Nucleotide sequence analysis of a region of adenovirus 5 DNA encoding a hitherto unidentified gene

W.Kruijer, F.M.A.van Schaik and J.S.Sussenbach

Laboratory for Physiological Chemistry, State University of Utrecht, Utrecht, The Netherlands

Received 23 October 1980

ABSTRACT

The nucleotide sequence of a region located on the adenovirus type 5 genome between coordinates 59.9 and 62.5 has been determined. This region comprises the 3'-termini of mRNAs coding for the hexon and the DNA binding protein. The sequence reveals the presence of an open reading frame in r-strand transcripts downstream from the hexon termination codon. The protein predicted from this region has a MW of 23,068 daltons. Based on the characteristics of the mutant Ad2ts1, which maps between coordinates 59.9 and 69.9, it is probable that the predicted protein is involved in processing of late proteins.

INTRODUCTION

The genome of the homologous adenovirus types 2 and 5 consists of a linear double-stranded DNA molecule with a MW of 23×10^6 daltons(1,2) and has a complex organization of early and late genes transcribed from both complementary strands (3-5). At some regions of the adenovirus genome early and late genes adjoin each other and it might be expected that these regions are of regulatory importance.

We have investigated the nucleotide sequence and the genetic organization of the region between the genes encoding the 72K DNA binding protein and the hexon gene of adenovirus type 5 (Fig. 1).

The gene for the DNA binding protein is transcribed early in infection from the l-strand. Heteroduplex studies with the corresponding Ad2 mRNA have shown that the 3'-termini of these mRNA molecules are located at position 61.6 (6-8).

The hexon gene is transcribed late in infection from the r-strand as part of a long precursor molecule. The transcription of this precursor is controlled by the major late promoter at coordinate 16.3 (9). Hexon mRNA as well as other late messengers are generated from this transcript by splicing (10, 11). The 3'-termini of hexon mRNA molecules have been mapped at coordinate

Figure 1 A map of the region on the Ad5 genome between coordinates 59.5 and 63.5. The cleavage sites for BamH I (59.5) and Bgl II (60.2 and 63.5) are indicated. Hexon and DNA binding protein mRNAs are transcribed from opposite directions and their 3'-termini adjoin at coordinate 61.6. r and l stands for rightward and leftward transcribed strands, respectively.

61.6 (12).

Thus, the 3'-termini of mRNAs of the hexon gene and of the DNA binding protein gene have both been mapped at coordinate 61.6, be it on opposite strands of the adenovirus genome. Akusjärvi and Pettersson have determined the nucleotide sequence at the carboxyl-terminal end of the gene coding for the Ad2 hexon (13). This sequence is located within a BamH I - Bgl II restriction fragment between coordinates 59.5 and 60.2. From the sequence can be deduced that the termination codon of the hexon gene is located at position 59.9, which is 1.7 map units removed from the position where the 3'-terminus of hexon mRNA has been mapped. This indicates that hexon mRNA contains a non-coding region of about 595 nucleotides long downstream from the termination codon.

In this report we present the nucleotide sequence of a region of Ad5 DNA located between coordinates 59.9 and 62.5. This sequence starts downstream from the termination codon of the hexon gene and terminates at a position within the carboxy-terminal end of the gene coding for the DNA binding protein. The sequence reveals the existence of a hitherto unidentified gene located between the genes coding for hexon and the DNA binding protein.

MATERIAL AND METHODS

Materials

Restriction endonucleases Alu I, Hae III, Mbo I, Rsa I, Hha I, were purchased from New England Biolabs (Beverly) and BamH I, Bgl II as well as T₄

polynucleotide kinase were from Boehringer (Mannheim). Bacterial alkaline phosphatase (BAP-F) was from Worthington (Freehold). γ - ^{32}P -ATP (specific activity > 2000 Ci/mmol) was purchased from the Radiochemical Centre (Amersham). Dimethylsulfate (98%) was obtained from Merck-Schuchardt (München), hydrazine (> 95%) from Eastman-Kodak (Rochester), formic acid (98-100%) from Merck (Darmstadt) and piperidine (> 99%) from Merck (Darmstadt). Acrylamide and bisacrylamide (purified for electrophoresis) were from BDH (Poole).

