Volume 9 Number 5 1981 Nucleic Acids Research

Nucleotide sequence of the EcoRI E fragment of adenovirus 2 genome

Jacques Hérissé and Francis Galibert

Laboratoire d'Hématologie Expérimentale, Centre Hayem, Hôpital Saint-Louis, 75475 Paris Cédex 10, France

Received 9 January 1981

SUMMARY

The entire nucleotide sequence of the Ad.2 EcoRI E fragment has been determined using the Maxam and Gilbert method. This sequence of 2222 bp, which maps between coordinate 83.4 and 89.7 contains information relative to the early 3 region and to the fiber gene. Altogether with fragment EcoRI D which has been recently sequenced, they cover the entire Early 3 region in which several mRNA were mapped. The aminoacid sequence of the 16K and 14K protein is deduced. The localization of the 14.5K mRNA directing the synthesis of the third E3 known protein is discussed, as well as the hypothetical existence of three other early 3 proteins, which would have a molecular weight of 11K.

The initiator ATG triplet of the fiber protein has been found at coordinate 86.1, it is followed up to the end of the fragment by an open reading frame allowing deduction of 80% of the aminoacid sequence of this protein.

Sequences known to be frequently present at the border of exon sequence were used to tentatively localize the additional "Z" late leader.

INTRODUCTION

Mapping of mRNA by electron microscopy experiments and S1 nuclease digestion (1,2), as well as nucleotide sequences analysis of cloned mRNA (3,4) and restriction DNA fragments (5-11) have been used to study the organization of the adenovirus genome.

These studies reveal a rather complex genome organization. Transcription is made at immediate early, early, intermediate and late time after infection (12-14) and contrary to the papovavirus genome organization (15,16), early and late regions are interspersed and scattered all along the adenovirus genome (17-19). The nucleotide sequences of the EcoRI F and D fragments (4417 bp) which map from coordinates 70.7 to 83.4 have been determined previously (10,11). These fragments cover the end of the late 4 region which codes for the 100K, 33K and pVIII proteins (1) and give rise to the additional x and y late leaders (14). They also code for the first half of the early 3 region (1), including the two leaders and the mRNA body of the 16K protein (20). On the opposite strand, the first leader of the early 2 region has also been mapped within the EcoRI F fragment at coordinate 75.1 (21). The present paper deals with the nucleotide sequence of the EcoRI E fragment which is next to EcoRI D fragment and maps from coordinates 83.4 to 89.7. This fragment codes for the remaining part of the early 3 region and for 80% of the fiber.

MATERIALS and METHODS

All materials used were as previously described (10,22).

Culture of HeLa cells, viral propagation and isolation of viral DNA were as described by Fraser and Ziff (23).

Cloning of the EcoRI E fragment, propagation of the recombinant and sequencing procedure using the Maxam and Gilbert method (24,25) were done as previously described (11).

RESULTS

A lambda WES/Ad.2 EcoRI E recombinant was constructed and used as starting material to determine the nucleotide sequence of the EcoRI E fragment. The cloned EcoRI E fragment was recovered after purification of the recombinant phage by hydrolysis with EcoRI restriction enzyme and sucrose gradient centrifugation. Five chemical reactions specific for G, AG, CT, C and AC were performed (24,25). As shown in fig.1, a large number of restriction fragments were used, and thus both DNA strands could be analysed independently all along the nucleotide sequence. Furthermore nucleotide sequence of all restriction cleavage sites used as starting points were analyzed as an internal part of another restriction fragment.

The EcoRI site mapped at 83.4 splits the E3 region into two halves. To check the absence of a small EcoRI fragment between the EcoRI D and E fragments the nucleotide sequence around EcoRI site 83.4 was analysed on a HinfI restriction fragment prepared from Ad.2 viral DNA.

Results obtained demonstrate that no additional fragment exists between EcoRI D and E fragments, allowing the nucleotide sequence of fragment E to be read directly after the nucleotide sequence of fragment D (11). Reading frames of EcoRI E fragment will therefore be defined as follows :
AAT, TCT correspond to reading frame 1 ; ATT, CTT to reading frame 2 and TTC, TTT to reading frame 3.

