(Reprinted from Nature, Vol. 273, No. 5658, pp. 113–120, May 11, 1978)

Complete nucleotide sequence of SV40 DNA

W. Fiers, R. Contreras, G. Haegeman, R. Rogiers, A. Van de Voorde, H. Van Heuverswyn, J. Van Herreweghe, G. Volckaert & M. Ysebaert

Laboratory of Molecular Biology, University of Ghent, Belgium

The determination of the total 5,224 base-pair DNA sequence of the virus SV40 has enabled us to locate precisely the known genes on the genome. At least 15.2% of the genome is presumably not translated into polypeptides. Particular points of interest revealed by the complete sequence are the initiation of the early t and T antigens at the same position and the fact that the T antigen is coded by two non-contiguous regions of the genome; the T antigen mRNA is spliced in the coding region. In the late region the gene for the major protein VP_1 overlaps those for proteins VP_2 and VP_3 over 122 nucleotides but is read in a different frame. The almost complete amino acid sequences of the two early proteins as well as those of the late proteins have been deduced from the nucleotide sequence. The mRNAs for the latter three proteins are presumably spliced out of a common primary RNA transcript. The use of degenerate codons is decidedly non-random, but is similar for the early and late regions. Codons of the type NUC, NCG and CGN are absent or very rare.

SIMIAN virus 40 (SV40) (and the distantly related polyoma virus) have been intensively studied both as models of eukaryote gene organisation and expression and because of their oncogenic potential (see refs 1–3 for recent reviews). SV40 infection in monkey cells follows a lytic cycle but in other cells such as mouse, rat, rabbit and human the infection is abortive and the SV40 genome or parts thereof can become integrated into the host cell genome resulting in transformation of the cell. SV40 can also cause tumours when inoculated into newborn animals. The determination of the complete sequence reported here now makes it possible to assign precise genomic locations to biological functions that have been discovered and analysed by many workers and which will be discussed below.

The SV40 virion contains, in addition to its genome wrapped in the host-derived histones H2A, H2B, H3 and H4, the major capsid protein VP_1 and two minor structural components, VP_2 and VP_3. SV40 DNA is a supercoiled, circular duplex of 5,224 base pairs. The single *Eco*RI restriction enzyme site is taken as zero point for the physical map (Fig. 1). The restriction enzymes *Hin*dII and III produce 13 fragments designated A to M[4,5], which are often used as a basis for further studies. The cleavage sites of many other restriction enzymes have been determined[1,2,6]. DNA replication initiates bidirectionally from around position 0.67 and terminates on the other side of the circle, where the two replication forks meet. The half-circle from 0.67 to 0.17, replicated and transcribed counterclockwise, corresponds almost exactly to the early region, while the segment 0.67 to 0.17 replicated and transcribed clockwise, corresponds to the late region (Fig. 1).

We report here the entire nucleotide sequence of SV40 DNA (strain 776). The earlier results were obtained by analysing highly labelled RNA transcribed *in vitro* from restriction fragments[7,8]. The introduction by Maxam and Gilbert[9] of the base-specific, partial chemical degradation of terminally labelled DNA, however, allowed much faster sequence analysis. All our more recent results were obtained by this approach, often applied on each of the two complementary strands. S. M. Weissman and his colleagues have pursued similar lines of investigation and their results are presented elsewhere (ref. 10 and references therein).

For representation of the sequence we have taken as a zero point the base pair present in the centre of the remarkable, 27-base-pair long palindrome located at position 0.663 (ref. 11). This centre lies at or very close to the origin of DNA replication. The single site on SV40 DNA for the restriction enzyme *Bgl*I is also around this position (the enzyme recognises the central part of the palindrome mentioned above). The nucleotide sequence of the early region is presented in Fig. 2 and the sequence of the late region in Fig. 3. For both regions, only the DNA strand with the same polarity as the messenger is given. This corresponds to a counterclockwise and a clockwise orientation respectively (see Fig. 1). The endpoint at the junction of the *Hin*d fragments C and B is somewhat arbitrary; it should be noted, however, that the end of the cytoplasmic late mRNA corresponds almost exactly to this position (see below).

The early region

Soon after infection of either permissive or non-permissive cells by SV40, virus-specific, poly(A)-containing 19S mRNA appears in the cytoplasm[12]. This early mRNA is transcribed counterclockwise and hybridises to about 48% of the genome from approximately 0.65 to 0.17 (refs 13, 14). The same region is also expressed in SV40-transformed cells, although transcription may proceed somewhat further in the 3′-end direction[15]. Temperature-sensitive mutants affected in an early function correspond to the complementation group A, and these mutations map in the *Hin*d fragments H, I, and occasionally B[16,17]. Gene *A* codes for

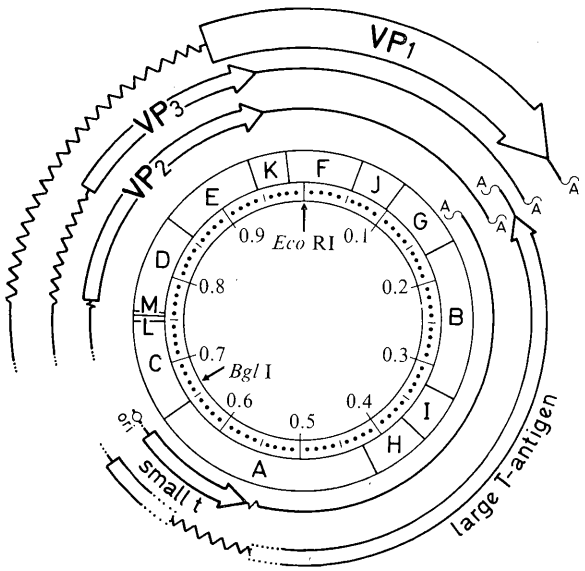

Fig. 1 Standard physical map of SV40 DNA and localisation of the main biological functions. The single cleavage site of the restriction enzyme *Eco*RI is used as a reference point for the physical map (inner circle). The next circle shows the position of the *Hin*dII and III restriction fragments A to M. These *Hin*d cleavage sites (and *Eco*RI) are also indicated on the total nucleotide sequence in Figs 2 and 3. The origin of DNA replication (ori) is at or close to 0.663; this site corresponds to a large palindromic sequence, part of which represents the recognition sequence of *Bgl*I. The centre of this palindrome is taken as the zero point for presenting the sequence of the early region (5′–3′ in counterclockwise orientation) in Fig. 2 and of the late region (5′–3′ in clockwise orientation) in Fig. 3. The five virus-coded proteins are indicated by blocked arrows; note that the T antigen is coded by two non-contiguous segments on the genome (there is some uncertainty in the position of the points as indicated by the dashes). Untranslated parts of the mRNA are shown as solid lines; dots indicate uncertainty as to the exact position of the 5′ end. Zigzag lines are used for the segments spliced out. A wavy line with an A illustrates the 3′-terminal polyA tail.

the T antigen, a protein with a molecular weight of about 90,000–100,000 (ref. 18). It is phosphorylated but presumably not glycosylated (at least not the nuclear T antigen[19]). T antigen is found in productive infections and in SV40-transformed cells. In the former, it is amongst other things essential for SV40 replication and for stimulating cellular functions; in non-permissive cells, T antigen is required for the initiation and presumably maintenance of transformation[1–3]. T antigen is an autoregulator; it shuts off its own synthesis at the level of transcription[20,21].

