Nucleotide sequence of the *sfaA* gene coding for the S-fimbrial protein subunit of *Escherichia coli*

Thomas Schmoll, Jörg Hacker and Werner Goebel

Institut für Genetik und Mikrobiologie, Universität Würzburg, F.R.G.

1. SUMMARY

The *sfaA* gene of the uropathogenic *Escherichia coli* O6 strain 536, which is responsible for the determination of the S fimbrial protein subunit, was sequenced. The structural gene codes for a polypeptide of 180 amino acids including a 24-residue N-terminal signal sequence. A size of 15.95 kDa was calculated for the processed SfaA protein. The nucleotide and deduced amino acid sequences show significant homology to those of the F1C fimbria and, to a lesser extent, of the mannose-sensitive hemagglutinating fimbria (FimA, PilA). Only week homology to P fimbriae subunits (F7$_2$, Pap) was found.

2. INTRODUCTION

Fimbriae or pili are protein appendices of the bacterial cell wall, 7 nm in diameter and up to 1 μm in length. A single fimbria consists of about 1000 fimbrillin subunits of identical protein mass [1,2]. Fimbriae, which are very often associated with an adhesive capacity, contribute to pathogenicity in both intestinal and extraintestinal *E. coli* infections [3–5].

Most fimbriae from pathogenic *E. coli* strains are able to agglutinate erythrocytes in a mannose-resistant manner (mannose-resistant hemagglutination, Mrh). These Mrh fimbriae can be further subdivided on the basis of their receptor specificity. The subgroups have been termed as P-, S-, M-, or X-fimbriae [6–9]. P-fimbriae, found predominantly among uropathogenic *E. coli* strains, recognize the α-D-Gal-(1-4)-β-Gal part of globosides [10]. P-fimbriae exhibit differences in their serological specificity (F groups F7 to F13, [11]). S-fimbrial adhesins (Sfa), which are associated with uropathogenic strains and, to a larger extent with strains causing newborn meningitis (NBM) interact with α-sialyl-β-(2-3)-Gal-containing receptors [12,13]. Another type of fimbria, F1C, which is devoid of demonstrable receptor specificity, is also associated with urinary tract infections (UTI) [14,15]. In contrast, mannose-sensitive hemagglutinating (Msh) fimbriae, which form another group of fimbriae and recognize α-mannose-containing receptors, can be detected on *E. coli* strains from different sources [11,16].

In this contribution we present the DNA se-

Correspondence to: Jörg Hacker, Institut für Genetik und Mikrobiologie, Univ. Würzburg, Röntgenring 11, D-8700 Würzburg, F.R.G.

230

quence and the deduced amino acid sequence of the S fimbrial protein subunit, termed SfaA. Surprisingly, the *sfaA* sequence shows only poor homology to other Mrh fimbriae of the P-type, but is closely related to that of the F1C and Msh fimbriae.

3. MATERIALS AND METHODS

3.1. Bacterial strains

The S-fimbrial adhesin (*sfa*, previously termed X) was cloned from the chromosome of the uropathogenic *E. coli* O6 strain 536 [17]. *E. coli* K-12 strains JM107 and JM103 were used as hosts for the M13 bacteriophage clones.

3.2. Cloning and nucleotide sequencing

Suitable fragments of the *sfa*-specific recombinant plasmids pANN801-13 and pANN801-15 (see [17,18]) were subcloned in the M13 derivatives as described. The fragments are indicated in Fig. 1. Nucleotide sequencing was performed by using

the Sanger technique with the M13mp18/mp19 system described previously [19].

3.3. Computer programs

The programs used for compiling the nucleotide sequence data were from J. Devereux (UWGCG).

