Volume 8 Number 9 1980 Nucleic Acids Research

Cloning of influenza cDNA into M13: the sequence of the RNA segment encoding the A/PR/8/34 matrix protein

Greg Winter and Stan Fields

MRC Laboratory of Molecular Biology, Hills Road, Cambridge, CB2 2QH, UK

Received 17 March 1980

ABSTRACT

A strategy has been developed for sequencing the single-stranded RNA genes of influenza virus. Restriction fragments derived from double-stranded cDNA copies of total influenza RNA were cloned into the bacteriophage M13mp2 and sequenced by the dideoxy technique. Sequences were extended and overlapped by using the virion RNA as template and priming with small restriction fragments. In the course of this strategy, the nucleotide sequence of segment 7 (1027 nucleotides) was completed and provides the primary structure of the matrix protein (27,861 daltons). In addition, there is a second long reading frame which partly overlaps the reading frame of the matrix protein.

INTRODUCTION

The genome of human influenza A virus is composed of eight single-stranded RNA segments [1] totalling about 14,000 nucleotides [2,3] and encoding at least nine polypeptides [4-6]. Exchange of RNA segments between different strains of influenza can occur naturally [7] and may well be involved in initiating new influenza pandemics [8]. A complete nucleotide sequence analysis of a human strain and its comparison with other strains should establish whether indeed simple segment exchange is the main mechanism by which pandemic strains arise and, if so, identify the parental strains. Sequencing should also reveal all the possible influenza polypeptide products and may even illuminate their early evolution. For example, are the three P polypeptides derived from a common ancestral polypeptide? Towards these ends we are sequencing an early human strain, A/PR/8/34.

The 5' and 3' ends of all the influenza segments have been sequenced directly [9,10] and the entire haemagglutinin gene has been sequenced after cloning the cDNA into a plasmid [11]. In order to sequence the total influenza genome using the rapid dideoxy method of sequencing [12] we chose a strategy based on the "shot-gun" approach of Sanger and colleagues for the mitochondrial genome [13]. Double stranded (ds)

cDNA from all the influenza segments was digested with restriction enzymes and ligated to the replicative form (RF) of the bacteriophage M13mp2 [14,15]. Single-stranded template was prepared from recombinant phages and sequenced using dideoxy nucleotides and a primer flanking the site of insertion [16-18]. Additionally, small restriction fragments derived from existing clones were used as primers for dideoxy sequencing on RNA [19-21] (virion RNA [vRNA] or influenza mRNA). These primings extended and overlapped existing blocks of sequence obtained from the shotgun approach.

Using this strategy, we have determined the sequence of segment 7, which encodes the most abundant polypeptide found in virions, the matrix (M) protein. This protein forms a continuous shell on the inner side of the lipid bilayer, but its function is unclear [22]. In addition, a second open reading frame in segment 7, overlapping the last 65 nucleotides coding for the M protein, suggests the possibility of another influenza polypeptide.

MATERIALS AND METHODS

Preparation of restriction fragments

The influenza strain A/PR/8/34 was grown in fertile hen eggs and purified. The RNA was extracted [10], polyadenylated [23] and single-stranded cDNA prepared with reverse transcriptase and $p(dT)_{12-18}$ primer (P.L. Biochemicals) [3]. The single strands were back-copied with reverse transcriptase to form ds cDNAs [24] and digested with MboI (Biolabs) or AluI (Biolabs).

Ligation into M13mp2 or M13mp2/Bam

EcoRI linkers (Collaborative Research) were ligated to the AluI fragments and then digested with EcoRI (Biolabs) [25]. These fragments were purified on a 1 ml Sepharose 4B column [25] and ligated into the EcoRI site of M13mp2 [15]. The MboI fragments were ligated directly into the BamHI site of M13mp2/Bam. The vector M13mp2/Bam is identical to M13mp2 except that a BamHI site has been removed from the vector and a small oligonucleotide containing a BamHI site has been introduced at the EcoRI site [26]. A typical ligation (10 µl) contained ds cDNA fragments derived from 1 µg influenza RNA, 20 ng M13mp2 vector, 0.25 units T4 ligase (Miles or a gift from Dr. A.R. MacLeod) in 50 mM Tris-Cl pH 7.5, 10 mM $MgCl_2$, 1 mM dithiothreitol (DTT), 100 µM ATP and was incubated for 24 hr at 4°C.