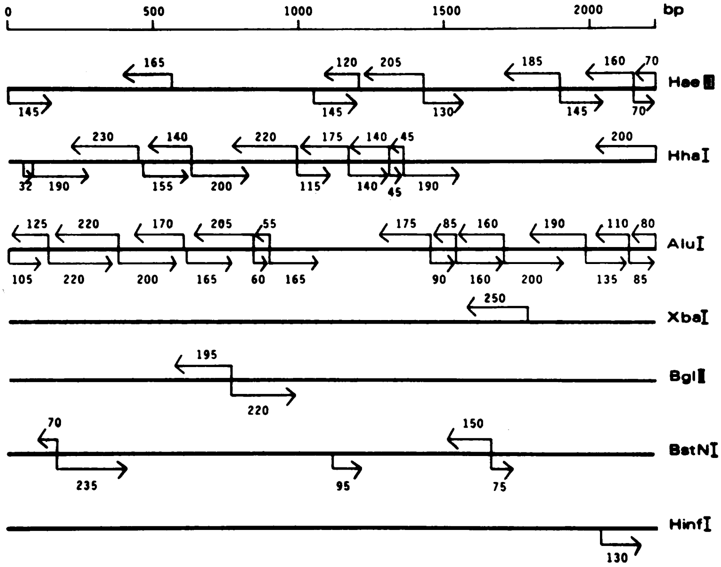

Fig.1 : Diagram of analysed DNA fragments. Vertical bars correspond to the position of the 5' labelled ends of restriction fragments used to determine the nucleotide sequence of the Ad.2 EcoRI E fragment. Numbers above each arrow indicate the length of nucleotide sequence analysed from the restriction sites.

In fig.2, is shown the nucleotide sequence of the EcoRI E fragment. This sequence is made of 2222 bp and is numbered from 2675 to 4896. The total length of the F, D and E fragments which account for 19% is therefore 6635 bp making 1% equal to 349 base pairs.

