Nucleotide sequence of the gene for the major structural protein of SV40 virus

J. Pan, V. B. Reddy, B. Thimmappaya and S. M. Weissman

Department of Human Genetics, Yale University School of Medicine, New Haven, CT 06510, USA

Received 13 June 1977

ABSTRACT

We have determined the sequence of the portion of Simian Virus 40 (SV40) that codes for the major structural protein of the virus. The gene contains 361 codons. Synonym codons for an amino acid are not used randomly. The dinucleotide CG occurs only once and there is 2 to 1 preference for uridylic acid in the third position of codons.

INTRODUCTION

Simian Virus 40 (SV40) is a small virus whose genome consists of a single supercoiled molecule of DNA that contains approximately 5,375 base pairs. The major structural protein of the virus, VP1, is produced in large amounts late in the infectious cycle and is a product of translation of 16S late mRNA (1). The amino acid composition of VP1 has been determined (2) and its amino terminal amino acid sequence has been analyzed and correlated with nucleotide sequences within SV40 DNA (3,4). We report here the nucleotide sequence of the region of SV40 DNA encoding the amino acids of VP1. This predicts the amino acid sequence of the protein and provides information about selective codon utilization by the virus. In the subsequent report (5) we compare nucleotide sequences in the DNA adjacent to and including the coding region for VP1 with the nucleotide sequences in the messenger RNA that directs the synthesis of VP1.

MATERIALS AND METHODS

These have been described in detail elsewhere (6,7,8). Briefly, restriction endonucleases Eco RI, Hind II,III, Alu I, Bam I, Mbo I and Pst¹ were obtained from commercial sources (New England Biolabs). The enzyme Hae III (9) and Hinf (10) were prepared by published procedures. The details of the methods for derivation of the sequence of Hind-II,III-G have been presented elsewhere (11). The remainder of the sequence was derived or confirmed predominantly by use of methods of DNA analysis. DNA fragments were prepared with radioactivity introduced separately at each restriction cleavage site

available within the segment of SV40 DNA extending from 0.945 to 0.170 fractional genome length relative to the single Eco RI site (see Fig. 1). Partial snake venom diesterase digests (12) of each labeled DNA fragment were prepared and the products fractionated two-dimensionally by electrophoresis and homochromatography. The pattern of oligonucleotides generally provided a sequence for approximately the 20 nucleotides adjacent to each restriction cleavage site. Each labeled fragment was also subjected to selected chemical degradation by the procedures of Maxam and Gilbert (13) and a series of gels was run with these products. This commonly provided information for sequences from about 3 to 100 nucleotides removed from each cleavage site. For the most part overlapping sequences were obtained from both strands of DNA in each segment of the gene.

Fig. 1 Restriction endonuclease cleavage sites used to derive the nucleotide sequence of the gene for VPI.

Top line: Hind II, III cleavage map of SV40 DNA, arranged so that the Hind II, III-G, B border forms the ends of the linear representation of the DNA.

Symbols to the left of each line represent restriction endonucleases (see Abbreviations) whose cleavage sites are shown by vertical bars through the line. Letters above the line represent longer restriction fragments. The bottom line shows location of DNA complementary to the translation initiation and termination codons for VPI.

RESULTS

The nucleotides encoding the amino acids of the VP1 have previously been located within the Hind II,III-K fragment 12 nucleotides from the junction with the Hind II,III-E fragment (Fig. 1). Similarly, the carboxy terminus of VP1 has been determined to be glutamine and the carboxy terminal peptides have been correlated with nucleotide sequences within the Hind II,III

Fig. 2 Products of partial snake venom diesterase digestion of the DNA fragment Hind II,III-F with radioactivity introduced at the end nearest Hind II, III-J. Letters indicate successive nucleotides removed by exonucleolytic digestion. Fractionation of products was from left to right by electrophoresis on Cellogel at pH 3.5 and from below upwards by chromatography on DEAE TLC plates.

fragments 70 nucleotides from its junction with Hind II,III-B (Zain, Thimmapaya, submitted for publication; Kempe, Beattie, Konigsberg, Weissman, unpublished results). To determine the sequence of the remainder of the gene we analyzed Hind II,III-K, F and J.

