Volume 8 Number 10 1980 Nucleic Acids Research

Nucleotide sequence of the EcoRI D fragment of adenovirus 2 genome

Jacques Hérissé, Geneviève Courtois and Francis Galibert

Laboratoire d'Hématologie Expérimentale, Centre Hayem, Hôpital Saint-Louis, Paris 10è, France

Received 17 March 1980

SUMMARY

The entire nucleotide sequence of the Ad. 2 EcoRI D fragment has been determined using the Maxam and Gilbert method. This sequence of 2678 bp contains informations relative to late mRNAs ending at position 78 and for which an AATAAA sequence corresponding to their 3' ends is found at residue number 833. Position of the PVIII mRNA is determined thus allowing deduction of the probable amino acid sequence of the PVIII protein. The position and the sequence of the first leader of early 3 mRNAs is determined as well as the sequence and position of the second early leader of region 3 mRNAs, which also correspond to the "y" leader of the fiber mRNA. Following the localization of an open reading frame in which an ATG could initiate protein synthesis it can be predicted that 3a, b, c mRNAs code for the 16K early protein and the probable amino acid sequence of this protein can be deduced. The CAGTTT sequence frequently present at the 5' end of a leader or of a mRNA body as well as the GGTGAG sequence which is found at the 3' end of several leaders were used to postulate the position of various early mRNAs of region 3 and to suggest the existence of an additional splicing event during the processing of mRNAs 3a, b and c. They were also used to predict the position of the additional "x" late leader.

The imbrication of informations concerning (i) the family of late mRNAs ending at position 78, (ii) the position of the "x" leader and the "y" leader and (iii) the beginning of early region 3 is also depicted.

INTRODUCTION

Adenovirus 2, a member of the non oncogenic human adenovirus subgroup C, has a genome made of a linear double stranded DNA molecule of approximately 35 000 base pairs. Its genome is transcribed into early and late mRNAs. Early messenger RNAs arise from 4 different regions, two on each DNA strand (1,2,3). Late messenger RNAs arise mainly by transcription of the r strand in a long transcript which gives rise by a very complicated pattern of splicing (4,5,6) and other maturation processes to numerous mRNAs identified by in vitro protein synthesis and electron microscopy (7,8).

In order to understand the organization of the adenovirus genome and to identify the short nucleotide sequences which are likely used as a signal

for the transcription and processing of mRNAs, we first focused our attention on the EcoRI F fragment which was thought to be at the junction between early regions E2 and E3 and was responsible for the transcription of several late mRNAs (1,9,10). Nucleotide sequence analysis of this fragment (11) allowed us to determine the position and the nucleotide sequence of the first leader of the 72K mRNA which is leftward transcribed (12). From the EcoRI F sequence we were also able to determine the reading frame corresponding to the 100K mRNA and to deduce the probable amino acid sequence of the C terminal part of this protein. The mapping of the 3' end of a part of the 100K message was also determined through the presence of an AATAAA sequence found 779 nucleotides downstream of the TAG stop codon. Informations concerning the PVIII protein led us to suggest that this protein could be translated from an ATG present near the end of the EcoRI F fragment up to a stop signal within the EcoRI D fragment.

The present paper deals with the nucleotide sequence of the EcoRI D fragment which lay between coordinates 75.9 and 83.4 (9). This sequence contains informations relative to the remaining part of the message coding for the PVIII protein (13). It also contains informations relative to several early and late leaders and to the body of several early mRNAs such as the 16K mRNA (3,14,15).

MATERIALS and METHODS

All materials used were as previously described (11,16).

Isolation of viral DNA.

Culture of Hela cells, viral propagation and isolation of viral DNA were as described by Fraser and Ziff (17).

Cloning of the EcoRI D fragment and propagation of recombinant.

Viral Ad.2 DNA was digested with EcoRI and resulting fragments were fractionated by electrophoresis on 1% agarose. After elution by electrophoresis, the DNA was phenol treated, alcohol precipitated and finally dissolved in distilled water.

Bacteriophage λgt WES λC (18) was propagated in E. coli strain C_{600} and recombinant were propagated in L3 physical comtainment facility. Phage were purified by centrifugation followed by banding in CsCl. DNA was isolated from banded phage by phenol extraction (16).

λDNA was digested with EcoRI and the λ arms were purified by sucrose gradient centrifugation. λ gt WES arms and purified Ad.2 EcoRI D fragment

were mixed and ligated with T4 DNA ligase in 50mM Tris HCl pH 7.8 10mM $MgCl_2$ 20mM dithiothreitol 1mM ATP and 50 µg/ml of bovine serum albumin at 12°C for 36 hours (19). E. coli strains C_{600} R^-K^-recBC were made competent by treatment with $CaCl_2$ and transfected according to Cameron et al (20). Several recombinant phages were recovered from the petri dish and grown in liquid culture. They were assayed for the presence of an Ad.2 EcoRI D fragment by restriction enzyme mapping (21). Selected recombinant phages were propagated in liquid culture infected with a multiplicity of 0.01.

Sequencing procedure

Sequence analysis was performed according to the method of Maxam and Gilbert (22) as revised by the same authors (23) and with the previously described slight modifications (11). Usually 10 to 20 picomoles of EcoRI D fragment were fully digested with one restriction enzyme according to the condition recommended by the manufacturers. Fragments were then dephosphorylated at 65°C with bacterial alkaline phosphatase which was then removed by alkaline treatment (24). Fragments were 5' end labelled with $^{32}P\gamma ATP$ (NEN above 3 000 Ci/mM) and polynucleotide kinase (22). Labelled fragments were fractionated by electrophoresis in agarose gel or acrylamide gel and eluted from agarose gel by electrophoresis (25) and from acrylamide gel as described by Maxam and Gilbert (22). Fragments smaller than 500 base pairs were denatured by heat treatment with 30% DMSO (23). Resulting single stranded fragments were separated by electrophoresis in acrylamide gel with a low ratio of bisacrylamide/acrylamide (1/50) as described by Maxam and Gilbert (23). Fragments larger than 500 base pairs were hydrolyzed with a second restriction enzyme (16).

Sequencing gels were 400 mm long and 1 mm thick. Usually a 25% acrylamide gel was run with the bromophenol blue at 9 cm from the top to determine the first 30 nucleotides. Further nucleotide sequences were obtained with several 16% acrylamide gels in which the phenol cyanol FF dye had migrated half way down, to the bottom, twice and three times the length of the gel.

RESULTS

A lambda WES/Ad.2 EcoRI D recombinant was constructed and used as starting material to determine the nucleotide sequence of the EcoRI D fragment. The cloned EcoRI D fragment was recovered after purification of the recombinant phage by hydrolysis with EcoRI restriction enzyme and sucrose gradient centrifugation. The chemical degradation method of Maxam and Gilbert was

used. As shown in the two examples of sequencing gel autoradiograms given in Fig. 1, five different reactions were used. In some circumstances few bases - mainly C and T - did not react as expected or did not give a clear visible band after electrophoresis. Nevertheless their existence was confirmed by analysis of the other DNA strand. As shown in fig. 2, a large number of restriction fragments were used, and thus both DNA strands could be analyzed independently all along the nucleotide sequence except for the 17 residues long HhaI fragment located between position 2029 and 2045, whose nucleotide sequence was analysed several times but on the r strand only. The nucleotide sequence of all restriction cleavage sites used as starting points were also analyzed as an internal part of another restriction fragment in order to detect the existence of two identical restriction sites closely following each other, as was noted for restriction sites AluI 2072 and AluI 2078 or HhaI 2284 and HhaI 2292.