We have localised the carboxyterminal end of the T antigen at nucleotide 2,549 of the early region (Fig. 2 and ref. 22). The exact position of the N-terminal starting signal of this gene, however, has only recently become clear[23,24]. Indeed, originally there was some confusion because the region between map positions 0.54 and 0.175 can code for, at most, a polypeptide of molecular weight 72,000 while the estimated size of T antigen exceeds 90,000. The reasons for placing the boundary at about map position 0.54—and not closer to the origin—were based on the isolation and characterisation of a series of non-defective deletion mutants which lack sequence information in the region 0.54–0.59 and which still produce T antigen of normal size[25,26]. Moreover, nucleotide sequence information in the region immediately preceding 0.54 clearly indicated that translation termination codons are present in all three possible reading frames, thereby precluding continuous protein synthesis through this region[6,27].

Progress in the characterisation of polyoma virus (which infects mice) has revealed that its molecular biology is in many aspects similar to that of SV40[1,2]. T. Benjamin and coworkers have isolated and characterised what seems to be a new class of early mutants[28,29]. These are host range mutants growing normally only on certain cell types; they can no longer transform cells. They have been mapped in the proximal part of the early region[30], in a segment which may approximately correspond to the 0.54–0.59 area of SV40 DNA. Indeed, the SV40 mutants mentioned above which have a deletion in this region cannot transform cells to anchorage-independent growth[31]. The protein responsible for this second early function was subsequently identified and called small t antigen[26,32,33]. It has a molecular weight of 15,000–20,000 and is absent or truncated in nearly all deletion mutants which map between 0.54–0.59 (refs 26, 34, 35). A remarkable fact is that T and t are related; they share immunogenic determinants and most of the methionine-containing tryptic peptides of t are also present in the tryptic fingerprint of T. Both proteins normally have a blocked N-terminus, which impedes further detailed analysis. Paucha *et al.*[23], however, were able to synthesise unblocked polypeptides *in vitro*. t And T were found to have an identical N-terminal sequence. The start codon for both proteins was identified by correlating the protein chemistry data with the nucleic acid sequence[23,24]. t Starts at nucleotide 80 of the early region (Fig. 2) and is read continuously until the stop codon UAA at position 602. We assume a colinearity between the gene and the protein which is likely since Paucha *et al.*[33] were able to synthesise apparently normal t (but not T) using complementary RNA transcribed *in vitro* from SV40 DNA with *Escherichia coli* RNA polymerase.

t Is 174 amino acids long and has a molecular weight of 20,503; the sequence is shown in Fig. 4. The N-terminal methionine is blocked by acetylation (A. Mellor and A. E. Smith, personal communication). t Contains 19 lysine and 8 arginine residues and 14 aspartic acid and 11 glutamic acid residues; hence it is slightly basic. Most noteworthy is the high content of S-containing amino acids: 10 methionine and 11 cysteine residues. The latter are present mainly in the second half of the molecule. Except for its essential role in transformation (possibly involving acetylation of histones amongst other things[36]), little is known as yet concerning the biological function or mechanism of action of t.

T starts at the same position on the genome as t[23,24] and both molecules are identical up to perhaps about the middle of t. Then, due to a crossover, the remaining part of T is coded by a region starting around map position 0.53 (see below) and terminating with the UAA codon at position 2,550 (Fig. 2). Such a mechanism whereby the synthesis of a polypeptide is directed by two non-contiguous segments on the genome is, of course, formally equivalent to the insertion of extraneous DNA into a structural gene[37,38]. The crossover in the t region cannot occur much before nucleotide 280 (Fig. 2) since otherwise the number of identical peptides in t and T could not be accounted for; on the other hand, the crossover cannot occur beyond nucleotide 380 since this region can be deleted in viable mutants which still make normal T but not t[25,26,34,35]. T can be coded again from nucleotide 672 onward, that is, a continuous reading frame is present until the termination codon at position 2,550. The deduced amino acid sequence is shown in Fig. 4. It is not known at which point in the nucleotide sequence translation actually picks up again, but we believe it to be not too far from the earliest possible position mentioned above. Indeed, the mutant dl-1001 lacks the *Hin*d fragments H and I and induces the synthesis of a shortened, T antigen-related protein with a molecular weight of 33,000 (ref. 18). Since these two *Hin*d fragments code for 20,400 molecular weight of protein, it seems likely that the deleted DNA is not a multiple of three nucleotides and, as a result, translation after the deletion must continue in another reading frame and stop soon after the *Hin*d B