4. RESULTS AND DISCUSSION

4.1. Localization of the sfaA gene coding for the S-fimbrillin

The localization of the *sfaA* gene coding for the S-fimbrial protein subunit was determined by the analysis of appropriate Tn5 mutants [17]. 3 mutants with an inserted Tn5 element within a sequence representing the 5'-end of the *sfa* determinant (between coordinates 7.0 and 8.5 kb in Fig. 1) abolished expression of fimbriae. A subcloned DNA (pANN801-15) consisting of the *Cla*I-*Eco*RV fragment of this region (coordinates 6.9–9.8 kb) and pACYC184 still expressed the
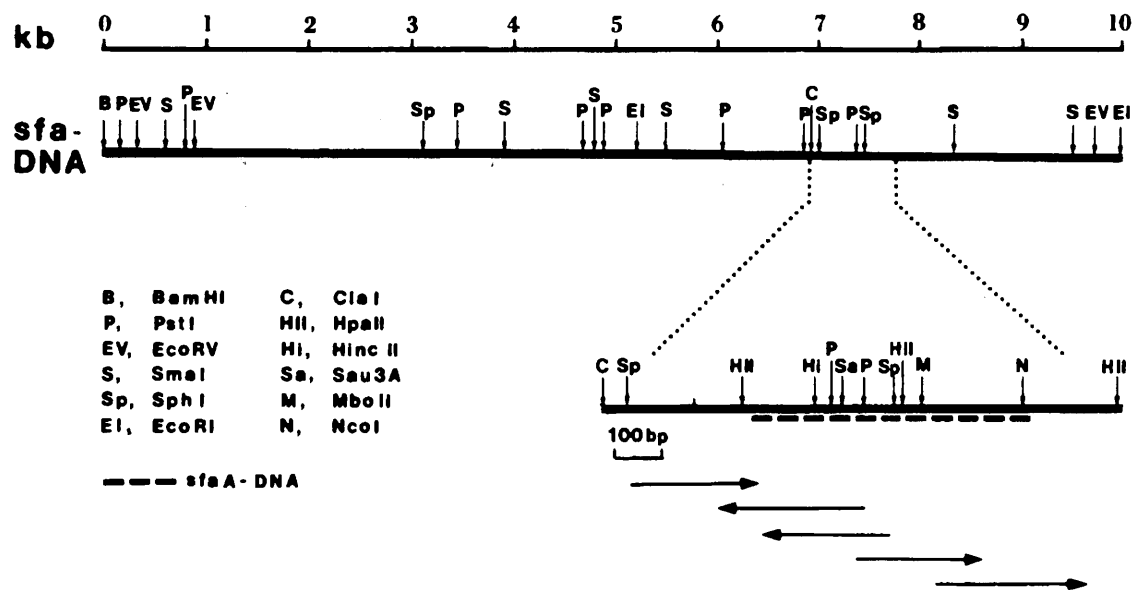
Fig. 1. Physical map of the *sfa* determinant and sequencing strategy. The black line represents the *sfa*-specific DNA region of the plasmid pANN801-13 [17]. The *sfa* determinant is located between coordinates 1.9 kb and 8.4 kb. The *sfaA* gene is indicated by the broken line between coordinates 7.8 kb and 7.2 kb. The main direction of transcription is from right to left [18] (Schmoll et al., in preparation). The direction of sequenced DNA fragments is shown by arrows.

fimbrial subunits as detected in Western blots (Ott and Hacker, unpublished results).

The nucleotide sequence between coordinates 7.0 kb and 8.8 kb was determined (Schmoll et al., in preparation) and only one polypeptide, of almost 16 kDa, which is comparable to what has been found previously for the S-fimbrial subunit [17], was encoded by this region between map positions 7.2 and 7.8 kb. The identity between this polypeptide and the S fimbrillin was further determined on the basis of the amino-acid composition. The amino-acid composition, deduced from the nucleotide sequence, was absolutely identical to the amino-acid composition of the S-fimbrillin determined by biochemical methods (Moch et al., submitted for publication).

4.2. Analysis of the nucleotide sequence of sfaA

The nucleotide sequence of the *sfaA*-containing DNA region is given in Fig. 2. This sequence contains a single reading frame (ORF) of 540 bp, starting at an ATG codon at bp 166–168 and ending with a TAA stop codon at bp 706–708. Putative consensus promoter sequences are found at positions 20–25 (-35 box) and 41–47 (-10 box). Using the λMu*lacZ* phage system [20] the *lacZ* gene was fused to the *sfaA* gene [18] (Schmoll et al., in preparation). The production of β-galactosidase by these *sfaA-lacZ* fusions and recent S1 mapping experiments (Jarchau et al., in preparation) give evidence that the predicted promoter in front of *sfaA* is used in vivo.

