JOURNAL OF VIROLOGY, Feb. 1980, p. 606-618
0022-538X/80/02-0606/13$02.00/0

Vol. 33, No. 2

Polyoma Virus DNA: Sequence from the Late Region That Specifies the Leader Sequence for Late mRNA and Codes for VP2, VP3, and the N-Terminus of VP1†

JOHN R. ARRAND,* EIICHI SOEDA,‡ JANE E. WALSH, NINA SMOLAR, AND BEVERLY E. GRIFFIN

Imperial Cancer Research Fund Laboratories, London WC2A 3PX, England

The DNA sequence of part of the late region of the polyoma virus genome is presented. This sequence of 1,348 nucleotide pairs encompasses the leader region for late mRNA and the coding sequence for the two minor capsid proteins VP2 and VP3. The coding sequence for the N-terminus of the major capsid protein overlaps the C-terminus of VP2/VP3 by 32 nucleotide pairs. From the DNA sequence the sizes and sequences of VP2 and VP3 could be predicted. Potential splicing signals for the processing of late mRNA's could be identified. Comparisons are made between the sequence of polyoma virus DNA and corresponding regions of simian virus 40 DNA.

Polyoma virus, as well as simian virus 40 (SV40) and the human adenoviruses, has attracted much study at both biological and biochemical levels as a model system for understanding eucaryotic cellular functions in general and oncogenesis in particular. A physical map of polyoma virus DNA has been derived (22) and has provided the key to the elucidation of the general functional organization of the polyoma viral genome (13, 21). The DNA is a covalently closed circular molecule, contains about 5,290 nucleotide pairs, and can be divided functionally into regions containing early and late genes. Late in productive infection (that is, after onset of viral DNA replication) three mRNA species which sediment at 16S, 19S, and 18S appear in the cytoplasm of infected cells. These mRNA's contain coding information for the viral capsid proteins VP1, VP2, and VP3. Each mRNA consists of an untranslated leader region, a body (i.e., coding region), and a polyadenylated 3′ terminus. A combination of genetic mapping (33), protein chemistry (26), mRNA mapping (30), and in vitro protein synthesis (44, 45) has enabled the organization of the late region to be established. Hybridization analysis (11, 31) indicated that there are intervening sequences between the 5′ region and the body sequences on the physical map of the late genes. Figure 1 shows the overall relationship.

The strong similarity in the genomic organizations of polyoma virus and SV40 (13) led to the suggestion that these viral species originated from a common ancestor. Heteroduplex analysis of the two viral DNAs (8) has indicated that the strongest sequence homology occurs in the region around the N-terminus of VP1 on both DNAs. In SV40, the C-terminus of the VP2/VP3 gene overlaps the N-terminus of the VP1 gene by 122 nucleotides, and two coding reading frames are used (5, 9, 37). This has led to speculation that the same may be true of polyoma virus and that the overlap may have placed severe evolutionary constraints on the DNA (35), thus conserving the sequence in this region of the two viruses.

As a step toward obtaining a more detailed understanding of the structural organization of the polyoma viral genome in relation to its biological function and its evolutionary relationship to other papovaviruses, we sequenced the DNA. This report presents the sequence of the part of the late region of polyoma virus DNA which specifies the leader sequence of late mRNA's and the coding sequence for VP2, VP3, and the N-terminus of VP1. The sequence indicated that, contrary to expectation, the overlap regions of the late genes appear not to have been subjected to a severe constraint during evolution.

MATERIALS AND METHODS

Polyoma virus DNA. The A2 strain of polyoma virus (14) was plaque purified twice on secondary mouse embryo cells. A single large plaque isolate was used to make virus stocks by standard methods (21). DNA for sequence analysis was prepared from a Hirt extract (27) of infected 3T6 cells by previously published procedures (12, 22).

Enzymes. All restriction enzymes were prepared by standard methods (38). Phage T4-induced polynu-

† This paper and the accompanying papers are dedicated to M. G. P. Stoker on the occasion of his retirement as Director of Research of the Imperial Cancer Research Fund.

‡ On leave from the National Institute of Genetics, Misima 411, Japan.

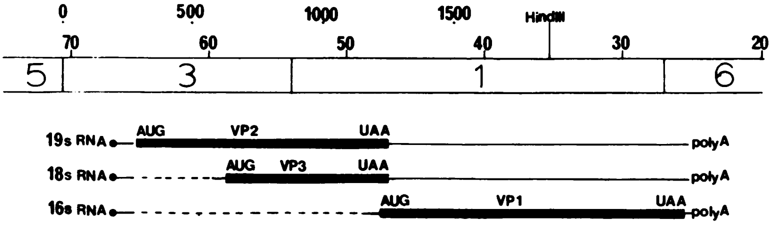

FIG. 1. *Topographical map of the polyoma virus late region. The HpaII restriction fragments (22) which define the late region are shown. The late region is defined as proceeding counterclockwise on the physical map of polyoma virus DNA from about 71 to 25 map units. In the lower part of the figure the three late mRNA species are aligned in their appropriate positions with respect to the DNA. Nucleotide numbers indicated at the top of this figure refer to the sequence (see Fig. 3). The dots represent cap structures. Mature mRNA species are represented by the narrow lines, and the coding regions are indicated by the heavy lines. The dotted lines indicate regions which are removed during mRNA maturation or splicing.*

cleotide kinase and *Escherichia coli* exonuclease III were purchased from P-L Biochemicals. The Klenow fragment of *E. coli* DNA polymerase I was obtained from Boehringer. T4-induced DNA polymerase was prepared by an unpublished procedure of R. Kamen, with minor modifications. Briefly, the cells were broken by sonication, and cell debris was removed by centrifugation. The supernatant was made 10% in polyethylene glycol and centrifuged. This supernatant was chromatographed on columns of DNA-cellulose, hydroxylapatite, DEAE-cellulose, and Bio-Gel AO.5M. Enzyme activity was followed using the assay described by Goulian et al. (19).

Preparation of fragments and sequence analysis by the chemical method. A prerequisite for sequencing by the Maxam-Gilbert method (32) is the generation of a suitable fragment of DNA labeled at only one end with ^{32}P. Polyoma virus DNA was digested with restriction enzymes under standard conditions at 37°C (1), except in the case of *Taq*I and *Bcl*I, when 65°C was used. The reaction mixture was phenol extracted, and the DNA was precipitated with ethanol. The fragments were then labeled at their 3' ends by using T4 polymerase and one α-^{32}P-labeled deoxyribonucleoside triphosphate in the presence of the three other deoxyribonucleoside triphosphates unlabeled (to minimize spurious labeling), as described previously (2). In some cases the fragments were labeled at their 5' ends using polynucleotide kinase and [γ-^{32}P]ATP as described by Maxam and Gilbert (32). All ^{32}P-labeled triphosphates were obtained from the Radiochemical Centre, Amersham, England.

Labeled fragments were separated by electrophoresis on 5% polyacrylamide gels (7) by using E buffer (42), and they were eluted either electrophoretically or by crushing and soaking (32).

To separate the labeled ends, the fragments were usually recleaved by a second restriction enzyme, and the subfragments were separated by electrophoresis on a 5 or 10% polyacrylamide gel; in some instances the two strands of the fragments were separated by denaturation and electrophoresis (32). The DNA fragments were eluted as described above and sequenced as described by Maxam and Gilbert (32). Polyacrylamide (20 or 25%) gels were used to fractionate the products of the chemical degradation reactions.

Enzymatic sequencing. The dideoxy sequencing method (40) was originally developed for use on single-stranded DNA. However, by treating linear, double-stranded DNA with exonuclease III, single-stranded extensions suitable for primed synthesis sequencing can be generated (46). Form I polyoma virus DNA was cleaved with *Bam*HI or *Hsu*I to give either full-length linear molecules or 56 and 44% fragments, respectively (22). After phenol extraction and ethanol precipitation, the resulting molecules were digested to the limit with exonuclease III by using the conditions described by Arrand et al. (1). The enzyme was inactivated by heating at 65°C for 10 min. This material was used directly as a template in the primed synthesis reaction, and an approximately threefold molar excess of double-stranded restriction fragment was used as primer. Conditions were essentially as described previously (40), and 50 μg of form I polyoma virus DNA yielded enough template for at least 30 reactions.

Primers were not removed before fractionation of the products of the sequencing reactions on an 8% polyacrylamide gel, as described by Sanger et al. (40).

RESULTS AND DISCUSSION

In this paper we present the sequence of the first 1,348 nucleotide pairs of the late region of polyoma virus DNA, from the junction of *Hpa*II fragments 5 and 3, which is conventionally defined as the boundary between the early and late regions (29, 48), through to the junction between *Hae*III fragments 16 and 4 (20). The sequence is numbered from the midpoint of the *Hpa*II recognition sequence at the junction of *Hpa*II fragments 3 and 5. Sequence data were obtained by using both the chemical degradation method of Maxam and Gilbert (32) and the enzymatic dideoxy inhibition procedure developed by Sanger et al. (40). A substantial portion of the DNA was sequenced either from both strands or by both methods. Regions of sequence which were obtained by using only one method on one strand were subjected to multiple determinations. Corroborative evidence comes from qualitative pyrimidine tract analysis (20), protein sequence data (24) (see below), and restriction enzyme mapping. Through the region under consideration here, about 40 restriction enzyme sites have been mapped independently of the sequence analysis (3, 6, 14, 20-22; Griffin, unpublished data; Arrand and Walsh, unpublished data; M. Fried, unpublished data). In every case the enzyme recognition site was found in the sequence at about the position predicted by the conventional mapping procedures. Figure 2 shows the restriction enzyme map of this region and indicates the strategy used to derive the final sequence, which is given in Fig. 3.