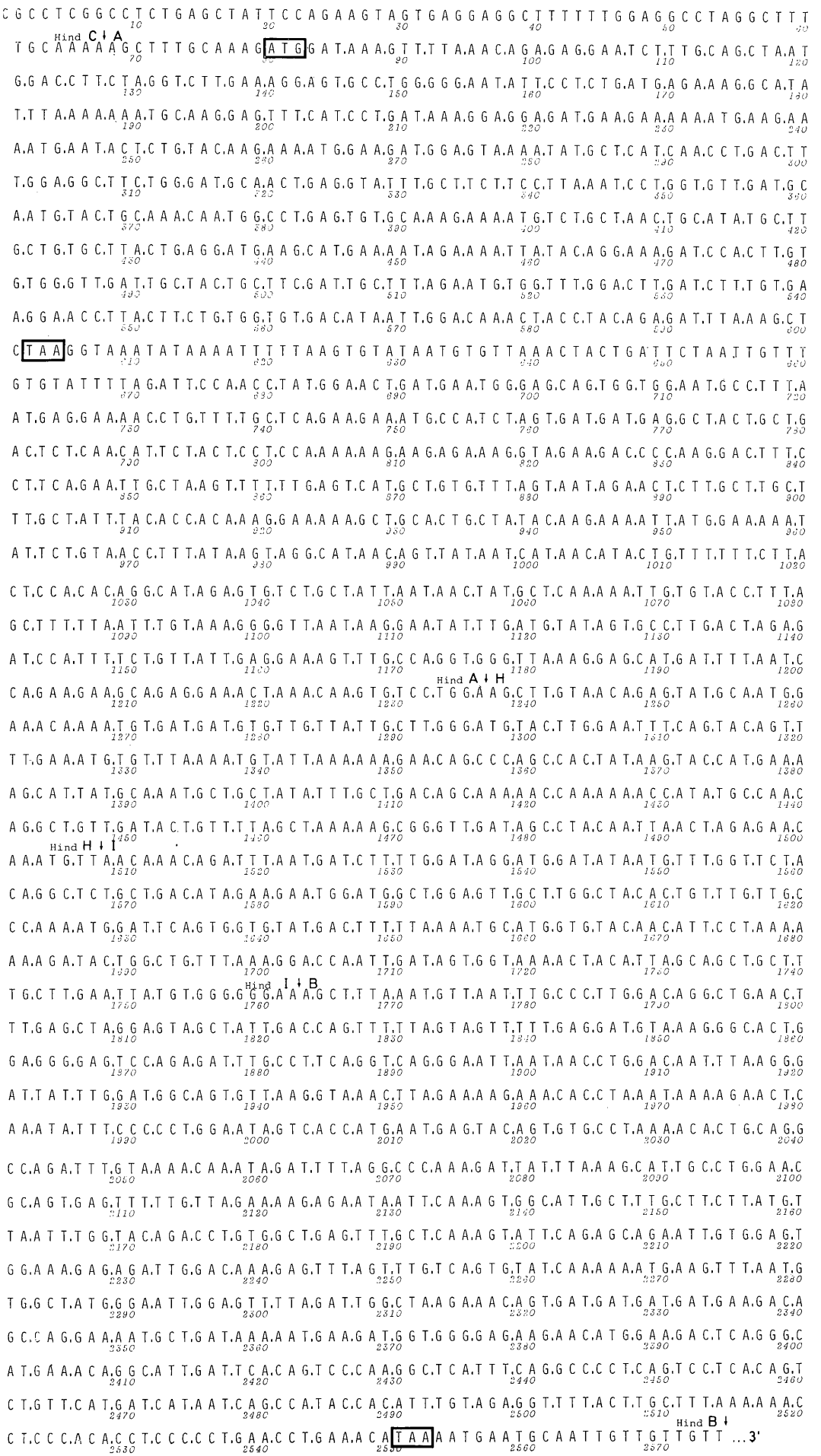

Fig. 2 Nucleotide sequence of the early region. Only the strand with the same polarity as the early mRNA is shown (in a 5′ to 3′ orientation). The sequence starts at the centre of the 27 base pair long palindrome at position 0.663 on the standard map (*BglI* site). The end is the junction between the *Hin*d fragments B and G. The different *Hin*d fragments are indicated. The initiation codon for t and T antigens is boxed, as are the respective termination codons. The reading frame for translation is indicated by dots (see text, and Figs 1 and 5 for the special role of the sequence between about nucleotide 380 and 672).

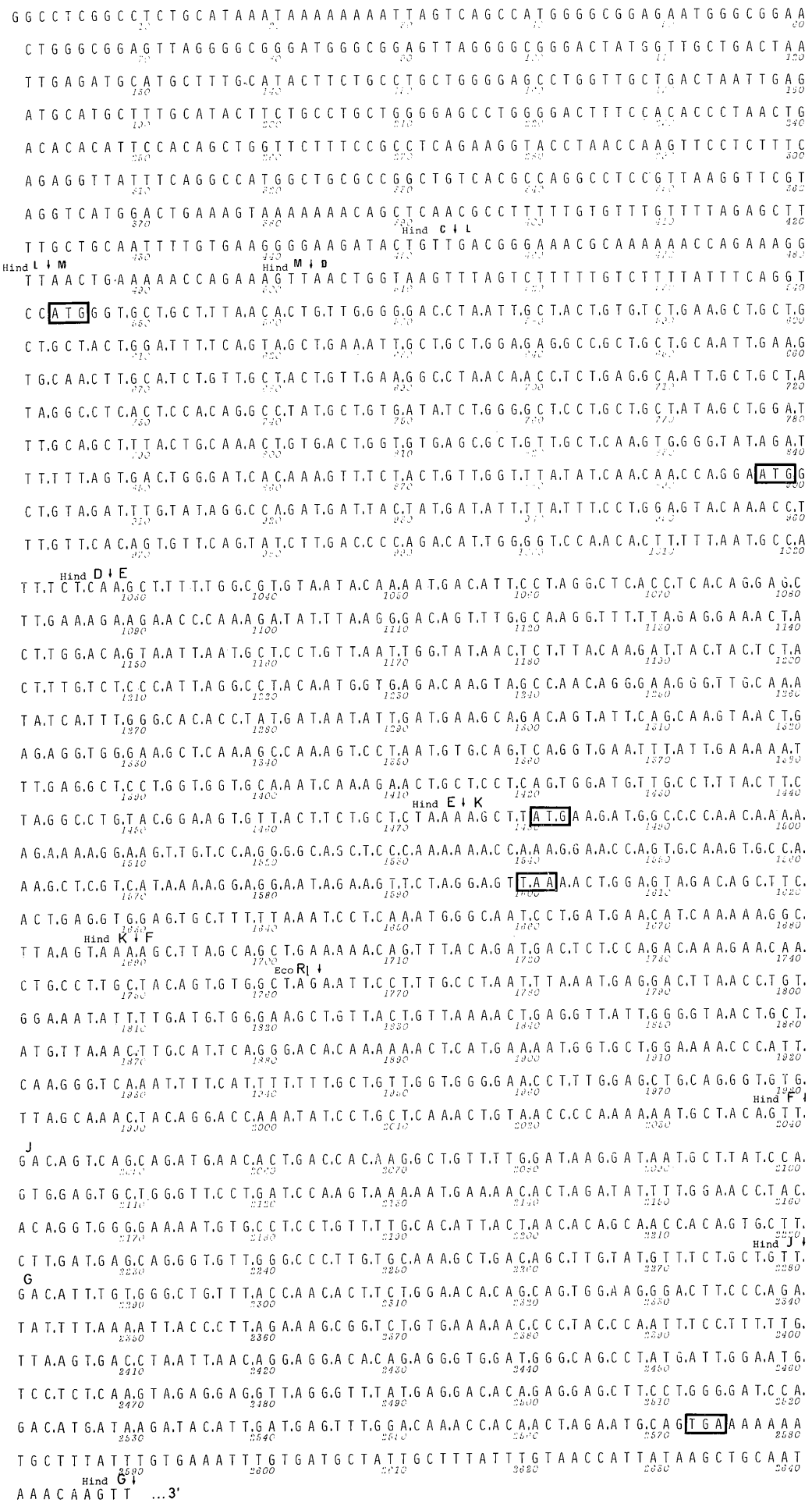

Fig. 3 Nucleotide sequence of the late region. Only the strand with the same polarity as late mRNA is shown (in a 5′ to 3′ orientation). The start and endpoint are the same as in Fig. 2. The cleavage sites of *Hin*dII and III and *Eco*RI restriction enzymes are shown (see also Fig. 1). The presumed initiation codons for VP_2, VP_3 and VP_1 are boxed, as well as the termination codons for VP_2/VP_3 and for VP_1. The reading frame in the translated region is indicated by dots (note that the last part of the VP_2/VP_3 gene overlaps the beginning of the VP_1 gene).

region is entered (this is evident from a close inspection of the sequence in Fig. 2). Hence, this explanation implies that about 300 amino acids of T must be coded for by the *Hin*d A fragment; about 100 amino acids may be shared with t and at most 188 may be coded for by the region before the junction to *Hin*d H (map position of 0.534–0.426). This would mean that translation must indeed resume rather soon after nucleotide 672.

Although the T antigen interacts specifically with SV40 DNA[39,40], it is not particularly basic. In certain areas S-containing amino acids are rather frequent[8,24] and, in general, many hydrophobic clusters are present. Particularly remarkable is the high proline content of the carboxy-terminal end: six out of ten residues are proline[22]. The latter aspect is confirmed by direct analysis of an adenovirus–SV40 fusion protein, in which the carboxyterminal sequence is derived from SV40 T antigen[41].

T is synthesised from a 19S mRNA[32,33]. As discussed above, the protein is coded for by two non-contiguous regions on the genome. In view of well-documented evidence from adenovirus mRNA and SV40 late mRNA[6,42–50], it is likely that the transposition or splicing-out of the inserted genetic information occurs at the mRNA level. Berk and Sharp[51] have provided direct evidence for an early SV40 messenger which corresponds approximately to the region 0.67–0.60 linked to 0.54–0.14 and is estimated to be 2,230 nucleotides long. It may be noted, however, that for the late SV40 messengers the splicing removes segments before the structural part of the gene, while in this early mRNA, the crossovers are in translated regions. We know that the segment deleted from late mRNA is present in (most of) the nuclear RNA, and hence that splicing in this case is a post-transcriptional event (our unpublished results). Whether this is also true for the early mRNA has not yet been established.

t is synthesised from an mRNA which sediments slightly faster than the mRNA coding for T (ref. 33). According to the interpretation discussed above, only about the first quarter of the RNA molecule would be translated and the remainder would constitute a link with the region beyond nucleotide 2,552 (Fig. 2) where there are presumably essential signals for the termination of transcription, processing, polyadenylation and/or transport (Fig. 1). According to recent evidence, even t mRNA is subject to splicing; a small segment, perhaps about 50 nucleotides, is removed from the region around 0.55–0.54 (ref. 51). It is noteworthy that this corresponds to a region which is unusually AT-rich: the 51 base pairs between nucleotides 591 and 641 are 82.4% AT and a middle segment of 18 base pairs is exclusively AT. Also the region following the common leader sequence of the two (or three) late SV40 mRNAs (nucleotide 509 to 535 in Fig. 3) is remarkably AT-rich. This feature could perhaps have a role in splicing.

