References and Notes

1. I. Newton, *Opticks*, Book I, Part I, Proposition VIII, Problem II (1730) (Dover, New York, 1952).
2. A. A. Michelson and F. G. Pease, *Astrophys. J.* **53**, 249 (1921).
3. R. Hanbury Brown, *The Intensity Interferometer* (Halsted, London, 1974).
4. A. Labeyrie, *Astron. Astrophys.* **6**, 85 (1970); D. V. Gezari, A. Labeyrie, R. V. Stachnik, *Astrophys. J.* **173**, L1 (1972).
5. H. A. McAlister, *Astrophys. J.* **215**, 159 (1977).
6. K. T. Knox and B. J. Thompson, *ibid.* **193**, L45 (1974).
7. P. Nisenson, D. C. Ehn, R. V. Stachnik, *Proc. Soc. Photo-Opt. Instrum. Eng.* **75**, 83 (1976).
8. R. V. Stachnik, P. Nisenson, D. C. Ehn, R. H. Hudgin, V. E. Schirf, *Nature (London)* **266**, 149 (1977).
9. T. M. Brown, *J. Opt. Soc. Am.*, in press.
10. H. W. Babcock, *Publ. Astron. Soc. Pac.* **65**, 229 (1953); *J. Opt. Soc. Am.* **48**, 500 (1958).
11. R. H. Dicke, *Astrophys. J.* **198**, 605 (1975).
12. J. W. Hardy, J. Feinlieb, J. C. Wyant, in "Digests of Technical Papers for the Topical Meeting on Optical Propagation through Turbulence," 9 to 11 July 1974, Boulder, Colo., paper TH-B1; J. W. Hardy, J. E. Lefebvre, C. L. Koliopoulos, *J. Opt. Soc. Am.* **67**, 360 (1977).
13. T. R. O'Meara, U.S. Patent No. 3,975,629 (1976).
14. R. A. Muller and A. Buffington, *J. Opt. Soc. Am.* **64**, 1200 (1974).
15. For partial descriptions of these systems see "Digests of Technical Papers for the Topical Meeting on Optical Propagation through Turbulence," 9 to 11 July 1974, in Boulder, Colo.; "Digests of Technical Papers for the Topical Meeting on Imaging in Astronomy," 18 to 21 June 1975, Cambridge, Mass.; "Imaging through the atmosphere," *Proc. Soc. Photo-Opt. Instrum. Eng.* **75** (1976); *J. Opt. Soc. Am.* **67** (No. 3) (1977).
16. S. L. McCall, T. R. Brown, A. Passner, *Astrophys. J.* **211**, 463 (1977).
17. A. Buffington, F. S. Crawford, R. A. Muller, A. J. Schwemin, R. G. Smits, *J. Opt. Soc. Am.* **67**, 298 (1977).
18. A. Buffington, F. S. Crawford, R. A. Muller, C. D. Orth, *ibid.*, p. 304.
19. For example, see J. Faulkner, *Phys. Rev. Lett.* **27**, 206 (1971).

20. A. J. Schwemin and R. G. Smits contributed much to the design and construction of the apparatus. The observatory measurements would have been impossible without the enthusiastic support of the staff of the Astronomy Department at Berkeley, the Lick Observatory staff, and the Mount Wilson Observatory staff. We are pleased with the continuing interest and support of L. W. Alvarez, H. W. Babcock, R. W. Birge, P. B. Boyce, and D. E. Osterbrock. D. D. Cudaback made several important contributions. We have enjoyed interesting conversations with L. V. Kuhi, J. E. Nelson, J. A. Tyson, G. Wallerstein, and E. J. Wampler. This work was supported by the Department of Energy and by grants from the National Science Foundation and the National Aeronautics and Space Administration.

The Genome of Simian Virus 40

V. B. Reddy, B. Thimmappaya, R. Dhar, K. N. Subramanian, B. S. Zain, J. Pan, P. K. Ghosh, M. L. Celma, S. M. Weissman

Summary. The nucleotide sequence of SV40 DNA was determined, and the sequence was correlated with known genes of the virus and with the structure of viral messenger RNA's. There is a limited overlap of the coding regions for structural proteins and a complex pattern of leader sequences at the 5′ end of late messenger RNA. The sequence of the early region is consistent with recent proposals that the large early polypeptide of SV40 is encoded in noncontiguous segments of DNA.

Simian virus 40 (SV40) is a small virus that replicates in the nucleus of host cells and may also transform cells from a variety of species (*1*). During the lytic cycle, gene expression is temporally regulated. Prior to viral DNA replication, one-half of one strand of the genome (the early region) is transcribed, forming cytoplasmic polyadenylated messenger RNA (mRNA). This mRNA, which is termed early RNA, is also present in most cells transformed by SV40 and directs the synthesis of this early mRNA. After the initiation of viral replication, the other strand of the other half of the DNA (the late region) is transcribed into mRNA (referred to here as late mRNA) that directs the synthesis of viral structural proteins. This virus has been the subject of intensive investigation as a model system for genes functioning in the nuclei of animal cells and for viral transformation of cells in culture. The genome of the virus consists of a circular DNA molecule containing more than 5200 base pairs (bp). In spite of the limited information in this DNA, the virus has a complex genetic structure and exhibits features of gene organization and expression different from those so far detected in prokaryotes.

In one approach to the detailed understanding of virus function, Fiers and his colleagues and our laboratory have separately determined the nucleotide sequence of the viral DNA. In this article, we present some of the principal features of this sequence (Fig. 1). More detailed reports of portions of the sequence have been published or are still in preparation (*2–14*).

This article represents work done by the authors while members of the Department of Human Genetics or the Department of Medicine (or both), at the Yale University School of Medicine, New Haven, Connecticut 06510. The present address of B. Thimmappaya is University of Connecticut Medical School, Farmington 02100; that of R. Dhar is National Institutes of Health, Bethesda, Maryland 20014; that of R. N. Subramanian is the Microbiology Department, University of Illinois Medical School, Chicago 60680; that of B. S. Zain is Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724; that of M. L. Celma is Departmento de Microbiologia, Centro Ramon y Cajal, Carretera de Colmenar, Room 9.1, Madrid 34, Spain.

Regions Know to Code for Identified Viral Proteins

SV40 virus codes for at least four proteins. The major viral structural protein, VP1, has a molecular weight, estimated by sodium dodecyl sulfate gel electrophoresis, of 43,500 to 48,000 (*15*). The virus also codes for two minor structural proteins, VP2 and VP3, whose estimated molecular weights are 39,000 and 27,000 (*15*), respectively, and one large "early" protein (the A protein or T antigen) (*16–18*), which is necessary for initiation of rounds of DNA replication and for cell transformation. The molecular weight of the A protein has been estimated by sodium dodecyl sulfate gel electrophoresis as approximately 94,000.

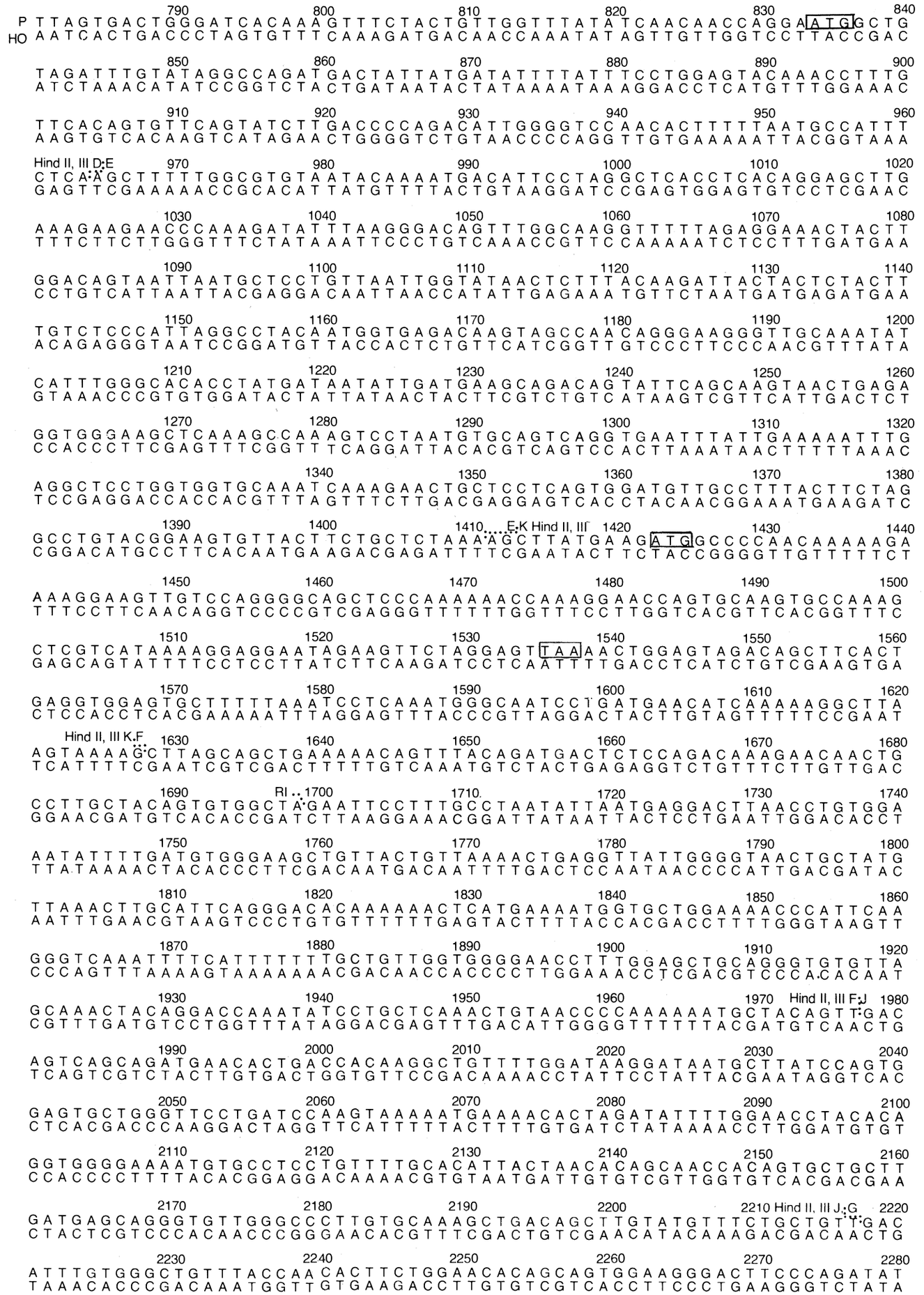
There are three stretches of DNA in the SV40 genome whose RNA transcripts contain an AUG base triplet (A, adenylic acid; U, uridylic acid; G, guanylic acid) followed by a long run of sense codons (Fig. 2). The first of these begins at residue 1423 of Fig. 1. The initial sequences of this region were determined by Fiers *et al.* (*12*) and found to agree fully with the amino acid sequence of the amino terminus of VP1. Proceeding from the AUG at position 1423 (Fig. 1), there are 361 subsequent sense codons in phase, followed by a single UGA termination codon. This sequence predicts that the carboxyl-terminal amino acid of VP1 is glutamine; and this prediction has been confirmed experimentally (*19*). The predicted amino acid composition for VP1 is in good agreement with the experimental values (*20*). Some tryptic peptides of VP1 have been

