Volume 10 Number 24 1982 Nucleic Acids Research

Comparison of the amino acid sequence of the major immunogen from three serotypes of foot and mouth disease virus

A.J.Makoff*, C.A.Paynter*, D.J.Rowlands+ and J.C.Boothroyd*§

*Department of Molecular Biology (formerly Immunochemistry), Wellcome Research Laboratories, Langley Court, Beckenham, Kent BR3 3BS and +Department of Biochemistry, Animal Virus Research Institute, Pirbright, Woking, Surrey GU24 0NF, UK

Received 29 September 1982; Revised and Accepted 16 November 1982

ABSTRACT

Cloned cDNA molecules from three serotypes of FMDV have been sequenced around the VP1-coding region. The predicted amino acid sequences for VP1 were compared with the published sequences and variable regions identified. The amino acid sequences were also analysed for hydrophilic regions. Two of the variable regions, numbered 129-160 and 193-204 overlapped hydrophilic regions, and were therefore identified as potentially immunogenic. These regions overlap regions shown by others to be immunogenic.

INTRODUCTION

Foot and mouth disease virus (FMDV) is a picornavirus causing a disease of domestic livestock which is of considerable economic importance (1). A complicating factor in control of the virus by vaccination is the existence of seven non-crossprotective serotypes:three European, A, O and C, three South African, SAT 1, SAT 2 and SAT 3, and one Asian, Asia 1. In addition each serotype is composed of a number of serologically related subtypes which are only partially cross-protective (2). With the development of recombinant DNA technology and rapid DNA sequencing, it is now possible to study in detail the extent of variation between the subtypes and serotypes. This is of particular relevance when applied to one of the major capsid polypeptides, VP1. This protein has been shown to be the major immunogen of the virus (3,4) and as such variation in its structure is likely to be responsible, at least in part, for the existence of the serotypes and subtypes. The complete primary structure (as predicted from the nucleotide sequence) of VP1 has recently been reported for two subtypes of serotype A [$A_{10}61$ (5) $A_{12}119$ (6)] and one of serotype O [O_1Kaufbeuren (O_1K) (7)]. We report here the construction of recombinant plasmids containing cDNA coding for VP1 for three more strains: A_{24}Cruzeiro, O_1 British Field Strain (O_1BFS) and C_3Indaial. Together with the previously published data, these results allow comparisons to be made between representatives of the three major European serotypes ($A_{10}61$, O_1BFS and C_3 Indaial), between three subtypes of the same serotype ($A_{10}61$, $A_{12}119$ and

0305-1048/82/1024-8285$2.00/0

A_{24}Cruzeiro), and between two isolates of the same subtype (O_1BFS and O_1K). The conclusions on the distribution of the regions of variability are discussed and compared with the recent results of Strohmaier et al (8) on the antigenicity of peptide fragments of VP1 from purified virus and those of Bittle et al (9) and Pfaff et al (10) on the ability of short synthetic peptides, corresponding to parts of VP1, to elicit the production of neutralizing antibody.

MATERIALS AND METHODS

(a) Construction of recombinants

Three strains of FMDV were used in this work:-A_{24}Cruzeiro, C_3Indaial and O_1BFS. The FMDV RNA was reverse transcribed using either oligo dT (11) or calf thymus DNA oligonucleotides (12) as primer. [The calf thymus DNA primed reverse transcription would be expected to generate cDNA molecules randomly distributed throughout the FMDV genome as distinct from the oligo dT primed reverse transcription which has tended to bias the cDNA population to the 3' end (11)]. The resulting double stranded cDNA was inserted into the Pst I site of the plasmid pAT153 by G:C tailing as described by Boothroyd et al (11). The recombinant plasmids were used to transform competent E. coli HB101 cells.

