Proc. Natl. Acad. Sci. USA
Vol. 78, No. 8, pp. 4887–4891, August 1981
Biochemistry

Molecular cloning of poliovirus cDNA and determination of the complete nucleotide sequence of the viral genome

(poliovirus RNA/translation frame/primer extension/reverse transcriptase)

VINCENT R. RACANIELLO AND DAVID BALTIMORE

Center for Cancer Research and Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139

Contributed by David Baltimore, May 11, 1981

ABSTRACT **The complete 7410 nucleotide sequence of the poliovirus type I genome was obtained from cloned cDNA. Double-stranded poliovirus cDNA was synthesized and inserted into the *Pst* I site of plasmid pBR322, and three clones were derived that together provided DNA copies of the entire poliovirus genome. Two of the clones contained inserts of 2.5 and 6.5 kilobases and represented all but the 5′ 115 bases of poliovirus RNA. A third clone was generated from primer-extended DNA and contained sequences from the 5′ end of the viral RNA. An open reading frame that was identified in the nucleotide sequence starting 743 bases from the 5′ end of the RNA and extending to a termination codon 71 bases from the 3′ end contained known poliovirus polypeptide sequence.**

Poliovirus, a member of the picornavirus group, has been studied for the last 25 years as the prototype for positive-strand RNA animal viruses (1, 2). Poliovirions contain a single strand of infectious RNA with a molecular weight of approximately 2.6×10^6 (3). This size is equivalent to 7.5 kilobases (kb), or sufficient nucleic acid to encode 2500 amino acids.

A variety of experimental results has suggested that the poliovirus genome is translated into a single polypeptide from which the functional viral proteins are derived by proteolysis (4–9). This hypothesis predicts a single, long, open reading frame in the poliovirus genome. To identify this open frame and to obtain precise information about the structure of poliovirus RNA, we have determined the nucleotide sequence of the viral genome. The value of having a DNA representation of the genome, coupled with the precision of sequence determination made possible when complementary strands of nucleic acid are independently analyzed, led us to derive molecular clones representing the viral RNA. We report here the complete 7410-nucleotide sequence of the molecularly cloned poliovirus genome in which an open reading frame 6597 nucleotides long is identified, showing that the genome can be copied into a single translation product.

MATERIALS AND METHODS

Synthesis and Cloning of Double-Stranded Poliovirus cDNA. Poliovirus double-stranded cDNA was synthesized from purified poliovirus RNA (10), using published conditions (11) with slight modifications. Double-stranded molecules were inserted into the *Pst* I site of plasmid pBR322 by using oligo(dG)·oligo(dC) tailing methods, as described (11). Tetracycline-resistant clones were screened for inserts by colony hybridization, using a calf thymus DNA-primed poliovirus cDNA probe (12, 13).

Primer Extension and Cloning of Primer Extended Material. A restriction fragment from clone pVR103 from bases 149–220 was prepared that was 5′-end-labeled at the *Bam*HI site only (position 220). This fragment was hybridized to poliovirus RNA and extended with reverse transcriptase (RNA-dependent DNA polymerase) by using published procedures (14, 15). The sequences of the extension products were determined chemically (16). For cloning, the extended fragment was purified by gel electrophoresis and tailed with oligo(dC). The fragment was then made double-stranded with the Klenow fragment of *Escherichia coli* DNA polymerase I in the presence of $(dG)_{12-18}$, tailed again with oligo(dC), and inserted into the *Pst* I site of pBR322. One clone, pVR105, contained sequences from the *Bam*HI site (position 220) up to and including the first base of the poliovirus genome.

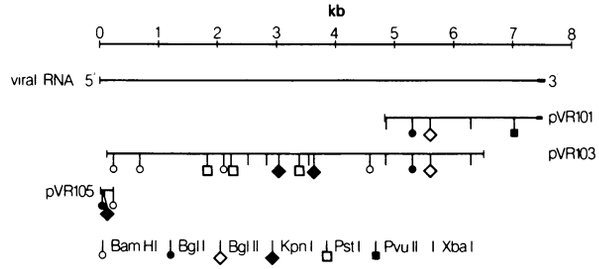

FIG. 1. Restriction map of poliovirus cDNA clones. The viral RNA is shown with the 3′ poly(A) (thicker line) at the right. Positions of the inserts from clones pVR101, pVR103, and pVR105 are shown beneath the viral RNA. Sites for rarely cutting restriction enzymes are indicated with symbols. The oligo(dG)·oligo(dC) tails at the ends of each insert are not shown.

