(Reprinted from Nature, Vol. 273, No. 5657, pp. 70–73, May 4, 1978)

Evidence for 'splicing' of SV40 16S mRNA

By the application of a wide variety of restriction enzymes and of recently developed DNA sequencing techniques, large segments of the SV40 genome have now been sequenced[1,2]. As a result, the expression of the viral genetic information can be studied in molecular detail. The major SV40-specific messenger RNA is a polyA-containing, cytoplasmic RNA which sediments at 16S and is only expressed after the onset of DNA replication[3]. *In vitro* translation studies have shown that it codes for VP1, the main structural protein of the virion (ref. 4 and A. Smith, personal communication). 16S mRNA hybridises mainly to the SV40 DNA *Hind* II+III restriction fragments K, F, J and G, which correspond to the region 0.95 to 0.17 on the standard map[5,6]. The localisation of these restriction fragments and of 16S mRNA is illustrated in Fig. 1.

We have shown previously that the codon corresponding to the N-terminal alanine of the VP1 protein is located in *Hind* II+III fragment K at 15–17 nucleotides clockwise from the *Hind* II+III E/K junction[7,8]. This pinpoints the beginning of the structural body of the VP1 gene. The termination codon of the VP1 gene, UGA, is located in *Hind* II+III fragment G at 78 nucleotides from the *Hind* II+III G/B junction[9] and is followed by an untranslated 3′-terminal sequence. The ribonuclease-T1 oligonucleotide to which the polyA tail is attached actually spans the *Hind* II+III G/B junction at map position 0.17 (ref. 10). The 16S mRNA is capped[1,11,12] and contains additional 6-methyl adenosine residues[13,14]. We have now characterised the 16S mRNA in further detail by comparing its fingerprint with the nucleotide sequence of the DNA. From the results we can conclude that the 5′-leader sequence of the RNA is at least 180 nucleotides long and is derived from a separate region on the genome, located at 0.72 to 0.76 map units. The splicing of RNA segments transcribed from non-contiguous regions on the DNA has recently also been observed in other laboratories both for adenovirus mRNAs and SV40 mRNA[15,16]. Our findings on the splicing of late SV40 mRNA were reported in brief previously[1].

SV40-specific late mRNA was labelled and isolated as follows: SV40-infected CV1 monkey cells (multiplicity of infection: 60 plaque-forming units (PFU) per cell) were labelled with ^{32}P-phosphate 24 h post-infection[11]. After a further incubation period of 24 h at 37 °C, the cells were lysed, the nuclei were removed and the total cytoplasmic RNA was recovered by several successive phenol–chloroform extractions. PolyA-containing RNA was isolated on an oligo dT–cellulose column and fractionated according to size by preparative electrophoresis on a 1.4% agarose gel. SV40-specific RNA was selected out by hybridisation of each gel fraction onto Sepharose-bound SV40 DNA[17]. Alternatively, the hybridisation step preceded the preparative electrophoresis (see Fig. 2). The main virus-specific species had a mobility corresponding to 16S mRNA. It was further characterised by digestion with RNase T1 and fractionation of the resulting products by a two-dimensional fingerprinting technique[18]. A typical pattern is shown in Fig. 3. All the oligonucleotides of chain length greater than four were eluted from the polyethyleneimine thin-layer plate and further identified by digestion with pancreatic RNase[19]. The resulting (partial) sequence of these T1-oligonucleotides was then compared with the actual DNA sequence of the late region of the SV40 genome.

