The Nucleotide Sequence of the Leftmost \textit{XhoI} Fragment (6\%) of Simian Adenovirus SA7P

By B. M. M. DEKKER, D. A. M. KONINGS, T. S. DENISOVA,\textsuperscript{1} R. A. GIBADULIN\textsuperscript{1} AND H. VAN ORMONDT*  

Department of Medical Biochemistry, Sylvius Laboratories, University of Leiden, Wassenaarseweg 72, 2333 AL Leiden, The Netherlands and \textsuperscript{1}D.I. Ivanovsky Institute of Virology, Academy of Medical Sciences U.S.S.R., Moscow, U.S.S.R.

(Accepted 13 July 1984)

SUMMARY

The DNA of simian adenovirus SA7P was cloned in pBR322. The nucleotide sequences of the leftmost 2238 bp and the rightmost 188 bp of the viral genome were determined. SA7P DNA has an inverted terminal repeat of 183 bp. The sequence at the left terminus exhibits extensive homology with that of the E1 regions of human adenovirus 5, 7 and 12 DNAs. Based on this homology, the RNA coordinates and coding regions could be deduced. The sequenced SA7P DNA contains the entire E1A and part of the E1B region.

INTRODUCTION

Virus SA7P is a simian adenovirus (Ad) highly oncogenic in hamsters (Gibadulin et al., 1982). This virus, originally characterized as SV20, was later found to be an antigenic variant of SA7 (Denisova & Gibadulin, 1982). Previous studies on transformation \textit{in vitro} with SA7P DNA fragments have shown that the transforming region of this virus, like that of human adenoviruses, is located within the left-terminal sequence (18\%) of the genome (Gibadulin et al., 1982).

The organization of the transforming region of human adenoviruses has been extensively studied (Tooze, 1981). The genes responsible for cell transformation are located in the leftmost 11\% of the linear viral genome, which is expressed in the early stage of productive infection (region E1). This region has been shown to consist of two separate transcription units: E1A (0 to 4.5\%) and E1B (4.6 to 11\%), each of which codes for a family of 5'- and 3'-coterminal mRNAs. These families differ among each other in the amount of information removed by RNA splicing. The expression of both E1 subunits seems to be required for complete transformation and oncogenicity: an E1A product is necessary for the expression of other early regions, e.g. E1B (Nevins, 1981; Bos & Ten Wolde-Kraamwinkel, 1983), and has also been shown to switch on certain cellular genes (Nevins, 1982). In cells transformed with Ad DNA fragments, the serotype origin of the fragment used in transfection determines their oncogenicity in immunocompetent hosts: Ad12 E1A suppresses the synthesis of class I major histocompatibility proteins, while Ad5 E1A does not (Bernards et al., 1983a; Schrier et al., 1983). The origin of E1B determines the growth potential \textit{in vivo} of transformed cells (Bernards et al., 1983b).

In the course of our studies on the transforming and oncogenic properties of SA7P, we have determined the nucleotide sequence of the leftmost \textit{XhoI} fragment of its DNA. We compared this sequence with those of the E1 regions of human Ad5, Ad7 and Ad12. On the basis of sequence homology, we could conclude that the SA7P sequence described here contains the entire E1A and part of the E1B region, and that the organization of the transforming region of the simian Ad SA7P is very similar to that of human adenoviruses.

METHODS

SA7P DNA was tailed with dCTP by terminal deoxynucleotidyl transferase and digested with EcoRI: the internal fragments were cloned in pBR322 linearized by EcoRI, and the terminal fragments in PstI-cut, dG-tailed...
**RESULTS AND DISCUSSION**

The overlapping tracts whose sequences were determined within the leftmost *XhoI* fragment (bp 1 to 2238) are schematically represented in Fig. 1. Both strands were sequenced virtually completely as is evident from this figure. The nucleotide sequence determined is shown in Fig. 2(a). We have given the sequence of the l strand, since this has the same polarity as the mRNAs transcribed from this region. In addition, RNA coordinates (5' and 3' termini, TATA boxes, splice sites) and initiation and stop codons, which could be deduced (see below) on the basis of sequence homology with human Ad DNAs, are indicated.

