Volume 7 Number 4 1979 Nucleic Acids Research

# Polyoma virus. The early region and its T-antigens

E.Soeda*, J.R.Arrand+ and Beverly E.Griffin+†

+Imperial Cancer Research Fund, Lincoln's Inn Fields, London, UK

Received 14 August 1979

## ABSTRACT

The DNA sequence of the early coding region of polyoma virus is presented. It consists of 2739 nucleotides. The sequence predicts that more than one reading frame can be used to code for the three known polyoma virus early proteins (designated small, middle and large T-antigens). From the DNA sequence, the 'splicing' signals used in the processing of viral RNA to functional messenger RNAs can be predicted, as well as the sizes and sequences of the three proteins. Other unusual aspects of the DNA sequence are noted. Comparisons are made between the DNA sequences and the predicted amino acid sequences of the respective large T-antigens of polyoma virus and the related virus Simian Virus (SV) 40.

## INTRODUCTION

The DNA tumour virus, polyoma, appears to code for at least three proteins which are synthesised early after infection, and which have been designated small, middle and large tumour (T) antigens. The best characterised of the three is large T-antigen, probably because its existence has been known for longest. Even so, it was only two years ago that large T-antigen was established as being at least in part, and probably wholly, virally coded (1). Because of its many functions, and the nature of these functions, it is a very interesting protein. It appears to be required for viral DNA replication, and for the initiation of cellular transformation. It may also stimulate replication of the host DNA and exercise control over viral transcription (for review, see ref. 2). These functions have been identified, but there may well be other, as yet unknown, functions associated with this protein. Unfortunately, it is present in lytically infected and transformed cells in such small amounts that it has not yet been isolated in sufficient quantity or with sufficient purity for further studies. In many of its functions, it is similar to the large T-antigen of the related virus, simian virus 40. The two antigens apparently differ, however, in that whereas in SV40, expression of large T-antigen is probably required for the maintenance of the transformed state of cells, a functional polyoma virus large T-antigen does not seem essential for maintenance in all cell lines (2, 3). In many of their functions, both proteins appear to resemble the somewhat smaller (60K) cistron A protein of the bacteriophage ØX 174 for which roles as varied as initiation of replication and strand separation and ligation have been identified (4).

The biological activities of the more recently isolated small and middle T-antigens of polyoma virus have not yet been identified, although middle T-antigen is known to be associated with membranes (5) and may be associated with a protein kinase activity (6). (There is no known SV40-coding equivalent of polyoma virus middle T-antigen). Studies with host-range transforming mutants (hr-t) suggest, however, that one or both of the polyoma virus proteins must play a role in transformation (7). The properties of several recently isolated early polyoma virus mutants with altered transformation characteristics support the premise that a functional middle T-antigen may be necessary for the full expression of transformation (8, 9). The small and middle T-antigens do not appear to affect viral or host DNA replication, but neither of the proteins has been studied in any detail.

We report here the sequence of the region of polyoma virus DNA which should contain the coding information for the viral early proteins. The DNA sequence allows predictions to be made about 'splicing' in the messenger species, and because such predictions can be made, about the amino acid sequence subsequently expected in each of the known T-antigens. The DNA sequence also predicts that, in addition to the three T-antigens, the genome could code for other early viral proteins. Some of these data have been presented elsewhere (10). A comparison at the molecular level can also be made between the early regions of polyoma virus and SV40 DNAs and their proteins. Although there are a great many similarities to be seen, there are also regions of non-homology between the two viral genomes, and these differences may be of critical biological significance.

## MATERIALS AND METHODS

Polyoma virus DNA (A2 large plaque strain) was grown and purified essentially as previously described (11). All DNA used for sequence determination between the EcoRI restriction site and the end of the early region came from a single preparation, using virus stocks made from twice plaque purified virus. Nucleotide sequence analysis was carried out essentially by the methods of Maxam and Gilbert (12) using either 5'- or 3'- labelled DNA fragments. The former were radioactively labelled using high specific activity $\gamma$-$^{32}$P-ATP (Radiochemical Centre, Amersham) and T4 induced polynucleotide kinase (P-L Biochemicals), and the latter using the appropriate high specific activity

$\alpha^{32}$P- deoxyribonucleotide triphosphates (Radiochemical Centre, Amersham) and T4 induced DNA polymerase, a gift from Dr. N. Smolar. The chemical degradation products were separated on 12% or 20% polyacrylamide gels, and subsequently visualised by exposure to Fuji medical X-ray film at - 70$^{o}$ with Fuji Mach-2 intensifying screens.

All restriction enzymes were made by standard procedures (36). Every region of the DNA sequence was determined more than once. In addition both strands of the DNA were sequenced over most of the polyoma virus early region (see Fig.1). Data, if desired, can be provided.

## RESULTS AND DISCUSSION

The Early coding sequence. One of the main reasons for determining the DNA sequence of the early region of polyoma virus was to ask whether the virus had enough information in this region to code for the large T-antigen, previously estimated to be between 80-105K in size and requiring nearly 3000 nucleotides to specify it. The answer is, it does. Another reason was to ask whether there was information available to code for any other protein(s). The answer to that is also affirmative. A third reason was to determine where on the genome the coding regions lie, and what protein sequences can be predicted from DNA sequence. This point is discussed subsequently. A fourth reason was to compare the polyoma virus sequence with the sequence of a related tumour virus, simian virus (SV) 40.

The DNA sequence for the coding part of the early region of the A2 (large

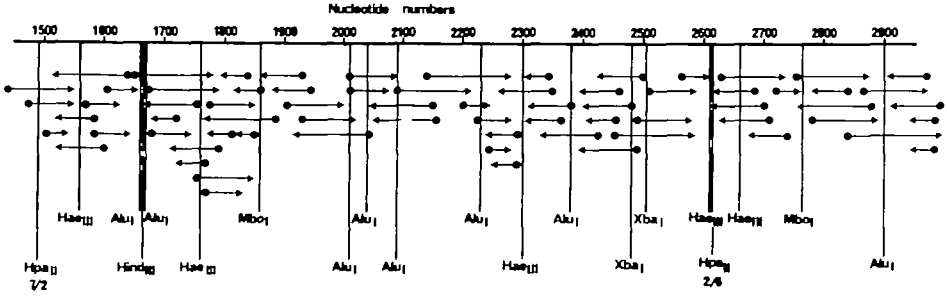

Figure 1: Sequencing the early region of polyoma virus DNA between nucleotide positions 1487 and 2919. Data for positions 1-1564 are given elsewhere (10). The arrows below the sequence numbers show the direction and extent of the sequence determination. Some of the restriction endonucleases used in sequence determination are indicated. Cleavage sites of the enzymes were all independently mapped using standard methods.

plaque) strain of polyoma virus is presented in Fig. 2. One extensive open coding frame is found in the sequence between positions 173 and 805, two open coding frames between positions 809 and 1496, and only one for the rest of the sequence given, up to position 2911. It may be relevant to point out that between positions 2583 and 2783 there is a second open reading frame. Although it is short, this frame contains an ATG as its first triplet and could potentially code for a protein which contains 67 amino acids and would be about 7.5K in size. Although the DNA sequence predicts that such a protein could exist, no such virus-coded protein has, as yet, been identified. Most studies aimed at studying polyoma virus early proteins would probably fail to reveal such a protein, however. Not only do they depend upon immunoprecipitation selection procedures (which are probably selecting proteins with sequence which corresponds to the regions around the N-termini of the T-antigens) (34) but also, for technical reasons, proteins smaller than about 12-15K are not observed (1, 13, 14). Data on coding frames is summarised in Fig. 3.