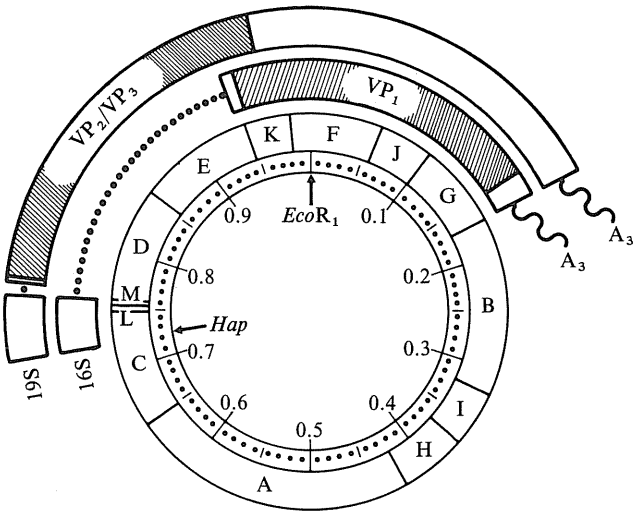

Fig. 1 *Hind* II+III restriction cleavage map of the SV40 genome and the localisation of the late 19S and 16S mRNA species. The single *Eco* RI restriction cleavage site is taken as reference point zero for the construction of the physical map. *Hind* II+III fragments are indicated by letters A through M. Striped blocks indicate the coding regions of the mRNAs, while clear blocks indicate the untranslated regions (that is, the 5′-leader and 3′-terminal segments); the wavy lines represent the 3′-terminal polyA tails. Dotted lines show the genome regions which are spliced away in the mRNA. This paper deals with the characterisation of the 16S mRNA, which codes for the major structural protein VP1 (the detailed characterisation of the late 19S mRNA will be documented elsewhere).

All the T1-oligonucleotides predicted from the DNA sequence of the *Hind* II+III fragments K, F, J and G were indeed identified in the fingerprint (Table 1), with the exception of the oligonucleotide UUAACAACAACAAUUG, which spans the *Hind* II+III G/B junction. We have characterised a part of the aforementioned oligonucleotide as being linked to the polyA tail; it was present in a smeared spot at the origin which was analysed as UU(AAC)$_{1-3}$polyA. This location of the 3′ end of the 16S mRNA on the SV40 DNA map confirms the assignment by Zain *et al.*[10].

We then attempted to pinpoint the 5′ end of the 16S mRNA. In the fingerprint (Fig. 3) only one unique (that is, present once per chain) T1-oligonucleotide, UUACUUCUG (spot No. 521), was clearly derived from the *Hind* II+III fragment E (ref. 20), while the other large, preceding ones were definitely missing. Their absence was not due to artefacts in the separation systems because these same oligonucleotides were clearly present in ribonuclease-T1 fingerprints of the larger, late 19S mRNA (manuscript in preparation). These data led to the conclusion that sequence representation in the 16S mRNA extends at most 40 nucleotides before the start codon of the VP1 protein (Fig. 4). However, it was not possible to locate the sequence AUU, which is directly linked to the cap at the 5′ terminus[1,11], into this preceding segment. Moreover, after correlation of all the potential G-terminal DNA-oligonucleotides from the *Hind* II+III fragments E, K, F, J and G with actual T1 products of the 16S mRNA, a set of discrete T1-oligonucleotides remained unaccounted for,

Table 1 T1-oligonucleotide content of SV40 16S mRNA

Fingerprint spot no.*	C.p.m./ nucleotide†	Calculated molarity‡	*Hind* II+III fragment§
004	374	1.2	K
013¶	622 or 2×311	1.0×2	L, J
014	287	0.9	F
016	220	0.7	F
023¶	990 or 4×248	0.8×4	K, K, G, G
024	316	1.0	G
026	209	0.7	M
034	274	0.9	J
037	235	0.8	L
044	276	0.9	J
057	230	0.7	K
104¶	927 or 3×309	1.0×3	K, F, J
112+103¶	1628 or 4×407	1.3×4	G, K, J, G
107	252	0.8	G
121¶	692 or 3×231	0.7×3	K, C, C
122¶	917 or 3×306	1.0×3	F, F, G
123¶	703 or 2×351	1.1×2	J, J
124	266	0.9	F
126	333	1.1	J
127+118¶	302	1.0	G, C
128	237	0.8	K
132	223	0.7	C
135	301	1.0	F
140	272	0.9	C
145	325	1.0	J
158	194	0.6	G
169	222	0.7	K
205	306	1.0	J
215	370	1.2	K
222	304	1.0	F
224+233¶	263	0.8	E–K, F
232	348	1.1	K
240	306	1.0	J
242	521 or 2×261	0.8×2	F, G
248+247 / 2410 ¶	239	0.8	F, F, F
303+312¶	2337 or 5×467	1.5×5	F, G, L–M, M–D, G
313	368	1.2	G
314+322¶	424	1.4	K, F, F
330	484	1.6	G
331	394	1.3	K
332	353	1.1	J
333	315	1.0	F
336+335¶	196	0.6	G, F
342	324	1.0	G
357	194	0.6	J
401+410¶	2188 or 5×438	1.4×5	F, G, C, J, C
412	324	1.0	C
413	283	0.9	F
420	581	1.9	J
426	224	0.7	G
502	364	1.2	J
504+522¶	465	1.5	F, F
520	303	1.0	C
521	309	1.0	E
525	297	1.0	F
611	946 or 2×473	1.5	G, G
654	281	0.9	G
846	269	0.9	K
878	218	0.7	G
938	310	1.0	G
1224**	297	1.0	F

* Numbers refer to the spots indicated in Fig. 3*b*. The code refers to the base composition, (see Fig. 3 and ref. 33). All spots corresponding to digestion products with chain length greater than four were cut out and counted by liquid scintillation.

