Volume 9 Number 16 1981 Nucleic Acids Research

Nucleotide sequence of adenovirus 2 DNA fragment encoding for the carboxylic region of the fiber protein and the entire E4 region

J.Hérissé, M.Rigolet, S.Dupont de Dinechin and F.Galibert

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

Received 1 July 1981

SUMMARY

The entire nucleotide sequence between coordinates 89.5 and 100% of the Ad 2 DNA genome has been determined using the Maxam and Gilbert method. This sequence of 3766 bp contains information relative to the carboxylic end of the fiber protein and to the entire E4 region.

The position within the nucleotide sequence of various open reading frames and of several consensus splicing sequences was correlated with the location by EM and S1 digestion of the E4 mRNA. This correlation allows to suggest an additional splicing event in the maturation process of i or f mRNA and to deduce the structure of most E4 mRNA. The aminoacid sequences of the corresponding proteins are deduced allowing the location of several glycosylation sites.

The presence of several open reading frames with a subtantial coding capacity permits to postulate on the existence of additional genes located at the 3' end of the fiber gene and the 3' end of the E4 region. The existence of these putative additional genes might explain that termination of transcription is several hundred nucleotides beyond the main known poly A addition sites of the L5 and E4 regions.

INTRODUCTION

Lytic infection of human cells by adenovirus proceeds through a cycle conveniently divided into two periods separated by the onset of viral DNA synthesis. Before viral DNA replication, at least five DNA regions are transcribed into mRNA that code for the early viral proteins (1-7). Viral messenger RNA corresponding to the early region E1A, E1B and E3 are transcribed on the r strand, mRNA of the E2 and E4 being transcribed from the l strand. At late time after infection, other transcripts are made, mainly from the r strand. By a very complicated pattern of splicing and other maturation processes these transcripts give rise to numerous mRNA identified by electron microscopy mapping and in vitro protein synthesis (8,9). However evidence for a more complex temporal pattern of transcription is derived from experiments showing that region E1B, E2 and E3 are still actively transcribed at intermediate time while the others become silent.

Further evidence is also provided by experiments showing that protein IX mRNA - which has a 3' terminus that coincides with that of the E1B mRNA - is made at the end of the early phase from a promotor different to that of the E1B region (6,10). Moreover it was recently shown that parts of the DNA sequence coding for late protein are already transcribed at early time (11).

Several years ago the analysis of the nucleotide sequence of the adenovirus 2 genome was undertaken, in order to draw a detailed functional map of the adenovirus genome where the various mRNA could be precisely located and their coding region delineated, and where regulatory sequences involved in the complex splicing mechanism utilized during the synthesis of the different viral mRNA might also be identified. During this work the EcoRI F, D and E fragments which map between coordinates 70.7 and 89.7 and cover the entire E3 region, the first leader of the E2 region, the 3' end of the L4 RNA family and 80% of the fiber mRNA were fully analyzed (12-15). The nucleotide sequence elucidated was then used to tentatively map the early mRNA corresponding to the 16, 14,5 and 14K E3 proteins and the late mRNA corresponding to the 100K, 33K, pVIII and fiber proteins.

In the present paper we report the nucleotide sequence of the remaining part of the right hand end of the genome, from coordinates 89.5 to 100%. This nucleotide sequence includes the entire E4 region and the 3' end of the fiber messenger RNA from which the carboxylic end of the protein can be deduced.

MATERIALS AND METHODS

All materials used were as previously described (12,16).

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

Cloning of the HindIII F fragment and propagation of the recombinant : Viral DNA was digested with HindIII endonuclease and the resulting fragments were fractionated by electrophoresis on agarose gel. Because of their identical size HindIII F and G fragments were eluted together from the agarose gel and both subjected to ligation. Eluted viral DNA fragments were ligated with T4 ligase to pBR 322 DNA digested with HindIII enzyme (18). E. coli strain $C_{600}Rk^{-}Mk^{+}$ made competent by $CaCl_2$ treatment (19) was transfected with the ligated DNA and ampicillin resistant - tetracyclin sensitive clones were selected. The recombinant DNA plasmid harbored by several bacterial clones was characterized by restriction mapping including digestion with HindIII, EcoRI, XbaI and HpaI. Propagation of the selected bacterial clone, extraction and purification of the plasmid DNA were done as previously described (20,21).

Preparation of the BglII viral DNA fragment : Viral DNA was fully digested with BglII restriction enzyme and fractionated by electrophoresis on agarose gel. BglII H fragment was eluted by electrophoresis and further purified from agarose gel contaminants by chromatography on hydroxylapatite. It was then used as starting material for nucleotide sequence analysis of the right hand end of the viral DNA genome.

Sequencing procedure : Sequence analysis was performed according to the method of Maxam and Gilbert (22,23). Five chemical reactions specific for G, AG, CT, C and AC were currently done, and fractionated on 25, 16 and 8% acrylamide gels. Sequencing gels were 0.8 mm thick and 400 or 800 mm long.

RESULTS

The two 5' ends of the adenovirus 2 DNA are covalently linked to a terminal protein (24,25). This linkage prevents the labelling of the DNA by the polynucleotide kinase and P^{32} ATP (26) and may alter the cloning of any terminal fragment within pBR 322. Therefore to analyze the nucleotide sequence at the right hand side of the EcoRI 89.7 site (without making too many viral preparations), 1) the HindIII F fragment (map coordinate 89.5-97.3) was cloned in pBR 322 and the cloned fragment used as starting material ; 2) the BglII H fragment (map coordinates 96-100%) was used directly, without cloning, as starting material for sequence analyses. Since these two fragments overlap each other and the HindIII F fragment overlaps the EcoRI site located between fragment EcoRI E and C, the absence of additional very small EcoRI fragments in between can be ascertained.

The nucleotide sequence of the HindIII F and BglII H fragments was determined using the chemical degradation method of Maxam and Gilbert on various restriction fragment subsets (22,23), as shown in fig.1. Due to the large number of restriction fragments used, the sequence on both chains of the HindIII F and BglII H fragments - except for the first three hundred nucleotides on chain r, whose 5' end could not be labelled with polynucleotide kinase (26) - could be derived independently.

The nucleotide sequence of the HindIII F and BglII H fragments is shown in fig.2. This sequence is made up of 3766 nucleotides, and according to previous results concerning the sequencing of the EcoRI E fragment (15), is

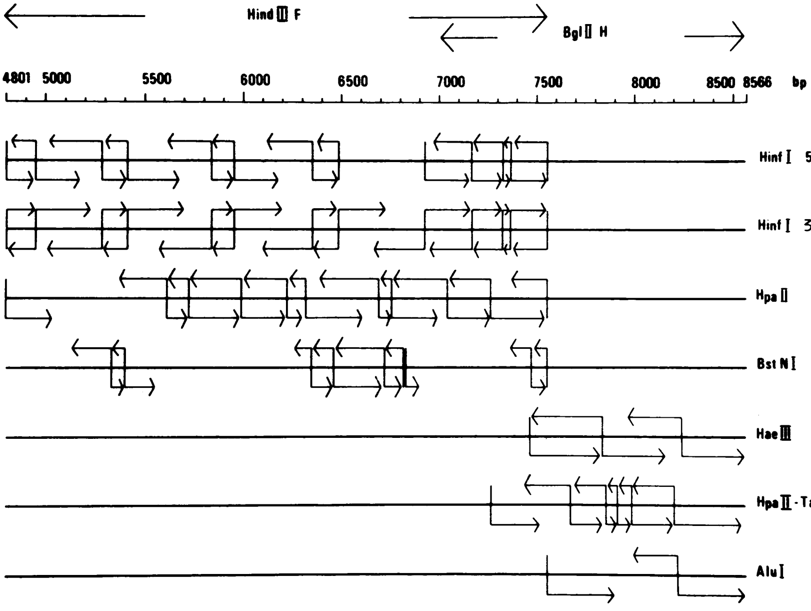

Fig.1 : Diagram of analysed DNA fragments. Vertical bars correspond to the position of the labelled end of restriction fragments used to determine the nucleotide sequence. Length of arrows is representative to the number of nucleotides analysed from a restriction site.

numbered from 4801 to 8566. Therefore the total length of the F, D, E (12,14,15) and C EcoRI fragments which accounts for 29.3% of the total adenovirus 2 genome is 10305 bp and thus 1% is equal to 351 in agreement with previous results.