(b) Screening of recombinants

The transformed cells were screened by the colony hybridization procedure described by Hanahan and Meselson (13). The C_3 and A_{24} recombinant plasmid transformed cells were probed with a Pvu II fragment of $pFA_{10}t$ 76 (5) which contains DNA sequences derived from $A_{10}61$, covering most of the regions coding for VP3 and VP1. The O_1 transformed cells were probed with a small Rsa I/HinfI fragment of pFO_1t351 [Fig 1(c)] which was derived from the plasmid pFO_1t251 (11), and comprises mainly the 3' end of the VP1 coding sequence.

The positive colonies were further screened by restriction mapping of their plasmid DNA. These maps were aligned with a linear map of the protein coding units of the FMDV genome, by analogy with the cDNA sequences of $A_{10}61$ (5) of $A_{12}119$ (6) and of O_1K (7). In some cases, Southern blotting of appropriate restriction fragments (14) was necessary to resolve ambiguities.

(c) Sequencing

All DNA sequencing was done by the method of Maxam and Gilbert (15) as described in Boothroyd et al (5).

(d) Analysis of Sequence Data

An Apple II microcomputer was used to search DNA sequences for restriction sites and to predict the corresponding protein sequences, using a programme written by Larson and Messing (16). It was also used to make hydrophilicity plots

using the programme in Kyte and Doolittle (17). These programmes were modified by H.D. Spence, who also wrote the programme which was used for the variability plot. A PDP 11 minicomputer was used to determine dinucleotide frequencies and codon usage, using programmes written by Staden (18, 19).

RESULTS AND DISCUSSION

VP1-Coding Sequences of the Three Serotypes

(i) $\underline{A}_{24}$. None of the colonies transformed with plasmids containing either random-primed or oligo dT-primed cDNA appeared to contain the entire VP1-coding region of A_{24}. One plasmid, $pFA_{24}t17$, from the oligo dT-primed clone bank, appeared to extend into the VP1 coding region from the 3' end but not quite as far as the 5' end. The VP1 region of this plasmid was sequenced using the strategy indicated in Fig. 1(a). The DNA sequence and predicted amino acid sequence are given in Fig. 2(a). Eleven codons at the 5' end of the VP1-coding region are presumed missing in this plasmid.

(ii) $\underline{C}_3$. The plasmid pFC_3r3, derived from the random-primed cDNA from C_3