† The c.p.m. per nucleotide is obtained by dividing the measured radioactivity of a spot (after subtraction of background radioactivity) by the (sum of) chain length(s) of the corresponding oligonucleotide(s) present in that spot.

‡ The molarity is calculated as the ratio between the effective c.p.m. per nucleotide for a given spot and the standard c.p.m. per nucleotide; the latter is the average of the c.p.m. per nucleotide values for all the listed oligonucleotides ($n>4$) present in the *Hind* II+III fragments E, K, F, J and G.

§ Each oligonucleotide present in the RNA could be unambiguously assigned to a *Hind* II+III fragment from the late region of the genome on the basis of the known DNA sequences (see Fig. 1); a dash connecting two fragments indicates that the oligonucleotide concerned overlaps the two *Hind* II+III fragments.

¶ Double-digestion analysis with pancreatic RNase revealed that several individual components or sequence isomers were present.

** This spot appears weaker in Fig. 3*a* because the radioactivity is spread over a larger area.

although they were also present in about molar amounts in the fingerprints of each SV40 16S mRNA preparation (Table 1). They could be correlated unambiguously, however, with a DNA segment more counterclockwise in the late region of the genome and spanning the *Hind* II+III fragments C (right end), L and M (map coordinates 0.72 to 0.76; see Figs 1 and 4). We also characterised the SV40 16S mRNA by digestion with pancreatic RNase and fingerprinting[18]. All the potential pyrimidine-terminal products were again identified, those from both the coding part of the gene and the 3′-untranslated segment as well as those from the approximately 40 nucleotide segment before the initiation codon and those from the 0.72–0.76 leader sequence region. These results further strengthened our conclusions on the sequence representation in the major, 16S mRNA.

No large oligonucleotides from *Hind* II+III fragment D were detected in 16S RNA, while all the products (T1 and pancreatic oligonucleotides) from *Hind* II+III fragment M were clearly present; hence we believe that the putative 3′-terminal 'splicing' site of the leader segment is located close to the *Hind* II+III restriction site between fragments M and D. Correct positioning of the 5′ end, and hence of the whole mRNA, was not yet possible, mainly because the sequence in that area lacks discriminative T1 and/or pancreatic oligonucleotides (Fig. 4). The T1 product UUAUUUCAG, preceding the *Hap* (or *Hpa* II) site by 14 nucleotides (ref. 21 and H. Van Heuverswyn, personal communication), was not present in the 16S RNA. However, as the aforementioned AUU sequence occurs in it, the oligonucleotide UUAUUUCAG may possibly correspond to the region where cap formation takes place[11]. Alternatively, the 5′-leader sequence may be generated by two splicing events and the 5′-terminal cap oligonucleotide would then be derived from a region still more counterclockwise on the genome. It is remarkable, however, that a number of viable

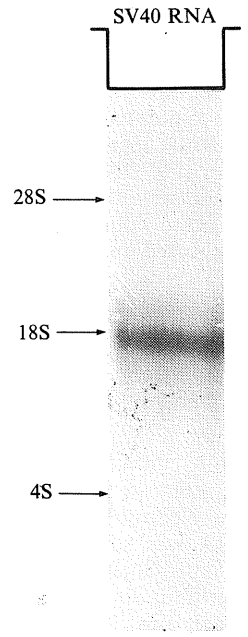

Fig. 2 Purification of ^{32}P-labelled late SV40 mRNA on 1.4% agarose gel. ^{32}P-labelled, polyA-containing RNA was extracted from SV40-infected CV1 cells and hybridised onto Sepharose-bound SV40 DNA. The virus-specific material was eluted, briefly heated and directly loaded onto a 1.4% agarose gel for electrophoresis in 20 mM Tris-citrate, 1 mM EDTA, *p*H 7.5, at 50 V for 3 h at 0 °C. The ^{32}P-labelled RNA was localised by autoradiography. The arrows indicate the position of the rRNA markers run in parallel lanes (note that rRNAs are relatively less retarded in agarose gel electrophoresis such that 18S rRNA moves to about the same position as 16S mRNA)[3].