RESULTS

Cloning of Poliovirus Double-Stranded cDNA. Three poliovirus-specific clones were characterized in detail. Plasmid pVR101 contained a poliovirus-specific insert 2.5 kb in length. Sequence analysis of this insert indicated that it contained about 40 As at one end followed by poliovirus-specific sequences (17). Fig. 1 shows the position of this insert at the 3′ end of the viral RNA. Plasmid pVR103 contained the longest poliovirus-specific insert, 6.5 kb. Restriction mapping with *Bgl* I, *Bgl* II, *Pvu* II, and *Xba* I revealed that pVR103 overlapped the 5′ end of pVR101 (Fig. 1). The total unique sequence cloned, approximately 7200 bases, was very close to the estimated 7500 base length of the poliovirus genome (3).

Because the double-stranded cDNA used to generate these clones included a "snap-back" step for second-strand synthesis, it was assumed that sequences at the very 5′ end of the viral RNA would not be found even in the longest clones. To deter-

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U. S. C. §1734 solely to indicate this fact.

Abbreviations: kb, kilobase(s); NCVP, noncapsid virion protein; VPg, genome-linked virion protein.

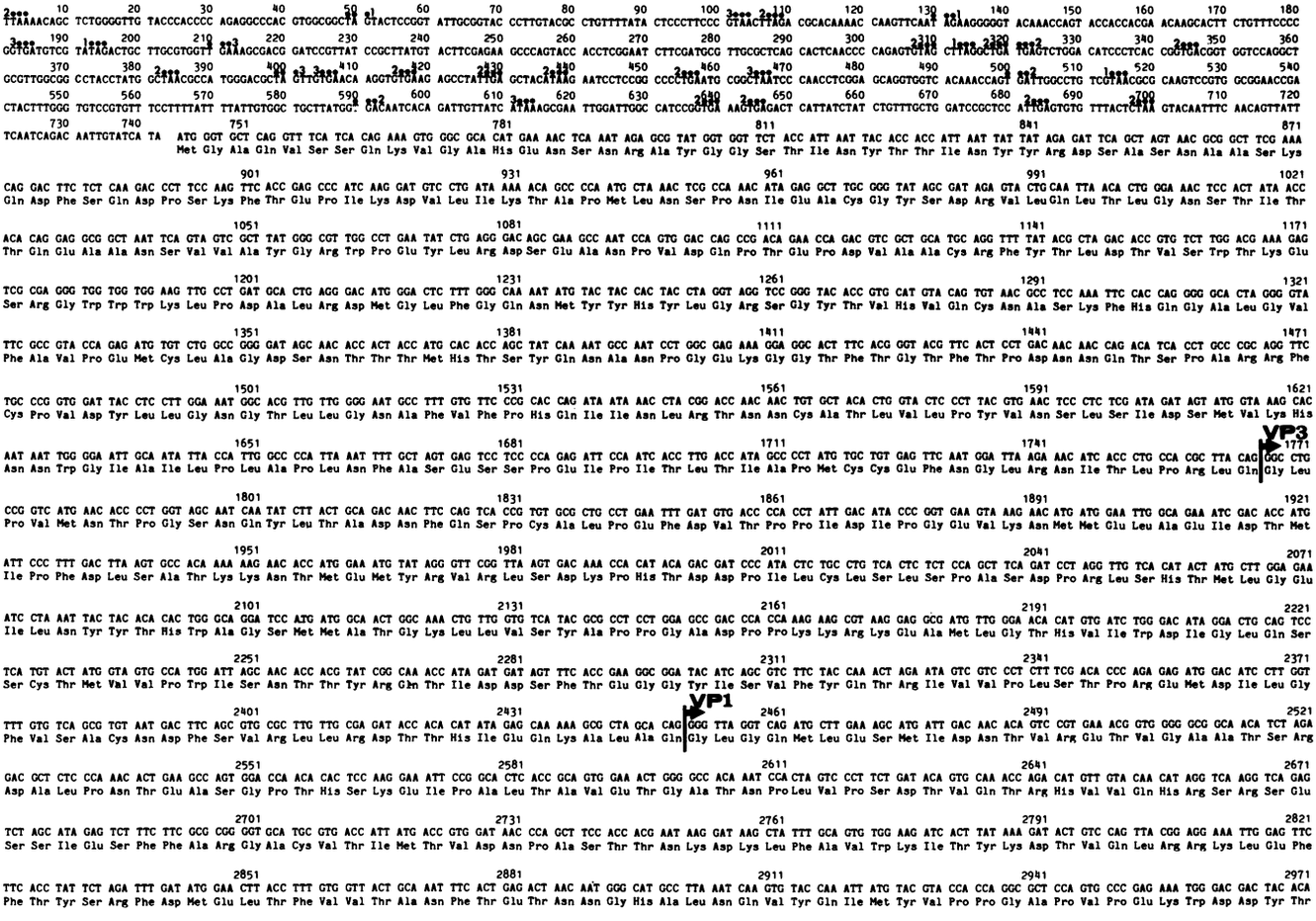

FIG. 2. (*Figure is continued on the next page.*)

FIG. 2. Nucleotide sequence of the poliovirus genome (U residues are shown as T). In the 5′ untranslated region, potential termination codons are marked by asterisks and the phase in which they occur is indicated. The sequence is translated into amino acids starting at base 743 through base 7339. The coding region and amino acid sequence of VPg, the protein linked to the 5′ end of the viral RNA, are indicated by underlining. The location of VPg is based on the published sequence (18). The positions of the virion proteins—VP1 and VP3—are indicated on the basis of preliminary amino acid sequence data (unpublished results).

mine the number of bases by which pVR103 fell short of the 5′ end of poliovirus RNA, a primer extension experiment was performed (see *Materials and Methods*). In this experiment, the longest extended product synthesized, which was obtained in quite good yield, was 220 bases long. Sequence analysis of the extended material (see below) showed that the primer extension had proceeded up to and including the 5′-terminal U of the viral RNA. This result indicated that pVR103 was missing the 5′ 115 bases of poliovirus RNA.

A portion of the primer-extended material was purified by electrophoresis, made double-stranded, and inserted into the *Pst* I site of pBR322. By using tails of oligo(dC) to clone this fragment, the *Bam*HI site, which had been phosphorylated for primer extension, was restored. Sequence analysis of the insert from one clone, pVR105, indicated that it represented bases 1–220 of the viral RNA and confirmed the nucleotide sequence of the 5′ end obtained by primer extension.