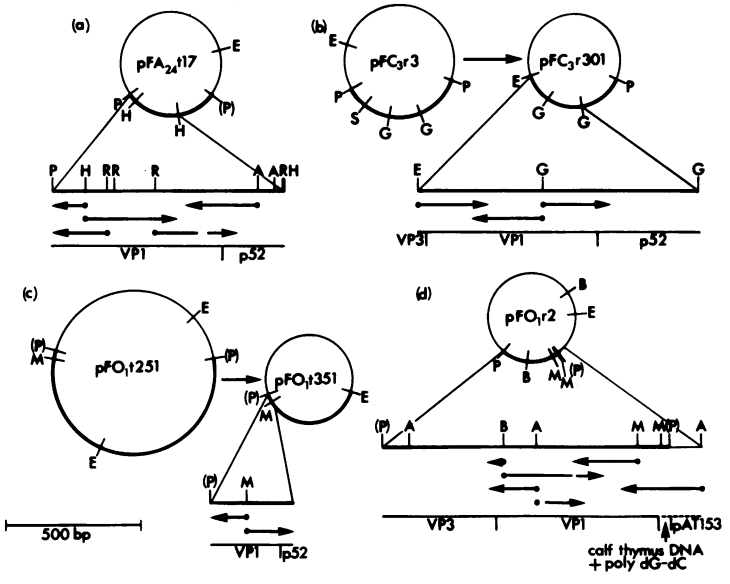
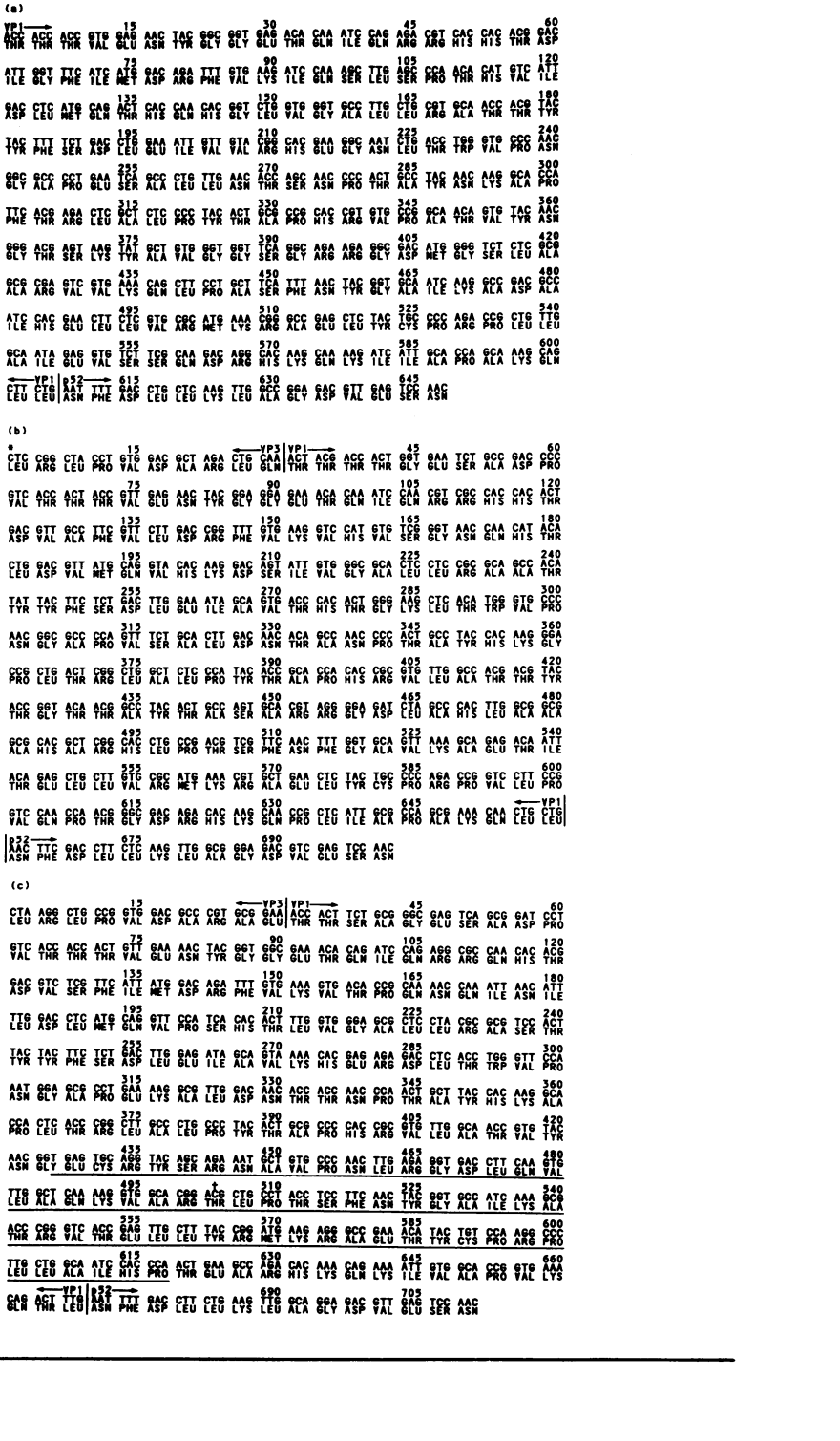


Fig. 1. Physical maps of the plasmids and sequencing strategy used. The fragments used for sequencing are indicated by •⟶ where • indicates the labelled end. Restriction enzyme sites used to generate the labelled end are shown. The abbreviations for the sites are:- A, Ava II; B, Bam HI; E, EcoRI; G, Bgl II; H, Hind III; M, Sma I; P, Pst I; (P), expected Pst I site which was lost presumably by slight exonuclease contamination; R, Rsa I, S, Sst I. The polypeptides encoded by each region are shown.