Were large T-antigen coded for by the entire polyoma virus early region, a protein with 913 amino acids (about 100K in size) could be made. This is not, however, the case. One striking bit of evidence which says that the simplistic idea of colinear coding from the entire early region, starting with a unique initiation codon and ending with a unique termination codon, is not happening comes from the appearance of a presumed termination codon (TAG) at position 806 in the first open coding frame encountered in the DNA. Further evidence comes from the fact that polyoma virus hr-t mutants, which have deletions in the early region of up to about 4% of the genome, make a full-sized large T-antigen (13, 14). The prototype of these mutants, NG-18, has a deletion of 187 base pairs which lies in the sequence given (Fig. 2) between positions 512 and 698 (15, 16). Moreover, studies on the total early mRNA population of polyoma virus show that all the messenger species are "spliced", that is, their sequences correspond to those found in non-contiguous regions of the DNA (17). Therefore the whole of the early region cannot be coding in a continuous manner for large T-antigen, neither need it be coding exclusively for large T-antigen. The DNA sequence consequently needs to be correlated with other results.

Data are available, which considered together with the DNA sequence, now allow for predictions to be made about the region of the polyoma genome which codes for large T-antigen, as well as for the two other known early viral proteins, middle and small T-antigens. Briefly, studies on the proteins themselves suggest that all three early antigens share common sequences, believed to lie at the N-terminal regions of the proteins (18, 19). Studies on early polyoma virus messenger RNAs (17), together with data evolving about sequences which surround spliced regions

