INFECTION AND IMMUNITY, Mar. 1992, p. 1254–1257
0019-9567/92/031254-04$02.00/0

Vol. 60, No. 3

Colonization Factor Antigen CFA/IV (PCF8775) of Human Enterotoxigenic *Escherichia coli*: Nucleotide Sequence of the CS5 Determinant

CHRISTOPHER A. CLARK,[1] MICHAEL W. HEUZENROEDER,[2] AND PAUL A. MANNING[1]*

Department of Microbiology and Immunology, The University of Adelaide, G.P.O. Box 498,[1] and Division of Clinical Microbiology, Institute of Medical and Veterinary Science,[2] Adelaide, South Australia 5001, Australia

Received 15 July 1991/Accepted 19 November 1991

Human enterotoxigenic *Escherichia coli* isolates expressing the colonization factor antigen CFA/IV (previously designated PCF8775) produce plasmid-encoded CS5 fimbriae. The nucleotide sequence of the region encoding the major CS5 fimbrial subunit was determined. The subunit is synthesized as a precursor of 203 amino acids (20.85 kDa) with a mature protein of 181 amino acids corresponding to a size of 18.6 kDa. The CS5 subunit shows homology to the corresponding component of porcine enterotoxigenic *E. coli* F41, particularly within the signal sequence and at the carboxy terminus.

Enterotoxigenic *Escherichia coli* (ETEC) strains are a major cause of acute infantile diarrhea in developing countries and are associated with traveller's diarrhea (13, 18). These bacteria cause disease by attaching to the small intestinal epithelium, where they release enterotoxins that can be heat labile, heat stable, or both. Colonization is mediated by colonization factor antigens (CFA), which are usually fimbrial in nature (3, 16). ETEC fimbriae are also protective antigens and have potential for use in vaccine development (8, 14, 17).

A number of distinct colonization factors for ETEC associated with human and animal disease have now been characterized. These include CFA/I (2), CFA/II (1a), CFA/III (7), CFA/IV (previously PCF8775) (21, 23), PCFO9 (4), and PCFO159 (20). CFA/II and CFA/IV both represent multiple fimbrial types (19, 22). In particular, CFA/IV produces fimbriae that are antigenically heterogeneous and have been shown to consist of three fimbrial antigens: CS4, CS5, and CS6 (22).

Manning et al. (11) described ETEC belonging to the O115:H40 and O115:H$^-$ serotypes in an outbreak of diarrhea among Australian aborigines. Molecular characterization of the isolates revealed a common 23-kDa protein that correlated with a fimbrial structure seen under the electron microscope. It was shown to be identical to the CS5 fimbriae of CFA/IV (5). In this paper we provide genetic characterization of CS5 and the nucleotide sequence of the major pilin.

The genes for CS5 biosynthesis were originally cloned in the cosmid vector pHC79, generating pPM1306 (5). A further derivative, pPM1312, which was still capable of mediating expression of CS5 fimbriae in *E. coli* K-12, was generated by sequential deletion with *Cla*I and *Mlu*I (5) (Fig. 1).

To determine the extent of the cloned DNA required for the synthesis of CS5 fimbriae, we subjected pPM1312 to