The sequence from positions 1 to 289 is believed to be a noncoding region (see below). A few unusual structural features which could form control signals are present. The sequence from position 1 through position 28 is probably part of the viral origin of DNA replication and has been discussed previously (48). There is a decanucleotide palindrome at positions 24 to 33. Other palindromes are tetra- or hexanucleotide sequences and appear at about the expected random frequency. The sequence between positions 62 and 129 can adopt a structure similar to that of a tRNA cloverleaf; the significance of this, if any, remains obscure. A curious feature is the frequent occurrence of tetranucleotide pairs of the form T_4 or A_4. Eight such sequences occur in a stretch of 208 nucleotide pairs between positions 82 and 290. The significance of this is not clear, although regions rich in adenine and thymine are often associated with promoter regions. This entire (noncoding) region seems fairly unusual in that it consists mainly of regions rich in adenine and thymine interspersed with regions rich in guanine and cytosine. There are no long repeating sequences. The coding region itself provides no remarkable features.

By correlating the results of other workers with those from DNA sequencing we have been able to (i) identify the genomic region which specifies the leader sequence and possible ribosome binding sites on late mRNA, (ii) suggest probable locations on the genome which specify the multiple caps found on late mRNA and the mRNA splicing signals, (iii) precisely define the translational limits of the three late proteins VP1, VP2, and VP3 and predict their sequences, (iv) make a comparison of these several functional entities with their counterparts in SV40 and BK virus, and (v) conclude that genetic drift in the gene overlap region of both polyoma virus and SV40 has not been severely constrained during evolution.

Coding region. Kamen and Shure (30) mapped the 5′ ends of the polyoma virus late 19S and 16S mRNA species at about 68 and 47 map units, respectively. In vitro protein synthe-

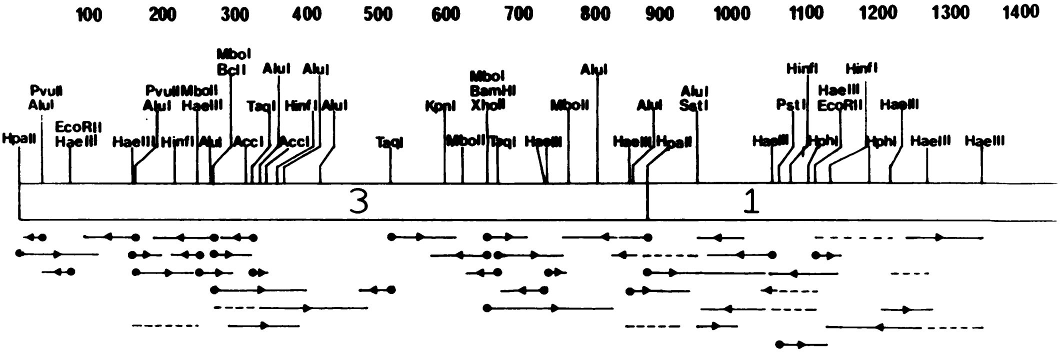

FIG. 2. *Detailed restriction map of the region between the junction of HpaII fragments 5 and 3 (70.7 map units) and the junction of HaeIII fragments 16 and 4 (45.3 map units) and a schematic diagram of the data from which the sequence (Fig. 3) was derived. The dots (●) show the positions of the ^{32}P label in fragments sequenced by the chemical method, and the solid lines represent the stretches of sequence which were obtained. The arrows indicate the polarity of reading from the gel (i.e., → indicates 5′ to 3′ and ← indicates 3′ to 5′). The dotted lines indicate primers used for enzymatic sequencing. The numbers at the top correspond to those in the sequence shown in Fig. 3.*

FIG. 3. *DNA sequence of the region of polyoma virus strain A2 shown in Fig. 2. The junction of HpaII fragments 5 and 3 is arbitrarily symmetrically divided and taken as the zero point. In keeping with the convention adopted for the early region, which has the same zero point (48), the nucleotide numbers in this paper are all negative, but for convenience the minus sign has been omitted. The sequence designated early strand has the same polarity as late mRNA (29). HpaII and HaeIII recognition sites are underlined and are designated 17/14, etc., according to the appropriate junctions on the physical map (20, 22). The initiation sites for the three late proteins VP1, VP2, and VP3 are also underlined. (In sequence studies carried out after submission of this manuscript, it was discovered that the sequence at position 88 should be TT, instead of T. This is in a noncoding part of the genome.)*

```
EARLY              10        20        30        40        50        60
STRAND 5'  GGCCTCTGCTTAATACTAAAAAAAACAGCTGTTGTCATAGTAATGATTGGGTGGAAACAT
LATE
STRAND 3'  CCGGAGACGAATTATGATTTTTTTTGTCGACAACAGTATCATTACTAACCCACCTTTGTA
           Hpa II 5/3

                   70        80        90       100       110       120
           TCTAGGCCTGGGTGGAGAGGCTTTTGCTCCTCTTGCAAAACCACACGCCCTCTGGAGGGC

           AGATCCGGACCCACCTCTCCGAAAACGAGGAGAACGTTTTGGTGTGCGGGAGACCTCCCG
               Hae III 17/14

                  130       140       150       160       170       180
           GTTGCCTAGCAACTAATTAAAAGAGGATGTCGCACGGCCAGCTGCGTCAGTTAGTCCACTT

           CAACGGATCGTTGATTAATTTTCTCCTACAGCGTGCCGGTCGACGCAGTCAATCAGGTGAA
                                              Hae III 14/15

                  190       200       210       220       230       240
           CCTGCTTAACTGACTTGACATTTTCTATTTTAAGAGTCGGGAGGAAAATTACTGTGTTGG

           GGACGAATTGACTGAACTGTAAAAGATAAAATTCTCAGCCCTCCTTTTAATGACACAACC

                  250       260       270       280       290       300
           AGGCCCTCCGCCATCTTCTGAAGCTGATCAAGTAAGTGAATTTTCAAAATGGGAGCCGCA

           TCCGGGAGGCGGTAGAAGACTTCGACTAGTTCATTCACTTAAAAGTTTTACCCTCGGCGT
           Hae III 15/5                                     VP2

                  310       320       330       340       350       360
           CTGACTATTCTAGTAGACCTCATCGAGGGATTAGCTGAAGTGTCTACCCTTACGGACTC

           CACTGATAAGATCATCTGGAGTAGCTCCCTAATCGACTTCACAGATGGGAATGCCCTGAG

                  370       380       390       400       410       420
           TCGGCAGAAGCTATTTTATCTGGAGAAGCCCTCGCTGCCCTTGATGGCGAAATTACAGCT

           AGCCGTCTTCGATAAAATAGACCTCTTCGGGAGCGACGGGAACTACCGCTTTAATGTCGA

                  430       440       450       460       470       480
           CTGACTTTGGAGGGTGTAATGAGTTCGGAGACAGCCCTAGCAACTATGGGTATTTCAGAG

           GACTGAAACCTCCCACATTACTCAAGCCTCTGTCGGGATCGTTGATACCCATAAAGTCTC

                  490       500       510       520       530       540
           GAGGTGTATGGGTTTGTCAGTACTGTGCCTGTGTTTGTAAGTCGAACAGCGGGGGCTATA

           CTCCACATACCCAAACAGTCATGACACGGACACAAACATTCAGCTTGTCGCCCCCGATAT

                  550       560       570       580       590       600
           TGGCTGATGCAGACAGTTCAAGGTGCCTCTACTATTTCCCTAGGAATACAGCGGTACCTA

           ACCGACTACGTCTGTCAAGTTCCACGGAGATGATAAAGGGATCCTTATGTCGCCATGGAT

                  610       620       630       640       650       660
           CACAACGAAGAGGTCCCTACTGTAAATAGAAATATGGCGTTGATACCATGGCGGGATCCA

           GTGTTGCTTCTCCAGGGATGACATTTATCTTTATACCGCAACTATGGTACCGCCCTAGGT
                                            VP3

                  670       680       690       700       710       720
           GCCCTTCTCGATATATACTTCCCAGGAGTTAATCAGTTTGCTCATGCTCTAAATGTAGTA

           CGGGAAGAGCTATATATGAAGGGTCCTCAATTAGTCAAACGAGTACGAGATTTACATCAT

                  730       740       750       760       770       780
           CATGATTGGGGCCATGGCCTACTTCATTCCGTGGGAAGATATGTGTGGCAAATGGTTGTG

           GTACTAACCCCGGTACCGGATGAAGTAAGGCACCCTTCTATACACACCGTTTACCAACAC
                    Hae III 5/  Hae III /12

                  790       800       810       820       830       840
           CAGGAAACACAACACAGACTGGAAGGAGCTGTAAGAGAACTAACTGTAAGACAGACACAT

           GTCCTTTCTGTTGTGTCTGACCTTCCTCGACATTCTCTTGATTGACATTCTGTCTGTGTA

                  850       860       870       880       890       900
           ACATTCCTGGATGGCCTAGCTAGGCTACTTGAAAACACCCGGTGGGTGGTTTCTAATGCT

           TGTAAGGACCTACCGGATCGATCCGATGAACTTTTGTGGGCCACCCACCAAAGATTACGA
                       Hae III 12/9               Hpa II 3/1

                  910       920       930       940       950       960
           CCTCAGTCAGCCATAGATGCAATCAACAGAGGTGCCTCATCAGTGAGCTCAGGGTACTCA

           GGAGTCAGTCGGTATCTACGTTAGTTGTCTCCACGGAGTAGTCACTCGAGTCCCATGAGT

                  970       980       990      1000      1010      1020
           TCACTAAGTGACTATTATAGGCAACTAGGTCTTAATCCACCACAGAGGAGAGCACTCTTT

           AGTGATTCACTGATAATATCCGTTGATCCAGAATTAGGTGGTGTCTCCTCTCGTGAGAAA

                 1030      1040      1050      1060      1070      1080
           AATCGCATAGAAGGGAGCATGGGAAATGGTGGGCCTACCCCTGCAGCACATATACAGGAT

           TTAGCGTATCTTCCCTCGTACCCTTTACCACCCGGATGGGGACGTCGTGTATATGTCCTA
                                           Hae III 9/18

                 1090      1100      1110      1120      1130      1140
           GAGTCAGGTGAGGTGATAAAGTTCTATCAGGCCCCAGGTGGTGCCCACCAAAGAGTCACT

           CTCAGTCCACTCCACTATTTCAAGATAGTCCGGGGTCCACCACGGGTGGTTTCTCAGTGA
                                        Hae III 18/13

                 1150      1160      1170      1180      1190      1200
           CCTGACTGGATGCTTCCTTTAATTCTAGGGCTGTACGGTGATATCACACCTACTTGGGCA

           GGACTGACCTACGAAGGAAATTAAGATCCCGACATGCCACTATAGTGTGGATGAACCCGT

                 1210      1220      1230      1240      1250      1260
           ACAGTCATAGAGGAAGATGGCCCCCAAAAGAAAAAGCGGCGTCTCTAAATGCGAGACAAA

           TGTCAGTATCTCCTTCTACCGGGGGTTTTCTTTTTCGCCGCAGAGATTTACGCTCTGTTT
                             Hae III 13/19
                           VP1

                 1270      1280      1290      1300      1310      1320
           ATGTACAAAGGCCTGTCCAAGACCCGCATCCGTTCCCAAACTGCTTATTAAAGGGGGTAT

           TACATGTTTCCGGACAGGTTCTGGGCGTAGGCAAGGGTTTGACGAATAATTTCCCCCATA
                    Hae III 19/16

                 1330      1340
           GGAGGTGCTGGATCTTGTGACAGGGCC

           CCTCCACGACCTAGAACACT GTCCCGG
                                  Hae III 16/4
```