4.3. Amino-acid sequence of the SfaA protein

The start codon ATG for SfaA is preceeded by

```
         1               20              40              60
GAAAATATTATCGGAGATAATGTCATAAATGCTGCCTGAGTGTATTTCTCACATTGCATTTATGAAGTTCTCCTG

        80              100             120             140             160
AAAAAAGATTCCCGTCGTTCGGGATATTGATTGTGTCTGTTGTGATGACAGATACGGTGTGCGTAGTTCAATTAAAAACAGGAATTAAAT

        180             200             220             240
ATGAAGTTAAAATTCATCTCCATGGCTGTATTTTCAGCCCTGACCTTGGGTGTTGCGACAAATGCGTCTGCTGTCACCACGGTTAATGGT
MetLysLeuLysPheIleSerMetAlaValPheSerAlaLeuThrLeuGlyValAlaThrAsnAlaSerAlaValThrThrValAsnGly
          -20                                                        -1 +1

        260             280             300             320             340
GGTACAGTTCATTTTAAGGGGGAAGTTGTTGATGCTGCATGTGCTGTAAACACTAATTCAGCAAATCAAACGTTTTCTGGGCAAGTTCGT
GlyThrValHisPheLysGlyGluValValAspAlaAlaCysAlaValAsnThrAsnSerAlaAsnGlnThrPheSerGlyGlnValArg
                                          20

        360             380             400             420
TCAGCTAAGTTGGCGAATGATGGAGAGAAGAGTTCCCCTGTTGGATTTAGTATTGAACTTAATGACTGTAGTTCTGCAACTGCCGGGCAT
SerAlaLysLeuAlaAsnAspGlyGluLysSerSerProValGlyPheSerIleGluLeuAsnAspCysSerSerAlaThrAlaGlyHis
          40                                                60

        440             460             480             500             520
GCATCAATTATCTTTGCAGGAAATGTTATTGCTACACACAATGATGTGCTGTCTCTACAGAATAGTGCTGCAGGTAGTGCAACAAATGTA
AlaSerIleIlePheAlaGlyAsnValIleAlaThrHisAsnAspValLeuSerLeuGlnAsnSerAlaAlaGlySerAlaThrAsnVal
                                          80

        540             560             580             600
GGTATTCAGATATTGGATCATACAGGTACTGCAGTTCAATTTGACGGAGTGACTGCATCTACACAATTTACATTAACAGATGGCACCAAT
GlyIleGlnIleLeuAspHisThrGlyThrAlaValGlnPheAspGlyValThrAlaSerThrGlnPheThrLeuThrAspGlyThrAsn
          100                                               120

        620             640             660             680             700
AAAATTCCTTTCCAGGCAGTTTATTATGCAACAGGTAAGTCAACGCCTGGTATTGCCAACGCCGACGCCACCTTTAAAGTTCAGTACCAG
LysIleProPheGlnAlaValTyrTyrAlaThrGlyLysSerThrProGlyIleAlaAsnAlaAspAlaThrPheLysValGlnTyrGln
                                          140                                        156

        720             740             760
TAATATCAGAACAGTGTAACGATATATACCCGGCCAGGAGGGCTGTTTTTATCATGC
End
```
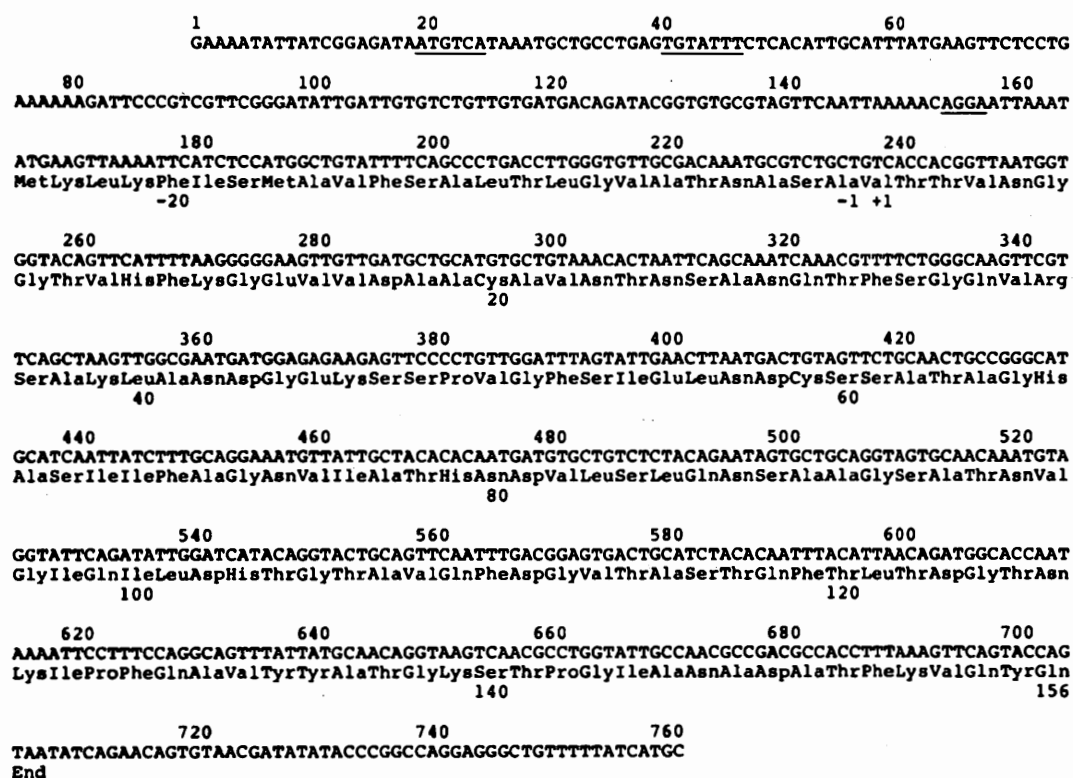
Fig. 2. Nucleotide sequence of the *sfaA* structural gene. The predicted amino-acid sequence of the SfaA protein is shown beneath the nucleotide sequence. Numbers above each line refer to the nucleotide position and numbers below each line to the amino acid position. The amino acids constituting the putative signal peptide are indicated by negative numbers, the amino acids of the mature protein are indicated by positive numbers. Potential ribosome-binding sites are underlined.