Figure 2

```
      173
Py    ATG  GAT  AGA  GTT  CTG  AGC  AGA  GCT  GAC  AAA  GAA  AGG  CTG  CTA  GAA  CTT  CTA  AAA  CTT  CCC 232
Py    MET  ASP  ARG  VAL  LEU  SER  ARG  ALA  ASP  LYS  GLU  ARG  LEU  LEU  GLU  LEU  LEU  LYS  LEU  PRO
SV40  MET  ASP  LYS  VAL  LEU  ASN  ARG  GLU  GLU  SER  LEU  GLN  LEU  MET  ASP  LEU  LEU  GLY  LEU  GLU

Py    AGA  CAA  CTA  TGG  GGG  GAT  TTT  GGA  AGA  ATG  CAG  CAG  GCA  TAT  AAG  CAG  CAG  TCA  CTG  CTA 292
Py    ARG  GLN  LEU  TRP  GLY  ASP  PHE  GLY  ARG  MET  GLN  GLN  ALA  TYR  LYS  GLN  GLN  SER  LEU  LEU
SV40  ARG  SER  ALA  TRP  GLY  ASN  ILE  PRO  LEU  MET  ARG  LYS  ALA  TYR  LEU  LYS  LYS  CYS  LYS  GLU

Py    CTG  CAC  CCA  GAC  AAA  GGT  GGA  AGC  CAT  GCC  TTA  ATG  CAG  GAA  TTG  AAC  AGT  CTC  TGG  GGA 352
Py    LEU  HIS  PRO  ASP  LYS  GLY  GLY  SER  HIS  ALA  LEU  MET  GLN  GLU  LEU  ASN  SER  LEU  TRP  GLY
SV40  PHE  HIS  PRO  ASP  LYS  GLY  GLY  ASP  GLU  GLU  LYS  MET  LYS  LYS  MET  ASN  THR  LEU  TYR  LYS

Py    ACA  TTT  AAA  ACT  GAA  GTA  TAC  AAT  CTG  AGA  ATG  AAT  CTA  GGA  GGA  ACC  GGC  TTC  CAG [GTA 412
Py    THR  PHE  LYS  THR  GLU  VAL  TYR  ASN  LEU  ARG  MET  ASN  LEU  GLY  GLY  THR  GLY  PHE  GLN
SV40  LYS  MET  GLU  ASP  GLY  VAL  LYS  TYR  ALA  HIS  GLN  PRO  ASP  PHE  GLY  ***  GLY  PHE  TRP  ASP

Py    AGA  AGG  CTA  CATGCGGATGGGTGGAATCTAAGTACCAAAGACACCTTTGGTGATAGATACTACCAGCGGTTCTGCAGAATGCCTCTTACCTGC 505
Py
SV40  ALA  THR  GLU

Py    CTAGTAAATGTTAAATACAGCTCATGTAGTTGTATATTATGCCTGCTTAGAAAGCAACATAGAGAGCTCAAAGACAAATGTGATGCCAGGTGCCTAGTA 604

Py    CTTGGAGAATGTTTTTGTCTTGAATGTTACATGCAATGGTTTGGAACACCAACCCGAGATGTGCTGAACCTGTATGCAGACTTCATTGCAAGCATGCCT 703

Py    ATAGACTGGCTGGACCTGGATGTGCACAGCGTGTATAATCCAAGTAAGTATCAAGAGGGCGGGTGGGTATTTACGGCCTATATTCTTACAG] 794
                          1          2 3           4  5             6  7                   a
Py    GGC  TCT  CCC b CCT  AGA 808
Py    GLY  SER  PRO   PRO  ARG

Py    ACG  GCG  GAG  CGA  GGA  ACT  GAG  GAG  AGC  GGC  CAC  AGT  CCA  CTA  CAC  GAT  GAC  TAC  TGG  TCA 868
Py    THR  ALA  GLU  ARG  GLY  THR  GLU  GLU  SER  GLY  HIS  SER  PRO  LEU  HIS  ASP  ASP  TYR  TRP  SER

Py    TTC  AGC  TAT  GGA  AGC  AAG  TAC  TTC  ACA  AGG  GAA  TGG  AAT  GAT  TTC  TTC  AGA  AAG  TGG  GAC 928
Py    PHE  SER  TYR  GLY  SER  LYS  TYR  PHE  THR  ARG  GLU  TRP  ASN  ASP  PHE  PHE  ARG  LYS  TRP  ASP

Py    CCC  AGC  TAC  CAG  TCG  CCG  CCT  AAG  ACT  GCC  GAG  TCT  TCT  GAG  CAA  CCC  GAC  CTA  TTC  TGT 988
Py    PRO  SER  TYR  GLN  SER  PRO  PRO  LYS  THR  ALA  GLU  SER  SER  GLU  GLN  PRO  ASP  LEU  PHE  CYS

Py    TAT  GAG  GAG  CCA  CTC  CTA  TCC  CCC  AAC  CCG  AGT  TCT  CCA  ACA  GAT  ACA  CCC  GCA  CAT  ACT 1048
Py    TYR  GLU  GLU  PRO  LEU  LEU  SER  PRO  ASN  PRO  SER  SER  PRO  THR  ASP  THR  PRO  ALA  HIS  THR

Py    GCT  GGA  AGA  AGA  CGA  AAT  CCT  TGT  GTT  GCT  GAG  CCC  GAT  GAC  AGC  ATA  TCC  CCG  GAC  CCC 1108
Py    ALA  GLY  ARG  ARG  ARG  ASN  PRO  CYS  VAL  ALA  GLU  PRO  ASP  ASP  SER  ILE  SER  PRO  ASP  PRO
Py     CCC AGA ACT CCT GTA TCC AGA AAG CGA CCA AGA CCA GCT GGA GCC ACT GGA GGA GGA GGA 1168
Py     PRO ARG THR PRO VAL SER ARG LYS ARG PRO ARG PRO ALA GLY ALA THR GLY GLY GLY GLY

Py     GGA GGA GTA CAT GCC AAT GGA GGA TCT GTA TTT GGA CAT CCT ACC GGG GGA ACA AGT ACC 1228
Py     GLY GLY VAL HIS ALA ASN GLY GLY SER VAL PHE GLY HIS PRO THR GLY GLY THR SER THR

Py     CCA GCT CAT CCC CCC CCC TAT CAT TCC CAG GGC GGG TCT GAG TCC ATG GGA GGG TCT GAT 1288
Py     PRO ALA HIS PRO PRO PRO TYR HIS SER GLN GLY GLY SER GLU SER MET GLY GLY SER ASP
SV40                                                   ILE PRO THR TYR GLY THR ASP GLU

Py     TCT TCG GGA TTT GCA GAG GGC TCA TTT CGA TCC GAT CCT AGA TGC GAG TCA GAG AAT GAG 1348
Py     SER SER GLY PHE ALA GLU GLY SER PHE ARG SER ASP PRO ARG CYS GLU SER GLU ASN GLU
SV40   TRP GLU GLN TRP TRP ASN ALA PHE ASN GLU GLU ASN LEU PHE CYS SER GLU GLU MET PRO

Py     AGC TAC TCA CAG AGC TGC TCT CAG AGC TCA TTC AAT GCA ACG CCA CCT AAG AAG GCT AGG 1408
Py     SER TYR SER GLN SER CYS SER GLN SER SER PHE ASN ALA THR PRO PRO LYS LYS ALA ARG
SV40   SER SER ASP ASP GLU ALA THR ALA ASP SER GLN HIS SER THR PRO PRO LYS LYS LYS ARG

Py     GAG GAC CCT GCT CCT AGT GAC TTT CCT AGC AGC CTT ACT GGG TAT TTG TCT CAT GCT ATT 1468
Py     GLU ASP PRO ALA PRO SER ASP PHE PRO SER SER LEU THR GLY TYR LEU SER HIS ALA ILE
SV40   LYS VAL GLU ASP PRO LYS ASP PHE PRO SER GLU LEU LEU SER PHE LEU SER HIS ALA VAL

Py     TAT TCT AAT AAA ACG TTC CCG GCA TTT CTA GTA TAC TCC ACC AAA GAG AAA TGC AAA CAA 1528
Py     TYR SER ASN LYS THR PHE PRO ALA PHE LEU VAL TYR SER THR LYS GLU LYS CYS LYS GLN
SV40   PHE SER ASN ARG THR LEU ALA CYS PHE ALA ILE TYR THR THR LYS GLU LYS ALA ALA LEU

Py     TTA TAT GAT ACC ATA GGG *** AAG TTC AGG CCC GAA TTC AAA TGC CTG GTC CAT TAT GAG 1585
Py     LEU TYR ASP THR ILE GLY *** LYS PHE ARG PRO GLU PHE LYS CYS LEU VAL HIS TYR GLU
SV40   LEU TYR LYS LYS ILE MET GLU LYS TYR SER VAL THR PHE ILE SER ARG HIS ASN SER TYR

Py     GAG GGG GGC ATG CTG TTC TTT CTA ACT ATG ACT AAG CAC AGG GTT TCA GCA GTT AAG AAT 1645
Py     GLU GLY GLY MET LEU PHE PHE LEU THR MET THR LYS HIS ARG VAL SER ALA VAL LYS ASN
SV40   ASN HIS ASN ILE LEU PHE PHE LEU THR PRO HIS ARG HIS ARG VAL SER ALA ILE ASN ASN

Py     TAT TGC TCT AAG CTT TGC CGC *** AGC TTC CTA ATG TGT AAG GCA GTC ACC AAG CCT ATG 1702
Py     TYR CYS SER LYS LEU CYS ARG *** SER PHE LEU MET CYS LYS ALA VAL THR LYS PRO MET
SV40   TYR ALA GLN LYS LEU CYS THR PHE SER PHE LEU ILE CYS LYS GLY VAL ASN LYS GLU TYR
```

```
Py    GAA TGC TAT CAA GTT GTA ACC GCA GCA CCA TTT CAG TTA ATA ACA GAA AAT AAG CCA GGC 1762
Py    GLU CYS TYR GLN VAL VAL THR ALA ALA PRO PHE GLN LEU ILE THR GLU ASN LYS PRO GLY
SV40  LEU MET TYR SER ALA LEU THR ARG ASP PRO PHE SER VAL ILE GLU GLU SER LEU PRO GLY

Py    *** CTC CAC CAA TTC GAG TTT *** ACA GAC GAG CCA GAA GAA CAG AAA GCA GTA GAC TGG 1816
Py    *** LEU HIS GLN PHE GLU PHE *** THR ASP GLU PRO GLU GLU GLN LYS ALA VAL ASP TRP