The 5′-end of the early mRNA has been reported to map around position 0.67 (ref. 52), but the exact location of the mRNA start is not known. The 3′-end was found to extend to about position 0.16 (ref. 53). This corresponds almost exactly to the termination codon of VP_1 (nucleotide 2,574 in Fig. 3). The signal AAUAAA is present in all eukaryotic mRNAs as well as in encephalomyocarditis virus RNA, 10 to 20 nucleotides before the poly(A) tail[54–56]. It is present twice in SV40 early mRNA and its complement can be seen at nucleotide 2,618–2,613 and again at nucleotide 2,589–2,584 in Fig. 3. Detailed comparison of the viral, cytoplasmic mRNAs with the DNA sequence indicates that the 3′-terminal, untranslated regions of the early and late mRNAs overlap for a distance of about 80–100 nucleotides[22,53]. As was found for parts of the regions spliced out of the early and late transcripts, the untranslated segment between the termination codons of the T antigen gene and the VP_1 gene is rather AT-rich (74.5% compared with 59.5% for the total SV40 genome).

The late region

Late transcription starts concomitantly with or following SV40 DNA replication, which itself requires a functional early gene *A* product. No late transcription occurs when DNA replication is blocked (as in non-permissive cells or after the addition of appropriate inhibitors). The 5′-end of late mRNA maps around 0.72, but its precise position is unknown[13,14,52]; the heteropolymeric 3′-end of the cytoplasmic mRNA to which the poly(A) tail is added, corresponds almost exactly to the *Hin*d G–*Hin*d B junction (see below; refs 50, 57 and Fig. 1).

The late region is known to code for only three structural proteins, VP_1, VP_2 and VP_3. The corresponding genes have been approximately localised on the genome by mapping temperature-sensitive mutations and DNA deletions which affect a particular protein[16,17,25,31,58]. All tryptic peptides derived from VP_3 are also present in a fingerprint of VP_2[59,60] and since deletions at the beginning of the VP_2 gene do not affect the size of VP_3, it is generally believed that translation of the VP_3 gene is initiated inside the VP_2 gene and from there on is read in the same reading frame[31,60].

The nucleotide sequence of the entire late region is shown in Fig. 3. Segments have been described before in detail: *Hin*d C[61], *Hin*d L and *Hin*d M[62], *Hin*d D[63,64], *Hin*d E[65], *Hin*d K[7,66], *Hin*d F[67,68], *Hin*d J[69] and *Hin*d G[70] (see ref. 10 for parallel results obtained by S. Weissman and colleagues; *Hin*d M and a segment around the *Hin*d C/A junction have been independently sequenced by R. Wu and coworkers[71,72]).

The origin of DNA replication is around 0.67 map units; Subramanian *et al.*[11] pointed out that this region contains a remarkable palindromic sequence of 27 base pairs. This

MET-ASP-LYS-VAL-LEU-ASN-ARG-GLU-GLU-SER-LEU-GLN-LEU-MET-ASP-LEU-LEU-GLY-LEU-GLU- 20
ARG-SER-ALA-TRP-GLY-ASN-ILE-PRO-LEU-MET-ARG-LYS-ALA-TYR-LEU-LYS-LYS-CYS-LYS-GLU- 40
PHE-HIS-PRO-ASP-LYS-GLY-GLY-ASP-GLU-GLU-LYS-MET-LYS-LYS-MET-ASN-THR-LEU-TYR-LYS- 60
LYS-MET-GLU-ASP-GLY-VAL-LYS-TYR-ALA-HIS-GLN-PRO-ASP-PHE-GLY-GLY-PHE-TRP-ASP-ALA- 80
THR-GLU-VAL-PHE-ALA-SER-SER-LEU-ASN-PRO-GLY-VAL-ASP-ALA-MET-TYR-CYS-LYS-GLN-TRP-100
PRO-GLU-CYS-ALA-LYS-LYS-MET-SER-ALA-ASN-CYS-ILE-CYS-LEU-LEU-CYS-LEU-LEU-ARG-MET-120
LYS-HIS-GLU-ASN-ARG-LYS-LEU-TYR-ARG-LYS-ASP-PRO-LEU-VAL-TRP-VAL-ASP-CYS-TYR-CYS-140
PHE-ASP-CYS-PHE-ARG-MET-TRP-PHE-GLY-LEU-ASP-LEU-CYS-GLU-GLY-THR-LEU-LEU-LEU-TRP-160
CYS-ASP-ILE-ILE-GLY-GLN-THR-THR-TYR-ARG-ASP-LEU-LYS-LEU 174