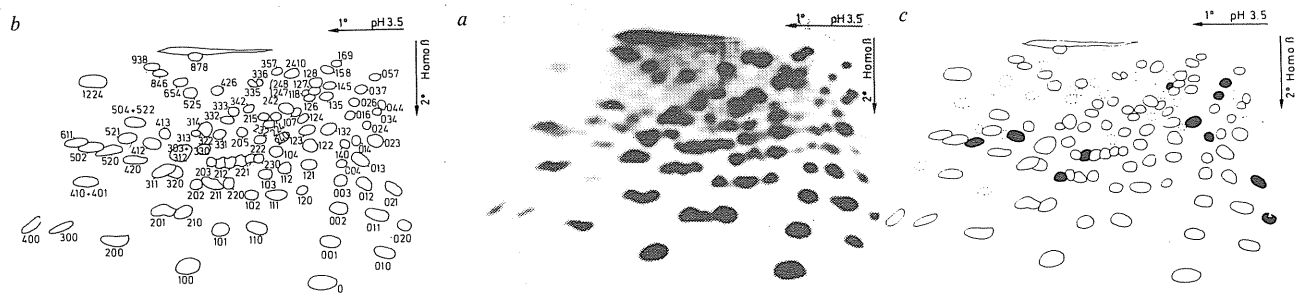

Fig. 3 Two-dimensional fractionation of a T1 digest of 16S SV40 mRNA. The RNA was prepared and isolated as briefly described in the text and digested with 3 units of ribonuclease T1 at 37 °C for 15 min. The hydrolysate was applied directly onto a pre-wetted cellulose acetate strip for high voltage electrophoresis at *p*H 3.5 in the presence of 5 M urea. The material was then transferred by reversed blotting onto a polyethyleneimine plate and developed in the second dimension with homomixture β at 60 °C (ref. 18). As the first dimension separates according to base composition and the second dimension according to chain length, the composition of nearly every spot can be deduced from its position. *a*, Autoradiograph of an original fingerprint. *b*, Schematic diagram identifying the oligonucleotides present. The latter are indicated by a standard nomenclature which is based on the nucleotide composition of the spots, as explained by De Wachter *et al.*[33]; the numerals refer to the U, C and A content respectively. *c*, Schematic diagram (corresponding to *b*) distinguishing between the digestion products derived from the body of the 16S mRNA (white spots) and the oligonucleotides which are unique to, and hence specific for, the 5′-terminal leader segment of the mRNA (black spots). Additional unique oligonucleotides, present only in the fingerprint of late 19S mRNA, are indicated by dashed circles. They are hardly visible on the original autoradiograph of the 16S fingerprint (*a*) and represent less than 10% of the radioactivity expected for a molar ratio.

deletion mutants have been reported which lack up to 50% and perhaps more of the 5′-leader fragment here characterised[22,23] (work is in progress in our laboratory to delineate more precisely the deletion of such mutants by actual DNA sequencing).

In conclusion, we have shown that the major SV40 mRNA is coded for by at least two separate regions on the genome. The 5′-leader sequence is about 180 nucleotides in length and is transcribed from the region 0.72 to 0.76 on the standard map. The body of the mRNA starts at most 40 nucleotides before the initiation codon for the VP1 protein and contains the whole structural VP1 gene followed by the untranslated 3′-terminal end; it thus extends clockwise from coordinates 0.94 to 0.17 (Fig. 1). There is no evidence that the two parts of the RNA molecule are joined by anything other than a normal phosphodiester linkage. It is at present not known whether the splicing occurs during transcription (a 'gliding' RNA polymerase) or is the result of a subsequent excision–ligation event. The former hypothesis is unlikely in view of the results with adenovirus-infected cells, where a full transcript appears to be the precursor to the late mRNAs[24–26], and considering that also in SV40 the 16S mRNA may be derived from a larger precursor[27]. Splicing by the latter mechanism may involve a base-paired stem, where the cutting-ligating enzyme(s) would act; however, we could not yet identify extensive base-pairing in the relevant region. The concept of splicing in the formation of eukaryotic mRNA may perhaps partly explain the complex processing events undergone by heterogeneous nuclear RNA and may form the basis of a new type of control for the expression of the genetic information.