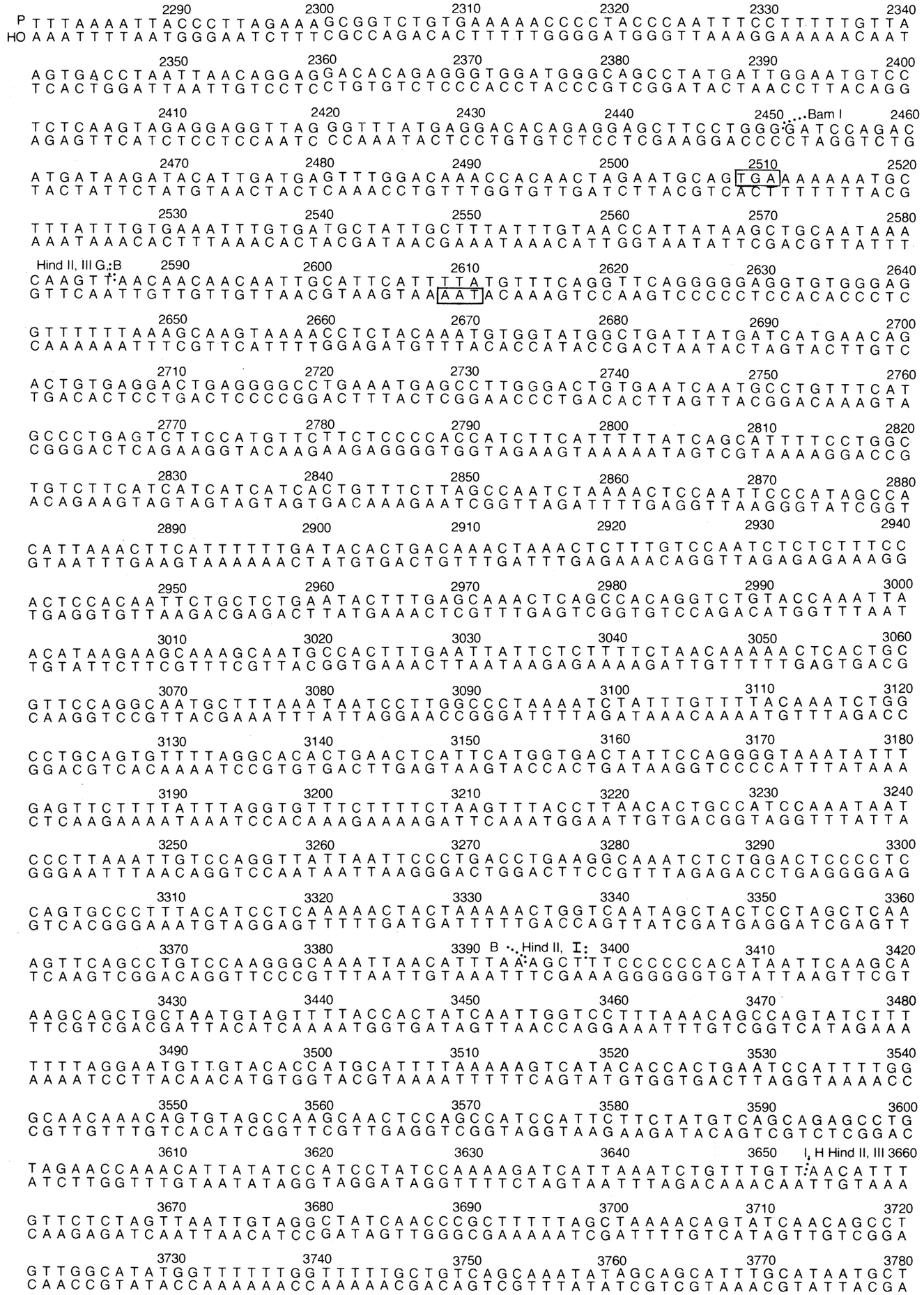
0036-8075/78/0505-0494$02.00/0

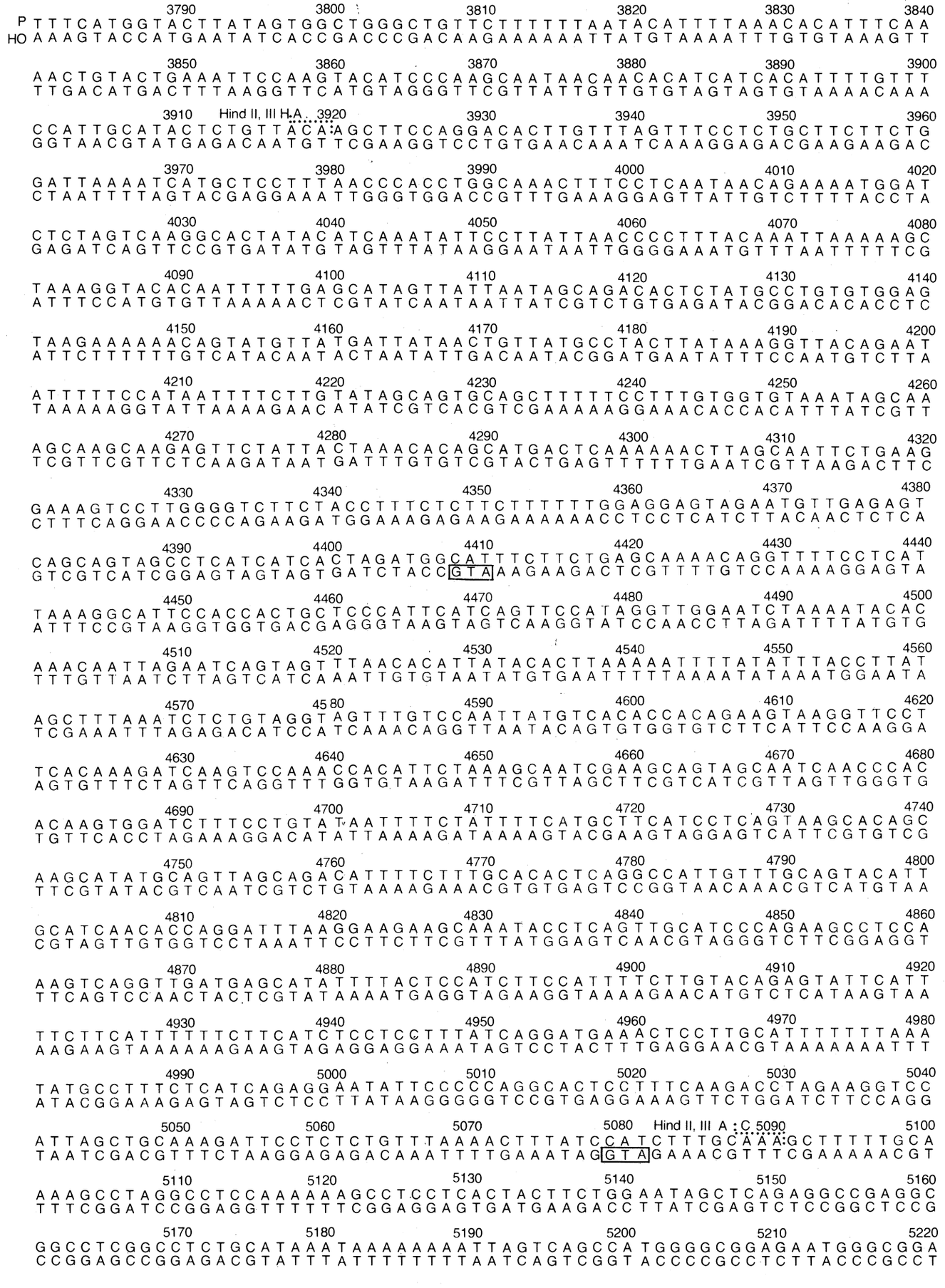
found to have sequences that agree with the amino acid sequences predicted by the nucleotide sequence of this region of the DNA. The 3′ terminus of SV40 late mRNA lies between residues 2585 and 2599, about 80 bases beyond the UGA referred to above (*2, 4*). Cells infected with mutant viruses with deletions of part of the DNA in this region synthesize a shorter protein that has some of the tryptic peptides of VP1 (*21*). Therefore, it seems quite certain that this region of DNA codes for VP1. The restriction endonuclease fragment Hind II/III-F, derived from this region, has an anomalously low electrophoretic mobility in acrylamide gels, relative to its molecular weight, as is predicted from the nucleotide sequence. After alkaline denaturation, the separated strands of this fragment move close to their predicted position in gels containing 8*M* urea.