The sequences of most parts of Hind II,III-K, F and J could be readily determined from the results of the various venom diesterase and gel "ladder" patterns obtained. In general, each gel pattern was obtained at least twice. Gels were generally run so that nucleotides from 3 to approximately 25, from 25 to 60 and from approximately 45 to 110 could be read from each of the available restriction enzyme cleavage sites. Representative patterns of oligonucleotides from limited venom diesterase digestion and ladder patterns from products of limited chemical degradation are shown in Figs. 2, 3 and 4. The derivation of the sequence of Hind II,III-G offered some additional problems and has been discussed in detail elsewhere (11).

When the complete sequence was derived (Fig. 5) it was apparent that the transcript corresponding to the late messenger RNA derived from this region of DNA would contain an AUG beginning 12 nucleotides within the Hind II, III-K fragment, followed by 361 sense codons followed by a termination codon UGA, within Hind II,III-G, 70 nucleotides from the junction with Hind II,III-B (Fig. 1). The nucleotide sequence for this region of DNA predicted an amino acid composition for VP1 (Table I). This composition is in good to excellent agreement with reported compositions for amino acid composition of VP1 purified by SDS acrylamide gel electrophoresis. The greatest discrepancy was that amino acid analyses showed more glycine and alanine than predicted. We presume this was due to contaminants in the protein. The estimated molecular weight of VP1 by sodium dodecyl sulfate gel electrophoresis is approximately 43,000. This is in fair agreement with the prediction of 361 amino acid residues, and the primary translation product of the VP1 gene may be modified in vivo. The overall composition of the protein is not very remarkable. The amino terminal sequence is somewhat rich in basic amino acids and, although

Fig. 3 Products of limited chemical degradation of SV40 DNA. The early strand of the DNA is labeled at the Eco RI cleavage site. Letters above the channels indicate selective cleavage at C, T (and C), G (and A), and A residues, respectively. Letters in the vertical column indicate successive nucleotides of the sequence.

- (a) Short electrophoretic run;
- (b) Long electrophoretic run.

Fig. 4 Products of limited chemical degradation of the DNA fragment Hind II, III-J with radioactivity introduced at the end nearest Hind II, III-F. The labeling of the figure is as in Fig. 3.

Sequence of SV40 VPI Gene

HindII, III E HindII, III K

GAA CUG CUC CUC AGU GGA UGU UGC CUU UAC UUC UAG GCC UGU ACG GAA GUG UUA CUU CUG CUC UAA AAG CUU

Met Ala Pro Thr Lys Arg Lys Gly Ser Cys Pro Gly Ala Ala Pro Lys Lys Pro Lys Glu Pro Val

AUG AAG AUG GCC CCA ACA AAA AGA AAA GGA AGU UGU CCA GGG GCA GCU CCC AAA AAA CCA AAG GAA CCA GUG

Gln Val Pro Lys Leu Val Ile Lys Gly Gly Ile Glu Val Leu Gly Val Lys Thr Gly Val Asp Ser Phe Thr

CAA GUG CCA AAG CUC GUC AUA AAA GGA GGA AUA GAA GUU CUA GGA GUU AAA ACU GGA GUA GAC AGC UUC ACU

Glu Val Glu Cys Phe Leu Asn Pro Gln Met Gly Asn Pro Asp Glu His Gln Lys Gly Leu Ser Lys Ser Leu

GAG GUG GAG UGC UUU UUA AAU CCU CAA AUG GGC AAU CCU GAU GAA CAU CAA AAA GGC UUA AGU AAA AGC UUA

Ala Ala Glu Lys Gln Phe Thr Asp Asp Ser Pro Asp Lys Glu Gln Leu Pro Cys Tyr Ser Val Ala Arg Ile