DISCUSSION

- Fiber mRNA and protein

On a cloned fiber mRNA, Zain et al (26,27) have determined a sequence of 62 nucleotides located downstream the ATG used as starting signal. An exactly identical sequence has been found within the EcoRI E fragment starting with ATG_{3658} (fig.2). This sequence open in reading frame 3 is located at coordinate 86.1 in accordance with previous results, and stays open up to the end of EcoRI E, while the two other frames are blocked by numerous nonsense triplets (fig.3). The coding capacity of that region account more probably for the 413 aminoacids located at the N-terminal end

```
 2675
 •                        2700
                          •
r    GAAATTAATACTTTGCCTCACAGTAAAAACAAAACGACTAAAAAACGCGGGATGGACACGAAACGAGGGTT
lAATTCTTTAATTATGAAACGGAGTGTCATTTTTGTTTTGCTGATTTTTTGCGCCCTACCTGTGCTTTGCTCCCAA
             MetLysArgSerValIlePheValLeuLeuIlePheCysAlaLeuProValLeuCysSerGln
             14.5K
 2750                                              2800
 •                                                 •
 TGGAGTCGCGGAGGGTTTTCTGTATAAAGGACGTCTAAGTGAGTTTATACCTTGTAAGGGTCGACGATGTTGTTT
 ACCTCAGCGCCTCCCAAAAGACATATTTCCTGCAGATTCACTCAAATATGGAACATTCCCAGCTGCTACAACAAA
 ThrSerAlaProProLysArgHisIleSerCysArgPheThrGlnIleTrpAsnIleProSerCysTyrAsnLys

                          2850
                          •
 GTCTCGCTAAACAGTCTTCGGACCAATATGCGGTAGTAGAGACAGTACCAAAAAACGTCATGGTAAAAACGGGAT
 CAGAGCGATTTGTCAGAAGCCTGGTTATACGCCATCATCTCTGTCATGGTTTTTTGCAGTACCATTTTTGCCCTA
 GlnSerAspLeuSerGluAlaTrpLeuTyrAlaIleIleSerValMetValPheCysSerThrIlePheAlaLeu

 2900                                              2950
 •                                                 •
 CGGTATATAGGTATGGAACTGTAACCGACCTTACGGTATCTACGGTACTTGGTGGGATGAAAGGGTCACGGGCGA
 GCCATATATCCATACCTTGACATTGGCTGGAATGCCATAGATGCCATGAACCACCCTACTTTCCCAGTGCCCGCT
 AlaIleTyrProTyrLeuAspIleGlyTrpAsnAlaIleAspAlaMetAsnHisProThrPheProValProAla

                          3000
                          •
 CAGTATGGTGACGTTGTCCAATAACGGGGTTAGTTAGTCGGAGCGGGGGGAAGAGGGTGGGGGTGACTCTAATCG
 GTCATACCACTGCAACAGGTTATTGCCCCAATCAATCAGCCTCGCCCCCCTTCTCCCACCCCCACTGAGATTAGC
 ValIleProLeuGlnGlnValIleAlaProIleAsnGlnProArgProProSerProThrProThrGluIleSer

 3050                                              3100
 •                                                 •
 ATGAAATTAAACTGTCCACCTCTACTGACTTAGAGATCTAGATCTTAACCTACCTTAATTGTGGCTTGTCGCGGAT
 TACTTTAATTTGACAGGTGGAGATGACTGAATCTCTAGATCTAGAATTGGATGGAATTAACACCGAACAGCGCCTA
 TyrPheAsnLeuThrGlyGlyAspAsp
                       MetThrGluSerLeuAspLeuGluLeuAspGlyIleAsnThrGluGlnArgLeu
                       14K
                         3150                                             3200
                         •                                                 •
 GATCTTTCCGCGTTCCGCCGCAGGCTCGCTCTTGCGGATTTTGTTCTTCAACTTCTGTACCAATTGGATGTGGTC
 CTAGAAAGGCGCAAGGCGGCGTCCGAGCGAGAACGCCTAAAACAAGAAGTTGAAGACATGGTTAACCTACACCAG
 LeuGluArgArgLysAlaAlaSerGluArgGluArgLeuLysGlnGluValGluAspMetValAsnLeuHisGln

                                                  3250
                                                  •
 ACATTTTCTCCATAGAAAACACACCAGTTCGTCCGGTTTGAATGGATGCTTTTTTGGTGATGGCCGTTGGCGGAG
 TGTAAAAGAGGTATCTTTTGTGTGGTCAAGCAGGCCAAACTTACCTACGAAAAAACCACTACCGGCAACCGCCTC
 CysLysArgGlyIlePheCysValValLysGlnAlaLysLeuThrTyrGluLysThrThrThrGlyAsnArgLeu

                         3300                                             3350
                         •                                                 •
 TCGATGTTCGATGGGTGGGTCGCGGTTTTTGACCACGAATACCACCCTCTTTTTGGATAGTGGCAGTGGGTCGTG
