JOURNAL OF BACTERIOLOGY, Dec. 1991, p. 7673–7683
0021-9193/91/237673-11$02.00/0

Vol. 173, No. 23

Sequence Analysis of the *clpG* Gene, Which Codes for Surface Antigen CS31A Subunit: Evidence of an Evolutionary Relationship between CS31A, K88, and F41 Subunit Genes

JEAN-PIERRE GIRARDEAU,* YOLANDE BERTIN, CHRISTINE MARTIN, MAURICE DER VARTANIAN, AND CHRISTIANE BOEUF

Laboratoire de Microbiologie, Institut National de la Recherche Agronomique, Centre de Recherche de Clermont-Ferrand—Theix, 63122 Saint Genes-Champanelle, France

Received 29 May 1991/Accepted 2 October 1991

The *clpG* gene coding for the CS31A subunit was localized on a 0.9-kb *Sph*I fragment from the recombinant plasmid pAG315. This was established by testing the ability of subclones to hybridize with a 17-meric oligonucleotide probe obtained from N-terminal analysis of the CS31A subunit. The nucleotide sequence of the region coding for CS31A was determined. From primer extension analysis, two initiation translation start sites were detected. Two possible promoterlike sequences were identified; the ribosome binding site and the translation terminator are proposed. Inverted repeat sequences leading to the formation of possible hairpin structures of the transcripts were found on the 5′ untranslated region of *clpG*. The deduced amino acid composition was in close agreement with the chemical amino acid composition and sequence match with the first 25 N-terminal amino acids from the published N-terminal sequence of the purified CS31A subunit. The *clpG* gene codes for a mature protein of 257 amino acids with a molecular size of 26,777 Da. An obvious homology was observed when the amino acid sequence of CS31A was compared with those of K88 and F41. This homology includes five different conserved sequences of up to 19 identical amino acids, which is associated with conserved proline. An extensive change in the CS31A region homologous to that identified to contain the K88 receptor binding site might be responsible for the functional divergence between CS31A and K88.

Fimbriae from many bacterial pathogens have been extensively characterized and shown to be important for virulence by promoting adhesion of bacteria to the host epithelial cells (10, 24, 25). Fimbriae are composed of structural subunits, referred to as pilins, that range in size from approximately 15 to 30 kDa. The K88 and F41 fimbrial adhesins produced at the surface of enterotoxigenic *Escherichia coli* strains mediate host-specific adhesion to the enterocytes and thereby enable the colonization of the epithelium of the small intestine (24, 25). The K88 and F41 operons contain various auxiliary proteins involved in processing, transport, and assembly of fimbrial subunits (31, 32). The genes required for the expression of K88 and F41 have been cloned and found to share extensive DNA sequence homology, except for the fimbrial subunit genes (4, 33).

Serological variants of K88 antigen have been described and reported as K88ab, K88ac, K88ad (14). Their primary structure revealed highly conserved, variable, and hypervariable regions (23). Variable regions are probably involved in the antigenic specificity of the three variants, and conserved regions are supposed to be involved in such common features as folding, processing, and functioning of the polypeptide (30). Jacobs et al. (22) demonstrated that two conserved tripeptides are involved in the K88 receptor binding domain and suggested that the receptor binding site could be a hydrophobic cleft which encompasses the two conserved amino acid sequences and which interacts with the ligand molecules.

In a previous study, we described a new K88-related, plasmid-encoded fibrillar protein on the surface of bovine enterotoxigenic and septicemic *E. coli* strains (13). Characterization of this antigen, reported as CS31A, revealed a 29-kDa polypeptide subunit which is correlated with the abundance of very fine granular organelles, arranged without apparent order, that form a wide capsulelike structure around bacterial cells. On the basis of this finding, CS31A clearly differs from typical fimbriae, but its structure is rather reminiscent of that of the adhesive protein capsule (34) or the nonfimbrial adhesins described for uropathogenic *E. coli* (19). CS31A was observed to have antigenic relatedness to and significant homology in the N-terminal amino acid sequence with subunit K88. A more distant homology was also observed with F41, but CS31A and F41 appeared immunologically unrelated. However, unlike F41 and K88, CS31A does not show any hemagglutinin activity or in vitro adhesive properties on enterocytes of various animals.

A recent investigation into the organization and expression of the CS31A determinants reported that homology between CS31A and K88 is extensive in regions coding for the accessory proteins, although the structural genes are distinct (28). Thus, in contrast to the extensive homology reported in DNA sequences required for the expression of auxiliary proteins of K88, F41, and CS31A operons, Southern blot analysis suggested that subunit genes are entirely nonhomologous (4). On the basis of this finding, it was proposed that, in the evolution of K88 and F41 operons, a subunit gene is replaced by an entirely new subunit gene (1). We were interested in examining whether the structural genes of CS31A, K88, and F41 operons were really entirely replaced or evolved from a common ancestral gene. In addition to this, we were also interested in knowing whether functional divergence between CS31A and K88 might be correlated with a modification of the amino acid sequence in the homologous region expected to contain the K88 receptor binding domain. With this aim, we investigated the localiza-

* Corresponding author.

TABLE 1. Plasmids used in the analysis of the 5'-end region and sequencing

Plasmid	Description
pAG315	8.5-kb *Eco*RI-*Hin*dIII fragment cloned into pBR322 and coding CS31A surface protein (32)
pSS15	4.5-kb *Sma*I fragment of pAG315 subcloned in pUC19
pPVS411[a]	4.6-kb *Pvu*II fragment of pAG315 subcloned in pBluescript SK
PSP15a[a]	0.9-kb *Sph*I fragment of pAG315 subcloned in pUC19
pSSP24	1.2-kb *Sma*I-*Sph*I fragment of pAG315 subcloned in pUC19
pSSP20	2.4-kb *Sph*I-*Sma*I fragment of pAG315 subcloned in pUC19
pH5[a]	*Hin*cII deletion derivative of pSP15a
PΔ23[a]	160-bp-long *Exo*III nuclease deletion derivative of pPV411
PΔ36[a]	650-bp-long *Exo*III nuclease deletion derivative of pPV411
pΔ35[a]	1.0-kb-long *Exo*III nuclease deletion derivative of pPV411

[a] Subclone used for sequencing.

tion, structure, and complete nucleotide sequence of the *clpG* gene coding for the CS31A structural subunit. The translated amino acid sequence was compared with those of the three K88 antigenic variants and with that of F41. Amino acid alignments, hydropathy patterns, and secondary-structure predictions are discussed.

MATERIALS AND METHODS

Bacterial strains, plasmids, and media. *E. coli* K-12 strain DH1 (16) was used as the host for plasmids in this study (Table 1). Strains were grown in LB medium at 37°C. Ampicillin (50 μg/ml), chloramphenicol (10 μg/ml), tetracycline (12.5 μg/ml), and kanamycin (50 μg/ml) were used for the selection of plasmid-containing strains.