A large number of A and T residues can be observed in this sequence at position 95.6 where there is a structure of 13A on chain r, or at position 97.9 where there are 28Ts out of 37 residues.

Previous sequence data on the right-hand extremity of the Ad 2 and Ad 5 DNA genome have been published by J. Arrand et al concerning the 103 bp

Fig.2 : Nucleotide sequence from coordinates 89.5 to 100% . The theoretical aminoacid sequence corresponding to the fiber protein and the various open reading frames of the E4 region are indicated. Putative glycosylation sites are underlined. Note that there is a Leucine indicated by a star at the N terminal position of region 5. This aminoacid would be coded by ligation of region 2 to region 5 i.e. T 7359 to TA 6628.

```
4801                                          4850
.        .        .         .        .         .        .         .
r 3'AACTGTTGAGTCCCCGGTAATGTTATCCTTTGTTTTTACTACTGTTTGAATGGGACACCTGTTGGGGTCTG
AGCTTTGACAACTCAGGGGCCATTACAATAGGAAACAAAAATGATGACAAACTTACCCTGTGGACAACCCCAGAC
SerPheAspAsnSerGlyAlaIleThrIleGlyAsnLysAsnAspAspLysLeuThrLeuTrpThrThrProAsp

                      4900                                           4950
    .         .        .         .        .        .         .        .
GGTAGAGGATTGACGTCTTAAGTAAGTCTATTACTGACGTTTAAATGAAACCAAGAATGTTTTACACCCTCAGTT
CCATCTCCTAACTGCAGAATTCATTCAGATAATGACTGCAAATTTACTTTGGTTCTTACAAAATGTGGGAGTCAA
ProSerProAsnCysArgIleHisSerAspAsnAspCysLysPheThrLeuValLeuThrLysCysGlySerGln

                                              5000
         .        .         .        .         .        .         .
CATGATCGATGACATCGACGAAACCGACATAGACCTCTAGAAAGTAGGTACTGTCCGTGGCAACGTTCACAATCA
GTACTAGCTACTGTAGCTGCTTTGGCTGTATCTGGAGATCTTTCATCCATGACAGGCACCGTTGCAAGTGTTAGT
ValLeuAlaThrValAlaAlaLeuAlaValSerGlyAspLeuSerSerMetThrGlyThrValAlaSerValSer

                      5050                                           5100
    .         .        .         .        .        .         .        .
TATAAGGAATCTAAACTGGTTTTGCCACAAGATTACCTCTTGAGGAGTGAATTTTTTGTAATGACCTTGAAATCT
ATATTCCTTAGATTTGACCAAAACGGTGTTCTAATGGAGAACTCCTCACTTAAAAAACATTACTGGAACTTTAGA
IlePheLeuArgPheAspGlnAsnGlyValLeuMetGluAsnSerSerLeuLysLysHisTyrTrpAsnPheArg

                                              5150
         .        .         .        .         .        .         .
TTACCCTTGAGTTGATTACGTTTAGGTATGTGTTTACGTCAACCTAAATACGGATTGGAAGATCGGATAGGTTTT
AATGGGAACTCAACTAATGCAAATCCATACACAAATGCAGTTGGATTTATGCCTAACCTTCTAGCCTATCCAAAA
AsnGlyAsnSerThrAsnAlaAsnProTyrThrAsnAlaValGlyPheMetProAsnLeuLeuAlaTyrProLys

                      5200                                           5250
    .         .        .         .        .        .         .        .
TGGGTTTCAGTTTGACGATTTTTATTGTAACAGTCAGTTCAAATGAACGTACCACTATTTTGATTTGGATACTAT
ACCCAAAGTCAAACTGCTAAAAATAACATTGTCAGTCAAGTTTACTTGCATGGTGATAAAACTAAACCTATGATA
ThrGlnSerGlnThrAlaLysAsnAsnIleValSerGlnValTyrLeuHisGlyAspLysThrLysProMetIle

                                              5300
         .        .         .        .         .        .         .
GAATGGTAATGTGAATTACCGTGATCACTTAGGTGTCTTTGATCGCTCCATTCGTGAATGAGATACAGAAAATGT
CTTACCATTACACTTAATGGCACTAGTGAATCCACAGAAACTAGCGAGGTAAGCACTTACTCTATGTCTTTTACA
LeuThrIleThrLeuAsnGlyThrSerGluSerThrGluThrSerGluValSerThrTyrSerMetSerPheThr

                      5350                                           5400
    .         .        .         .        .        .         .        .
ACCAGGACCCTTTCACCTTTTATGTGGTGACTTTGAAAACGATGGTTGAGAATGTGGAAGAGGATGTAACGGGTC
TGGTCCTGGGAAAGTGGAAAATACACCACTGAAACTTTTGCTACCAACTCTTACACCTTCTCCTACATTGCCCAG
TrpSerTrpGluSerGlyLysTyrThrThrGluThrPheAlaThrAsnSerTyrThrPheSerTyrIleAlaGln

                                              5450
         .        .         .        .         .        .         .
CTTATTTCTTAGCACTTGGACAACGTACAATACAAAGTTGCACAAATAAAAAGTTAACGTCTTTTAAAGTTCAGT
GAATAAAGAATCGTGAACCTGTTGCATGTTATGTTTCAACGTGTTTATTTTTCAATTGCAGAAAATTTCAAGTCA
Glu COOH

                      5500                                           5550
    .         .        .         .        .        .         .        .
                                                   Region 7 COOH LeuVal
AAAAAGTAAGTCATCATATCGGGGTGGTGGTGTATCGAATATAACTAGTGGCATGGAATTAGTTTGAGTGTCTTG
TTTTTCATTCAGTAGTATAGCCCCACCACCACATAGCTTATATTGATCACCGTACCTTAATCAAACTCACAGAAC
```

```
                                                5600
         .         .         .         .         .         .         .
ArgThrAsnLeuArgGlyGlyGlyGluTrpCysValSerTyrValThrArgGluGlyArgSerAlaLysPheLeu
GGATCATAAGTTGGACGGTGGAGGGAGGGTTGTGTGTCTCATGTGTCAGGAAAGAGGGGCCGACCGGAATTTTTC
CCTAGTATTCAACCTGCCACCTCCCTCCCAACACACAGAGTACACAGTCCTTTCTCCCCGGCTGGCCTTAAAAAG

                       5650                                              5700
    .         .         .         .         .         .         .         .
MetMetAspHisThrValSerMetAsnLysProThrIleAsnTrpValThrGluGlnArgAlaLeuArgGluAsp
GTAGTATAGTACCCATTGTCTGTATAAGAATCCACAATATAAGGTGTGCCAAAGGACAGCTCGGTTTGCGAGTAG
CATCATATCATGGGTAACAGACATATTCTTAGGTGTTATATTCCACACGGTTTCCTGTCGAGCCAAACGCTCATC

                                                5750
         .         .         .         .         .         .         .
ThrIleAsnIlePheGluGlyProLeuGluSerLeuAsnMetAspSerAspLeuGlnGlnAlaValProGlnGln
TCACTATAATTATTTGAGGGGCCCGTCGAGCGAATTCAAGTACAGCGACAGGTCGACGACTCGGTGTCCGACGAC
AGTGATATTAATAAACTCCCCGGGCAGCTCGCTTAAGTTCATGTCGCTGTCCAGCTGCTGAGCCACAGGCTGCTG

                       5800                                              5850
    .         .         .         .         .         .         .         .
                                  Region 6 COOH MetProThrSerAspTyrAspHisMet
GlyValGlnProGlnGluValProProSerProSerThrTrpAla NH2
AGGTTGAACGCCAACGAGTTGCCCGCCGCTTCCCCTTCAGGTGCGGATGTACCCCCATCTCAGTATTAGCACGTA
TCCAACTTGCGGTTGCTCAACGGGCGGCGAAGGGGAAGTCCACGCCTACATGGGGGTAGAGTCATAATCGTGCAT

                                                5900
         .         .         .         .         .         .         .
LeuIleProArgHisHisGlnLeuLeuAlaArgIlePheGlnGlnArgArgArgGluThrArgCysSerTyrLeu
GTCCTATCCCGCCACCACGACGTCGTCGCGCGCTTATTTGACGACGGCGGCGGCGAGGCAGGACGTCCTTATGTT
CAGGATAGGGCGGTGGTGCTGCAGCAGCGCGCGAATAAACTGCTGCCGCCGCCGCTCCGTCCTGCAGGAATACAA

                       5950                                              6000
    .         .         .         .         .         .         .         .