 AGCTACAAGCTACCCACCCAGCGCCAAAAACTGGTGCTTATGGTGGGAGAAAAACCTATCACCGTCACCCAGCAC
 SerTyrLysLeuProThrGlnArgGlnLysLeuValLeuMetValGlyGluLysProIleThrValThrGlnHis

                                                  3400
                                                  •
 AGCCGTCTTTGTCTCCCGACGGACGTGAAGGGGATAGTCCCAGGTCTCCTGGAGACGTGAGAATAATTTTGGTAC
 TCGGCAGAAACAGAGGGCTGCCTGCACTTCCCCTATCAGGGTCCAGAGGACCTCTGCACTCTTATTAAAACCATG
 SerAlaGluThrGluGlyCysLeuHisPheProTyrGlnGlyProGluAspLeuCysThrLeuIleLysThrMet

                        3450                                              3500
                          •                                                •
 ACACCATAATCTCTAGAATAAGGTAAGTTGATTGTATTTGTGTGTTATTTAATGAATGAATTTTAGTCAGTCGTTT
 TGTGGTATTAGAGATCTTATTCCATTCAACTAACATAAACACACAATAAATTACTTACTTAAAATCAGTCAGCAAA
 CysGlyIleArgAspLeuIleProPheAsn
```

```
                                               3550
                                                •
AGAAACAGGTCGAATAAGTCGTAGTGGAGGAAAGGAAGGAGGGTTGAGACCATAGAGTCGGCGGAAAATCGACGTT
TCTTTGTCCAGCTTATTCAGCATCACCTCCTTTCCTTCCTCCCAACTCTGGTATCTCAGCCGCCTTTTAGCTGCAA

                     3600                                               3650
                      •                                                  •
TGAAAGAGGTTTCAAATTTACCCTACAGTTTAAGGAGTACAAGAACAGGGAGGCGTGGGTGATAGAAGTATAACAA
ACTTTCTCCAAAGTTTAAATGGGATGTCAAATTCCTCATGTTCTTGTCCCTCCGCACCCACTATCTTCATATTGTT

                                             3700
                                              •
CGTCTACTTTGCGCGGTCTGGCAGACTTCTGTGGAAGTTGGGGCACATAGGTATACTGTGTCTTTGGCCCGGAGGT
GCAGATGAAACGCGCCAGACCGTCTGAAGACACCTTCAACCCCGTGTATCCATATGACACAGAAACCGGGCCTCCA
   MetLysArgAlaArgProSerGluAspThrPheAsnProValTyrProTyrAspThrGluThrGlyProPro
   Fiber
                   3750                                              3800
                    •                                                 •
TGACACGGGAAAGAATGGGGAGGTAAACAAAGTGGGTTACCAAAGGTTCTTTCAGGGGGACCTCAAGAGAGAGAT
ACTGTGCCCTTTCTTACCCCTCCATTTGTTTCACCCAATGGTTTCCAAGAAAGTCCCCCTGGAGTTCTCTCTCTA
ThrValProPheLeuThrProProPheValSerProAsnGlyPheGlnGluSerProProGlyValLeuSerLeu

                                            3850
                                             •
GCGCAGAGGCTTGGAAACCTGTGGAGGGTGCCGTACGAACGCGAATTTTACCCGTCGCCAGAATGGGATCTGTTC
CGCGTCTCCGAACCTTTGGACACCTCCCACGGCATGCTTGCGCTTAAAATGGGCAGCGGTCTTACCCTAGACAAG
ArgValSerGluProLeuAspThrSerHisGlyMetLeuAlaLeuLysMetGlySerGlyLeuThrLeuAspLys

                   3900                                              3950
                    •                                                 •
CGGCCTTTGGAGTGGAGGGTTTTACATTGGTGACAATGAGTCGGTGAATTTTTTTGTTTCAGTTTGTATTCAAAC
GCCGGAAACCTCACCTCCCAAAATGTAACCACTGTTACTCAGCCACTTAAAAAAACAAAGTCAAACATAAGTTTG
AlaGlyAsnLeuThrSerGlnAsnValThrThrValThrGlnProLeuLysLysThrLysSerAsnIleSerLeu

                                            4000
                                             •
CTGTGGAGGCGTGGTGAATGTTAATGGAGTCCGCGGGATTGTCACCGTTGGTGGCGAGGAGACTATCAATGATCG
GACACCTCCGCACCACTTACAATTACCTCAGGCGCCCTAACAGTGGCAACCACCGCTCCTCTGATAGTTACTAGC
AspThrSerAlaProLeuThrIleThrSerGlyAlaLeuThrValAlaThrThrAlaProLeuIleValThrSer

                   4050                                              4100
                    •                                                 •
CCGCGAGAATCGCATGTCAGTGTTCGGGGTGACTGGCACGTTCTGAGGTTTGATTCGTAACGATGATTTCCCGGG
GGCGCTCTTAGCGTACAGTCACAAGCCCCACTGACCGTGCAAGACTCCAAACTAAGCATTGCTACTAAAGGGCCC
GlyAlaLeuSerValGlnSerGlnAlaProLeuThrValGlnAspSerLysLeuSerIleAlaThrLysGlyPro

                                            4150
                                             •