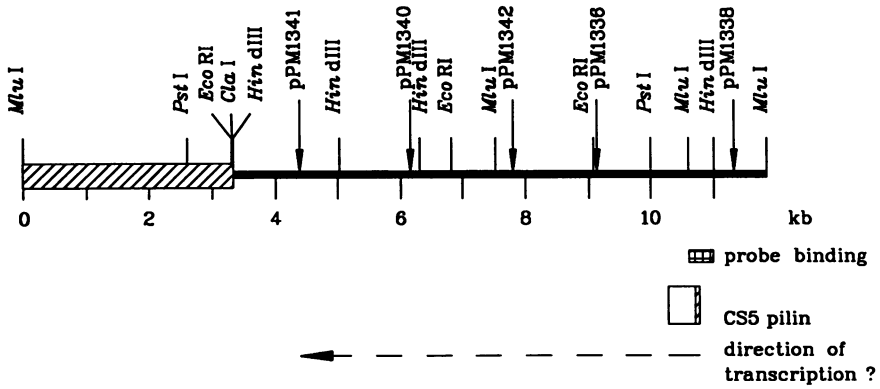

FIG. 1. Physical map of the cloned CS5 determinant. Transposon mutagenesis with pRU669 ($R_{ts}1$::Tn*1725*) (24) was performed as previously described (10). The vertical arrows correspond to the sites of insertion of Tn*1725*, and the relevant plasmids are indicated. The hatched box corresponds to residual vector (pHC79) DNA. The region corresponding to the segment of DNA that bound to a synthetic DNA probe based on the N-terminal sequence of the purified fimbrial subunit is shown. The boxes in the lower part of the figure represent the region of DNA required to encode the indicated proteins. The box corresponding to the CS5 major fimbrial subunit (CS5 pilin) is precisely located based on the DNA sequence data. The N-terminal signal sequence coding region is shown by the cross-hatching.

* Corresponding author.

```
1    TT CAC AAA TAA ATT CGT GTT TAT TGT AAG ACA AAG AAG GAT GAG AAT      47

48   AAA ATG AAG AAA AAT TTA CTG ATA ACT TCA GTG TTG GCA ATG GCA ACC     95
1        Met Lys Lys Asn Leu Leu Ile Thr Ser Val Leu Ala Met Ala Thr     15

96   GTA TCA GGT TCT GTT TTG GCT GCT GTT ACA AAT GGC CAA CTC ACA TTT     143
16   Val Ser Gly Ser Val Leu Ala Ala Val Thr Asn Gly Gln Leu Thr Phe     31

144  AAT TGG CAG GGA GTG GTT CCT TCC GCT CCC GTT ACT CAG AGC AGC TGG     191
32   Asn Trp Gln Gly Val Val Pro Ser Ala Pro Val Thr Gln Ser Ser Trp     47

192  GCT TTT GTG AAC GGA TTG GAT ATA CCG TTT ACT CCT GGT ACT GAA CAG     239
48   Ala Phe Val Asn Gly Leu Asp Ile Pro Phe Thr Pro Gly Thr Glu Gln     63

240  TTG AAT ATC ACC CTT GAT TCA AAT AAA GAT ATC ACG GCC CGT TCG GTT     287
64   Leu Asn Ile Thr Leu Asp Ser Asn Lys Asp Ile Thr Ala Arg Ser Val     79

288  AAG CCT TAT GAT TTT TTC ATT GTT CCA GTT TCT GGA AAC GTA ACT CCT     335
80   Lys Pro Tyr Asp Phe Phe Ile Val Pro Val Ser Gly Asn Val Thr Pro     95

336  GGA GCG CCG GTT ACG CGT GAC ACG TCA GCT AAT ATA AAC AGT GTG AAC     383
96   Gly Ala Pro Val Thr Arg Asp Thr Ser Ala Asn Ile Asn Ser Val Asn    111

384  GCT TTT CTA TCA AGT GTA CCC GTT TCT AAT GGT TTT GTT GGC AAC AAG     431
112  Ala Phe Leu Ser Ser Val Pro Val Ser Asn Gly Phe Val Gly Asn Lys    127

432  CAG TTA ACC CTG AGT ACC GCA GTA GAA GCA GCT AAG GGG GAA GTC GCA     479
128  Gln Leu Thr Leu Ser Thr Ala Val Glu Ala Ala Lys Gly Glu Val Ala    143

480  ATC ACT TTA AAT GGT CAA GCG CTT AAA GTG GGG AGC GCT AGT CCA ACA     527
144  Ile Thr Leu Asn Gly Gln Ala Leu Lys Val Gly Ser Ala Ser Pro Thr    159

528  GTT GTT ACT GTG GCT AGT AAT AAA AAA GAG TCT CAT ATT TCT ATT GAT     575
160  Val Val Thr Val Ala Ser Asn Lys Lys Glu Ser His Ile Ser Ile Asp    175

576  ATG AAT GCC AAG GCA GCT GCT GCG GAT GTG GCA GAG GGG GCA GCT ATT     623
176  Met Asn Ala Lys Ala Ala Ala Ala Asp Val Ala Glu Gly Ala Ala Ile    191

624  AAC TTT GTA GCT CCG GTA ACA TTT GCT GTT GAT ATT TAA TCT GCA TTA     671
192  Asn Phe Val Ala Pro Val Thr Phe Ala Val Asp Ile ***                 204

672  TTT TTA TAC CAA AGG AGG GGG GGG CCC TCC TTT GCC GGA ATA GTT TTT     719

720  ATG AAG ATT CTG TAT TCT TTT TTG TTG TTA CCT TTT TTT TCT TGC GCC     767
1    Met Lys Ile Leu Tyr Ser Phe Leu Leu Leu Pro Phe Phe Ser Cys Ala     16

768  TTC AGT GTT GAT TCA ATG ATA AAG TTT TCA GGC GAA GAT GAC TTT TTT     815
17   Phe Ser Val Asp Ser Met Ile Lys Phe Ser Gly Glu Asp Asp Phe Phe     32

816  CTT GTA AAT GGA AAT AGC AAG GAA AGA GAG TAT ATC TAT GTA ACG CTT     863
33   Leu Val Asn Gly Asn Ser Lys Glu Arg Glu Tyr Ile Tyr Val Thr Leu     48

864  TCT GAA CTA ATT AGC GAG AAA AAC AAT AGG CGC GAT GAA ATA TTT TAC     911
49   Ser Glu Leu Ile Ser Glu Lys Asn Asn Arg Arg Asp Glu Ile Phe Tyr     64
                                   PstI
912  AAC GCA GAC AAT GTG CCT CTA TGG CCT ATA TCT GCA G                  948
65   Asn Ala Asp Asn Val Pro Leu Trp Pro Ile Ser Ala                    76
```