MetAlaThrThrGluGluAlaIleIleArgValAlaArgLeuMetLeuArgArgThrArgArgAlaCysCysArg
GTACCGTCACCAGAGGAGTCGCTACTAAGCGTGGCGGGCGTCGTACTCTGCGGAACAGGAGGCCCGTGTCGTCGC
CATGGCAGTGGTCTCCTCAGCGATGATTCGCACCGCCCGCAGCATGAGACGCCTTGTCCTCCGGGCACAGCAGCG

                                                6050
         .         .         .         .         .         .         .
ValArgIleGluSerLeuAspAlaCysTyrSerCysCysLeuValValIleAsnAsnLeuIleGlyCysHisLeu
GTGGGACTAGAGTGAATTTAGTCGTGTCATTGACGTCGTGTCGTGGTGTTATAACAAGTTTTAGGGTGTCACGTT
CACCCTGATCTCACTTAAATCAGCACAGTAACTGCAGCACAGCACCACAATATTGTTCAAAATCCCACAGTGCAA

                       6100                                              6150
    .         .         .         .         .         .         .         .
AlaSerTyrGlyPheSerMetAlaProValValSerGlyValHisGlyAspTyrTrpLeuArgLeuTyrIleLeu
CCGCGACATAGGTTTCGAGTACCGCCCCTGGTGTCTTGGGTGCACCGGTAGTATGGTGTTCGCGTCCATCTAATT
GGCGCTGTATCCAAAGCTCATGGCGGGGACCACAGAACCCACGTGGCCATCATACCACAAGCGCAGGTAGATTAA

                                                6200
         .         .         .         .         .         .         .
HisArgGlyArgMetPheValSerSerMetPheMetValGluLysProMetAsnTyrAsnValValGluArgTyr
CACCGCTGGGGAGTATTTGTGCGACCTGTATTTGTAATGGAGAAAACCGTACAACATTAAGTGGTGGAGGGCCAT
GTGGCGACCCCTCATAAACACGCTGGACATAAACATTACCTCTTTTGGCATGTTGTAATTCACCACCTCCCGGTA

                       6250                                              6300
    .         .         .         .         .         .         .         .
TrpIlePheArgGlnAsnPheMetAlaGlyAspValValMetArgPheTrpSerAlaLeuValGlnGlyGlyAla
GGTATATTTGGAGACTAATTTGTACCGCGGTAGGTGGTGGTAGGATTTGGTCGACCGGTTTTGGACGGGCGGCCG
CCATATAAACCTCTGATTAAACATGGCGCCATCCACCACCATCCTAAACCAGCTGGCCAAAACCTGCCCGCCGGC
```

```
                                                6350
         .         .         .         .         .         .         .
IleCysGlnLeuSerGlyProSerSerCysHisCysHisLeuAlaTrpSerGluTyrGlyHisIleMetMetSer
ATACGTGACGTCCCTTGGCCCTGACCTTGTTACTGTCACCTCTCGGGTCCTGAGCATTGGTACCTAGTAGTACGA
TATGCACTGCAGGGAACCGGGACTGGAACAATGACAGTGGAGAGCCCAGGACTCGTAACCATGGATCATCATGCT

                       6400                                              6450
    .         .         .         .         .         .         .         .
ThrMetIleAspIleAsnAlaCysCysLeuCysValHisMetCysLysArgLeuIleValLeuGluGluArgThr
GCAGTACTATAGTTACAACCGTGTTGTGTCCGTGTGCACGTATGTGAAGGAGTCCTAATGTTCGAGGAGGGCGCA
CGTCATGATATCAATGTTGGCACAACACAGGCACACGTGCATACACTTCCTCAGGATTACAAGCTCCTCCCGCGT

                                                6500
         .         .         .         .         .         .         .
LeuValMetAspTrpProValValTrpGluGlnIleLeuThrPheGlyValSerCysProLeuGlyArgValTyr
GTCTTGGTATAGGGTCCCTTGTTGGGTAAGGACTTAGTCGCATTTAGGGTGTGACGTCCCTTCTGGAGCGTGCAT
CAGAACCATATCCCAGGGAACAACCCATTCCTGAATCAGCGTAAATCCCACACTGCAGGGAAGACCTCGCACGTA

                       6550                                              6600
    .         .         .         .         .         .         .         .
      Region 5 COOH LeuThrValAsnProCysCysArgIleIleArgTrpTyrProLeuAlaProArgGln
SerValAsnHisMetThrLeuThrAsnCysGluProLeuLeuProHisAspGluLeuIleThrAlaArgThrGlu
TGAGTGCAACACGTAACAGTTTCACAATGTAAGCCCGTCGTCGCCTACTAGGAGGTCATACCATCGCGCCCAGAG
ACTCACGTTGTGCATTGTCAAAGTGTTACATTCGGGCAGCAGCGGATGATCCTCCAGTATGGTAGCGCGGGTCTC

                                                6650
         .         .         .         .         .         .         .
              Region 4 COOH GlnValSerHisAlaSerValValSerIleThrAsnThrThrThrAsp
  ArgLeuLeuLeuTyrAlaIleGlyLeu* NH2
ThrGluPheProProLeuArgAspArgSerTyrProThrArgArgSerLeuArgSerArgThrProArgLeuThr
ACAGAGTTTTCCTCCATCCGCTAGGGATGACATGCCTCACGCGGCTCTGTTGGCTCTAGCACAACCAGCATCACA
TGTCTCAAAAGGAGGTAGGCGATCCCTACTGTACGGAGTGCGCCGAGACAACCGAGATCGTGTTGGTCGTAGTGT

                       6700                                              6750
    .         .         .         .         .         .         .         .
 HisTrpIleSerArgArgValTyrAspTyrLysArgPheCysPheTrpThrArgAlaHisCysValSerArgArg
MetGlyPheProValGlySerThrThrMet NH2
GTACGGTTTACCTTGCGGCCTGCATCAGTATAAAGGACTTCGTTTTGGTCCACGCCCGCACTGTTTGTCTAGACGC
CATGCCAAATGGAACGCCGGACGTAGTCATATTTCCTGAAGCAAAACCAGGTGCGGGCGTGACAAACAGATCTGCG

                                               6800
        .         .         .         .         .         .         .
ArgArgAspArgArgLysAlaArgGluThrTyrTyrAsnTyrTyrIleTrpGluArgLeuAlaAspLeuArgGly
AGAGGCCAGAGCAGCGAATCGAGCGAGACACATCATCAACATCATATAGGTGAGAGAGTTTCGTAGGTCCGCGGG
TCTCCGGTCTCGTCGCTTAGCTCGCTCTGTGTAGTAGTTGTAGTATATCCACTCTCTCAAAGCATCCAGGCGCCC

                      6850                                              6900
   .         .         .         .         .         .         .         .
ArgAlaGluProGluIleTyrValGlyGluHisAlaAlaAlaArgIleValAspValValAlaSerTyrAlaVal
GGACCGAAGCCCAAGATACATTTGAGGAAGTACGCGGCGACGGGACTATTGTAGGTGGTGGCGTCTTATTCGGTG
CCTGGCTTCGGGTTCTATGTAAACTCCTTCATGCGCCGCTGCCCTGATAACATCCACCACCGCAGAATAAGCCAC

                                               6950
        .         .         .         .         .         .         .
GlyLeuTrpGlyValCysGluAsnGlnSerAspCysValProProAlaProLeuAlaProLeuValMet NH2
TGGGTCGGTTGGATGTGTAAGCAAGACGCTCAGTGTGTGCCCTCCTCGCCCTTCTCGACCTTCTTGGTACAAAAAAA
ACCCAGCCAACCTACACATTCGTTCTGCGAGTCACACACGGGAGGAGCGGGAAGAGCTGGAAGAACCATGTTTTTTT
```

```
                     7000                                             7050
 .         .         .         .         .         .         .         .
Region 3
  COOH GluLeuLeuAsnAspLeuValGluPheHisLeuAspIleLeuHisValArgGluGlyGlyThrAlaHis
AAAAAATAAGGTTTTCTAATAGGTTTTGGAGTTTTACTTCTAGATAATTCACTTGCGCGAGGGGAGGCCACCGCAC
TTTTTTATTCCAAAAGATTATCCAAAACCTCAAAATGAAGATCTATTAAGTGAACGCGCTCCCCTCCGGTGGCGTG

                                             7100
     .         .         .         .         .         .         .
AspPheGluValAlaLeuSerCysIleIleAlaAsnThrLeuHisGlnValIleAlaGluLeuLeuCysValAla
CAGTTTGAGATGTCGGTTTCTTGTCTATTACCGTAAACATTCTACAACGTGTTACCGAAGGTTTTCCGTTTGACG
GTCAAACTCTACAGCCAAAGAACAGATAATGGCATTTGTAAGATGTTGCACAATGGCTTCCAAAAGGCAAACTGC

                     7150                                             7200
 .         .         .         .         .         .         .         .
ArgValAspLeuHisValTyrLeuSerPheGlyGluProHisIleGluGluIlePheMetGlyAlaGlyGluVal
GGAGTGCAGGTTCACCTGCATTTCCGATTTGGGAAGTCCCACTTAGAGGAGATATTTGTAAGGTCGTGGAAGTTG
CCTCACGTCCAAGTGGACGTAAAGGCTAAACCCTTCAGGGTGAATCTCCTCTATAAACATTCCAGCACCTTCAAC