Transformation and purification of recombinant DNA

E. coli JM101 [Δ (lacpro) supE, thi, F' tra D 36 proAB lac I^q ZΔM15], provided by J. Messing [27] was transformed by the calcium chloride technique

[28], the white recombinant plaques were tooth-picked into 1 ml 2 x TY [29] containing early exponential JM101 and grown for 4½ hr at 37°C. The culture was centrifuged for 5 min in an Eppendorf microfuge and the phage precipitated from the supernatant by addition of 0.2 ml 2.5 M NaCl, 20% PEG (6000). After a further spin, the phage pellet was dissolved in 100 µl 10 mM Tris-Cl pH 7.5, 1 mM EDTA, extracted with phenol and ether, and the single-stranded DNA was precipitated from ethanol. The DNA was dissolved in 50 µl 10 mM Tris-Cl pH 7.5, 1 mM EDTA and 5 µl used for dideoxy sequencing.

The cloning and growth of recombinant M13 strains was carried out under Category II* containment conditions as advised by the U.K. Genetic Manipulation Advisory Group. M13mp2 contained an amber mutation (R. Cortese, unpublished) as did M13mp2/Bam (G.W., unpublished).

Nucleotide sequencing

Single strand M13 template was sequenced by the dideoxy technique [12] using the Klenow fragment of DNA polymerase (Boehringer) and a "universal primer" [16-18]. Restriction fragments prepared as primers for dideoxy sequencing on RNA [19-21] were pretreated with exonuclease III (Biolabs)[30]. Direct RNA sequencing on 5' labelled RNA was similar to that described by Simoncsits et al. [31] except that $MgCl_2$ was added to the formamide [32].

Clone turn-around

The RF of clone C4 was purified from infected cells by cleared lysis [33] and a caesium chloride gradient. Alternatively, small quantities of RF (<1 µg) may be conveniently prepared for clone turn-around on 1% low gelling temperature (LGT) agarose gels. The RF was cut with EcoRI and the digestion checked for completeness by running an aliquot on an agarose gel. Ten ng cut RF was re-ligated as described above except in 2 µl at 14°C for 6 hr. After transformation, the white plaques were grown up and the phage precipitated with PEG (as above) and redissolved in 50 µl water. Clones with inserts in an orientation opposite to that of the original clone were identified by hybridisation, the hybrid having a lower mobility on agarose gels than the parent single strands. A capillary containing 2 µl aliquots of the test clones was incubated at 67°C for 1 hr with 2 µl of the original reference clone, 4 µl (100 mM Tris-Cl pH 7.5, 100 mM $MgCl_2$, 500 mM NaCl) and 1 µl (80% glycerol, 1% SDS and 0.2% bromophenol blue). The capillaries were cooled in ice and the contents loaded directly on a 1% agarose gel. The screening of hybridised single strands on agarose gels has been developed independently [34].