There is a second, long, continuous stretch of coding triplets in the late region of SV40. The triplet AUG at positions 480 to 482 in the late strand transcript is followed by 351 sense triplets, followed by a termination codon UAA at positions 1536 to 1538. This set of triplets includes almost the entirety of fragment Hind II/III-D and the entirety of Hind II/III-E, and overlaps the first 38 coding triplets of VP1. The analysis of proteins made in cells infected with deletion mutants of SV40 has provided evidence that VP3 and VP2 are coded for in part by the fragment Hind II/III-E, and that VP2 but not VP3 is coded for by parts of Hind II/III-D (*22*). The molecular weight of VP2 agrees fairly well with a coding sequence of 351 triplets. There is a single AUG at positions 834 to 836 in this sequence, which would be translated as an internal methionine within the amino acid sequence of VP2. This triplet is followed

Fig. 1. Nucleotide sequence of SV40 DNA. The upper line shows the sequence proceeding in the 5′ to 3′ direction from sequences near the origin of DNA replication and represents the strand of DNA complementary to early mRNA. The clevage sites for the *Haemophilus influenzae* strain d restriction endonucleases II and III (Hind II/III) (*66*) are indicated above the DNA sequence. Base triplets corresponding to initiation and termination codons referred to in the text are boxed in. The DNA sequence was analyzed by the method of Maxam and Gilbert (*67*) and Hind II/III fragments G and large parts of A, B, C, and D, analyzed by transcription of viral DNA fragments (*68*) and by limited digestion with snake venom diesterase as described by Maniatis *et al.* (*69*). Hind II/III-H sequences are based in large part on the results of Fiers and co-workers (*14*). An earlier report (*4*) had placed a CTG between residues 4877 and 4888 instead of a T·G and T at separate positions. We cannot reproduce the results on which this had been based, and repeated analyses indicate the present sequence to be correct for this region. Residue 1814 is a C rather than T as previously reported (*9*). We have confirmed this, and also noted a small Alu I fragment beginning at residue 530 that we had missed in our initial analyses, but that had been shown in the Alu I restriction endonuclease cleavage maps published by Fiers and his colleagues (*70*) and by Jay and Wu (*71*). We are grateful to Dr. J. Lautenberger for alerting us to errors in our initial version of the sequence about residues 1180 to 1280. The portions of the sequence not completely confirmed by RNA analyses are probably more than 99 percent accurate, but occasionally a base was misidentified by the chemical degradation method. Both residues 4758 and 4768 were identified as deoxycytidylic acid by both repeated runs of the Maxam-Gilbert procedure and by analysis of radioactively labeled RNA transcripts. However, analysis of the DNA of our present stock shows by the Maxam-Gilbert procedure that the residue is T and the base in the complementary strand is A. Residue 5194 is a C in the original virus stock but an A in the present stock, and residue 1674 is an A in the original strain and a C in the new strain. The sequence from 4560 to 4641 is from the new virus, and appears to have some base changes from the original strain such as a C instead of an A at 4571.


790 800 810 820 830 840
P TTAGTGACTGGGATCACAAAGTTTCTACTGTTGGTTTATATCAACAACCAGGAATGGCTG
HO AATCACTGACCCTAGTGTTTCAAAGATGACAACCAAATATAGTTGTTGGTCCTTACCGAC
850 860 870 880 890 900
TAGATTTGTATAGGCCAGATGACTATTATGATATTTTATTTCCTGGAGTACAAACCTTTG
ATCTAAACATATCCGGTCTACTGATAATACTATAAAATAAAGGACCTCATGTTTGGAAAC
910 920 930 940 950 960
TTCACAGTGTTCAGTATCTTGACCCCAGACATTGGGGTCCAACACTTTTTAATGCCATTT
AAGTGTCACAAGTCATAGAACTGGGGTCTGTAACCCCAGGTTGTGAAAAATTACGGTAAA
Hind II, III D;E 970 980 990 1000 1010 1020
CTCAAGCTTTTTGGCGTGTAATACAAAATGACATTCCTAGGCTCACCTCACAGGAGCTTG
GAGTTCGAAAAACCGCACATTATGTTTTACTGTAAGGATCCGAGTGGAGTGTCCTCGAAC
1030 1040 1050 1060 1070 1080
AAAGAAGAACCCAAAGATATTTAAGGGACAGTTTGGCAAGGTTTTTAGAGGAAACTACTT
TTTCTTCTTGGGTTTCTATAAATTCCCTGTCAAACCGTTCCAAAAATCTCCTTTGATGAA
1090 1100 1110 1120 1130 1140
GGACAGTAATTAATGCTCCTGTTAATTGGTATAACTCTTTACAAGATTACTACTCTACTT
CCTGTCATTAATTACGAGGACAATTAACCATATTGAGAAATGTTCTAATGATGAGATGAA
1150 1160 1170 1180 1190 1200
TGTCTCCCATTAGGCCTACAATGGTGAGACAAGTAGCCAACAGGGAAGGGTTGCAAATAT
ACAGAGGGTAATCCGGATGTTACCACTCTGTTCATCGGTTGTCCCTTCCCAACGTTTATA
1210 1220 1230 1240 1250 1260
CATTTGGGCACACCTATGATAATATTGATGAAGCAGACAGTATTCAGCAAGTAACTGAGA
GTAAACCCGTGTGGATACTATTATAACTACTTCGTCTGTCATAAGTCGTTCATTGACTCT
1270 1280 1290 1300 1310 1320
GGTGGGAAGCTCAAAGCCAAAAGTCCTAATGTGCAGTCAGGTGAATTTATTGAAAAATTTG
CCACCCTTCGAGTTTCGGTTTTCAGGATTACACGTCAGTCCACTTAAATAACTTTTTAAAC
1340 1350 1360 1370 1380
AGGCTCCTGGTGGTGCAAATCAAAGAACTGCTCCTCAGTGGATGTTGCCTTTACTTCTAG
TCCGAGGACCACCACGTTTAGTTTCTTGACGAGGAGTCACCTACAACGGAAATGAAGATC
1390 1400 1410 E;K Hind II, III 1420 1430 1440
GCCTGTACGGAAGTGTTACTTCTGCTCTAAAAGCTTATGAAGATGGCCCCAACAAAAAGA
CGGACATGCCTTCACAATGAAGACGAGATTTTCGAATACTTCTACCGGGGTTGTTTTTCT
1450 1460 1470 1480 1490 1500
AAAGGAAGTTGTCCAGGGGCAGCTCCCAAAAAACCAAAGGAACCAGTGCAAGTGCCAAAG
TTTCCTTCAACAGGTCCCCGTCGAGGGTTTTTTGGTTTCCTTGGTCACGTTCACGGTTTC
1510 1520 1530 1540 1550 1560
CTCGTCATAAAAGGAGGAATAGAAGTTCTAGGAGTTAAAACTGGAGTAGACAGCTTCACT
GAGCAGTATTTTCCTCCTTATCTTCAAGATCCTCAATTTTGACCTCATCTGTCGAAGTGA
1570 1580 1590 1600 1610 1620
GAGGTGGAGTGCTTTTTAAATCCTCAAATGGGCAATCCTGATGAACATCAAAAAGGCTTA
CTCCACCTCACGAAAAATTTAGGAGTTTACCCGTTAGGACTACTTGTAGTTTTTCCGAAT
Hind II, III K;F 1630 1640 1650 1660 1670 1680
AGTAAAAGCTTAGCAGCTGAAAAACAGTTTACAGATGACTCTCCAGACAAAGAACAACTG
TCATTTTCGAATCGTCGACTTTTTGTCAAATGTCTACTGAGAGGTCTGTTTCTTGTTGAC
1690 RI 1700 1710 1720 1730 1740
CCTTGCTACAGTGTGGCTAGAATTCCTTTGCCTAATATTAATGAGGACTTAACCTGTGGA
GGAACGATGTCACACCGATCTTAAGGAAACGGATTATAATTACTCCTGAATTGGACACCT
1750 1760 1770 1780 1790 1800
AATATTTTGATGTGGGAAGCTGTTACTGTTAAAACTGAGGTTATTGGGGTAACTGCTATG
TTATAAAACTACACCCTTCGACAATGACAATTTTGACTCCAATAACCCCATTGACGATAC
1810 1820 1830 1840 1850 1860
TTAAACTTGCATTCAGGGACACAAAAAACTCATGAAAATGGTGCTGGAAAACCCATTCAA
AATTTGAACGTAAGTCCCTGTGTTTTTTGAGTACTTTTACCACGACCTTTTGGGTAAGTT
1870 1880 1890 1900 1910 1920
GGGTCAAATTTTCATTTTTTTTGCTGTTGGTGGGGAACCTTTGGAGCTGCAGGGTGTGTTA
CCCAGTTTAAAAGTAAAAAAACGACAACCACCCCTTGGAAACCTCGACGTCCCACACAAT
1930 1940 1950 1960 1970 Hind II, III F;J 1980
GCAAACTACAGGACCAAATATCCTGCTCAAACTGTAACCCCAAAAAATGCTACAGTTGAC
CGTTTGATGTCCTGGTTTATAGGACGAGTTTGACATTGGGGTTTTTTACGATGTCAACTG
1990 2000 2010 2020 2030 2040
AGTCAGCAGATGAACACTGACCACAAGGCTGTTTTGGATAAGGATAATGCTTATCCAGTG
TCAGTCGTCTACTTGTGACTGGTGTTCCGACAAAACCTATTCCTATTACGAATAGGTCAC
2050 2060 2070 2080 2090 2100
GAGTGCTGGGTTCCTGATCCAAGTAAAAATGAAAACACTAGATATTTTGGAACCTACACA
CTCACGACCCAAGGACTAGGTTCATTTTTACTTTTGTGATCTATAAAACCTTGGATGTGT
2110 2120 2130 2140 2150 2160
GGTGGGGAAAATGTGCCTCCTGTTTTGCACATTACTAACACAGCAACCACAGTGCTGCTT
CCACCCCTTTTACACGGAGGACAAAACGTGTAATGATTGTGTCGTTGGTGTCACGACGAA
2170 2180 2190 2200 2210 Hind II, III J;G 2220
GATGAGCAGGGTGTTGGGCCCTTGTGCAAAGCTGACAGCTTGTATGTTTCTGCTGTTGAC
CTACTCGTCCCACAACCCGGGAACACGTTTCGACTGTCGAACATACAAAGACGACAACTG
2230 2240 2250 2260 2270 2280
ATTTGTGGGCTGTTTACCAACACTTCTGGAACACAGCAGTGGAAGGGACTTCCCAGATAT
TAAACACCCGACAAATGGTTGTGAAGACCTTGTGTCGTCACCTTCCCTGAAGGGTCTATA