FIG. 2. Nucleotide sequence of the region containing the gene for the CS5 major fimbrial subunit. The nucleotide sequence was generated from cloned DNA fragments in pUC18 or pUC19 by using an Applied Biosystems model 373a automated DNA sequencer. The data were analyzed by using the LKB programs DNASIS V6.0 and PROSIS V6.0, and SeqEd (Applied Biosystems). The CS5 major fimbrial subunit is encoded by the sequence from nucleotides 51 to 659. The relevant *Pst*I and *Mlu*I sites are indicated. The region underlined corresponds to the N-terminal sequence of the purified CS5 major fimbrial subunit as determined by sequential Edman degradations (5); residues corresponding to amino acids 47 and 48 of the derived sequence were ambiguous. The signal peptidase cleavage site is indicated by the arrowhead. A potential stem-loop structure after the CS5 pilin gene is shown. The N-terminal coding region for the gene after that for CS5 pilin is shown.

```
CS5  MKKNLLITSVLAMATVSGSVLAAV-TNGQLTFNWQGVVPSAPVTQSSWAFVNGLDIPFTP
F41  MKKTLIALAVAASAAVSGSVMAADWTEGQPGDIIIGGEITSPSVKWLWKTGEGLS-SFSN
     ***.*.  .* * *.*****.**  *.** .    *   ..* ..  *   .**. .*.

CS5  GTEQ-----LNITLDSNKDITARSVKP--YDFFI----VP-----------VSGNVTPGA
F41  TTNEIVKRKLNISVPTDELFLAAKMSDGIKGVFVGNTLIPKIEMASYDGSVITPSFTSNT
     .*..     ***.. ... . * ...    . *.    .*           .. . *...

CS5  PVTRDTSANINSVNAFLSSVPVSNGFVGN---------KQLTLSTAVEAAKGEVAITLNG
F41  AMDIAVKVKNSGDNTELGTLSVPLSFGAAVATIFDGDTTDSAVAHIIGGSAGTVFEGLVN
     ... ..... .. *. *....*. .* ..         .. ...  .... *.*  .* .

CS5  QA-------------------------LKVGSASPTVVTV-----------ASNKKESH
F41  PGRFTDQNIAYKWNGLSKAEMAGYVEKLMPGQSASTSYSGFHNWDDLSHSNYTSANKASY
     ..                          *. * ...*  .             .*..*.*.

CS5  ISIDMNAKAAAADVAE-----GAAINFVAPVTFAVDI-
F41  LSYGSGVSAGSTLVMNLNKDVAGRLEWVAPVTITVIYS
     .* . ...*... * .     .. ...*****..*
```