Nucleotide Sequence of the Poliovirus Genome. The sequences of plasmids pVR101 and pVR103 were determined by using the Maxam–Gilbert method (16). The sequence of about 92% of the entire genome length was determined on both strands and the remainder was determined on one strand. Many areas were analyzed two or three times to ensure accuracy. The sequence of the 5′ 115 bases was obtained both from the DNA synthesized onto the 5′-^{32}P-labeled primer and also from direct sequence analysis of the 5′ clone, pVR105. The 5′ sequence overlapped and agreed with the published 5′ 85 nucleotides (18), showing that primer extension occurred up to the initial U residue on the polioviral RNA.

The complete nucleotide sequence of the poliovirus genome is shown in Fig. 2.

The in-phase translation of the nucleotide sequence is also shown in Fig. 2. An open reading frame beginning at base 671 is followed by a methionine codon at position 743 and continues until a termination codon 71 bases from the 3′ end. Limited amino acid sequence of virion proteins VP1 and VP3 (unpublished data), as well as the published sequence of VPg (19), could be located in the predicted amino acid sequence (Fig. 2), providing strong confirmation of the identified open reading frame. The predicted amino acid sequence, however, must be considered tentative until further protein sequence is obtained.

DISCUSSION

Knowledge of the complete poliovirus RNA sequence now allows us to determine whether the open reading frame is long enough to encode a single polyprotein containing the sequences of all the poliovirus polypeptides. The open frame reported here is 6597 bases in length, or enough to encode a 245,000-dalton protein. This size is near the predicted length of a precursor that

could give rise to all of the picornavirus proteins (20). As might be expected for a cytoplasmic virus, there is no apparent need for RNA splicing to produce poliovirus mRNA.

The 5′ Sequence Before the Putative Initiating AUG. An interesting feature at the 5′ end of the poliovirus RNA molecule is the very long sequence that precedes the major open reading frame. This region contains 29 termination codons distributed among the three phases (see Fig. 2). Some of this 5′ sequence may encode short peptides. For instance, the longest open frame in the 5′ region, starting from the AUG at base 586 and terminating at base 781, could encode a 65-amino acid peptide. A long 5′ untranslated stretch in picornavirus RNA is not without precedent. In foot and mouth disease virus RNA, the poly(C) tract precedes the region where translation begins and lies 400 bases from the 5′ end of the molecule (21). The 5′ untranslated region in encephalomyocarditis virus RNA may be 1 kb in length (21).

