Volume 12 Number 5 1984 Nucleic Acids Research

The complete nucleotide sequence of the RNA coding for the primary translation product of foot and mouth disease virus

A.R.Carroll*, D.J.Rowlands* and B.E.Clarke*

Animal Virus Research Institute, Ash Road, Pirbright, Woking, Surrey, UK

Received 10 January 1984; Revised and Accepted 16 February 1984

ABSTRACT

The complete nucelotide sequence of the coding region of foot and mouth disease virus RNA (strain $A_{10}61$) is presented. The sequence extends from the primary initiation site, approximately 1200 nucleotide from the 5' end of the genome, in an open translational reading frame of 6,999 nucleotides to a termination codon 93 nucleotides from the 3' terminal poly (A). Available amino acid sequence data correlates with that predicted from the nucleotide sequence. The amino acid sequence around cleavage sites in the polyprotein shows no consistency, although a number of the virus-coded protease cleavage sites are between glutamate and glycine residues.

INTRODUCTION

Foot and mouth disease (FMD) is a highly contagious viral disease of cloven-hooved animals and is of paramount economic importance. The virus (genus: aphthovirus; family: Picornaviridae) has a single stranded positive sense RNA genome of approximately 8,500 nucleotides. The genome is 3' polyadenylated, can serve directly as a messenger and contains a poly (C) tract (100-200 nucleotides long) located approximately 400 nucleotides from the 5' end (1).

Unlike most eukaryotic mRNAs, FMDV RNA is not capped at its 5' end but has a small viral-coded protein (VPg) covalently attached to the terminal 5' uridine residue (2, 3). The VPg protein is not required for translation and, by correlation with poliovirus is probably involved in replication (4, 5). The RNA of FMDV, like other picornaviruses, appears to have a long 5' untranslated leader sequence. Removal of the poly C tract of FMDV and all nucleotides to its 5' side has no effect on the spectrum of proteins generated by in vitro translation. This indicates that the site for initiation of translation is located to the 3' side of the poly (C) tract (6).

The initial translational product of FMDV RNA would be a polyprotein of about 250,000 daltons. However, this protein is not observed as it is nascently processed into the viral structural and non-structural proteins as shown in Figure 1.

We describe here the complete nucleotide sequence of the region coding for this poly protein from one strain of FMDV.

MATERIALS AND METHODS

Recombinant Plasmids

The construction of recombinant plasmids has been described previously (7). Briefly double stranded cDNA was synthesised using RNA from FMDV strain A_{10} 61 and inserted by G:C tailing into the Pst I site of pAT 153. Two of the resulting recombinants (pFA76 and pFA206) contained inserts corresponding to ≈85% of the FMDV genome. Detailed restriction maps of these recombinants were generated by standard procedures. pFA206 was subcloned into two more manageable size clones pFA206α or pFA206β. Initially pFA206 was digested with EcoRI, recircularised using T4 DNA ligase and used to transform E. coli MC1061 by standard procedures (8, 9). This resulted in the clone pFA206α representing those sequences to the 3' side of the EcoRI site in pFA206 (see Figure 1). pFA206β was generated by digestion of pFA206 with EcoRI and PstI, gel purification of the required band and ligation into EcoRI/PstI digested pAT 153.

DNA sequencing

All DNA sequencing was carried out by the method of Maxam and Gilbert (10). Recombinant DNA was sequenced using uniquely 5' or 3' ^{32}P labelled restriction fragments as described previously (11). Primer extension sequencing was carried out as described by Rowlands et al (12). Briefly, a 5' ^{32}P-labelled primer (Cell Tech. Ltd., UK) was used to direct the synthesis of cDNA using reverse transcriptase and viral RNA as a template. The resulting 5' labelled cDNA was size fractionated and sequenced by the method of Maxam and Gilbert (10).

Analysis of Sequence Data

DNA sequence was analysed using an Apple II microcomputer as described previously (13).