sis showed that the 19S message codes for VP2 (45) which, as shown by peptide mapping, shares common sequences with VP3 (24). Similarly, the 16S message was shown to code for the structurally distinct major capsid protein, VP1. Subsequently, Siddell and Smith (44) showed that a third messenger species, 18S, codes for VP3. Examination of the appropriate areas of the DNA sequence for potential ATG initiation codons followed by long, in-phase stretches of uninterrupted coding sequence suggested that the ATG at positions 290 to 292 could be the initiation codon for VP2 and that the ATG at positions 635 to 637 could initiate VP3. The ATG at positions 1,218 to 1,220 seemed a likely initiator for VP1. Comparison with sequences from SV40 (9, 37) suggested also that these might be the initiation points for translation. However, final proof was provided by aligning the partial N-terminal amino acid sequences of VP1, VP2, and VP3 synthesized in vitro (25) with the amino acid sequences predicted from the DNA sequence. This correspondence is shown in Fig. 4.

In SV40, the N-terminal alanine of mature VP1 is immediately preceded by an ATG codon. However, another ATG occurs two codons before this (9), and, at least in vitro, this latter codon is used as the initiator (A. Mellor, R. Hewick, and A. E. Smith, personal communication). Polyoma virus, on the other hand (Fig. 3), contains only the second ATG, i.e., the one immediately adjacent to the (presumably) N-terminal alanine.

From the start of VP2 at position 290 there is a continuous reading frame open to positions 1,247 through 1,249, where reading is then closed by a TAA termination codon. The two alternative reading frames are closed by multiple termination codons (Fig. 5). This gives lengths of 319 amino acids for VP2 and 204 amino acids for VP3, corresponding to molecular weights of about 34,600 and 22,800, respectively. These figures are in good agreement with the molecular weights of 35,000 and 23,000 obtained by sodium dodecyl sulfate-polyacrylamide gel electrophoresis (24). The predicted amino acid compositions are also in excellent agreement with those directly determined by R. Hewick (Ph.D. thesis, Council of National Academic Awards, London, England, 1977).

A comparison between the nucleotide and amino acid sequences of VP2/VP3 of polyoma virus and the sequences of VP2/VP3 of SV40 is shown in Fig. 6. It is apparent that the hydrophobic N-terminal region of VP2 is quite highly conserved in the two viruses. Through this part of the molecule (amino acids 1 through 122 [Fig. 6]) 48% of the DNA sequence shows homology,

```
In vitro sequence            Ala  Pro  Lys  -    Lys  -    -    -    -    Lys  -    -
VP1 Predicted SequenceMet    Ala  Pro  Lys  Arg  Lys  Ser  Gly  Val  Ser  Lys  Cys  Glu
DNA sequence          A T G, G C C, C C C, A A A, A G A,A A A, A G C, G G C, G T C, T C T, A A A, T G C, GA G,

 -    Lys  -    -    Lys  Ala  -    Pro
 Thr  Lys  Cys  Thr  Lys  Ala  Cys  Pro
A C A, A A A, T G T, A C A, A A G, G C C, T G T, C C A.


VP2 Met  -    -    -    Leu  -    -    Leu  -    -    Leu
    Met  Gly  Ala  Ala  Leu  Thr  Ile  Leu  Val  Asp  Leu
 A T G, G G A, G C C, G C A, C T G, A C T, A T T, C T A, G T A, G A C, C T C


VP3       -    Leu  -    Pro  -    -    -    Pro  -    Leu  Leu  -    -    -    -
     Met  Ala  Leu  Ile  Pro  Trp  Arg  Asp  Pro  Ala  Leu  Leu  Asp  Ile  Tyr  Phe
    A T G, G C G, T T G, A T A, C C A, T G G, C G G, G A T, C C A, G C C, C T T, C T C, G A T, A T A, T A C, T T C,


     Pro  -    -    -    -    -    -    -    -    Leu
     Pro  Gly  Val  Asn  Gln  Phe  Ala  His  Ala  Leu
    C.C A,G G A, G T T, A A T, C A G, T T T, G C T, C A T, G C T, C T A
```

FIG. 4. *Correlation of in vitro protein sequence with the predicted amino acid sequences for the N-termini of the three late proteins. The DNA sequences (taken from Fig. 3) and the amino acid sequences predicted from them are aligned with the partial N-terminal sequences obtained by Hewick et al. (25) from in vitro-synthesized late proteins. The sequences are in perfect agreement for 20, 11, and 25 sequencer cycles of VP1, VP2, and VP3, respectively.*

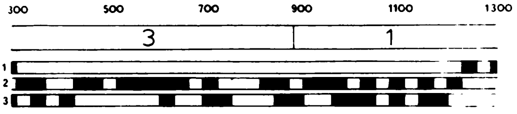

FIG. 5. *Schematic diagram showing part of the late region of polyoma virus DNA divided into three coding frames, which are shown relative to the HpaII physical map. The numbers at the top correspond to the nucleotide numbers in Fig. 3. Whenever a termination codon occurs in the early strand sequence (i.e., the strand having the same polarity as late mRNA) within 10 triplet codons, a solid block is drawn to indicate this. Frame 1 starts at position 290, frame 2 starts at position 291, and frame 3 starts at position 292. VP2 starts at position 290 in frame 1 and terminates with a TAA codon at positions 1,247 to 1,249. VP1 starts at position 1,218 in frame 2.*

whereas 35% of the amino acids are homologous. If, however, conservative amino acid changes are scored as homologous (e.g., valine to leucine), then this figure becomes 56%. Another feature which probably serves to maintain the overall conformation and spatial arrangement of the charged groups in the protein molecule, but which does not score in the linear comparison described above, is the occasional appearance of amino acid inversions; e.g., in SV40 VP2, Ala-Glu at positions 27 and 28 (Fig. 6) reads Glu-Ala in polyoma virus. Moreover, the same codons are used in the specification of these amino acids. A similar phenomenon occurs within VP3 at positions 130 and 131 and, nearer its C-terminus, at positions 193 and 194.

It is noteworthy that of the 115 amino acids unique to VP2 (that is, not found in VP3), over one-half have aliphatic side chains (alanine, glycine, isoleucine, leucine, methionine, and valine). The hydrophobicity of this region is also increased by the presence of tryptophan (one residue), phenylalanine (two residues), and tyrosine (two residues). The N-terminal region of VP2 in SV40 is characterized by a very high alanine content. In polyoma virus the preponderance is not so acute, there being only 14 alanine residues, compared with 28 in SV40. Hydrophobic N-terminal regions are often found in precursors of proteins intended for secretion through membranes. In these proteins the hydrophobic signal sequence is subsequently processed away by proteolytic cleavage. As far as is known, however, polyoma virus VP2 maintains its hydrophobic N-terminus when it is incorporated as a viral structural protein. It may be that the long hydrophobic tail of VP2 is associated with some anchoring function onto cellular membranes.

The central portion of VP2 (i.e., from the start of VP3 onward) is of much more average amino acid composition. The homology with SV40 progressively decreases and then suddenly increases toward the C-terminus, which is quite basic. (Three of the four lysines found in VP2 occur in the last six residues.) When the proteins are aligned to maximize homology, the C-terminus of SV40 VP2/VP3 continues 28 amino acids beyond the termination of polyoma virus VP2/VP3 and is significantly more basic. It has been suggested (5) that the basic tail of SV40 VP2/VP3 may be involved in interaction with DNA during virion maturation. The same could also apply to the polyoma virus proteins. From amino acid 266 to the end of polyoma virus VP2, there is a high degree of homology with SV40. A total of 64% of the amino acids are identical; when conservative substitutions are counted, this figure is 73%. This is indeed a remarkable degree of homology (Fig. 6). The DNAs show 62% homology.