                                             7250
     .         .         .         .         .         .         .
MetGlyLeuTyrAsnGluAspArgTrpArgIleLeuIleAspArgLeuLeuAspArgIleAsnLeuGlyAlaMet
GTACGGGTTTATTAAAAGTAGAGCGGTGGAATAGTTATACAGAGATTCGTTTAGGGCTTATAATTCAGGCCGGTA
CATGCCCAAATAATTTTCATCTCGCCACCTTATCAATATGTCTCTAAGCAAATCCCGAATATTAAGTCCGGCCAT

                     7300                                             7350
 .         .         .         .         .         .         .         .
ThrPheIleGlnGluLeuAlaGlyGluValLysLeuArgLeuCysArgIleMet NH2
                                       Region 2 COOH SerGlnLeuPheGluProGluGlu
ACATTTTTAGACGAGGTCTCGCGGGAGGTGGAAGTCGGAGTTCGTCGCTTAGTACTAACGTTTTTAAGTCCAAGGAG
TGTAAAAATCTGCTCCAGAGCGCCCTCCACCTTCAGCCTCAAGCAGCGAATCATGATTGCAAAAATTCAGGTTCCTC

                                             7400
     .         .         .         .         .         .         .         .
CysValGlnIleLeuAsnLeuLeuProValAsnValPheIleGlyArgAspArgLeuAspArgArgLeuAlaLeu
TGTCTGGACATATTCTAAGTTTTCGCCTTGTAATTGTTTTTATGGCGCTAGGGCATCCAGGGAAGCGTCCCGGTC
ACAGACCTGTATAAGATTCAAAAGCGGAACATTAACAAAAATACCGCGATCCCGTAGGTCCCTTCGCAGGGCCAG

                   7450                                             7500
         .         .         .         .         .         .         .
GlnValTyrAspHisLeuAspAlaArgValLeuAlaAlaValGluGlyGlyProValMetValPheSerGlyVal
GACTTGTATTAGCACGTCCAGACGTGCCTGGTCGCGCCGGTGAAGGGGCGGTCCTTGGTACTGTTTTCTTGGGTG
CTGAACATAATCGTGCAGGTCTGCACGGACCAGCGCGGCCACTTCCCCGCCAGGAACCATGACAAAAGAACCCAC

                                             7550
     .         .         .         .         .         .         .         .
SerIleIleValArgMetSerProAlaIleSerValLeuThrAlaGlyIleTyrAlaGlnGlnMetProProSer
TGACTAATACTGTGCGTATGAGCCTCGATACGATTGGTCGCATCGGGGATACATTCGAACAACGTACCCGCCGCT
ACTGATTATGACACGCATACTCGGAGCTATGCTAACCAGCGTAGCCCCTATGTAAGCTTGTTGCATGGGCGGCGA

                   7600                                             7650
         .         .         .         .         .         .         .
IlePheHisLeuThrSerSerLeuPheAspProLeuAlaGluArgLeuPheAlaLeuValAspTyrAspHisGlu
ATATTTTACGTTCCACGACGAGTTTTTTAGTCCGTTTCGGAGCGCGTTTTTTCGTTCGTGTAGCATCAGTACGAG
TATAAAATGCAAGGTGCTGCTCAAAAAATCAGGCAAAGCCTCGCGCAAAAAAGCAAGCACATCGTAGTCATGCTC

                                             7700
     .         .         .         .         .         .         .         .
HisLeuTyrLeuCysThrLeuGluProValValValSerPheSerValMetLysArgGluPheMetAspAlaPro
TACGTCTATTTCCGTCCATTCAAGGCCTTGGTGGTGTCTTTTTCTGTGGTAAAAAGAGAGTTTGTACAGACGCCC
ATGCAGATAAAGGCAGGTAAGTTCCGGAACCACCACAGAAAAAGACACCATTTTTCTCTCAAACATGTCTGCGGG
```

```
                  7750                                              7800
        .         .         .         .         .         .         .
GluGlnMet NH2                  Region 1 COOH ValAsnSerAlaGlnArgValValProPhe
AAGGACGTAATTTGTGTTTTATTTTATTGTTTTTTTTTTGTAAATTTGTAATCTTCGGACAGAATGTTGTCCTTT
TTCCTGCATTAAACACAAAATAAAATAACAAAAAAAAAACATTTAAACATTAGAAGCCTGTCTTACAACAGGAAA

                                           7850
   .         .         .         .         .         .         .         .
ValValArgIleLeuMetLeuArgValValAlaMetGlyAlaHisGlyTyrPhePheGlnAspGlyHisAsnPhe
TTGTTGGGAATATTCGTATTCTGCCTGATGCCGGTACGGCCGCACTGGCATTTTTTTGACCAGTGGCACTAATTT
AACAACCCTTATAAGCATAAGACGGACTACGGCCATGCCGGCGTGACCGTAAAAAAACTGGTCACCGTGATTAAA

                  7900                                              7950
        .         .         .         .         .         .         .
LeuValValSerLeuGluGluThrMetAspProThrMetIleTyrSerGluThrPheValAspProGlnAsnVal
TTCGTGGTGGCTGTCAAGGAGCCAGTACAGGCCTCAGTATTACATTCTGAGCCATTTGTGTAGTCCAACCAATTG
AAGCACCACCGACAGTTCCTCGGTCATGTCCGGAGTCATAATGTAAGACTCGGTAAACACATCAGGTTGGTTAAC

                                           8000
   .         .         .         .         .         .         .         .
AspThrLeuAlaLeuPheArgGlyPheTyrGlyProProIleCysValArgLeuArgLeuSerLeuMetValAla
TAGCCAGTCACGATTTTTCGCTGGCTTTATCGGGCCCCCTTATGTATGGGCGTCCGCATCTCTGTTGTAATGTCG
ATCGGTCAGTGCTAAAAAGCGACCGAAATAGCCCGGGGGAATACATACCCGCAGGCGTAGAGACAACATTACAGC

                  8050                                              8100
        .         .         .         .         .         .         .
GlyMetProProIleValPheAsnIleProSerPhePheValTyrValGlySerPheGlyGluGlnArgProLeu
GGGGTATCCTCCATATTGTTTTAATTATCCTCTCTTTTTGTGTATTTGTGGACTTTTTGGGAGGACGGATCCGTT
CCCCATAGGAGGTATAACAAAATTAATAGGAGAGAAAAACACATAAACACCTGAAAAACCCTCCTGCCTAGGCAA

                                           8150
   .         .         .         .         .         .         .         .
IleAlaGlyGluArgGluLeuValValTyrLeuAlaGluValAlaAlaAlaMet NH2
TTATCGTGGGAGGGCGAGGTCTTGTTGTATGTCGCGAAGGTGTCGCCGTCGGTATTGTCAGTCGGAATGGTCATTTT
AATAGCACCCTCCCGCTCCAGAACAACATACAGCGCTTCCACAGCGGCAGCCATAACAGTCAGCCTTACCAGTAAAA

                8200                                              8250
      .         .         .         .         .         .         .         .
TTTGGATAATTTTTTGTGGTGAGCTGTGCCGTGGTCGAGTTAGTCAGTGTCACATTTTTCCCGGTTCATGTCTCGCT
AAACCTATTAAAAAACACCACTCGACACGGCACCAGCTCAATCAGTCACAGTGTAAAAAGGGCCAAGTACAGAGCGA

                                        8300
          .         .         .         .         .         .         .
CATATATATCCTGATTTTTTACTGCATTGCCAATTTCAGGTGTTTTTTGTGGGTCTTTTGGCGTGCGCTTGGATGCG
GTATATATAGGACTAAAAAATGACGTAACGGTTAAAGTCCACAAAAAACACCCAGAAAACCGCACGCGAACCTACGC

             8350                                              8400
   .         .         .         .         .         .         .         .
GGTCTTTGCTTTCGGTTTTTTGGGTGTTGAAGGAGTTTAGAAGTGAAGGCAAAAGGGTGCTATGCAGTGAAGGGTAA
CCAGAAACGAAAGCCAAAAAACCCACAACTTCCTCAAATCTTCACTTCCGTTTTCCCACGATACGTCACTTCCCATT

                                      8450
        .         .         .         .         .         .         .         .
AATTTTTTTGATGTTAAGGGTTATGTACGTTCAATGAGGCGGGATTTTGGATGCAGTGGGCGGGGCAAGGTGCGGG
TTAAAAAAACTACAATTCCCAATACATGCAAGTTACTCCGCCCTAAAACCTACGTCACCCGCCCCGTTCCCACGCCC

        8500                                              8550            8566
        .         .         .         .         .         .         .
GCGCGGTGCAGTGTTTGAGGTGGGGGAGTAATAGTATAACCGAAGTTAGGTTTTATTCCATATAATAACTACTAC5'
CGCGCCACGTCACAAACTCCACCCCCTCATTATCATATTGGCTTCAATCCAAAATAAGGTATATTATTGATGATG
                                                                           1 3'
```