RESULTS AND DISCUSSION

Sequencing Strategy

The construction of cDNA clones from FMDV RNA (serotype A_{10} 61) has been described elsewhere (7). These clones represented in total about 85% of the viral genome (Figure 1). The nucleotide sequence of the region coding for the structural protein (VPs 1-4) has been previously reported (11). As shown in Figure 1 a large clone (pFA206) having a cDNA insert of 5.4Kb represented the major part of the genome coding for the non-structural proteins. To simplify sequence analysis this insert was sub-cloned into two halves (pFA206β and pFA206α) as described in methods. Detailed restriction analysis of these clones (see Figure 1) allowed a sequencing strategy to be carried out so that most regions of the genome were sequenced in both strands. Those regions which were not sequenced in both strands are clearly indicated in Figure 1 and these regions were determined from

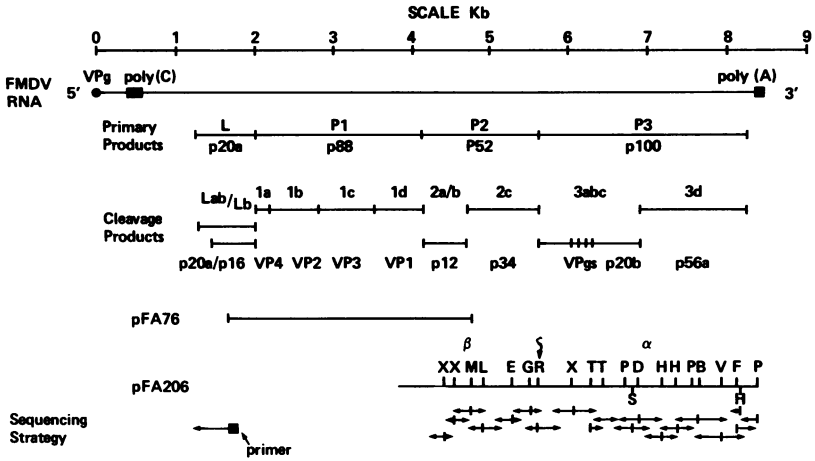
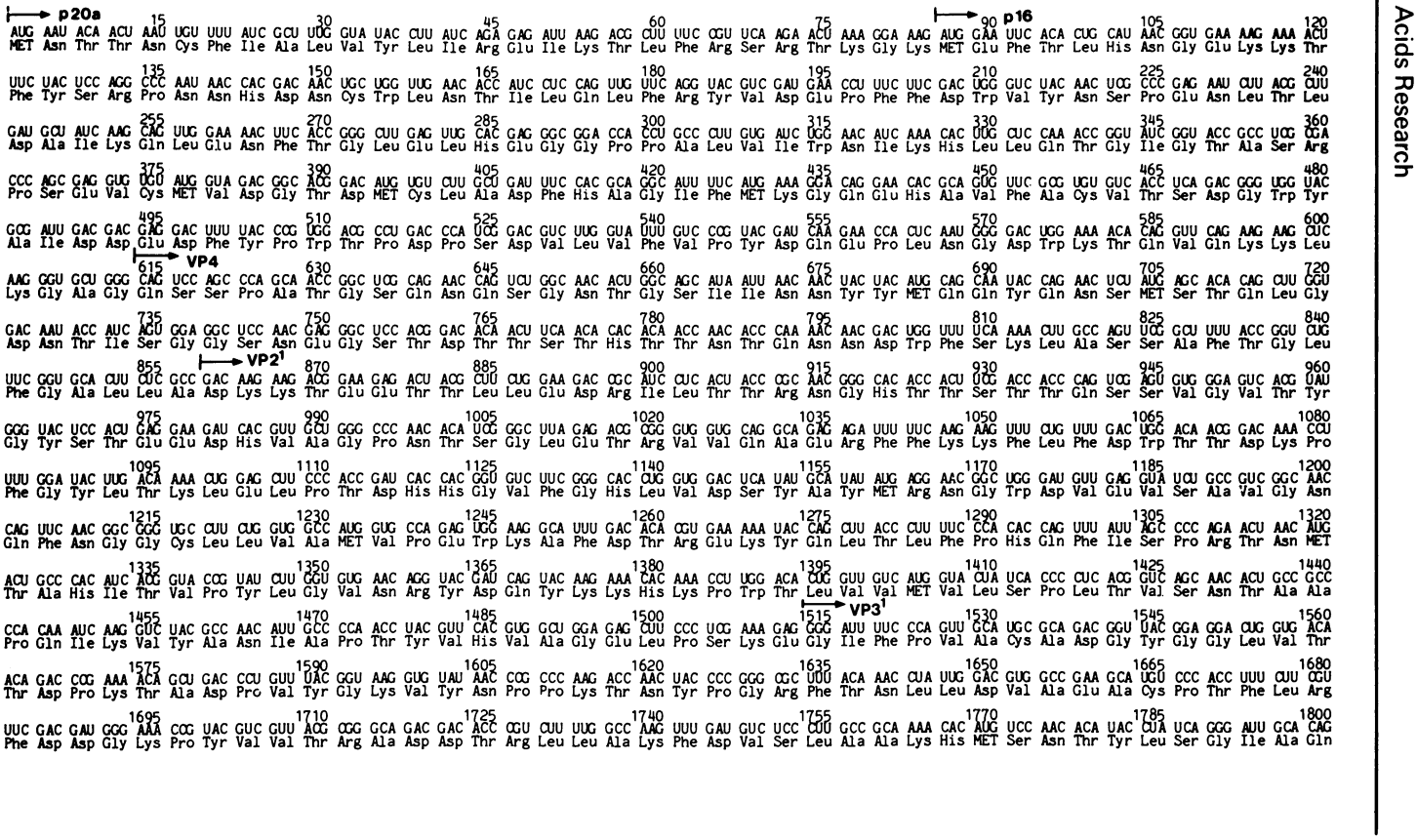