The nucleotide sequence analysis of fragments F and D has revealed the existence

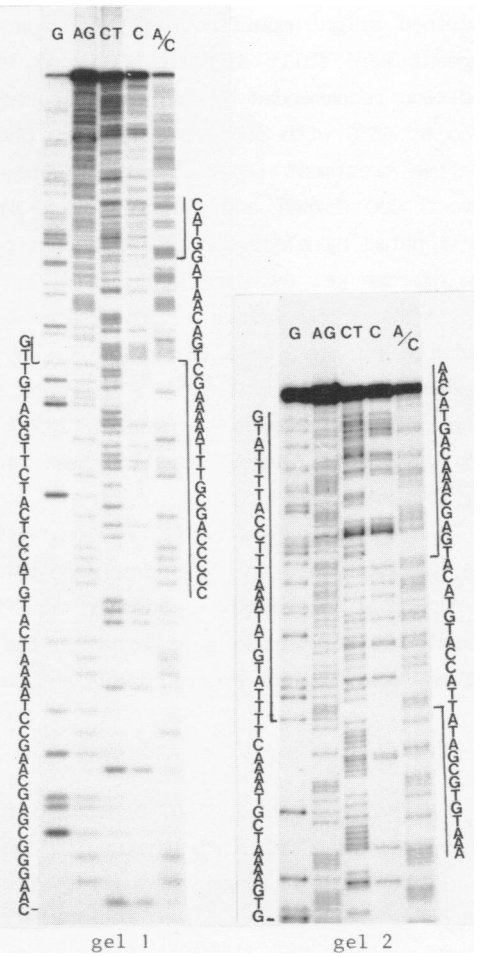


Fig. 1

Autoradiograms of sequencing gels. Products obtained by the chemical degradation method of Maxam and Gilbert were fractionated by electrophoresis on 16% polyacrylamide gels. Gel 1 (Hae II A fragment labelled on the r chain) covers the sequence from nucleotide 1473 to nucleotide 1397 and gel 2 (HindIII B fragment labelled on the l chain), from nucleotide 1690 to nucleotide 1771. The position of analysed fragments are indicated on fig. 2.

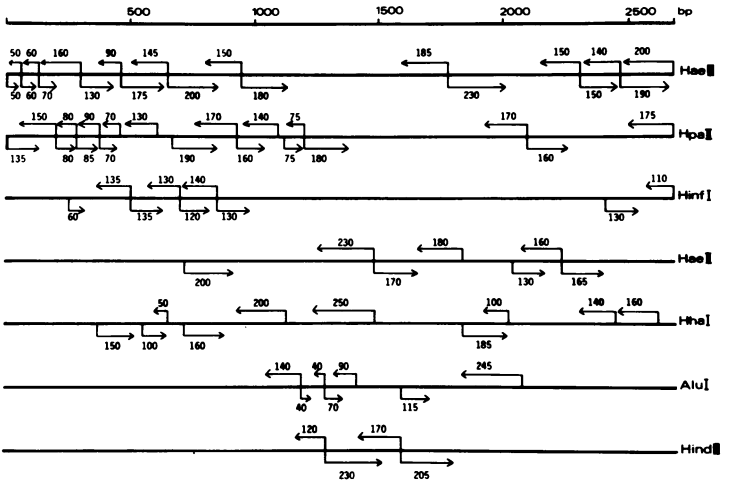


Fig. 2

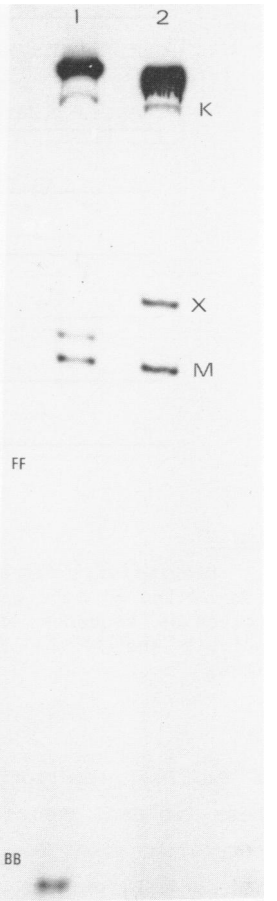
Diagram of analysed DNA fragments. Vertical bars correspond to the position of the 5' labelled ends of restriction fragments used to determine the nucleotide sequence of the Ad.2 Eco D fragment. Numbers above each arrow indicate the lenght of nucleotide sequence analysed from the restriction sites.

of two SmaI restriction sites, instead of one, mapping at position 76.8. These two SmaI restriction sites are located on either side of the EcoRI restriction site 75.9. One is located at residue 1706 in fragment EcoRI F i.e. at 33 residues upstream of the EcoRI site (11) and the other, in fragment EcoRI D, at residue 197 downstream of the EcoRI site (this work). This kind of result prompted us to analyse the nucleotide sequence around the EcoRI site 75.9 as an internal part of an overlapping fragment to be sure that no small EcoRI fragment had been missed between the EcoRI F and EcoRI D fragments (9).

Ad.2 DNA was phenol extracted from purified virion and fully digested with SmaI restriction endonuclease. Resulting fragments were 5' end labelled with polynucleotide kinase and fractionated by electrophoresis on a 6% acrylamide gel side by side with a double digest SmaI-EcoRI to characterize the SmaI fragment overlapping the EcoRI site 75.9. This fragment named X in fig. 3 was eluted. After hydrolysis with HaeIII the longest obtained subfragment was used for nucleotide sequence analysis. Results obtained demonstrate that no additional fragment exists between fragments EcoRI F and

Fig. 3

Fractionation on a 6% acrylamide gel of Ad.2-SmaI and EcoRI double digest (lane 1) and Ad.2-SmaI digest (lane 2). BB and FF indicate the position of the bromophenol and cyanol FF dyes.



EcoRI D, allowing the nucleotide sequence of fragment D to be read directly after the nucleotide sequence of EcoRI F. Therefore, the reading frame of the nucleotide sequence of EcoRI D will not be defined from the EcoRI cleavage site, as we did for the EcoRI F but according to the reading frame of the EcoRI F fragment : reading frame 3 will be AAT, TCT ; reading frame 1 will be ATT, CTC and reading frame 2 TTC, TCC (Fig. 4).