Chemicals. Restriction enzymes were purchased from Bethesda Research Laboratories, Inc., and Boehringer Mannheim Biochemicals (Mannheim, Germany). They were used in the manner specified by the suppliers. The internal primers and the N-terminal oligonucleotide probe were synthesized on a Biosearch 8600 apparatus by the Laboratoire de Virologie et d'Immunologie Moléculaire (Institut National de la Recherche Agronomique, Jouy-en-Josas, France).

DNA manipulation and analysis. Purification of recombinant plasmids was achieved by the alkaline procedure of Birnboim and Doly (2). The method of Humphreys et al. (20) was used for large-scale purification, and purification by cesium chloride density gradient centrifugation as described by Maniatis et al. (27) was done before DNA sequencing.

Agarose gel electrophoresis and isolation of restriction enzyme-generated DNA fragments were performed as described by Maniatis et al. (27). Ligations were carried out with T4 DNA ligase in ligase buffer (20 mM Tris-HCl [pH 7.5], 10 mM $MgCl_2$, 1 mM ATP), and the reaction mixture was allowed to incubate overnight at 4°C. Unidirectional deletions of plasmids were performed by use of the EXO III/Mung-Bean nuclease deletion kit (Stratagene) as recommended by the manufacturer.

Synthesis of an N-terminal oligonucleotidic probe and Southern hybridization. The previously described NH_2-terminal amino acid sequence of subunit CS31A (13) was studied for a stretch of amino acids where the codon degeneracy was minimal. On the basis of the codon usage in the nucleotide sequence of the structural gene for the K88 and F41 fimbrial subunits (1, 8), the amino acid sequence Ile-Thr-Ala-Asp-Ala-Tyr-Lys was chosen and the 21-meric oligonucleotide sequence 5'-ATC-ACT-GCA-GAT-GCA-TAT-AAA-3' was synthesized by using a Biosearch 8600 apparatus. The synthetic oligonucleotide was 5' end labeled with [γ-^{32}P]ATP by using T4 polynucleotide kinase (Boehringer Mannheim). Southern blot hybridization at high stringency was as described by Maniatis et al. (27). On the basis of two possible mismatches, hybridization was carried out at a T_m of 41°C, which was 10°C below the hybridization temperature of the oligonucleotide.

DNA sequencing and sequence analysis. DNA fragments to be sequenced were subcloned into sequencing vector pUC19 or Bluescript and transformed in *E. coli* DH1. Sequencing of double-stranded DNA templates was performed by the dideoxy chain termination method of Sanger et al. (35) with Sequenase-modified T7 polymerase enzyme and a sequencing kit with [α-^{35}S]dATP (U.S. Biochemical Corp., Cleveland, Ohio).

Determination of the 5' end of the *clpG* transcript by primer extension analysis. Primer extension analysis was carried out as described by Maniatis et al. (27) with [γ-^{32}P]dATP as the radioactive label and total cellular RNA prepared by hot phenol extraction (9). The transcriptional initiation site was determined by the primer extension method by using reverse transcriptase. The synthetic 30-meric oligonucleotide used is complementary to the region from nucleotides 430 to 459.

Computer analysis. Sequence analyses were performed on a VAX 8530 computer by using the software of the Centre Inter Universitaire d'Informatique à Orientation Biomédicale (CITI 2). They were compared with the GenBank genetic sequence data and EMBL nucleotide sequence library data. The deduced protein sequence was compared with sequences of the SWISS-PROT and NBRF protein sequence data banks. Hydrophobicity pattern and prediction of antigenic determinants were determined by using the algorithm developed by Hopp and Woods (18), by averaging over six amino acid residues. Secondary-structure predictions were made by using the method of Gibrat et al. (12).

Nucleotide sequence accession number. The GenBank accession number for the *clpG* gene sequence is M55389.

RESULTS

Determination of the location of the 5' end of the *clpG* gene. The cloning of CS31A determinants in an 8.5-kb *Eco*RI-*Hin*dIII fragment in the recombinant plasmid pAG315 and the structural organization of CS31A genes were previously described (28). Strong similarities between the structural organization of the CS31A and K88 gene clusters were observed, but the CS31A subunit gene was not clearly located. To define the region which codes for the CS31A subunit, we have subcloned various restriction enzyme fragments from the plasmid pAG315 into pUC19 and pBluescript and tested each of them for hybridization with the N-terminal oligonucleotide probe to localize the 5' end of the CS31A structural gene. Table 1 lists the plasmids used, and Fig. 1A shows their relative positions on a restriction map of pAG315. The probe hybridized by Southern blotting with pPVS411, pSS15, and an 0.9-kb *Sph*I fragment subcloned in pSP15a. No reaction occurred with pSSP24 and pSSP20. This localized the 5'-end region within the two *Sph*I sites of

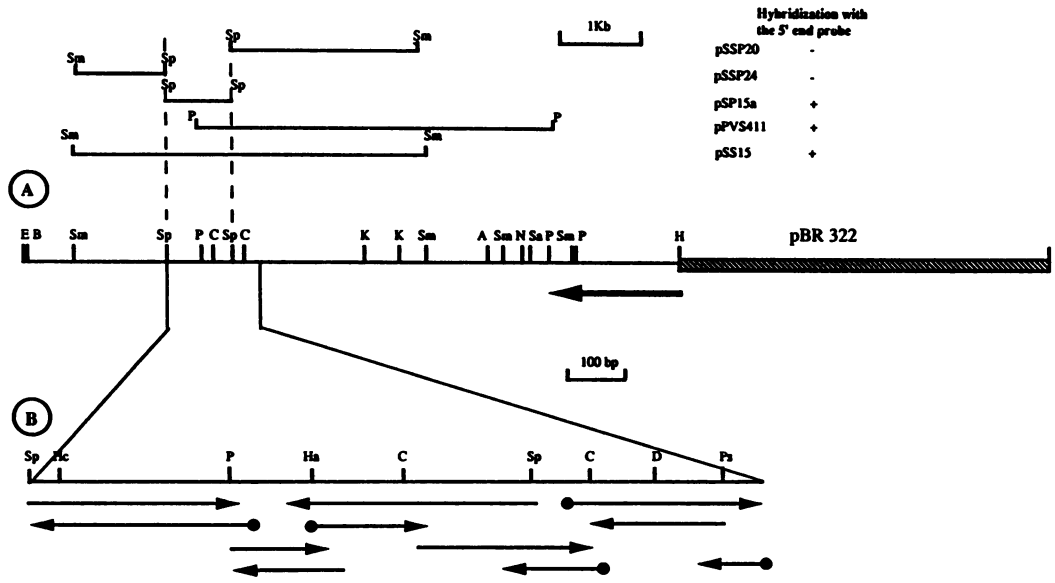

FIG. 1. (A) Restriction map of the 8.5-kb *Eco*RI-*Hin*dIII DNA fragment cloned in pBR322 that codes for production of CS31A. The bars above the restriction map denote various subcloned DNA fragments described in Table 1 that were used to localized the 5'-end region of *clpG*. Hybridization of subclones with the 5'-end probe are indicated. The direction of transcription is indicated by an arrow. (B) Restriction map and sequencing strategy of the 1.3-kb DNA fragment containing *clpG*. Horizontal arrows indicate the direction and extent of each sequenced clone. Horizontal arrows with a circle mark positions where a synthetic oligonucleotide was used for sequencing.

the *Sma*I fragment of pSS15 plasmid. Hybridization with the *Pvu*II fragment (pPVS411) suggested that the 5' end of the structural gene was located within the 0.5-kb *Sph*I-*Pvu*II fragment. A more precise localization was not possible because of the lack of suitable restriction sites within this fragment.