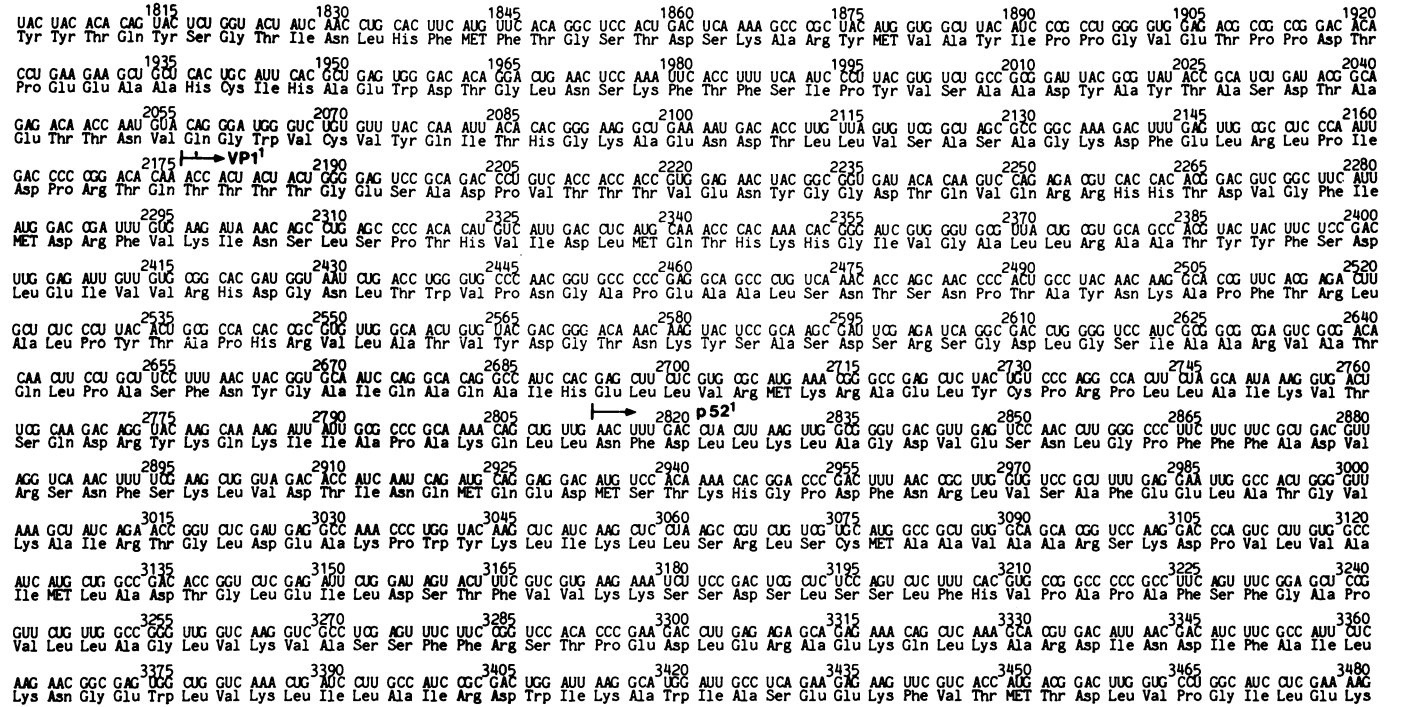
FIGURE 1 Organisation of the FMDV genome. The positions of the gene products is based on previously published work (1). The naming of the polypeptides has been done using both existing names and the recommendations agreed at the 3rd European Study Group on moleculear biology of picornaviruses, Urbino, Italy, September, 1983. The new nomenclature is indicated above the corresponding polypeptides and the old nomenclature is indicated below. The alignment of two cDNA clones with respect to the viral genome is shown. The Eco R1 site was used to sub-clone pFA206 into α and β as indicated. The sequence of 2820 nucleotides from the 5' end of the clone pFA76 has been reported previously (7). The sequence strategy is illustrated by the bars and arrows indicating the restriction sites used and the extent of the sequence obtained. Primer extension sequencing was used to obtain sequence to the 5' side of pFA76 and the location of the primer and sequence obtained is shown. The abbreviations used for restriction enzymes are: B, Bam HI; D, Hind III: E, Dde I; F, Hin FI; G, Bgl II; H, Hpa II; L, Bgl I; M, Sma I; P, Pst I; R, Eco RI; S, Sal I; T, Taq I; V, Pvu II; X, Xho I.