SV40  GLY LEU LYS GLU HIS ASP PHE ASN PRO GLU GLU ALA GLU GLU THR LYS GLN VAL SER TRP

Py    ATT ATG GTA GCA GAC TTT GCA CTA GAA AAC AAC CTT GAT GAT CCC CTG TTA ATT ATG GGG 1876
Py    ILE MET VAL ALA ASP PHE ALA LEU GLU ASN ASN LEU ASP ASP PRO LEU LEU ILE MET GLY
SV40  LYS LEU VAL THR GLU TYR ALA MET GLU THR LYS CYS ASP ASP VAL LEU LEU LEU LEU GLY

Py    TAT TAT CTT GAT TTT GCC AAA GAG GTT CCT TCA TGC ATA AAG TGT AGC AAA GAG GAA ACC 1936
Py    TYR TYR LEU ASP PHE ALA LYS GLU VAL PRO SER CYS ILE LYS CYS SER LYS GLU GLU THR
SV40  MET TYR LEU GLU PHE GLN TYR SER PHE GLU MET CYS LEU LYS CYS ILE LYS LYS GLU GLN

Py    CGC CTC CAA ATA CAT TGG AAA AAC CAT AGA AAG CAT GCA GAG AAT GCA GAC CTC TTC CTG 1996
Py    ARG LEU GLN ILE HIS TRP LYS ASN HIS ARG LYS HIS ALA GLU ASN ALA ASP LEU PHE LEU
SV40  PRO *** *** SER HIS TYR LYS TYR HIS GLU LYS HIS TYR ALA ASN ALA ALA ILE PHE ALA

Py    AAT TGT AAA GCT CAA AAG ACA ATC TGT CAG CAG GCA GCT GCG AGT *** CTG GCA TCC AGG 2053
Py    ASN CYS LYS ALA GLN LYS THR ILE CYS GLN GLN ALA ALA ALA SER *** LEU ALA SER ARG
SV40  ASP SER LYS ASN GLN LYS THR ILE CYS GLN GLN ALA VAL ASP THR VAL LEU ALA LYS LYS

Py    AGA CTG AAA TTA GTA GAG TGT ACC CGC AGC CAG CTA TTA AAG GAG AGA TTG CAA CAG TCT 2113
Py    ARG LEU LYS LEU VAL GLU CYS THR ARG SER GLN LEU LEU LYS GLU ARG LEU GLN GLN SER
SV40  ARG VAL ASP SER LEU GLN LEU THR ARG GLU GLN MET LEU THR ASN ARG PHE ASN ASP LEU

Py    CTC CTC AGG CTA AAA GAA CTT GGC TCC TCC *** *** *** *** GAT GCT CTA CTC TAC CTA 2161
Py    LEU LEU ARG LEU LYS GLU LEU GLY SER SER *** *** *** *** ASP ALA LEU LEU TYR LEU
SV40  LEU ASP ARG MET ASP ILE MET PHE GLY SER THR GLY SER ALA ASP ILE GLU GLU TRP MET

Py    GCA GGT GTC GCT TGG TAC CAG TGT CTT TTA GAC GAC TTT CCT CAA ACC CTG TTT AAG ATG 2221
Py    ALA GLY VAL ALA TRP TYR GLN CYS LEU LEU GLU ASP PHE PRO GLN THR LEU PHE LYS MET
SV40  ALA GLY VAL ALA TRP LEU HIS CYS LEU LEU PRO LYS MET ASP SER VAL VAL TYR ASP PHE

Py    CTT AAG CTG CTA ACA GAA AAT GTG CCA AAA CGA CGC AAC ATA CTT TTT AGA GGA CCA GTT 2281
Py    LEU LYS LEU LEU THR GLU ASN VAL PRO LYS ARG ARG ASN ILE LEU PHE ARG GLY PRO VAL
SV40  LEU LYS CYS MET VAL TYR ASN ILE PRO LYS LYS ARG TYR TRP LEU PHE LYS GLY PRO ILE

Py    AAT TCA GGA AAG ACA GGC CTA GCA GCC GCG CTT ATT AGC CTG TTA GGA GGC AAG TCT CTC 2341
Py    ASN SER GLY LYS THR GLY LEU ALA ALA ALA LEU ILE SER LEU LEU GLY GLY LYS SER LEU
SV40  ASP SER GLY LYS THR THR LEU ALA ALA ALA LEU LEU GLU LEU CYS GLY GLY LYS ALA LEU
```

```
Py    AAC  ATA  AAT  TGC  CCT  GCA  GAT  AAA  CTT  GCT  TTT  GAG  CTT  GGT  GTG  GCA  CAG  GAC  CAG  TTT 2401
Py    ASN  ILE  ASN  CYS  PRO  ALA  ASP  LYS  LEU  ALA  PHE  GLU  LEU  GLY  VAL  ALA  GLN  ASP  GLN  PHE
SV40  ASN  VAL  ASN  LEU  PRO  LEU  ASP  ARG  LEU  ASN  PHE  GLU  LEU  GLY  VAL  ALA  ILE  ASP  GLN  PHE

Py    GTG  GTC  TGT  TTT  GAA  GAT  GTA  AAG  GGT  CAA  ATA  GCC  TTG  AAC  AAA  CAA  CTG  CAG  CCA  GGG 2461
Py    VAL  VAL  CYS  PHE  GLU  ASP  VAL  LYS  GLY  GLN  ILE  ALA  LEU  ASN  LYS  GLN  LEU  GLN  PRO  GLY
SV40  LEU  VAL  VAL  PHE  GLU  ASP  VAL  LYS  GLY  THR  GLY  GLY  GLU  SER  ARG  ASP  LEU  PRO  SER  GLY

Py    ATG  GGA  GTG  GCT  AAT  CTA  GAT  AAT  CTC  AGG  ACT  ACC  TGG  AAT  GGG  AGT  GTA  AAG  GTC  AAT 2521
Py    MET  GLY  VAL  ALA  ASN  LEU  ASP  ASN  LEU  ARG  THR  THR  TRP  ASN  GLY  SER  VAL  LYS  VAL  ASN
SV40  GLN  GLY  ILE  ASN  ASN  LEU  ASP  ASN  LEU  ARG  ASP  TYR  LEU  ASP  GLY  SER  VAL  LYS  VAL  ASN

Py    CTA  GAA  AAA  AAG  CAC  AGC  AAC  AAA  AGG  TCA  CAA  CTC  TTT  CCA  CCC  TGT  GTG  TGT  ACA  ATG 2581
Py    LEU  GLU  LYS  LYS  HIS  SER  ASN  LYS  ARG  SER  GLN  LEU  PHE  PRO  PRO  CYS  VAL  CYS  THR  MET
SV40  LEU  GLU  LYS  LYS  HIS  LEU  ASN  LYS  ARG  THR  GLN  ILE  PHE  PRO  PRO  GLY  ILE  VAL  THR  MET

Py    AAT  GAA  TAT  CTC  CTA  CCA  CAA  ACA  GTA  TGG  GCC  CGG  TTT  CAC  ATG  GTG  TTG  GAT  TTC  ACC 2641
Py    ASN  GLU  TYR  LEU  LEU  PRO  GLN  THR  VAL  TRP  ALA  ARG  PHE  HIS  MET  VAL  LEU  ASP  PHE  THR
SV40  ASN  GLU  TYR  SER  VAL  PRO  LYS  THR  LEU  GLN  ALA  ARG  PHE  VAL  LYS  GLN  ILE  ASP  PHE  ARG

Py    TGC  AAA  CCC  CAT  CTG  GCC  CAA  TCT  CTT  GAA  AAG  TGT  GAA  TTT  TTG  CAA  AGG  GAA  AGA  ATT 2701
Py    CYS  LYS  PRO  HIS  LEU  ALA  GLN  SER  LEU  GLU  LYS  CYS  GLU  PHE  LEU  GLN  ARG  GLU  ARG  ILE
SV40  PRO  LYS  ASP  TYR  LEU  LYS  HIS  CYS  LEU  GLU  ARG  SER  GLU  PHE  LEU  LEU  GLU  LYS  ARG  ILE

Py    ATT  CAG  AGT  GGA  GAT  ACC  CTT  GCC  CTA  TTA  CTC  ATA  TGG  AAT  TTC  ACT  TCA  GAT  GTA  TTT 2761
Py    ILE  GLN  SER  GLY  ASP  THR  LEU  ALA  LEU  LEU  LEU  ILE  TRP  ASN  PHE  THR  SER  ASP  VAL  PHE
SV40  ILE  GLN  SER  GLY  ILE  ALA  LEU  LEU  LEU  MET  LEU  ILE  TRP  TYR  ARG  PRO  VAL  ALA  GLU  PHE

Py    GAT  CCT  GAT  ATT  CAG  CGT  CTT  GTG  AAG  GAG  GTT  CGT  GAC  CAG  TTT  GCT  AGT  GAG  TGC  TCC 2821
Py    ASP  PRO  ASP  ILE  GLN  GLY  LEU  VAL  LYS  GLU  VAL  ARG  ASP  GLN  PHE  ALA  SER  GLU  CYS  SER
SV40  ALA  GLN  SER  ILE  GLN  SER  ARG  ILE  VAL  GLU  TRP  LYS  GLU  ARG  LEU  ASP  LYS  GLU  PHE  SER

Py    TAC  AGT  TTG  TTT  TGT  GAT  ATA  CTT  TGT  AAT  GTG  CAA  GAA  GGC  GAC  GAC  CCC  TTG  AAG  GAC 2881
Py    TYR  SER  LEU  PHE  CYS  ASP  ILE  LEU  CYS  ASN  VAL  GLN  GLU  GLY  ASP  ASP  PRO  LEU  LYS  ASP
SV40  LEU  SER  VAL  TYR  GLN  LYS  MET  LYS  PHE  ASN  VAL  ALA  MET  GLY  ILE  GLY  VAL  LEU  ASP  TRP

Py    ATA  TGT  GAT  ATA  GCT  GAA  TAC  ACA  GTT  TAT 2911 TGA  (ATAAA)2919
Py    ILE  CYS  ASP  ILE  ALA  GLU  TYR  THR  VAL  TYR
SV40  LEU  ARG  ASN  SER  ASP  ASP  ASP  ASP  GLU  ASP  SER  GLN  GLU  ASN  ALA  ASP  LYS  ASN  GLU  ASP

SV40  GLY  GLY  GLU  LYS  ASN  MET  GLU  ASP  SER  GLY  HIS  GLU  THR  GLY  ILE  ASP  SER  GLN  SER  GLN

SV40  GLY  SER  PHE  GLN  ALA  PRO  GLN  SER  SER  GLN  SER  VAL  HIS  ASP  HIS  ASN  GLN  PRO  TYR  HIS

SV40  ILE  CYS  ARG  GLY  PHE  THR  CYS  PHE  LYS  LYS  PRO  PRO  THR  PRO  PRO  PRO  GLU  PRO  GLU  THR  TAA
```