Corrected adenovirus 2 genome SmaI map

The conventional SmaI map of adenovirus 2 indicates existence of three SmaI fragments named K, L, M with a length smaller than 1000 base pairs and mapping respectively at position 99%, 18% and 11% (data compiled by M. Zabeau). Therefore according to our sequence data which revealed two SmaI sites - one on either side of the EcoRI site 75.9 - four SmaI fragments were expected after electrophoresis on a 6% acrylamide gel. Furthermore this fourth fragment was expected to be the smallest one as it was absent from the SmaI physical map. As can be seen on fig. 3 we isolated only three small fragments : (i) the X fragment which is 230 bp and maps around position 75.9. (ii) A smaller fragment of slightly less than 200 bp according to its electrophoretic mobility. This fragment more probably corresponds to the M fragment mapping at 11% and whose nucleotide sequence and position on the Ad.5 genome has recently been determined by Maat et al (26). (iii) A larger fragment of about 700 bp. This fragment corresponds to the K fragment, since it has an electrophoretic mobility identical to one of the two fragments obtained by labelling the entire Ad.2 DNA prior digestion with SmaI (data not shown). Finally, fractionation on a 1% agarose gel of an Ad.2 SmaI digest does not reveal the existence of an additional fragment

```
                                                  50
r  3'GAGGAGCTTGTCCGCCGATAATGGTGGTGTGGAGCATTATTGGAATTAGGGGCATCAACCGGGCGACGGGAC
lAATTCTCCTCGAACAGGCGGCTATTACCACCACACCTCGTAATAACCTTAATCCCCGTAGTTGGCCCGCTGCCCTG
  IleLeuLeuGluGlnAlaAlaIleThrThrThrProArgAsnAsnLeuAsnProArgSerTrpProAlaAlaLeu
                 100                                                  150
  CACATGGTCCTTTCAGGGCGAGGGTGGTGACACCATGAAGGGTCTCTGCGGGTCCGGCTTCAAGTCTACTGATTG
  GTGTACCAGGAAAGTCCCGCTCCCACCACTGTGGTACTTCCCAGAGACGCCCAGGCCGAAGTTCAGATGACTAAC
  ValTyrGlnGluSerProAlaProThrThrValValLeuProArgAspAlaGlnAlaGluValGlnMetThrAsn
                                  200
  AGTCCCCGCGTCGAACGCCCGCCGAAAGCAGTGTCCCACGCCAGCGGGCCCGTCCCATATTGAGTGGACTTTTAG
  TCAGGGGCGCAGCTTGCGGGCGGCTTTCGTCACAGGGTGCGGTCGCCCGGGCAGGGTATAACTCACCTGAAAATC
  SerGlyAlaGlnLeuAlaGlyGlyPheArgHisArgValArgSerProGlyGlnGlyIleThrHisLeuLysIle
                 250                                                  300
  TCTCCCGCTCCATAAGTCGAGTTGCTGCTCAGCCACTCGAGGAGAGAACCAGAGGCAGGCCTGCCCTGTAAAGTC
  AGAGGGCGAGGTATTCAGCTCAACGACGAGTCGGTGAGCTCCTCTCTTGGTCTCCGTCCGGACGGGACATTTCAG
  ArgGlyArgGlyIleGlnLeuAsnAspGluSerValSerSerSerLeuGlyLeuArgProAspGlyThrPheGln
                                  350
  TAGCCGCCGCGACCGGCGAGAAGTAAATGCGGGGCAGTCCGCTAGGATTGAGACGTCTGGAGCAGGAGCCTCGGC
  ATCGGCGGCGCTGGCCGCTCTTCATTTACGCCCCGTCAGGCGATCCTAACTCTGCAGACCTCGTCCTCGGAGCCG
  IleGlyGlyAlaGlyArgSerSerPheThrProArgGlnAlaIleLeuThrLeuGlnThrSerSerSerGluPro
                 400                                                  450
  GCGAGGCCTCCGTAACCTTGAGATGTTAAATAACTCCTCAAGCACGGAAGCCAAATGAAGTTGGGGAAAAGACCT
  CGCTCCGGAGGCATTGGAACTCTACAATTTATTGAGGAGTTCGTGCCTTCGGTTTACTTCAACCCCTTTTCTGGA
  ArgSerGlyGlyIleGlyThrLeuGlnPheIleGluGluPheValProSerValTyrPheAsnProPheSerGly
                                  500
  GGAGGGCCGGTGATGGGCCTGGTCAAATAAGGGTTGAAACTGCGCCACTTTCTGAGCCGCCTGCCGATGCTGACT
  CCTCCCGGCCACTACCCGGACCAGTTTATTCCCAACTTTGACGCGGTGAAAGACTCGGCGGACGGCTACGACTGA
  ProProGlyHisTyrProAspGlnPheIleProAsnPheAspAlaValLysAspSerAlaAspGlyTyrAsp
                 550                                                  600
  TACTGGTCACCTCTCCGTCTCGCTGACGCGGACTGTGTGGAGCTGGTGACGGCGGCGGTGTTCACGAAACGGGCG
  ATGACCAGTGGAGAGGCAGAGCGACTGCGCCTGACACACCTCGACCACTGCCGCCGCCACAAGTGCTTTGCCCGC
                                  650
  CCGAGGCCACTCAAAACAATGAAACTTAACGGGCTTCTCGTATAGCTCCCGGGCCGCGTGCCGCAGGCCGAGTGG
  GGCTCCGGTGAGTTTTGTTACTTTGAATTGCCCGAAGAGCATATCGAGGGCCCGGCGCACGGCGTCCGGCTCACC
                 700                                                  750
  TGGGTCCATCTCGAATGTGCATCGGACTAAGCCCTCAAATGGTTCGCGGGGGACGATCACCTCGCCCTCGCCCCA
  ACCCAGGTAGAGCTTACACGTAGCCTGATTCGGGAGTTTACCAAGCGCCCCCTGCTAGTGGAGCGGGAGCGGGGT
                                  800
  GGGACACAAGACTGGCACCAAACGTTGACAGGATTGGGACCTAATGTAGTTCTAGAAACAACAGTAGAGACACGA
  CCCTGTGTTCTGACCGTGGTTTGCAACTGTCCTAACCCTGGATTACATCAAGATCTTTGTTGTCATCTCTGTGCT
                 850                                                  900
  CTCATATTATTTATGTCTTTAATCTTAGATGACCCCGAGGACAGCGGTAGGACACTTGCGGTGGCAAAAATGGGT
  GAGTATAATAAATACAGAAATTAGAATCTACTGGGGCTCCTGTCGCCATCCTGTGAACGCCACCGTTTTTACCCA
                                  950
  GGGTTTCGTCTGGTTTCGTTTGGAGTGGAGGCCAAACGTGTTCGCCCGGTTATTCATGGAATGGACCATGAAATT
  CCCAAAGCAGACCAAAGCAAACCTCACCTCCGGTTTGCACAAGCGGGCCAATAAGTACCTTACCTGGTACTTTAA
                 1000                                                 1050
  GCCGAGAAGTAAACATTAAATGTTGTCAAAGGTCGCTCTGCTTCATTCAAACGGTGTGTTGGAAGAGCCGAAGTT
  CGGCTCTTCATTTGTAATTTACAACAGTTTCCAGCGAGACGAAGTAAGTTTGCCACACAACCTTCTCGGCTTCAA
                                  1100
  GATGTGGCAGTTCTTTTTGTGGTGGTGGTGGGAGGAGTGGACGGCCCTTGCATGCTCACGCAGTGGCCAACG
