open box. (B) Major open reading frames in spliced transcripts. (C) Major open reading frames in unspliced transcripts. The solid arrows indicate the extent of open reading frames, the vertical bars represent overlapping ATG triplets, and the broken vertical lines represent the terminator triplets.

additional open reading frame originating with a nonoverlapping initiator AUG triplet at nucleotide 3,924 could also code for a protein of approximately 20,000 daltons. Buller and Rose (10) have detected a nonstructural viral protein of similar molecular weight in infected cells. There are at least four other nonoverlapping open reading frames in the right half of the genome, but none can code for polypeptides of >6,000 daltons.

Finally, the analysis was extended to determine the frequency of codon usage by the putative gene for the viral capsid proteins. Such an analysis is shown in Table 1. It is noteworthy that CGN triplet codons are used at a much lower frequency, an observation consistent with other known eucaryotic genes (27, 41). Data shown in Table 2 compare the amino acid composition of the viral capsid proteins derived from the nucleotide sequence and that determined

TABLE 1. Codon usage of genes for viral structural proteins

First nucleotide in codon	Second nucleotide in codon				Third nucleotide in codon
	U	C	A	G	
U	Phe 15	Ser 9	Tyr 4	Cys 2	U
	Phe 12	Ser 7	Tyr 18	Cys 4	C
	Leu 0	Ser 8	Term.[a] 1	Term. 0	A
	Leu 1	Ser 4	Term. 0	Trp 10	G
C	Leu 6	Pro 10	His 3	Arg 4	U
	Leu 15	Pro 7	His 13	Arg 2	C
	Leu 1	Pro 8	Gln 12	Arg 5	A
	Leu 8	Pro 5	Gln 23	Arg 1	G
A	Ile 8	Thr 4	Asn 13	Ser 7	U
	Ile 8	Thr 21	Asn 25	Ser 11	C
	Ile 0	Thr 8	Lys 6	Arg 9	A
	Met 10	Thr 9	Lys 10	Arg 2	G
G	Val 4	Ala 7	Asp 6	Gly 5	U
	Val 9	Ala 6	Asp 19	Gly 11	C
	Val 5	Ala 10	Glu 9	Gly 20	A
	Val 8	Ala 5	Glu 7	Gly 5	G

[a] Term., Terminator.

TABLE 2. Amino acid composition of AAV 2 capsid proteins

Amino acid	% Total residues	
	Observed[a]	Calculated
Phenylalanine	4.8	5.4
Leucine	6.2	6.2
Isoleucine	3.3	5.2
Methionine	1.7	1.9
Valine	5.3	5.2
Serine	8.3	9.1
Proline	6.6	5.9
Threonine	8.6	8.3
Alanine	4.5	5.6
Tyrosine	4.6	4.4
Histidine	2.1	3.2
Lysine	3.8	3.2
Cysteine	0.3	1.2
Glycine	9.9	8.1
Arginine	4.8	4.6
Glutamine + glutamic acid	10.6	10.1
Aspargine + aspartic acid	14.7	12.5
Tryptophan	—[b]	1.9

[a] Values were taken from Rose et al. (38).
[b] —, Not done.

experimentally by Rose et al. (38). The similar values obtained indicate the likelihood that the viral structural proteins are indeed coded for by the mRNA transcripts that map on the right half of the AAV genome.

DISCUSSION

Mammalian DNA viruses have provided powerful model systems for studies on eucaryotic genome organization and gene expression. By virtue of its small size, the AAV genome is readily amenable to the type of detailed study described in this paper. We determined the complete nucleotide sequence of the AAV genome to define the intricacies involved in viral replication, transcription, and gene expression at the molecular level.

The inverted terminal repeats in the single-stranded viral genome are interesting in several respects. They have been demonstrated to be at junctions between viral and cellular sequences in AAV latently infected cells (18) and seem to be important in the rescue of AAV DNA from the integrated state (Samulski et al., unpublished data). These data suggest that the AAV genome may function as an insertional element, all of which contains terminally inverted repeats (11). Additionally, in its most common form, the genome is flanked by TG and CA, which is also of interest because all known eucaryotic insertional elements have the same dinucleotides at their termini. The relationship between the arrangement of terminal sequences in AAV DNA and the process of AAV DNA replication is well established (6, 39, 40).