repeat found at both ends of the Ad 2 genome (26), Shinagawa et al concerning the SmaI K fragment (map coordinate : 98.5-100) covering the last five hundred nucleotides of Ad 2 (27) and Steenbergh and Sussenbach concerning the HindIII K fragment (map coordinate : 97.3-100) of the Ad 5 DNA genome (28). The sequence presented in fig.2 is in complete agreement with these previous results on Ad 2 except for two base changes at position 8075 and 8131 where we observed an A and a T instead of a G and a C. It is worthwile noting that these two positions are among the few differences observed between the Ad 2 and Ad 5 sequence (27,28). But contrary to what was expected the nucleotides found at position 8075 and 8131 in the Ad 2 sequence shown in fig.2 are identical to the nucleotides found in the Ad 5 sequence (28) and not to those found in the Ad 2 sequence determined by Shinagawa (27).

DISCUSSION

Fiber mRNA and other transcripts of the r strand : A nucleotide sequence corresponding to the body of the fiber mRNA (29) was located within the EcoRI E fragment (14). This sequence starts with ATG 3658 located at position 86.15. It contains an open reading frame which extends from this ATG up to the end of the EcoRI E fragment, and codes for 413 aminoacids of the N terminal end of the fiber protein. As shown in fig.2, the same reading frame stays open in fragment HindIII F up to TAA 5404 located at position 91.2.

Altogether from ATG 3658 up to TAA 5404, this sequence codes for a protein of 582 aminoacids with a theoretical molecular weight of 61 925 daltons, in very good agreement with the estimated 62 000 daltons of the fiber protein (30).

As shown in table 1 the aminoacid composition which can be derived from this DNA sequence is in no way peculiar and the aminoacid sequence as well is rather banal. However, in agreement with the glycosylation nature of the fiber protein and probably its antigenic property, as many as 9 glycosylation sites Asn - X - Ser/Thr can be observed (31), evenly distributed all along the nucleotide sequence.

Most if not all eukaryotic mRNA exhibit a sequence AAUAAA near by the site of polyadenylation (32-34). In the case of the fiber mRNA it is interesting to note that codon TAA 5404 which closes the reading frame is embedded in the sequence AATAAA of which the two first A belong to the last codon GAA coding for Glu.

However, two other AATAAA sequences can be observed downstream at position 92.1 and 92.5 (nucleotides 5710 and 5884) which could also play a role in the polyadenylation of the fiber mRNA. But, because of the very good agreement so far observed between EM mapping of mRNA and sequence data it is more probable that the AATAAA 5402 sequence is the actual poly A addition site of the fiber mRNA (35).

Table 1 : Aminoacid composition of the fiber protein.

	%	Absolute number
Ala	6,3	37
Arg	2	10
Asn	8,1	47
Asp	5	29
Cys	0,5	3
Gln	3	18
Glu	3	18
Gly	8	46
His	0,9	5
Ile	5,5	32
Leu	9,6	56
Lys	5,8	34
Met	2,1	12
Phe	2,9	17
Pro	5,8	34
Ser	10,7	62
Thr	11,7	68
Trp	0,7	4
Tyr	2,9	17
Val	5,7	33

The fiber mRNA is the last known transcript made on the r strand, besides the fact that transcription goes on up to map coordinate 98.2 (36). As shown in fig.3, three other open reading frames with a substantial coding capacity can be observed on the l strand sequence. The translation of these open reading frames could start with ATG 6331, 6584, 7097 and stops with the nonsense triplets TAG 6616, TAA 7025, TGA 7433, making three proteins of respectively 10 000, 16 000 and 12 000 daltons. At present there is no argument indicating that these open reading frames indeed correspond to some mRNA and proteins but they could provide an explanation as to why the transcription of the r strand greatly overpasses the end of the fiber message located at map coordinate 91.2 and goes up to map position 98.2 making transcripts 2.5 kilobases longer than apparently needed.

A sequence AATAAA is observed starting with nucleotide 7750 (map coordinate 98) which could correspond to the polyadenylation signal of these potential transcripts, unless this sequence which is centered in a region unusually rich in A residues somehow plays a role in the stopping of the r strand transcription by the RNA polymerase.