a potential ribosome binding site (–AGGA–) [21]. The N-terminal part of the 180-amino-acid-long SfaA protein has the characteristics of a prokaryotic signal peptide [22]. As shown in Figs. 3 and 4, the N-terminal part of the SfaA protein has the same amino-acid composition as the F1C fimbrillin [23]. Therefore, the signal cleavage site is proposed to be between amino acid –1 (alanine) and amino acid +1 (valine; for coordinates see Fig. 2) as determined for F1C by sequencing the N-terminal end of the mature fimbrillin [24]. The immediate N-terminal end of the signal sequence is positively charged (2 lysine residues) and is followed by a typical highly hydrophobic domain (Fig. 3). With the exception of a few codons (e.g., those coding for phenylalanine) the codon usage for SfaA is compatible with normal genes present in E. coli (not shown; [25]), as also reported for P- and Msh-fimbriae [26,27].

The molecular weight of the mature SfaA protein was determined from the deduced amino-acid sequence as 15 950d which is in good agreement with previously published data [17]. The processed SfaA protein contains 2 cysteine residues at amino acid positions +20 and +59, which may form a intramolecular disulfide bridge similar to what has been shown for other subunit proteins [24,26]. The mature S-fimbrillin subunit further contains alanine [21], tyrosine [18], glycine [15], valine [15] and serine [14] as predominant amino acids. Tryptophan and methionine do not occur in the sequence (Fig. 2). The hydrophilicity profile, calculated according to the method of Hopp and Woods [28] shows several hydrophobic regions for the SfaA protein at amino-acid positions 32–35, 73–78, 83–103, 117–125 and 151–160 (these amino-acid positions are given for the precursor protein, as indicated in Fig. 3). The hydrophilic maxima which should be located in or immediately adjacent to the antigenic dominant parts of the protein [28] were found around amino-acid positions 41 and 70 and, to a lesser extent, around amino acids 147 and 163.

4.4. Comparison of the sfaA gene with other fimbrial genes

The nucleotide sequence and the deduced amino acid sequence of the S fimbrial subunit were compared with the corresponding sequences of other fimbrial genes from E. coli involved in extraintestinal and intestinal infections [23,26,27,29–32]. No significant homology was found between the sequences of sfaA and of the genes coding for the fimbrial protein subunits of plasmid-encoded K88ab and K99 adhesins, which are associated with intestinal infections in animals (for details, see Table 1). Surprisingly, homology between sfaA and P-fimbrial determinants ($F7_2$, papA), which both belong to the group of Mrh fimbriae, is poor on the DNA and on the protein level. As indicated in Fig. 4, only the N-terminal and the C-terminal parts of the S- and P-fimbrillin proteins show homology, and also, the 2 cysteine residues are located at exactly the same place.