The TAA termination codon for VP2/VP3 occurs at positions 1,247 to 1,249, and the initiation codon for VP1 lies at positions 1,218 to 1,220. Thus, there is an out-of-phase overlap of 32 nucleotide pairs between the C-terminal end of the VP2/VP3 gene and the N-terminal end of the VP1 gene. This is in contrast to SV40, which has a larger overlap between the two genes (122 nucleotides) (5, 9, 37).

Electron microscopic examination of heteroduplexes between polyoma virus and SV40 DNAs revealed a single region of relatively strong homology between the two species, which was located at 47 to 52 map units in polyoma virus and 0.93 to 0.98 map units in SV40 (8). In SV40 this area includes the 122-nucleotide pair overlap between VP2/VP3 and VP1. It has been argued that since in the overlap region any mutation in one gene would also affect the other, this would impose a severe evolutionary constraint to conserve this region and thus explain the observed homology (9, 35). However, it was found that the overlap region in polyoma virus consists of only 32 nucleotide pairs, and when the sequences of polyoma virus and SV40 DNAs were aligned (47), the region of polyoma viral DNA which corresponded to the SV40 overlap region showed only very limited homology. Therefore, in this case the presence of overlapping genes does not appear to restrict severely evolutionary divergence.

A similar alignment of polyoma virus and SV40 sequences near the C-termini of the large-T antigens and the VP1 species reveals a much stronger homology than that seen near the C-terminus of VP2/VP3 (47, 48). Therefore, at present we cannot understand why the experimental observation placed the maximum homology in the overlap region.

Almost all possible codons are used in the specification of VP2/VP3 (the exceptions being

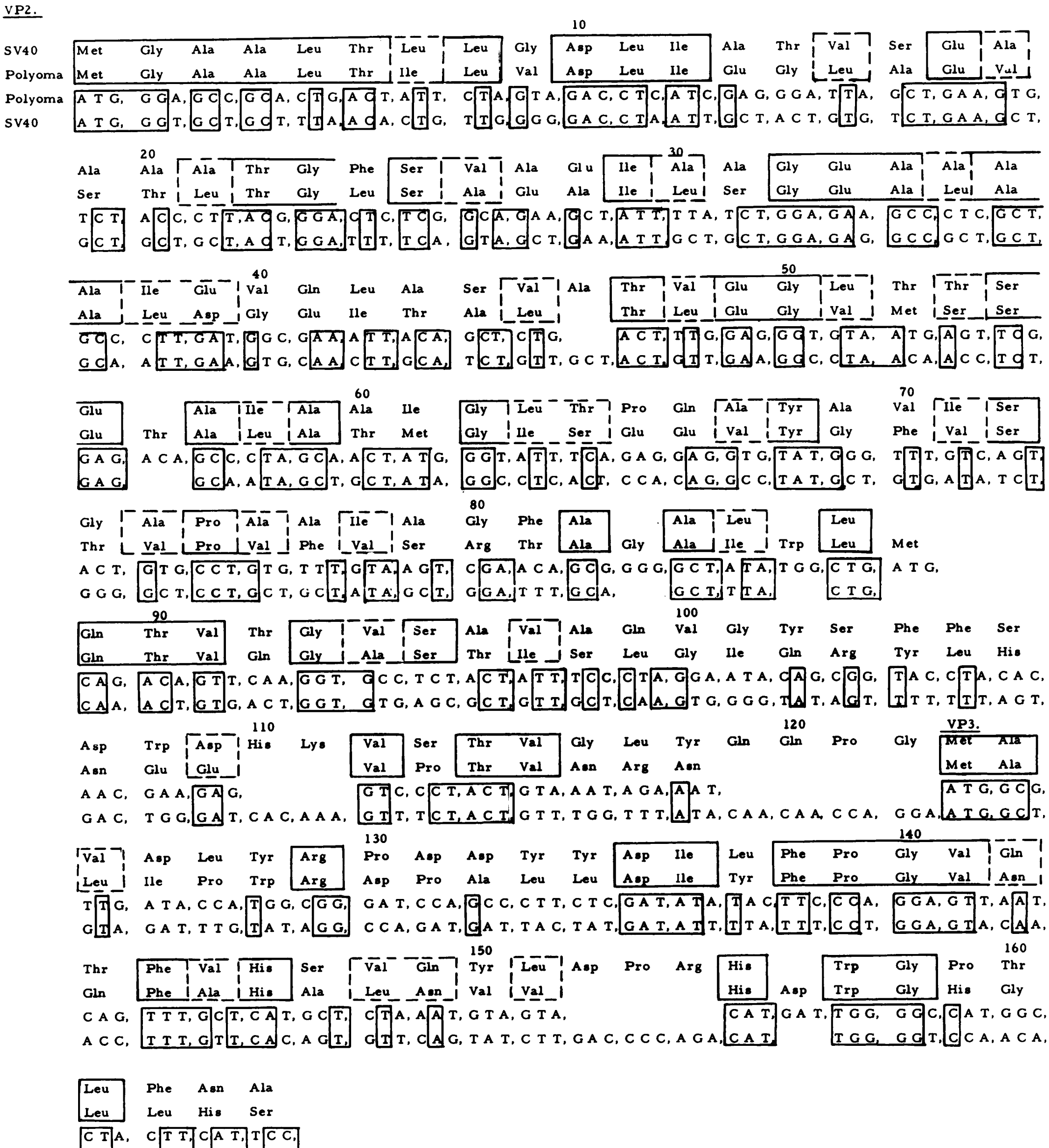
VP2.
10
SV40 Met Gly Ala Ala Leu Thr Leu Leu Gly Asp Leu Ile Ala Thr Val Ser Glu Ala
Polyoma Met Gly Ala Ala Leu Thr Ile Leu Val Asp Leu Ile Glu Gly Leu Ala Glu Val
Polyoma ATG, GGA, GCC, GCA, CTG, ACT, ATT, CTA, GTA, GAC, CTC, ATC, GAG, GGA, TTA, GCT, GAA, GTG,
SV40 ATG, GGT, GCT, GCT, TTA, ACA, CTG, TTG, GGG, GAC, CTA, ATT, GCT, ACT, GTG, TCT, GAA, GCT,
20
30
Ala Ala Ala Thr Gly Phe Ser Val Ala Glu Ile Ala Ala Gly Glu Ala Ala Ala
Ser Thr Leu Thr Gly Leu Ser Ala Glu Ala Ile Leu Ser Gly Glu Ala Leu Ala
TCT, ACC, CTT, ACG, GGA, CTC, TCG, GCA, GAA, GCT, ATT, TTA, TCT, GGA, GAA, GCC, CTC, GCT,
GCT, GCT, GCT, ACT, GGA, TTT, TCA, GTA, GCT, GAA, ATT, GCT, GCT, GGA, GAG, GCC, GCT, GCT,
40
50
Ala Ile Glu Val Gln Leu Ala Ser Val Ala Thr Val Glu Gly Leu Thr Thr Ser
Ala Leu Asp Gly Glu Ile Thr Ala Leu Thr Leu Glu Gly Val Met Ser Ser
GCC, CTT, GAT, GGC, GAA, ATT, ACA, GCT, CTG, ACT, TTG, GAG, GGT, GTA, ATG, AGT, TCG,
GCA, ATT, GAA, GTG, CAA, CTT, GCA, TCT, GTT, GCT, ACT, GTT, GAA, GGC, CTA, ACA, ACC, TCT,
60
70
Glu Ala Ile Ala Ala Ile Gly Leu Thr Pro Gln Ala Tyr Ala Val Ile Ser
Glu Thr Ala Leu Ala Thr Met Gly Ile Ser Glu Glu Val Tyr Gly Phe Val Ser
GAG, ACA, GCC, CTA, GCA, ACT, ATG, GGT, ATT, TCA, GAG, GAG, GTG, TAT, GGG, TTT, GTC, AGT,
GAG, GCC, ATA, GCT, GCT, ATA, GGC, CTC, ACT, CCA, CAG, GCC, TAT, GCT, GTG, ATA, TCT,
80
Gly Ala Pro Ala Ala Ile Ala Gly Phe Ala Ala Leu Leu
Thr Val Pro Val Phe Val Ser Arg Thr Ala Gly Ala Ile Trp Leu Met
ACT, GTG, CCT, GTG, TTT, GTA, AGT, CGA, ACA, GCG, GGG, GCT, ATA, TGG, CTG, ATG,
GGG, GCT, CCT, GCT, GCT, ATA, GCT, GGA, TTT, GCA, GCT, TTA, CTG,
90
100
Gln Thr Val Thr Gly Val Ser Ala Val Ala Gln Val Gly Tyr Ser Phe Phe Ser
Gln Thr Val Gln Gly Ala Ser Thr Ile Ser Leu Gly Ile Gln Arg Tyr Leu His
CAG, ACA, GTT, CAA, GGT, GCC, TCT, ACT, ATT, TCC, CTA, GGA, ATA, CAG, CGG, TAC, CTA, CAC,
CAA, ACT, GTG, ACT, GGT, GTG, AGC, GCT, GTT, GCT, CAA, GTG, GGG, TAT, AGT, TTT, TTT, AGT,
110
120
VP3.
Asp Trp Asp His Lys Val Ser Thr Val Gly Leu Tyr Gln Gln Pro Gly Met Ala
Asn Glu Glu Val Pro Thr Val Asn Arg Asn Met Ala
AAC, GAA, GAG, GTC, CCT, ACT, GTA, AAT, AGA, AAT, ATG, GCG,
GAC, TGG, GAT, CAC, AAA, GTT, TCT, ACT, GTT, TGG, TTT, ATA, CAA, CAA, CCA, GGA, ATG, GCT,
130
140
Val Asp Leu Tyr Arg Pro Asp Asp Tyr Tyr Asp Ile Leu Phe Pro Gly Val Gln
Leu Ile Pro Trp Arg Asp Pro Ala Leu Leu Asp Ile Tyr Phe Pro Gly Val Asn
TTG, ATA, CCA, TGG, CGG, GAT, CCA, GCC, CTT, CTC, GAT, ATA, TAC, TTC, CCA, GGA, GTT, AAT,
GTA, GAT, TTG, TAT, AGG, CCA, GAT, GAT, TAC, TAT, GAT, ATT, TTA, TTT, CCT, GGA, GTA, CAA,
150
160
Thr Phe Val His Ser Val Gln Tyr Leu Asp Pro Arg His Trp Gly Pro Thr
Gln Phe Ala His Ala Leu Asn Val Val His Asp Trp Gly His Gly
CAG, TTT, GCT, CAT, GCT, CTA, AAT, GTA, GTA, CAT, GAT, TGG, GGC, CAT, GGC,
ACC, TTT, GTT, CAC, AGT, GTT, CAG, TAT, CTT, GAC, CCC, AGA, CAT, TGG, GGT, CCA, ACA,
Leu Phe Asn Ala
Leu Leu His Ser
CTA, CTT, CAT, TCC,
CTT, TTT, AAT, GCC,