appeared to have all of VP1 and some of the flanking regions as well. DNA between the unique Sst I site (situated at the 3' end of the VP3 coding region) and the EcoRI site of the vector was removed. The remainder of the plasmid was treated with DNA polymerase I large fragment to repair the EcoRI cohesive end and to hydrolyse the Sst I cohesive end. The blunt ends were ligated together to recreate an EcoRI site generating the plasmid pFC_3r301 [Fig. 1(b)]. The VP1-coding region of this plasmid was sequenced using the strategy shown. The DNA sequence and predicted amino acid sequence are given in Fig. 2(b).

The first base shown, cytosine, was actually read as a thymine in the fragment labelled at the EcoRI site of pFC_3r301 and is the fourth base of the EcoRI recognition sequence GAATTC. In the parental plasmid pFC_3r3, there was an Sst I site at this position reflecting the original sequence in the FMDV genome and therefore the corresponding hexamer would be the Sst I recognition sequence GAGCTC.

(iii) O_1. The plasmid pFO_1t251 was reported previously (10) and has an insert of 4.8 Kb, one end of which is very near to the 3' end of the FMDV genome. Restriction mapping suggested that the other end extends some way into the VP1-coding region. The plasmid was subcloned by removing the DNA between the two EcoRI sites and ligating the two remaining EcoRI cohesive ends, recreating an EcoRI site and generating the smaller plasmid pFO_1t351 [Fig. 1(c)]. The VP1-coding region was sequenced using the strategy shown in the figure. The plasmid contains 243 bp at the 3' end of the VP1 coding sequence [see below and Fig. 2 (c)].

The plasmid pFO_1r2 derived from the random-primed cDNA from O_1 has DNA from the 5' end of the VP1-coding region of FMDV genome. It would be expected to overlap with the VP1-coding DNA in pFO_1t351 since this DNA was used as the probe during colony hybridization with which there was a positive response. The presence of the same Sma I site in both plasmids and in the O_1K sequence (7) confirmed this. However, the presence of an additional Sma I site approximately 80 bp downstream of the first Sma I site in pFO_1r2 does not correspond to a site in either O_1BFS or O_1K [Figs. 1(c),(d)]. [Sequencing showed that this Sma I site was located on a 21 bp stretch of DNA between the FMDV sequence and the artificially introduced poly dC tract (data not shown). This 21 bp sequence does not correspond

Fig. 2. Nucleotide and predicted amino acid sequences for the VP1 regions of (a) A_{24} Cruzeiro, (b) C_3Indaial, (c) O_1BFS.
(a) The first eleven codons of VP1 in A_{24} are presumed missing.
(b) * The C was read as a T in the plasmid pFC_3301 but is presumed to be a C in the parental plasmid pFC_3r3 (see text).
(c) The underlined sequence was derived from both plasmids pFO_1t351 and pFO_1r2. †The C was derived from pFO_1r2 and was replaced by a T in pFO_1t351, which would predict a methionine residue (see text).

to any known O_1BFS sequence, and since it is at the 3' end of the FMDV insert was presumed to be calf thymus DNA resulting from the priming of reverse transcriptase in the original cDNA preparation.]

The relevant region of pFO_1r2 was sequenced according to the strategy shown in Fig. 1(d). The FMDV insert does not extend as far as the 3' end of the VP1 coding region, and therefore the O_1 sequences from both pFO_1r2 and pFO_1351 have been combined to give the complete VP1 coding sequence presented in Fig. 2(c). The region from bases 426 to 618 is common to both plasmids and is underlined. There is a discrepancy between the two plasmids. Position 503 is cytosine in pFO_1r2 and thymine in pFO_1t351. Since the O_1K sequence has a cytosine at this position and replacement by a thymine leads to a threonine being replaced by a methionine in the amino acid sequence it was thought more likely that the original base in the virus was a cytosine. [This discrepancy could have arisen at a number of places: in the FMDV RNA which was obtained from an uncloned preparation, in the cDNA during reverse transcription or in the plasmid during replication of the bacterium. We are unable to comment on which possibility is the most likely.]