Early 4 mRNA : The early 4 region correspond to a family of leftward transcripts which have been mapped between coordinates 99.2 and 91.3 (1,37). According to in vitro translation data (7,38), these mRNA would code for at least 9 proteins with a molecular weight of 35K, 24K, 21K, 19K, 17K, 16K, 14K, 13K and 11K. Besides the fact that the synthesis of the E4 proteins starts about two hours after infection, reaches a maximum around three and then declines, these proteins seem to be non essential for DNA replication

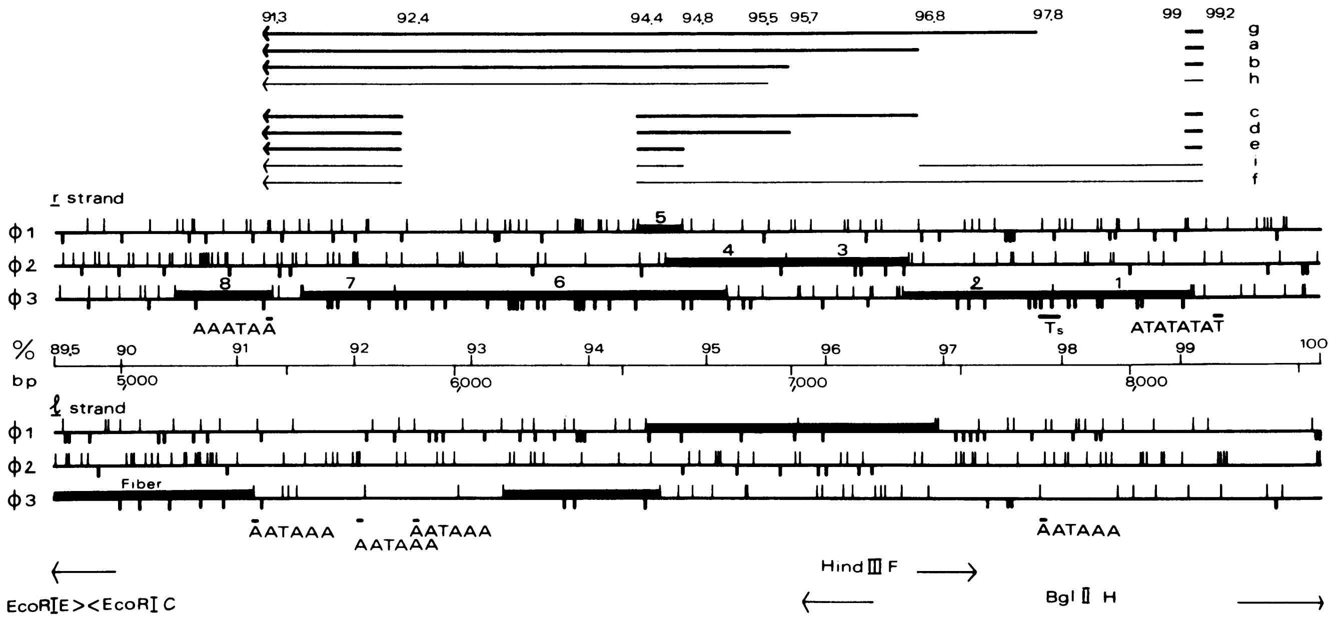

Fig.3 : Diagram showing the localisation of initiator and stop codons between coordinates 89.5 and 100%. The position of the various mRNA are from Chow et al (1) and Berk and Sharp (37). ϕ1, ϕ2 and ϕ3 correspond to the different reading frames as defined in the text. Upper vertical bars are for nonsense codons. Underneath vertical bars correspond to ATG triplets. E4 mRNA are transcribed from the l strand (leftward transcribed strand). They correspond in sequence and polarity to the rightward strand. Therefore open reading frames corresponding to these mRNA are indicated by thick lines on the r strand. On the contrary the open reading frame corresponding to the fiber protein is indicated by a thick line on the l strand.

and their role is at present unknown.

Recently Ziff et al have determined a short DNA sequence of the Ad 2 genome encoding the TATA box and the cap site of the E4 mRNAs (39). This sequence of 40 nucleotides is located between nucleotides 8271 and 8232 (fig.2) at map coordinates 99.2 in accordance with previous results (1). This sequence is followed by the leader sequence which is 62 nucleotides long and stops at residue number 6176 with the common donor sequence GGTAAG (13,14,40,41). This leader sequence is devoided of ATG able to play a role during the initiation of the translation. Therefore such a signal would have to be located in the body of the various mRNA spliced to this leader sequence. At the other end of the E4 region an AATAAA sequence is encountered which indicates the position of the poly A addition site (32-34). This sequence is located at nucleotide 5449, i.e. at map coordinate 91.3 in complete agreement with previous electron microscopy mapping of the 3' end of the E4 mRNA (1). The poly A addition site of this early RNA family is then slightly downstream of the poly A addition site of the fiber mRNA, whose AATAAA signal is located on the 1 chain at nucleotide 5402. The fact that these two regions exist and are transcribed at early and late time after infection, in opposite direction and ending at a very close position to each other, is reminiscent of a situation encountered in the SV40 and polyoma viruses (42).

Nine different E4 mRNA have been mapped on the Ad 2 virus chromosome by electron microscopy observation or by S1 digestion of RNA-DNA duplex (1,37) : RNA a, b, g and h, with only one intron sequence, located between the leader (coordinate 99) and the various bodies (coordinates : 96.8 ; 95.7 ; 97.8 ; 95.5) have been observed by EM and/or S1 digestion. On the contrary, mRNA c, d, e, f, i have an intron sequence located between coordinates 94.4 and 92.4 and have only been observed by EM. Apart for the e mRNA, transcripts of this subclass seem to be scarce.

As shown in fig.3 the uneven distribution of the stop codons in the r strand, creates in the lefward transcripts, several open reading regions called 1 to 8, able to code for the E4 proteins (7) from one of the ATGs located at their beginning.

a and b mRNA : It is interesting to note that in most cases there is good correlation between the map coordinate of the 5' end of the mRNA body and the beginning of an open reading frame. This is particularly noticeable for the a and b mRNA. The body of these messengers starts at map coordinates 96.8 and 95.7 just in front of the open reading regions 3 and 4. Translation of these two open reading regions would give two proteins of 13K by using the first ATG as initiator or 13 and 11K by using the second ATG for region 3.

The aminoacid composition of the 11K protein would reveal a noticeable excess of acidic over basic residues (10 glutamic acid and 7 aspartic acid as compared to 6 arginine and 6 histidine) indicating that this protein would probably have an acidic PIE and therefore could correspond to the 11K acidic protein obtained by in vitro translation of the E4 mRNA. Interestingly enough, non published data by J. Lewis (38) have suggested that this 11K acidic protein should be coded by the a mRNA, in agreement with DNA sequence prediction.

c and d mRNA : Two other mRNA, called c and d, also have the 5' end of their body at coordinate 96.8 and 95.7 like the a and b transcripts. These two mRNA have an intron sequence between coordinates 94.4 and 92.4, that is downstream stop codon TAA 6985 and TAG 6628 closing the open reading regions 3 and 4. They should therefore code for the same 11K and 13K proteins as the a and b mRNA. Since c and d mRNA are scarce (1), they were not seen by S1 mapping (37), they could be the products of some abnormal processing of the a and b mRNA - due to the presence in these transcripts of the donor and acceptor sequences which are recognized during the processing of the e mRNA.

e mRNA : This mRNA is by far the most abundant species of the subclass of mRNA having an intron between coordinates 94.4 and 92.4 (1). This mRNA is composed of a leader, a first exon going from 94.8 to 94.4 and a second exon from 92.4 to 91.3. Close to the 5' end of the first exon which falls around nucleotide 6700 (fig.2) there are two AG's, at position 6712 and 6630 which could determine the position of the 5' end of the first exon (29,43,44). More AG's can be found upstream and downstream but AG's 6712 and 6630 are the only ones which can be included in a sequence able to pair to some extent with the 5' end of the UI RNA (45). Further down, a sequence GTGAG corresponding to the consensus donor sequence is located at nucleotide 6531. Interestingly enough such a sequence is the only one around in the analyzed DNA and its map position coïncides with that observed by EM for the 3' end of the first exon. Furthermore placing the 5' end of this exon at AG's 6712 or 6630 would determine a length for this exon of 181 or 99 nucleotides, compared with the 145 nucleotides as estimated by EM examination of DNA-RNA heteroduplex.