FIG. 6. *Comparison of the nucleotide and predicted amino acid sequences of polyoma virus and SV40 VP2 proteins. Gaps have been left to maximize homology. The sequence is numbered by amino acids as a composite of the polyoma virus and SV40 sequences. The solid boxes show perfect homologies, and the dotted boxes indicate conservative amino acid substitutions. The initiating methionine of VP3 at position 123 is overlined. Overall, the amino acid sequences show 34% homology (48% counting conservative substitutions). The nucleotide sequences are 42% homologous. In this context, it should be noted that homologies between amino acid sequences and DNA sequences cannot be directly compared. That is, the expected homology between any two unrelated species would be 25% at the DNA level and 5% at the amino acid level. Thus, in terms of their protein sequences, there is a remarkable conservation of amino acids between polyoma virus and SV40. There appears to be a contour of homology in that, although the N- and C-terminal regions are well conserved, the central parts of the VP2 proteins show more divergence.*

the codons for cysteine, which does not appear in the sequence, and the CCG proline codon). As has been observed in other systems (9, 18, 34, 37, 39), the frequency of usage of different codons which specify the same amino acid appears to be nonrandom (Fig. 7). However, in polyoma

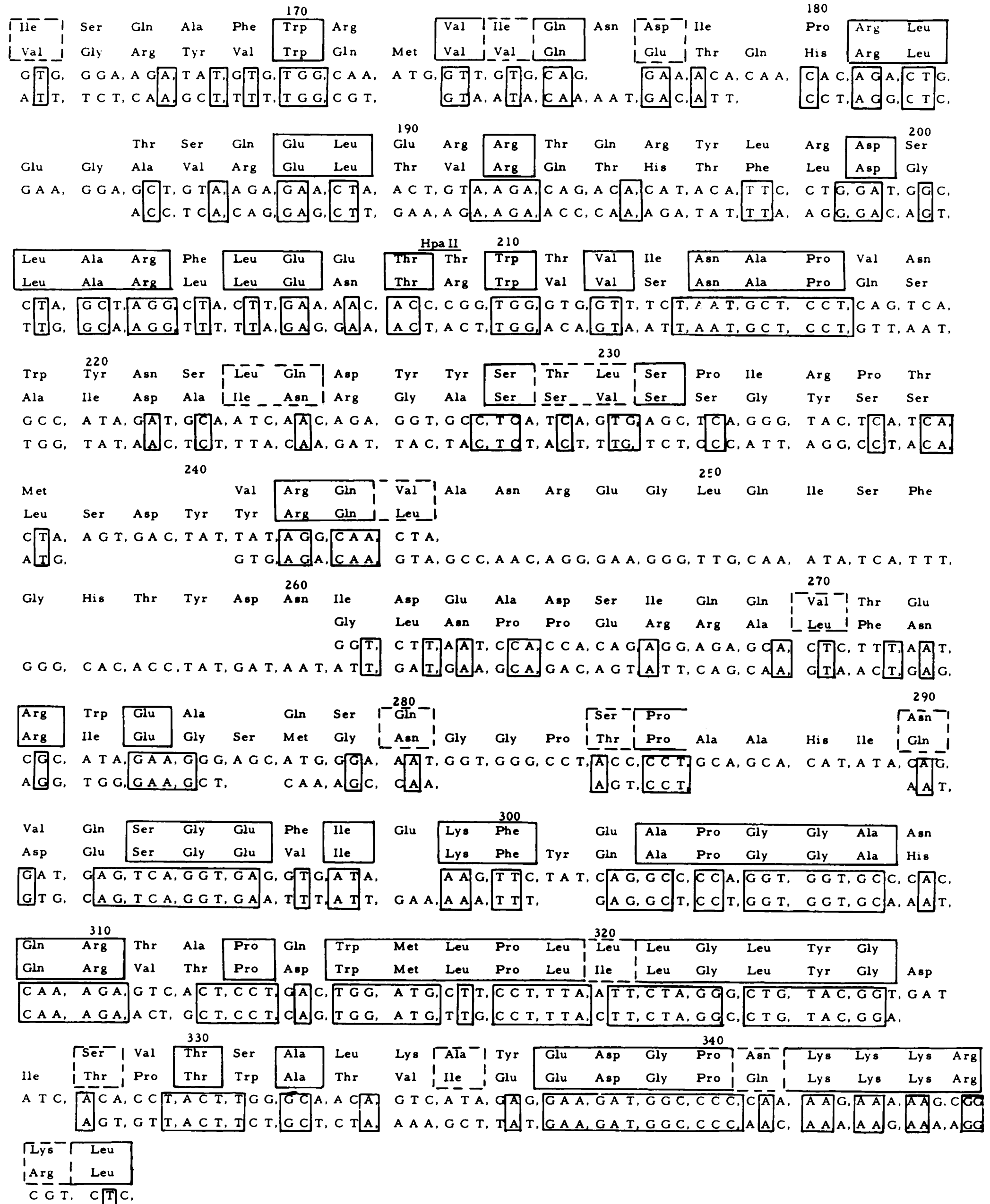
170 180
Ile Ser Gln Ala Phe Trp Arg Val Ile Gln Asn Asp Ile Pro Arg Leu
Val Gly Arg Tyr Val Trp Gln Met Val Val Gln Glu Thr Gln His Arg Leu
GTG, GGA, AGA, TAT, GTG, TGG, CAA, ATG, GTT, GTG, CAG, GAA, ACA, CAA, CAC, AGA, CTG,
ATT, TCT, CAA, GCT, TTT, TGG, CGT, GTA, ATA, CAA, AAT, GAC, ATT, CCT, AGG, CTC,
190 200
Thr Ser Gln Glu Leu Glu Arg Arg Thr Gln Arg Tyr Leu Arg Asp Ser
Glu Gly Ala Val Arg Glu Leu Thr Val Arg Gln Thr His Thr Phe Leu Asp Gly
GAA, GGA, GCT, GTA, AGA, GAA, CTA, ACT, GTA, AGA, CAG, ACA, CAT, ACA, TTC, CTG, GAT, GGC,
ACC, TCA, CAG, GAG, CTT, GAA, AGA, AGA, ACC, CAA, AGA, TAT, TTA, AGG, GAC, AGT,
HpaII 210
Leu Ala Arg Phe Leu Glu Glu Thr Thr Trp Thr Val Ile Asn Ala Pro Val Asn
Leu Ala Arg Leu Leu Glu Asn Thr Arg Trp Val Val Ser Asn Ala Pro Gln Ser
CTA, GCT, AGG, CTA, CTT, GAA, AAC, ACC, CGG, TGG, GTG, GTT, TCT, AAT, GCT, CCT, CAG, TCA,
TTG, GCA, AGG, TTT, TTA, GAG, GAA, ACT, ACT, TGG, ACA, GTA, ATT, AAT, GCT, CCT, GTT, AAT,
220 230
Trp Tyr Asn Ser Leu Gln Asp Tyr Tyr Ser Thr Leu Ser Pro Ile Arg Pro Thr
Ala Ile Asp Ala Ile Asn Arg Gly Ala Ser Ser Val Ser Ser Gly Tyr Ser Ser
GCC, ATA, GAT, GCA, ATC, AAC, AGA, GGT, GCC, TCA, TCA, GTG, AGC, TCA, GGG, TAC, TCA, TCA,
TGG, TAT, AAC, TCT, TTA, CAA, GAT, TAC, TAC, TCT, ACT, TTG, TCT, CCC, ATT, AGG, CCT, ACA,
240 250
Met Val Arg Gln Val Ala Asn Arg Glu Gly Leu Gln Ile Ser Phe
Leu Ser Asp Tyr Tyr Arg Gln Leu
CTA, AGT, GAC, TAT, TAT, AGG, CAA, CTA,
ATG, GTG, AGA, CAA, GTA, GCC, AAC, AGG, GAA, GGG, TTG, CAA, ATA, TCA, TTT,
260 270
Gly His Thr Tyr Asp Asn Ile Asp Glu Ala Asp Ser Ile Gln Gln Val Thr Glu
Gly Leu Asn Pro Pro Glu Arg Arg Ala Leu Phe Asn
GGT, CTT, AAT, CCA, CCA, CAG, AGG, AGA, GCA, CTC, TTT, AAT,
GGG, CAC, ACC, TAT, GAT, AAT, ATT, GAT, GAA, GCA, GAC, AGT, ATT, CAG, CAA, GTA, ACT, GAG,
280 290
Arg Trp Glu Ala Gln Ser Gln Ser Pro Asn
Arg Ile Glu Gly Ser Met Gly Asn Gly Gly Pro Thr Pro Ala Ala His Ile Gln
CGC, ATA, GAA, GGG, AGC, ATG, GGA, AAT, GGT, GGG, CCT, ACC, CCT, GCA, GCA, CAT, ATA, CAG,
AGG, TGG, GAA, GCT, CAA, AGC, CAA, AGT, CCT, AAT,
300
Val Gln Ser Gly Glu Phe Ile Glu Lys Phe Glu Ala Pro Gly Gly Ala Asn
Asp Glu Ser Gly Glu Val Ile Lys Phe Tyr Gln Ala Pro Gly Gly Ala His
GAT, CAG, TCA, GGT, GAG, GTG, ATA, AAG, TTC, TAT, CAG, GCC, CCA, GGT, GGT, GCC, CAC,
GTG, CAG, TCA, GGT, GAA, TTT, ATT, GAA, AAA, TTT, GAG, GCT, CCT, GGT, GGT, GCA, AAT,
310 320
Gln Arg Thr Ala Pro Gln Trp Met Leu Pro Leu Leu Leu Gly Leu Tyr Gly
Gln Arg Val Thr Pro Asp Trp Met Leu Pro Leu Ile Leu Gly Leu Tyr Gly Asp
CAA, AGA, GTC, ACT, CCT, GAC, TGG, ATG, CTT, CCT, TTA, ATT, CTA, GGG, CTG, TAC, GGT, GAT
CAA, AGA, ACT, GCT, CCT, CAG, TGG, ATG, TTG, CCT, TTA, CTT, CTA, GGC, CTG, TAC, GGA,
330 340
Ser Val Thr Ser Ala Leu Lys Ala Tyr Glu Asp Gly Pro Asn Lys Lys Lys Arg
Ile Thr Pro Thr Trp Ala Thr Val Ile Glu Glu Asp Gly Pro Gln Lys Lys Lys Arg
ATC, ACA, CCT, ACT, TGG, GCA, ACA, GTC, ATA, GAG, GAA, GAT, GGC, CCC, CAA, AAG, AAA, AAG, CGG
AGT, GTT, ACT, TCT, GCT, CTA, AAA, GCT, TAT, GAA, GAT, GGC, CCC, AAC, AAA, AAG, AAA, AGG
Lys Leu
Arg Leu
CGT, CTC,
AAG, TTG,