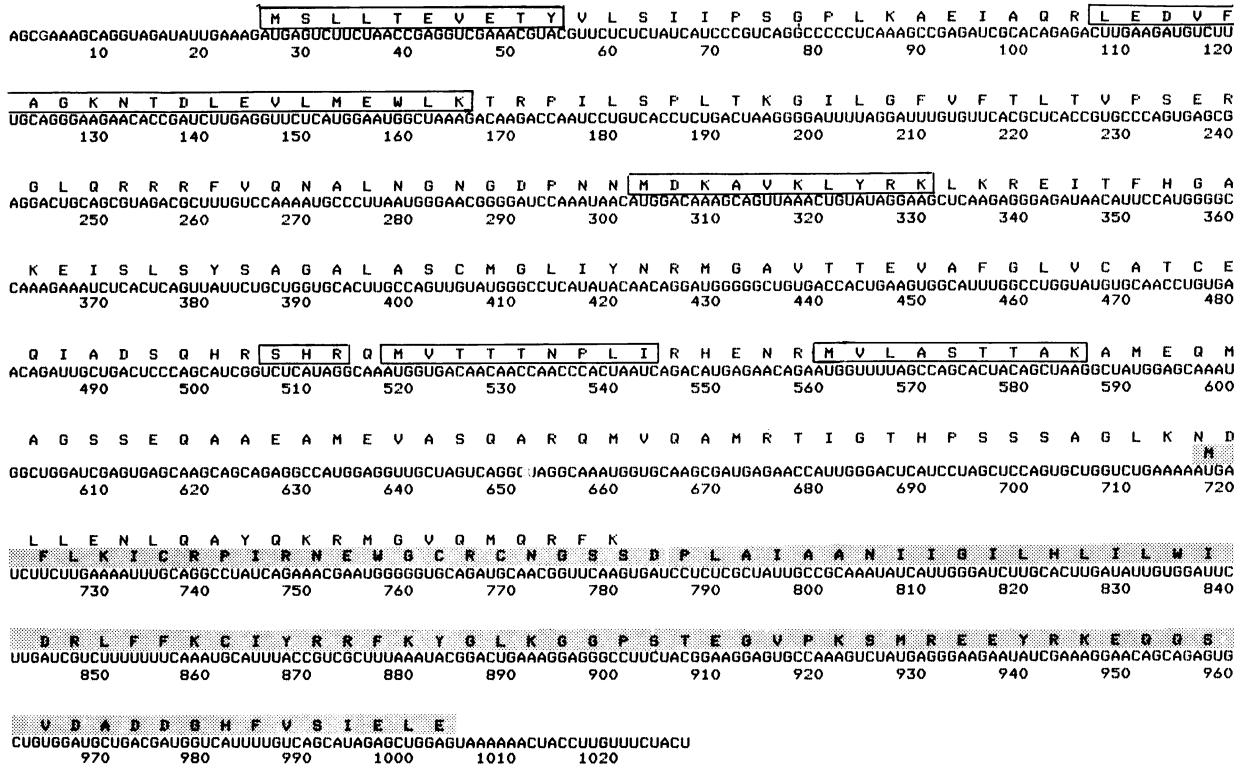

Figure 1. The nucleotide sequence of influenza segment 7 (strain A/PR/8/34) in the mRNA sense (complementary to vRNA). The amino acid sequences predicted by the two open reading frames are depicted in the one letter amino acid code; the second frame is cross-hatched. Peptides resembling those described in [35] and [36] are boxed.

RESULTS

From shotgun cloning of AluI and MboI restriction fragments of total ds cDNA, a master file of sequences derived from the entire genome was created. One clone (C4) containing a poly(A) tail matched and extended the sequence at the 3' end of segment 7 as determined by Both and Air [35]. Discrepancies between the two sequences (bases 78-110 in Fig. 1) were double checked by sequencing the clone in its opposite orientation (Fig. 2).