Codon usage and dinucleotide frequencies

Both codon usage and dinucleotide frequencies were determined for the three new sequences (data not shown). The codon usage showed a similar bias of C or G over U or A in the third position as reported by Boothroyd et al (5) for A_{10} and Kurz et al (7) for O_1K. The dinucleotide frequencies were very similar to those found in A_{10} (5).

Comparison of nucleotide sequences

The nucleotide sequences from Fig. 2 were compared with the three published sequences for A_{10} (5), A_{12} (6) and O_1K (7). The variation was very similar to that found when comparing the predicted amino acid sequences (see below). Since the variation in the amino acid sequences is more likely to be biologically significant our analysis has been confined to this variation.

Comparison of amino acid sequences

The three predicted amino acid sequences from Fig. 2 were compared with the three published sequences. They were aligned to give maximum homology which necessitated the introduction of gaps (indicated by dashes) but these were kept to a minimum (Fig. 3). They are arranged to enable comparisons to be made between (b) different serotypes (A_{10}, C_3, O_1BFS), (a) different subtypes of the same serotype (A_{10}, A_{12}, A_{24}) and (c) different isolates of the same subtype (O_1BFS, O_1K). The variations are shown by boxes being drawn around all amino acids which differ from the others within the comparison group. Some of the flanking sequence is included, but only VP1 is numbered. In order to maintain consistency between

Fig. 3. Comparison of the predicted amino acid sequences between (a) A_{12}, A_{24} and A_{10}, (b) A_{10}, C_3 and O_1BFS, (c) O_1BFS and O_1K. The amino acids were aligned for maximum homology with the minimum introduction of gaps, each indicated by a -. Each position in VP1 is numbered from 1 to 216, including the gaps. All differences within a comparison group are boxed. † The threonine residue was predicted in pFO_1r2 but a methionine was predicted in pFO_1t351 (see legend to Fig. 2(c) and text). The abbreviations for the amino acids are:- A, ala; C, cys; D, asp; E, glu; F, phe; G, gly; H, his; I, ile; K, lys; L, leu; M, met; N, asn; P, pro; Q, gln; R, arg; S, ser; T, thr; V, val; W, trp; Y, tyr.

different sequences, the gaps have been included in the numbering system.

As expected, the variation between different serotypes is much greater than the variation between different subtypes. The levels of amino acid sequence homology for VP1 between the different serotypes are 72% ($A_{10}vC_3$), 69% (C_3vO_1) and 69% (O_1vA_{10}). The levels of homology between different subtypes are 88% ($A_{12}vA_{24}$), 87% ($A_{24}vA_{10}$) and 88% ($A_{10}vA_{12}$) (assuming that the missing amino acids at the N terminus of A_{24} vary from A_{10} as much as the analogous region does between A_{10} and A_{12}). These values are similar to those obtained by comparing the nucleotide sequences of the corresponding VP1-coding regions (data not shown). Both sets of values are similar to the levels of nucleotide sequence homology for the entire genome, estimated by competition hybridization between different serotypes and subtypes of FMDV (20). The O_1 sequence is highly conserved between the two