TAATGTCACAGTCTACCTTTCGATCGGGACGTTTGTAGTCGGGGGAGAGACCGTCACTGTCGCTGTGGGAATGA
ATTACAGTGTCAGATGGAAAGCTAGCCCTGCAAACATCAGCCCCCTCTCTGGCAGTGACAGCGACACCCTTACT
IleThrValSerAspGlyLysLeuAlaLeuGlnThrSerAlaProLeuSerGlySerAspSerAspThrLeuThr

                   4200                                              4250
                    •                                                 •
CATTGACGTAGTGGGGGCGATTGATGACGGTGCCCATCGAACCCGTAATTGTACCTTCTAGGATAAATACATTTA
GTAACTGCATCACCCCCGCTAACTACTGCCACGGGTAGCTTGGGCATTAACATGGAAGATCCTATTTATGTAAAT
ValThrAlaSerProProLeuThrThrAlaThrGlySerLeuGlyIleAsnMetGluAspProIleTyrValAsn

                                            4300
                                             •
TTACCTTTTTATCCTTAATTTTATTCGCCAGGAAACGTTCATCGTGTTTTGAGGCTATGTGATTGTCATCAATGA
AATGGAAAAATAGGAATTAAAATAAGCGGTCCTTTGCAAGTAGCACAAAACTCCGATACACTAACAGTAGTTACT
AsnGlyLysIleGlyIleLysIleSerGlyProLeuGlnValAlaGlnAsnSerAspThrLeuThrValValThr

                   4350                                              4400
                    •                                                 •
CCTGGTCCACAGTGGCAACTTGTTTTGAGGGAATCTTGGTTTCAACGTCCTCGATAACCAATACTAAGTAGTTTG
GGACCAGGTGTCACCGTTGAACAAAACTCCCTTAGAACCAAAGTTGCAGGAGCTATTGGTTATGATTCATCAAAC
GlyProGlyValThrValGluGlnAsnSerLeuArgThrLysValAlaGlyAlaIleGlyTyrAspSerSerAsn
```

```
                                                 4450
                                                   •
TTGTACCTTTAATTTTGCCCGCCACCGTACGCATATTTATTGTTGAACAATTAAGATCTACACCTAATGGGTAAA
AACATGGAAATTAAAACGGGCGGTGGCATGCGTATAAATAACAACTTGTTAATTCTAGATGTGGATTACCCATTT
AsnMetGluIleLysThrGlyGlyGlyMetArgIleAsnAsnAsnLeuLeuIleLeuAspValAspTyrProPhe

                   4500                                                 4550
                     •                                                    •
CTACGAGTTTGTTTTGATGCAGAATTTGACCCCGTCCCTGGGGACATATAATTACGTAGAGTATTGAACCTGTAT
GATGCTCAAACAAAACTACGTCTTAAACTGGGGCAGGGACCCCTGTATATTAATGCATCTCATAACTTGGACATA
AspAlaGlnThrLysLeuArgLeuLysLeuGlyGlnGlyProLeuTyrIleAsnAlaSerHisAsnLeuAspIle

                                                 4600
                                                   •
TTGATATTGTCTCCGGATATGGAAAAATTACGTAGTTTGTTATGATTTTTTGACCTTCAATCGTATTTTTTTAGG
AACTATAACAGAGGCCTATACCTTTTTAATGCATCAAACAATACTAAAAAACTGGAAGTTAGCATAAAAAAATCC
AsnTyrAsnArgGlyLeuTyrLeuPheAsnAlaSerAsnAsnThrLysLysLeuGluValSerIleLysLysSer

                   4650                                                4700
                     •                                                    •
TCACCTGATTTGAAACTATTATGACGGTATCGATATTTACGTCCTTTCCCAGACCTCAAACTATGTTTGTGTAGA
AGTGGACTAAACTTTGATAATACTGCCATAGCTATAAATGCAGGAAAGGGTCTGGAGTTTGATACAAACACATCT
SerGlyLeuAsnPheAspAsnThrAlaIleAlaIleAsnAlaGlyLysGlyLeuGluPheAspThrAsnThrSer

                                                 4750
                                                   •
CTCAGAGGTCTATAGTTGGGTTATTTTTGATTTTAACCGAGACCGTAACTAATGTTACTTTTGCCACGGTACTAA
GAGTCTCCAGATATCAACCCAATAAAAACTAAAATTGGCTCTGGCATTGATTACAATGAAAACGGTGCCATGATT
GluSerProAspIleAsnProIleLysThrLysIleGlySerGlyIleAspTyrAsnGluAsnGlyAlaMetIle

                   4800                                                 4850
                     •                                                    •
TGATTTGAACCTCGCCCAAATTCGAAACTGTTGAGTCCCCGGTAATGTTATCCTTTGTTTTTACTACTGTTTGAA
ACTAAACTTGGAGCGGGTTTAAGCTTTGACAACTCAGGGGCCATTACAATAGGAAACAAAAATGATGACAAACTT
ThrLysLeuGlyAlaGlyLeuSerPheAspAsnSerGlyAlaIleThrIleGlyAsnLysAsnAspAspLysLeu


TGGGACACCTGTTGGGGTCTGGGTAGAGGATTGACGTCTTAA r chain
ACCCTGTGGACAACCCCAGACCCATCTCCTAACTGCAG 3'  l chain
ThrLeuTrpThrThrProAspProSerProAsnCysArgIle
```