Fig. 2(b) gives the sequence of the left and right termini of the SA7P as well as the inverted terminal repeat (ITR) of simian Ad SA7 as far as it was determined by Tolun *et al.* (1979). It appears that the length of the SA7P ITR is 183 bp. Our sequence around the *EcoRI* site at position 565 is in disagreement with the one published previously (Denisova *et al.*, 1982). By sequencing the fragment (*ClaI–XhoI*, positions 544 to 2333) which overlaps the *EcoRI* site we found that the sequence downstream from the *EcoRI* site was totally different from the one published by Denisova *et al.* (1982).

The nucleotide sequence described in this paper covers the leftmost 2238 nucleotides of the SA7P genome. As will be shown, this encompasses the ITR, the complete E1A region, the entire coding sequence for a 21K E1B protein and the codons for the amino-terminal portion of a proposed E1B 55K polypeptide. The SA7P ITR is the longest of the Ad ITR sequences yet published (Van Ormondt & Galibert, 1984). Within this ITR, positions 9 to 50 coincide exactly with the consensus sequence noted by Stillman *et al.* (1982) for Ad ITRs. Another conserved tract in the 5' non-transcribed region is the one proposed to be required for encapsidation of viral DNA (positions 306 to 332; Brinckmann *et al.*, 1983).
Simian adenovirus SA7P region E1A sequence

E1A RNA coordinates

We assigned the RNA coordinates of the fragment on the basis of their homology with the corresponding sequences of the human adenoviruses. These sites are indicated in the nucleotide sequence (Fig. 2a) and in the bottom part of Fig. 5. The 5' cap site of the E1A mRNA can be deduced from the alignment with the corresponding regions of Ad5, Ad7 and Ad12 (shown in Fig. 3), from the E1A TATA box to the presumed E1A initiation codon. The strong conservation and the distance to the TATA box (30 bp) makes it highly probable that the cap site is located at the A in position 440 (A440). At position A1445 we found the unique sequence AATAAA usually associated with the 3' poly(A) tail of eukaryotic RNAs. By comparison with the human Ad DNAs, we expect the 3' poly(A) addition site to be encoded between nucleotides 1465 and 1472.

The E1A region of human adenoviruses encodes three 5'- and 3'-coterminal mRNAs (termed '9S', '12S' and '13S'). The strong conservation of the regions flanking the splice sites in the human Ad E1A RNAs (Fig. 4) permits us to suggest the possible donor and acceptor candidates for analogous SA7P mature mRNAs. As is shown in the alignment, the potential common splice acceptor (position G1176) is preceded by an A/T-rich region and followed by a stretch with high homology. The splice donor for the '13S' mRNA can also be easily deduced on the basis of its consensus intron sequence and its relative position. The analogous '12S' and '9S' splice donors cannot be deduced unambiguously from the nucleotide alignment. For the '12S' mRNA, two potential donor residues are T964 (aligning with the Ad12 and Ad7 donors) and T973, neither of them optimal in their neighbouring intron sequence. Further upstream, nucleotides G838 and G758 are followed by the near-consensus tracts GTGAGC and GTGAGA, respectively; splicing at the first point would result in an mRNA only slightly shorter than the Ad5 '12S' species. There is also more than one candidate donor residue for the '9S' mRNA: T576 which aligns with the '9S' splice site of Ad12, and G588 which is adjacent to an optimal 5' intron sequence, GTGAGT.

E1B RNA coordinates

Downstream from the postulated AATAAAA sequence signalling the 3' poly(A) addition site for the E1A mRNA, we located a TATA box at position 1509. From this we infer that, in analogy with the human Ad DNAs, the 5' cap site for the SA7P E1B mRNA is located at position 1538. The splice donor used late in infection in all three human Ad E1B mRNAs proved to be less than 30 nucleotides downstream from the stop codon for the 21K E1B protein product (Van Ormondt & Hesper, 1983). Since there is no strong sequence homology in this region, we presume that in SA7P E1B RNA the analogous splice donor is at position G2119 which is the only potential splice site within 100 nucleotides downstream from the 21K stop codon.