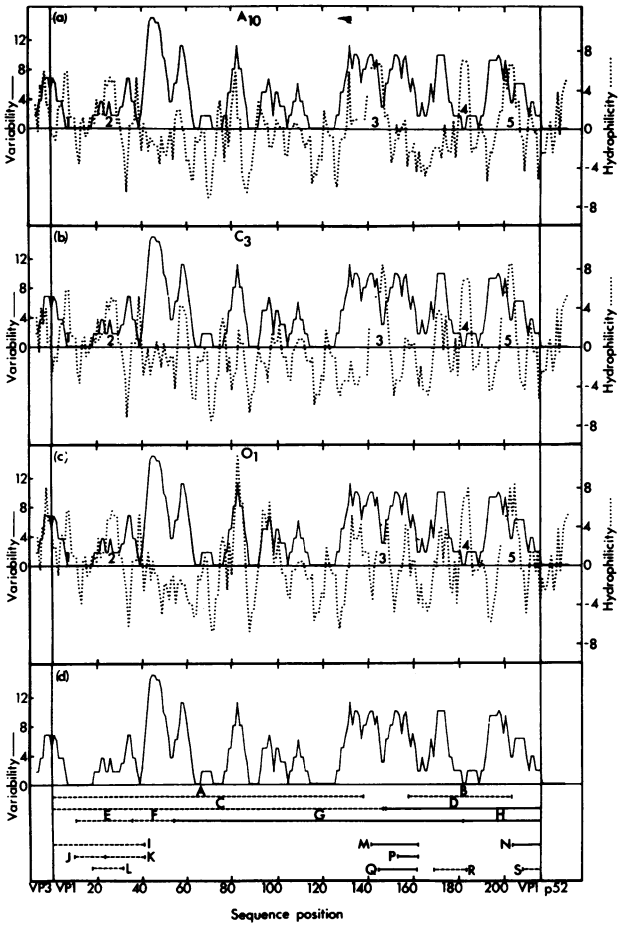


Fig. 4. Variability of A_{10}61, C_3Indaial and O_1BFS plotted with (a-c) hydrophilicity of A_{10}, C_3 or O_1 or with (d) linear representations of peptides tested for immunogenicity (8-10).
(a-c) Both type of plot is generated by combining the scores of five positions in the sequence, each point representing the middle (i.e. third) position of each group of five. The variability plot (——) uses a score of 0 for every identical comparison for each position and a score of 1 for every different comparison for each position. This gives a maximum score of 3 for each position and 15 for each group of 5 positions. The hydrophilicity plots (·····) use the scoring system of Hopp and Woods (22) for each amino acid, the scores ranging from -3.4 to +3.0. This gives a maximum score of 15 and a minimum score of -17 for each group of 5 positions. In order to align the two plots, gaps were introduced in the hydrophilicity plots corresponding to gaps in the appropriate sequence in in Fig. 3 which were used in the variability plot. The major hydrophilicity peaks are numbered 1-5 (see text).
(d) ——immunogenic peptide,--- non-immunogenic peptide. (A, B) peptides generated by trypsin (8); (C, D) peptides generated by mouse submaxillary gland protease (8); (E-H) peptides generated by cyanogen bromide (8). (I-S) peptides produced by chemical synthesis (9, 10).

different isolates O_1K and O_1BFS. The differences were limited to three amino acids and seven nucleotides (99% homology in both cases).

The variation (particularly between different serotypes) is not uniformly distributed throughout the sequence but concentrated in a number of discrete regions. The virus is under strong selective pressure to evade the host immune system and so variation in a number of places is not surprising. Four classes of variation would be expected:-

(1) Variation in the immunogenic sites.

(2) Variation elsewhere leading to conformational changes in the immunogenic sites.

(3) Variation elsewhere necessary to compensate for any deleterious conformational changes due to (1) or (2).

(4) Variation due to random genetic drift.

The variable regions which appear the most significant are:- 42-51, 56-61, 79-85, 95-102, 129-147, 151-160, 170-176, 193-204. Some of these regions would be expected to contain immunogenic regions.