DNA sequencing strategy. Figures 1B and 2 show the sequencing strategy and DNA sequence of the *clpG* gene, respectively. On the basis of the molecular size of the CS31A subunit (29 kDa) and the amino acid composition, it has been previously shown that the CS31A subunit consists of approximately 243 amino acids (13), assuming that a minimal region of 800 bp is needed to code for the structural gene. The 0.9-kb *Sph*I-*Sph*I fragment concluded to code for the 5'-end region was subjected to nucleotide sequencing by the dideoxy method. Sequence information obtained from both ends by using the M13 (−20) and M13 reverse primers and compared with the nucleotide sequence deduced from the N-terminal amino acid sequence of the CS31A polypeptide indicated that the subunit gene is transcribed from right to left (Fig. 1) and starts close to the right *Sph*I site of the 0.9-kb fragment of the pSP15a plasmid.

In agreement with the structural organization of the CS31A gene cluster previously described (28), we further concluded that this fragment coded the entire *clpG* gene and confirmed the high relatedness in structural organization with the K88 operon. A series of overlapping subclones of this region was constructed by restriction enzyme deletions in pUC19, and their nucleotide sequences were determined. The nucleotide sequence that extended beyond the right *Sph*I cleavage site was obtained from the 4.6-kb *Pvu*II-*Pvu*II fragment of pPVS411 reduced in size by *Exo*III digestion to a 1.0-kb fragment (pΔ35) carrying the conserved *Pvu*II end site. Regions without suitable restriction sites were sequenced by using five internal 17-meric oligonucleotide primers.

Nucleotide sequence analysis. Nucleotide sequence analysis of 1,318 bp revealed the presence of a single open reading frame (ORF) of 834 bp starting at an ATG codon located 60 nucleotides upstream of the right *Sph*I site and capable of translating into 278 amino acids of a 28,780-Da polypeptide (Fig. 2). At the 3' end, three stop codons terminated the ORF at positions 1218, 1227, and 1245, and two regions of dyad symmetry at positions 1223 to 1242 and 1272 to 1316 were observed. The transcript of this region has the potential to form a stable tandem of stem-loop structures supposed to be the transcription terminator of the *clpG* gene.

At the 5'-end region, the translation initiation codon was preceded by a sequence coding for a putative ribosome binding site located 8 nucleotides upstream of the ATG codon. Computer analysis of the nucleotide sequence in the region surrounding *clpG* did not reveal an ORF in the sequence extending beyond 300 bp of the start codon. Two large inverted repeat sequences would be predicted to form stem-loop structures (Fig. 3) with stabilities of −32 and −13 kcal (1 cal = 4.184 J), according to the algorithm of Zuker and Stiegler (37). The first potential stem-loop was located within the large intercistronic region at positions 240 to 314, whereas the second inverted repeat sequence flanked the putative ribosome binding site of the *clpG* gene.

The amino acid sequence of the peptide Ile-Thr-Ala-Asp-Ala-Tyr-Lys chosen for the N-terminal probing was found at positions 16 to 22. On the basis of the codon usage in the nucleotide sequence of the fimbrial subunit gene, the deduced oligonucleotide sequence of the 21-meric probe contained three mismatches (ATC to ATT, GCA to GCT, and GCA to GCG). However, such changes did not significantly affect the specificity of the N-terminal probe used in the Southern blot hybridization. The amino-terminal sequence of the purified CS31A protein (13), previously determined by Edman degradation, was identical through the first 25 residues, with only one change at the first residue, from glycine,