ILE-PRO-THR-TYR-GLY-THR-ASP-GLU-TRP-GLU-GLN-TRP-TRP-ASN-ALA-PHE-ASN-GLU-GLU-ASN- 20
LEU-PHE-CYS-SER-GLU-GLU-MET-PRO-SER-SER-ASP-ASP-GLU-ALA-THR-ALA-ASP-SER-GLN-HIS- 40
SER-THR-PRO-PRO-LYS-LYS-LYS-ARG-LYS-VAL-GLU-ASP-PRO-LYS-ASP-PHE-PRO-SER-GLU-LEU- 60
LEU-SER-PHE-LEU-SER-HIS-ALA-VAL-PHE-SER-ASN-ARG-THR-LEU-ALA-CYS-PHE-ALA-ILE-TYR- 80
THR-THR-LYS-GLU-LYS-ALA-ALA-LEU-LEU-TYR-LYS-LYS-ILE-MET-GLU-LYS-TYR-SER-VAL-THR-100
PHE-ILE-SER-ARG-HIS-ASN-SER-TYR-ASN-HIS-ASN-ILE-LEU-PHE-PHE-LEU-THR-PRO-HIS-ARG-120
HIS-ARG-VAL-SER-ALA-ILE-ASN-ASN-TYR-ALA-GLN-LYS-LEU-CYS-THR-PHE-SER-PHE-LEU-ILE-140
CYS-LYS-GLY-VAL-ASN-LYS-GLU-TYR-LEU-MET-TYR-SER-ALA-LEU-THR-ARG-ASP-PRO-PHE-SER-160
VAL-ILE-GLU-GLU-SER-LEU-PRO-GLY-GLY-LEU-LYS-GLU-HIS-ASP-PHE-ASN-PRO-GLU-GLU-ALA-180
GLU-GLU-THR-LYS-GLN-VAL-SER-TRP-LYS-LEU-VAL-THR-GLU-TYR-ALA-MET-GLU-THR-LYS-CYS-200
ASP-ASP-VAL-LEU-LEU-LEU-LEU-GLY-MET-TYR-LEU-GLU-PHE-GLN-TYR-SER-PHE-GLU-MET-CYS-220
LEU-LYS-CYS-ILE-LYS-LYS-GLU-GLN-PRO-SER-HIS-TYR-LYS-TYR-HIS-GLU-LYS-HIS-TYR-ALA-240
ASN-ALA-ALA-ILE-PHE-ALA-ASP-SER-LYS-ASN-GLN-LYS-THR-ILE-CYS-GLN-GLN-ALA-VAL-ASP-260
THR-VAL-LEU-ALA-LYS-LYS-ARG-VAL-ASP-SER-LEU-GLN-LEU-THR-ARG-GLU-GLN-MET-LEU-THR-280
ASN-ARG-PHE-ASN-ASP-LEU-LEU-ASP-ARG-MET-ASP-ILE-MET-PHE-GLY-SER-THR-GLY-SER-ALA-300
ASP-ILE-GLU-GLU-TRP-MET-ALA-GLY-VAL-ALA-TRP-LEU-HIS-CYS-LEU-LEU-PRO-LYS-MET-ASP-320
SER-VAL-VAL-TYR-ASP-PHE-LEU-LYS-CYS-MET-VAL-TYR-ASN-ILE-PRO-LYS-LYS-ARG-TYR-TRP-340
LEU-PHE-LYS-GLY-PRO-ILE-ASP-SER-GLY-LYS-THR-THR-LEU-ALA-ALA-ALA-LEU-LEU-GLU-LEU-360
CYS-GLY-GLY-LYS-ALA-LEU-ASN-VAL-ASN-LEU-PRO-LEU-ASP-ARG-LEU-ASN-PHE-GLU-LEU-GLY-380
VAL-ALA-ILE-ASP-GLN-PHE-LEU-VAL-VAL-PHE-GLU-ASP-VAL-LYS-GLY-THR-GLY-GLY-GLU-SER-400
ARG-ASP-LEU-PRO-SER-GLY-GLN-GLY-ILE-ASN-ASN-LEU-ASP-ASN-LEU-ARG-ASP-TYR-LEU-ASP-420
GLY-SER-VAL-LYS-VAL-ASN-LEU-GLU-LYS-LYS-HIS-LEU-ASN-LYS-ARG-THR-GLN-ILE-PHE-PRO-440
PRO-GLY-ILE-VAL-THR-MET-ASN-GLU-TYR-SER-VAL-PRO-LYS-THR-LEU-GLN-ALA-ARG-PHE-VAL-460
LYS-GLN-ILE-ASP-PHE-ARG-PRO-LYS-ASP-TYR-LEU-LYS-HIS-CYS-LEU-GLU-ARG-SER-GLU-PHE-480
LEU-LEU-GLU-LYS-ARG-ILE-ILE-GLN-SER-GLY-ILE-ALA-LEU-LEU-LEU-MET-LEU-ILE-TRP-TYR-500
ARG-PRO-VAL-ALA-GLU-PHE-ALA-GLN-SER-ILE-GLN-SER-ARG-ILE-VAL-GLU-TRP-LYS-GLU-ARG-520
LEU-ASP-LYS-GLU-PHE-SER-LEU-SER-VAL-TYR-GLN-LYS-MET-LYS-PHE-ASN-VAL-ALA-MET-GLY-540
ILE-GLY-VAL-LEU-ASP-TRP-LEU-ARG-ASN-SER-ASP-ASP-ASP-ASP-GLU-ASP-SER-GLN-GLU-ASN-560
ALA-ASP-LYS-ASN-GLU-ASP-GLY-GLY-GLU-LYS-ASN-MET-GLU-ASP-SER-GLY-HIS-GLU-THR-GLY-580
ILE-ASP-SER-GLN-SER-GLN-GLY-SER-PHE-GLN-ALA-PRO-GLN-SER-SER-GLN-SER-VAL-HIS-ASP-600
HIS-ASN-GLN-PRO-TYR-HIS-ILE-CYS-ARG-GLY-PHE-THR-CYS-PHE-LYS-LYS-PRO-PRO-THR-PRO-620
PRO-PRO-GLU-PRO-GLU-THR 626

Fig. 4 Amino acid sequence of the early region of SV40 DNA. The sequence in the top part corresponds to t antigen (174 amino acids). T antigen starts at the same position and shares an estimated 100 amino acids with t; the lower part gives the maximal sequence which can follow the common N-terminal segment. For reasons discussed in the text, it is likely that indeed most of this coding potential is actually used.

segment is also rather well conserved in polyoma DNA (refs 73, 74; B. Griffin, personal communication). The first structural gene, which codes for VP_2, presumably starts at position 543 (Fig. 3). This has not been firmly established by correlation with protein sequence data but is supported by various arguments[63]. As discussed above, the first initiating codon in the early region of the genome is at position 80 (Fig. 2). Hence, in between these divergently directed structural genes there is a 622 base pair (542+79+1) DNA segment which contains information not only for initiation of DNA replication, but presumably also for early transcription, late transcription, for ribosome binding and initiation of translation, and for the processing of mRNAs, as discussed below (some of these biological signals may of course overlap each other or part of the structural genes). This DNA sequence contains many remarkable features, such as palindromes, long repeats, so-called 'true' palindromes, blocks with an exceptionally high AT-content which are moreover asymmetric (that is T residues in one strand and A residues in the other) and unusual alternations of AT-rich and GC-rich blocks[11,61-63,72,75]. The biological significance of these features is largely unknown, but at least some of these sequences can be deleted. In particular, a segment of about 90 nucleotides from map position 0.675 to 0.692 is redundant[76]. Other deletions in this area are still viable but result in a decreased plaque size and/or burst size[77,78].

We believe that the VP_2 gene starts at position 543 and the VP_3 gene at position 897 and that both end with a UAA at 1,599 (refs 63–66). Initiation elsewhere would either involve GUG as a start codon (which is unlikely) or result in an unrealistic polypeptide size. VP_2 contains 352 amino acids before processing and VP_3, 234 amino acids. The corresponding molecular weights of 38,533 and 26,967, respectively, are in reasonable agreement with direct size estimates based on the electrophoretic mobility of these proteins in SDS-polyacrylamide gels (see ref. 79 for example). The amino acid sequence has been published[7,63-66] and can be deduced from the information given in Fig. 3. It should be noted that this sequence refers to the primary translation product; that may be subsequently processed and is presumably blocked at the N-terminus[08]. The composition deduced corresponds moderately well with the directly determined values[81] and in particular shows the absence of cysteine residues in VP_3 (ref. 82). The aminoterminal sequence of VP_2 is exceptionally rich in alanine and is very hydrophobic; it contains no basic residue in the first 98 amino acids. Conceivably, it may be involved in an interaction with membranes. The carboxy-terminus of VP_2 and VP_3 is, however, unusually basic; 15 of the last 34 residues are either lysine or arginine. This basic tail may perhaps anchor the protein to the DNA[83,84]. It is of interest that the VP_2/VP_3 gene overlaps the VP_1 gene over a distance of 122 nucleotides[65]. Since the common segment is read in two different translation frames, it may be less malleable evolutionarily and this may be why it is precisely this region which cross-hybridises with the polyoma genome[85].