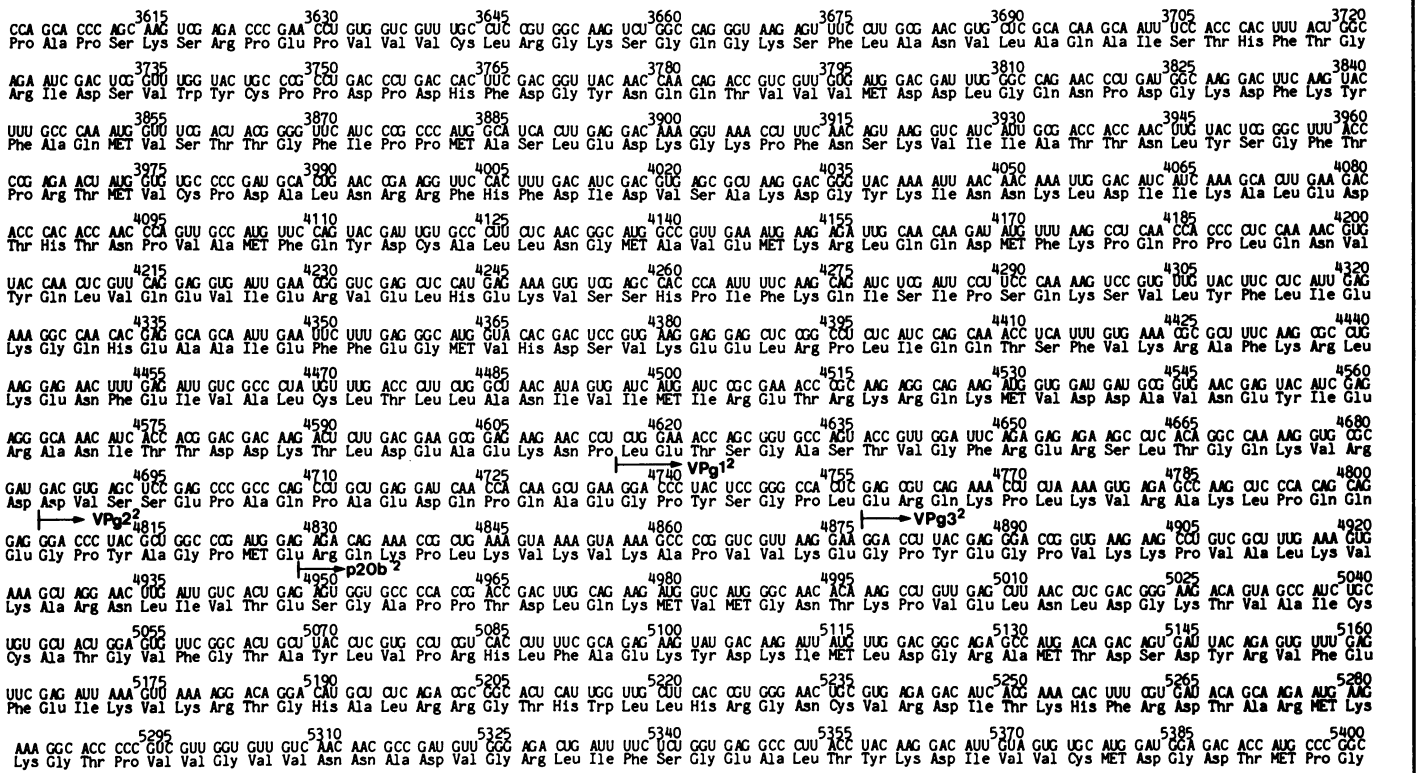
several independent sequencing experiments.