FIG. 3. Homology between CS5 pilin and F41 pilin. Data base searches were performed via electronic mail with the EMBL MAILFASTA server. The amino acid sequences of the major fimbrial subunits of CS5 and F41 were aligned by using the CLUSTAL program of Higgins and Sharp (6). The dots below the line correspond to similar amino acids, and the asterisks correspond to identical amino acids.

mutagenesis with Tn*1725* as previously described (10, 24) (Fig. 1). Few insertions were obtained in the cloned DNA, suggesting a form of hot spotting by Tn*1725*, although this has not occurred in other cases where we have extensively used this transposon (9, 10, 12). Initial studies indicated that the minimal coding region is delimited by the noninactivating transposon insertions in plasmids pPM1341 and pPM1338, a distance of approximately 7.0 kb (5). The insertions between these two points, namely, those in plasmids pPM1336, pPM1340, and pPM1342, all eliminate fimbrial expression but not the fimbrial subunit (Fig. 1; data not shown).

Using a synthetic oligodeoxynucleotide probe based on the N-terminal sequence of the mature CS5 major subunit, we previously showed that the N-terminal coding region lies on an *Mlu*I-*Hin*dIII fragment (Fig. 1) (5). Initial sequencing in this region confirmed this but revealed that most of the coding region was to the left of the *Mlu*I site as shown in Fig. 1. The complete sequence of this stretch of DNA was subsequently determined (Fig. 2).

An open reading frame encoding a protein of 203 amino acid residues was identified. The first 22 amino acids correspond to a typical signal sequence (25, 26), and residues 23 to 46 and 49 to 52 agree with the N-terminal analysis previously obtained with the purified CS5 fimbrial subunit (5). Protein sequence analyses of residues 25 and 26 were ambiguous. Thus, it is clear that the open reading frame was that for the gene encoding the major CS5 subunit.

Based on the sequence data, the precursor for the CS5 subunit would have a size of 20.85 kDa and the mature protein would be 18.6 kDa. Previous estimates based on sodium dodecyl sulfate-polyacrylamide gel electrophoresis suggested that the mature subunit had a size of about 23 kDa. The basis of this minor discrepancy is unknown but it is worth noting that there is an unusual distribution of charged residues within the subunit, and there are no charged residues within the first 30 residues of the mature subunit. Also, CS5 is the most hydrophobic of any of the ETEC pili examined and is readily precipitated with less than 5% ammonium sulfate (5).

Following the CS5 major subunit gene is a potential stem-loop structure that may function as a transcriptional terminator or attenuator to reduce the expression of genes downstream. It is relatively stable thermodynamically, with a free energy of −21.1 kCal (ca. −88.3 kJ)/mol (23). This is followed by the first 76 residues of the next open reading frame. Of these residues, the first 18 also are typical of a signal sequence, suggesting that the product is an exported protein. There are also several long stretches of hydrophobic or neutral amino acids within the protein. These regions and the organization of the predicted turns in the protein structure suggest that there is much intrinsic information associated with self-assembly into the marked helical structure that has been observed within the fimbriae (5, 11).

We have sought proteins related to the CS5 major subunit in the various data bases. The only protein with any significant homology was the major fimbrial subunit of the F41 fimbriae of porcine ETEC (15). The CS5 subunit shows a high degree of homology within the signal sequence and at the extreme ends of the mature protein (Fig. 3). The homology of CS5 with the F41 fimbriae of porcine ETEC is not dramatic but is statistically significant. This suggests that the two proteins have a common origin and functional signal sequence (and presumably C-terminal) domains that are highly conserved, whereas the remainder of the protein was subject to considerable evolutionary pressures, affecting its antigenicity and probably also reflecting adaption to different host epithelial receptors. F41 has also been shown to be related to K88, another porcine ETEC (15); another human fimbrial type, PCFO9 (4), is even more closely related (1).