2290 2300 2310 2320 2330 2340
P TTTAAAATTACCCTTAGAAAGCGGTCTGTGAAAAACCCCTACCCAATTTCCTTTTTGTTA
HO AAATTTTAATGGGAATCTTTCGCCAGACACTTTTTGGGGATGGGTTAAAGGAAAAACAAT
2350 2360 2370 2380 2390 2400
AGTGACCTAATTAACAGGAGGACACAGAGGGTGGATGGGCAGCCTATGATTGGAATGTCC
TCACTGGATTAATTGTCCTCCTGTGTCTCCCACCTACCCGTCGGATACTAACCTTACAGG
2410 2420 2430 2440 2450 Bam I 2460
TCTCAAGTAGAGGAGGTTAGGGTTTATGAGGACACAGAGGAGCTTCCTGGGGATCCAGAC
AGAGTTCATCTCCTCCAATCCCAAATACTCCTGTGTCTCCTCGAAGGACCCCTAGGTCTG
2470 2480 2490 2500 2510 2520
ATGATAAGATACATTGATGAGTTTGGACAAACCACAACTAGAATGCAGTGAAAAAAATGC
TACTATTCTATGTAACTACTCAAACCTGTTTGGTGTTGATCTTACGTCACTTTTTTTACG
2530 2540 2550 2560 2570 2580
TTTATTTGTGAAATTTGTGATGCTATTGCTTTATTTGTAACCATTATAAGCTGCAATAAA
AAATAAACACTTTAAACACTACGATAACGAAATAAACATTGGTAATATTCGACGTTATTT
Hind II, III 2590 2600 2610 2620 2630 2640
CAAGTTAACAACAACAATTGCATTCATTTTATGTTTCAGGTTCAGGGGGAGGTGTGGGAG
GTTCAATTGTTGTTGTTAACGTAAGTAAAATACAAAGTCCAAGTCCCCCTCCACACCCTC
2650 2660 2670 2680 2690 2700
GTTTTTTAAAGCAAGTAAAACCTCTACAAATGTGGTATGGCTGATTATGATCATGAACAG
CAAAAAATTTCGTTCATTTTGGAGATGTTTACACCATACCGACTAATACTAGTACTTGTC
2710 2720 2730 2740 2750 2760
ACTGTGAGGACTGAGGGGCCTGAAATGAGCCTTGGGACTGTGAATCAATGCCTGTTTCAT
TGACACTCCTGACTCCCCGGACTTTACTCGGAACCCTGACACTTAGTTACGGACAAAGTA
2770 2780 2790 2800 2810 2820
GCCCTGAGTCTTCCATGTTCTTCTCCCCACCATCTTCATTTTTATCAGCATTTTCCTGGC
CGGGACTCAGAAGGTACAAGAAGAGGGGTGGTAGAAGTAAAAATAGTCGTAAAAGGACCG
2830 2840 2850 2860 2870 2880
TGTCTTCATCATCATCATCACTGTTTCTTAGCCAATCTAAAACTCCAATTCCCATAGCCA
ACAGAAGTAGTAGTAGTAGTGACAAAGAATCGGTTAGATTTTGAGGTTAAGGGTATCGGT
2890 2900 2910 2920 2930 2940
CATTAAACTTCATTTTTTGATACACTGACAAACTAAACTCTTTGTCCAATCTCTCTTTCC
GTAATTTGAAGTAAAAAACTATGTGACTGTTTGATTTGAGAAACAGGTTAGAGAGAAAGG
2950 2960 2970 2980 2990 3000
ACTCCACAATTCTGCTCTGAATACTTTGAGCAAACTCAGCCACAGGTCTGTACCAAATTA
TGAGGTGTTAAGACGAGACTTATGAAACTCGTTTGAGTCGGTGTCCAGACATGGTTTAAT
3010 3020 3030 3040 3050 3060
ACATAAGAAGCAAAGCAATGCCACTTTGAATTATTCTCTTTTCTAACAAAAACTCACTGC
TGTATTCTTCGTTTCGTTACGGTGAAAACTTAATAAGAGAAAAGATTGTTTTTGAGTGACG
3070 3080 3090 3100 3110 3120
GTTCCAGGCAATGCTTTAAATAATCCTTGGCCCTAAAATCTATTTGTTTTACAAATCTGG
CAAGGTCCGTTACGAAATTTATTAGGAACCGGGATTTTAGATAAACAAAATGTTTAGACC
3130 3140 3150 3160 3170 3180
CCTGCAGTGTTTTAGGCACACTGAACTCATTCATGGTGACTATTCCAGGGGTAAATATTT
GGACGTCACAAAATCCGTGTGACTTGAGTAAGTACCACTGATAAGGTCCCCATTTATAAA
3190 3200 3210 3220 3230 3240
GAGTTCTTTTATTTAGGTGTTTCTTTTCTAAGTTTACCTTAACACTGCCATCCAAATAAT
CTCAAGAAAATAAATCCACAAAGAAAAGATTCAAATGGAATTGTGACGGTAGGTTTATTA
3250 3260 3270 3280 3290 3300
CCCTTAAATTGTCCAGGTTATTAATTCCCTGACCTGAAGGCAAATCTCTGGACTCCCCTC
GGGAATTTAACAGGTCCAATAATTAAGGGACTGGACTTCCGTTTAGAGACCTGAGGGGAG
3310 3320 3330 3340 3350 3360
CAGTGCCCTTTACATCCTCAAAAACTACTAAAAACTGGTCAATAGCTACTCCTAGCTCAA
GTCACGGGAAATGTAGGAGTTTTTGATGATTTTTGACCAGTTATCGATGAGGATCGAGTT
3370 3380 3390 B Hind II, 3400 3410 3420
AGTTCAGCCTGTCCAAGGGCAAATTAACATTTAAAGCTTTCCCCCCACATAATTCAAGCA
TCAAGTCGGACAGGTTCCCGTTTAATTGTAAATTTCGAAAGGGGGGTGTATTAAGTTCGT
3430 3440 3450 3460 3470 3480
AAGCAGCTGCTAATGTAGTTTTACCACTATCAATTGGTCCTTTAAACAGCCAGTATCTTT
TTCGTCGACGATTACATCAAAATGGTGATAGTTAACCAGGAAATTTGTCGGTCATAGAAA
3490 3500 3510 3520 3530 3540
TTTTAGGAATGTTGTACACCATGCATTTTAAAAAGTCATACACCACTGAATCCATTTTGG
AAAATCCTTACAACATGTGGTACGTAAAATTTTTCAGTATGTGGTGACTTAGGTAAAACC
3550 3560 3570 3580 3590 3600
GCAACAAACAGTGTAGCCAAGCAACTCCAGCCATCCATTCTTCTATGTCAGCAGAGCCTG
CGTTGTTTGTCACATCGGTTCGTTGAGGTCGGTAGGTAAGAAGATACAGTCGTCTCGGAC
3610 3620 3630 3640 3650 I, H Hind II, III 3660
TAGAACCAAACATTATATCCATCCTATCCAAAAGATCATTAAATCTGTTTGTTAACATTT
ATCTTGGTTTGTAATATAGGTAGGATAGGTTTTCTAGTAATTTAGACAAACAATTGTAAA
3670 3680 3690 3700 3710 3720
GTTCTCTAGTTAATTGTAGGCTATCAACCCGCTTTTTAGCTAAAACAGTATCAACAGCCT
CAAGAGATCAATTAACATCCGATAGTTGGGCGAAAAATCGATTTTGTCATAGTTGTCGGA
3730 3740 3750 3760 3770 3780
GTTGGCATATGGTTTTTTGGTTTTTGCTGTCAGCAAATATAGCAGCATTTGCATAATGCT
CAACCGTATACCAAAAAACCAAAAACGACAGTCGTTTATATCGTCGTAAACGTATTACGA