Lazarides, Files and Weber[80] determined the first seven amino acids of VP_1. With this information we could establish that the codon for the N-terminal alanine starts at position 1,489 (refs 7, 86). It is preceded by an AUG codon, but another AUG appears two codons before. Recent *in vitro* translation experiments suggest that the latter is used for initiation (A. Mellor, R. Hewick, and A. E. Smith, personal communication). This would mean that a Met-Lys-Met sequence is subsequently processed away. The VP_1 polypeptide is 361 amino acids long and its entire sequence has been published[6,87]. It has a molecular weight of 39,708 which is considerably less than the 42,000 to 47,000 estimated by its mobility in SDS gels (see ref. 79 for example). This discrepancy may possibly be related to the high proline content. The VP_1 gene is terminated by a single UGA signal starting at position 2,572 in *Hin*d G.

Once the lytic infection is fully under way, the amount of late virus-specific RNA in both the cytoplasm and the nucleus is 10 to 20-fold higher than the RNA derived from the early region. Two late poly(A)-containing mRNAs —19S and 16S—can readily be resolved, coding for VP_2 and VP_1 respectively. In the case of polyoma it is known that VP_3 is made on a separate 18S mRNA (ref. 88; B. Kamen, personal communication; T. Hunter, personal communication), and the same may be true for SV40. The bulk of the 16S mRNA hybridises to the restriction fragments *Hin*d K, F, J and G[89,90], which, as we have seen, contain the structural information for VP_1. Late 19S mRNA hybridises, in addition, to the *Hin*d fragments D and E, which together with *Hin*d K, contain the VP_2 gene[91]. Obviously, 19S mRNA (and the putative 18S mRNA) carries the information for making VP_1 but does not express it (Fig. 1). The initiating AUG for VP_1 is possibly too far away from the capped 5′-terminus, or is perhaps buried in a secondary structure; a hairpin model for this region has been proposed[65].

Zain *et al.*[57] have reported that the T_1-oligonucleotide (G)UUAACAACAACAAUUG, which corresponds to the junction of the fragments *Hin*d G–*Hin*d B, is no longer present in the late mRNAs, although all preceding T_1-oligonucleotides can be identified. We have confirmed this and found that the 3′-end of both late mRNAs is in the form (G)UU(AAC)1-3 poly(A). Counting from the U-A junction (the last nucleotide in Fig. 3), the 3′-untranslated sequence of 16S mRNA is 78 nucleotides (including the terminator UGA). It contains the sequence AAUAAA at a distance of 12–7 nucleotides from the U-A junction (this common signal present in 3′-untranslated regions was discussed above for the early mRNA). Under certain conditions, single-stranded SV40 DNA forms hairpins which can be fixed and visualised by electron microscopy[92,93]. One hairpin occurs around map position 0.17 and may correspond to a hairpin structure which can be proposed for the 3′-untranslated segment[70]. It is possible that in the nucleus the transcription runs somewhat further and is terminated in *Hin*d B[14,53]. This sequence would correspond to anti-early mRNA (see Fig. 2).

The 5′-end of SV40 mRNA is a 7-methyl G cap[94-96]. Further investigation has revealed that both 19S and 16S late mRNAs start with either 7-methyl Gppp 6,2′-dimethyl ApU or 7-methyl Gppp 6,2′dimethyl Ap 2′-methyl UpU, the latter being relatively enriched in the 16S mRNA[94]. It remains to be established whether both types of cap structure come from the same place on the genome and whether they are directly derived from the 5′-end of the primary transcript.

A number of laboratories have recently reported that the 5′-leader sequence of adenovirus mRNA is coded for by several non-contiguous regions which precede and are separated from the structural body of the gene by a considerable distance[42-46]. A similar observation for SV40 was made by hybridisation of purified late mRNA to *Eco*RI-digested DNA or to appropriate restriction fragments[47,48]. By comparison of the fingerprints of the late mRNAs with the known DNA sequence of the genome it was possible to conclude independently that these messengers are spliced[6,49,50]. Since the T_1-oligonucleotides between positions 344 and 501 are all present in both the late 19S fingerprint and in the 16S fingerprint, the common 5′-leader sequence must be at least 161 nucleotides long and the capped oligonucleotides must be derived from residue 308 or closer to the origin. The segment spliced out of the 16S RNA must start somewhere between residues 502 to 517 and terminate in the region of residues 1,431 to 1,458, that is, at least 21 residues before the presumptive AUG initiation

	U		C		A		G	
U	Phe	22, 1	Ser	16, 4, 6	Tyr	16, 9	Cys	3, 4	U
	Leu	16, 16			Ochre 1		Opal 1		C
					Amber		Trp 10		A
									G
C	Leu	10, 3, 6, 6	Pro	21, 8, 15	His	6, 5	Arg	2, 1	U
					Gln	26, 16			C
									A
									G
A	Ile	23, 8	Thr	25, 15, 16	Asn	19, 11	Ser	16, 6	U
	Met	2+11			Lys	28, 8	Arg	15, 17	C
									A
									G
G	Val	24, 1, 11, 18	Ala	45, 6, 12	Asp	18, 17	Gly	12, 7, 19, 15	U
					Glu	22, 19			C
									A
									G

Fig. 5 Codons used in the SV40 late genes (VP_1 and VP_2). The numbers refer to the frequency with which each triplet is used. The data for VP_1 and VP_2 are combined (a separate table for VP_1 has previously been published in ref. 6). The codeword table for the early genes is relatively similar.

codon of VP_1 (ref. 50). Since the T_1-oligonucleotides between residues 513 and 538 are missing from the late 19S fingerprint, a smaller fragment of 10 to 41 nucleotides must be spliced out of the 19S mRNA. The reason for this splicing and how it comes about is not known.

General characteristics and use of code words

SV40 DNA is 5,224 nucleotides long (2,574+2,649+1). At least 15.2% of this genetic material does not code directly for protein synthesis (this is a minimum percentage as the exact size of the region spliced out of the T gene is not known). A segment of 122 nucleotides (2.3% of the total genome) located in the late region, is read in two different frames and fulfils a dual function by coding for parts of three different proteins, VP_3/VP_2 and VP_1. A considerable portion of the genome directs the synthesis of two different proteins albeit read in the same frame (t and T antigens; VP_3 and VP_2).

Earlier studies had indicated[97], and confirmed[6,87,98], that in SV40 DNA, as in all vertebrate DNA, the dinucleotide sequence CG is very infrequent. It is of interest to note, however, that this drastic deficiency applies mainly to the translated regions and much less to the control region between the initiation codons of the early and late genes (map position 0.648 to 0.767, corresponding to nucleotide 80 in Fig. 2 and nucleotide 543 in Fig. 3). In this control region, CG occurs 18 times relative to an expected value of 36 to 37, whereas in the remainder of the genome it occurs only 9 times (expected 183). Our previous conclusions on the non-random utilisation of codons in SV40 DNA[6,7] become strengthened now that the total sequence is known. A codeword table for the late region is given in Fig. 5, but we stress that approximately the same relative distribution is valid for the early genes. Particularly striking is the deficiency of codons containing the sequence CG: the codons of the type NCG for serine, proline, threonine, or alanine are not used at all and only 3 out of 35 codons for arginine are of the CGN type (Fig. 5). Of the 97 codons ending with C, only 4 are followed by a code word beginning with G (1 occurs in the overlap region between the VP_2/VP_3 gene and the VP_1 gene). In the translated early region, the sequence CG occurs only 3 times.

Another striking feature of the codeword table is the low frequency of codons of the type NUC; in fact, the codon for isoleucine, AUC does not occur at all in the whole genome. Also statistically highly significant is the fact that codons of the type UUPu are preferred over CUN for leucine, and the codon AAA is preferred over AAG for lysine[6]. Although relative codon utilisation is remarkably consistent over the entire SV40 genome, it cannot reflect a general property of higher animal cells. Indeed, in other systems, highly non-random, but quite different patterns of selective codon utilisation are encountered (ref. 56 and references therein). Presumably, as in prokaryotes[99], various factors such as the tRNA composition, modulation effects, codon–anticodon interaction energy, and structural requirements of the mRNA influence the use made of the degeneracy of the code.