Hydrophilicity of VP1

It has been demonstrated that antigenic determinants are found on the surface of proteins (21, 22). There are methods available for predicting surface regions of proteins from the primary sequence, based on the solvent partition of each amino acid or calculations from crystallographic data or both (17, 23, 24). The hydrophilicity values of Hopp and Woods (23) were applied to the VP1 sequences of A_{10}, C_3 and O_1BFS. These values were preferred since they have been successfully applied in predicting the majority of antigenic determinants of twelve proteins (23) and do not suffer from the disadvantage of subjective adjustments which were made by Kyte and Doolittle (17). Fig. 4(a-c) shows the results for the three sequences using overlapping groups of 5 amino acids (dotted lines) plotted on the same axes as a plot of variability between A_{10}, C_3 and O_1BFS (solid lines). In order to obtain a meaningful alignment between the variability plot and each hydrophilicity plot, gaps were introduced in the hydrophilicity plots coresponding to the gaps in the sequences in Fig. 3.

There are several major hydrophilic regions common to all 3 serotypes, and some which are only hydrophilic for 1 or 2 serotypes. Since gross morphological differences between the serotypes are unlikely, only those hydrophilic regions which are common to all 3 serotypes have been considered. Consequently the number of potentially surface regions is a minimum estimate. However, some of these regions might be located on the inner surface of the virus capsid.

Peaks 1, 2 and 4 are all in conserved regions. Peak 3 overlaps both variable regions 129-147 and 151-160, and therefore the combined region 129-160 might be

expected to contain at least one immunogenic region. Peak 5 overlaps some of the variable region 193-204 and so this region is also likely to be immunogenic. It is interesting that the most variable region, 42-51 does not overlap a hydrophilic region in any of the 3 serotypes sequenced. This strongly suggests that this region has variation of either class (2) or (3) outlined above.

Hydrophilic plots were also made for the same three sequences but using the values of Kyte and Doolittle (17). Although there were differences in detail from the above analysis, the conclusions about the major hydrophilic regions were the same (data not shown).

Immunogenic Regions

We have identified two regions 129-160 and 193-204 which are both variable and overlap hydrophilic regions. We have argued that this makes them likely to be immunogenic. This conclusion provides a rational basis for the observations made in related work of others. Strohmaier et al (8) treated intact virus particles (strain O_1K) with proteolytic enzymes. Only VP1 was cleaved and the resulting peptides were purified and tested for their ability to raise neutralizing antibodies. The peptides (A-D) are represented on Fig. 4(d) as solid lines if immunogenic and dashed lines if non-immunogenic. These workers also cleaved isolated VP1 with cyanogen bromide, generating 4 peptides (E-H). Their results suggest the presence of at least two immunogenic regions: one between 55 and 182 and one between 183 and 216 (using our numbering system). The authors implied that 147-156 and 204-216 were likely to define the immunogenic regions more precisely because they did not overlap non-immunogenic peptides. We think this interpretation is an oversimplification because it ignores the possibility that an immunogenic region may behave non-immunogenically if the peptide containing it fails to adopt the same conformation as in vivo. Also it is conceivable that a part of an immunogenic region may not be immunogenic when it is cleaved from the rest of the site.

Bittle et al (9) chemically synthesized various peptides [Fig. 4(d), I-P] using the published sequence from O_1K. Three peptides M-P were immunogenic; M giving a much higher response than the other two. Since both peptides M (141-162) and P (153-162) lie almost completely within the variable region 129-160 and P lies within M, the finding that M is much more immunogenic than P is entirely consistent with our conclusions. Peptide N (203-216) barely overlaps the variable region 193-204, but both lie within peptide H (183-216) found to be immunogenic by Strohmaier et al (8). Our conclusions would therefore predict that a peptide extending further towards the N-terminal end of VP1 would be more immunogenic than N.

Pfaff et al (10) also chemically synthesized peptides [Fig 4(d), Q-S] using the published sequence from O_1K. Peptide Q (144-161) which is almost identical to

peptide M (141-162) of Bittle et al (9), also raised neutralizing antibodies. Peptide S (208-216) is much smaller than the poorly immunogenic peptide N (203-216) and, unlike N, does not overlap the variable region 193-204. Peptide R (169-183) overlaps the variable region 170-176, which is only slightly hydrophilic in C_3 and O_1 and not at all in A_{10}. It also overlaps the hydrophilic peak 4 which is not in a variable region [Fig. 4]. The failure of both peptides R and S to raise neutralizing antibodies is consistent with our conclusions.