  CTACACCGTCAAGAAAAACACCACCACCACCCTCCTCACCTGCCGGGAACGTACGAGTGCGTCACCGGTTGC
                 1150                                                 1200
  ACGCGGGTGTGGATGTCGGACTCGCATTGGTCTGTAATGAGGGTAAAAAGGTTTTGTCCTCCACTCGAGTTGAGG
  TGCGCCCACACCTACAGCCTGAGCGTAACCAGACATTACTCCCATTTTTCCAAAACAGGAGGTGAGCTCAACTCC
```

```
                                        1250.
GCCTTGAGTCCAGTTTTTTCGTAAAACGCCCCACGACCCTAAAAAATTAATTCATATACTCGTTAAGTTCATTGA
CGGAACTCAGGTCAAAAAAGCATTTTGCGGGGTGCTGGGATTTTTTAATTAAGTATATGAGCAATTCAAGTAACT
          1300.                                                  1350.
GATGTTCGAACAGATTAAAAAGACCTTAACCCCAGCCCCAATAGGAATGAGAACATTAAGACAAATAAGAATATG
CTACAAGCTTGTCTAATTTTTCTGGAATTGGGGTCGGGGTTATCCTTACTCTTGTAATTCTGTTTATTCTTATAC
                                        1400.
ATCGTGAAGACACGGAATCCCAACGGCGGACGACGTGCGTGCAAACATGGATAACAGTCGAAAAATTTGCGACCC
TAGCACTTCTGTGCCTTAGGGTTGCCGCCTGCTGCACGCACGTTTGTACCTATTGTCAGCTTTTTAAACGCTGGG
          1450.                                                  1500.
CCGTTGTAGGTTCTACTCCATGTACTAAAATCCGAACGAGCGGGAACGCCGTCAGACGTCGCGACGGTTTTTCCAA
GGCAACATCCAAGATGAGGTACATGATTTTAGGCTTGCTCGCCCTTGCGGCAGTCTGCAGCGCTGCCAAAAAGGTT
                MetArgTyrMetIleLeuGlyLeuLeuAlaLeuAlaAlaValCysSerAlaAlaLysLysVal
                                        1550.
CTCAAATTCCTTGGTCGAACGTTACAATGTAAATTTAGTCTTCGATTACTTACGTGATGAGAATATTTTACGTGG
GAGTTTAAGGAACCAGCTTGCAATGTTACATTTAAATCAGAAGCTAATGAATGCACTACTCTTATAAAATGCACC
GluPheLysGluProAlaCysAsnValThrPheLysSerGluAlaAsnGluCysThrThrLeuIleLysCysThr
          1600.                                                  1650.
TGTCTTGTACTTTTCGAATAATAAGCGGTGTTTCTGTTTTAACCGTTCATACGACATATACGATAAACCGTCGGT
ACAGAACATGAAAAGCTTATTATTCGCCACAAAGACAAAATTGGCAAGTATGCTGTATATGCTATTTGGCAGCCA
ThrGluHisGluLysLeuIleIleArgHisLysAspLysIleGlyLysTyrAlaValTyrAlaIleTrpGlnPro
                                        1700.
CCACTGTGATTGCTGATATTACAGTGTCAGAAGGTTCCACTTTTAGCATTTTGAAAATACATATTTAAAGGTAAA
GGTGACACTAACGACTATAATGTCACAGTCTTCCAAGGTGAAAATCGTAAAACTTTTATGTATAAATTTCCATTT
GlyAspThrAsnAspTyrAsnValThrValPheGlnGlyGluAsnArgLysThrPheMetTyrLysPheProPhe
          1750.                                                  1800.
ATACTTTACACGCTATAATGGTACATGTACTCGTTTGTCATGTTCAACACCGGGGGTGTTTTCACAAATCTCTTG
TATGAAATGTGCGATATTACCATGTACATGAGCAAACAGTACAAGTTGTGGCCCCCACAAAAGTGTTTAGAGAAC
TyrGluMetCysAspIleThrMetTyrMetSerLysGlnTyrLysLeuTrpProProGlnLysCysLeuGluAsn
                                        1850.
TGACCGTGGAAAACAAGGTGGCGAGACGAATAATGTCGCGAACGAAACCATACATGGAATGAAATAGAGTTTATG
ACTGGCACCTTTTGTTCCACCGCTCTGCTTATTACAGCGCTTGCTTTGGTATGTACCTTACTTTATCTCAAATAC
ThrGlyThrPheCysSerThrAlaLeuLeuIleThrAlaLeuAlaLeuValCysThrLeuLeuTyrLeuLysTyr
          1900.                                                  1950.
TTTTCGTCTGCGTCAAAATAACTACTTTTCTTTTACGGAACTAAAAGGCGAACGAACATAAGGGGACCTGTTAAA
AAAAGCAGACGCAGTTTTATTGATGAAAAGAAAATGCCTTGATTTTCCGCTTGCTTGTATTCCCCTGGACAATTT
LysSerArgArgSerPheIleAspGluLysLysMetPro
                                        2000.
TGAGATACACCCTATACGAGGTCCGCCCGTTCTAATATGGGTGTTGGAAGTTTAGTTTGAAAGGACCTGCAATCG
ACTCTATGTGGGATATGCTCCAGGCGGGCAAGATTATACCCACAACCTTCAAATCAAACTTTCCTGGACGTTAGC
          2050.                                                  2100.
CGGACTAAAGACGGTCGCGGACGTGACGTTTAAACTAGTTTGGGTCGAAGTCGAACGGACGAGGTCTCTACTGGC
GCCTGATTTCTGCCAGCGCCTGCACTGCAAATTTGATCAAACCCAGCTTCAGCTTGCCTGCTCCAGAGATGACCG
                                        2150.
CGAGTTGGTAGCGCGGGTGTTGCCTGATAGCGTTGTGGTGACGATGGCCTGATTGTAGACGGGATTTAAATGGGG
GCTCAACCATCGCGCCCACAACGGACTATCGCAACACCACTGCTACCGGACTAACATCTGCCCTAAATTTACCCC
          2200.                                                  2250.
TTCAAGTACGGAAACAGTTACTGACCCGCTCGAACCTGTACACCACCAAAAGGTATCGCGAATACAAACAAACGG
AAGTTCATGCCTTTGTCAATGACTGGGCGAGCTTGGACATGTGGTGGTTTTCCATAGCGCTTATGTTTGTTTGCC
                                        2300.
AATAATAATACACCGAATAAACAACGGATTTCGCGTCTGCGCGGTCTGGGGGGTAGATATCCGGATAGTAACACG
TTATTATTATGTGGCTTATTTGTTGCCTAAAGCGCAGACGCGCCAGACCCCCCATCTATAGGCCTATCATTGTGC
          2350.                                                  2400.
AGTTGGGTGTGTTACTTTTTTAAGTATCTAACCTGCCAGACTTTGGTACAAGAGAAGAAAATGTCATACTAATTT
TCAACCCACACAATGAAAAAATTCATAGATTGGACGGTCTGAAACCATGTTCTCTTCTTTTACAGTATGATTAAA
```

```
                                                    2450.
ACTCTGTACTAAGGAGCTCAAGAATATAATAACTGGGAACAACGCGAAAAAGACACGCACGAGATGTAACCGGCGC
TGAGACATGATTCCTCGAGTTCTTATATTATTGACCCTTGTTGCGCTTTTCTGTGCGTGCTCTACATTGGCCGCG

          2500.                                                     2550.
CAGCGAGTGTAGCTTCATCTAACGTAGGGTGGAAAGTGTCAAATGGACGAAATGCCTAAACAGTGGGAATAGGAG
GTCGCTCACATCGAAGTAGATTGCATCCCACCTTTCACAGTTTACCTGCTTTACGGATTTGTCACCCTTATCCTC

                                                    2600.
TAGACGTCGGAGCAGTGACATCAGTAGCGGAAGTAAGTCAAGTAACTGACCCAAACACACGCGTAACGCATGGAG
ATCTGCAGCCTCGTCACTGTAGTCATCGCCTTCATTCAGTTCATTGACTGGGTTTGTGTGCGCATTGCGTACCTC

          2650.
TCCGTGGTAGGCGTTATGTCTCTGTCCTGATATCGACTAGAAGAGTCTTAA r chain
AGGCACCATCCGCAATACAGAGACAGGACTATAGCTGATCTTCTCAG 3'  1 chain
```