The process of AAV genome transcription has been described in relatively great detail. Studies on in vivo transcription indicate that only the negative strand of AAV DNA is transcribed by the host cell RNA polymerase II (13). Whether there is regulation of the relative amounts of the three major RNA transcripts that are derived from overlapping regions representing approximately 92% of the viral genome has not been rigorously determined. Although the shortest transcript is the most abundant species late in a productive infection (26), this is not the case when cells containing functional Ad E1a and b are infected by AAV alone. Under these conditions, the longer species are more abundant (M. A. Labow and K. I. Berns, unpublished data). Little information is available on AAV-specific protein synthesis. Three major capsid proteins (approximate molecular weights, 85,000, 72,000, and 61,000) and two nonstructural proteins (approximate molecular weights, 25,000 and 16,000) have been observed in infected cells (10). All three of the capsid proteins are believed to be derived from the smallest RNA transcript, initiating at 0.385 map unit. This presents an apparent paradox, because the open reading frame can potentially code for a protein with a maximum molecular weight of 63,000. Several explanations are possible. (i) The sequence is incorrect, and there is an earlier ATG in phase. We think this is unlikely, because Green and Roeder (22) sequenced much of this control region and have also failed to find an ATG triplet. (ii) A triplet other than ATG (or GTG) is used for initiation; this would be unique. (iii) The molecular weight determined for the proteins may be incorrect because of post-translational modifications or unusual conformations, which might alter the gel mobilities. (iv) The putative TAA stop codon might be suppressed in some instances. However, this alone would be insufficient to account for the size discrepancy unless transcripts were copied off multimeric or circular forms of the genome. (v) There could be unsuspected processing of either RNA transcripts or proteins. Indeed, F. W. Studier (personal communication) has found that the translation of the mRNA of bacteriophage T7 gene 10 involves a phase change leading to the synthesis of two coat proteins of different sizes with overlapping amino acid sequences. Data to support any of the above hypotheses have yet to emerge. In addition, the reason for the presence of a rather large leader sequence upstream of the first AUG triplet codon in the spliced form of the smallest transcript is still unclear, although a similar observation has been made in an immediate-early gene transcript (IE mRNA-5) of herpes simplex virus type 1 (42).

The large open reading frames on the left half of the genome suggest the existence of nonstructural proteins. Although no proteins known to be coded for by the two large transcripts mapping on the left half of the genome have yet been detected in infected cells, at least two nonstructural viral proteins could conceivably be derived from these transcripts. Putative functions for such nonstructural proteins could include a role in DNA replication or in the integration process in latent infection. A similar situation exists in the gene expression of the autonomous parvovirus, minute virus of mice, wherein one of the nonstructural proteins that is found associated with the 5′ ends of the viral DNA is believed to be coded for by the transcript mapping in the left half of the genome (D. Ward, personal communication). McPherson and Rose (Abstr. Annu. Meet. Am. Soc. Microbiol. 1982, S74, p. 247) have presented some evidence for an approximately 72,000-molecular-weight protein, which was seen after AAV-human Ad coinfection of African green monkey kidney cells. Under these conditions, AAV coat protein synthesis is suppressed. One possible function for a nonstructural protein might be in DNA replication. Laughlin et al. (28) have suggested a possible "rep" function for AAV, based on their studies with AAV defective interfering particles. Another possibility is that a nonstructural protein might be involved in AAV latent infection. The notion of an early protein is supported by the observation of Ostrove and Berns (34) of AAV RNA transcripts in the absence of normal AAV DNA replication.

The organization of the AAV genome parallels that of the autonomous parvovirus genome in several major aspects. (i) Both have palindromic terminal sequences, and in both the 3′ terminus is believed to serve as a primer for DNA replication. (ii) The 5′-terminal sequences in both are found in two orientations (5). (iii) In both, there are several overlapping transcripts and three major open reading frames. (iv) In both, all three coat proteins seem to be coded for by the open reading frame in the right half of the genome.

In summary, we obtained the complete nucleotide sequence of the AAV genome, mapped all the major mRNA transcripts on the genome, and assigned segments of the genome to code for putative nonstructural viral proteins and the major viral capsid proteins. Further work involving the identification of other viral proteins is under way.

ACKNOWLEDGMENTS

We thank W. W. Hauswirth, N. Muzyczka, B. J. Flanegan, and R. J. Samulski for useful criticism. The expert technical assistance of Theresa Korhnak and computer analysis by Wei Chang are gratefully acknowledged. We also thank Patrice Boyd, Sandy Ostrofsky, and Grace Pedersen for preparing the manuscript.