ACKNOWLEDGEMENTS

We would like to thank L.P. Dicken for technical assistance with much of the sequencing, H.D. Spence for computer programming and P.E. Highfield for subcloning plasmid pFO_1t251. We would also like to thank F. Brown and P.E. Highfield for helpful comments.

§Present address: Department of Medical Microbiology, Stanford University, Stanford, CA 94305, USA

REFERENCES

1. Sangar, D.V. (1979) J. gen Virol 45, 1-13.
2. Pereira, H.G. (1977) International Symposium on foot-and-mouth disease. Develop. biol. Standard, vol. 35, pp 167-174.
3. Wild, T.F., Burroughs, J.N. and Brown, F. (1969) J. gen. Virol. 4, 313-320.
4. Burroughs, J.N., Rowlands, D.J., Sangar, D.V., Talbot, P. and Brown, F. (1971) J. gen. Virol. 13, 73-84.
5. Boothroyd, J.C., Harris, T.J.R., Rowlands, D.J. and Lowe, P.A. (1982) Gene 17, 153-161.
6. Kleid, D.G., Yansura, D., Small, B., Dowbenko, D., Moore, D.M., Grubman, M.J., McKercher, P.D., Morgan, D.O., Robertson, B.H. and Bachrach, H.L. (1981) Science 214, 1125-1129.
7. Kurz, C., Forss, S., Küpper, H., Strohmaier, K. and Schaller, H. (1981) Nucleic Acids Research 9, 1919-1931.
8. Strohmaier, K., Franze, R. and Adam, K.-H. (1982) J. gen. Virol. 59, 295-306.
9. Bittle, J.L. Houghten, R.A., Alexander, H., Shinnick, T.M., Sutcliffe, J.G., Lerner, R.A., Rowlands, D.J. and Brown, F. (1982) Nature 298, 30-33.
10. Pfaff, E., Mussgay, M., Böhm, H.O., Schultz, G.E. and Schaller, H. (1982) The EMBO J. 1, 869-874.
11. Boothroyd, J.C., Highfield, P.E., Cross, G.A.M., Rowlands, D.J., Lowe, P.A., Brown, F. and Harris, T.J.R. (1981) Nature 290, 800-802.
12. Taylor, J.M., Illmensee, R. and Summers, J. (1976) Biochim. Biophys. Acta 442, 324-330.
13. Hanahan, D. and Meselson, M. (1980) Gene 10, 63-67.
14. Southern, E.M. (1975) J. mol. Biol. 98, 503-517.
15. Maxam, A. and Gilbert, W. (1980) in Methods in Enzymology Vol. 65, pp. 499-560, Grossman, L. and Moldave, K. (Eds.), Academic Press, N.Y.
16. Larson, R. and Messing, J. (1982) Nucleic Acids Research 10, 39-49.
17. Kyte, J. and Doolittle, R.F. (1982) J. mol. Biol. 157, 105-132.
18. Staden, R. (1977) Nucleic Acids Research 4, 4037-4051.
19. Staden, R. (1980) Nucleic Acids Research 8, 3673-3694.
20. Robson, K.J.H., Harris, T.J.R. and Brown, F. (1977) J. gen. Virol. 37, 271-276.
21. Atassi, M.Z. (1975) Immunochemistry 12, 423-438.
22. Reichlin, M. (1975) Adv. Immunol. 20, 71-123.
23. Hopp, T.P. and Woods, K.R. (1981) Proc. Natl. Acad. Sci. U.S.A. 78, 3824-3828.
24. Chothia C. (1976) J. mol. Biol. 105, 1-14.