Previous mapping studies using defined T_1 oligonucleotide probes had indicated that the clone pFA206 did not include sequences extending to the poly (A) at the 3' end of the genome (7). The sequence data show however that this clone does in fact contain the whole of the 3' end of the genome including 33 bases of the poly (A) tract.

None of the cDNA clones hybridised with T_1 oligonucleotides derived from the 5' end of the genome suggesting that this region was not represented in our clone banks. In order to obtain 5' end sequence beyond the region represented in the clones a primer extension method was employed. A synthetic oligonucleotide (Cell Tech Ltd., UK) whose sequence was derived from the 5' proximal region of pFA76 (see Figure 1) was used to prime cDNA transcripts which were sequenced as described in methods. Approximately 450 bases were sequenced using this primer.


p20a
p16
VP4
VP2'
VP3'


VP1'
p52'


VPg1²
VPg2²
VPg3²
p20b²

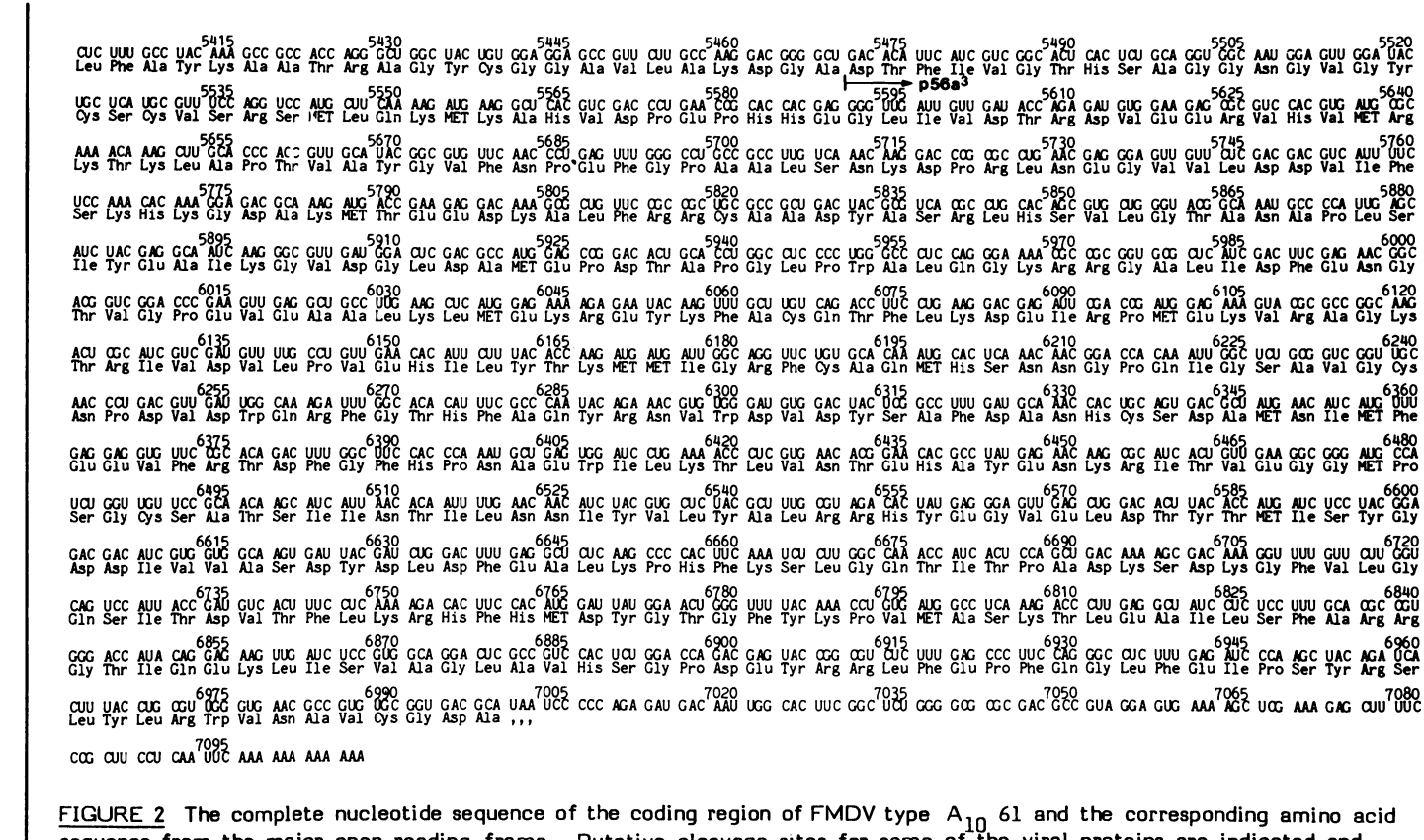


FIGURE 2 The complete nucleotide sequence of the coding region of FMDV type A_{10} 61 and the corresponding amino acid sequence from the major open reading frame. Putative cleavage sites for some of the viral proteins are indicated and those for which there is supportive amino acid sequence data are indicated with a superfix referring to the following references: 1) 11, and references therein; 2) 2, 21. 3) 14.

Nucleotide Sequence

The nucleotide sequence determined from the recombinant clones and by primer extension sequencing is shown in Figure 2. The sequence includes an open translational reading frame of 2,333 codons terminating at a UAA codon 93 bases from the 3' poly (A) sequence. Two lines of evidence suggest that this represents the sequence coding for the FMDV polyprotein. Firstly, no other reading frame of significant length is found in the sequence. Secondly, protein sequence data is available for several viral proteins (11 and refs. therein, 14) and these sequences correlate with the major open reading frame. Further sequence towards the 5' end from the major open reading frame has termination codons in all three reading frames (15).

Examination of our sequence gels allows us to estimate the poly (C) tract to be at least 500 nucleotides from the initiation site for translation. This would mean that FMDV RNA has a 5' untranslated region in excess of 1,000 nucleotides long and a total genome length of at least 8,100 nucleotides.