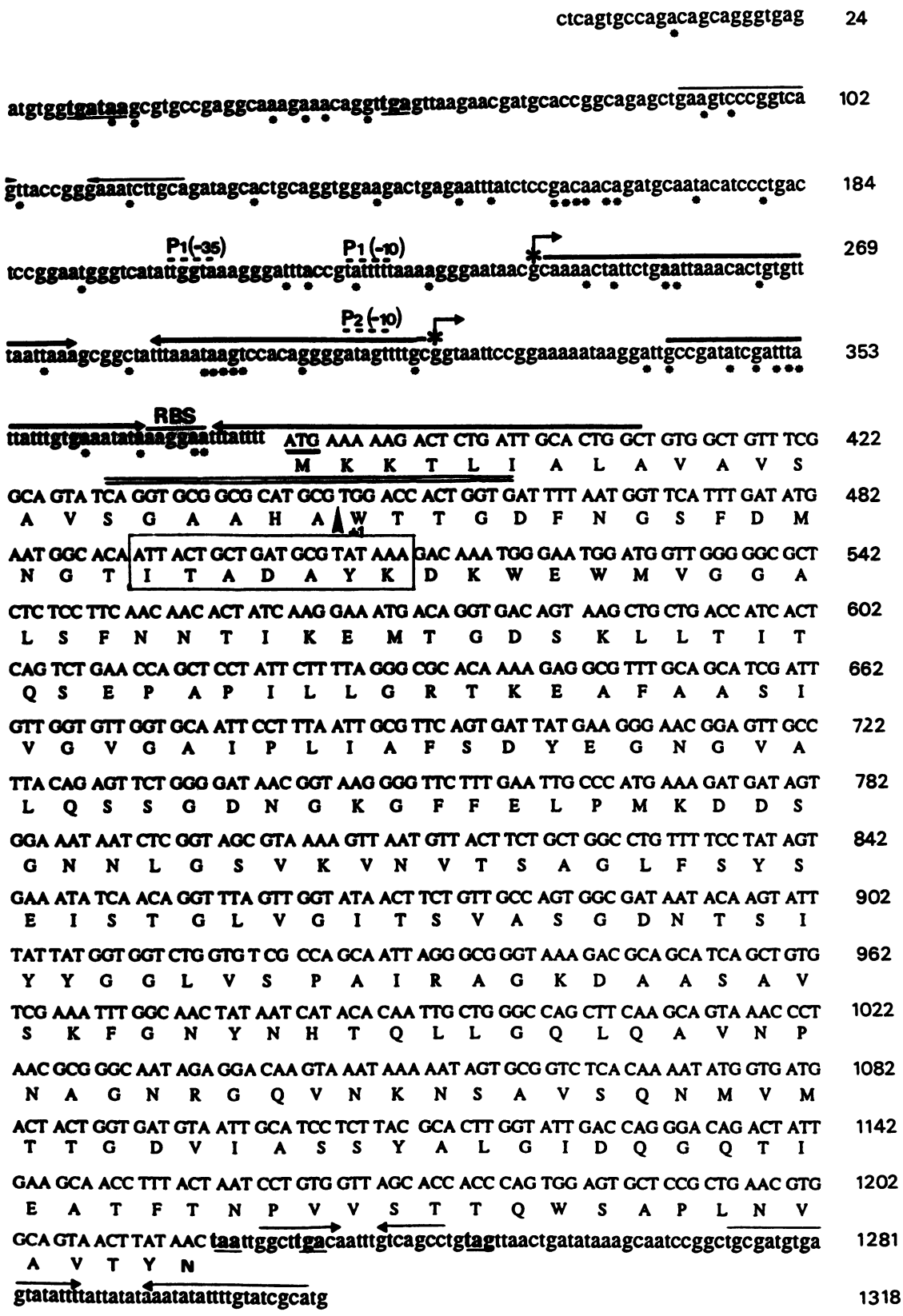

FIG. 2. Nucleotide and deduced amino acid sequences of the *clpG* gene. A putative ribosome-binding site (RBS) is indicated. The start codon is underlined. Thick arrows above the sequence denote the two tandemly inverted repeat sequences in the untranslated 5′ region of *clpG*. Thin arrows after the translational stop indicate inverted repeat sequences which could function as the terminator of transcription. The putative −35 and −10 sequences of promoters P1 and P2 are underlined. The two start sites of transcription are indicated by asterisks (*). The vertical arrowhead indicates the site of leader peptide cleavage. The double line above nucleotides 430 through 459 indicates the position of the primer used in the primer extension experiment, and the sequence chosen as the N-terminal probe is boxed. The asterisks (✸) in the 5′ untranslated region indicate base changes between CS31A and K88 nucleotide sequences.

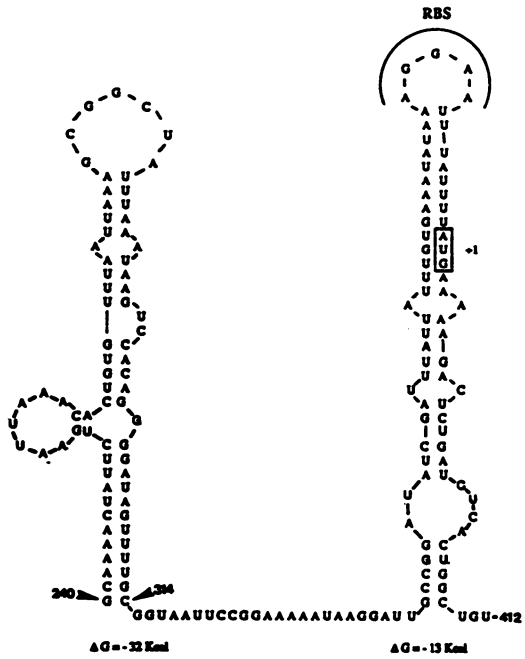

FIG. 3. Secondary-structure model for the leader region of the larger *clpG* mRNA. The first stem-loop derived from nucleotide positions G*-240 (corresponding to the start site of the larger transcript) to C*-314 (corresponding to the start site of the shorter transcript) is predicted to have a stability of −32 kcal. The second putative stem-loop structure, derived from nucleotides 388 to 409, is predicted to have a stability of −13 kcal. The AUG initiation codon for *clpG* and the ribosome binding site (RBS) are indicated.

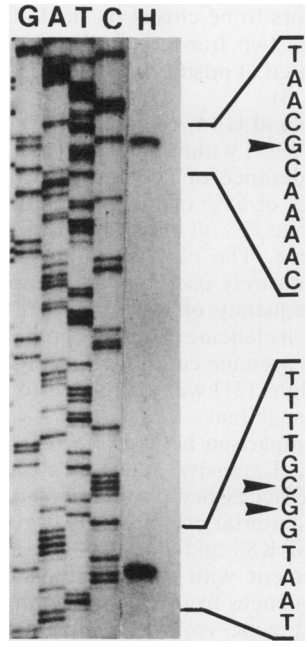

FIG. 4. Primer extension analysis of *clpG* transcription. A γ-^{32}P-labeled 30-meric oligonucleotide complementary to the sequence underlined in Fig. 2 was hybridized with the total RNA of *E. coli* DH1(pAG315) and then extended by reverse transcriptase. The resultant products were analyzed on a sequencing gel in parallel with sequence obtained with the same primer. Lanes G, A, T, and C, dideoxy sequencing reactions; lane H, primer extension products. The sequence in the region of the transcription start sites is shown, and initiation sites are indicated by horizontal arrows.

as determined by chemical analysis, to tryptophan, as determined by nucleotide sequencing. As contamination by a glycine-containing buffer is often responsible for uncertain chemical determination when glycine is found as the first residue, tryptophan, deduced from nucleotide sequencing, was assumed to be the first amino acid of the CS31A polypeptide subunit.

Since the CS31A polypeptide is exported at the bacterial surface, it can be expected that it is produced in a precursor form. The 30-kDa polypeptide previously identified in minicell analysis has been envisaged as this precursor (28). These previous results indicated the presence of a signal sequence needed for the transmembrane secretion of the polypeptide. A signal peptidase cleavage site was deduced from the N-terminal amino acid analysis to exist between residues Ala and Trp + 1. Cleavage at this site would result in a mature polypeptide of 257 amino acids, assuming a molecular size of 26,777 Da, with a signal sequence of 21 amino acid residues showing properties of a typical prokaryotic signal sequence. The 21-residue peptide included two positively charged residues (Lys-Lys) which were followed by a stretch of hydrophobic amino acids, ending at an alanine (Fig. 2).

The predicted amino acid composition of the encoded protein was comparable to the previously reported amino acid composition of the purified CS31A polypeptide (13), except for the aspartic acid determination, which resulted in a net overvalue in chemical analysis. The rich content of the deduced protein in asparagine can explain this discrepancy, since acidic hydrolysis resulted in an asparagine-to-aspartic-acid change. In agreement with the chemical determination, no cysteine was found in the CS31A subunit polypeptide.