Figure 2: The early region of the polyoma virus genome (A2 strain) and a comparison with SV40. Line 1 of each horizontal column represents the nucleotide sequence of polyoma virus (Py) early DNA and has been divided into the triplets that would appear in the coding frame for large T-antigen (Fig. 3). The sequence given has the same polarity as early mRNAs (31). Line 2 is the predicted amino acid sequence of polyoma virus large T-antigen. Line 3 (where appropriate) is the SV40 large T-antigen sequence, data taken from Fiers et al. (30) and Reddy et al. (32). Gaps in sequence (indicated by *) are used in order to maximise homology. The DNA sequence which corresponds to that shown here between positions 399 and 833 has been reported for a different strain of polyoma virus, made by marker rescue of the hr-t mutant, NG-18 (16). That sequence differed by a single nucleotide (G) insert at position 812, relative to the sequence we earlier reported (10). A resequencing of this area indicated that the G was present.

The numbering of nucleotides in the DNA sequence is that previously adopted (10). The reason for suggesting that the ATG at position 173 is the initiation codon for polyoma virus T-antigens has been put forward in the text. Within the coding sequence (line 1), two presumed termination codons (underlined) appear, TAG at position 806 and TGA at position 2912. These are discussed in the text. Sequences underlined (dashed lines) between positions 725 and 809 are discussed in the text and in Fig. 5. A predicted splice in the mRNA coding for polyoma virus large T-antigen is put within brackets. The intervening sequence lies between positions 410-794 and includes about 7.5% of the genome (from about 78.5 to 85.8 units on the physical map (11) of polyoma virus DNA) (see Fig. 3).

Homologies between the early regions of polyoma virus and SV40 are indicated in two ways. Dots placed above a nucleotide in Line 1 of each horizontal column indicate DNA sequence homology. Sequence homology between the amino acids of the large T-antigens coded for by the two viruses is indicated by boxes. The major differences in the early regions of the two viruses appear in the polyoma virus DNA sequence between positions 724-1264 (about 10% of the polyoma virus genome), for which there is no analogy in SV40, and at the end of the SV40 early region (last three lines in this figure), for which there is no analogy in polyoma virus.

in other mRNAs (20, 21) allow for the following predictions (which are discussed subsequently):

- DNA sequence between nucleotide position 173 and 409, and between positions 795 and 2911 probably codes for polyoma virus large T-antigen. The actual codons used by this gene, as predicted from the sequence, are given in Fig. 4. From

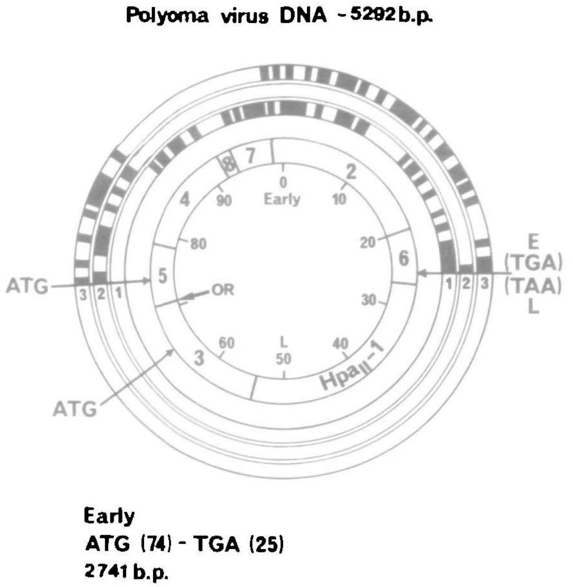

Figure 3: The early region of polyoma virus has been divided into its coding frames and related to the standard Hpa II physical map (11) of the viral genome. The initiation codon (ATG) for the early proteins is presumed to lie at nucleotide 173 (see Fig. 2, and 10, 34). Beginning at this site, the DNA sequence has been divided into three potential coding frames, frame 1 beginning at position 173, frame 2 at 174 and frame 3 at 175. Wherever a termination codon appears within twenty-seven nucleotides (equivalent to nine amino acids) within the sequence (Fig. 2), a solid bar is drawn to indicate this. Thus, the coding potential of the early region is apparent. 'Splicing' events would theoretically allow any two (or more) open areas on any frame to be joined together to produce a mRNA and subsequently a protein. The sequences (Fig. 3) thought to correspond to the three known polyoma virus T-antigens are discussed in the text. The N-terminus of all three proteins is thought to lie at the methionine triplet (ATG) at 74 map units; the presumptive TGA termination codon for large T-antigen is shown at 25.5 map units.

the DNA sequence (see Fig. 2) large T-antigen can be predicted to be a protein containing 785 amino acids, with a calculated molecular weight of 88,000 daltons, which is somewhat smaller than the size (about 100K) previously estimated from SDS-polyacrylamide gels (8, 14).