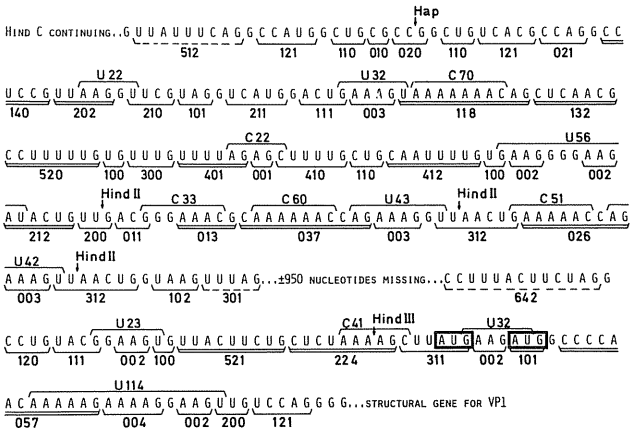

Fig. 4 Partial nucleotide sequence of the untranslated 5′-terminal leader sequence of SV40 16S mRNA. The RNA sequence has been deduced from the corresponding DNA sequences of the different *Hin*d II+III restriction fragments of SV40 DNA (reading in clockwise direction): the last 149 nucleotides of *Hin*d II+III fragment C (ref. 34 and H. Van Heuverswyn, personal communication), *Hin*d II+III fragments L and M[35], the first 15 nucleotides of *Hin*d II+III fragment D[36], the last 44 nucleotides of *Hin*d II+III fragment E[20], and the first 48 nucleotides of *Hin*d II+III fragment K[37]. The numbers of the T1 products refer to their base composition (see legend to Fig. 3, Table 1 and ref. 33). Doubly underlined oligonucleotides are unique and therefore discriminative products certainly present in the fingerprint (Table 1); singly underlined oligonucleotides are also observed but their presence is not discriminative. Dotted lines denote characteristic oligonucleotides which are predicted from the DNA sequencing data but which were not detected in the mRNA fingerprint. An analogous system of numbering is used for the pancreatic RNase oligonucleotides[38]; characteristic products certainly present in the fingerprint are indicated by lines above the sequence. Initiation of translation of the VP1 gene can start at either of the two AUG triplets shown in boxes[8], but recent *in vitro* translation experiments favour the 5′-proximal AUG triplet (A. Mellor, R. Hewick, M. Waterfield and A. Smith, personal communication).

After this manuscript was submitted results were published which were obtained by nucleic acid sequencing[28], nucleic acid hybridisation[29,30] and electron microscopy[31] and which lead to the same general conclusions as reported here. More detailed information, however, is revealed by our characterisation of the mRNA nucleotide sequence, as summarised in Fig. 4. Also, the complete nucleotide sequence of SV40 DNA is now known[21,32].

Guy Haegeman holds a fellowship from the Nationaal Fonds voor Wetenschappelijk Onderzoek (Belgium). This research was supported by grants from the Algemene Spaar- en Lijfrentekas (A.S.L.K.) and from the Fonds voor Geneeskundig Wetenschappelijk Onderzoek (F.G.W.O.) of Belgium.

Guy Haegeman
Walter Fiers

Laboratory of Molecular Biology,
Ledeganckstraat, 35,
B-9000 Ghent, Belgium

Received 22 August 1977; accepted 13 March 1978.

1. Van de Voorde, A. *et al.* in *Early Proteins of Oncogenic DNA Viruses* (eds May, P., Monier, M. & Weil, R.) 17–30 (INSERM, Paris, 1977).
2. Dhar, R., Subramanian, K., Pan, J. & Weissman, S. *Proc. natn. Acad. Sci. U.S.A.* **74**, 827–831 (1977).
3. Weinberg, R., Ben-Ishai, Z. & Newbold, J. *J. Virol.* **13**, 1263–1273 (1974).