FIG. 6. *Continued*

virus VP2/VP3 the bias toward certain codons and against others is not as dramatic as that found, for example, in SV40 VP2/VP3. An examination of the three possible reading frames reveals that the three termination codons occur with about equal frequency.

Polyoma virus DNA, as well as the DNA of many other mammalian viruses, contains the CG dinucleotide sequence relatively infrequently (49). It is, however, found in the coding sequence for VP2/VP3. Whereas codons for serine, proline, threonine, and alanine show a strong bias against the NCG triplet, 7 of the 19 arginine codons were found to be CGN (Fig. 7). However, the overall distribution of CG (i.e., in both coding and noncoding frames) appears to be nonrandom. For example, from the junction of *Hpa*II fragments 5 and 3 through the apparently un-

	U	C	A	G	
U	Phe 4 (2)	Ser 4 (1)	Tyr 5 (4)	Cys 0 (0)	U
	3 (3)	2 (1)	4 (3)	0 (0)	C
	Leu 3 (1)	8 (7)	Ochre	Opal	A
	2 (1)	2 (0)	Amber	Trp 7 (6)	G
C	Leu 7 (5)	Pro 8 (6)	His 6 (6)	Arg 1 (1)	U
	6 (3)	1 (1)	3 (2)	1 (1)	C
	13 (8)	6(6)	Gln 7 (6)	1 (0)	A
	6 (3)	0 (0)	9 (7)	4 (3)	G
A	Ile 6 (1)	Thr 9 (3)	Asn 8 (6)	Ser 4 (3)	U
	3 (2)	3 (2)	3 (2)	2 (2)	C
	9 (7)	9 (5)	Lys 1 (1)	Arg 8 (7)	A
	Met 8 (4)	1 (0)	3 (3)	3 (3)	G
G	Val 4 (3)	Ala 10 (5)	Asp 9 (8)	Gly 10 (7)	U
	4 (2)	10 (5)	3 (2)	5 (4)	C
	7 (4)	8 (5)	Glu 10 (5)	9 (4)	A
	10 (6)	2 (1)	9 (3)	6 (4)	G

FIG. 7. *Frequency of codon usage for the VP2 protein of polyoma virus. The numbers in parentheses refer to VP3.*

transcribed region, the leader region (see below), and the region of VP2 that is not part of VP3, CG occurs 18 times. Through the VP3 coding sequence (an equivalent length of sequence) it occurs only seven times. The significance of this, if any, is not known.

An examination of the sequence of the complementary strand for potential coding regions revealed an ATG codon at positions 38 to 36, which is followed by 52 in-phase codons. This potential coding sequence includes the region which contains the origin of replication and is closed by a TAA termination codon at positions 122 to 124 in the early region sequence (48). It is not known at present whether this has any physiological significance. Similar potential short coding regions have been identified in other areas of polyoma virus DNA and also in SV40 DNA, but again the potential putative polypeptides have not been identified.

Late leader region. In studies of polyoma virus late mRNA's, the three late identified messages have been shown to share a common leader, which in itself consists of a three- to fivefold imperfect repeat of a fundamental sequence (31). Moreover, it has been found that each species of polyoma viral late mRNA contains multiple cap structures (11). These structural features of the messages were mapped between 66 and 71 map units on the polyoma virus genome, which roughly corresponds to the first 267 nucleotides in the sequence given (Fig. 3).

Four large T1 RNase oligoribonucleotides from the reiterated leader region have been characterized and partially sequenced (31). Examination of the appropriate region of the DNA sequence allows the leader sequence (as defined by the four T1 oligonucleotides) to be located within a continuous stretch, from nucleotide positions 226 to 273 (Fig. 8). Inspection of the surrounding sequence fails to reveal a duplication of the genomic DNA. It has been suggested that the generation of late mRNA may involve the splicing of tandem full-length transcripts of the entire DNA to form the mature functional mRNA molecules. This could lead to reiteration of a sequence not found in DNA in the mRNA. It is not clear where the limits of the reiteration lie since the T1 oligonucleotides immediately to the left of position 225 and to the right of 273 are small and cannot be quantitated with accuracy. However, the large T1 oligonucleotides from positions 199 to 215 and positions 274 to 292 (which contains the VP2 initiation codon) were not found to be reiterated in the mRNA. Although this further defines the reiteration limits, the exact definition must await RNA sequencing data.

One potentially significant difference between the RNA data and the DNA sequence is apparent. One of the reiterated T1 oligonucleotides from late leader RNA was characterized as AUC(A)$_n$G, which can be aligned with the DNA between positions 268 and 273. However, data on the RNA estimate $n = 3$ (31), whereas the DNA sequence predicts $n = 2$.

The interesting possibility arises that splices involved in the generation of late mRNA's occur at this point in such a fashion that RNA molecules containing a stretch of adenine residues longer than that found in the DNA are produced. For this to occur and for the rule which says that intervening sequences (that is, those removed during splicing) contain GT at their 5′ ends and AG at their 3′ ends to be obeyed, (the

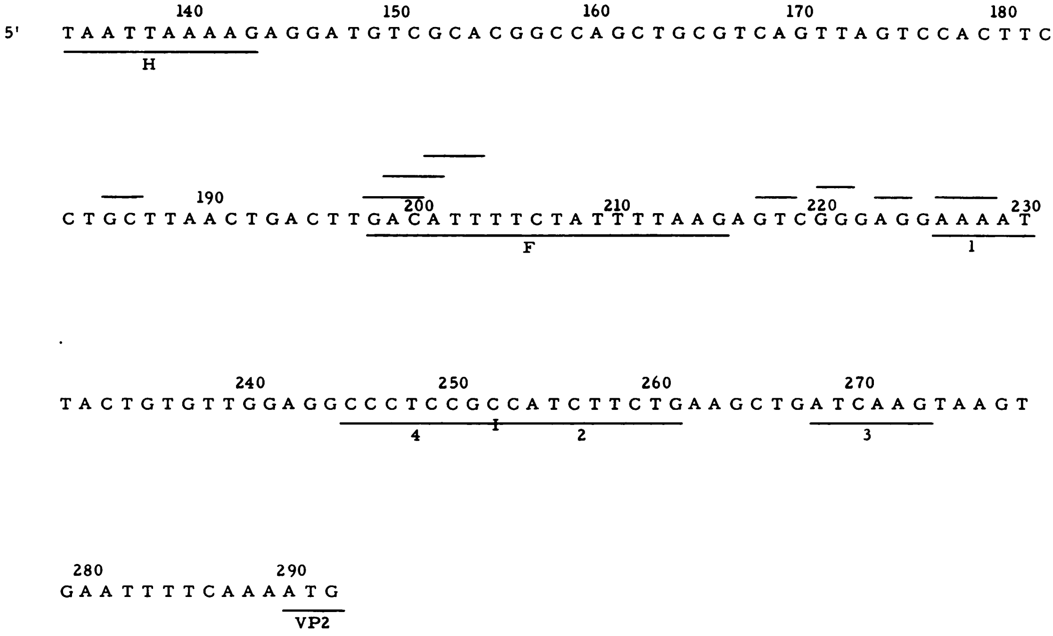
140 150 160 170 180
5' TAATTAAAAGAGGATGTCGCACGGCCAGCTGCGTCAGTTAGTCCACTTC
H
190 200 210 220 230
CTGCTTAACTGACTTGACATTTTCTATTTTAAGAGTCGGGAGGAAAAT
F
1
240 250 260 270
TACTGTGTTGGAGGCCCTCCGCCATCTTCTGAAGCTGATCAAGTAAGT
4
2
3
280 290
GAATTTTCAAAATG
VP2