CODON USAGE FOR LARGE T-ANTIGEN

|  | U |  | C |  | A |  | G |  |  |
|---|---|---|---|---|---|---|---|---|---|
| U | Phe | 22 + 2 = 24 | Ser | 15 + 0 = 15 | Tyr | 3 + 1 = 14 | Cys | 15 + 0 = 15 | U |
|  |  | 15 + 1 = 16 |  | 11 + 0 = 11 |  | 9 + 1 = 10 |  | 11 + 0 - 11 | C |
|  | Leu | 8 + 1 = 9 |  | 10 + 1 = 11 | Term | 0 + 0 = 0 | Term | 1 + 0 = 1 | A |
|  |  | 7 + 1 = 8 |  | 2 + 0 = 2 |  | 0 + 0 = 0 | Trp | 9 + 2 = 11 | G |
| C | Leu | 15 + 2 = 17 | Pro | 15 + 0 = 15 | His | 11 + 1 = 12 | Arg | 1 + 0 = 1 | U |
|  |  | 12 + 1 = 13 |  | 16 + 1 = 17 |  | 6 + 1 = 7 |  | 4 + 0 = 4 | C |
|  |  | 17 + 5 = 22 |  | 15 + 1 = 16 | Gln | 15 + 1 = 16 |  | 5 + 0 = 5 | A |
|  |  | 11 + 5 = 16 |  | 4 + 0 = 4 |  | 17 + 6 = 23 |  | 1 + 0 = 1 | G |
| A | Ile | 7 + 0 = 7 | Thr | 10 + 1 = 11 | Asn | 20 + 2 = 22 | Ser | 9 + 1 = 10 | U |
|  |  | 1 + 0 = 1 |  | 12 + 1 = 13 |  | 8 + 1 = 9 |  | 15 + 2 = 17 | C |
|  |  | 12 + 0 = 12 |  | 12 + 1 = 13 | Lys | 18 + 4 = 22 | Arg | 13 + 5 = 18 | A |
|  | Met | 11 + 4 = 15 |  | 3 + 0 = 3 |  | 27 + 1 = 28 |  | 9 + 1 = 10 | G |
| G | Val | 8 + 1 = 9 | Ala | 15 + 1 = 16 | Asp | 21 + 2 = 23 | Gly | 4 + 1 = 5 | U |
|  |  | 5 + 0 = 5 |  | 9 + 1 = 10 |  | 17 + 2 = 19 |  | 10 + 1 = 11 | C |
|  |  | 12 + 1 = 13 |  | 19 + 1 = 20 | Glu | 18 + 4 = 22 |  | 21 + 5 - 26 | A |
|  |  | 8 + 0 = 8 |  | 3 + 0 = 3 |  | 28 + 0 = 28 |  | 9 + 1 = 10 | G |

Figure 4: Codons predicted to be used by polyoma virus in coding for large T-antigen. The first number in each vertical column gives codon usage after the splice in the mRNA and the second number that before the splice (see Fig. 2). The total is also given. The rare use of codons CGN or NCG is apparent. The infrequent use of AUC (Ile) is surprising. This codon is not used at all in the coding sequence of SV40 (30,32).

- DNA sequence between nucleotide positions 173 and 746, and between 809 and 1496 may code for middle T antigen using a different coding frame from that used for large T-antigen. Middle T-antigen could then be predicted to contain 429 amino acids and to have a molecular weight of 49,500 daltons. This is somewhat smaller than the size (55K) estimated by other assays (13, 18). The predicted protein sequence has been discussed elsewhere (10).

- DNA sequence between nucleotide positions 173 and 746, and between 795 and 805 (for alternatives, see below) may code for small T-antigen. Thus, small T-antigen could be predicted to contain 195 amino acids (see Fig. 5) and to have a molecular weight of 22,000 daltons. This is consistent with the size reported for this protein (13, 18).

Some of the data given above for the proteins were derived from "splicing rules". These say that in the messenger RNAs (a) repeated sequences (although

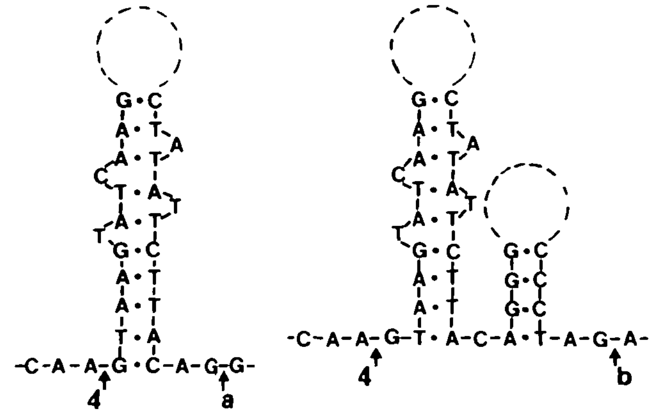

Figure 5: Secondary structure which could exist in DNA corresponding to intervening sequences in unprocessed polyoma virus mRNAs. The location of 4, a and b on polyoma virus early DNA is given in Fig. 2. Sequence 'spliced out' between positions 4 and a could give rise to a messenger RNA coding for small T-antigen using only one reading frame of the DNA (frame 1). Sequence 'spliced out' between positions 4 and b could produce a messenger RNA coding for middle T-antigen using frame 1 sequence on its 5'-side and frame 3 sequence on its 3'-side (see Fig. 3). Alternative splicing sites for both proteins are discussed in the text. Although splices suggested in this figure are in agreement with current splicing 'rules', final assignments of splicing junctions must ultimately rest on sequencing of either the relevant mRNAs or the viral proteins.

often short) are found at the junctions between coding and intervening sequences, (b) the sequence GT appears at the 5'-side of the intervening sequence and AG at the 3'-side, the 'GT-AG' rule (22), (c) no other AG dinucleotide occurs within 13 nucleotides prior to the terminal AG of the intervening sequence (21), and (d) pyrimidine-rich sequences frequently are found at the 3'-side of the intervening sequence (23). Gannon et al (20) suggest that as a general pattern, the sequence TCAGGTA appears around the 3'-junction of intervening sequences and TNCAGG around the 5'-junction. (An 'intervening sequence' is defined as that sequence present in the DNA, but missing in the corresponding mRNA, and presumably 'spliced out' during messenger processing). Other data were taken from studies on the polyoma virus messenger RNAs which appear to have spliced junctions at points which correspond to the DNA sequence at about positions 410, 750, 790 and 810 (R. Kamen, manuscript in preparation).

From these considerations, the intervening sequence for large T-antigen is predicted to have at its 5'-side the sequence CCAGGTA (which lies between nucleotides 406-412) and at its 3'-side the sequence TACAGG (from position 790-795), the underlined portions of these sequences being absent in the mRNA. This splice results in a frame-shift which removes the TAG termination codon found in the sequence between positions 806-808 and moves the body of large T-antigen to a different coding frame, designated frame 2 in Fig. 3. It is noteworthy that the corresponding splicing sequences in SV40 have been found to be TGAGGTA and TTTAGA, respectively (23). If the other predictions are correct, there is an interesting aspect of the sequence in the mature mRNA species. It is apparent from the DNA sequence that a heptanucleotide with sequence GAGGAAC would appear within the coding sequence ten base pairs before the 5'-spliced junction and ten base pairs after the 3'-junction (between positions 393-399 and 820-826, respectively). This may be fortuitous, but because of the length of this oligonucleotide and its position it may also play some role in splicing, such as, for example, preventing any secondary structure being formed between these two particular parts of the RNA during the process leading to maturation.