Four ATG's can be observed in this region : ATG 6705, 6678 and 6540 are located in the open reading frame 3, ATG 6556 in frame 2 which is closed by TAA 6553. Therefore splicing the leader sequence to nucleotide 6628 would mean skipping the two first ATG, probably relegating the initiation of translation to ATG 6540 at a position very close to the border of the first exon. According to this hypothesis, only 3 aminoacids could be coded by this exon. Moreover a control mechanism would have to prevent initiation at ATG 6556 which is located in a closed reading frame.

On the contrary, ATG 6705 could be used as initiator triplet in the translation of open reading region 6 by splicing the leader to nucleotide 6710 (fig.3). Such an initiation at a position very close to the beginning of the 5' end of the mRNA body has often been observed in the E3 region(14,15).

The 5' end of the second exon has been mapped by EM at coordinate 92.4 i.e. around nucleotide 5820. Nearby this position several AG's can be observed which could correspond to the end of the intron. This is particularly true for AG's 5822 and 5792 because : 1) of their position ; 2) they belong to a sequence able to pair with the UI RNA 5' end (45) ; 3) they allow the remaining part of the e mRNA to be read on phase 3, the only one open downstream that position (open reading region 7 in fig.3). Translation would then continue on that exon up to TGA 5544. According to this hypothesis, a protein of 17 000 daltons might be synthesized from ATG 6705 up to TGA 5544 and could correspond to the 17K protein observed by in vitro translation (7,38). This protein would be rather rich in proline (10%), arginine (10%) and threonine (11%) and would also exhibit 4 glycosylation sites (31), all located in its second half.

In another connection from ATG 6540, located just in front of the second splice, a protein of 10K could be synthesized with this e mRNA.

i and f mRNAs : mRNA i and f are the only ones with no intron starting at map coordinate 99. For these two messenger RNA the first leader is still attached to the mRNA body. Slightly downstream this leader sequence, the 3' end of which has been located at nucleotide 8176, there is an ATG (residue number 8160) which is followed in phase 3 by an open reading frame extending up to TAA 7776 (open reading region 1 in fig.3), and able to code for a protein of 14 000 daltons.

While no splice has been detected in this region by EM or S1 mapping (1,37), it is interesting to note that a sequence GTAAG (residue number 7797) is present. From this sequence, one could suspect the existence of a short intron splicing out stop codon TAA 7776 and a region very rich in T (28 thymine over 37 residues) as already observed. This hypothetical intron could extend up to one of the various acceptor sequences located at nucleotides 7736, 7622 or 7433 allowing translation to continue in open frame 3 (open reading region 2 in fig.3), up to stop codon TGA 7332 at coordinate 96.7. According to the length of this hypothetical intron a protein of 28, 24 or 17K could be made by translation of mRNA f from ATG 8160 up to TGA 7332.

An intron sequence starting at coordinate 96.8 has been observed for mRNA i. In agreement with this result sequence GTGAG 7358 might well indicate the position of the donor splice site. The second exon of mRNA i has been located by EM at coordinates 94.8-94.4 in the same position as the first exon of e mRNA. Therefore the same reasoning can be applied to i mRNA. If the acceptor sequence is determined by AG 6712, sequence GAAA 6710 would be spliced to sequence GTCT 7362 creating a TGA stop codon in reading frame 3 (open reading region 2 in fig.3). This TGA (residue number for T : 7359 and residue number for G : 6710) would then block the reading of mRNA i at the splice junction. On the contrary if the acceptor sequence is determined by AG 6630, sequence TAGG 6628 would be spliced to sequence GTCT 7362, allowing the reading to continue along the second exon sequence in frame 1 up to stop codon TGA 6545 which is located upstream the next donor sequence GTGAG 6531.

In short, depending on the various possible splicing events, which may occur during the maturation of i mRNA, proteins with a molecular weight ranging from 14 to 31K could be made.

g mRNA : mRNA g is a rather abundant transcript which has only been observed by S1 mapping (37). It has a leader sequence at coordinate 99.2 and its body extends from 97.8 to 91.3. Several ATGs located in open reading frame 3 are encountered at the beginning of its body. Starting from one of these, translation could go up to stop codon TGA 7332 (open reading region 2 in fig.3). These proteins, of molecular weight in the range of 14K, would therefore possess community of structure with the proteins coded by mRNA i or f which are also read in frame 3.

h mRNA : This messenger has not been observed by EM and from time to time only by S1 mapping (37). Like most of the other E4 messengers it has a leader at coordinate 99.2. According to its length estimated as 1500 nucleotides by gel electrophoresis, its body should extend from coordinates 95.5 up to 91.3. As shown in fig.3, this messenger, in spite of its paucity, is the best candidate for translating the long open reading frame located between coordinates 95.2 and 92.4 and able to code from ATG 6705 to stop codon 5820 for a protein of 34 000 daltons with only one glycosylation site (open reading region 6, fig.3). A protein of 35K has been observed in some circumstances by Green et al (38) and could correspond well with the protein coded by mRNA h.

The various hypotheses made concerning the translation capacity of the E4 mRNA are summarized in the diagram shown in fig.4. From this, it appears that proteins coded by a and b mRNA would have a particular structure with no relationship between themselves or the others. On the contrary, proteins made by f, g and i messenger RNA could have a rather long aminoacid sequence in common as suspected after fingerprint analysis (38).

Transcription beyond the early region 4 poly A site : Termination of

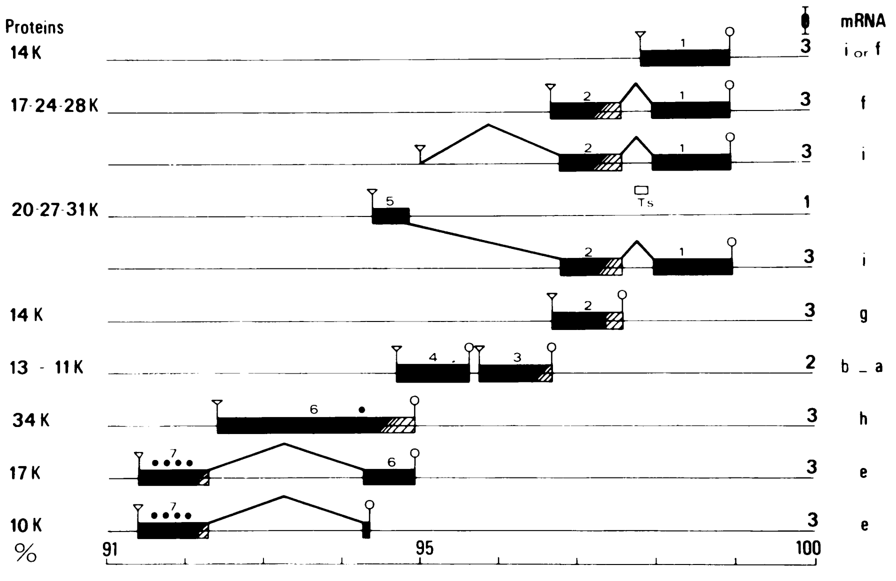

Fig.4 : Hypothetical translation pattern of E4 mRNA. The putative structure of the various E4 mRNA is derived from their position on the genome as determined by EM or S1 digestion, the localisation on the sequence of the open reading frames and consensus sequences found at the border of the introns. Numbers 1 to 7 correspond to the different open reading regions. ⚲ corresponds to initiator ATG triplet. Y corresponds to stop codon closing a reading frame. Blackened areas correspond to translated regions. Stripped areas correspond to regions where there are several possible ATG initiator triplets or acceptor sequences leading to some ambiguity in determining where the translation starts. Closed dots indicate potential glycosylation sites.

transcription of the late transcription unit occurs near map position 99 (36), distal to any sequences included within known mRNA molecules. Likewise it has been recently shown that transcription continues past the poly A addition site in early 2 and 4 transcription units (46). These results would thus indicate that transcription proceeds some distance beyond the poly A addition site, thereby requiring an RNA chain cleavage to generate the poly A addition site.