Fig.2 : Ad.2 EcoRI E nucleotide sequence. The theoretical amino-acid sequences corresponding to the 14,5K, 14K and Fiber proteins are indicated. r and l stand for rightward and leftward transcribed chains.

of the fiber protein. Therefore according to a molecular weight of 62 000 daltons i.e approximately 560 aminoacids, 80% of the fiber protein would be encoded within fragment EcoRI E. Translation of this coding region shows the existence of 6 glycosylation sites (Asn - X - Ser/Thr) evenly distributed (28). This high proportion of glycosylation sites may be in relation with the antigenic properties of the fiber protein as often observed with viral enveloppe proteins.

From nucleotide sequences analysis of the region surrounding the splice point no single sequence emerges. Nevertheless some features such as the presence of a GT and AG at the 5' and 3' end of intron seem to be a

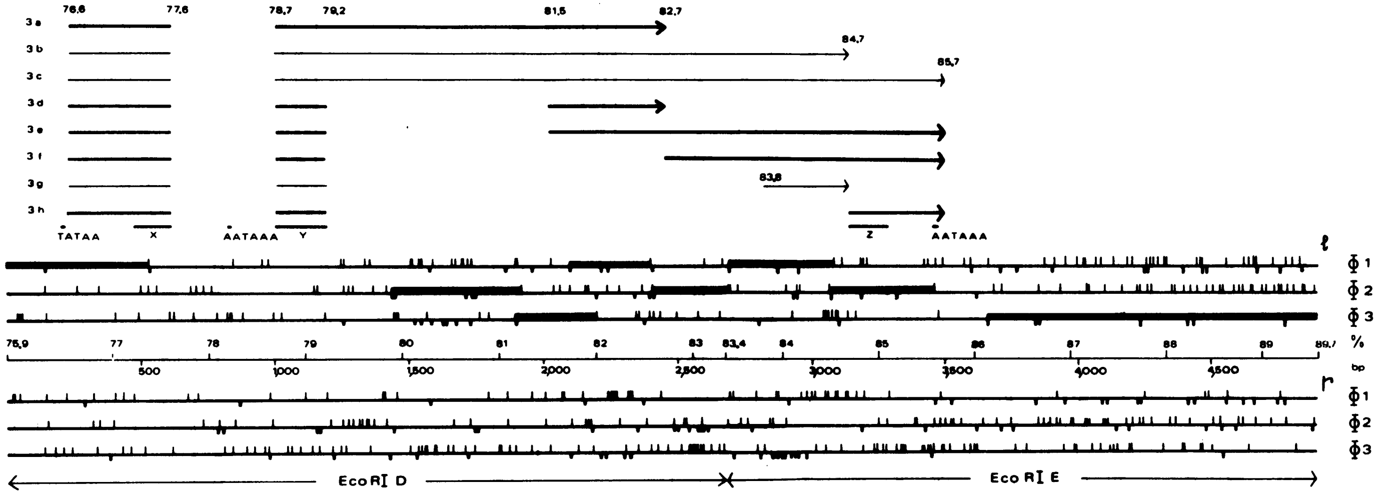

Fig.3 : Diagram showing the localisation of initiator and stop codons within the EcoRI D and E fragments. The position of the 3a to 3h mRNA and that of the additional x, y and z late leaders of the Fiber mRNA are from Chow et al (14). TATAA stands for the 5' end of the early 3 mRNA , while the two sequences AATAAA at residue number 796 and 3470 correspond respectively to the 3' end of the late mRNA ending at 78, and to the 3' end of various early 3 mRNA . ϕ1, ϕ2 and ϕ3 correspond to the different reading frames as defined in the text. Upper vertical bars are for the nonsense codons. Underneath vertical bars correspond to the ATG triplets. E3 mRNA are transcribed from the r strand (rightward transcribed strand). They correspond in sequence and polarity to the leftward strand. Therefore, open reading frames corresponding to these mRNA are indicated by thick lines on the l strand.

rather general rule (29,30). Moreover sequences more or less identical to CAGTTT and GGTGAG have been found at the border of several adenovirus 2 leaders (21,27,31,32). More recently it has been observed that nucleotide sequences at the border of the intron could be paired with the 5' terminal sequence of the U1 nuclear RNA suggesting a mechanism for the splicing (33). Between coordinates 84 and 85 where the additional Z leader has been mapped, several potential splicing sequences can be detected. Taking into account the length of the Z leader which has been estimated to be 0.4% of the genome length, two sets of sequences shown in fig.4 seem to have a better chance of being the actual Z leader splicing sequences. Chow et al have placed by EM the Z leader at the 5' end of the 3 h mRNA body (14). As it will be described later on this mRNA could correspond to the open reading frame starting with residue number 3064. This reinforces the hypothetical placement of the Z leader between nucleotide 3066 and 3225.

- Early region 3

This region codes for proteins which are non essential for the replication (20) of the virus. By in vitro translation of early mRNA hybridizing to EcoRI D and E fragments three proteins with a molecular weight of 16K, 14.5 and 14K have been obtained. By electron microscopy experiments, it has

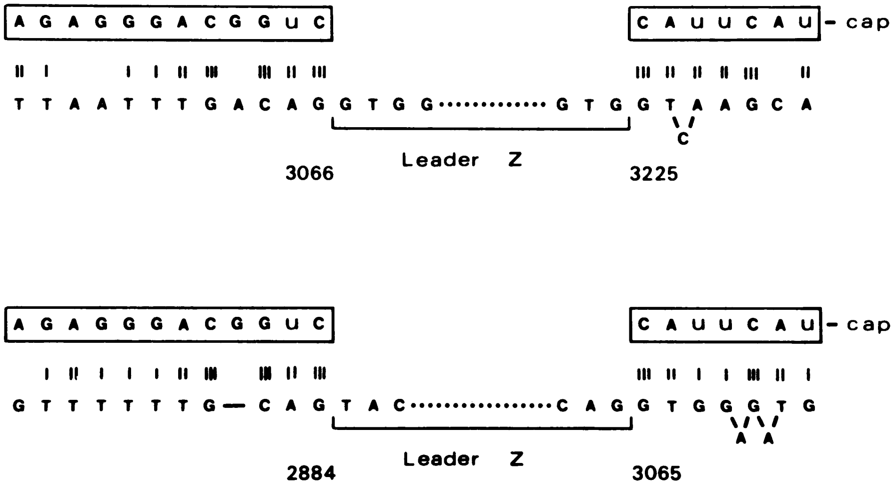
A G A G G G A C G G U C
C A U U C A U - cap
T T A A T T T G A C A G G T G G ············ G T G G T A A G C A
C
Leader Z
3066
3225
A G A G G G A C G G U C
C A U U C A U - cap
G T T T T T T G — C A G T A C ················ C A G G T G G G T G
A A
Leader Z
2884
3065

Fig.4 : Potential splicing sequences for the Z leader. Boxed sequences correspond to the 5' terminal sequence of the U1 RNA. EcoRI E nucleotide sequences resembling to the nucleotide sequences more often found around splice points are tentatively paired with the U1 RNA sequence.

been shown that E3 mRNA end up at coordinates 82.7, 84.7 or 85.7 (14). The first value falls within the EcoRI D fragment, the two others within the EcoRI E fragment. An AATAAA sequence is present at nucleotide 3470, coordinate 85.6, positioning without ambiguity the 3' end of 3c, e, f and h mRNA (14,34-36). On the contrary no such sequence is found elsewhere in the EcoRI D and E fragments apart from the AATAAA sequence which belongs to the mRNA of the L4 family. Could it be possible that the 3a, b, d and g messages (14) have a particular structure without the AATAAA sequence, a situation which has already been observed with the human hepatitis B virus (37) ? Another explanation would be an additional splicing between the 3' observed ends of 3a, b, d, and g messenger RNA and a short strand of RNA transcribed from region 85.7. The short size of this last exon could explain why it was not observed by EM heteroduplex analysis.