The replicative form of clone C4 was isolated from infected cells, cut with

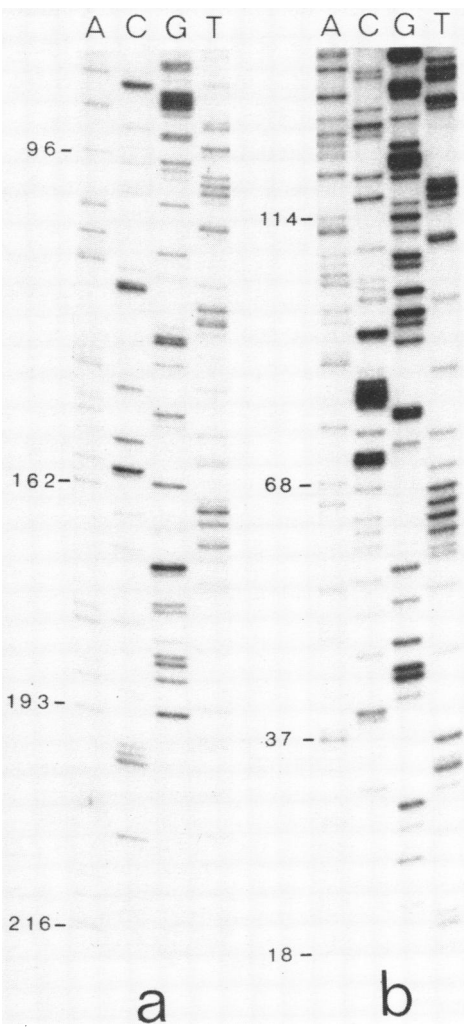

Figure 2. Sequences from clone C4 in both orientations. Dideoxy primings with a 30-nucleotide long primer [18]; (a) on clone 4, and (b) on clone 4 turned around were fractionated on a 6% polyacrylamide, 7 M urea thin gel. Sequence positions correspond to Fig. 1. In (a) the sequence read from the gel is the complement of that in Fig. 1, whereas in (b) the sequence is the same as that in Fig. 1.

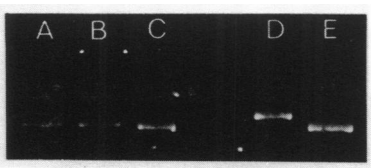

Figure 3. Screening for clone turn-around. Incubations of annealed clones (see Methods) were separated on a 1% agarose gel and stained with ethidium bromide. A, clone C4 reference; B, first test clone; C, second test clone, D, clone C4 and first test clone; E, clone C4 and second test clone. The first test clone was in the opposite orientation to clone C4.

EcoRI and re-ligated. The inserted sequence was thus cut out and on re-ligation to the vector may adopt the opposite orientation It is important to completely cut the RF since otherwise a high proportion of the newly derived clones will be in the original orientation Single strands from re-ligated clones were screened on agarose gels by hybridisation with single-stranded DNA from a clone in the original orientation (Fig. 3). Of the 12 white plaques screened, four were in the opposite orientation. Clone C4 and C4 turned around, as well as MboI clones whose sequences were contained within C4, provided the sequence of the 3' 331 nucleotides of the vRNA (Fig. 4).

The 3' sequence was extended toward the 5' end by priming on the vRNA (Fig. 5) with a fragment (bases 290-331) prepared by a BamHI/EcoRI digest of the purified RF of clone C4. The sequence of about 200 residues from this priming overlapped with a block of sequence (bases 524-894) already determined (Fig. 4). This block had been obtained from AluI and MboI clones. In turn, the sequence was extended to the very 5' end of the vRNA by priming with a restriction fragment prepared by a Sau3A/EcoRI digest of the RF of clone C17. Bases 951 and 954 were difficult to identify since there were bands in all four tracks. These bases were assigned by direct enzymic sequencing [31] on 5' labelled segment 7 vRNA (results not shown).

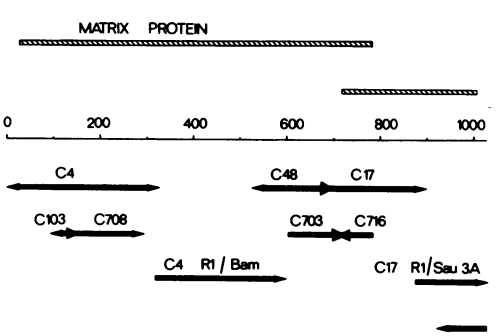

Figure 4. Summary of evidence. A, first reading frame; B, second reading frame; C, sequence in mRNA sense (as in Fig. 1); D, AluI clones; E, MboI clones; F, primings on vRNA; G, direct RNA sequencing. The arrows indicate the directions of sequencing.