GCA GCU GAA AAA CAG UUU ACA GAU GAC UCU CCA GAC AAA GAA CAA CUG CCU UGC UAC AGU GUG GCU AGA AUU

Pro Leu Pro Asn Ile Asn Glu Asp Leu Thr Cys Gly Asn Ile Leu Met Trp Glu Ala Val Thr Val Lys Thr

CCU UUG CCU AAU AUU AAU GAG GAC UUA ACC UGU GGA AAU AUU UUG AUG UGG GAA GCU GUU ACU GUU AAA ACU

Glu Val Ile Gly Val Thr Ala Met Leu Asn Leu His Leu Gly Thr Gln Lys Thr His Glu Asn Gly Ala Gly Lys

GAG GUU AUU GGG GUA ACU GCU AUG UUA AAC UUG CAU UUA GGG ACA CAA AAA ACU CAU GAA AAU GGU GCU GGA AAA

Pro Ile Gln Gly Ser Asn Phe His Phe Phe Ala Val Gly Gly Glu Pro Leu Glu Leu Gln Gly Val Leu Ala

CCC AUU CAA GGG UCA AAU UUU CAU UUU UUU GCU GUU GGU GGG GAA CCU UUG GAG CUG CAG GGU GUG UUA GCA

Asn Tyr Arg Thr Lys Tyr Pro Ala Gln Thr Val Thr Pro Lys Asn Ala Thr Val Asp Ser Gln Gln Met Asn

AAC UAC AGG ACC AAA UAU CCU GCU CAA ACU GUA ACC CCA AAA AAU GCU ACA GUU GAC AGU CAG CAG AUG AAC

Thr Asp His Lys Ala Val Leu Asp Lys Asp Asn Ala Tyr Pro Val Glu Cys Trp Val Pro Asp Pro Ser Lys

ACU GAC CAC AAG GCU GUU UUG GAU AAG GAU AAU GCU UAU CCA GUG GAG UGC UGG GUU CCU GAU CCA AGU AAA

Asn Glu Asn Thr Arg Tyr Phe Gly Thr Tyr Thr Gly Gly Glu Asn Val Pro Pro Val Leu His Ile Thr Asn

AAU GAA AAC ACU AGA UAU UUU GGA ACC UAC ACA GGU GGG GAA AAU GUG CCU CCU GUU UUG CAC AUU ACU AAC

Thr Ala Thr Thr Val Leu Leu Asp Glu Gln Gly Val Gly Pro Leu Cys Lys Ala Asp Ser Leu Tyr Val Ser Ala

ACA GCA ACC ACA GUG CUG CUU GAU GAG CAG GGU GUU GGG CCC UUG UGC AAA GCU GAC AGC UUG UAU GUU UCU GCU

Val Asp Ile Cys Gly Leu Phe Thr Asn Thr Ser Gly Thr Gln Gln Trp Lys Gly Leu Pro Arg Tyr Phe Lys

GUU GAC AUU UGU GGG CUG UUU ACC AAC ACU UCU GGA ACA CAG CAG UGG AAG GGA CUU CCC AGA UAU UUU AAA

Ile Thr Leu Arg Lys Arg Ser Val Lys Asn Pro Tyr Pro Ile Ser Phe Leu Leu Ser Asp Leu Ile Asn Arg

AUU ACC CUU AGA AAG CGG UCU GUG AAA AAC CCC UAC CCA AUU UCC UUU UUG UUA AGU GAC CUA AUU AAC AGG

Arg Thr Gln Arg Val Asp Gly Gln Pro Met Ile Gly Met Ser Ser Gln Val Glu Glu Val Arg Val Tyr Glu

AGG ACA CAG AGG GUG GAU GGG CAG CCU AUG AUU GGA AUG UCC UCU CAA GUA GAG GAG GUU AGG GUU UAU GAG

Asp Thr Glu Glu Leu Pro Gly Asp Pro Asp Met Ile Arg Tyr Ile Asp Glu Phe Gly Gln Thr Thr Thr Arg

GAC ACA GAG GAG CUU CCU GGG GAU CCA GAC AUG AUA AGA UAC AUU GAU GAG UUU GGA CAA ACC ACA ACU AGA

Met Gln

AUG CAG UGA

Greenaway and Levine found no cysteine residues in their composition, our sequence contains 7 cysteine codons. It has been suggested that, in the related virus, polyoma, the capsid protein subunits may be held together by disulfide bonds (14), presumably between cysteines of the protein monomers.