In contrast, the sequences of the Msh fimbrial subunits (FimA, PilA) and the F1C fimbrillin reveal striking homology with the SfaA sequence. These results indicate an evolutionary relationship between sfaA and these non-Mrh determinants. The extent of homology between the SfaA protein and the F1C fimbrillin is 70.9%. On the DNA level, homology is 73.3%. These results argue for a very close relatedness of both determinants. The similarity is further confirmed by the hydrophilicity profiles of the encoded proteins (Fig. 3). Sequence alterations, however, were found in those parts of the determinants which code for hydrophilic domains of the proteins (between the sec-

Table 1

Comparison of the nucleotid and amino acid sequence of different fimbrial subunits with the S-fimbrial protein subunit determinant sfaA

Fimbrial subunit	HA-type [a]	Nucleotide sequence homology (%)	Amino-acid sequence homology (%)	Ref.
F1C	–	73.3	70.9	[23]
PilA	Msh	61.2	56.6	[29]
FimA	Msh	59.0	52.0	[27]
$F7_2$	Mrh (P)	41.8	21.7	[30]
PapA	Mrh (P)	39.3	19.6	[26]
K99	Mrh	36.4	12.6	[31]
K88ab	Mrh	35.4	11.6	[32]

[a] HA, hemagglutination; Msh, mannose-sensitive hemagglutination; Mrh, mannose-resistant hemagglutination; P,P-specific hemagglutination.
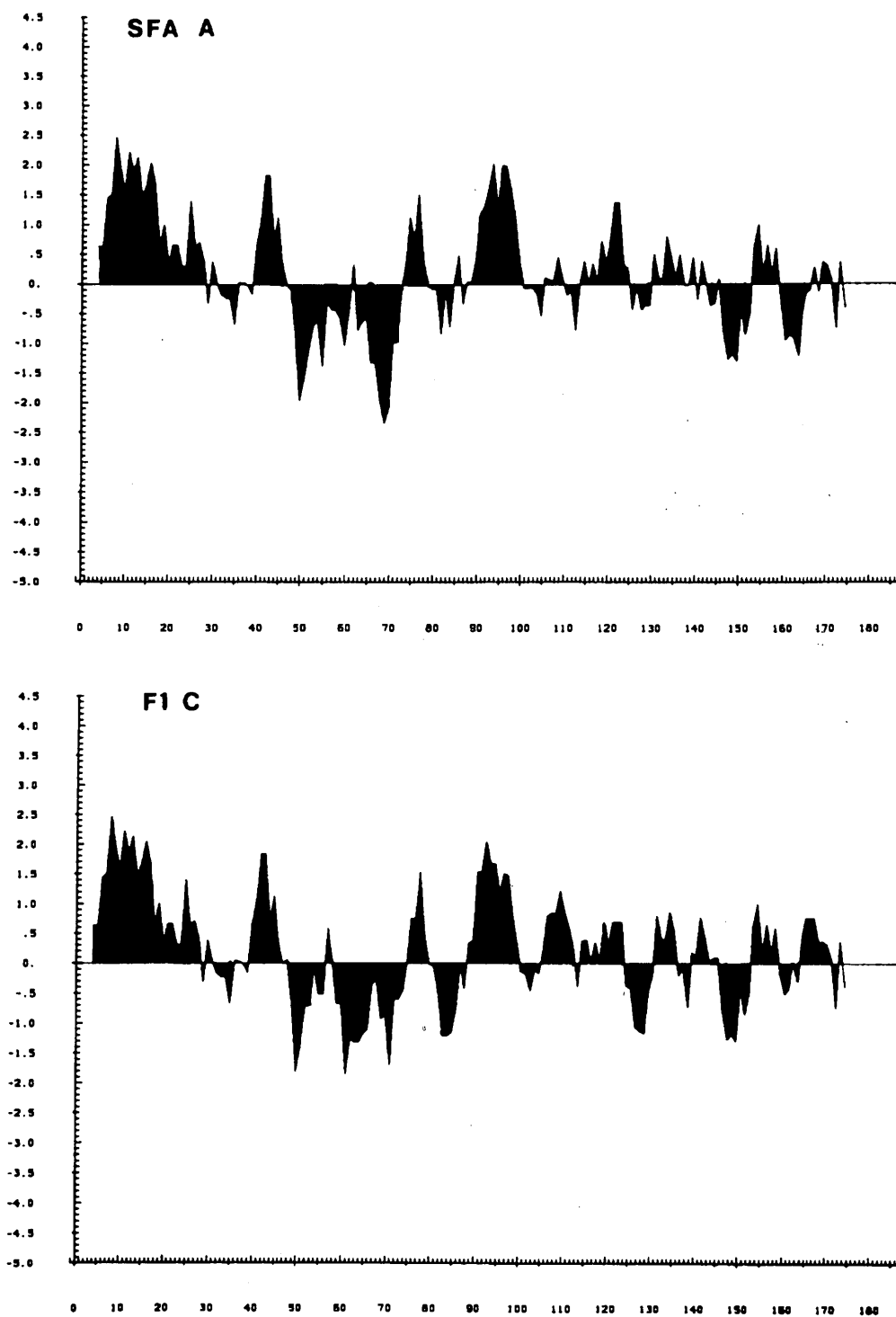
Fig. 3. Hydrophilicity profile of the SfaA fimbrial subunit, compared with the hydrophilicity profile of the F1C protein [23]. The amino acid residues are numbered from the N-terminal end of the signal peptide.
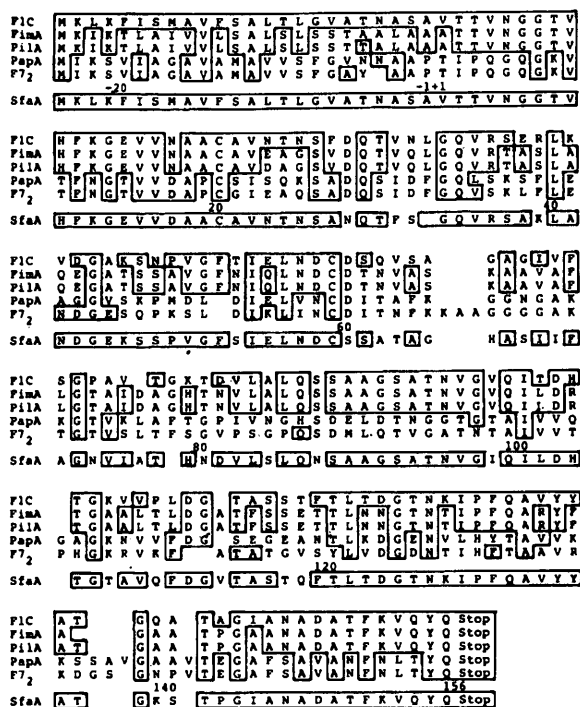
234