Methods

Adenovirus type 5 (Ad5) was isolated from infected HeLa cells and viral DNA was isolated from purified virions as described previously (14).

Ad5 DNA was digested with Bgl II and the resulting fragments were separated by electrophoresis on a 3% polyacrylamide (PAA) gel. Fragment L was excised from the gel and isolated as described before (15). The purified L fragment was digested with various restriction enzymes and from the resulting fragments the 5'-terminal phosphate groups were removed by digestion with BAP-F in 50 mM Tris-HCl pH 8.0, at 37°C or 55°C. Subsequently the dephosphorylated 5'-ends were phosphorylated using γ - ^{32}P -ATP and T₄ polynucleotide kinase (16).

The separation of both ^{32}P -labeled ends of each fragment was achieved in either of two ways: 1) by enzymatic cleavage of the fragments with a second restriction enzyme and electrophoresis of the cleavage products on a 5 or 6% PAA gel, 2) by denaturation of the fragments in 30% DMSO and electrophoresis of the strands on a 5% low ratio bisacrylamide/acrylamide gel (16).

Then, the 5'-endlabeled fragments were subjected to chemical degradation according to Maxam and Gilbert (16). Specific cleavage at guanine residues was achieved by methylation with dimethylsulfate, at adenine residues by ring opening with alkali, at purine residues by partial depurination at pH 2.0, at pyrimidine residues by hydrazinolysis and at cytosine residues by hydrazinolysis in 1.0 M NaCl, in all cases followed by strand scission with piperidine. The cleavage products were separated by electrophoresis on thin (40 x 30 x 0.045 cm) 8 and 20% PAA gels with 8.3 M urea (16, 17). Gels were subjected to autoradiography, using Fuji medical X-Ray films at -20°C or, when necessary, at -70°C with Ilford Fast Tungstate intensifying screens.

RESULTS

In order to obtain restriction fragments suitable for sequence determination, a physical map of the Bgl II L fragment (coordinates 60.2-63.5) was

constructed for a number of restriction enzymes using the method developed by Smith and Birnstiel (18). The position of various restriction sites was confirmed in the determined sequence and, in addition, new cleavage sites were found that could be used to generate other fragments suitable for sequence analysis.

All sequences were determined according to the method of Maxam and Gilbert (16) as indicated in Materials and Methods. Fig. 2 shows the regions of the r- and l-strands which have been sequenced. The complete primary structure of both complementary strands was determined, except at positions 690-700 and 866-876 on the r-strand, where the sequence could be read unambiguously for the l-strand only.

The complete nucleotide sequence determined is shown in Fig. 3. Several regulatory signals are present in the sequence between the recognition sequences for BamH I (59.5) and Bgl II (63.5) and a schematic representation of these signals is summarized in Fig. 4.

The sequence between residues 613 and 685 has a remarkable symmetry and due to self-complementary sequences in the r- and l-strands around the axis of symmetry between residue 650-651, a hairpin can be drawn as shown in Fig. 5. Fig. 6 shows a sequencing gel from this region.

Figure 2 Diagram of analyzed DNA fragments located between cleavage sites for BamH I (59.5) and Bgl II (63.5). Vertical bars represent restriction endonucleases cleavage sites and arrows indicate the direction and length of the restriction fragments, which were used to determine the r-(lower part) and l- (upper part) strand sequences after 5'-endlabeling.