Fig. 4

Ad.2-EcoRI D nucleotide sequence. The theoretical amino acid sequence corresponding to the C terminal part of the PVIII protein and to the 16K protein are indicated. r and l stand for rightward and leftward transcribed chains.

migrating between the J and K fragments. From all these results we suggest : 1) the X fragment of fig. 3 corresponds to the L fragment ; 2) the L fragment does not map at position 18 but at position 75.9.

In fig. 4 the nucleotide sequence of the EcoRI D fragment is shown. This sequence is made of 2678 base pairs. The total length of fragments F and D is therefore 4417. This accounts for 12,7 per cent of the adenovirus 2 genome, making 1% equal to 347 base pairs. This is in very good agreement with the 350 estimate.

DISCUSSION

100K and PVIII mRNA

Chow et al, by electron microscopy examination, have located the 3' end of a family of late mRNAs at position 78.3 (13) and Fraser and Ziff have shown by fingerprinting analysis that several characteristic large T1 oligonucleotides should be present near by the 3' end of late mRNAs hybridizing to the EcoRI D fragment (17). DNA sequence analysis reveals the presence of the characteristic sequence AATAAA at position 833 on the l strand which has the same polarity as the late mRNAs. Around this AATAAA sequence on the r strand there is a DNA nucleotide sequence able to give rise by transcription to several large T1 oligonucleotides with the following nucleotide sequences: a TCCTAACCCTG; b ATTACATCAAG; c TCATCTCTG; d TATAATAAATACAG; e AAATTAG. These predicted spots aligned in the putative mRNA from 5' to 3' in the order a to e would give pancreatic products nearly identical to those

obtained by Fraser and Ziff (17) with oligonucleotides 13-10-16-5-15 : spot 5 containing the characteristic AAUAAA, spot 15 being the closest to the 3' end. An additional large T1 oligonucleotide can be predicted from the nucleotide sequence TAGAAAC (residue 804 to 810) located between the nucleotide sequence giving rise by transcription to spot 10 and 16. This AUCUUUG oligonucleotide which should migrate according to its base composition more rapidly than spot 16 in both dimensions of T1 fingerprint was not analyzed. No other large spot, predicted from the DNA sequence is absent from the RNA fingerprint if the 3' end of the mRNA is localized after the sequence AAATTAG (residue 844 and 850) corresponding to spot 15 and before the end of the AATCTACTG (residue 851 and 859). This would place the 3' end of this family of late mRNAs at map coordinate 78.3 between the 12th and 21st residue after the AAUAAA sequence, as already observed with several eukaryotic messenger RNAs (27,28,29). This very good agreement between previously published data and nucleotide sequence analysis clearly indicates the position of this family of mRNAs on the adenovirus genomic map.

Late mRNAs selected by hybridization to the EcoRI D fragment are able to direct the in vitro synthesis of the 100K and PVIII proteins (30). Actually looking at the nucleotide sequence of the EcoRI F fragment we were able to determine the end of the open reading frame corresponding to the C terminal part of the 100K and 33K protein and the AATAAA sequence located at the 3' end of part of the corresponding mRNA (11). Therefore the only known protein coded by the nucleotide sequence around EcoRI site 75.9 should be the PVIII protein. Looking for the presence of an open reading frame downstream position 74, where the 5' end of the PVIII mRNA maps (8), we previously suggested that this mRNA might be translated from one of the five ATGs present at the end of the EcoRI F fragment (11). These five ATGs are all located in the open reading frame 1. This reading frame continues beyond EcoRI site 75.9 within the EcoRI D fragment free of terminator codon up to stop codon TGA at position 524. The two other frames in EcoRI D fragment are blocked by several stop codons in that region (Fig. 5). The open reading frame from ATG_{1582} within EcoRI F to TGA_{524} within EcoRI D is able to code for a protein 227 amino acids long with a theoretical molecular weight of 24,703 daltons. This very good agreement with the 26,000 daltons molecular weight of the PVIII protein (31) is further indicative of the translating capacity of this nucleotide sequence into the PVIII protein. Nevertheless it is still possible that the actual ATG present at the N terminal end of this protein could be one of the following ATGs located at residue 1609 ;

1630 ; 1681 ; 1699 in fragment EcoRI F. In this event the PVIII protein would have a molecular weight equal to 23,656 ; 22,730 ; 20,946 or 20,366. If the 26,000 daltons molecular weight, determined by gel electrophoresis, is more in favour of the first ATG as starting triplet, one should note for the PVIII protein, according to the nucleotide sequence, the rather large amount of proline residues (20 i.e. 8,8%) which are known to slightly reduce the mobility of a protein during gel electrophoresis giving rise to a higher calculated molecular weight.

The primary structure of the PVIII protein as deduced from the nucleotide sequence shows several characteristics. As noted above there is a large number of prolines preventing the formation of a large portion of α helical structure. The amount of arginine 7.5% and the absence of cysteine residue are also noticeable. The absence of the sequence Asn-X-Thr (or Ser) suggests this protein is not glycosylated (32).

Early 3 region

Fragment EcoRI D codes for part of the early region 3 which maps from coordinates 76.6 within EcoRI D to coordinate 86 within EcoRI E (9,10).

Electron microscopy mapping experiments have mainly described two classes of early mRNAs from the E3 region (3). Messenger RNAs of the first class have a leader mapping at coordinates 76.6-77.6 and a body hybridizing from coordinate 78.6 up to coordinates 82.7, 84.7 and 86. Messenger RNAs of the second class have identical over all map coordinates but a second deletion between coordinates 79.1 and 81.5, 82.7 or 84.7 creates a second leader (78.6-79.1). This second leader is probably identical to the fourth late leader named "y" and found in some fiber mRNA (13).

Recently Baker and Ziff have determined the nucleotide sequence of a stretch of DNA 42 residues long (unpublished results). This sequence contains the putative TATAA recognition sequence for the RNA polymerase and the capped nucleotide of region E3 mRNA's. An exactly identical sequence is found in the EcoRI D fragment between residue 203 and 244. This places the capped G nucleotide at residue 237 and locates the beginning of the E3 region at coordinates 76.6 in complete agreement with previously published data (3).

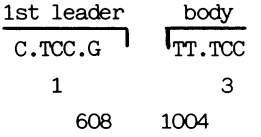
At the 3' end of several leaders such as those found in rabbit and mouse β globin mRNA (33) first late Ad.2 (34) first early 2 Ad.2 (12) fourth late Ad.2 (14) a GGGTGAG sequence has been observed. Therefore we looked for a similar sequence downstream the beginning of E3 region at a distance corresponding to the estimated length of the leader. A GGTGAG sequence is

found at position 608 (coordinate 77.6) defining from the G cap a DNA segment 372 nucleotides long and no similar sequence is found around. This very good agreement between the electron microscopy and nucleotide sequence data suggest very strongly that the DNA stretch delineated by residue 237 and the GGTGAG sequence (residue number 608) indeed corresponds to the first leader of the E3 region.

By electron microscopy analysis it has been shown that the body of the first class of E3 mRNAs, the second leader of the second class of E3 mRNAs and the fourth leader of the fiber mRNA start on the genome at coordinate 78.6 (3). Recently Zain et al (14) have determined the nucleotide sequence of the "y" leader which is 181 residues long. This sequence is identified within the EcoRI D nucleotide sequence. It maps at coordinates 78.7 - 79.2 and starts with nucleotide 1002 and stops with nucleotide 1185. The EcoRI D sequence differs from the previously published nucleotide sequence of the "y" leader by the addition of three base pairs. A GC, an AT and a CG pair are observed respectively at position 1094, 1124 and 1125.