Mapping of the 5′ end of the *clpG* mRNA. Primer extension analysis was carried out to map the 5′ end of the *clpG* mRNA. A γ-^{32}P-labeled 30-meric synthetic oligonucleotide complementary to the region extending from nucleotides 430 to 459 of the sequence shown in Fig. 2 was hybridized with the total RNA of *E. coli* DH1(pAG315) and then extended by reverse transcriptase. The resultant products were analyzed on a sequencing gel run in parallel with a sequence of this region obtained with the same primer. As shown in Fig. 4, two primer extension products were detected corresponding to a transcription initiation starting at the G at position 240 for the larger transcript and at C or G at positions 314 and 315 for the shorter transcript. This suggested that transcription of the *clpG* gene originates from two distinct start sites 76 bp apart and that initiation of transcription at these positions results in two transcripts with leader regions 79 and 155 bases in length preceding the coding region. Although examination of the nucleotide sequence upstream from each start point of transcription revealed only the larger transcript sequences that were homologous with the consensus sequences for the −35 and −10 regions of *E. coli* promoters, it was proposed that two promoters, P1 and P2 (Fig. 2), exist for the *clpG* gene. Otherwise, the nucleotide sequence extending within the two transcription initiation sites has the potential to form the extensive secondary structure, described above, at positions 240 to 314. The high stability (−32 kcal) predicted by the algorithm of Zuker and Stiegler (37) suggests that the larger transcript has the potential to form a stable stem-loop structure. Intriguingly, this possible

structure appears to be closed by the G-C base pairs corresponding to the two transcription initiation sites described above and located at positions 240 and 314 on the nucleotide sequence (Fig. 3).

Codon usage and G+C content of the *clpG* gene. The high A+T content (58%) within the *clpG* gene was due primarily to the preponderance of adenine- and thymine-terminated codons. A total of 65% of the variable third-base positions were A+T, while *E. coli* only uses A+T-terminated codons 43% of the time. The *clpG* gene makes extensive use of codons that are rarely used by highly expressed genes in *E. coli* (codon frequency of less than 1%). This was notably observed for isoleucine (ATA), proline (CCC), glycine (GGA), and all arginine codons (AGA-AGG). The frequency of optimal codons (21) was calculated to be 0.61, indicating a low-codon-usage bias.

Sequence comparison between CS31A, K88, and F41 fimbrial subunits. Extensive nucleotide sequence homology throughout the accessory protein-coding genes for CS31A, K88, and F41 fimbrial proteins was previously reported by comparisons of K88 and F41 (33) and of CS31A and K88 (28). In agreement with these results, comparison of the nucleotide sequences from the region immediately upstream of the subunit gene reveals a high degree of homology between CS31A and F41 and between CS31A and K88 (93 and 88%, respectively). The homology extends into nucleotides of the region encompassing the beginning of the coding sequence. In such, a sequence of CS31A was identical through the first 11 codons of the signal sequence of F41 and to the first 7 codons of K88. In contrast with general diversity in fimbrial signal sequence, a highly conserved (80%) signal sequence was observed for CS31A, K88, and F41, with an identical cleavage site for CS31A and K88 prepilins (Fig. 5A).

The nucleotide sequence of the region coding for the mature CS31A polypeptide showed a similarity of 30 and 60% with F41 and K88 structural genes, respectively. However, in spite of this nucleotide sequence homology, a negative reaction was observed in Southern blot analysis when heterologous subunit genes were used as probes (4, 28).

When compared, the predicted amino acid sequence of the mature CS31A fimbrial subunit showed obvious homology with those of the K88 and F41 subunits and yielded the alignment shown in Fig. 5B. By using the program ALIGN (PIR [7, 11]), global homologies of 46 and 24% of amino acid sequence identity were noted with K88 and F41, respectively. Alignment scores of 29 with K88 and of 8.82 with F41 indicate significant structural similarity with the CS31A subunit. The frequency of amino acid change between CS31A and K88 revealed highly conserved and variable regions (Fig. 5B). As commonly observed within primary structures of numerous fimbrial subunits, a high degree of homology was observed in the N- and C-terminal regions (60 and 65%, respectively), suggesting that structural constraints are imposed on these regions.

When the amino acid sequence of the CS31A subunit was compared with that of K88, five highly conserved regions, noted as P1 to P5 on Fig. 5B and containing as many as 7 to 19 identical (plus conserved) amino acid residues, were found at the same position within each of the sequences. Except for the conserved amino acid cluster P3, the same conserved regions were found for the CS31A and F41 amino acid sequences. Moreover, when the distribution of the different conserved amino acid clusters was examined, each of the clusters was observed to be associated with a conserved proline for the CS31A, K88, and F41 fimbrial subunit sequences (Fig. 5b and see Fig. 7), suggesting that these regions are extremely important for protein folding or stability.

Outside of the terminal region and conserved clusters, three large, variable regions noted as V1, V2, and V3, containing as many as 18, 30, and 35 amino acid residues located at positions 31 to 49, 123 to 151, and 186 to 221, respectively, were found when CS31A and K88 sequences were compared (see Fig. 7). The extensive changes in the central variable regions (V2) of fimbrial subunits of CS31A and K88 contrast with the adjacent highly conserved regions P3 and P4. No significant homology is apparent between the amino acid sequences of CS31A, K88, and F41 in the V2 region. However, the octapeptide sequence VTSAGLFS, located within the V2 region of CS31A at position 123 to 130, is similar to the sequence at position 139 to 147 in the homologous region of the three K88 variants (Fig. 6). Although the nucleotide sequence of this octapeptide was highly conserved (90%) in the fimbrial subunit gene of CS31A and K88, it was not possible to know whether this correspondence is fortuitous or not. Interestingly, the two tripeptides Ser-Leu-Phe and Ala-Ile-Phe involved in the K88 binding activity (22) were not found in the primary structure of CS31A.

Immediately downstream of this large unconserved region, at positions 152 to 157 extends a strong hydrophobic hexapeptide IYYGGL which is highly conserved among CS31A and K88 and associated with proline at position 160. The tyrosine tandem (positions 153 and 154) in the primary structure of CS31A and conservation of tyrosine 154 among CS31A and the three K88 variants suggest the involvement of this area in the maintenance or functions shared by CS31A and K88 subunits. In the F41 sequence, tyrosine 154 was lost but might be conservatively substituted at this position by a phenylalanine residue.

Prediction of hydropathy pattern and secondary structure. We have applied the hydrophobicity analysis of Hopp and Woods (18) to the sequence coding for the CS31A subunit and compared its pattern with patterns obtained for K88 and F41. Despite large changes in amino acid sequences, hydropathy profiles of CS31A and K88 were remarkably alike (Fig. 7). The order and location of hydrophobic and hydrophilic regions were highly conserved, especially at the half N-terminal part where major hydrophilic domains are clustered. Furthermore, hydropathy analysis revealed that all five of the proline-associated conserved regions (P1 to P5) are located within or immediately adjacent to hydrophobic domains and suggests that these regions form the common hydrophobic core of the two proteins (3).