The first 742 bases of the viral RNA before the presumed initiator AUG codon contain eight other AUG methionine codons, approximating the frequency with which AUG appears in the coding region of the genome. It has been suggested that eukaryotic translation involves a sequence of events in which ribosomes bind to the 5′ terminus of mRNA, migrate in a 3′ direction, and initiate translation when they encounter the first AUG (22). This model is clearly not applicable to poliovirus RNA because the ninth AUG is the one on which initiation appears to occur. Picornavirus RNAs also differ from other eukaryotic mRNAs by their lack of a 5′ 7-methyl-GTP cap (10, 23). Perhaps translation initiation on picornavirus RNAs occurs by a mechanism quite different from that used by other RNAs. For instance, ribosomes might bind to the RNA very near the initiating AUG, thus bypassing the cap-dependent binding process.

In contrast to the 5′-end sequence, the untranslated sequence at the 3′ end is quite short. The open reading frame ends with two consecutive termination codons starting at position 7340, leaving 71 untranslated bases before the poly(A) tail. As reported previously (21, 24), the putative polyadenylylation signal A-A-U-A-A-A is absent from the 3′ end of the viral RNA.

It has been reported that, under certain conditions, translation of poliovirus RNA *in vitro* may initiate at a second site, yielding a protein of 5,000–10,000 molecular weight (25, 26). In the poliovirus sequence reported here, there are several regions that overlap the main reading frame that could encode peptides of this size. One example of an open frame that could code for a peptide of 7200 molecular weight was discussed above. The region from the AUG at base 5308 through base 5487 could encode a peptide of 6600 molecular weight. The absence of termination codons in these and other areas, of course, may simply be due to chance.

Organization of the Coding Sequence. Mapping all the various poliovirus proteins on the viral RNA will require further amino acid sequence data. Preliminary data on VP1 and VP3 have allowed us to position the virion proteins in the 5′ half of the genome (Fig. 2). The sequence of the 5′-terminal protein found on poliovirus RNA, VPg, has been partially determined (19), and its coding region can be precisely identified in our sequence between nucleotides 5342 and 5407 (see Fig. 2). Between VPg and the end of the translatable part of the RNA, there are 2002 bases, which could encode about 72,000 molecular weight of protein. This end of the genome is known to contain the gene for NCVP-2, a protein with an estimated molecular weight of 77,000 (8). Thus VPg would be located at the 5′ edge of the NCVP-2 coding region. Pallansch *et al.* (27) suggested that the encephalomyocarditis virus VPg is not part of the protein equivalent to poliovirus NCVP-2 but is part of its precursor. In addition, it has been postulated that a protease is encoded in the region at the 3′ edge of VPg (28). Our data are consistent with this positioning, given the accuracy of molecular weight estimates. The known product cleaved from NCVP-2 is the RNA replicase, p63 (29). There is certainly sufficient coding sequence between VPg and the end of the open reading frame to encode p63 plus a second protein, possibly the protease mentioned above.

Dyad Symmetries in Poliovirus RNA. In a previous report on the 5′-terminal sequence of poliovirus RNA, a stem and loop structure was suggested (18). A computer search of the total genome sequence has identified many much longer potential stem structures, but their significance at this time is uncertain. The 5′ potential stem and loop did not present a significant barrier to the reverse transcriptase.

Frequency of CpG and Codon Usage. The frequency of the dinucleotide CpG in poliovirus RNA is lower than expected. In the coding region CpG occurs at a frequency of 2.6%, whereas a frequency of 5.3% would be expected on a random basis considering base composition. This deficiency in CpG is reflected in codon usage (see Table 1). In poliovirus, as in cellular mRNAs (30), there is a bias against CpG-containing serine, proline, threonine, and alanine codons. Furthermore, arginine codons of the type AGA and AGG occur more frequently than CGN codons (Table 1).