Coding capacity

To interpret the nucleotide sequence in terms of its coding capacity, we have arranged the nonsense codons present in the SA7P DNA fragment according to the three reading frames (Fig. 5). The putative E1A region is shown to contain three stretches open for translation, one in each frame. The 5'-proximal AUG in the E1A mRNA (position 487) is in reading frame 1 (nucleotides 445 to 1063). The alignment in Fig. 3 shows that the 5' untranslated leader is shorter than those of Ad5, Ad7 and Ad12. The open tract in frame 1 is separated from the other two by an A/T-rich region containing nonsense codons in two phases. As a result of RNA splicing, open reading frame 1 can be linked to either of the other E1A open frames, the position of the predicted splice sites (see above) determining which of them is used in the translation of the 3' exon.

This organization allows the coding of at least three E1A polypeptides, in analogy with the human adenoviruses. The amino acid sequence specified by the '13S' mRNA terminates at TAG 1403 and would have a molecular weight of 28983. It is shown (Fig. 6) in a comparison with the corresponding human Ad E1A proteins. This comparison is based on an alignment of the encoding nucleotides (not presented) and of functionally related amino acids, with the help
Fig. 2. (a) The primary structure of SA7P DNA between the left terminus and the XhoI site at position 2333. The sequence given is that of the l strand. The following features are indicated: the end of the inverted terminal repeat (ITR), the conserved encapsidation sequence (306 to 322; underlined), two TATA boxes, a 3'-terminal polyadenylation signal AATAAA, suspected coordinates and signals for initiation and termination of polypeptide synthesis. The latter were deduced from sequence homology with human Ad DNAs. Above the DNA sequence, the amino acid sequences of the predicted proteins are given in one-letter code (see legend to Fig. 6).
Fig. 2 continued

<table>
<thead>
<tr>
<th></th>
<th>CATCATCAAT AATATACCTT ATTTGGGAAC GGTGCCAATA TGCTAATGAG GTGGGCGGAG SA7P</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>---ATCAAT AATATACCTT ATTTGGGAAC GGTGCCAATA TGCTAATGAG GTGGGCGGAG SA7</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>TTTGGTGACG TATGCCGGAAG TGGGCGGAGC AAGGGGCGGG GCGAGAGGCG GGGCTTTTGG SA7P</th>
</tr>
</thead>
<tbody>
<tr>
<td>61</td>
<td>TTTGGTGACG TATGCCGGAAG TGGGCGGAGC AAGGGGCGGG GCGAGAGGCG GGGCTTTTGG SA7P</td>
</tr>
<tr>
<td></td>
<td>*   **  *   ***  *  ****</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>TAGGCGTGGC CGGCCGTGGG AACGGAAGTG ACGTCGGGGG CGCGCCGGAC GTGACGTGTT SA7P</th>
</tr>
</thead>
<tbody>
<tr>
<td>121</td>
<td>TAGGCGTGGC CGGCCGTGGG AACGGAAGTG ACGTCGGGGG CGCGCCGGAC GTGACGTGTT SA7P</td>
</tr>
</tbody>
</table>

Fig. 2. (b) The inverted terminal repeat of SA7P DNA. The sequences of the 183 5'-terminal residues of the l and r strands (left and right termini, respectively) are identical and given as one row of capitals. The diverging nucleotides beyond the end of the ITR are printed in lower case. The ITR of SA7P DNA is compared with that of simian Ad SA7, as far as was determined by Tolun et al. (1979). Differences between the two species are indicated by asterisks. The consensus sequence found by Stillman et al. (1982) to occur in all the human Ad ITRs determined is underlined.