Fig. 4. Comparison of the amino acid sequence of the SfaA protein with those of F1C, common type I (Msh; FimA, PilA), Pap, and F7₂ fimbriae [23,26,27,29,30]. The amino-acid sequences are given in single letter code. Homologies are indicated by the boxes. Numbering applies only to SfaA fimbrial protein. Gaps have been introduced to obtain maximum homology between the 2 proteins.

ond cystein at position 83 and amino-acid position 143; coordinates of the precursor). These differences were already discussed on the basis of a restriction site polymorphism, observed in the corresponding nucleotide sequences [33]. It is interesting to note that these hydrophilic parts of the proteins seem to determine the antigenic characters of both fimbriae [28]. It is suggested therefore, that the observed sequence alterations may account for the incomplete antigenic identity of the two fimbriae, as demonstrated in a quantitative ELISA assay (A. Pere, unpublished data). Otherwise, it has been shown [33] that homology between the *sfa* and F1C determinants exists over the entire coding region, which is in contrast to what has been found for the *sfa* and the *fim* and

pil determinants, respectively (Ott and Hacker, unpublished results). It will be interesting to determine more precisely the differences in those parts of the *sfa* and F1C determinants which code for the specific adhesive property of the S-fimbriae.

ACKNOWLEDGEMENTS

REFERENCES

[1] Jones, G.W. and Isaacson, R.E. (1983) CRC Crit. Rev. Microbiol. 10, 229–260.

[2] Klemm, P. (1985) Rev. Infect. Dis. 7, 321–340.

[3] Gaastra, W. and de Graaff, F.K. (1982) Microbiol. Rev. 46, 129–161.

[4] Svanborg-Edén, C., Hagberg, L., Hanson, L.A., Jodal, U., Leffler, H., Lomberg, H. and Straube, E. (1983) Progr. Allergy 33, 175–189.

[5] Marre, R., Hacker, J., Henkel, W. and Goebel, W. (1986) Infect. Immun. 54, in press.