However, more recently, it has been shown that early region 2 would code, not only for the 72K mRNA, whose poly A addition site has been mapped at coordinate 61.6 (1) but for some other mRNA whose main bodies extend between coordinates 11.2 and 31.5 (4). We have already noticed (this work) that downstream the poly A addition site of the fiber mRNA there are three open reading frames with a substantial coding capacity, the mRNA of which could end at position 98 with AATAAA 7750. In that respect, it is interesting to note that downstream the poly A addition site of the E4 region there is an open reading frame (region 8 in fig.3). This region which starts with TGA 5454 and stops with TAG 5163 can be read in frame 3. By hybridization and UV transcription mapping, Nevins et al (46) have calculated that transcription of the early region 4 transcription unit terminates at approximately 88.4, i.e. in the EcoRI E fragment. Nucleotide sequence around this position reveals the presence of a unique ATTAAA sequence (residue number 4584) at map coordinate 88.8 (15). This sequence which ressembles the more usual AATAAA poly A addition site signal has already been observed at the 3' end of the 3 a and d mRNA of the E3 region (ATTAAA 2397) (14) and also at the 3' end of several interferon mRNA (47). Therefore we would like to suggest that termination of transcription in the late transcription unit and the Early 4 transcription unit, several hundred nucleotides beyond the main known poly A addition site could be due to the existence of some other genes or exons not detected until now - as recently shown for the early 2 region which codes also for the terminal protein (4).

If region 8 is indeed translated, translation could start with ATG 5427 and region 8 would have to be spliced with the first leader of the E4 region. Translation could also be initiated in an other upstream exon and continues in region 8 after splicing. Several possible acceptor sequences can be observed at the beginning of region 8, which can hybridize with the UI RNA sequence (45) : AG 5367, 5355 and 5332. In that respect it could be interesting to note that at the end of region 7 there is a putative donor sequence GTGAG 5545 which by ligation to one of the acceptor sequences determined by AG 5355 or 5367 could allow translation to pass beyond the splice point.

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

ACKNOWLEGMENT

This work was supported by Grant from CNRS and INSERM.

REFERENCES

1. Chow, L.T., Broker, T.R. and Lewis, J.B. (1979) J. Mol. Biol. 134, 265-303
2. Berk, A.J., Lee, F., Harrison, T., Williams, J. and Sharp, P.A. (1979) Cell 17, 935-944
3. Lewis, J.B. and Mathews, M.B. (1980) Cell 21, 303-313
4. Stillman, B.W., Lewis, J.B., Chow, L.T., Mathews, M.B. and Smart, J.E. (1981) Cell 23, 497-508
5. Persson, H., Monstein, H., Akusjärvi, G. and Philipson L. (1981) Cell 23, 485-496
6. Broker, T.R. and Chow, L.T. (1979) ICN-UCLA Symposia on Molecular and cellular Biology 14, 611-635
7. Harter, M.L. and Lewis, J.B. (1978) J. Virol. 26, 736-749
8. Nevins, J.R. and Darnell, J.E. (1978) J. Virol. 25, 811-823
9. Chow, L.T. and Broker, T.R. (1978) Cell 15, 497-510
10. Alestrom, P., Akusjärvi, G., Perricaudet, M., Mathews, M., Klessig, D., and Pettersson, U. (1980) Cell 19, 671-681
11. Nevins, J.R. and Wilson, M.C. (1981) Nature 290, 113-118
12. Galibert, F., Hérissé, J. and Courtois, G. (1979) Gene 6, 1-22
13. Baker, C.C., Hérissé, J., Courtois, G., Galibert, F. and Ziff, E. (1979) Cell 18, 569-580
14. Hérissé, J., Courtois, G. and Galibert, F. (1980) Nucleic Acids Res. 8, 2173-2192
15. Hérissé, J. and Galibert, F. (1981) Nucleic Acids Res. 9, 1229-1240
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. 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
19. Carmeron, J.R., Panascuko, S.M., Lehman, I.R. and Davis, R.W. (1975) Proc. Natl. Acad. Sci. U.S.A. 72, 3416-3420
20. Birnboim, H.C. and Doly, J. (1979) Nucleic Acids Res. 7, 1513-1523
21. Hampe, A., Therwath, A., Soriano, P. and Galibert, F. (1981) Gene 14, 11-21
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. (1980) In Methods in Enzymology 65, part I, 499-560
24. Robinson, A.J. and Bellett, A.J.D. (1974) Cold Spring Harbor Symp. Quant. Biol. 39, 523-531
25. Rekosh, D.M.K., Russell, W.C., Bellett, A.J.D. and Robinson, A.J. (1977) Cell 11, 283-295
26. Arrand, J.R. and Roberts, R.J. (1979) J. Mol. Biol. 128, 577-594
27. Shinagawa, M., Padmanabhan, R.V. and Padmanabhan, R. (1980) Gene 9, 99-114

28. Steenbergh, P.H. and Sussenbach, J.S. (1979) Gene 6, 307-318
29. Zain, S., Sambrook, J., Roberts, R.J., Keller, W., Freed, M. and Dunn, A.R. (1979) Cell 16, 851-861
30. Lewis, J.B., Anderson, C.W., Atkins, J.F. and Gesteland, R.F. (1974) Cold Spring Harbor Symp. Quant. Biol. 39, 581-590
31. Struck, D.K., Lennartz, W.J. and Brew, K. (1978) J. Biol. Chem. 253, 5786-5794
32. Mc Reynolds, L., O'Malley, B.W., Nisbet, A.D., Fothergill, J.E., Givol, D., Fields, S., Robertson, M. and Browlee, G.G. (1978) Nature 273, 723-728
33. Efstratiadis, A. and Kafatos, F.C. (1977) Cell 10, 571-585
34. Proudfoot, N.J. (1977) Cell 10, 559-570
35. Chow, L.T., Roberto, J.M., Lewis, J.B. and Broker, T.R. (1977) Cell 11, 819-836
36. Fraser, N.W., Nevins, J.R., Ziff, E. and Darnell, J.E. (1979) J. Mol. Biol. 129, 643-656
37. Berk, A.J. and Sharp, P.A. (1978) Cell 14, 695-711
38. EMBO Adenovirus meeting at Peebles, Scotland (June 1980)
39. Baker, C.C. and Ziff, E.B. (1981) J. Mol. Biol. in press
40. Ziff, E. and Evans, R. (1978) Cell 15, 1463-1475
41. Aküsjarvi, G. and Petterson, U. (1979) J. Mol. Biol. 134, 143-158
42. Reddy, V.B., Thimmappaya, B., Dhar, R., Subramanian, K.N., Zain, B.S., Pan, J., Ghosh, P.K., Celma, M.L. and Weissman, S.M. (1978) Science 200, 494-502
43. Breathnach, R., Benoist, C., O'Hare, K., Gannon, F. and Chambon, P. (1978) Proc. Natl. Acad. Sci. U.S.A. 75, 4853-4857
44. Catterall, J.F., O'Malley, W.O., Robertson, M.A., Staden, R., Tanaka, R. and Brownlee, G.G. (1978) Nature 275, 510-513
45. Avvedimento, V.E., Vogeli, G., Yamada, Y., Maizel, J.V., Ira Pastan, Jr. V. and Benoit de Crombrugghe, B. (1980) Cell 21, 689-696
46. Nevins, J.R., Blanchard, J.M. and Darnell, Jr.J.E. (1980) J. Mol. Biol. 144, 377-386
47. Goeddel, D.V., Leung, D.W., Dull, T.J., Gross, M., Lawn, R.M., Mc Candliss, R., Seeburg, P.H., Ullrich, A., Yelverton, E. and Gray, P.W. (1981) Nature 290, 20-26