of the program TRIALS of Hogeweg & Hesper (1984). The alignment reveals that the SA7P E1A nucleotide and amino acid sequences show the same pattern of divergence and convergence as that expected by the human Ad E1A regions (Van Ormondt et al., 1980). Across the whole region, SA7P and Ad12 are the most closely related. The central domain, i.e. the amino acids encoded by the region between the '13S' and '12S' splice donors in human Ad E1A RNA (in SA7P: Cys-163 to Ser-191) is highly conserved, which is also a feature of the recently published E1A sequence of Tupaia Ad DNA (Brinckmann et al., 1983). In addition, our sequence exhibits a second main region of homology which is located in the amino-terminal part of the protein and extends from Ser-40 to Gly-75. Interestingly, the alignment reveals that in the heterologous region preceding the central domain the SA7P sequence shows a striking resemblance to the Ad12 sequence in having a string of alanine residues. This may be significant in that both adenoviruses are highly oncogenic and that this region is located within the exon that is specific for oncogenesis (J. L. Bos, personal communication).

The proposed E1B region has two open reading frames (Fig. 5). One of these allows the synthesis of a 21279 mol. wt. protein initiating at the 5'-proximal AUG (position 1569) of the predicted E1B mRNA. Comparison of the deduced SA7P E1B polypeptide with the corresponding human Ad proteins revealed a similar homology pattern as that shown by these proteins among themselves (Van Ormondt & Hesper, 1983); there was no resemblance to any one of them in particular (not shown). The second AUG codon downstream from the E1B cap site (position 1874) is preceded by a stretch of nucleotides homologous to those flanking the E1B 55K start codon of the human Ad analogues. As far as the determined sequence up to the XhoI site goes, this AUG codon should determine the start of a polypeptide resembling the human Ad E1B 55K protein.

The above discussion makes it clear that simian adenovirus SA7P is closely related to the human adenoviruses, at least in the transforming region of its DNA. Not only is the organization quite similar, but also the gene products show considerable homology. In contrast, the corresponding region of prosimian (Tupaia) adenovirus (Brinckmann et al., 1983) is notably less homologous, especially in the carboxy terminus of the E1A protein and in the interval between E1A and E1B.
Fig. 3. Comparison of SA7P DNA residues 409 to 490 with the 5'-terminal region of human Ad E1A RNA. Positions occupied by identical nucleotides in human Ad5, Ad7 and Ad12 DNAs are boxed. The asterisks denote positions occupied in SA7P DNA by residues identical to the boxed ones. TATA boxes, RNA 5' termini and initiation codons are indicated.
Fig. 4. Splice sites of E1A '13S' RNA. Nucleotides 1044 to 1194 of SA7P DNA compared with the splice donor and acceptor regions for the E1A '13S' mRNAs of human Ad12, Ad7 and Ad5. Homologies in the human Ad DNAs are boxed; the asterisks denote positions occupied in SA7P DNA by residues identical to the boxed ones.

Fig. 5. Open reading frames and deduced RNA coordinates. The nonsense codons (TAA, TGA and TAG; indicated as vertical marks) in the left terminus of the 1 strand of SA7P DNA arranged according to their reading frames. In the longer open frames the presumed initiation triplets (ATG) are indicated. The coding regions are represented as solid bars. In the bottom part, the RNA coordinates as deduced by sequence homology with human Ad DNAs are given: the continuous lines represent exons, the connecting interrupted carets introns. The scale at the top indicates nucleotide numbers.
Fig. 6. Comparison of the SA7P E1A '13s' RNA translation product with those of human Ad12, Ad7 and Ad5. The dashes in the sequences denote gaps introduced in order to align them. Positions occupied by identical residues in the three human Ad proteins are boxed. The asterisks indicate SA7P–Ad12 homologies. The amino acid sequences are given in the one-letter code: A, Ala; C, Cys; D, Asp; E, Glu; F, Phe; G, Gly; H, His; I, Ile; K, Lys; L, Leu; M, Met; N, Asn; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp; Y, Tyr.
B. M. M. Dekker and Others

This work was supported by the Netherlands Organization for the Advancement of Pure Research (ZWO) through the Foundation of Chemical Research in the Netherlands (SON).

REFERENCES


(Received 11 June 1984)