The elucidation of the total SV40 DNA sequence has revealed the complex and efficient organisation of the genes and has allowed us to deduce the almost complete amino acid sequence of all five SV40-coded proteins. Further study of the relationship between DNA structure and function may reveal any additional genetic information. The more difficult task remains, however, of explaining the complex biological functions and effects of this oncogenic virus in terms of the primary structure of its gene products.

We thank our colleagues who helped us at different stages of this project with advice, information and materials, and particularly Drs W. Gilbert, B. Hirt, K. Kleppe, A. Maxam, P. May and D. Nathans. Drs K. Danna, E. Soeda, R. Thijs and R. Yang contributed to earlier parts of this research project. We thank José Van der Heyden for assistance. Our research was supported by a grant from the Kankerfonds of the Algemene Spaar-en Lijfrentekas (ASLK) of Belgium. R.C. and G.H. hold fellowships from the NFWO.

Received 23 January; accepted 22 March 1978.

1. Fried, M. & Griffiin, B. *Adv. Cancer Res.* **24**, 67–113 (Academic Press, New York, 1977).
2. Fareed, G. C. & Davoli, D. *A. Rev. Biochem.* **46**, 471–522 (1977).
3. Kelly, T. J. & Nathans, D. *Adv. Virus Res.* **21**, 86–173 (Academic Press, New York, 1977).
4. Danna, K. & Nathans, D. *Proc. natn. Acad. Sci. U.S.A.* **68**, 2913–2917 (1971).
5. Yang, R., Danna, K., Van de Voorde, A. & Fiers, W. *Virology* **68**, 260–265 (1975).
6. Van de Voorde, A. *et al. INSERM-EMBO Workshop, July 1977* (eds May, P., Monier, R. & Weil, R.) **69**, 17–30 (1977).
7. Fiers, W. *et al. Proc. 10th FEBS Meeting, Paris*, 20–25 *July*, 1975, 17–33 (1975).
8. Volckaert, G., Contreras, R., Soeda, E., Van de Voorde, A. & Fiers, W. *J. molec. Biol.* **110**, 467–510 (1977).
9. Maxam, A. M. & Gilbert, W. *Proc. natn. Acad. Sci. U.S.A.* **74**, 560–564 (1977).
10. Reddy, V. B. *et al. Science* (in the press).
11. Subramanian, K. N., Dhar, R. & Weissman, S. M. *J. biol. Chem.* **252**, 355–367 (1977).
12. Weinberg, R. A., Warnaar, S. O. & Winocour, E. *J. Virol.* **10**, 193–201 (1972).
13. Khoury, G., Martin, M. A., Lee, T. N. H., Danna, K. J. & Nathans, D. *J. molec. Biol.* **78**, 377–389 (1973).
14. Sambrook, J., Sugden, B., Keller, W. & Sharp, P. A. *Proc. natn. Acad. Sci. U.S.A.* **70**, 3711–3715 (1973).
15. Ozanne, B., Sharp, P. A. & Sambrook, J. *J. Virol.* **12**, 90–98 (1973).
16. Lai, C. D. & Nathans, D. *Virology* **60**, 466–475 (1974).
17. Mantei, N., Boyer, H. W. & Goodman, H. M. *J. Virol.* **16**, 754–757 (1975).
18. Rundell, K., Collins, J. K., Tegtmeyer, P., Ozer, H. L., Lai, H. L. and Nathans, D. *J. Virol.* **21**, 636–646 (1977).
19. Tegtmeyer, T., Rundell, K. & Collins, J. K. *J. Virol.* **21**, 647–657 (1977).
20. Tegtmeyer, T., Schwartz, M., Collins J. K. & Rundell, K. J. *J. Virol.* **16**, 168–178 (1975).
21. Reed, S. I., Stark, G. R. & Alwine, J. C. *Proc. natn. Acad. Sci. U.S.A.* **73**, 3083–3087 (1976).
22. Van Heuverswyn, H., Van de Voorde, A. & Fiers, W. *Eurn. J. Biochem.* (in the press).
23. Paucha, E., Mellor, A., Harvey, R., Smith, A. E., Hewick, R. M. and Waterfield, M. D., *Proc. natn. Acad. Sci. U.S.A.* (in the press).
24. Volckaert, G., Van de Voorde, A. & Fiers, W. *Proc. natn. Acad. Sci. U.S.A.* (in the press).
25. Shenk, T. E., Carbon, J. & Berg, P. *J. Virol.* **18**, 664–671 (1976).
26. Crawford, L. V. *et al. Proc. natn. Acad. Sci. U.S.A.* **75**, 117–121 (1978).
27. Thimmappaya, B. & Weissman, S. M. *Cell* **11**, 837–843 (1977).
28. Staneloni, R. J., Fluck, M. M. & Benjamin, T. L. *Virology* **77**, 598–609 (1977).
29. Fluck, M. M., Staneloni, R. J. & Benjamin, T. L. *Virology* **77**, 610–624 (1977).
30. Feunteun, J., Sompayral, L., Fluck, M. & Benjamin, T. *Proc. natn. Acad. Sci. U.S.A.* **73**, 4169–4273 (1976).
31. Cole, C., Landers, T., Goff, S., Manteuil-Brutlag, S. & Berg, P. *J. Virol.* **24**, 277–294 (1977).
32. Prives, C., G lboa, E., Revel, M. & Winocour, E. *Proc. natn. Acad. Sci. U.S.A.* **74**, 457–461 (1977).
33. Paucha, E., Harvey, R., Smith, R. & Smith, A. E. *INSERM-EMBO Workshop, July 1977* (eds May, P., Monier, R. & Weil, R.) **69**, 189–198 (1977).
34. Sleigh, M., Topp, W., Hanich, R. & Sambrook, J. *Cell* (in the press).
35. Feunteun, J., Kress, M., Gardes, M. & Monier, R. *P.N.A.S.* (in the press).
36. Schaffhausen, B. S. & Benjamin, T. L. *Proc. natn. Acad. Sci. U.S.A.* **73**, 1092–1096 (1976).
37. Williamson, B. *Nature* **270**, 295–297 (1977).
38. Breathnach, R., Mandel, S. L. & Chambon, P. *Nature* **270**, 314–319 (1977).
39. Jessel, D., Hudson, J., Landau, T., Tenen, D. & Livingston, D. N. *Proc. natn. Acad. Sci. U.S.A.* **72**, 1960–1964 (1975).
40. Reed, S. I., Ferguson, J., Davis, R. W. & Stark, G. R. *Proc. natn. Acad. Sci. U.S.A.* **72**, 1605–1609 (1975)
41. Fey, G., Lewis, J. B. & Bothwell, A. (submitted).
42. Chow, L. T., Gelinas, R. E., Broker, T. R. & Roberts, R. J. *Cell* **12**, 1–8 (1977).
43. Klessig, D. F. *Cell* **12**, 9–22 (1977).
44. Dunn, A. R. & Hassell, J. A. *Cell* **12**, 23–36 (1977).
45. Berget, S. M., Moore, C. & Sharp, P. A. *Proc. natn. Acad. Sci. U.S.A.* **74**, 3171–3175 (1977).
46. Kitchimgnan, G. R., Lai, S. D. & Westphal, H. *Proc. natn. Acad. Sci. U.S.A.* **74**, 4392–4395 (1977).
47. Aloni, Y., Dhar, R., Laub, O., Horowitz, N. & Khoury, M. *Proc. natn. Acad. Sci. U.S.A.* **74**, 3686–3690 (1977).
48. HSU, M.-T. & Ford, J. *Proc. natn. Acad. Sci. U.S.A.* **74**, 4982–4985 (1977).