3790 3800 3810 3820 3830 3840
P T T T C A T G G T A C T T A T A G T G G C T G G G C T G T T C T T T T T T A A T A C A T T T T A A A C A C A T T T C A A
HO A A A G T A C C A T G A A T A T C A C C G A C C C G A C A A G A A A A A A T T A T G T A A A A T T T G T G T A A A G T T
3850 3860 3870 3880 3890 3900
A A C T G T A C T G A A A T T C C A A G T A C A T C C C A A G C A A T A A C A A C A C A T C A T C A C A T T T T G T T T
T T G A C A T G A C T T T A A G G T T C A T G T A G G G T T C G T T A T T G T T G T G T A G T A G T G T A A A A C A A A
3910 Hind II, III H A 3920 3930 3940 3950 3960
C C A T T G C A T A C T C T G T T A C A A G C T T C C A G G A C A C T T G T T T A G T T T C C T C T G C T T C T T C T G
G G T A A C G T A T G A G A C A A T G T T C G A A G G T C C T G T G A A C A A A T C A A A G G A G A C G A A G A A G A C
3970 3980 3990 4000 4010 4020
G A T T A A A A T C A T G C T C C T T T A A C C C A C C T G G C A A A C T T T C C T C A A T A A C A G A A A A T G G A T
C T A A T T T T A G T A C G A G G A A A T T G G G T G G A C C G T T T G A A A G G A G T T A T T G T C T T T T A C C T A
4030 4040 4050 4060 4070 4080
C T C T A G T C A A G G C A C T A T A C A T C A A A T A T T C C T T A T T A A C C C C T T T A C A A A T T A A A A A G C
G A G A T C A G T T C C G T G A T A T G T A G T T T A T A A G G A A T A A T T G G G G A A A T G T T T A A T T T T T C G
4090 4100 4110 4120 4130 4140
T A A A G G T A C A C A A T T T T T G A G C A T A G T T A T T A A T A G C A G A C A C T C T A T G C C T G T G T G G A G
A T T T C C A T G T G T T A A A A A C T C G T A T C A A T A A T T A T C G T C T G T G A G A T A C G G A C A C A C C T C
4150 4160 4170 4180 4190 4200
T A A G A A A A A A C A G T A T G T T A T G A T T A T A A C T G T T A T G C C T A C T T A T A A A G G T T A C A G A A T
A T T C T T T T T T G T C A T A C A A T A C T A A T A T T G A C A A T A C G G A T G A A T A T T T C C A A T G T C T T A
4210 4220 4230 4240 4250 4260
A T T T T T C C A T A A T T T T C T T G T A T A G C A G T G C A G C T T T T T C C T T T G T G G T G T A A A T A G C A A
T A A A A A G G T A T T A A A A G A A C A T A T C G T C A C G T C G A A A A A G G A A A C A C C A C A T T T A T C G T T
4270 4280 4290 4300 4310 4320
A G C A A G C A A G A G T T C T A T T A C T A A A C A C A G C A T G A C T C A A A A A A C T T A G C A A T T C T G A A G
T C G T T C G T T C T C A A G A T A A T G A T T T G T G T C G T A C T G A G T T T T T T G A A T C G T T A A G A C T T C
4330 4340 4350 4360 4370 4380
G A A A G T C C T T G G G G T C T T C T A C C T T T C T C T T C T T T T T T G G A G G A G T A G A A T G T T G A G A G T
C T T T C A G G A A C C C C A G A A G A T G G A A A G A G A A G A A A A A A A C C T C C T C A T C T T A C A A C T C T C A
4390 4400 4410 4420 4430 4440
C A G C A G T A G C C T C A T C A T C A C T A G A T G G C A T T T C T T C T G A G C A A A A C A G G T T T T C C T C A T
G T C G T C A T C G G A G T A G T A G T G A T C T A C C G T A A G A A G A C T C G T T T T G T C C A A A A G G A G T A
4450 4460 4470 4480 4490 4500
T A A A G G C A T T C C A C C A C T G C T C C C A T T C A T C A G T T C C A T A G G T T G G A A T C T A A A A T A C A C
A T T T C C G T A A G G T G G T G A C G A G G G T A A G T A G T C A A G G T A T C C A A C C T T A G A T T T T A T G T G
4510 4520 4530 4540 4550 4560
A A A C A A T T A G A A T C A G T A G T T T A A C A C A T T A T A C A C T T A A A A A T T T T A T A T T T A C C T T A T
T T T G T T A A T C T T A G T C A T C A A A T T G T G T A A T A T G T G A A T T T T T T A A A A T A T A A A T G G A A T A
4570 45 80 4590 4600 4610 4620
A G C T T T A A A T C T C T G T A G G T A G T T T G T C C A A T T A T G T C A C A C C A C A G A A G T A A G G T T C C T
T C G A A A T T T A G A G A C A T C C A T C A A A C A G G T T A A T A C A G T G T G G T G T C T T C A T T C C A A G G A
4630 4640 4650 4660 4670 4680
T C A C A A A G A T C A A G T C C A A A C C A C A T T C T A A A G C A A T C G A A G C A G T A G C A A T C A A C C C A C
A G T G T T T C T A G T T C A G G T T T G G T G T A A G A T T T C G T T A G C T T C G T C A T C G T T A G T T G G G T G
4690 4700 4710 4720 4730 4740
A C A A G T G G A T C T T T C C T G T A T A A T T T T C T A T T T T C A T G C T T C A T C C T C A G T A A G C A C A G C
T G T T C A C C T A G A A A G G A C A T A T T A A A A G A T A A A A G T A C G A A G T A G G A G T C A T T C G T G T C G
4750 4760 4770 4780 4790 4800
A A G C A T A T G C A G T T A G C A G A C A T T T T C T T T G C A C A C T C A G G C C A T T G T T T G C A G T A C A T T
T T C G T A T A C G T C A A T C G T C T G T A A A A G A A A C G T G T G A G T C C G G T A A C A A A C G T C A T G T A A
4810 4820 4830 4840 4850 4860
G C A T C A A C A C C A G G A T T T A A G G A A G A A G C A A A T A C C T C A G T T G C A T C C C A G A A G C C T C C A
C G T A G T T G T G G T C C T A A A T T C C T T C T T C G T T T A T G G A G T C A A C G T A G G G T C T T C G G A G G T
4870 4880 4890 4900 4910 4920
A A G T C A G G T T G A T G A G C A T A T T T T A C T C C A T C T T C C A T T T T C T T G T A C A G A G T A T T C A T T
T T C A G T C C A A C T A C T C G T A T A A A A T G A G G T A G A A G G T A A A A G A A C A T G T C T C A T A A G T A A
4930 4940 4950 4960 4970 4980
T T C T T C A T T T T T T C T T C A T C T C C T C C T T T A T C A G G A T G A A A C T C C T T G C A T T T T T T T A A A
A A G A A G T A A A A A A A G A A G T A G A G G A G G A A A T A G T C C T A C T T T G A G G A A C G T A A A A A A A T T T
4990 5000 5010 5020 5030 5040
T A T G C C T T T C T C A T C A G A G G A A T A T T C C C C C A G G C A C T C C T T T C A A G A C C T A G A A G G T C C
A T A C G G A A A G A G T A G T C T C C T T A T A A G G G G G T C C G T G A G G A A A G T T C T G G A T C T T C C A G G
5050 5060 5070 5080 Hind II, III A C 5090 5100
A T T A G C T G C A A A G A T T C C T C T C T G T T T A A A A C T T T A T C C A T C T T T G C A A A G C T T T T T G C A
T A A T C G A C G T T T C T A A G G A G A G A C A A A T T T T G A A A T A G G T A G A A A A C G T T T C G A A A A A C G T
5110 5120 5130 5140 5150 5160
A A A G C C T A G G C C T C C A A A A A A G C C T C C T C A C T A C T T C T G G A A T A G C T C A G A G G C C G A G G C
T T T C G G A T C C G G A G G T T T T T T C G G A G G A G T G A T G A A G A C C T T A T C G A G T C T C C G G C T C C G
5170 5180 5190 5200 5210 5220
G G C C T C G G C C T C T G C A T A A A T A A A A A A A A A T T A G T C A G C C A T G G G G C G G A G A A T G G G C G G A
C C G G A G C C G G A G A C G T A T T T A T T T T T T T T T A A T C A G T C G G T A C C C C G C C T C T T A C C C G C C T
A C T G G G
T G A C C C

by 233 additional coding triplets. The predicted molecular weight and amino acid composition for this protein are both in fairly good agreement with the values reported for analyses of VP3, except for the large excess of glycine reported in the amino acid analyses for VP3 (*18*). Amino acid sequence data are not available for VP3, but it seems likely that this sequence of triplets does code for VP3. Gibson *et al.* have already proved that parts of VP2 and VP3 are coded for by the same segment of DNA (*23*), but VP3 may not be derived from cleavage of VP2 (*22, 24*).

A third long sequence of sense triplets would be transcribed from the early strand of SV40 DNA, within the early region. This extends from positions 4506 to 2611, but the sequence does not begin with an initiator codon. The first potential initiator codon AUG lies at positions 4411 to 4409 (descending numbers indicate a reading from right to left) and is followed by 599 sense triplets. The region covered by the sense triplets includes the entire area in which temperature-sensitive mutants in the gene A product of SV40 have been mapped (*25, 26*). Mutant virus with deletions in this region may produce smaller peptides derived from the A protein (*16, 27*). The carboxyl-terminal sequence of the protein coded for by this DNA would be rich in proline, and this is consistent with observations on SV40 A protein (*28*) and on an SV40 directed peptide coded for by an adenovirus 2 SV40 hybrid virus (*29*). However, it is unlikely that this is the total sequence coding for the A protein. A deletion mutant of SV40 removes Hind II/III-H and Hind-I, fusing Hind II/III-A and -B at the Hind III cleavage site. The mutant virus produces a protein with a molecular weight of 33,000, which shares tryptic peptides with the early protein with a molecular weight of 94,000 produced by wild-type SV40 (*16, 30*). As was pointed out to us (*31*), this deletion would place translation termination triplets immediately beyond the end of the coding sequence in Hind II/III-A. Therefore, Hind II/III-A contains long coding sequences for the SV40 A protein, even though the longest continuous stretch of sense codons in Hind II/III-A corresponds to a protein of molecular weight less than 22,000.

The discrepancy between the molecular weight of the A protein estimated by sodium dodecyl sulfate gel electrophoresis (approximate molecular weight, 94,000) and the protein that could be coded for by the largest contiguous set of coding sequences in the early region of SV40 (approximate molecular weight, 70,000) has been discussed (*10, 32*). This discrepancy is even more pronounced if one assumes that translation initiates with an AUG codon. The A protein synthesized in vitro by SV40 early mRNA isolated from infected cells has been reported to have the same molecular weight as that directly isolated from infected cells (*27, 30, 33*), so that posttranslational addition of material to the A protein or modification of the cellular translation apparatus are unlikely explanations for this discrepancy. The A protein is also produced in transformed cells containing a single copy of SV40 DNA (*34*), and we suspect that covalent rearrangements of viral DNA are not a likely explanation. SV40 complementary RNA (cRNA), transcribed in vitro from viral DNA, directs the synthesis of a polypeptide that reacts with antiserum from animals bearing SV40-induced tumors (that is, T antiserum) and has a molecular weight of about 60,000, which is consistent with the long coding sequence in the early region of viral DNA (*27, 33, 35*). The most acceptable explanation is that the nucleotide sequence of early mRNA is not identical to the DNA sequence of the early region of SV40 (*5, 26*) (see below).