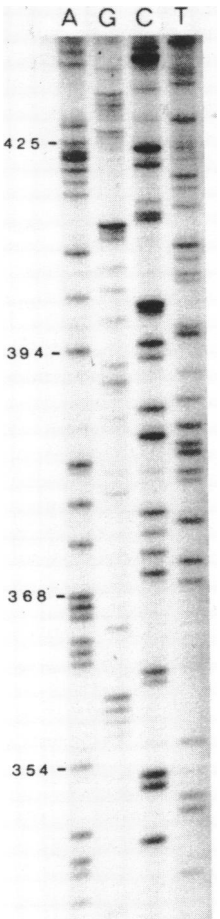

Figure 5. Sequence from priming on vRNA with BamHI/EcoRI clone C4 primer. Dideoxy primings with a 43-long primer derived from clone C4 were fractionated on a 6% polyacrylamide, 7 M urea thin gel. Sequence positions correspond to Fig. 1.

DISCUSSION

The M13 strategy we have adopted for the influenza genome provides an extremely rapid method for sequencing RNA. Cloning was used to provide single-stranded DNA templates corresponding to many different parts of the RNA, with no need to characterise recombinant clones before sequencing. With the ease of sequencing by a universal primer, sequences from all over the genome were readily obtained. These sequences were stored in a computer master file to which the sequence derived from each new clone was compared. New sequences which overlapped with the master file were added in at the position of overlap [36,37].

The use of the shotgun strategy made it inevitable that some fragments would be cloned more than once. Within limits, such repetition of sequence is useful, providing a cross-check on both accuracy and on possible microheterogeneity in the RNA population. However, as the master file increased in size, random cloning returned increasing yields of purely redundant sequences. Some order was therefore introduced into the random approach. Firstly, long inserts in M13 were turned around to allow sequencing from the other end. Secondly, small restriction fragments derived from the RF of sequenced clones were used as primers on the virion RNA, and also on influenza messenger RNA purified from infected cells (results not shown). Thus a single primer could be extended in both directions.

Using this combination of random cloning, clone turn-around and priming on RNA, we have completed the structure of the segment encoding the influenza matrix protein.

When the sequence in its messenger sense is translated, it shows two long reading frames (bases 26-781 and 717-1004). The first reading frame encodes the matrix protein as is apparent from the similarity in composition and/or sequence of tryptic, chymotryptic [35] and cyanogen bromide [38] peptides with those predicted from the nucleotide sequence. The predicted molecular weight of 27,861 daltons is in agreement with the 25,000 daltons [39] determined experimentally, as is the amino acid composition (Table 1). The analysis is rich in arginine and methionine and sparse in aspartic acid (Table 1). This results in a somewhat hydrophobic protein soluble in chloroform/methanol [42] with a high net positive charge of +9.5 at pH 7.5. It is tempting to speculate that basic residues, perhaps as clusters, line the inner face of the virion shell and interact with the negative phosphate groups of the RNA wound on the nucleoprotein (NP) cores [43]. A secondary

Table 1. Amino acid composition (mole %) of the matrix protein as determined experimentally [40] is compared with that predicted by the nucleotide sequence. The average amino acid composition of proteins is included for comparison [41].

	Matrix protein (experimental)	Matrix protein (predicted)	Average protein
Cys	-	1.2	2.9
Asp	7.3	2.4	5.5
Asn		4.4	4.3
Thr	6.6	7.1	6.1
Ser	6.8	7.1	7.0
Glu	13.2	6.8	6.0
Gln		6.0	3.9
Pro	3.7	3.2	5.2
Gly	<7.6	6.4	8.4
Ala	10.5	9.9	8.6
Val	6.9	6.4	6.6
Met	3.9	5.6	1.7
Ile	4.1	4.4	4.5
Leu	10.3	10.3	7.4
Tyr	1.8	2.0	3.4
Phe	2.7	2.8	3.6
Trp	-	0.4	1.3
His	1.7	2.0	2.0
Lys	5.7	5.2	6.6
Arg	7.0	6.8	4.9

structure prediction of the protein by Chou-Fasman rules [44] supplemented by use of the helix wheel [45] reveals a cluster of basic residues on one side of an α-helix (Lys-94, Lys-97, Arg-100, Lys-101, Lys-103, Arg-104) which may fulfil this function. The solubility of the protein in an organic solvent suggests that the hydrophobic residues of the protein may not only interact with the lipid coat on the surface of the virion, but also comprise the intersubunit contacts.