1
 3' -GACGGCGGTACCCGAGGTCACCTCGTCTTGTACTTTTCGTAACAGTTTCTAGAACCAACCCCGGTATAAAAAC
 5' -CTGCCGCATGGGCTCCAGTGAGCAGGAACTGAAAGCATTGTCAAAGATCTTGGTTGGGGCCATATTTTTTG
 MetGlySerSerGluGlnGluLeuLysAlaIleValLysAspLeuGlyCysGlyProTyrPheLeu
 100
 CCGTGGATACTGTTTCGCGAAAGGTCGAAACAAAGAGGTGTGTTTCGAGCGGACGCGGTATCAGTTATGCCGGCCTA
 GGCACCTATGACAAGCGCTTTCCAGGCTTTGTTTCTCCACACAAGCTCGCCTGCGCCATAGTCAATACGGCCGGT
 GlyThrTyrAspLysArgPheProGlyPheValSerProHisLysLeuAlaCysAlaIleValAsnThrAlaGly
 150
 GCGCTCTGACCCCGCATGTGACCTACCGGAAACGGACCTTGGGCGTGAAGTTTTGTACGATGGAGAACTCGGG
 CGCGAGACTGGGGCGTACACTGGATGGCCTTTGCCTGGAACCCGCACTCAAAACATGCTACCTCTTTGAGCCC
 ArgGluThrGlyGlyValHisTrpMetAlaPheAlaTrpAsnProHisSerLysThrCysTyrLeuPheGluPro
 200
 AAACCGAAAGACTGGTCTGAGTTCTGCTCAAATGGTCAAACCTCATGCTCAGTGAGGACGCGCATCGCGGTAA
 TTTGGCTTTTCTGACCAGCGACTCAAGCAGGTTTACAGTTTGTAGTACGAGTCACTCCTGCGCCGTAGCGCCATT
 PheGlyPheSerAspGlnArgLeuLysGlnValTyrGlnPheGluTyrGluSerLeuLeuArgSerAlaIle
 250
 CGAAGAAGGGGCTGGGCAATATTCGACCTTTTCAGTGGGTTTCGATGTCCCCGGGTTGAGCGCGCGGACA
 GCTTCTTCCCCCGACCGCTGTATAACGCTGGAAAGTCCACCAAGCGTACAGGGGCCAACTCGGCCGCTGT
 AlaSerSerProAspArgCysIleThrLeuGluLysSerThrGlnSerValGlnGlyProAsnSerAlaAlaCys
 300
 CCTGATAAGACGACGTACAAAGAGGTGCGGAAACGGTTGACCGGGTTTGAGGGTACCTAGTGTGGGGTGGTAC
 GGACTATTCTGCTGCATGTTTCTCCAGCCTTTGCCAATGGCCCCAACTCCCATGGATCACAACCCACCATTG
 GlyLeuPheCysCysMetPheLeuHisAlaPheAlaAsnTrpProGlnThrProMetAspHisAsnProThrMet
 350
 TTGGAATAATGGCCCATGGGTTGAGGTACGAGTTGTCAAGGGTCCATGTCCGGTGGGACGACGCTTGGTCTCT
 AACCTTATTACGGGGTACCAACTCCATGCTCAACAGTCCCAGGTACAGCCACCTGCGTCGCAACCAGGAA
 AsnLeuIleThrGlyValProAsnSerMetLeuAsnSerProGlnValGlnProThrLeuArgArgAsnGlnGlu
 400
 GTCGAGATGTCGAAGGACCTCGCGGTGAGCGGGATGAAGCGCTGGTGTACCGCTCTAATCCTCGCGGTGAAGA
 CAGCTCTACAGTTCCTGGAGCGCCACTCGCCCTACTTCCGACGCCACAGTGCAGATTAGGAGCGCCACTTCT
 GlnLeuTyrSerPheLeuGluArgHisSerProTyrPheArgSerHisSerAlaGlnIleArgSerAlaThrSer
 450
 AAAACAGTGAACCTTTTGTACATTTTATACATGATCTCTGTGAAAGTTATTTCCGTTTACGAAAATAAACAT
 TTTTGTCACTTGAATAATGTAAATAATGTACTAGAGACACTTCAATAAAGGCAATGCTTTTATTGTGA
 PheCysHisLeuLysAsnMet
 500
 GTGAGAGCCCACTAATAATGGGGTGGGAACGGCAGACGCGGCAAAATTTTGTATTTCCCAAGACGGCGGTAG
 CACTCTCGGGTATTATTTACCCCCACCCTTGCCTCTGCGCCGTTTAAAAATCAAAGGGGTTCTGCCGCGCATC
 550
 SerHisAlaValProLeuSerValAsnArgTyrGlnHisLysThrSerTrpLysPheGluProValValMetArg
 CGATACGCGGTGACCGTCCCTGTGCAACGCTATGACCACAAATCACGAGGTGAATTTGAGTCCGTGTTGGTAGGC
 GCTATGCGGCACTGCGAGGACACGTTGCGATACTGGTGTTAGTGCTCCACTTAACTCAGGCACAACCATCCG
 600
 ProLeuGluThrPheAsnGluSerTrpLeuSerArgValMetValLeuAlaAsnLeuLeuAspProAlaSerIle
 GCCGTCGAGCCACTTCAAAAGTGAGGTGTCGACGCGTGGTAGTGGTTCGCAAAATCGTCCAGCCCGCGGCTATA
 CGGCAGCTCGGTGAAGTTTCACTCCACAGGCTGCGCACCATCACCAACGCGTTTACGAGGTGGGCGCGGATAT
 650
 LysPheAspCysAsnPro
 GAACCTCAGCGTCAACCCCG-5' r-strand
 CTTGAAGTCGAGTTGG -3' l-strand
 700
 750
 800
 850