Prediction of the splicing junctions for small and middle T-antigens is difficult. It can be seen (Fig. 2) that between positions 720 and 773 there are seven potential heptanucleotides that could serve for the 5'-junction of intervening sequences, although not all of them may be equally valid. The 'GT' part of each of the sequences is underlined and given numbers 1 through 7 in Fig. 2. Sequences which obey the 'rules' for the 3'-junction lie between positions 789-794 (TACAGG, labelled a in Fig. 2) and 804-809 (CCTAGA, labelled b in Fig. 2). Splices which occur between the two junctions at 4 and a (designated 4a, see Fig. 5) would give rise to a messenger that could code for a protein the (approximate) size of small T-antigen. Splices between 1,2, or 7 and b would also lead to a protein which is compatible with the size of small T-antigen. In 4a, the small T-antigen would terminate in frame 1, using the termination codon TAG which lies between positions 806-808, which gives rise to a protein with 195 amino acids (see above). The SV40 DNA sequences which correspond to sequences around the splice regions for small T-antigen have been found to be TAAGGTA and TTTAGA (23). Proposed sequences (CCAAGTA and CCTAGA) for polyoma virus are shown in Fig. 5. Similar processing events could conceivably occur in the mRNA for middle T-antigen. Thus, splicing between positions 4 and b (see Fig. 5) or between positions 3, 5, or 6 and a could give rise to messengers which would involve sequences from two different reading frames and would code for middle T-antigens which differ only by a few internal amino acids. Sequences around the potential splicing junctions between 4 and a and b are shown (Fig. 5) because they fit the existing data regarding splicing in polyoma virus mRNAs and allow for the formation of intervening sequences with fairly stable secondary structures between the 5'- and 3'-ends. Although much has already been written about splicing (20-23), the exact sequence or structural requirements are still largely a matter of conjecture and comparison between mRNA species. Secondary structures may be important, for example, in substrate recognition by splicing enzymes, but so may be a number of other factors. As more data become available from different systems, they may help in elucidating the requirements for splicing.

It can readily be seen from this discussion that a judicious choice of splicing sites, or even mistakes in splicing, could lead to a number of proteins which have only slightly different primary structures. A small virus, like polyoma virus, may acquire additional coding potential and functional flexibility by using different splicing signals to modify parts of its proteins. There is to date no evidence to suggest that the virus does, or does not, avail itself of this flexibility.

Another interesting aspect of the DNA sequence is the presence of the sequence AATAAA twice within the early region. The sequence AAUAAA has been found near the 3'-end of a large number of eukaryotic mRNAs and is thought to be involved in the processing of messengers, either in cleaving a primary transcript to a functional messenger or adding a poly-A tail to its 3'-end (24). The corresponding sequence AATAAA appears twice in the early region of polyoma virus DNA, between nucleotide positions 1475-1480 and 2914-2919 (or at about 98.4 and 25.6 units on the physical map (11) of polyoma virus) (see Fig. 3). Polyoma early mRNAs with sedimentation coefficients about 20S have been previously described (25). These would correspond to species transcribed from practically the whole of the early region and would be expected to use the AATAAA signal at 25.6 map units. Polyadenylated messenger species about half this size which appear to use the signal at 98.4 map units have recently been identified but not yet correlated with any known protein (R. Kamen, personal communication). Both of the AATAAA sequences found in polyoma virus DNA appear as part of larger symmetrical sequences (Fig. 6a, b), capable of forming either hairpin loops or four-stranded species such as those previously described as possibly existing around the viral origin of replication (10). The symmetrical sequence shown in Fig. 6b contains in its top strand the termination codon TGA that is presumably used for large T-antigen. The most interesting aspect of these sequences is that, in addition to their symmetry, they contain both the AATAAA signal and termination codons on both strands of the DNA.

## Comparison with SV40

It can be seen from a comparison of the DNA sequences and the predicted amino acid sequences (Fig. 2) that polyoma virus and SV40 have regions in which considerable homology is observed and other regions which appear unique to each

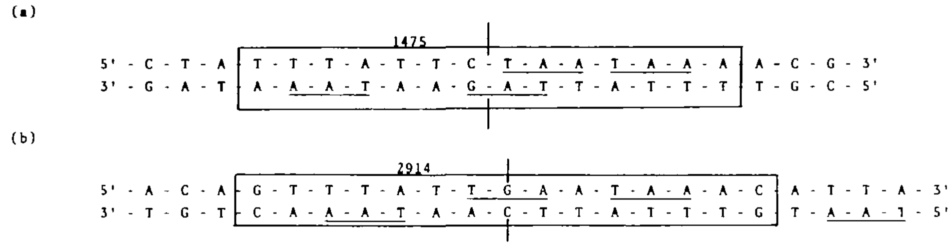
(a)
1475
5' - C - T - A
T - T - T - A - T - T - C - T - A - A - T - A - A - A
A - C - G - 3'
3' - G - A - T
A - A - A - T - A - A - G - A - T - T - A - T - T - T
T - G - C - 5'
(b)
2914
5' - A - C - A
G - T - T - T - A - T - T - G - A - A - T - A - A - A - C
A - T - T - A - 3'
3' - T - G - T
C - A - A - A - T - A - A - C - T - T - A - T - T - T - G
T - A - A - T - 5'

Figure 6: Symmetry in the regions of the DNA which contain the sequence corresponding to the AAUAAA postulated to be a processing signal in eukaryotic mRNAs (19). Top lines,

a) The AATAAA sequence which occurs about half-way through the early region DNA.

b) The AATAAA sequence which is found at the end of the early region. The TGA is thought to be the termination codon for large T-antigen; the TAA on the opposite strand outside the symmetrical sequence is the postulated termination codon for the capsid protein VP1 (E. Soeda, J.R. Arrand and B.E. Griffin, manuscript submitted).

It can be seen in the bottom line (a and b) that anti-early (or late) sequences also contain AATAAA and termination codons. If the hydrogen bonds forming the double-stranded structures are broken, each strand of the DNA can be folded into a hairpin loop. Alternatively, if a double-stranded DNA is bent about the two-fold axis of symmetry, a four-stranded helix can be formed. The possible significance of the latter type structures for interacting with proteins has been discussed elsewhere (7).

virus. A large amount of homology is seen in the DNAs in the regions which code for the N-terminal portions of the proteins. This point has already been discussed (10). In the early coding region of polyoma virus, between nucleotide numbers 809 to 1264 (or about 9% of the polyoma genome), there appears to be no counterpart in the SV40 sequence. The "extra" sequences in polyoma virus can account for much of the apparent size difference between the polyoma virus and SV40 large T-antigens (polyoma virus T-antigen being larger than its SV40 counterpart). There is limited homology between the two viral genomes in what could be considered the "internal part" of the polyoma virus DNA sequence (position 1265 to about 2221), and more extensive homology towards the end (positions 2222 to 2776). The maximum homology is seen in a region which spans the DNA from nucleotide positions 2222 to 2740 and includes about 10% of the genome (from 12 to 22 map units) (see Fig.

3). Throughout this region there is nearly 60% homology between the two DNAs and over 60% homology between predicted amino acid sequences. In a portion of this region, between positions 2504-2566, there is about 75% DNA homology and 19 out of 21 predicted amino acids are the same. These sequence data are in agreement with recent nucleic acid hybridisation studies, carried out on polyoma virus and SV40 DNAs under relatively non-stringent conditions, which show that DNA from 12 to 59 map units on the polyoma virus physical map can hybridise to SV40 DNA (26). It is tempting to speculate that a domain of large T-antigen encoded between 12 and 22 map units plays some role in DNA replication. Two pieces of data allow for this speculation:

- It has been found that the T-antigen isolated from adenovirus-2/SV40 (Ad2/SV40) hybrids binds to polyoma virus DNA and protects specifically a region around the origin of replication (R. Tjian and B.E. Griffin, unpublished).