From the nucleotide sequence of the EcoRI D fragment (11), we have suggested that 3a mRNA would code for the 16K protein (20) either from ATG_{1440} or $_{1449}$ to end up with TGA_{1917} (11). By in vitro protein synthesis in the presence of various radio labelled aminoacids, H. Persson et al (38) have confirmed that 3a mRNA directed the synthesis of the 16K protein from ATG_{1440} (11). They have also shown that 3h mRNA mainly selected by hybridization to EcoRI E fragment and whose body maps at coordinates 84.7-85.7 would code for a 14K protein (38). Upstream $AATAAA_{3470}$ (coordinate : 85.7), a sequence, open in reading frame 2 could be translated from ATG_{3072} up to TAA_{3456} into a protein of 128 aminoacids with a molecular weight of 14 762 daltons. The very good correlation between the map coordinates of the body of 3h mRNA and the map coordinates of this open reading frame, as well as the good correlation between the observed and calculated molecular weight strongly suggest that the 14K protein is indeed coded by this open reading sequence. The primary structure of this protein, as it can be deduced from the nucleotide sequence, reveals a fairly large number of lysine and arginine residues able to give rise, by trypsin digestion, to a large number of peptides which could be useful for characterizing this gene product. The aminoacid sequence does not exhibit any glycosylation site (28).

A third protein synthesized in vitro by the E3 mRNA complex is called 14.5K (39). Its synthesis is directed by an mRNA which is mainly selected by hybridization to the EcoRI E fragment (38). This RNA could then be either the 3e, f or g mRNA (14). Overlapping these mRNA one can find in the l strand (the antisense strand) several open reading frames (fig.3). The largest starts with ATG_{2687} and goes up to TGA_{3077}. It could code for a polypeptide of 130 aminoacids with a molecular weight of 14 529. It is therefore tempting to suggest this protein exists and corresponds to the 14.5K protein. The 14.5K protein is a minor in vivo product (38). In accordance with this, it could be synthesized by translation of 3g mRNA which is also a minor species (14), the map coordinates of which coincide with that of this open reading frame. Three other sequences free of stop codons are overlapped by the body of E3 mRNA. These sequences are defined as follows : (I) from ATG_{1900} to TGA_{2197} in frame 3 (II) from ATG_{2096} to TAA_{2399} in frame 1 (III) from ATG_{2409} to TGA_{2682} in frame 2 (see fig.3). They could respectively code for a polypeptide 99, 101 and 91 aminoacids long. Their calculated molecular weights do not favor the idea that one of them corresponds to the 14.5K protein, reinforcing the localization of the reading frame for the latter at ATG_{2687}. On the other hand these open reading frames could correspond to the 3d, e and f mRNA.

Chow et al have located the 5' end of the body of 3d and 3e mRNA at coordinate 81.5 (14). However, Kitchingman and Westphal located the 5' end of these mRNA at a slightly different position, one at 80.1, and the other at 81.3 (40). If this is true, the various bodies of E3 mRNA would start at six different positions along the genome, corresponding respectively to the 3a, d, e, f, g and h messengers (14). Therefore it is striking to observe within the DNA sequence six different open reading frames in which the position of the first ATG coincides very well with the beginning of the body of these mRNA (11, fig.3). From this, we would like to suggest that the E3 region could code for six different proteins instead of the three usually observed (39,41). The three postulated additional proteins, which would have a molecular weight of 11K, would have been missed after a one dimentional electrophoresis gel, because of the large amount of globin present in the reticulocyte system.

In agreement with this hypothesis, we would like to suggest that the first ATG found in the body of the various mRNA codes for the N terminal methionine, relegating the control of the expression of the message within the first or second leader. This hypothesis is further substantiated by the existence of a potential binding site for the ribosomes, within the second early leader (11), and the suggested existence of an additional splicing event for the 3a mRNA (11,40), eliminating an intron sequence between residue numbers 1188 and 1410 (11), and consequently ATG_{1258}.

- Coding capacity of the leftward strand

Hybridization experiments have suggested that no leftward transcripts are made, apart from the leaders of the 72K mRNA, between coordinate 91, the 3' end of the E4 mRNA, and coordinate 66.5 where the body of the 72K mRNA begins (1,14). The distribution of the nonsense codons and ATG triplets in the r strand from coordinate 89.7, the right end of the E fragment down to 70.7 the left end of the F fragment suggests that the l strand has a very limited coding capacity between these coordinates (10,11, fig.3). Nevertheless the presence of three regions, open from ATG_{4232}, $_{3839}$ and $_{3184}$ and closed respectively with TGA_{3920}, TAA_{3482} and TAG_{2899} indicates that the leftward strand could code for proteins of 11K, 13K and 10K.