Figure 3 The determined nucleotide sequence. The sequence starts with residue -8 at the position where Alu I cuts the BamH I (59.5) - Bgl II (60.2) fragment and continues to the position where Sau96 I cuts the Bgl II L fragment. The amino acids for the 23K and the DNA binding protein which can be derived from the DNA sequence are also indicated.

Figure 4 Schematic representation of regulatory sequences in the mRNAs transcribed from the r- and l-strands as deduced from the sequence. The coding parts of these mRNAs are illustrated by the horizontal bars.

DISCUSSION

Organization of the r-strand

The position of the hexon termination codon has been located exactly at coordinate 59.9 on the Ad2 genome (13). The AATAAA sequence, which is found in most polyA-containing mRNAs 14-30 nucleotides upstream from the polyA tail (19, 20), is not found close to this termination codon. Instead such a sequence is found 679 nucleotides downstream from this termination codon between

Figure 5 Possible secondary structure in the region between residues 613-685. Inverted sequences around the axis of symmetry between residue 650-651 are indicated in boxes.

Figure 6 Autoradiogram of a 20% PAA sequencing gel showing the sequence of the region between residues 660-625 in the r-strand.

residues 640 and 645 in the l-strand, which has the same polarity as mRNAs transcribed from the r-strand. This suggests that the 3'-end of hexon mRNA is located between residues 654-670, about 700 nucleotides downstream from the hexon stopcodon. This position corresponds with coordinate 61.9 on the Ad5 genome and is in agreement with the results from heteroduplex studies with hexon mRNA which located the 3'-end of this mRNA at coordinate 61.6 (12).

Between the termination codon of the hexon gene and the sequence AATAAA⁶⁴⁰ an open reading frame is found in the l-strand, starting with ATG¹ and terminating with TAA⁶¹³. This suggests that a hitherto unknown gene is located 36 nucleotides downstream from the termination codon of the hexon gene. The predicted gene comprises 612 nucleotides and codes for a protein of 204 amino acids with a molecular weight of 23,068 daltons. The amino acid composition of this 23K protein is shown in Table 1.

Between map positions 60.0 and 61.7 no separate class of mRNAs with a coding capacity for a protein of 204 amino acids has been described in the

Table 1.