When CS31A and F41 sequences were compared, a more distant relatedness was observed in hydropathy profiles. However, despite the extensive change in amino acid sequences, hydropathy analysis revealed a conserved order of hydrophobic and hydrophilic domains within homologous regions. This finding was likely observed at the N-terminal part of CS31A and F41 in a large region extending from position 31 to position 70 which includes variable (V1) and conserved (P1) regions (Fig. 7). Despite extensive differences in the amino acid sequences in V1, the order of hydrophilic domains seems similar. This suggests that this region contains a possible site for antigenic determinants involved in antigenic specificity of CS31A and F41.

The empirical secondary-structure predictions of Gibrat et al. (12) were used to estimate the secondary-structure similarity of CS31A, K88, and F41 structural subunits. To

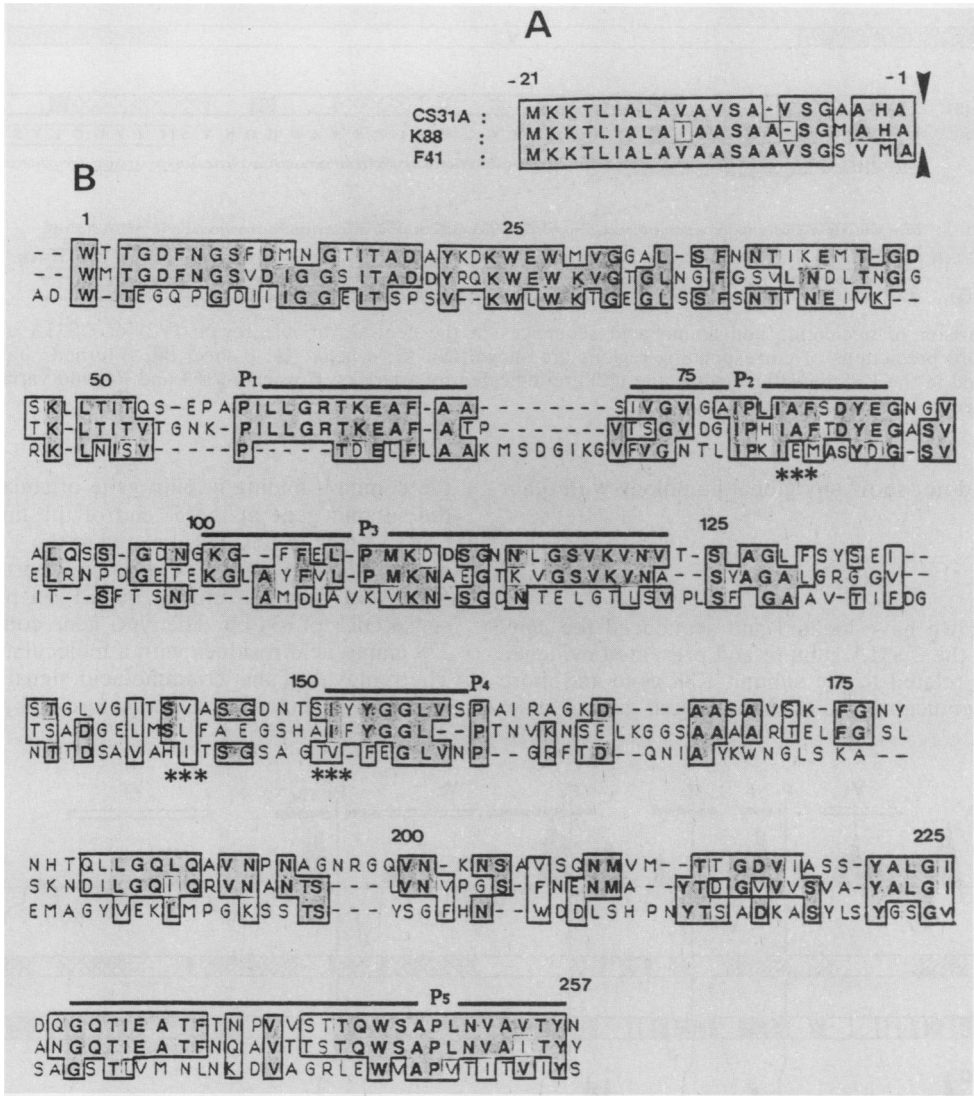

FIG. 5. (A) Amino acid sequences of the signal peptides of CS31A, K88, and F41 prepilins. Vertical arrowheads indicate the site of the leader peptide cleavage. (B) Comparison of the primary structure of the mature CS31A fimbrial subunit with K88 and F41 fimbrial subunits. Identical plus functionally similar amino acids are boxed, and gaps have been introduced to obtain maximal fitting. Identical residues are identified by shading. Numbering corresponds only to the CS31A sequence. The five conserved regions associated with conserved proline are noted (P1 to P5). Asterisks indicate positions of the tripeptides involved in the K88 receptor binding site. Residues ST, MVLI, and WFY are assumed to be functionally similar (3).

improve the accuracy of the sequence alignment and significance of sequence similarity, rather than folding information, the number and order of the secondary-structure elements were considered. Despite extensive differences in primary sequences, high similarity in overall structure was observed on the basis of the secondary-structure prediction analysis (Fig. 7). Except for the central variable region V2, all three proteins could be aligned in a similar secondary-structure prediction pattern that followed the conserved hydropathy profiles and conservation of amino acid clusters associated with the proline residue. Sequence predictions for the central region V2 suggest for CS31A an extensive β-sheet configuration interspersed by a β-turn that contrasts with a high potential for an α-helical configuration predicted for the K88 homologous region (Fig. 6). The rich serine-threonine content (40%) of the CS31A V2 region may contribute to the structural diversity predicted in this region. Despite considerable sequence variations, a common β-sheet prediction may contribute to structural identity in the C-terminal region of the three proteins.

Homology with other fimbrial subunits. Comparison of the sequence of the first 24 N-terminal amino acid residues of the CS31A subunit with those of adhesins recently described (Fig. 8) revealed a striking homology (71%) with NFA-4, a fibrillar adhesin newly characterized on human uropathogenic *E. coli* (19), and a more distant but significant relatedness (30%) with PCF-09, a newly described adhesin on human enteropathogenic *E. coli* (17). Relatedness with CS31A was highly supported by similar molecular weights and morphology shared by NFA-4 and PCF-09.