[6] Källenius, G., Möllby, R., Svenson, S.B., Winberg, J., Lundberg, A. and Svensson, S. (1980) FEMS Microbiol. Lett. 7, 297–300.

[7] Parkkinen, J., Finne, J., Achtman, M., Väisänen, V. and Korhonen, T.K. (1983) Biochem. Biophys. Res. Commun. 111, 456–461.

[8] Väisänen, V., Korhonen, T.K., Jokinen, M., Gahmberg, C.G. and Ehnholm, C. (1982) Lancet 1, 1192.

[9] Väisänen-Rhen, V. (1984) Infect. Immun. 46, 401–407.

[10] Bock, K., Breimer, M.E., Brignole, A., Hansson, G.C., Karlsson, K.-A., Larson, G., Leffler, H., Samuelsson, B.E., Strömberg, N., Svanborg-Edén, C. and Thurin, J. (1985) J. Biol. Chem. 260, 8545–8551.

[11] Ørskov, I. and Ørskov, F. (1983) Prog. Allergy 33, 80–105.

[12] Korhonen, T.K., Valtonen, M.V., Parkkinen, J., Väisänen-Rhen, V., Finne, J., Ørskov, F., Ørskov, I., Svenson, S.B. and Mäkelä, P.H. (1985) Infect. Immun. 48, 486–491.

[13] Parkkinen, J., Rogers, G.N., Korhonen, T.K., Dahr, W. and Finne, J. (1986) Infect. Immun. 54, 37–42.

[14] Schmitz, S., Abe, C., Moser, I., Ørskov, I., Ørskov, F., Jann, B. and Jann, K. (1986) Infect. Immun. 51, 54–59.

[15] Pere, A., Leinonen, M., Väisänen-Rhen, V., Rhen, M. and Korhonen, T.K. (1985) J. Gen. Microbiol. 131, 1705–1711.

[16] Duguid, J.P., Smith, I.W., Dempster, G. and Edmunds, P.N. (1955) J. Pathol. Bacteriol. 70, 335–348.

[17] Hacker, J., Schmidt, G., Hughes, C., Knapp, S., Marget, M. and Goebel, W. (1985) Infect. Immun. 47, 434–440.

[18] Hacker, J., Jarchau, T., Knapp, S., Marre, R., Schmidt, G., Schmoll, T. and Goebel, W. (1986) in Protein-Carbohydrate Interactions in Biological Systems (Lark, D., Ed.) Academic Press, London, pp. 125–133.

[19] Hu, N. and Messing, (1982) Gene 17, 271–277.

[20] Bremer, E., Silhavy, T.J. and Weinstock, G.M. (1985) J. Bacteriol. 162, 1092–1099.

[21] Gold, L., Pribnow, D., Schneider, T., Shinedling, S., Singer, B.S. and Stormo, G. (1981) Annu. Rev. Microbiol. 35, 365–403.

[22] Inouye, M. and Hageloua, S. (1980) Crit. Rev. Biochem. 7, 339–371.

[23] Van Die, I., Van Geffen, B., Hoekstra, W. and Bergmans, H. (1984) Gene 34, 187–196.

[24] Klemm, P., Ørskov, I. and Ørskov, F. (1982) Infect. Immun. 36, 462–468.

[25] Ikemura, T. (1981) J. Mol. Biol. 151, 389–409.

[26] Baga, M., Normark, S., Hardy, J., O'Hanley, P., Lark, D., Olsson, O., Schoolnik, G. and Falkow, S. (1984) J. Bacteriol. 157, 330–333.

[27] Klemm, P. (1984) Eur. J. Biochem. 143, 395–399.

[28] Hopp, T.P. and Woods, K.R. (1981) Proc. Natl. Acad. Sci. USA 78, 3824–3828.

[29] Orndorff, P.E. and Falkow, S. (1985) J. Bacteriol. 162, 454–457.

[30] Van Die, I. and Bergmans, H. (1984) Gene 32, 83–90.

[31] Roosendaal, B., Gaastra, W. and de Graaf, F.K. (1984) FEMS Microbiol. Lett. 22, 253–258.

[32] Gaastra, W., Mooi, F.R., Stuitje, A.R. and de Graaf, F.K. (1981) FEMS Microbiol. Lett. 12, 41–46.

[33] Ott, M., Hacker, J., Schmoll, T., Jarchau, T., Korhonen, T.K. and Goebel, W. (1986) Infect. Immun. 54, 646–653.