The nucleotide composition and dinucleotide frequency of the coding region are consistent with the data previously obtained for the region of the genome coding for the structural proteins (11). In particular the base composition shows a bias in favour of C+G over A+U. This is also reflected in chemical analysis of the RNA (16). The relatively high frequency of CpG is also maintained throughout the genome being 83% of the expected frequency compared to 37% for eukaryotes in general (17) and 47% for poliovirus (18).

Predicted Amino Acid Sequence

The processing of picornaviral proteins is a complex process and appears highly variable between different members of the family. However, the general genomic organisation and the primary cleavage products derived from the polyprotein are similar, (1, 19) with one important exception. The region to the 5' side of the genome coding for the structural genes in FMDV and encephalomyocarditis virus code for additional proteins which have no identifiable counterparts in poliovirus. It has been shown that two proteins are coded for in this part of the genome in FMDV; namely p20a (Lab) and p16 (Lb) (for polypeptide nomenclature see Figure 1). Analysis of these proteins has shown them to have similar tryptic peptide maps and limited proteolysis patterns (6, 15). Both of these proteins can be labelled *in vitro* using N-formyl (^{35}-S) methionine t RNA_f^{met} (6) and studies involving a protease inhibitor suggest that the C-termini of these proteins are the same (20). These data suggest that there are two initiation sites in FMDV RNA. Examination of the sequence data presented here shows the presence of a second AUG codon located at nucleotide position 85 and in the same reading

frame as the initial AUG at position 1 (see Figure 2). Initiation at these two sites would result in two similar proteins with a molecular weight difference of 3,800, i.e. the difference in molecular weight between p20a (Lab) and p16 (Lb). Beck et al (21) have also recently reported the presence of these two putative initiation sites in two other serotypes of the virus.

Although no protein sequence data are available to confirm the predicted sequence of the three VPg proteins, their charge, amino acid composition and tryptic peptide composition (2) are entirely consistent with the sequences presented here. These data also agree with those of Forss and Schaller (22) on the predicted sequence of the VPgs from FMDV serotypes O and C. Although in general FMDV and polio show a high degree of similarity in genome organisation the VPg genes are an example where they differ significantly. Polio has a single VPg gene whereas FMDV has three, which, moreover are highly conserved in three serotypes (22). All three VPgs from FMDV are equally represented in RNA extracted from virus particles (2).