A variety of (potential) fimbrial colonization factors have been identified in ETEC of human origin. By analogy with similar structures in animal ETEC, these too probably represent the important protective determinants for inclusion in vaccines. Thus, to develop a broad-based vaccine, the relative incidence of the different colonization factors needs to be established. The availability of the nucleotide sequences of individual genes specific for each of the different colonization factors also has the potential to facilitate the development of suitable epidemiological reagents. For example, by making a cocktail of synthetic oligodeoxynucleotide primers, a multiplex polymerase chain reaction could be set up so that several factors could be looked for simultaneously and rapidly. This has advantages over antibody-related tests in that cross-reactions between some human CFAs have been observed. Also, gene probing consumes considerably more time and reagents. The polymerase chain reaction also enables a high degree of stringency in recognition.

Nucleotide sequence accession number. The nucleotide sequence for the CS5 fimbrial subunit determinant has been submitted to EMBL under accession number X63411.

ACKNOWLEDGMENTS

We thank the National Health and Medical Research Council of Australia and the Clive and Vera Ramaciotti Foundations for support. C.A.C. is supported by a Commonwealth Postgraduate Research Award.

REFERENCES

1. **Clark, C. A., and P. A. Manning.** Unpublished data.

1a. **Evans, D. G., D. J. Evans, S. Glegg, and I. Parsley.** 1979. Purification and characterization of CFA/II antigens of enterotoxigenic *Escherichia coli*. Infect. Immun. **25:**738–748.