Recent analysis of mRNA or DNA genes from prokaryotic and eukaryotic sources have shown that there is non-random selection of synonym codons used in individual messenger RNAs. This has been especially striking in the analyses of the messenger RNAs for rabbit (15) and human hemoglobin (16). The prominent features of these RNAs include the relative bias against the dinucleotide CG. The deficiency of CG in eukaryotic cell nuclear DNA has been known for a long time and in this respect. The dinucleotide CG occurs only

TABLE I

Amino Acid Composition of VP1 (moles per 100 moles of amino acid)

	Predicted from DNA sequence	Reported by Greenaway & LeVine ⁽²⁾		Predicted from DNA sequence	Reported by Greenaway & LeVine
Lys	6.98	6.51	Ala	5.03	6.41
His	1.68	1.55	Val	8.38	9.16
Arg	3.63	3.62	Met	2.23	1.77
Asp	10.89	10.61	Ile	4.19	4.07
Thr	8.66	7.56	Leu	7.82	9.14
Ser	4.75	5.68	Tyr	3.07	2.57
Glu	11.73	11.45	Phe	3.07	3.72
Pro	7.82	6.45	Cys	1.96	ND
Gly	8.10	9.68			

one time in the gene for VP1. Recently, Sanger and his colleagues have reported the nucleotide sequence of small DNA phage ϕ X 174 (17). They noted that there was a substantial preference for U over C in the third position of the codons of the various mRNAs of ϕ X 174. The preference expressed itself overall as approximately 2.1 uridylic acids for each cytidylic acid in the third position of codons. The mRNA for VP1 shows a quite similar result with approximately 2.35 uridylic acids for each cytidylic acid in the third position of codons (Table 2). This is even true for the first 39 nucleotides of VP1, although these nucleotides are also used in the second phase as the codons for the terminal amino acids of VP2 minor structural proteins of SV40. We have also observed a similar bias in the sequences for VP2, VP3 and the early genes.

The globin mRNA also exhibited other regularities, such as a marked preference for the codon CUG from among the six possible synonym codons shown for leucine. The viral mRNA shows no suggestion of this preference. Seventeen out of the twenty-eight codons for leucine in the viral mRNA belong to the UU purine family, although this codon is not used at all in beta globin mRNA. There are no AUC codons out of 15 isoleucine codons of VP1 mRNA and only one UUC codon out of 11 phenylalanine codons. The other messenger RNAs of SV40 also show a similar bias against these two codons above what might be expected from the preference for uridylic above cytidylic acid in the third position of codons. The codon AAC appears to occur almost as frequently as the codon AAU, so that the selection does not seem to be explicable simply on the basis of overall nucleotide composition of the triplet. The codon GUC occurs only once, compared with 15 uses of the codon GUU and a total of 30 valine codons in the VP1 region.

TABLE II
Use of Codons in VP1 Gene

1st Position	3rd Position	2nd Position			
		C	A	G	U
C	C	5	2	0	1
	A	11	8	0	2
	G	0	10	1	4
	U	12	4	0	4
A	C	9	8	3	0
	A	11	19	7	3
	G	0	6	5	8
	U	12	11	6	12
G	C	1	11	2	1
	A	4	11	12	4
	G	0	13	10	10
	U	13	9	5	15
U	C	2	5	4	1
	A	1	0	0	8
	G	0	0	3	9
	U	5	6	3	10

One curious feature of the mRNA is the occurrence of the run of six uridylic acids in the middle of the coding region of the viral message. SV40 DNA contains at least 15 runs of six or more deoxyadenylic acids that would be transcribed into runs of uridylic acids. These sequences are of some interest as this is a feature common to a number of strong transcription termination signals in prokaryotes. In most other sites except the VP1 mRNA, the sequence of six uridylic acids would either occur in anti-sense RNA or in a strategically located position, such as between two genes in the early region of SV40 or preceding the start of the VP3 gene in the late region of SV40. Except for features of the Hind II,III-G fragment that have been discussed elsewhere, there are no readily apparent long repeats or other symmetry elements within the sequence coding for VP1, nor within the nucleotides immediately preceding the codons for this protein. W. Fiers and his colleagues have recently published the nucleotide sequence of a portion of Hind-F (18) and also part of Hind-K (4 et op. cit.).