The poliovirus VPg has been found in a precursor molecule of molecular weight 12,000 of which the VPg protein itself is contained in the C-terminal portion. The non-VPg sequences in this precursor contain a hydrophobic region of 22 amino acids situated 7 amino acids to the N-terminal side of VPg. This hydrophobic area is thought to be responsible for the association of the VPg precursor with cellular membranes, since the VPg precursor is found in membrane bound complexes (23). Comparison of hydrophilicity plots representing the VPg region from FMDV and polio show no hydrophobic region in FMDV comparable to that in polio. The functional precursor, (if any), containing the FMDV VPgs is at present unknown.

Comparison of our sequence with the recently published sequence of the FMDV A_{12} polymerase (p56a; 3d) (14) show them to be very similar. There are 16 amino acid changes (out of 470 total) which are distributed throughout the protein.

Cleavage sites

The proteolytic cleavage sites involved in the processing of the poliovirus polyprotein into the final products show a high degree of uniformity not seen with FMDV. In polio eight of the identified cleavage sites are between gln-gly pairs; of the remaining cleavages one is between tyr-gly and one is between asn-ser (24). The latter cleavage is between VP4 and VP2 and occurs during the final stages of virus maturation. With FMDV, on the other hand, little homology is apparent in the sequences around the known cleavage sites. The possible cleavage sites involved in processing the polyprotein of FMDV are indicated in Figure 2. Those sites for which some amino acid sequence data are available are also indicated. Both host

and viral specified proteases are involved in the processing of FMDV proteins (20).

The cleavages thought to be caused by a viral enzyme (secondary cleavages) indicate a preference for glu-gly bonds, for example between VP2 and VP3, at the N termini of the VPgs and between p20b (3c) and p56a (3d, polymerase). Of the twelve putative cleavage sites (both primary and secondary) five occur between glutamate and glycine.

The putative host specified cleavage site (primary cleavage) between p20a/p16 and p88 (L and P1) has been revised from previous reports (7). This cleavage is now proposed to occur after amino acid 204 between gly and gln on the basis of homology between FMDV and EMC VP4 proteins (A. Palmenberg, personal communication), and our recent observation that FMDV VP4 contains proline (15). With the revised cleavage site VP4 contains a single proline at amino acid position 208 (see Figure 2). The N-terminus of VP4 is known to be refractory to Edman degradation in all picornaviruses which have been examined (11, 19, 24, 25) and with FMDV and EMC this blockage may be due to deamination and cyclisation of a glutamine.

CONCLUDING REMARKS

At present we do not have clones which span the poly (C) tract and so it has not been possible to analyse the sequence of the 5' non coding region of the FMDV genome. The reason for the lack of 5' end clones is not clear. It could be simply due to the low frequency of cDNA transcripts which extend to the extreme 5' end, or alternatively the unusual structure of the poly (C) tract may be unstable in *E. coli* and eliminated during plasmid replication. Direct RNA sequencing of the extreme 5' end of the genome has been reported (26) and shows that considerable homology exists between the first 27 nucleotides from all seven serotypes of FMDV. This suggests that this region has a critical role in virus replication.

Sequence data from the 3' untranslated region of FMDV have also been reported (14, 27). Porter *et al* (27) analysed the nucleotide sequence adjacent to the poly (A) for five serotypes and found ≈75% homology in the first 20 nucleotides and a highly conserved sequence of eleven nucleotides at position -7 to -17 from the poly (A) tract. One of these serotypes was identical to the virus used in this study and the sequence of 42 nucleotides reported agrees exactly with that reported here. Recently the 3' sequence of a second type A virus (A_{12}) has been reported (14) and maintains the sequence of the eleven highly conserved nucleotides, however the remainder of the sequence up to the poly (A) is considerably different and includes three additional nucleotides. In the A_{12} virus there is a single UAA termination codon located 96 nucleotides from the poly (A)

tract in contrast to the single UAA termination codon located 93 nucleotides from the poly (A) reported here from the A_{10} virus.