Comparison, by the program FASTP (26), of the encoded amino acid sequence of CS31A fimbrial subunit with those of other proteins included in the NBRF and SWISS-PROT data

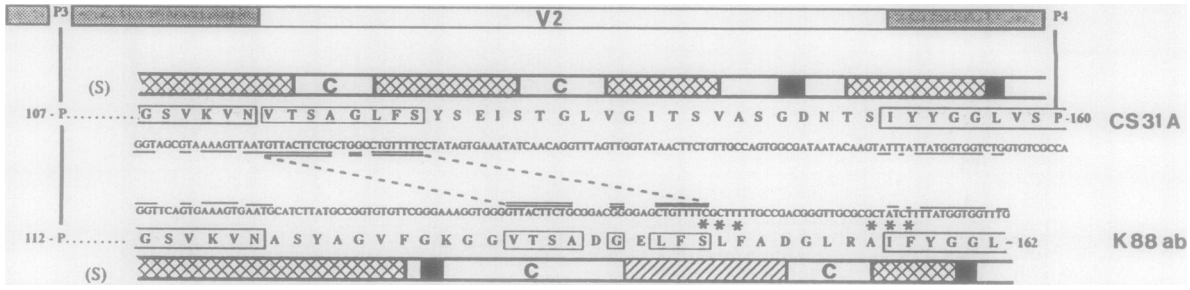

FIG. 6. Comparison of nucleotide and amino acid sequences in the central variable region (V2) of CS31A and K88ab subunits. Secondary-structure predictions of corresponding regions are shown (S): ▨, α-helix; ▩, β-sheet; ■, β-turn; C, random coil. The two tripeptides involved in the K88 receptor binding site (22) are indicated by asterisks. Conserved (P3 and P4) and variable (V2) regions are indicated by shaded and open bars, respectively.

base libraries did not show any global homology with other sequences.

DISCUSSION

In this study, we have located and sequenced the *clpG* gene coding for the CS31A subunit and presented evidence that it is highly related to the subunit K88 gene and more distantly, but significantly, to the F41 subunit gene. Unlike the common finding in pilin gene organization that located the subunit gene at the 5′ end of the fimbrial operon, the CS31A fimbrial gene was located at the 3′ end of the CS31A operon. Sequencing of the 5′ end portion of the 0.9-kb *Sph*I-*Sph*I fragment cloned in pSP15a plasmid revealed a single ORF of 834 bp. The *clpG* gene codes for a protein of 278 amino acid residues with a molecular weight of 28,784. The removal of the 21-amino-acid signal sequence peptide would be expected to yield a mature fimbrial protein subunit

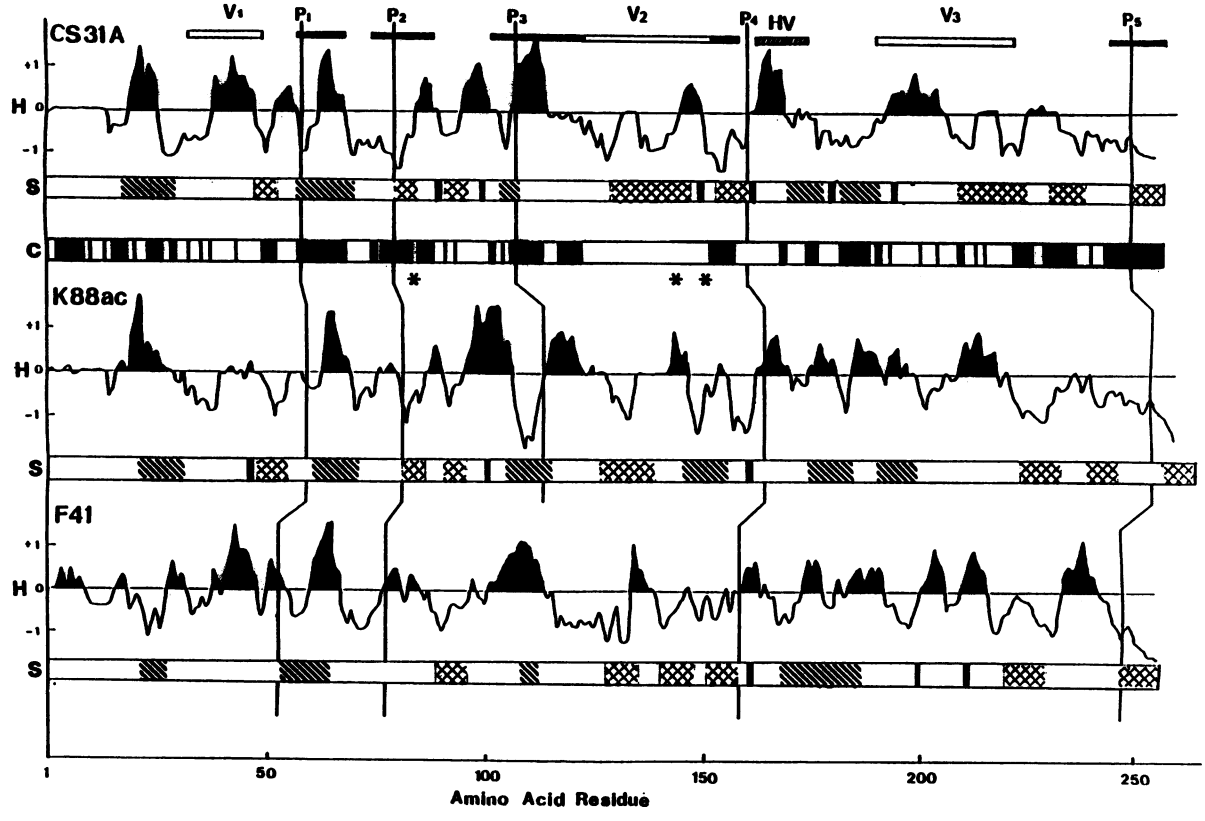

FIG. 7. Protein sequence analysis of CS31A, K88ac, and F41 fimbrial subunits. H, hydropathy profiles determined by using the Hopp and Woods algorithm (18), with positive and negative indices representing the hydrophilic and hydrophobic domains, respectively. S, possible secondary structures predicted by the program GOR III by using the algorithm of Gibrat et al. (12). ▧, α-helice; ▩, β-sheet; ■, β-turn. Only regions with predictions of high potential secondary structures are noted. C, location of identical amino acids in CS31A and K88 subunits (indicated by black bars). Vertical lines are positioned at each of the five conserved proline residues. Conserved (P1 to P5), variable (V1, V2, V3), and hypervariable (HV) regions are noted by closed, open, and dashed bars, respectively. Asterisks indicate the positions of the tripeptides involved in the K88 receptor binding site.