FIG. 8. *Features of the leader region. The four reiterated T1 oligonucleotides characterized by Legon et al. (31) as defining the reiterated late leader sequence are underlined and designated 1 through 4. F indicates the large T1 oligonucleotide identified as being at the 5′ end of a staggered set of late mRNA's (10). The overlinings indicate all the cap I and cap II sequences identified by Flavell et al. (11). The only cap sequences which have been positively positioned are the three at the 5′ end of oligonucleotide F; the rest represent a minimum number of potential capping points within a minimum length of the genome. H indicates the sequence which resembles the Hogness box (see text). Note the frequent occurrence of sequences of the form T_4 or A_4 within this region.*

so-called GT-AG rule [4, 15]) requires the sequence AGAG at the 3′ splicing junction. S1 gel mapping studies suggest that splicing junctions occur around nucleotides at positions 200, 580, and 1,150 (28). Since two of these sites (those at positions 580 and 1,150) are probably used in the maturation processes that lead to the 18S and 16S late mRNA's (see below), this leaves only the sequence around position 200 as a splicing point for the generation of the T1 oligoribonucleotide under consideration. An attractive explanation is that this oligonucleotide arises from a sequence generated by a "leader-to-leader" splice, which results in an mRNA with the reiterated sequence. From a consideration of the DNA sequence, such a splice could be postulated to occur between ribonucleotides corresponding to the DNA sequence between positions 272 and 215. The leader-to-leader splice would be as follows: 5′-TCAA/GTAAGT (269-278)....-TTTTCTATTTTAAG/AG-3′ (202-217). For this splice, not only is the 3′ site of the intervening sequence preceded by a long pyrimidine-rich tract (4, 16, 17), but the sequence can be correlated with S1 mapping data. A more recent correlation of accumulated splicing data has led to a consensus sequence of YYNYAG/ for 3′ junctions (41), and the sequence between positions 210 and 215 is in general accord with this. Moreover, the sequences of these regions are found to be remarkably similar to those found in a leader-to-leader splice in the late region of SV40 (16). However, the final proof of splicing junctions in polyoma virus late mRNA's must await sequence studies of the messengers themselves.

In a separate study, Legon (30a) has shown that the four T1 oligonucleotides which define the reiterated leader region are found in association with ribosomal subunits. One possibility that has been advanced is that the reiterated leader sequence constitutes a "landing platform" for ribosomes before the initiation of protein synthesis. However, there is no proof for this, and without further evidence it remains only an attractive hypothesis. The leader putative ribosome binding site contains the sequence 5′-UCUUCUG-3′ at positions 255 through 261. Following the postulates of Shine and Dalgarno (43), this (allowing for GU base pairs) could form a complex between the leader region of the mRNA and the 3′ end of mouse 18S rRNA, whose sequence has been determined to be 5′-AACCUGCGGAAGGAUCAUUG (23). Similar base pairing associations are also possible close to the initiation codons for the polyoma virus early proteins (48), and, as shown in an accompanying paper, they are possible, but not probable, in the case of the major capsid protein, VP1 (47). In the latter case, the leader sequence itself may supply the sequence for base pairing associations.

In their study of the capping of late polyoma viral mRNA's, Flavell et al. (11) found eight cap I and four cap II structures at the 5′ ends of the three late messages. These 12 sequences can be accommodated within the first 105 nucleotides nearest the VP2 initiation codon, the region which also encompasses the leader (Fig. 8). Thus, this stretch of sequence (nucleotides 185 through 292 [Fig. 8]) defines the minimum region of the genome which is required to accommodate all of the structural features found to date within the 5′-terminal regions of the three polyoma late mRNA's.

Subsequently, Flavell et al. (10) identified in late mRNA's large, capped T1 oligonucleotides which probably arise from the DNA sequence immediately preceding position 215. This defines the termini of a staggered set of capped mRNA species which have the 5′ sequences GACAUUUUC, ACAUUUUC, and AUUUUC (Fig. 8).

Single-stranded nuclease (S1) mapping studies have shown that the leader region is spliced onto the bodies of the messages for VP3 (18S) and VP1 (16S) somewhere in the region of positions 580 and 1,150, respectively (28) (i.e., about 55 to 75 nucleotides before the initiation codons [Fig. 3]). From the DNA sequence, good candidates for splice junctions which can lead to functional mRNA's appear as follows: for 18S, 5′-TCAA/GTAAGT (269-278)....CCCTAG/GA-3′ (579-586) and for 16S, 5′-TCAA/GTAACT (269-278)....TTCTAG/GG-3′ (1164-1171). These sequences are in good agreement with current ideas about sequences which specify splicing in RNAs (see above) and with the S1 mapping data. Moreover, Seif et al. (41) point out that in a large number of cases examined to date, no AG dinucleotide sequence occurs within the 13 nucleotides which precede the AG at the 3′ side of the intervening sequence. In the sequences considered here, this is also the case. If these hypotheses are correct, then in the mRNA for VP3 (18S) the 3′ end of the splice occurs 50 nucleotides before the initiation codon, and in the mRNA for VP1 (16S) it occurs 48 nucleotides before the initiation codon. It has been observed (D. Hogness, unpublished data) that the canonical sequence TATAAATA occurs about 23 nucleotides before the base which specifies the capped 5′ end in many eucaryotic mRNA's. Moreover, it has been suggested that this sequence could be the eucaryotic equivalent of the procaryotic "Pribnow box" (36). Examination of the appropriate region of polyoma virus DNA reveals only one sequence which resembles the "Hogness box," that is, TAATTAAAAG at positions 134 to 143 (Fig. 8). This is 53 to 56 nucleotides before the three capped 5′ ends previously identified (10). This sequence is very similar to the TATAAAAG sequence found in adenovirus 2 DNA (51). In the noncoding sequence of the polyoma virus early region (48), the sequence TATAATTA has been found to occur 45 nucleotides away from the initiation codon of early mRNA's (at position 173). In confirmation of a possible significance for this sequence (R. Kamen, personal communication), the 5′ end of an early mRNA has been mapped at about position 150. However, it is too soon to say whether these sequences play any role in transcription.

In their studies on the sequence of BK virus DNA, Yang and Wu (50) noted an 80% homology between BK virus and SV40 DNAs in the region which specifies the late leader sequences. A comparison of the polyoma virus late leader region with the corresponding sequences in SV40 and BK virus reveals little homology. This is somewhat surprising since the leader region is sandwiched between the viral origin of DNA replication, for which sequencing reveals a high degree of homology among the three viruses (48), and the coding sequence for VP2/VP3, which (see above) also shows significant homology. A computation of the rates of evolution of polyoma virus, SV40, and BK virus based on a comparison of their (relatively homologous) amino acid and nucleotide coding region sequences gives rise to a phylogenetic tree which corresponds to that derived for their host organisms on the basis of the fossil record (E. Soeda, T. Maruyama, J. R. Arrand, and B. E. Griffin, submitted for publication). Thus, the observation of little homology in the leader regions of polyoma virus versus SV40 and BK virus suggests that these mRNA controlling regions are evolving more rapidly than the coding sequences and replication origins or, alternatively, that on an evolutionary time scale they are a recent development.

ACKNOWLEDGMENTS

We thank D. Hogness for permission to cite his unpublished data. We are also very grateful to R. Kamen for access to his data before publication and for many helpful discussions. We thank C. Maddock for much of the work done in preparing polyoma virus DNA and S. Barrett for restriction enzymes.