4. Prives, C., Aviv, H., Gilboa, E., Revel, M. & Winocour, E. *Cold Spring Harb. Symp. quant. Biol.* **39**, 309–316 (1974).
5. May, E., Kopecka, H. & May, P. *Nucleic Acids Res.* **2**, 1995–2005 (1975).
6. Khoury, G., Carter, B., Ferdinand, F., Howley, P., Brown, M. & Martin, M. *J. Virol.* **17**, 832–840 (1976).
7. Fiers, W. *et al. FEBS Symp.* **39**, 17–33 (1975).
8. Van de Voorde, A., Contreras, R., Rogiers, R. & Fiers, W. *Cell* **9**, 117–120 (1976).
9. Van Heuverswyn, H., Van de Voorde, A. & Fiers, W. *Eur. J. Biochem.* (in the press).
10. Zain, B., Weissman, S., Dhar, R. & Pan, J. *Nucleic Acids Res.* **1**, 577–593 (1974).
11. Haegeman, G. & Fiers, W. *J. Virol.* (in the press).
12. Groner, Y., Carmi, P. & Aloni, Y. *Nucleic Acids Res.* **4**, 3959–3968 (1977).
13. Lavi, S. & Shatkin, A. *Proc. natn. Acad. Sci. U.S.A.* **72**, 2012–2016 (1975).
14. Aloni, Y. *FEBS Lett.* **54**, 363–368 (1975).
15. Berget, S., Moore, C. & Sharp, P. *Proc. natn. Acad. Sci. U.S.A.* **74**, 3171–3175 (1977).
16. Klessig, D. *Cell* **12**, 9–21 (1977).
17. Gilboa, E., Prives, C. & Aviv, H. *Biochemistry* **14**, 4215–4220 (1975).
18. Volckaert, G., Min Jou, W. & Fiers, W. *Analyt. Biochem.* **72**, 433–446 (1976).
19. Volckaert, G. & Fiers, W. *Analyt. Biochem.* **83**, 228–239 (1977).
20. Contreras, R., Rogiers, R., Van de Voorde, A. & Fiers, W. *Cell* **12**, 529–538 (1977).
21. Fiers, W. *et al. Nature* (in the press).
22. Mertz, J. & Berg, P. *Proc. natn. Acad. Sci. U.S.A.* **71**, 4879–4883 (1974).
23. Shenk, T., Carbon, J. & Berg, P. *J. Virol.* **18**, 664–671 (1976).
24. Wall, R., Philipson, L. & Darnell, J. *Virology* **50**, 27–34 (1972).
25. Bachenheimer, S. & Darnell, J. *Proc. natn. Acad. Sci. U.S.A.* **72**, 4445–4449 (1975).
26. McGuire, P., Swart, C. & Hodge, L. *Proc. natn. Acad. Sci. U.S.A.* **69**, 1578–1582 (1972).
27. Aloni, Y., Shani, M. & Reuveni, Y. *Proc. natn. Acad. Sci. U.S.A.* **72**, 2587–2591 (1975).
28. Celma, M., Dhar, R., Pan, J. & Weissman, S. *Nucleic Acids Res.* **4**, 2549–2559 (1977).
29. Aloni, Y., Dhar, R., Laub, O., Horowitz, M. & Khoury, G. *Proc. natn. Acad. Sci. U.S.A.* **74**, 3686–3690 (1977).
30. Lavi, S. & Groner, Y. *Proc. natn. Acad. Sci. U.S.A.* **74**, 5323–5327 (1977).
31. Hsu, M. & Ford, J. *Proc. natn. Acad. Sci. U.S.A.* **74**, 4982–4985 (1977).
32. Reddy, V., Thimmappaya, B., Dhar, R., Subramanian, K., Zain, B., Pan, J., Celma, C. & Weissman, S. *Science* (in the press).
33. De Wachter, R., Merregaert, J., Vandenberghe, A., Contreras, R. & Fiers, W. *Eur. J. Biochem.* **22**, 400–414 (1971).
34. Van Heuverswyn, H., Van de Voorde, A. & Fiers, W. *Nucleic Acids Res.* **4**, 1015–1024 (1977).
35. Ysebaert, M., Thys, F., Van de Voorde, A. & Fiers, W. *Nucleic Acids Res.* **3**, 3409–3421 (1976).
36. Ysebaert, M., Van Heuverswyn, H., Van de Voorde, A. & Fiers, W. *Eur. J. Biochem.* (in the press).
37. Rogiers, R., Van de Voorde, A., Soeda, E. & Fiers, W. *Eur. J. Biochem.* (in the press).
38. Vandenberghe, A., Min Jou, W. & Fiers, W. *Proc. natn. Acad. Sci. U.S.A.* **72**, 2559–2562 (1975).

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