This work was supported by grants from the American Cancer Society and the National Institutes of Health.

REFERENCES

- 1 Prives, C. L., Aviv, H., Gilboa, E., Revel, M. and Winocour, E. (1974) Cold Spring Harbor Symp. Quant. Biol. **39**, 309-316.
- 2 Greenaway, P. S., and LeVine, D. (1973) Biochem. Biophys. Res. Comm. **52**, 1221-1227.
- 3 Lazarides, R., Files, J. A., and Weber, K. (1974) Virology **60**, 584-587.
- 4 Van de Voorde, A., Contreras, R., Rogiers, R. and Fiers, W. (1976) Cell, **9**, 117-120.
- 5 Celma, M.L., Dhar, R., Pan, J., and Weissman, S. M. Accompanying manuscript.
- 6 Marotta, C.A., Lebowitz, P., Dhar, R., Zain, B.S., and Weissman, S. M., in Methods in Enzymology, **29**, part E, Grossman, L., and Moldave, K. (ed.) (Academic Press, New York, 1974), 254-272.
- 7 Dhar, R., Reddy, V. B., and Weissman, S. M. Submitted for publication.
- 8 Thimmappaya, B., Dhar, R., Zain, B. S., and Weissman, S. M. Submitted for publication.
- 9 Middleton, J. H., Hutchison, C. A., III (1972) J. Virol. **10**, 40-50.
- 10 Roberts, R. J., Myers, P. A., Morrison, A., Murray, K. (1976) J. Mol. Biol. **103**, 199-208.
- 11 Zain, B. S., Thimmappaya, B., and Weissman, S. M. Submitted for publication.
- 12 Maniatis, T., Jeffrey, D., and Kleid, D. G. (1975) Proc. Natl. Acad. Sci. USA **72**, 1184-1188.
- 13 Maxam, A., and Gilbert, W. (1977) Proc. Natl. Acad. Sci. USA **74**, 557-564.
- 14 Walter, G., and Deppert, W. (1974) Cold Spring Harbor Symp. Quant. Biol. **39**, 255-257.
- 15 Efstratiadis, A., Kafatos, F. D., and Maniatis, T. (1977) Cell, in press.
- 16 Marotta, C. A., Wilson, J. T., Forget, B. G., and Weissman, S. M. (1977) J. Biol. Chem., in press.
- 17 Sanger, F., Air, G. M., Barrell, G. G., Brown, N. L., Coulson, A. R., Fiddes, J. C., Hutchison, C. A., Slocumbe, P. M., and Smith, M. (1977) Nature **265**, 687-694.
- 18 Contreras, R., Volckaert, G., Thys, F., Van de Voorde, A., and Fiers, W. (1977) Nucl. Acid Res. **4**, 1001-1015.

1

ABBREVIATIONS

Restriction Endonucleases:

<u>Eco</u> RI	Enzymes from plasmid-bearing <u>E. coli</u> .
<u>Hind</u> II,III	<u>Hemophilus influenzae</u> strain d enzyme.
<u>Alu</u> I	<u>Arthrobacter luteus</u> enzyme.
<u>Bam</u> I	<u>Bacillus amyloliquefaciens</u> .
<u>Mbo</u> I	<u>Moraxella bovis</u> enzyme.
<u>Pst</u>	<u>Providentia stuartii</u> .
<u>Hae</u> III	<u>H. aegyptius</u> enzyme.
<u>Hinf</u>	<u>H. influenzae</u> strain f.