Early mRNAs selected by hybridization with EcoRI D direct the in vitro synthesis of 3 polypeptides with a molecular weight of 13,000, 14,000 and 15,500 daltons (15). The translating capacity of the EcoRI D sequence was therefore tested by determined the number and the location of the various stop codons and potential ATG initiator triplets (Fig. 5).

The first ATG found downstream the G cap (residue number 237) is within the first leader at position 527. This ATG is in phase 1 which is open downstream up to the end of this leader. As indicated above the nucleotide sequence spliced to this leader is known. If this sequence is spliced to the leader according to previous established rules (35,36), leaving GT and AG at the border of the intron sequence, the open reading phase 1 of the leader is followed by frame 3 which is blocked twice at stop codons, TAA_{1021} and TGA_{1189}.



On the contrary if the body of the mRNA is spliced according to the "y" nucleotide sequence as determined by Zain et al (14), the reading frame stays open and can be read through from the first ATG_{527} to stop codon TAA_{1247} and a protein of 12,000 daltons can be made.

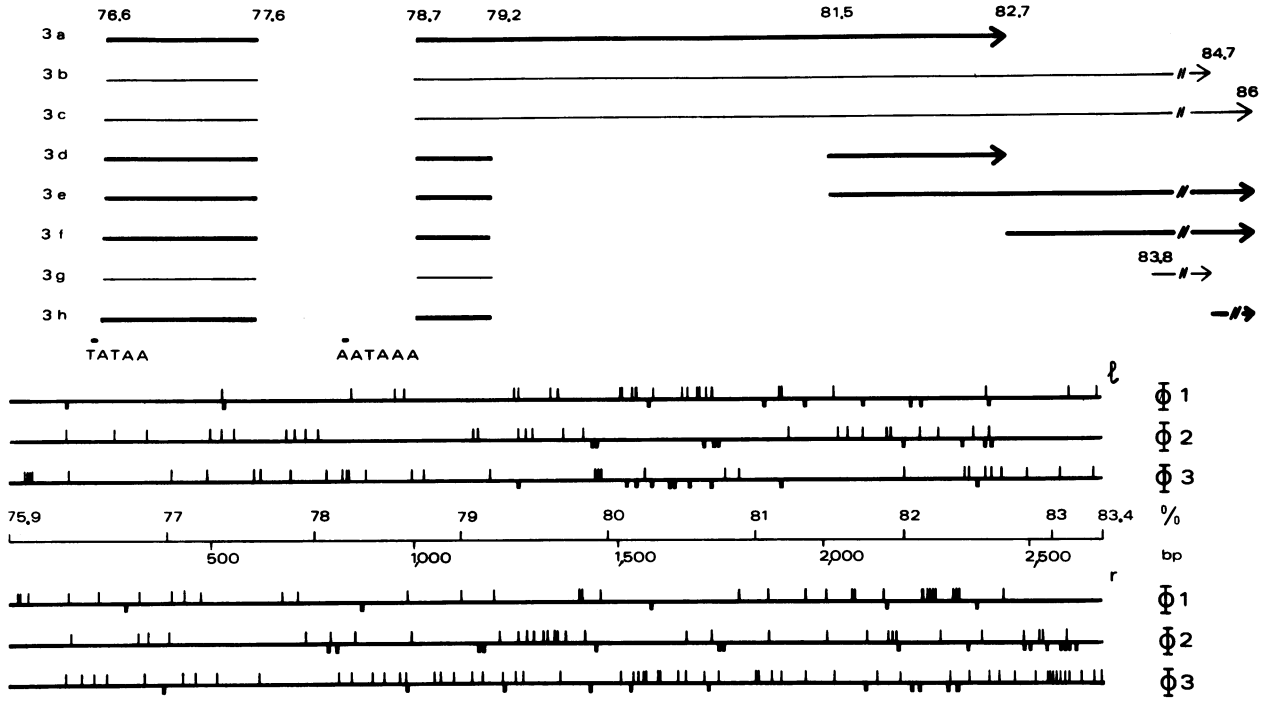
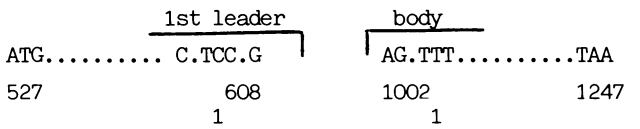


Fig. 5

Diagram showing the localisation of initiator and stop codons. The position of the 3a to 3h mRNAs are from Chow et al (3). The nucleotide sequences TATAA and AATAAA correspond respectively to the 5' end of the early 3 mRNAs and to the 3' end of late mRNAs ending at 78. $\phi 1$, $\phi 2$ and $\phi 3$ correspond to the various reading frames as defined in the text. Upper vertical bars are for the nonsense codons. Underneath vertical bars correspond to the ATG triplets.



Nevertheless the use of the ATG_{527} as initiator codon and the synthesis of this 12,000 daltons protein does not seem to occur for the following reason. Since all the early messages of that region have an identical first leader and lose during processing an identical first intron (3), the use of ATG_{527} as initiator codon stands for all mRNAs. It follows that all the early proteins coded by region 3 would have an identical N terminal sequence, 87 amino acids long. This common 87 amino acids sequence added in front to the amino acid sequences translated by the other open reading frame present downstream, will give rise to proteins with a molecular weight not compatible with the result of in vitro protein synthesis (15).

Furtherdown in the sequence there is an ATG at position 1258.As can be seen in fig. 5, this ATG does not seem to be able to initiate protein synthesis, since it is located in phase 3 which is blocked by TGA_{1441}. On the contrary one of the two ATGs 1440 and 1449 (map coordinate 79.9) both in phase 2, open up to TGA_{1917}, could initiate protein synthesis. This open reading phase from ATG_{1440} to TGA_{1917} could code for a protein of 159 amino acids with a theoretical molecular weight of 18,439. This protein does probably exist and corresponds to the 15,500 molecular weight protein obtained by in vitro synthesis (15) and more probably to the in vivo synthesized 18,500 molecular weight glycosylated protein recently mapped by Ross and Levine (37) between coordinates 79.8 and 81.

According to the nucleotide sequence this protein should have a rather large amount of lysine residue (11.3%). A sequence Asn-Val-Thr is present twice in agreement with the potential glycosylation of this protein (32,37).

This protein is probably coded by mRNA 3a (fig. 5) which is a major species. This mRNA according to Chow et al ends up at position 82.7 (3). However the sequence AATAAA which has been found in many circumstances near the position of poly A addition (27,28,29) is not observed in the nucleotide sequence of that region. Could it be possible that this message does indeed end up within the EcoRI E fragment which is only at 0.7% ? This possibility is not as yet borne out by the very good agreement within 0.2 or 0.3% between coordinates obtained by EM and nucleotide sequence ; or could it be possible that this messenger RNA has a particular structure at its 3' end without the common AAUAAA sequence - a situation which has already been

observed with the HBs gene of the hepatitis B virus (38).The minor species 3b and 3c messenger RNAs which have an identical 5' structure to 3a mRNA should code for the same protein.

On the DNA sequence in front of ATG_{1440} a CAGCTT sequence is found at position 1408 (map coordinate 79.8). This sequence which greatly resembles the CAGTTT sequence found at the 5' end of several leaders (14,35,39) or mRNA bodies (33) could suggest an additional splicing point to link the second early leader to the messenger RNA body. It this case, class I early 3 messengers could have a similar structure to class II early 3 messengers with the same two leaders in front of the body of the messenger RNAs. This intron less than 1% in size (1187-1408) could have been missed by electron microscopy examination because of its small size.