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

ACKNOWLEGMENT

We thank Stéphane Dupont de Dinechin for expert technical assistance. This work was supported by Grants from CNRS and INSERM.

REFERENCES

1. Chow L.T., Roberts J.M., Lewis J.B. and Broker T.R. (1977) Cell 11, 819-836.
2. Berk A.J. and Sharp P.A. (1977) Cell 12, 721-732.
3. Perricaudet M., Akusjärvi G., Virtanen A. and Pettersson U. (1979) Nature 281, 694-696.
4. Perricaudet M., Le Moullec J.M., Tiollais P. and Pettersson U. (1980) Nature 288, 174-176.
5. Van Ormondt H., Maat J., De Waard A. and Van der Eb A.J. (1978) Gene 4, 309-328.
6. Maat J. and Van Ormondt H. (1979) Gene 6, 75-90.
7. Dijkema R., Dekker B.M.M. and Van Ormondt H. (1980) Gene 9, 141-156.
8. Maat J., Van Beveren C.P. and Van Ormondt H. (1980) Gene 10, 27-38.
9. Akusjarvi G. and Pettersson U. (1978) Virology 91, 477-480.
10. Galibert F., Hérissé J. and Courtois G. (1979) Gene 6, 1-22.
11. Hérissé J., Courtois G. and Galibert F. (1980) Nucleic Acids Res. 8, 2173-2192.
12. Lewis J.B. and Mathews M.B. (1980) Cell 21, 303-313.
13. Flint J. (1977) Cell 10, 153-166.
14. Chow L.T., Broker T.R. and Lewis J.B. (1979) J. Mol. Biol. 134, 265-303.
15. Fiers W., Contreras R., Hageman G., Rogiers R., Vande Woorde A., Van Henverswyn H., Van Herreweghe J., Volckaert G. and Ysetaert M. (1978) Nature 273, 113-120.
16. Reddy V.B., Thimmappaya B., Dhar R., Subramanian K.N., Zain B.S., Pan J., Ghosh P.K., Celma M.L., Weissman S.M. (1978) Science 200, 494-502.
17. Berk A.J. and Sharp P.A. (1978) Cell 14, 695-711.
18. Nevins J.R. and Darnell J.E. (1978) J. Virol. 25, 811-823.
19. Chow L.T. and Broker T.R. (1978) Cell 15, 497-510.
20. Ross S.Z. and Levine A.J. (1979) Virology 99, 427-430.

21. Baker C.C., Hérissé J., Courtois G., Galibert F. and Ziff E. (1979) Cell 18, 569-580.
22. Hérissé J., Courtois G. and Galibert F. (1978) Gene 4, 279-294.
23. Fraser N. and Ziff E. (1978) J. Mol. Biol. 124, 27-51.
24. Maxam A.M. and Gilbert W. (1977) Proc. Natl. Acad. Sci. U.S.A. 74, 560-564.
25. Maxam A.M. and Gilbert W. (1980) In Methods in Enzymology 65, part I, 499-560.
26. Zain Sayeeda B. and Roberts R.J. (1979) J. Mol. Biol. 131, 341-352.
27. Zain Sayeeda B., Sambrook J., Robert R.J., Keller W., Fried M. and Dunn A.R. (1979) Cell 16, 851-861.
28. Struck D.K., Lennartz W.J. and Brew K. (1978) J. Biol. Chem. 253, 5786-5794.
29. Breathnach R., Benoist C., O'Hare K., Gannon F. and Chambon P. (1978) Proc. Natl. Acad. Sci. U.S.A. 75, 4853-4857.
30. Catterall J.F., O'Malley W.O., Robertson M.A., Staden R., Tanaka R. and Brownlee G.G. (1978) Nature 275, 510-513.
31. Akusjarvi G. and Petterson U. (1979) J. Mol. Biol. 134, 143-158.
32. Ziff E. and Evans R. (1978) Cell 15, 1463-1475.
33. Avvedimento V.E., Vogeli G., Yamada Y., Maizel J.V., Ira Pastan Jr.V. and Benoit de Crombrugghe B. (1980) Cell 21, 689-696.
34. 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.
35. Efstratiadis A. and Kafatos F.C. (1977) Cell 10, 571-585.
36. Proudfoot N.J. (1977) Cell 10, 559-570.
37. Galibert F., Mandart E., Fitoussi F., Tiollais P. and Charnay P. (1979) Nature 281, 646-650.
38. Persson H., Jörnvall H. and Zabielski J. (1980) Proc. Natl. Acad. Sci. U.S.A., in press.
39. Harter M.L. and Lewis J.B. (1978) J. Virol. 26, 736-749.
40. Kitchingman G.R. and Westphal H. (1980) J. Mol. Biol. 137, 23-48.
41. Persson H., Jansson M. and Philipson L. (1980) J. Mol. Biol. 136, 375-394.