The 3′ end of early SV40 mRNA lies close to the region between residues 2520 and 2500 (Figs. 1 and 2). The 3′ terminal sequences include a set of four consecutive AUG triplets, followed by 91 sense triplets, reading in a frame different from that in which the long coding stretch of the A gene is translated, but there is no evidence that indicates that an additional peptide is coded for by this region in vivo.

Sequences Outside of the Main Coding Regions of SV40

A remarkable and unanticipated feature of SV40 virus is that almost 23 percent of the DNA is not part of the three coding regions that obviously direct the synthesis of long regions of peptide chains in known proteins. This 23 percent of the DNA can be divided into several domains, including the origin of replication, a cluster of tandem repeated sequences, and segments of DNA that could potentially code for peptides.

After the original demonstrations that SV40 DNA replication initiated at a unique site and proceeded bidirectionally (*36*), there has been an extensive effort to define the minimal SV40 sequences necessary on a circular DNA molecule in order that it be replicated in the presence of helper SV40 virus. This function has now been confined to a region of less than 100 nucleotides between about 5119 and 5198 (*22, 37*) (Fig. 2). A detailed derivation and a discussion of the sequence embedding the origin of replication have been presented (*3*). Two of the most outstanding features of the sequence are the presence of certain symmetric regions, including a long twofold rotational symmetry between residues 5147 and 5172 and the presence of a long, continuous A·T (T, thymidylic acid) stretch, including eight consecutive deoxyadenylic acids on one strand of DNA. A related papovavirus, BK virus, also has a long, although imperfect, symmetry region adjacent to a long A·T-rich stretch with nine consecutive deoxyadenylic acids near or at the origin of replication (*38*). Polyoma is a somewhat more distantly related papovavirus and may show less striking symmetry at the origin of replication but has a long A·T stretch with eight consecutive deoxyadenylic acids and a sequence that shows partial homology with SV40 (*39*). The recently reported nucleotide sequence about the origin of replication of bacteriophage lambda (*40*) also contains a long region of imperfect symmetry adjacent to a series of 16 residues, 14 of which are deoxyadenylic acids. Various mechanisms may be used to initiate DNA replication, but replication of papovavirus DNA resembles that of host cell DNA in that replication initiates on double-stranded DNA, and neither discontinuities in parental strand DNA nor covalent protein DNA linkages have been detected. In view of the diversity of mechanisms of synthesis for the initiation of RNA, it seems unlikely that all origins of DNA replication will be readily recognized from the DNA sequence alone (*41, 42*), but it also seems unlikely that some more striking features of the origin of replication of the papovaviruses do not have functional significance directly in the process of initiation of DNA replication or in interrelations between this process and the control of transcription. The presence of similar features in sequences of several replication origins is consistent with the suggestion that some initiation sites for replication of host cell DNA may have features similar to the SV40 origin (*43*).

Another remarkable feature of the SV40 sequence that has been the source of previous comment is that approximately 3 percent of the genome adjacent to the origin of replication is occupied by a set of three tandem repeats of sequences. The most extensive repeat comprises more than 1 percent of the viral genome. Analysis of the composition of deletion mutants indicates that major

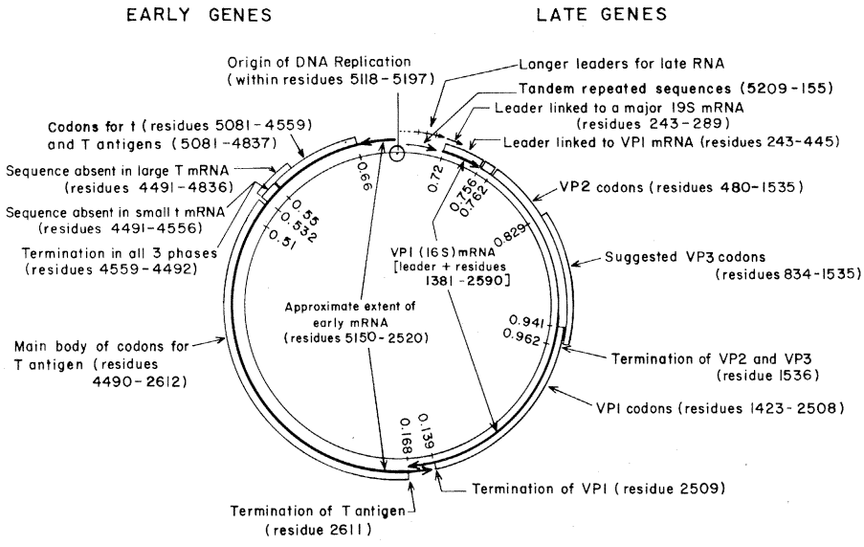


Fig. 2. Schematic representation of SV40 genome indicating the location of some of the principal features. The genome is represented as a circle with the origin of replication at the top. Numbers within the inner circle refer to locations in fractional genome length relative to the single Eco RI restriction endonuclease cleavage site within SV40 DNA. Residue numbers and labels refer to the numbers used in the sequence shown in Fig. 1. Arcs with arrowheads indicate positions of mRNA with arrowheads pointing in the 5′ to 3′ direction. Outer boxes indicate continuous coding segments within the viral mRNA.

parts of single copies of each repeat can be removed and in at least one case both copies of a repeating sequence can be removed in SV40 and leave a viable virus (*44*). Repeating sequences are not present in ϕX174 and no long tandem repeated sequence has been reported in bacteriophage genomes. However, reiterated sequences of varying length and perfection have been recognized for some time as a property of the DNA of higher eukaryotes, and it has been suggested (*11*) that there may be some undefined analogy between the reiterated sequences of SV40 and of host cell DNA. Each repeat represents a local twofold translational symmetry, and the total expanse of this bloc of reiterated sequences is about 180 nucleotides. Since the completion of these analyses, we have found that a portion of our viral stock has an insertion of approximately 37 nucleotides at about position 100 of the present sequence (Fig. 1). The sequence appears to be a copy of the sequence between residues 154 to 192, so that each copy of the tandem repeat is 91 bases long. This may represent a property of the ancestral line of the virus whose DNA sequence shows in Fig. 1.

Much of the remaining sequence of SV40 DNA is composed of blocs of nucleotides whose transcripts would include an AUG, followed by termination codons. It is not inconceivable that some or all of these blocs could be translated to form short peptides. Sequences corresponding to these regions of SV40 are present in cytoplasmic polyadenylylated RNA, probably on polysomes. The most accessible of these regions is a sequence derived from the right-hand portion of the Hind II/III-C fragment, the adjacent fragments of Hind II/III-L, -M, and the first few nucleotides of Hind II/III-D. The cytoplasm of SV40 infected cells contains large amounts of RNA transcribed from this region (*4*). At the end of the region, there is a segment of the DNA sequence whose transcript was not abundant in cytoplasmic polyadenylated RNA (*4*). This DNA is followed by the DNA whose sequence corresponds to late RNA coding for the amino acids of VP2. Most of the RNA transcript of the initial part of the region is not covalently linked to the mRNA for VP2 (*4, 7*). A large part (extending from residues 243 to 445) has been observed to be linked to the main body of mRNA that encodes VP1 (*7, 45*); this mRNA is transcribed from the region of SV40 DNA beginning at residue 1382 (Fig. 2). The VP1 mRNA also contains 7-methylguanylic acid (*4, 46*) attached to the 5′ end of the RNA. The sequence preceding the initiator codon for the short coding sequence overlapping Hind II/III-C, -L, -M, and -D is similar to that preceding the initiator codon of VP2, so that it is possible that at some stage in the infection it might be translated into protein. Alternatively, local secondary structure, such as the possible base pairing between residues 246 to 258 and 274 to 283 or between residues 244 to 263 and 295 to 313, might hinder initiation of translation at the first AUG in the RNA sequence. A principal form of 19*S* late mRNA has as its main body a continuous transcript beginning at nucleotide 474, and a 5′ terminal sequence of 47 nucleotides transcribed from DNA beginning at position 243. Some late polyadenylated RNA has longer 5′ terminal sequences, transcribed from templates extending further into Hind II/III-C, partly to a region that appears to end at residue 182. There are late RNA species whose 5′ end extends to each of at least three additional positions in the Hind II/III-C fragment that is closer to the Hind II/III-A fragment. The most prominent of these start points for late RNA lies near or within the longest A·T stretch at the origin of DNA replication. Unlike the 16*S* mRNA, some of the molecules of 19S to 20*S* late polyadenylated RNA are a faithful transcript of a continuous segment of viral DNA, without any gaps.

Until recently, mRNA from the early portion of the SV40 genome was not investigated as extensively because of the relatively small amounts of early mRNA made in infected cells. Studies of deletion mutants indicated that this region can itself be subdivided into two domains: a region between approximately 4500 and 4800 that can be deleted without affecting the viability of the virus (*27, 32*) and a region from about 4800 up to and including the origin of replication that must be maintained largely intact (*22*) for viral A protein formation or function. Several workers (*27, 47, 48*) have found a second early peptide in infected or transformed cells. This peptide (termed small t antigen) also reacts with serums from tumor-bearing animals. Several of the methionine-containing peptides in tryptic digests of small t antigen are also present in digests of the A protein (*27, 47*). The simplest possibility is that small t antigen is coded for by the 174 codons between residues 5081 and 4559 in the sequence shown in Fig. 1. Between about 4500 and 4630 there is an A·T-rich DNA sequence whose early strand transcript would include at least three termination codons in each possible reading frame. The codons between 5081 and 4559, and the codons between 4490 and 2611 of SV40 could together code for a peptide with a molecular weight about 90,000—about the size of the A protein. Berk and Sharp (*49*) have evidence that SV40 early mRNA contains sequences that probably derive from noncontiguous regions of the DNA. We have noted that oligonucleotide maps of SV40 early mRNA complementary to Hind II/III-A contain all the expected products plus one product not present in cRNA prepared in vitro. Our analyses show that a substantial part of early 19*S* mRNA consists of transcripts from residue 5150 to residue 4837, joined

to transcripts of residues 4491 onward, while a smaller part of the early 19S mRNA is a more complete transcript of the early region that still lacks 66 of the bases predicted from the DNA sequence between residues 4491 and 4556 (*50*). Therefore, a situation similar to but perhaps even more complex than the joining of RNA segments in the late mRNA of SV40 also occurs in early SV40 mRNA, with the additional effect that fusion occurs within coding regions. Therefore, the conclusion is that the A protein gene is not a contiguous DNA sequence, but is interrupted by a segment of 346 bp whose transcript is not present in A protein mRNA (*7, 27*). Deletion mutants lacking the region from about 4500 to 4800 make full-sized A protein (*32*) but not small t antigen (*27, 51*). This result and the results of mRNA analysis both indicate that not all the codons from small t contribute to the A protein. Further studies of early mRNA and the amino acid sequence of the early proteins are needed to exclude more complex arrangements of coding regions. The extensive studies of splicing in adenovirus mRNA show that complex fusion events may occur, although direct evidence of fusion within an active coding region of adenovirus mRNA has not yet been published. A DNA insert has been found in clones of mouse β-globin genes, a finding that was accompanied by the suggestion that this sequence may be deleted from RNA during or after transcription of the globin genes (*52*). If this suggestion is correct, then animal cell genes may commonly be encoded in noncontiguous segments of DNA and the same segment of DNA may contain codons for shared amino acid sequences present in more than one protein.