LITERATURE CITED

1. **Arrand, J. R., W. Keller, and R. J. Roberts.** 1974. Extent of terminal repetition of adenovirus 2 DNA. Cold Spring Harbor Symp. Quant. Biol. **39**:401-407.
2. **Arrand, J. R., and R. J. Roberts.** 1979. The nucleotide sequences at the termini of adenovirus 2 DNA. J. Mol. Biol. **128**:577-594.
3. **Berkner, K. L., and W. R. Folk.** 1979. A map of the sites in the polyoma genome cleaved by endonuclease *Alu*I. Virology **92**:482-494.
4. **Breathnach, R., C. Benoist, K. O'Hare, F. Gannon, and P. Chambon.** 1978. Ovalbumin gene: evidence for a leader sequence in mRNA and DNA sequences at the exon-intron boundaries. Proc. Natl. Acad. Sci. U.S.A. **75**:4853-4857.
5. **Contreras, R., R. Rogiers, A. Van de Voorde, and W. Fiers.** 1977. Overlapping of the VP2-VP3 gene and the VP1 gene in the SV40 genome. Cell **12**:529-538.
6. **Crawford, L. V., and A. K. Robbins.** 1976. The cleavage of polyoma virus DNA by restriction enzymes *Kpn*I and *Pst*I. J. Gen. Virol. **31**:315-321.
7. **De Wachter, R., and W. Fiers.** 1971. Fractionation of RNA by electrophoresis on polyacrylamide slabs. Methods Enzymol. **21**:167-178.
8. **Ferguson, J., and R. W. Davis.** 1975. An electron microscopic method for studying and mapping the region of weak sequence homology between simian virus 40 and polyoma DNAs. J. Mol. Biol. **94**:135-149.
9. **Fiers, W., R. Contreras, G. Haegeman, R. Rogiers, A. Van de Voorde, H. Van Heuverswyn, J. Van Herreweghe, G. Volckaert, and M. Ysebaert.** 1978. Complete nucleotide sequence of SV40 DNA. Nature (London) **273**:113-120.
10. **Flavell, A. J., A. Cowie, J. R. Arrand, and R. Kamen.** 1980. Localization of three major capped 5' ends of polyoma virus late mRNA's within a single tetranucleotide sequence in the viral genome. J. Virol. **33**:902-908.
11. **Flavell, A. J., S. Legon, A. Cowie, and R. Kamen.** 1979. Multiple 5'-terminal cap structures in late polyoma virus RNA. Cell **16**:357-371.
12. **Fried, M.** 1974. Isolation and partial characterization of different defective DNA molecules derived from polyoma virus. J. Virol. **13**:939-949.
13. **Fried, M., and B. E. Griffin.** 1977. Organization of the genomes of polyoma virus and SV40. Adv. Cancer Res. **24**:67-113.
14. **Fried, M., B. E. Griffin, E. Lund, and D. L. Robberson.** 1974. Polyoma virus—a study of wild type, mutant and defective DNAs. Cold Spring Harbor Symp. Quant. Biol. **39**:45-52.
15. **Gannon, F., K. O'Hare, F. Perrin, J. P. LePennec, C. Benoist, M. Cochet, R. Breathnach, A. Royal, A. Garapin, B. Cami, and P. Chambon.** 1979. Organisation and sequences at the 5' end of a cloned complete ovalbumin gene. Nature (London) **278**:428-434.
16. **Ghosh, R. K., V. B. Reddy, J. Swinscoe, P. Lebowitz, and S. M. Weissman.** 1978. Heterogeneity and 5'-terminal structures of the late RNAs of simian virus 40. J. Mol. Biol. **126**:813-846.
17. **Ghosh, R. K., V. B. Reddy, J. Swinscoe, P. Lebowitz, and S. M. Weissman.** 1978. The 5'-terminal leader sequence of late mRNA from cells infected with simian virus 40. J. Biol. Chem. **253**:3643-3647.
18. **Godson, G. N., B. G. Barrell, R. Staden, and J. C. Fiddes.** 1978. Nucleotide sequence of bacteriophage G4 DNA. Nature (London) **276**:236-247.
19. **Goulian, M., Z. J. Lucas, and A. Kornberg.** 1968. Enzymatic synthesis of deoxyribonucleic acid. XXV. Purification and properties of deoxyribonucleic acid polymerase induced by infection with phage T4. J. Biol. Chem. **243**:627-638.
20. **Griffin, B. E.** 1978. Fine structure of polyoma virus DNA. J. Mol. Biol. **117**:447-471.
21. **Griffin, B. E., and M. Fried.** 1976. Structural mapping of the DNA of an oncogenic virus (polyoma viral DNA). Methods Cancer Res. **12**:49-86.
22. **Griffin, B. E., M. Fried, and A. Cowie.** 1974. Polyoma DNA: a physical map. Proc. Natl. Acad. Sci. U.S.A. **71**: 2077-2081.
23. **Hagenbuchle, O., M. Santer, J. A. Steitz, and R. J. Mans.** 1978. Conservation of the primary structure at the 3' end of 18S rRNA from eukaryotic cells. Cell **13**: 551-563.
24. **Hewick, R. M., M. Fried, and M. D. Waterfield.** 1975. Nonhistone virion proteins of polyoma: characterisation of the particle proteins by tryptic analysis by use of ion-exchange columns. Virology **66**:408-419.
25. **Hewick, R. M., A. Mellor, A. E. Smith, and M. D. Waterfield.** 1980. Partial amino-terminal sequences of the polyoma nonhistone proteins VP1, VP2, and VP3 synthesized in vitro. J. Virol. **33**:631-636.
26. **Hewick, R. M., M. D. Waterfield, L. K. Miller, and M. Fried.** 1977. Correlation between genetic loci and structural differences in the capsid proteins of polyoma virus plaque morphology mutants. Cell **11**:331-338.
27. **Hirt, B.** 1967. Selective extraction of polyoma DNA from infected mouse cell cultures. J. Mol. Biol. **26**:365-369.
28. **Kamen, R., J. Favaloro, and J. Parker.** 1980. Topography of the three late mRNA's of polyoma virus which encode the virion proteins. J. Virol. **33**:637-651.
29. **Kamen, R., J. Sedat, and E. Ziff.** 1976. Orientation of the complementary strands of polyoma virus DNA with respect to the DNA physical map. J. Virol. **17**:212-218.
30. **Kamen, R., and H. Shure.** 1976. Topology of polyoma virus messenger RNA. Cell **7**:361-371.

30a. **Legon, S.** 1979. The binding of ribosomes to polyoma virus RNA. Possible role of the leader region in initiation site recognition. J. Mol. Biol. **134**:219-240.

31. **Legon, S., A. J. Flavell, A. Cowie, and R. Kamen.** 1979. Amplification in the leader sequence of 'late' polyoma virus mRNAs. Cell **16**:373-388.
32. **Maxam, A. M., and W. Gilbert.** 1977. A new method for sequencing DNA. Proc. Natl. Acad. Sci. U.S.A. **74**:560-564.
33. **Miller, L. K., and M. Fried.** 1976. Construction of the genetic map of the polyoma genome. J. Virol. **18**:824-832.
34. **Min Jou, W., G. Haegeman, M. Ysebaert, and F. Fiers.** 1972. Nucleotide sequence of the gene coding for the bacteriophage MS2 coat protein. Nature (London) **237**:82-88.
35. **Miyata, T., and T. Yasunaga.** 1978. Evolution of overlapping genes. Nature (London) **272**:532-535.
36. **Pribnow, D.** 1975. Nucleotide sequence of an RNA polymerase binding site at an early T7 promoter. Proc. Natl. Acad. Sci. U.S.A. **72**:784-788.
37. **Reddy, V. B., B. Thimmappaya, R. Dhar, K. N. Subramanian, B. S. Zain, J. Pan, P. K. Ghosh, M. L. Celma, and S. M. Weissman.** 1978. The genome of simian virus 40. Science **200**:494-502.
38. **Roberts, R. J.** 1976. Restriction endonucleases. Crit. Rev. Biochem. **4**:123-164.
39. **Sanger, F., G. M. Air, B. G. Barrell, N. L. Brown, A. R. Coulson, J. C. Fiddes, C. A. Hutchison III, P. M. Slocombe, and M. Smith.** 1977. Nucleotide sequence of ϕ X 174 DNA. Nature (London) **265**:687-695.
40. **Sanger, F., S. Nicklen, and A. R. Coulson.** 1977. DNA sequencing with chain terminating inhibitors. Proc. Natl. Acad. Sci. U.S.A. **74**:5463-5467.
41. **Seif, I., G. Khoury, and R. Dhar.** 1979. BKV splice sequences based on analysis of preferred donor and acceptor sites. Nucleic Acids Res. **6**:3387-3398.
42. **Sharp, P. A., B. Sugden, and J. Sambrook.** 1973. Detection of two restriction endonuclease activities in *Haemophilus parainfluenzae* using analytical agarose-

ethidium bromide electrophoresis. Biochemistry **12**: 3055-3063.

43. **Shine, J., and L. Dalgarno.** 1974. The 3′ terminal sequence of *Escherichia coli* 16S ribosomal RNA: complementarity to nonsense triplets and ribosome binding sites. Proc. Natl. Acad. Sci. U.S.A. **71**:1342-1346.
44. **Siddell, S. G., and A. E. Smith.** 1978. Polyoma virus has three late mRNA's: one for each virion protein. J. Virol. **27**:427-431.
45. **Smith, A. E., R. Kamen, W. F. Mangel, H. Shure, and T. Wheeler.** 1976. Location of the sequences coding for capsid proteins VP1 and VP2 on polyoma virus DNA. Cell **9**:481-487.
46. **Smith, A. J. H.** 1979. The use of exonuclease III for preparing single stranded DNA for use as a template in the chain terminator sequencing method. Nucleic Acids Res. **6**:831-848.
47. **Soeda, E., J. R. Arrand, and B. E. Griffin.** 1980. Polyoma virus DNA: complete nucleotide sequence of the gene which codes for polyoma virus capsid protein VP1 and overlaps the VP2/VP3 genes. J. Virol. **33**:619-630.
48. **Soeda, E., J. R. Arrand, N. Smolar, and B. E. Griffin.** 1979. Sequence from early region of polyoma virus DNA containing viral replication origin and encoding small, middle and (part of) large T antigens. Cell **17**:357-370.
49. **Subak-Sharpe, J. H.** 1967. Base doublet frequency patterns in the nucleic acid and evolution of viruses. Br. Med. Bull. **23**:161-168.
50. **Yang, R. C. A., and R. Wu.** 1978. BK virus DNA: cleavage map and sequence analysis. Proc. Natl. Acad. Sci. U.S.A. **75**:2150-2154.
51. **Ziff, E. B., and R. M. Evans.** 1978. Coincidence of the promoter and capped 5′-terminus of RNA from the adenovirus 2 major late transcription unit. Cell **15**: 1463-1475.