The second open reading frame in segment 7 suggests a possible undiscovered polypeptide (11,033 daltons). Overlapping genes have been found in several viruses such as ØX174, G4 [46], SV40 [47,48] and have been tentatively identified in segment 8 of influenza [5,6]. The presence of such overlapping genes in influenza would suggest that strong selection pressure in favour of a compact genome is also operating on influenza. This view is supported by the nucleotide sequence of the segment encoding the haemagglutinin, in which only 21 bases precede the initiation codon and 29 bases follow the termination codon [11].

ACKNOWLEDGEMENTS

We deeply appreciate the facilities and encouraging advice offered by Dr. G.G. Brownlee throughout the course of this work. We thank Drs. F. Sanger and R. Cortese for practical suggestions, Mrs. M.A. Robertson for supplying virus and Dr. A.D. McLachlan for the Chou-Fasman prediction. Dr. B.W.J. Mahy kindly provided facilities for growing the virus.

REFERENCES

1. McGeoch, D., Fellner, P. and Newton, C. (1976) Proc. Nat. Acad. Sci. USA 73, 3045-3049.
2. Desselberger, U. and Palese, P. (1978) Virology 88, 394-399.
3. Sleigh, M.J., Both, G.W. and Brownlee, G.G. (1979) Nucleic Acids Res. 6, 1309-1321.
4. Inglis, S.C., McGeoch, D.J. and Mahy, B.W.J. (1977) Virology 78, 522-536.
5. Inglis, S.C., Barrett, T., Brown, C.M. and Almond, J.W. (1979) Proc. Nat. Acad. Sci. USA 76, 3790-3794.
6. Lamb, R.A. and Choppin, P.W. (1979) Proc. Nat. Acad. Sci. USA 76, 4908-4912.
7. Young, J.F. and Palese, P. (1979) Proc. Nat. Acad. Sci. USA 76, 6547-6551.
8. Laver, W.G. and Webster, R.G. (1979) British Medical Bulletin 35, 29-33.
9. Skehel, J.J. and Hay, A.J. (1978) Nucleic Acids Res. 5, 1207-1218.
10. Robertson, J.S. (1979) Nucleic Acids Res. 6, 3745-3757.
11. Porter, A.G., Barber, C., Carey, N.H., Hallewell, R.A., Threlfall, G. and Emtage, J.S. (1979) Nature 282, 471-477.
12. Sanger, F., Nicklen, S. and Coulson, A.R. (1977) Proc. Nat. Acad. Sci. USA 74, 5463-5467.
13. Sanger, F., Coulson, A.R., Barrell, B.G., Smith, A.J.H. and Roe, B.A.