In vesicular stomatitis mRNAs, a deficiency in CpG has been observed within the coding triplets, at the border of codons (at adjacent codons of the type NNC/GNN) and in the noncoding region (31). In poliovirus the CpG deficiency is not observed at codon borders. On the basis of the composition of the coding region, random assortment would generate 116 CpGs between adjacent codons, and 95 are found. Furthermore, in the poliovirus 5′ noncoding segment, the observed frequency of CpG (5.4%) is close to the expected value of 6.2%.

Vertebrate cellular DNA as well as eukaryotic mRNA has a lower than expected frequency of CpG sequences (30, 32). It has been hypothesized that the low incidence of CpG in DNA is due to methylation of C residues in the dinucleotide, which drives mutation to TpG (33). However, this explanation could not apply to poliovirus RNA, which lacks methyl groups (34).

Comparison with Other Sequence Data. It is of interest to compare our sequence, derived from cloned DNA, to polioviral

Table 1. Codon usage in the poliovirus open reading frame

	U		C		A		G	
U	Phe 38	Ser	20	Tyr	40	Cys	22	U
	42		32		58		19	C
	Leu 23		48	Term	0	Term	0	A
	38		10		0	Trp	28	G
C	Leu 25	Pro	30	His	18	Arg	7	U
	26		20		33		7	C
	33		56	Gln	38		3	A
	32		13		43		7	G
A	Ile 52	Thr	53	Asn	47	Ser	21	U
	48		53		70		20	C
	33		50	Lys	63	Arg	48	A
	Met 67		13		60		23	G
G	Val 22	Ala	50	Asp	54	Gly	37	U
	29		35		63		26	C
	28		58	Glu	63		46	A
	59		18		48		33	G

Codon first position is at left, second position at top, third position at right. The total occurrence of each codon through the open reading frame of the poliovirus genome is shown. Term, chain termination.

RNA sequences obtained by using other methods. Porter *et al.* (17) synthesized end-labeled poliovirus cDNA and determined the sequence of 156 bases at the 3′ end of the viral genome. Compared to our sequence, theirs contains a base substitution at position 7380 and an extra G immediately preceding the poly(A) stretch. This additional G was also reported by Kitamura and Wimmer (24). The lack of an extra G at that position was confirmed by us in an independent 3′ clone. The sequence reported by Kitamura and Wimmer also contained three other differences from our sequence: a missing C at position 6814, and extra Ts at positions 6947 and 7228. The missing C and the extra Ts in Kitamura and Wimmer's sequence result in a 3′ untranslated region 561 nucleotides long. In contrast, according to our sequence, the 3′ untranslated region is 71 bases in length. The sequence of pVR101 in this region was confirmed on both strands.

It should be possible to use plasmids pVR101, pVR103, and pVR105 to construct a single, long clone that represents the entire poliovirus genome. The availability of a complete cloned copy of the poliovirus genome should prove useful for future studies on structure and function of the viral RNA.

Note Added in Proof. A reexamination of our sequencing gels indicates that 30 bases should be inserted following nucleotide 2104:

CTG AAG TTC ACG TTT CTG TTC TGT GGA TTC
Leu- Lys- Phe- Thr- Phe- Leu- Phe- Cys- Gly- Ser

A comparison of our poliovirus type I sequence with a recently published sequence of the same virus strain (35) reveals a number of single-base deletions, insertions, and substitutions. We have verified our sequence in all the areas that differ from the sequence reported by Kitamura *et al.*

The assistance of Mr. Robert Sege and Mr. Arthur Lee in computer analysis of our sequence and the technical assistance of Mr. Philip Hollingshead are appreciated. This work was supported by Grant AI-08388 from the National Institutes of Allergy and Infectious Diseases and Grant CA-14051 from the National Cancer Institute (core grant to S. E. Luria). V.R.R. is supported by a postdoctoral fellowship from the National Institutes of Allergy and Infectious Diseases. D.B. is an American Cancer Society Research Professor.