Selection of Codons in SV40 mRNA

Table 1 shows the codons used in the regions of SV40 DNA that are known to code for proteins. The selective use of codons in these regions of SV40 DNA, as well as in other eukaryotic and prokaryotic messages has been discussed (*6, 9, 14, 41, 53–57*). The selective use of codons in some regions of SV40 DNA is very striking when the entire sequence is examined. Table 1 includes the relatively small number of codons that derive from the overlap of VP1, VP2, and VP3 genes. The overlap would have some slight tendency to reduce the selective use of codons so that the results in Table 1 probably represent a minimum estimate of the selection pressure for codons in SV40 genes.

Sanger *et al.* (*41*) have shown that there is an excess of uridylic acid over other nucleotides in the third position of the codons in the genes of bacteriophage φX174 (*41*). In SV40, the bias against cytidylic acid in the third position of codons is especially marked when there is a uridylic acid in the second position in the codon. There are only nine codons (0.6 percent of all codons) that have U-C as their second and third residues, and the entire virion contains no AUC codons for isoleucine. In contrast, the AUA codon is rare both in the coat protein gene of small RNA bacteriophages (*53*) and in the genes of φX174 (*41*), but it is used with considerable frequency by SV40. The isoleucine codon AUC is not used by SV40 but is used at five positions in sea urchin histone H2B mRNA (*54*). The dinucleotide UC occurs in coding regions of SV40 more often in positions 1 and 2 or 3 and 1 than in positions 2 and 3 of coding triplets, presumably as a consequence of selection exerted during mRNA translation.

The DNA functioning in mammalian nuclei has been known for some time to have a marked deficiency in the dinucleotide C-G, and this deficiency has been demonstrated also in papovavirus DNA. The deficiency of C-G in the coding region of SV40 occurs in all three reading phases and is more marked than that for the whole viral DNA since the region of SV40 DNA near the origin of replication has a number of C-G sequences. Deficiency of codons containing C-G was also noted in rabbit (*55*) and human (*56*) β-globin mRNA, although it is less marked in insulin mRNA (*57*), human α-globin mRNA (*58*), or sea urchin histone H2B mRNA (*54*). It results in a marked preference for AG purine codons for arginine, compared with CGN codons (N, any nucleotide), a preference opposite to the situation in MS2 and φX174. The globin mRNA's show a marked preference for the CUG codon for leucine, and this preference is also shown in the partial sequence for the insulin mRNA. However, in SV40, UUA and UUG codons are used considerably more frequently than CUN codons and in particular there is no preferred use of the CUG codon.

Table 1. Codon usage in SV40 mRNA. The left column represents the first nucleotide of a codon; the second column shows the nucleotides in the third position of each codon. The four columns labeled C, A, G, and U are the four possible second-position nucleotides. For example, codon CUA occurs 16 times; CAU occurs 21 times; and GCU occurs 72 times. All codons for VP1, VP2, and the stretches of sense triplets that include probable codons for the A protein and little t antigen are included.

Position 1	Position 3	Position 2			
		C	A	G	U
C	C	14	10	1	3
	A	25	42	0	16
	G	0	31	2	20
	U	38	21	2	23
A	C	24	25	12	0
	A	26	70	36	21
	G	0	35	26	40
	U	40	44	32	45
G	C	12	31	15	2
	A	21	60	35	21
	G	0	41	24	31
	U	72	58	20	38
U	C	9	23	18	4
	A	15	3	1	40
	G	0	0	26	39
	U	29	32	14	60

The A·T-Rich Regions of SV40 DNA

The early studies of SV40 DNA demonstrated that there were certain sequences that selectively bound the DNA binding protein coded for by gene 32 of phage T4 (*59*), and were selectively modified by water-soluble carbodiimides (*60*). It was suggested that these regions might be particularly A·T-rich. The SV40 sequence shows that one of these A·T-rich regions lies near position 4500 on the SV40 genome and corresponds to the segment containing termination codons preceding the continuous coding sequence of the A protein. Other A·T-rich regions less clearly defined by the above studies include regions on either side of the imperfect repeats of A·T-rich sequences occurring in DNA coding for the 3′ end of the sequences preceding the coding region of VP1 mRNA and other A·T-rich repeats in the early region within the segment persumably coding for small t antigen (*4*). The sequence of the templates for late mRNA does not obviously provide support for mechanisms of RNA fusion that would require extensive similarity or complementarity between the 3′ end of the donor and the 5′ end of the acceptor RNA.

Runs of six or more uridylic acids are a prominent feature of some transcription termination sequences in prokaryotes and perhaps in RNA polymerase III transcripts in higher cells (*61*). There are six consecutive uridylic acids in the RNA transcript of the late strand about 40 bases beyond the 3′ end of cytoplasmic late mRNA, and in the late mRNA preceding the initiator codon for VP3, and seven consecutive uridylic acids near the 3′ end of early mRNA. There are seven

consecutive uridylic acids (residues 1875 to 1881) in the middle of the VP1 mRNA and two sets of six consecutive uridylic acids (residues 4299 to 4304 and 4145 to 4150) within the coding region for the A protein. With the exception of the sequence at the end of late mRNA, the consecutive uridylic acids are not preceded by RNA sequences rich in guanylic acid or precisely self-complementary regions such as are found in prokaryotic termination signals. The hexanucleotide AAUAAA is found within the 25 bases preceding the poly(A) of many eukaryotic mRNA's, including SV40 early and late mRNA (*8, 62*), and may be part of the signal determining the location of the 3′ end of polyadenylated RNA. However, it can also be found internally in SV40 early mRNA (residues 3188 to 3193) and hence is not in itself a sufficient signal.

RNA polymerase II is responsible for the synthesis of the bulk of mRNA in animal cells. Unfortunately, identification of promoters for RNA polymerase II has not been possible because of the lack of rigorous demonstration of precise and faithful initiation of transcription in vitro by this enzyme, and of the limited genetic analysis available. The A·T-rich sequences and the rotationally symmetric sequences adjacent at the origin of DNA replication are promising candidates for the SV40 early promoters since the 5′ end of the bulk of SV40 early mRNA lies just downstream from these sequences (*63*). The 5′ terminal nucleotide of the body of early mRNA is transcribed from residue 5143. The A protein binds to SV40 DNA very near this position (*64*), and functional A protein inhibits early mRNA synthesis in vivo (*65*). The relation between the protein binding site on the DNA and the position of the 5′ end of the mRNA resembles that between the repressor and the *lac* mRNA in *Escherichia coli*.

In summary, the sequence analysis of SV40 DNA has shown a remarkably intricate structure, including features such as overlapping genes, reminiscent of recent discoveries in gene structure of small DNA phages, but other features not noted in prokaryotic systems. Perhaps the most remarkable result is the probability that genes coding for single peptides may be composite, including segments derived from noncontiguous regions of DNA.

Note added in proof: After this article was submitted for publication we learned that A. Smith and his colleagues have obtained partial amino acid sequence data for amino terminal regions of small and large T peptides synthesized in vitro. The sequences of both peptides match the codons from position 5100 onward in the DNA sequence. We also learned that a similar suggestion about the origin of the two early proteins of polyoma had been considered by Ito and his colleagues (*72*).