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

LITERATURE CITED

1. **Atchison, R. W.** 1970. The role of herpesviruses in adeno-associated virus replication *in vitro*. Virology **42**:155–162.
2. **Atchison, R. W., B. C. Casto, and W. McD. Hammon.** 1965. Adenovirus-associated defective virus particles. Science **194**:754–756.
3. **Berns, K. I., and S. Adler.** 1972. Separation of two types of adeno-associated virus particles containing complementary polynucleotide chains. Virology **9**:394–396.
4. **Berns, K. I., A. Cheung, J. Ostrove, and M. Lewis.** 1982. Adeno-associated virus latent infection, p. 249–265. *In* B. W. J. Mahy, A. C. Minson, and G. K. Darby (ed.), Virus persistence. Cambridge University Press, New York.
5. **Berns, K. I., and W. W. Hauswirth.** 1979. Adeno-associated viruses. Adv. Virus Res. **25**:407–449.
6. **Berns, K. I., W. W. Hauswirth, K. H. Fife, and E. W. Lusby.** 1979. Adeno-associated virus DNA replication. Cold Spring Harbor Symp. Quant. Biol. **43**:781–787.
7. **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.
8. **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.
9. **Buller, R. M., E. Janik, E. D. Sebring, and J. A. Rose.** 1981. Herpes simplex virus types 1 and 2 completely help adenovirus-associated virus replication. J. Virol. **40**:241–247.
10. **Buller, R. M., and J. A. Rose.** 1978. Characterization of adeno-associated virus polypeptides synthesized *in vivo* and *in vitro*. p. 399–410. *In* D. C. Ward and P. Tattersall (ed.), Replication of mammalian parvoviruses. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.
11. **Calos, M. P., and J. H. Miller.** 1980. Transposable elements. Cell **20**:579–595.
12. **Carter, B. J.** 1975. Intracellular distribution and polyadenylate content of adeno-associated virus RNA sequences. Virology **73**:273–285.
13. **Carter, B. J.** 1978. Parvovirus transcription, p. 33–52. *In* D. C. Ward and P. Tattersall (ed.), Replication of mammalian parvoviruses. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.
14. **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.
15. **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.
16. **Carter, B. J., and J. A. Rose.** 1972. Adenovirus-associated virus multiplication. VIII. Analysis of in vivo transcription induced by complete and partial helper virus. J. Virol. **10**:9–16.
17. **Cavalier-Smith, T.** 1974. Palindromic base sequences and replication of eukaryotic chromosome ends. Nature (London) **250**:467–470.
18. **Cheung, A., M. D. Hoggan, W. W. Hauswirth, and K. I. Berns.** 1980. Integration of the adeno-associated virus genome into cellular DNA in latently infected Detroit 6 cells. J. Virol. **33**:739–748.
19. **Fife, K. H., K. Murray, and K. I. Berns.** 1977. Structure and nucleotide sequence of the terminal regions of adeno-associated virus DNA. Virology **78**:475–487.
20. **Fitzgerald, M., and T. Shenk.** 1981. The sequence AAUAAA forms part of the recognition site for polyade-

nylation of late SV40 messenger RNAs. Cell 24:251–260.
21. **Gerry, H. W., T. J. Kelly, and K. I. Berns.** 1973. Arrangement of nucleotide sequence in adeno-associated virus DNA. J. Mol. Biol. 79:207–224.
22. **Green, M. R., and R. G. Roeder.** 1980. Definition of a novel promotor for the major adeno-associated virus mRNA. Cell 22:231–242.
23. **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.
24. **Green, M. R., S. E. Straus, and R. G. Roeder.** 1980. Transcripts of the adeno-associated virus genome: multiple polyadenylated RNAs including a potential primary transcript. J. Virol. 35:560–565.
25. **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.
26. **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.
27. **Kaptein, J. S., and D. P. Nayak.** 1982. Complete nucleotide sequence of the polymerase 3 gene of human influenza virus A/WSN/33. J. Virol. 42:55–63.
28. **Laughlin, C. A., M. W. Myers, D. L. Risin, and B. J. Carter.** 1979. Defective interfering particles of the human parvovirus adeno-associated virus. Virology 94:162–174.
29. **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.
30. **Lusby, E. W., and K. I. Berns.** 1982. Mapping of the 5′ termini of two adeno-associated virus 2 RNAs in the left half of the genome. J. Virol. 41:518–526.
31. **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.
32. **Maxam, A., and W. Gilbert.** 1980. Sequencing end labeled DNA with base specific chemical cleavage. Methods Enzymol. 65:499–560.
33. **Mayor, H. D., K. Torikai, J. L. Melnick, and M. Mandel.** 1969. Plus and minus single-stranded DNA separately encapsidated in adeno-associated satellite virions. Science 166:1280–1282.
34. **Ostrove, J., and K. I. Berns.** 1980. Adenovirus early region 1b gene function required for rescue of latent adeno-associated virus. Virology 104:502–505.
35. **Reddy, V. B., P. K. Ghosh, P. Lebowitz, M. Piatak, and S. M. Weissman.** 1979. Simian virus 40 early mRNAs. 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.
36. **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.
37. **Rose, J. A., and F. J. Koczot.** 1971. Adeno-associated virus multiplication. VI. Base composition of the DNA species and strand-specific in vivo transcription. J. Virol. 8:771–777.
38. **Rose, J. A., J. V. Maizel, J. K. Inman, and A. J. Shatkin.** 1971. Structural proteins of adenovirus-associated viruses. J. Virol. 8:766–770.
39. **Samulski, R. J., K. I. Berns, M. Tan, and N. Muzyczka.** 1982. Cloning of adeno-associated virus into pBR322: rescue of intact virus from the recombinant plasmid in human cells. Proc. Natl. Acad. Sci. U.S.A. 79:2077–2081.
40. **Straus, S. E., E. D. Sebring, and J. A. Rose.** 1976. Concatemers of alternating plus and minus strands are intermediates in adenovirus-associated virus DNA synthesis. Proc. Natl. Acad. Sci. U.S.A. 73:742–746.
41. **Swartz, M. N., T. A. Trautner, and A. Kornberg.** 1962. Enzymatic synthesis of deoxyribonucleic acid: further studies on nearest neighbor base sequences in deoxyribonucleic acid. J. Biol. Chem. 237:1961–1967.
42. **Watson, R. J., and G. F. Van de Woude.** 1982. DNA sequence of an immediate early gene (IE mRNA-5) of herpes simplex virus type I. Nucleic Acids Res. 10:979–991.