1. Baltimore, D. (1971) *Bacteriol. Rev.* **35**, 235–241.
2. Baltimore, D. (1969) in *The Biochemistry of Viruses*, ed. Levy, H. B. (Dekker, New York), pp. 101–176.
3. Granboulan, N. & Girard, M. (1969) *J. Virol.* **4**, 475–479.
4. Jacobson, M. F. & Baltimore, D. (1968) *Proc. Natl. Acad. Sci. USA* **61**, 77–84.
5. Holland, J. A. & Kiehn, E. D. (1968) *Proc. Natl. Acad. Sci. USA* **60**, 1015–1022.
6. Summers, D. F. & Maizel, J. V., Jr. (1968) *Proc. Natl. Acad. Sci. USA* **59**, 966–971.
7. Jacobson, M. F., Asso, J. & Baltimore, D. (1970) *J. Mol. Biol.* **49**, 657–669.
8. Taber, R., Rekosh, D. M. & Baltimore, D. (1971) *J. Virol.* **8**, 395–401.
9. Saborio, J. L., Pong, S.-S. & Koch, G. (1974) *J. Mol. Biol.* **85**, 195–211.
10. Flanegan, J. B., Pettersson, R. F., Ambros, V., Hewlett, M. & Baltimore, D. (1977) *Proc. Natl. Acad. Sci. USA* **74**, 961–965.
11. Bothwell, A. L. M., Paskind, M., Reth, M., Imanishi-Kari, T., Rajewsky, K. & Baltimore, D. (1981) *Cell* **24**, 625–637.
12. Taylor, J. M., Illmensee, R. & Summers, J. (1976) *Biochim. Biophys. Acta* **442**, 324–330.
13. Grunstein, M. & Hogness, D. S. (1975) *Proc. Natl. Acad. Sci. USA* **72**, 3961–3965.
14. Casey, J. & Davidson, N. (1977) *Nucleic Acids Res.* **4**, 1539–1552.
15. Lamb, R. A. & Lai, C. J. (1980) *Cell* **21**, 475–485.
16. Maxam, A. M. & Gilbert, W. (1980) *Methods Enzymol.* **65**, 499–559.
17. Porter, A. G., Fellner, P., Black, D. N., Rowlands, D. J., Harris, T. J. R. & Brown, F. (1978) *Nature (London)* **276**, 298–301.
18. Larsen, G. R., Semler, B. L. & Wimmer, E. (1981) *J. Virol.* **37**, 328–335.
19. Kitamura, N., Adler, C. J., Rothberg, P. G., Martinko, J., Nathenson, S. G. & Wimmer, E. (1980) *Cell* **21**, 295–302.
20. Baltimore, D., Jacobson, M. F., Asso, J. & Huang, A. S. (1969) *Cold Spring Harbor Symp. Quant. Biol.* **34**, 741–746.
21. Fellner, P. (1979) in *The Molecular Biology of Picornaviruses* ed. Perez-Bercoff, R. (Plenum, New York), pp. 25–47.
22. Kozak, M. (1978) *Cell* **15**, 1109–1123.
23. Lee, Y. F., Nomoto, A., Detjen, B. M. & Wimmer, E. (1977) *Proc. Natl. Acad. Sci. USA* **74**, 59–63.
24. Kitamura, N. & Wimmer, E. (1980) *Proc. Natl. Acad. Sci. USA* **77**, 3196–3200.
25. Celma, M. L. & Ehrenfeld, E. (1975) *J. Mol. Biol.* **98**, 761–780.
26. Humphries, S., Knauert, F. & Ehrenfeld, E. (1979) *J. Virol.* **30**, 481–488.
27. Pallansch, M. A., Kew, O. M., Palmenberg, A. C., Golini, F., Wimmer, E. & Rueckert, R. (1978) *J. Virol.* **35**, 414–419.
28. Palmenberg, A. C., Pallansch, M. S. & Rueckert, R. R. (1979) *J. Virol.* **32**, 770–778.
29. Flanegan, J. B. & Baltimore, D. (1979) *J. Virol.* **29**, 352–360.
30. Grantham, R., Gautier, C., Gouy, M., Mercier, R. & Pave, A. (1980) *Nucleic Acids Res.* **8**, r49–r62.
31. Rose, J. K. & Gallione, C. J. (1981) *J. Virol.*, in press.
32. Russell, G. J., Walker, P. M. B., Elton, R. A. & Subak-Sharpe, J. H. (1976) *J. Mol. Biol.* **108**, 1–23.
33. Salser, W. (1978) *Cold Spring Harbor Symp. Quant. Biol.* **42**, 985–1002.
34. Fernandez-Munoz, R. & Darnell, J. E. (1976) *J. Virol.* **18**, 719–726.
35. Kitamura, N., Semler, B. L., Rothberg, P. G., Larsen, G. R., Adler, C. J., Dorner, A. J., Emini, E. A., Hanecak, R., Lee, J. J., van der Werf, S., Anderson, C. W. & Wimmer, E. (1981) *Nature (London)* **291**, 547–553.