2. **Evans, D. G., R. P. Silver, D. J. Evans, Jr., W. S. D. G. Chase, and S. L. Gorbach.** 1975. Plasmid-controlled colonization factor associated with virulence in *Escherichia coli* enterotoxigenic for humans. Infect. Immun. **12:**656–667.
3. **Gaastra, W., and F. K. de Graaf.** 1982. Host-specific fimbrial adhesins of noninvasive enterotoxigenic *Escherichia coli* strains. Microbiol. Rev. **46:**129–161.
4. **Heuzenroeder, M. W., T. R. Elliot, C. J. Thomas, R. Halter, and P. A. Manning.** 1990. A new fimbrial type (PCFO9) on enterotoxigenic *Escherichia coli* O9:H^-LT^+ isolated from a case of infant diarrhea in Central Australia. FEMS Microbiol. Lett. **66:**55–60.
5. **Heuzenroeder, M. W., B. L. Neal, C. J. Thomas, R. Halter, and P. A. Manning.** 1989. Characterization and molecular cloning of the PCF8775 CS5 antigen from an enterotoxigenic *Escherichia coli* O115:H40 isolated in Central Australia. Mol. Microbiol. **3:**303–310.
6. **Higgins, D. G., and P. M. Sharp.** 1988. CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. Gene **73:**237–244.
7. **Honda, T., M. Arita, and T. Miwatani.** 1984. Characterization of new hydrophobic pili of human enterotoxigenic *Escherichia coli*—a possible new colonization factor. Infect. Immun. **43:** 959–965.
8. **Levine, M. M., J. B. Kaper, R. E. Black, and M. L. Clements.** 1983. New knowledge on pathogenesis of bacterial enteric pathogens as applied to vaccine development. Microbiol. Rev. **47:**510–550.
9. **Manning, P. A., E. J. Bartowsky, D. I. Leavesley, J. A. Hackett, and M. W. Heuzenroeder.** 1985. Molecular cloning using immune sera of a 22000 dalton minor outer membrane protein of *Vibrio cholerae*. Gene **34:**95–103.
10. **Manning, P. A., M. H. Brown, and M. W. Heuzenroeder.** 1984. Cloning of the structural gene (*hly*) for the haemolysin of *Vibrio cholerae* El Tor strain O17. Gene **31:**225–231.
11. **Manning, P. A., G. D. Higgins, R. Lumb, and J. A. Lanser.** 1986. Colonization factor antigens and a new fimbrial type, CFA/V on O115:H40 and H^- strains of enterotoxigenic *Escherichia coli* in Central Australia. J. Infect. Dis. **156:**841–844.
12. **Manning, P. A., K. N. Timmis, and G. Stevenson.** 1985. Colonization factor antigen II (CFA/II) of enterotoxigenic *Escherichia coli*: molecular cloning of the CS3 determinant. Mol. Gen. Genet. **200:**322–327.
13. **Merson, M. H., R. E. Black, M. U. Khan, and I. Huq.** 1980. Enterotoxigenic *Escherichia coli* diarrhea: acquired immunity and transmission in an endemic area, p. 34–45. *In* O. Ouchterlony and J. Holmgren (ed.), Cholera and related diseases. 43rd Nobel Symposium. S. Karger, Basel.
14. **Morgan, R. L., R. E. Isaacson, H. W. Moon, C. C. Brinton, and C. C. To.** 1978. Immunization of suckling pigs against enterotoxigenic *Escherichia coli*-induced diarrheal disease by vaccinating dams with purified 987P or K99 pili: protection correlates with pilus homology of vaccine and challenge. Infect. Immun. **22:**771–777.
15. **Moseley, S. L., G. Dougan, R. A. Schneider, and H. W. Moon.** 1986. Cloning of chromosomal DNA encoding the F41 adhesin of ETEC and genetic homology between adhesins F41 and K88. J. Bacteriol. **167:**799–804.
16. **Parry, S. H., and D. M. Rooke.** 1985. Adhesins and colonization factors of *Escherichia coli*, p. 79–155. *In* M. Sussman (ed.), The virulence of *Escherichia coli*. Academic Press, London.
17. **Rutter, J. M., and G. W. Jones.** 1973. Protection against enteric disease caused by *Escherichia coli* a model for vaccination with a virulence determinant? Nature (London) **242:**531–532.
18. **Sack, R. B.** 1975. Human diarrheal disease caused by enterotoxigenic *Escherichia coli*. Annu. Rev. Microbiol. **29:**333–353.
19. **Smyth, C. J.** 1982. Two mannose-resistant haemagglutinins on enterotoxigenic *Escherichia coli* of serotype O6:K15:H16 or H^- isolated from travellers and infantile diarrhoea. J. Gen. Microbiol. **128:**2081–2096.
20. **Tacket, C. O., D. R. Maneval, and M. M. Levine.** 1987. Morphology and genetics of a new fimbrial putative colonization factor of enterotoxigenic *Escherichia coli* O159:H4. Infect. Immun. **55:**1063–1069.
21. **Thomas, L., A. Cravioto, S. M. Scotland, and B. Rowe.** 1982. New fimbrial antigen type (E8775) that may represent a new colonization factor in enterotoxigenic *Escherichia coli* in humans. Infect. Immun. **35:**1119–1124.
22. **Thomas, L. V., M. M. McConnell, B. Rowe, and A. M. Field.** 1985. The possession of three novel *coli* surface antigens by enterotoxigenic *Escherichia coli* strains positive for the putative colonization factor PCF8775. J. Gen. Microbiol. **131:**2319–2326.
23. **Tinoco, I., Jr., P. N. Borer, B. Dengler, M. D. Levine, O. C. Uhlenbeck, D. M. Crothers, and J. Gralla.** 1973. Improved estimation of secondary structure in ribonucleic acids. Nature (London) New Biol. **246:**40–41.
24. **Ubben, D., and R. Schmitt.** 1986. Tn1721 derivatives for transposon mutagenesis, restriction mapping and nucleotide sequence analysis. Gene **41:**145–152.
25. **von Heijne, G.** 1984. How signal sequences maintain cleavage specificity. J. Mol. Biol. **173:**243–251.
26. **von Heijne, G.** 1985. Signal sequences: the limits of variation. J. Mol. Biol. **184:**99–105.