A knowledge of the coding sequence of a virus is a necessary pre-requisite for the full understanding of the functions of the encoded proteins in virus structure and replication. Moreover, comparisons of the complete sequences of different viruses from within the picornavirus family will indicate evolutionary relationships between the virus groups. Finally the sequence data is essential for experiments involving site specific mutagenesis of an infectious clone of virus RNA (28) in order to modify specific regions of the genome.

ACKNOWLEDGEMENTS

We would like to thank Dr A Makoff for computer analysis and Dr F Brown for helpful discussion and critical reading of the manuscript. We would like to thank Dr M Eaton (Cell Tech Ltd, UK) for kindly providing the synthetic oligonucleotide primer.

*Present address: Wellcome Biotechnology Limited, Ash Road, Pirbright, Woking, Surrey, UK

REFERENCES

1. Sangar DV. (1979) J. Gen. Virol. 45, 1-13.
2. King AMQ, Sangar DV, Harris TJR and Brown F. (1980) J. Virol. 34, 627-634.
3. Sangar DV, Rowlands DJ, Harris TJR and Brown F. (1977) Nature 268, 648-650.
4. Flanegan JB, Petersson RF, Ambros V, Hewlett MJ and Baltimore D. (1977) PNAS USA 74, 961-965.
5. Nomoto A, Detjen B, Pozzatti R and Wimmer E. (1977) Nature 268, 208-213.
6. Sangar DV, Black DN, Rowlands DJ, Harris TJR and Brown F. (1980) J. Virol. 33, 59-68.
7. Boothroyd JC, Highfield PE, Cross GAM, Rowlands DJ, Lowe PA, Brown F and Harris TJR. (1981) Nature 290, 800-802.
8. Wensink PC, Finnegan DJ, Donelson JE and Hogness DS. (1974) Cell 3, 315-325.
9. Woods DE, Crampton JM, Clarke BE and Williamson R. (1980) Nucleic Acids Research 8, 5157-5168.
10. Maxam A and Gilbert W. (1980) in Methods in Enzymology Vol. 65, pp499-560, Grossman L and Moldave K (Eds), Academic Press, N.Y.
11. Boothroyd JC, Harris TJR, Rowlands DJ and Lowe PA. (1982) Gene 17, 153-161.
12. Rowlands DJ, Clarke BE, Carroll AR, Brown F, Nicholson BH, Bittle JL, Houghten RA and Lerner RA. (1983) Nature 306, 694-697.
13. Makoff AJ, Paynter CA, Rowlands DJ and Boothroyd JC. (1982) Nucleic Acids Research 10, 8285-8295.
14. Robertson BH, Morgan DO, Moore DM, Grubman MJ, Card J, Fischer T, Weddell G, Dowbenko D and Yansura D. (1983) Virology 126, 614-623.
15. Clarke BE (1984) Manuscript in preparation.
16. Newman JFE, Rowlands DJ and Brown F. (1973) J. Gen. Virol. 18, 171-180.
17. Nussinov R. (1981) J. Mol. Biol. 149, 125-131.
18. Racaniello VR and Baltimore D. (1981) PNAS USA 78, 4887-4891.
19. Rueckert RR. (1976) in Comprehensive Virology, Fraenkel-Conrat H and

Wagner RR (Eds) Vol. 6 pp131-213, Plenum Press, New York.
20. Burroughs JN, Sangar DV, Clarke BE and Rowlands DJ. (1984) J. Virol. In press.
21. Beck E, Forss S, Strebel K, Cattaneo R and Feil G. (1983) Nucleic Acids Research 11, 7873-7885.
22. Forss S and Schaller H. (1982) Nucleic Acids Research 10, 6441-6450.
23. Semler BL, Anderson CW, Kitamura N, Rotherberg PG, Wishart WL and Wimmer E. (1981) PNAS USA 78, 3463-3468.
24. Kitamura N, Semler BL, Rothberg PG, Larsen GR, Adler CJ, Dorner AJ, Emini EA, Hanecak R, Lee JJ, van der Werf S, Anderson CW and Wimmer E. (1981) Nature 291, 547-553.
25. Scraba DG. (1979) in The Molecular Biology of Picornaviruses, Perez-Bercoff R. (Ed) pp1-23, Plenum Press, New York.
26. Harris TJR. (1980) J. Virol. 36, 659-664.
27. Porter AG, Fellner P, Black DN, Rowlands DJ, Harris TJR and Brown F. (1978) Nature 276, 298-301.
28. Racaniello VR and Baltimore D. (1981) Science 214, 916-919.