References and Notes

1. J. Tooze, *The Molecular Biology of Tumor Viruses* (Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y., 1974).
2. R. Dhar, S. M. Weissman, B. S. Zain, J. Pan, A. M. Lewis, Jr., *Nucl. Acid Res.* **1**, 595 (1974); K. N. Subramanian, R. Dhar, J. Pan, B. S. Zain, S. M. Weissman, in *Molecular Mechanisms in the Control of Gene Expression* (Academic Press, New York, 1976).
3. K. N. Subramanian, R. Dhar, S. M. Weissman, *J. Biol. Chem.* **252**, 355 (1977).
4. R. Dhar, K. N. Subramanian, J. Pan, S. M. Weissman, *Proc. Natl. Acad. Sci. U.S.A.* **74**, 827 (1977).
5. R. Dhar, V. B. Reddy, S. M. Weissman, *J. Biol. Chem.* **253**, 612 (1978).
6. V. B. Reddy, R. Dhar, S. M. Weissman, *ibid.*, p. 621.
7. B. Thimmappaya *et al. Cold Spring Harbor Symp. Quant. Biol.*, in press.
8. B. Thimmappaya, B. S. Zain, R. Dhar, S. M. Weissman, *J. Biol. Chem.*, in press; B. S. Zain, B. Thimmappaya, R. Dhar, S. M. Weissman, *ibid.*, in press.
9. J. Pan, V. B. Reddy, B. Thimmappaya, S. M. Weissman, *Nucl. Acid Res.* **4**, 2539 (1977).
10. B. Thimmappaya and S. M. Weissman, *Cell* **11**, 837 (1977).
11. K. N. Subramanian, V. B. Reddy, S. M. Weissman, *ibid.* **10**, 497 (1977).
12. W. Fiers *et al.*, *FEBS Fed. Eur. Biochem. Soc. Proc. Meet.* **10**, 17 (1975); A. Van de Voorde, R. Contreras, R. Rogiers, W. Fiers, *Cell* **9**, 117 (1976); R. Contreras, G. Volckaert, F. Thys, A. Van de Voorde, W. Fiers, *Nucl. Acid Res.* **4**, 1001 (1977); H. Van Heuverscuyn, A. Van de Voorde, W. Fiers, *ibid.*, p. 1015.
13. E. Jay, R. Roychoudary, R. Wu, *Biochem. Biophys. Res. Commun.* **69**, 678 (1976).
14. G. Volckaert, R. Contreras, E. Soeda, A. Van de Voorde, W. Fiers, *J. Mol. Biol.* **110**, 467 (1977).
15. N. P. Salzman and G. Khoury, *Comprehensive Virology*, H. Fraenkel-Conrat and R. R. Wagner, Eds. (Plenum, New York, 1974), pp. 63–141.
16. K. Rundell, J. K. Collins, P. Tegtmeyer, H. L. Ozer, C.-J. Lai, D. Nathans, *J. Virol.* **21**, 636 (1976).
17. R. B. Carroll and A. E. Smith. *Proc. Natl. Acad. Sci. U.S.A.* **73**, 2254 (19 C. Ahmad-Zadeh, B. Allet, S. Greenblatt, R. Weil, *ibid.*, p. 1097.
18. J. L. Anderson, C. Chang, P. Mora, R. B. Martin, *J. Virol.* **21**, 459 (1977).
19. T. Kempe, W. Beathe, W. Konigsberg, S. M. Weissman, unpublished results.
20. P. S. Greenaway and D. LeVine, *Biochem. Biophys. Res. Commun.* **52**, 1221 (1973); W. Beattie, T. Kempe, W. Konigsberg, S. M. Weissman, unpublished results.
21. C.-J. Lai and D. Nathans, *Virology* **75**, 335 (1976).
22. C. Cole, T. Landers, S. Goffe, S. Manteuil-Brutlag, P. Berg, *J. Virol.* **24**, 277 (1977).
23. W. Gibson, T. Hunter, B. Cogen, W. Eckhart, *Virology*, in press.
24. A. Smith, personal communication.
25. C.-J. Lai and D. Nathans, *Virology* **60**, 466 (1974).
26. ———, *ibid.* **66**, 70 (1975); J. Mertz, thesis, Stanford University (1975).
27. L. V. Crawford, C. N. Cole, A. E. Smith, E. Paucha, P. Tegtmeyer, K. Rundell, P. Berg, *Proc. Natl. Acad. Sci. U.S.A.* **75**, 117 (1978).
28. D. Livingston, personal communication.
29. A. Bothwell and G. Fey, in preparation.
30. K. Rundell, J. K. Collins, P. Tegtmeyer, H. L. Ozer, C. J. Lai, D. Nathans, *J. Virol.* **21**, 633 (1977).
31. E. Cole, personal communication.
32. T. Shenk, J. Carbon, P. Berg, *J. Virol.* **18**, 664 (1976).
33. C. Prives and H. Aviv, *INSERM* **47** (1975); A. Smith, S. T. Bayley, T. Wheeler, W. F. Mangel, *ibid.*, p. 331.
34. B. Steinberg, R. Pollack, W. Topp, M. Botchan, *Cell*, in press.
35. B. E. Roberts, K. J. Danna, M. Gorecki, R. C. Mulligan, A. Rich, S. Rozenblatt, *INSERM* **47**, 313 (1975); F. Greenblatt, B. Allet, R. Wu, C. Ahmad-Zadeh, *J. Mol. Biol.* **109**, 361 (1976).
36. D. Nathans and K. Danna, *Nature (London) New Biol.* **236**, 200 (1972); M. M. Thoren, E. D. Sebring, N. P. Salzman, *J. Virol.* **10**, 462 (1972).
37. M. Gutai and D. Nathans, personal communication; T. Shenk, K. N. Subramanian, P. Berg, unpublished results.
38. R. Dhar, C.-J. Lai, G. Khoury, *Cell* **13**, 345 (1978).
39. E. Soeda, K. Miura, A. Nakaso, G. Kimura, *FEBS Lett.* **79**, 383 (1977); T. Friedman, personal communication.
40. K. Denniston-Thompson, D. D. Moore, K. E. Kruger, M. E. Furth, F. R. Blattner, *Science* **198**, 1051 (1977).
41. F. Sanger, G. M. Air, G. G. Barrell, N. L. Brown, A. E. Coulson, J. C. Fiddes, C. Hutchison, P. M. Slocombe, M. Smith, *Nature (London)* **265**, 681 (1977).
42. D. Bastia, *Nucl. Acid Res.* **4**, 3123 (1977).
43. R. G. Martin and A. Oppenheim, *Cell* **11**, 859 (1977).
44. K. Subramanian, T. Shenk, P. Berg, personal communication.
45. M. L. Celma, R. Dhar, J. Pan, S. M. Weissman, *Nucl. Acids Res.* **4**, 2549 (1977); Y. Aloni, *Cold Spring Harbor Symp. Quant. Biol.*, in press; Ming Ta Hsu, *ibid.*, in press; P. K. Ghosh, V. B. Reddy, J. Swinscoe, P. V. Choudary, P. Lebowitz, S. M. Weissman, *J. Biol. Chem.*, in press.
46. S. Lavi, personal communication; W. Fiers, personal communication; P. Ghosh, J. Swinscoe, P. Lebowitz, S. M. Weissman, unpublished observations.
47. C. Prives, E. Gilboa, M. Revel, E. Winocur, *Proc. Natl. Acad. Sci. U.S.A.* **74**, 457 (1977); D. Simmons and M. Martin, in preparation.
48. W. Topp, personal communication.
49. A. Berk and P. Sharp, personal communication.
50. P. K. Ghosh and V. B. Reddy, unpublished observation; P. Sharp, personal communication.
51. W. Topp, personal communication.
52. S. M. Tilghman, D. C. Tiemeier, F. Polsky, M. H. Edgell, J. G. Seidman, A. Leder, L. W. Enquist, B. Norman, P. Leder, *Proc. Natl. Acad. Sci. U.S.A.* **74**, 4406 (1977).
53. W. Fiers *et al.*, *Nature (London)* **260**, 500 (1976).
54. M. L. Birnstiel, W. Schaffner, H. O. Smith, *ibid.* **266**, 603 (1977).
55. A. Efstratiadis, F. P. Kafatos, T. Maniatis, *Cell* **10**, 511 (1977).
56. C. A. Marotta, J. T. Wilson, B. G. Forget, S. M. Weissman, *J. Biol. Chem.* **252**, 5040 (1977).
57. A. Ullrich, J. Shine, J. Chirgwin, R. Pictet, E. Tischer, W. J. Rutter, H. M. Goodman, *Science* **196**, 1313 (1977).
58. J. Wilson, L. Wilson, C. A. Marotta, B. G. Forget, S. M. Weissman, unpublished observations.
59. P. Beard, J. Morrow, P. Berg, *J. Virol.* **12**, 1303 (1973).
60. M. Chen, J. Lebowitz, N. P. Salzman, *J. Virol.* **18**, 211 (1976).
61. M. L. Celma, J. Pan, S. M. Weissman, *J. Biol. Chem.*, in press.
62. R. Dhar, K. N. Subramanian, B. S. Zain, A. Levine, C. Patch, S. M. Weissman, *INSERM* **47**, 25 (19 N. J. Proudfoot and G. G. Brownlee, *Nature (London)* **252**, 359 (19 C. Milstein, G. G. Brownlee, E. M. Cartwright, J. M. Jarvis, N. J. Proudfoot, *ibid.*, p. 354.
63. R. Dhar, K. N. Subramanian, J. Pan, S. M. Weissman, *J. Biol. Chem.* **252**, 368 (1977); S. I. Reed, G. R. Stark, J. C. Alwine, *Proc. Natl. Acad. Sci. U.S.A.* **73**, 3083 (1976).
64. S. I. Reed, J. Ferguson, R. W. Davis, G. R. Stark, *Proc. Natl. Acad. Sci. U.S.A.* **72**, 1605 (1975); D. Jessel, T. Landau, J. Hudson, T. Lalor, D. Tenen, D. M. Livingston, *Cell* **8**, 535 (1976); R. Tjian, personal communication.
65. P. Tegtmeyer, M. Schwartz, J. K. Collins, K. Rundell, *J. Virol.* **16**, 168 (1975).
66. K. J. Danna, G. H. Sack, Jr., D. Nathans, *J. Mol. Biol.* **78**, 363 (1973).
67. A. Maxam and W. Gilbert, *Proc. Natl. Acad. Sci. U.S.A.* **74**, 557 (1977).
68. C. A. Marotta, P. Lebowitz, R. Dhar, B. S. Zain, S. M. Weissman, *Methods Enzymol.* **29** (part E), 254 (1974).
69. T. Maniatis, A. Jeffrey, D. Kleid, *Proc. Natl. Acad. Sci. U.S.A.* **72**, 1184 (1975).
70. R. Yang, A. Van de Voorde, W. Fiers, *Eur. J. Biochem.* **61**, 119 (1975).
71. E. Jay and R. Wu, *Biochemistry* **15**, 3612 (1976).
72. Y. Ito, J. A. Brocklehurst, N. Spurr, M. Thiveral-Grignon, J. M. Griffiths, M. Fried, *EMBO INSERM Workshop on Early Proteins of Oncogenic DNA Viruses* (1977), in press.

73. Supported by NCI grant 16038 from the National Cancer Institute and by a grant from the American Cancer Society. We thank Drs. L. Crawford, C. Cole, R. Martin, P. Tegtmeyer, and W. Topp for valuable discussions about the structure of the early region of SV40, and Drs. P. Berg, P. Tegtmeyer, D. Nathans, and C. Cole for their critical reading of this manuscript.