An advantage offered by this kind of splicing might be elimination of the ATG_{1258}, located in a close reading frame, allowing the recognition site for the ribosomes to be the same for all the early mRNAs of region 3 and to be located upstream in the "y" leader or at the end of the first leader dowstream ATG_{527}, each successive intron in the various early mRNAs allowing in turn the expression of the genes located downstream.

Further down in the nucleotide sequence one can find two open phases. One starting at position 80.9 (residue number 1798) is open in frame 3 and possesses an ATG at residue number 1900. The translating capacity from the ATG_{1900} to the stop codon TGA_{2197} corresponds to a protein of 99 amino acids. The other open reading phase, in frame 1, starts at position 81.5 (residue number 2027). Translation from the ATG_{2096} to the stop codon TAA_{2399} could correspond to a protein of 100 amino acids. There is no indication that these proteins exist nevertheless they could correspond to the translation products of the two major species, 3d and 3e mRNAs, which have been respectively mapped by electron microscopy from 81.5 to 82.7 and 81.5 to 86 (3). If this is true, presence of a CAGTTT sequence is worth noting at position 1889 in front of ATG 1900 and of a CAGCTT sequence repeated twice at position 2071 and 2077 in front of ATG 2096. On the other hand, if the position of the "y" leader (residue number 1002 to 1185) which maps according to Chow et al (3) at coordinates 78.6-79.1 is used as a reference, map coordinate 81.5 falls between ATG_{1900} and ATG_{2096} leaving ATG_{2096} as the only possible initiator codon.

Another early mRNA has been mapped by EM at position 82.7. This mRNA called 3f ends up at position 86 within the EcoRI E fragment (3). Upon considering the nucleotide sequence around residue number 2430 $\mp$ 100,

several observations can be made to estimate the position of the 5' end of the body of 3f mRNA and the possibility for the protein synthesis to start within the EcoRID fragment. Between ATG_{2340} and TGA_{2410} as shown in fig. 4 and 5 there are 5 ATGs and 7 stop codons. The first 3 ATGs are located in a closed reading frame. The other two, ATG_{2402} and ATG_{2409}, are respectively located in reading frame 1 open up to TGA_{2597} and in frame 2 open up to the end of the EcoRI D fragment.

If we still consider, as already pointed out, that the recognition site for the ribosome is located at the end of the first leader or within the second one, it follows that the first ATG encountered within the body of the messenger RNA is used as initiator triplet. This would place the 5' end of 3f mRNA slightly downstream residue 2399 in order to use ATG_{2402} as initiator or downstream nucleotide 2403 to use ATG_{2409} as initiator triplet. Further down in the sequence, at position 2515, the already familiar sequence CAGTTT is observed, which could indicate the beginning of the body of the 3f mRNA. Downstream this sequence, reading frame 1 and 3 are closed and reading frame 2 is open up to the end of the EcoRI D fragment but without the presence of any ATG to initiate the protein synthesis relegating the coding ability of 3f mRNA within its EcoRI E counter part. No argument allowing a choice between these various possibilities can be advanced.

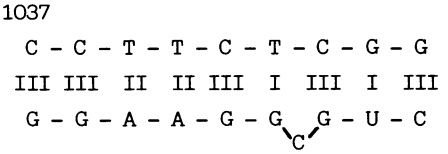
The "x" and "y" late leaders

All late Ad.2 mRNAs have a tripartite leader which maps at position 16.5 ; 19.5 ; 26.5 (5,6). In addition the fiber mRNA has two other leaders named "x" and "y" (13). Leader "y" which has been mapped at coordinates 78.6-79.1 by electron microscopy and whose nucleotide sequence has been determined from a cloned mRNA (14) corresponds to the nucleotide sequence delineated by residue 1002 and 1185, in very good agreement with electron microscopy data. Leader "x" has been mapped by electron microscopy at position 77.2-77.6 (13) and its 3' end coincides with the 3' end of the first leader of region 3 early mRNAs (3). As discussed above this 3'end could well correspond to the nucleotide sequence GGTGAG (residue number 608). Upon considering the EcoRI D nucleotide sequence, upstream of that

position and at a distance corresponding to the expected length of the "x" leader, a CAGTTT sequence (residue number 473) is found. This sequence which has been already observed at the 5' end of several leaders (14,35,39) and very close to the position where the 5' end of the body of 3 d and f mRNAs have been mapped (this work) could indicate the 5' end of the "x" leader. Endonucleolytic cleavage leaving GT and AG at both ends of the lost sequence (intron) would give rise to a leader 133 residues long corresponding to the nucleotide sequence 476 to 608, which fits perfectly well with a length of 0.4% estimated by electron microscopy. This leader contains an ATG at position 527 which could indicate that the ribosome recognition site for the late mRNAs is not in the tripartite leader. Otherwise this ATG_{527} would function as an initiator triplet. The nucleotide sequence downstream and the "y" leader would then be translated into the N terminal amino acids sequence of the fiber. This is not probable since the fiber mRNA has a heterogeneous 5' structure - some have no "x" leader and some no "y" leader while others have both leaders (13).

Binding site of the ribosome

Inconsistent results have been obtained concerning the putative hybridization of the 3' end conserved part of the 18S ribosomal RNA to a messenger RNA nucleotide sequence located upstream the ATG initiator triplet (34,38,40,41). A sequence able to base pair with the 18S rRNA has been observed in several cases, but a computer analysis made on all published mRNA sequences to determine the presence of a sequence able to hybridize to the 18S RNA does not seem to bear out this idea (42). On the other hand, this hybridization has been demonstrated in procaryotic cells and it would be surprising if nothing similar but more complex does not occur in more evolved living cells. Upon considering the EcoRI D nucleotide sequence for a possible base pairing between mRNA and the 3' end of the 18S ribosomal RNA, several sequences could be noted. Among them a sequence located at residue number 1037 could pair in the following way



The intriguing point is the location of this sequence within the second early leader where the ribosome binding site might be, as was stated above.

The CAGTTT sequences

As already pointed out a sequence identical or very similar to CAGTTT has been observed at the 5' end of several leaders and mRNA bodies (14,33,34,39). In the EcoRI D fragment this sequence was used to define the beginning of the "x" leader. It was also found in many places where a 5' end of mRNA body was positioned by electron microscopy. Since there is no explanation for the presence of this sequence and no argument proving its requirement it is worth noting that this sequence occurs 11 times in the 1 strand which has a polarity identical to the mRNA's. From a random distribution this sequence would only be expected 2 to 3 times. On the contrary there is no similar sequence in the other strand which does not correspond in that region to any mRNA. In Table I are shown the nucleotide sequence surrounding the various "CAGTTT" found within the EcoRI D fragment. Upon considering these nucleotide sequences one may note the presence of identical short sequences such as A_nG, AAAT and TTTG whose presence does not seem to be only explanable by a random distribution.