AMINO ACID COMPOSITION OF THE 23K PROTEIN

AMINO ACID	NUMBER	MOLE%	AMINO ACID	NUMBER	MOLE%
<i>Ala</i>	14	6.9	<i>Leu</i>	17	8.5
<i>Arg</i>	11	5.4	<i>Lys</i>	8	3.9
<i>Asn</i>	10	4.9	<i>Met</i>	7	3.4
<i>Asp</i>	5	2.5	<i>Phe</i>	14	6.9
<i>Cys</i>	8	3.9	<i>Pro</i>	14	6.9
<i>Gln</i>	12	5.9	<i>Ser</i>	20	9.8
<i>Glu</i>	9	4.4	<i>Thr</i>	11	5.4
<i>Gly</i>	12	5.9	<i>Trp</i>	3	1.5
<i>His</i>	8	3.9	<i>Tyr</i>	7	3.4
<i>Ile</i>	6	2.9	<i>Val</i>	8	3.9

literature. However, there is evidence that an unidentified gene is located between the genes for hexon and the DNA binding protein: Firstly, an mRNA species has been isolated after infection which hybridized with adenoviral sequences between coordinates 59.7 and 61.8 (Chow and Broker, personal communication). The 5'-terminus of this mRNA contained the tripartite leader sequence and the 3'-terminus mapped at the same position as the hexon mRNA viz. at coordinate 61.6. This short mRNA was present in the infected cell in very low quantities. This mRNA covers the open reading frame between ATG¹ and TAA⁶¹³ and it is attractive to assume that ATG¹ is used to start translation. Secondly, the temperature-sensitive mutant Ad2ts1 has been mapped between coordinates 59.0 and 69.0 using Ad2/Ad5 interserotypic heteroduplex mapping (21). DNA replication proceeds normally at the restrictive temperature and this mutant is efficiently complemented by hexon and 100K mutants, indicating that the mutation does not effect the genes coding for the DNA binding protein, the hexon and 100K protein, respectively. This implies that Ad2ts1 is affected in an unidentified gene, located between coordinates 59.0 and 69.0 (22). A very attractive candidate for this gene is the above mentioned 23K protein located between the hexon and DNA binding protein gene.

Mutant Ad2ts1 is defective in the processing of precursor polypeptides pVI, pVII, pVIII and as a result of this defect non-infectious virions accumulate (23, 24). If the Ad2ts1 gene and the 23K gene are identical it is likely that this gene encodes an endopeptidase. This is in agreement

with the observation of Bhatti and Weber (25) that wild type young virions contain an adenovirus associated protease specific for cleavage of core polypeptide pVII to polypeptide VII, while this protease is absent in virions produced at 39°C in Ad2ts1 infected cells.

Organization of the l-strand

The sequence of both strands between residues 613-711 encodes no amino acids. The codon TAA⁶¹³ in the l-strand shuts off, as outlined before, the gene for the 23K protein while the sequence TAA⁷¹³ terminates an open reading frame in the r-strand. The sequence of the r-strand has the same polarity as l-strand transcripts. The only known l-strand transcripts from the region between coordinates 59.9 and 62.5 encode the DNA binding protein. Therefore, the open reading frame preceeding TAA⁷¹³ in the r-strand must specify the C-terminal amino acids of the DNA binding protein.

The sequence AATAAA is present in the r-strand between residues 656-661. Assuming that polyA addition takes place 14-30 nucleotides downstream from the AATAAA sequence the 3'-termini of mRNAs of the DNA binding protein gene are located at coordinate 61.8 which is in good agreement with results derived from electronmicroscopic and S1 nuclease studies (7, 8).

Border region between r- and l-strand transcripts

Since the AATAAA sequences in r- and l-strands are only 10 nucleotides apart, it might be expected that the 3'-ends of the mRNAs for hexon and 23K protein and the mRNA for the DNA binding protein partially overlap. In this regard it is noteworthy that the region between residues 613 and 685, which partially encodes the 3'-non-translated regions of r- and l-strands transcripts, possesses remarkable symmetry due to inverted repeated sequences (Fig. 5). The axis of symmetry is located between residues 650-651 and due to self-complementary sequences a hairpin can be drawn in which both AATAAA sequences are involved. It remains to be established whether the secondary structure at the 3'-ends of transcripts originating from the region between coordinates 59.9 and 62.5 has any biological significance.