(1980) manuscript submitted for publication.
14. Messing, J., Gronenborn, B., Müller-Hill, B. and Hofschneider, P.H. (1977) Proc. Nat. Acad. Sci. USA 74, 3642-3646.
15. Gronenborn, B. and Messing, J. (1978) Nature 272, 375-377.
16. Schreier, P.H. and Cortese, R. (1979) J. Mol. Biol. 129, 169-172.
17. Heidecker, G., Messing, J. and Gronenborn, B. (1980) Gene, in press.
18. Anderson, S., Gait, M.J., Mayol, L. and Young, I.G. (1980) manuscript submitted for publication.
19. McReynolds, L., O'Malley, B.W., Nisbet, A.D., Fothergill, J.E., Givol, D., Fields, S., Robertson, M.A. and Brownlee, G.G. (1978) Nature 273, 723-728.
20. Hamlyn, P.H., Brownlee, G.G., Cheng, C., Gait, M.J. and Milstein, C. (1978) Cell 15, 1067-1075.
21. Zimmern, D. and Kaesberg, P. (1978) Proc. Nat. Acad. Sci. USA 75, 4257-4261.
22. Lenard, J. and Compons, R.W. (1974) Biochim. Biophys. Acta 344, 51-94.
23. Sippel, A.E. (1973) Eur. J. Biochem. 37, 31-40.
24. Sleigh, M.J., Both, G.W. and Brownlee, G.G. (1979) Nucleic Acids Res. 7, 879-893.
25. Maniatis, T., Hardison, R.C., Lacy, E., Lauer, J , O'Connell, C., Quon, D., Sim, G.K. and Efstratiadis, A. (1978) Cell 15, 687-701.
26. Rothstein, R.J., Lall, L.F., Bahl, C.P., Narang, S.A. and Wu, R. (1980) Methods in Enzymology, in press.
27. Messing, J. (1979) Recombinant DNA Technical Bulletin 2, 43-48.
28. Cohen, S.N., Chang, A.C.Y. and Hsu, L. (1972) Proc. Nat. Acad. Sci. USA 69, 2110-2114.
29. Miller, J. (1972) in "Experiments in Molecular Genetics" Cold Spring Harbor, New York.
30. Zain, B.S. and Roberts R.J. (1979) J. Mol. Biol. 131, 341-352.
31. Simoncsits, A., Brownlee, G.G., Brown, R.S., Rubin, J.R. and Guilley, H. (1977) Nature 269, 833-836.
32. Winter, G. and Brownlee, G.G. (1978) Nucleic Acids Res. 5, 3129-3139.
33. Clewell, D.B. and Helinski, D.R. (1969) Proc. Nat. Acad. Sci. USA 62, 1159-1166.
34. Herrmann, R., Neugebauer, K., Pirkl, E., Zentgraf, H. and Schaller, H. (1980) Mol. Gen. Genet. 177, 231-242.
35. Both, G.W. and Air, G.M. (1979) Eur. J. Biochem. 96, 363-372.
36. Staden, R. (1977) Nucleic Acids Res. 4, 4037-4051.
37. Staden, R. (1979) Nucleic Acids Res. 6, 2601-2610.
38. Robertson, B.H., Bhown, A.S., Campans, R.W. and Bennett, J.C. (1979) J. Virol. 30, 759-766.
39. Skehel, J.J. (1972) Virology 49, 23-36.
40. Erikson, A.H. and Kilbourne, E.D. (1980) Virology 100, 34-42.
41. Dayhoff, M.O., Hunt, L.T. and Hurst-Calderone, S. (1978) in Atlas of Protein Sequence and Structure, Vol. 5, Supplement 3, Chapter 25, (ed. M.O. Dayhoff) National Biomedical Research Foundation.
42. Gregoriades, A (1973) Virology 54, 369-383.
43. Wrigley, N.G. (1979) British Medical Bulletin 35, 35-38.
44. Chou, P.Y. and Fasman, G.D. (1974) Biochemistry 13, 211-245.
45. Schiffer, M. and Edmundson, A.B. (1967) Biophys. J. 7, 121-135.
46. Fiddes, J.C. and Godson, G.N. (1979) J. Mol. Biol. 133, 19-43.
47. Reddy, V.B., Thimmappaya, B., Dhar, R., Subramanian, K.N., Zain, B.S., Pan, J., Ghosh, P.K., Celma, M.L. and Weissman, S.M. (1978) Science 200, 494-502.
48. Fiers, W., Contreras, R., Haegeman, G., Rogiers, R., Van de Voorde, A., Van Heuverswyn, H., Van Herreweghe, J., Volckaert, G. and Ysebaert, M. (1978) Nature 273, 113-120.