As already pointed out the organization of the Ad. 2 genome is quite complex. Several unrelated informations use the same nucleotide sequence for different purposes. While the AATAAA ending the 100K and PVIII messagers falls between the two leaders of early messenger RNA, the promoter and the 5' end of early mRNAs use a nucleotide sequence also used to translate the PVIII protein. Moreover part of the non coding region of the PVIII messenger

161 TCAGATGACTAACTCAGGGGCGCAGCTTGCGGGCGGCTTTCGTCACAGGGTGC
473 ACCTCCCGGCCACTACCCGGACCAGTTTATTCCCAACTTTGACGCGGTGAAAG
1001 GCTCTTCATTTGTAATTTACAACAGTTTCCAGCGAGACGAAGTAAGTTTGCCA
1408 CACGCACGTTTGTACCTATTGTCAGCTTTTTAAACGCTGGGGGCAACATCCAA
1477 TTAGGCTTGCTCGCCCTTGCGGCAGTCTGCAGCGCTGCCAAAAAGGTTGAGTT
1516 AAAAAGGTTGAGTTTAAGGAACCAGCTTGCAATGTTACATTTAAATCAGAAGC
1678 GACACTAACGACTATAATGTCACAGTCTTCCAAGGTGAAAATCGTAAAACTTT
1889 ATCTCAAATACAAAAGCAGACGCAGTTTTATTGATGAAAAGAAAATGCCTTGA
2071 GCACTGCAAATTTGATCAAACCCAGCTTCAGCTTGCCTGCTCCAGAGATGACC
2515 GTAGATTGCATCCCACCTTTCACAGTTTACCTGCTTTACGGATTTGTCACCCT
2589 CACTGTAGTCATCGCCTTCATTCAGTTCATTGACTGGGTTTGTGTGCGCATTG

Table 1

Nucleotide sequence surrounding the various "CAGTTT" sequences found in the 1 chain of the EcoRI D fragment. Sequences with only one C residue replacing any of the 3 Ts are also take in account. Figure indicate the position of the "CAGTTT".

RNA is also found in the first leader for early messenger RNAs and in the "x" leader of the fiber mRNA (Fig. 5). On the contrary electron microscopy heteroduplex experiments have shown that no mRNA are able to hybridize to the l strand. As a matter of facts the distribution of the non sens codons and ATG triplets in the r strand (Fig. 5) suggest that the l strand have no coding capacity in the EcoRI D region.

Biohazards associated with the experiments described in this publication have been examined previously by the French National Control Committee.

ACKNOWLEDGEMENT

This work was supported by Grant from the Delegation Generale à la Recherche Scientifique et Technique (DGRST) and by l'Institut National de la Santé et de la Recherche Medicale (INSERM).

Note : While this paper was submitted for publication. Persson, Jörnvall and Zabielski confirmed by partial aminoacid sequences analysis our prediction concerning the start of the 16K protein at ATG_{1440} (Proc. Nat. Acad. Sci., submitted for publication).

REFERENCES

1. Flint J. (1977) Cell 10, 153-166.
2. Berk A.J. and Sharp P.A. (1977) Cell 12, 721-732.
3. Chow L.T., Broker T.R. and Lewis J.B. (1979) J. Mol. Biol. 134, 265-303
4. Klessig D.F. (1977) Cell 12, 9-22.
5. Berget S.M., Moore C. and Sharp P.A. (1977) Proc. Natl. Acad. Sci. U.S.A. 74, 3171-3175.
6. Chow L.T., Gelinas R.E., Broker T.R. and Roberts R.J. (1977) Cell 12, 1-8.
7. Nevins J.R. and Darnell J.E. (1978) J. Virol 25, 811-823.
8. Chow L.J., Roberts R.J., Lewis J.B. and Broker T.R. (1977) Cell 11, 819-836.
9. Mulder C., Arrand J.R., Delius H., Keller W., Pettersson U., Roberts R.J. and Sharp P.A. (1974) Cold Spring Harbor Symp. Quant. Biol. 33, 397-400.
10. Pettersson U., Tibbetts C. and Philipson L. (1976) J. Mol. Biol. 101, 497-502.
11. Galibert F., Hérissé J. and Courtois G. (1979) Gene 6, 1-22.
12. Baker C.C., Hérissé J., Courtois G., Galibert F. and Ziff E. (1979) Cell 18, 569-580.
13. Chow L.T. and Broker T.R. (1978) Cell 15, 497-510.
14. Zain S., Sambrook J., Roberts R.J., Keller W., Fried M. and Dunn A. (1979) Cell 16, 851-861.
15. Harter M.L. and Lewis J.B. (1978) J. Virol. 26, 736-749.
16. Hérissé J., Courtois G and Galibert F. (1978) Gene 4, 279-294.
17. Fraser N. and Ziff E. (1978) J. Mol. Biol. 124, 27-51.
18. Leder P., Tiemeier D. and Enquist L. (1977) Science 196, 175-177.
19. Cordell D., Bell G., Tisher E., De Noto F.M., Ullrich A., Pictet R., Rulter W.J. and Goodman H.M. (1979) Cell 18, 533-543.

20. Cameron J.R., Panasenko S.M., Lehman I.R. and Davis R.W. (1975) Proc. Natl. Acad. Sci. U.S.A. 72, 3416-3420.
21. Mc Parland R.H., Engelking H.M. and Pearson G.D. (1978) Biochem. Biophys. Acta 518, 413-423.
22. Maxam A.M. and Gilbert W. (1977) Proc. Natl. Acad. Sci. U.S.A. 74, 560-564.
23. Maxam A.M. and Gilbert W. (1979) Methods in Enzymology in press.
24. Kroecker W.D. and Laskowski M.S.R. (1977) Analyt. Biochem. 79, 63-73.
25. Galibert F., Sedat J. and Ziff E. (1974) J. Mol. Biol. 87, 377-402.
26. Maat J., Van Beveren C.P. and Van Ormondt H. (1980) Gene, in press.
27. Mc Reynolds L., O'Malley B.W., Nisbet A.D., Fothergill J.E., Givol D., Fields S., Robertson M. and Brownlee G.G. (1978) Nature 273, 723-728.
28. Efstratiadis A. and Kafatos F.C. (1977) Cell 10, 571-585.
29. Proudfoot N.J. (1977) Cell 10, 559-570.
30. Lewis J.B., Anderson C.W. and Atkins J.F. (1977) Cell 12, 37-44.
31. Lewis J.B., Anderson C.W., Atkins J.F. and Gesteland R.F. (1974) Cold Spring Harbor Symp. Quant. Biol. 39, 581-590.
32. Struck D.K., Lennartz W.J. and Brew K. (1978) J. Biol. Chem. 253, 5786-5794.
33. Konkel D.A., Tilghman S.M. and Leder P. (1978) Cell 15, 1125-1132.
34. Ziff E. and Evans R. (1978) Cell 15, 1463-1475.
35. Breathnach R., Benoist C., O'Hare K., Gannon F. and Chambon P. (1978) Proc. Natl. Acad. Sci. U.S.A. 75, 4853-4857.
36. Catterall J.F., O'Malley W.O., Robertson M.A., Staden R., Tanaka R. and Brownlee G.G. (1978) Nature 275, 510-513.
37. Ross S.Z. and Levine A.J. (1979) Virology 99, 427-430.
38. Galibert F., Mandart E., Fitoussi F., Tiollais P. and Charnay P. (1979) Nature 281, 646-650.
39. Akusjarvi G. and Pettersson U. (1979) J. Mol. Biol. 134, 143-158.
40. Hagenbüchle O., Santer M., Steitz J.A. and Mans R.J. (1978) Cell 13, 551-563.
41. Alberty H., Raba M. and Gross H.J. (1978) Nucleic Acid Res. 5, 425-434.
42. De Wachter R. (1979) Nucleic Acid Res. 7, 2045-2054.