The genes for proteins IX and IV_{a2} constitute another example of genes transcribed from opposite strands (26). However, in this case the corresponding AATAAA sequences in the r-strand and l-strand are separated by 45 nucleotides and no extensive secondary structure is found.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge stimulating discussions with drs H.S. Jansz, P.Ch. van der Vliet and thank dr. P.H. Steenbergh for introduction of W.K. and F.M.A. van S. in sequence analysis.

This work was supported in part by the Netherlands Foundation for Chemical Research (SON) with financial aid from the Netherlands Organisation for the Advancement of Pure Research.

REFERENCES

1. Van der Eb, A.J. and Van Kesteren, L.W. (1966). *Biochim. Biophys. Acta* **129**, 441-444.
2. Green, M., Pina, M., Kimes, R., Wensink, P.C., Machattie, L.A. and Thomas, C.A. (1967). *Proc. Natl. Acad. Sci USA* **57**, 1302.
3. Philipson, L. (1979). *Adv. Virus Res.* **25**, 347-405.
4. Flint, S.J. (1977) *Cell* **10**, 153-166.
5. Flint, S.J., Berget, S.M. and Sharp, P.A. (1976) *Virology* **72**, 443-455.
6. Kitchingman, G.R., Lai, S-Pg. and Westphal, H. (1977). *Proc. Natl. Acad. Sci. USA* **74**, 4392-4395.
7. Berk, A.J. and Sharp, P.A. (1978) *Cell* **14**, 695-711.
8. Chow, L.T., Broker, T.R. and Lewis, J.B. (1979) *J. Mol. Biol.* **134**, 265-303.
9. Evens, R.M., Fraser, N., Ziff, E., Weber, J., Wilson, M. and Darnell, J.E. (1977). *Cell* **12**, 733-739.
10. Akusjärvi, G. and Pettersson, U. (1979). *J. Mol. Biol.* **134**, 143-158.
11. Nevins, J.R. and Darnell, J.E. (1978) *Cell* **15**, 1477-1493.
12. Chow, L.T. and Broker, T.R. (1978) *Cell* **15**, 497-510.
13. Akusjärvi, G. and Pettersson, U. (1978) *Virology* **91**, 477-480.
14. Sussenbach, J.S. (1971) *Virology* **46**, 969-972.
15. Steenbergh, P.H., Maat, J., Van Ormondt, H. and Sussenbach, J.S. *Nucl. Acids Res.* **4**, 4371-4389.
16. Maxam, A.M. and Gilbert, W. (1989) In: *Methods in Enzymology*, Colowick, S.P. and Kaplan, N.O. eds. Vol. 69 pp 499-559, Academic Press, New York.
17. Sanger, F. and Coulson, A.R. (1978) *FEBS letters* **87**, 107-110.
18. Smith, H.O. and Birnstiel, M.L. (1976). *Nucl. Acids Res.* **3**, 2387-2398.
19. Proudfoot, N.J. and Brownlee, G.G. (1976). *Nature* **263**, 211-214.
20. Benoist, C., O'Hare, K., Breathnach, R. and Chambon, P. (1980). *Nucl. Acids Res.* **8**, 127-142.
21. Hassell, J.A. and Weber, J. (1978). *J. Virol.* **28**, 671-678.
22. Weber, J. and Hassell, J.A. (1979). *J. Gen. Virol.* **44**, 639-655.
23. Weber, J. (1976). *J. Virol.* **17**, 462-471.
24. Mirza, M.M.A. and Weber, J. (1977). *Virology* **80**, 83-97.
25. Bhatti, A.R. and Weber, J. (1979). *Virology* **96**, 478-485.
26. Aleström, P., Akusjärvi, G., Perricaudet, M., Mathews, M.B., Klessig, D.F. and Pettersson, U. (1980). *Cell* **19**, 671-681.