49. Celma, L., Dhar, R., Pan, J. & Weissman, S. M. *Nucleic Acids Res.* **4**, 2549–2559 (1977).
50. Haegeman, G. & Fiers, W. *Nature* **273**, 70–73 (1978).
51. Berk, A. J. & Sharp, P. A. *P.N.A.S.* (in the press).
52. Dhar, R., Subramanian, K. N., Pan, J. & Weissman, S. M., *J. biol., Chem.* **252**, 368–376 (1977).
53. Subramanian, K. N. *et al. Prog. Nucleic Acid Res. molec. Biol.* **19**, 157–164 (1976).
54. Proudfoot, H. J. & Brownlee, G. G. *Nature* **263**, 211–214 (1976).
55. Merregaert, J. *et al. Eur. J. Biochem.* **82**, 55–63 (1978).
56. Seeburg, P. H., Shine, J., Martial, J. A., Baxter, S. & Goodman, H. M. *Nature* **270**, 486–494 (1977).
57. Zain, B. S., Weissman, S. M., Dhar, R. & Pan, S. *Nucleic Acids Res.* **1**, 577–594 (1974).
58. Shenk, T. E., Rhodes, C., Rigby, P. W. J. & Berg, P. *Proc. natn. Acad. Sci. U.S.A.* **72**, 989–993 (1974).
59. Fey, G. & Hirt, B. *Cold Spring Harbour Symp. quant. Biol.* **39**, 235–241 (1974).
60. Rozenblatt, S., Mulligan, R., Gorecki, M., Roberts, S. & Rich, A. *Proc. natn. Acad. Sci. U.S.A.* **73**, 2743–2751 (1976).
61. Van Heuverswyn, H., Van de Voorde, A. & Fiers, W. *Nucleic Acids Res.* **4**, 1015–1024 (1977).
62. Ysebaert, M., Thys, F., Van de Voorde, A. & Fiers, W. *Nucleic Acids Res.* **3**, 3409–3421 (1976).
63. Ysebaert, M., Van Heuverswyn, H., Van de Voorde, A. & Fiers, W. *Eur. J. Biochem.* **85**, 195–204 (1978).
64. Ysebaert, M., Van de Voorde, A. & Fiers, W. (in the press).
65. Contreras, R., Rogiers, R., Van de Voorde, A. & Fiers, W. *Cell* **12**, 529–538 (1977).
66. Rogiers, R., Van de Voorde, A., Soeda, E. & Fiers, W. *Eur. J. Biochem.* **85**, 205–224 (1978).
67. Contreras, R., Volckaert, G., Thys, F., Van de Voorde, A. & Fiers, W. *Nucleic Acids Res.* **4**, 1001–1014 (1977).
68. Contreras, R., Van de Voorde, A. & Fiers, W. *Eur. J. Biochem.* (in the press).
69. Van Heuverswyn, H., Van de Voorde, A. & Fiers, W. (submitted).
70. Van Heuverswyn, H., Van de Voorde, A. & Fiers, W. *Eur. J. Biochem.* (in the press).
71. Tu, C-P., Roychoudhury, R. & Wu, R. *Fed. Proc.* **35**, 1595 (1976).
72. Jay, E., Roychoudhury, R. & Wu, R. *Biochem. biophys. Res. Commun.* **69**, 678–686 (1976).
73. Soeda, E., Kimura, G. & Miura, K-I. *Proc. natn. Acad. Sci. U.S.A.* **75**, 162–166 (1978).
74. Friedman, T. *INSERM-EMBO Workshop, July 1977* (eds May, P., Monier, R. & Weil, R.) **69**, 31–38 (1977).
75. Dhar, R., Subramanian, K. N., Pan, J. & Weissman, S. M. *Proc. natn. Acad. Sci. U.S.A.* **74**, 827–831 (1977).
76. Shenk, T. *J. molec. Biol.* **113**, 503–515 (1977).
77. Mertz, J. E. & Berg, P. *Proc. natn. Acad. Sci. U.S.A.* **71**, 4879–4883 (1974).
78. Carbon, J., Shenk, T. E. & Berg, P. *Proc. natn. Acad. Sci. U.S.A.* **72**, 1392–1396 (1975).
79. Estes, M. K., Huang, E.-S. & Pagano, J. S. *J. Virol.* **7**, 635–641 (1971).
80. Lazarides, E., Files, J. G. & Weber, K. *Virology* **60**, 584–587 (1974).
81. Greenaway, C. J. & Levine, D. *Biochem. biophys. Res. Commun.*, **52**, 1221–1227 (1973).
82. Pett, D. N., Estes, M. K. & Pagano, J. S. *J. Virol.* **15**, 379–385 (1975).
83. Huang, E.-S., Estes, M. K. & Pagano, J. S. *J. Virol.* **9**, 923–929 (1972).
84. Christiansen, G., Landers, T., Griffith, J. & Berg, P. *J. Virol.* **21**, 1079–1084 (1977).
85. Ferguson, J. & Davis, R. W. *J. molec. Biol.* **94**, 135–149 (1975).
86. Van de Voorde, A., Contreras, R., Rogiers, R. & Fiers, W. *Cell* **9**, 117–120 (1976).
87. Pan, J., Thimmappaya, B., Reddy, V. B. & Weissman, S. M. *Nucleic Acids Res.* **4**, 2539–2548 (1977).
88. Saddell, S. J. & Smith, A. E. (submitted).
89. Khoury, G., Carter, B. J., Ferdinand, F.-J., Howley, P. M., Brown, N. & Martin, M. A. *J. Virol.* **17**, 832–840 (1976).
90. May, E., Kopecka, H. & May, P. *Nucleic Acids Res.* **2**, 1995–2005 (1975).
91. May, E., Maizel, J. V. & Salzman, N. P. *Proc. natn. Acad. Sci. U.S.A.* **74**, 496–500 (1977).
92. Shen, C. K. J. & Hearst, J. E. *Proc. natn. Acad. Sci. U.S.A.* **74**, 1363–1367 (1977).
93. Hsu, M. T. & Jelinek, W. R. *Proc. natn. Acad. Sci. U.S.A.* **74**, 1631–1634 (1977).
94. Haegeman, G. & Fiers, W. *J. Virol.* **25**, 824–830 (1978).
95. Lavi, S. & Shatkin, A. *Proc. natn. Acad. Sci. U.S.A.* **72**, 2012–2016 (1975).
96. Groner, Y., Carmi, P. & Aloni, Y. *Nucleic Acids Res.* **4**, 3959–3968 (1977).
97. Morrison, J. M., Keir, H., Subak-Sharpe, H. & Crawford, L. V. *J. gen. Virol.* **1**, 101–108 (1967).
98. Fiers, W. *et al. Cold Spring Harbor Symp. quant. Biol.* **39**, 179–186 (1974).
99. Grosjean, H., Sankoff, D. Min Jou, W. & Fiers, W. (submitted).

Printed in Great Britain by Henry Ling Ltd., at the Dorset Press, Dorchester, Dorset