- The region between 12 and 22 map units, although not unusually rich in basic amino acids, nonetheless contains many more basic than acidic amino acids. It is interesting to note that all of the polyoma virus (tsa) mutants which are temperature sensitive for replication could have lesions confined within this region (27,28), but more precise mutant mapping is needed to confirm that this is indeed the case.

The C-terminal end of the SV40 large T-antigen sequence extends well beyond the end of the homology region. This C-terminal sequence, which apparently codes for the SV40 helper function (29), ends in a very proline-rich (six out of eleven) stretch of amino acids (30). Polyoma virus also has two stretches of sequence which would code for relatively proline rich areas in a protein. These lie, however, within the region which appears to have no homology with SV40. In the DNA between positions 1099 and 1144, six out of fifteen amino acids encoded are proline residues and between 1208 and 1246, five out of twelve amino acids are prolines. Whether this represents some rearrangement of sequence between the viruses, or is entirely fortuitous, is not known.

At the moment the major conclusions that appear to be allowed by our sequence studies are:

- Polyoma virus contains enough information for a protein the size of large T-antigen to be encoded entirely within the viral genome. Excluding amino acid modifying groups, a protein about 88K in size is predicted to be made.

- Polyoma virus can also code for the two other known early proteins, middle and small T-antigens, using more than one coding frame over a part of the genome.

- The potential exists for coding for additional early viral proteins, not yet identified.

- The spliced junctions in the mRNAs coding for the early proteins may be predicted and the tentative sequences of large, middle and small T-antigens obtained.

- Polyoma virus and SV40 DNAs have both notable sequence similarities and differences within their early regions.

## ACKNOWLEDGEMENTS

We thank Dr. R. Kamen for many helpful discussions. We also acknowledge the skilled technical help of Christine Maddock, Jane Walsh and Steve Barrett, and are grateful to Dr. Nina Smolar for preparation of T4-induced DNA polymerase.

Footnote: After this work was completed and the manuscript written, a paper concerning a similar part of polyoma virus appeared in press (35). Our sequence is very similar to that reported, but not identical. However, it should be noted that we use a different strain of polyoma virus (A2 strain) and strain variation in terms of sequence has already been commented upon (10).

*On leave from: The National Institute of Genetics, Misima 411, Japan.

†To whom reprint requests should be sent.

## REFERENCES

1. Ito, Y., Spurr, N. and Dulbecco (1977) Proc. Natl. Acad. Sci. USA, 74, 1259-1263.
2. Weil, R. (1978) Biochem. et Biophys. Acta 516, 301-388.
3. Seif, R. and Cuzin, F. (1977) J. Virol. 24, 721-728.
4. Eissenberg, S., Griffith, J. and Kornberg, A. (1977) Proc. Natl. Acad. Sci. USA, 74, 3198-3203.
5. Ito, Y., Brocklehurst, J.R. and Dulbecco, R. (1977) Proc. Natl. Acad. Sci. USA, 74, 4666-4670.
6. Smith, A.E., Smith, R., Griffin, B.E. and Fried, M. Submitted for publication.
7. Lania, L., Griffiths, M., Cooke, B., Ito, Y. and Fried, M. Cell, submitted for publication.
8. Ito, Y., Spurr, N. and Griffin, B.E. Submitted for publication.
9. Griffin, B.E. and Maddock, C. (1979) J. Virol., in press.

10. Soeda, E., Arrand, J.R., Smolar, N. and Griffin, B.E. (1979) Cell, 17, 357-370.

11. Griffin, B.E., Fried, M. and Cowie (1974) Proc. Natl. Acad. Sci. USA, 71, 2077-2081.

12. Maxam, A.M., and Gilbert, W. (1977), Proc. Natl. Acad. Sci. USA, 74, 560-564.

13. Ito, Y., Brocklehurst, J.R., Spurr, N., Griffiths, M., Hurst, J. and Fried, M. (1977) in INSERM colloquium, ed. May, P., Monier, R. and Weil, R., 69, 145-152.

14. Schaffhausen, B.S., Silver, J.E. and Benjamin, T.L. (1978) Proc. Natl. Acad. Sci. USA, 75, 79-83.

15. Soeda, E. and Griffin, B.E. (1978) Nature, 276, 294-298.

16. Hattori, J., Carmichael, G.G. and Benjamin, T.L. (1979) Cell, 16, 505-513.

17. Kamen, R., Favaloro, J., Parker, J., Treisman, R., Flavell, A.J., Cowie, A. and Legon, S. (1979) Differentiation, in press.

18. Hutchinson, M.A., Hunter, T. and Eckhart, W. (1978) Cell, 15, 65-77.

19. Smart, J.E., and Ito, Y. (1978) Cell, 15, 1427-1437.

20. Gannon, F., O'Hare, K., Perrin, F., LePennec, J.P., Benoist, C., Cochet, M., Breathnach, R., Royal, A., Garapin, A., Cami, B. and Chambon, P. (1979) Nature, 278, 428-434.

21. Seif, I., Khoury, G. and Dhar, R. (1979) Nucl. Acids Res. 6, 3387-3398.

22. Breathnach, R., Benoist, C., O'Hare, K., Gannon, F. and Chambon, P. (1978) Proc. Natl. Acad. Sci. USA, 75, 4853-4857.

23. Ghosh, P.K., Reddy, V.B., Swinscoe, J., Lebowitz, P. and Weissman (1978) J. Mol. Biol. 126, 813-846.

24. Proudfoot, N.J. and Brownlee, G.G. (1976) Nature, 263, 211-214.

25. Kamen, R., and Shure, H. (1976) Cell, 7, 361-371.

26. Howley, P.M., Israel, M.A., Law, M-F. and Martin, M.A. (1979) J. Biol. Chem., in press.

27. Miller, L.K. and Fried, M. (1976) J. Virol., 18, 824-832.

28. Feunteun, J., Sompayrac, L., Fluck, M. and Benjamin, T. (1976) Proc. Natl. Acad. Sci. USA, 73, 4169-4173.

29. Lebowitz, P., Kelly, T.J., Nathans, D., Lee, T.N. and Lewis, A.M. (1974) Proc. Natl. Acad. Sci. USA, 71, 441-445.

30. Fiers, W., Contreras, R., Haegeman, G., Rogiers, R., Van de Voorde, A., Van Heuverswyn, H., Van Herreweghe, J., Volckaert, G. and Ysebaert, M., Nature 273, 113-120.

31. Kamen, R., Sedat, J. and Ziff, E. (1976) J. Virol., 17, 212-218.

32. 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.

33. Van Heuverswyn, H. and Fiers, W. Submitted for publication.

34. Smart, J. and Ito, Y. (1978) Cell 15, 1427-1437.

35. Friedmann, T., Esty, A., LaPorte, P. and Deininger, P. (1979) Cell 17, 715-724.

36. Roberts, R.J. (1976) C.R.C. Crit. Rev. Biochem. 4, 123-164.