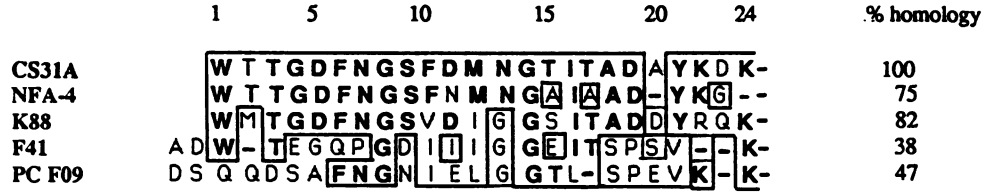

FIG. 8. Comparison of N-terminal sequences of CS31A, K88, F41, NFA-4, and PCF-09 (17, 19). Identical plus functionally similar amino acids are boxed and represent the global homology determined when these sequences were compared with the CS31A sequence.

of 257 residues with a molecular weight of 26,777. This value is approximately 10% smaller than that estimated for the 29.5-kDa protein by sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE) that was previously reported. However, the abnormal migration in SDS-PAGE appeared as a common feature in CS31A, K88, and F41 fimbrial subunits. It might be explained by the reduction of the charge/mass ratio of the protein-SDS complex or by local preferential conformation adopted by the dissociated subunits.

The results presented here might indicate that the *clpG* gene is monocistronic. Transcription initiated at two start sites 79 or 155 bp upstream of the translational start, and two sequences resembling transcriptional terminators were located at positions 925 to 942 and 972 to 1017 immediately downstream of three translational stop codons. The primer extension analysis revealed two apparent transcription start sites 76 bp apart, and it would seem that transcription proceeds from two tandemly located promoters P1 and P2. The larger cDNA is initiated 34 bases downstream from an *E. coli* consensus −35 promoter sequence (P1). Although the shorter cDNA appeared more abundant, no recognizable −35 promoter sequence was observed at the appropriate distance from its putative start site. However, Northern (RNA) blot analysis used to identify the mRNA transcribed in vivo from the pAG315 plasmid in strain DH1 detected two transcripts which are in size in good agreement with the locations of the two transcription initiation sites observed with the primer extension experiment (unpublished data). From these results, it would be expected that a tandem of promoters was involved in transcription of the *clpG* gene. Furthermore, start sites on G-240 and on C-314 for the larger and shorter transcripts, respectively, appear at positions expected to close the putative stem-loop structure which extends from position 240 to position 314 (Fig. 3). A second putative stem-loop structure which extends from position 332 to position 404 involved the ribosome-binding site and the translational initiation codon of the *clpG* gene. Regions of dyad symmetry sequestering the ribosome-binding site and the ATG codon are present on CFA/I, K99, K88, and PAP operons and are thought to be involved in the regulation of translation of a gene coding for auxiliary proteins or in temperature-dependent regulation of translation (15, 29). However, whether or not the secondary structural features of the leader region are involved in regulation of *clpG* at the transcription or/and translational level is not known. Studies on the regulation of the expression are currently in progress.

Comparison of nucleotide sequences from the region immediately upstream of the three K88-related subunit genes reveals a high homology (80%) in the 400 nucleotides that precede the ORF. This was in agreement with our previous data which indicate that genes coding for CS31A, K88, and F41 determinants are entirely homologous (28). Furthermore, conservation of the signal sequence in concert with high homology (70%) in the first 30 residues of the subunit indicate extreme functional pressure to conserve this sequence. This suggests that the presubunit region might act as a topologic signal by which nascent presubunit molecules are recognized for transfer by a recognition system homologous to CS31A, K88, and F41 determinants.

Comparison of the predicted amino acid sequences of mature CS31A and K88 fimbrial subunits indicates a high similarity (46%) and reveals the presence of highly conserved, variable, and hypervariable regions. The hydrophobic carboxy-terminal part was the most highly conserved region (65%) between CS31A and K88, with conservation in both sequences of the penultimate tyrosine residue, assessed as essential for the processing and periplasmic transport of mature fibrillar subunit (36). Despite extensive changes in the internal region, four different conserved clusters of up to 19 identical residues were observed between CS31A and K88. Thus, conservation of terminal parts, in concert with the occurrence of conserved clusters and their association with each of the conserved proline residues, suggests a common pattern of folding for CS31A and K88. On the basis of hydropathy and amino acid sequence analysis, the conserved cluster P1 may be the common continuous antigenic determinant previously detected on CS31A and K88 by Western blot (immunoblot) analysis with antibodies raised against denatured CS31A subunit (13). Regions which are predicted to be antigenic on the basis of their flexibility and hydrophilicity coincide with clusters of changes at positions 38 to 48, 94 to 101, and 192 to 202 and probably contribute to the antigenic specificity of CS31A and K88. This was in agreement with the absence of cross-reactivity observed by Western blot analysis between CS31A and specific antibodies anti-b, -c, and -d determinants of the respective K88 antigenic variants (unpublished data).

On the basis of the absence of hybridization in Southern blot analysis with the heterologous internal subunit probe (4) and the absence of immunological relatedness between F41 and K88, it was concluded that there was a complete lack of homology between these fimbrial proteins. From this, authors have suggested that entire subunit genes have been replaced during the evolution of K88 and F41 (24, 33). Furthermore, automatic alignment between K88 and F41 or CS31A and F41 was impossible because of the extensive change in amino acid sequence. Manual alignment based on the position of the conserved proline residue yielded a weak sequence identity value (24%) which cannot support or rule out structural homology between F41 and the K88-related proteins. However, despite extensive substitutions in the primary sequence, our results, based on (i) the significant alignment score (8.82) obtained with the program ALIGN (after manual alignment), (ii) the conservation of all amino acid clusters associated with conserved proline, (iii) the similar pattern of hydropathy in the N terminus, and (iv) the similarity of secondary-structure patterns, form evidence to

indicate that F41 has evolved with K88 and CS31A from a common ancestral gene.

Disulfide bridges in many pilins presumably maintain the local structural integrity in disulfide loops which often show extensive differences between serotypic variants. For the three K88-related subunits, which are characterized by the absence of cysteine, a comparison of amino acid sequences suggests that the conserved hydrophobic amino acid clusters associated with proline (P1 to P5) may have a similar function to that of disulfide bridges. Presumably, they reflect structural constraints in regions that are supposed to form the hydrophobic core of the proteins. Despite extensive substitutions, in concert with the conserved terminal regions, they contribute to maintain the local folding and the structural integrity of the molecules. In addition to the conservation of the hydrophobic core, the similarity of the secondary-structure prediction patterns suggests a common structure for the CS31A and K88 structural subunits.

Extensive substitutions in variable regions (V1, V2, and V3) presumably contribute to serotypic and/or functional differences. Extensive differences in the V2 region result in the CS31A subunit primary sequence losing the two tripeptides Ser-148–Leu–Phe-150 and Ala-156–Ile–Phe-156 which are involved in the K88 receptor binding domains (22). The possibly low structural constraints on the central region V2 may have contributed to the observed diversity of the K88-related subunits and to the apparent absence of an adhesin function in CS31A.

E. coli strains with the surface antigen CS31A appear to be pathogenic in an animal model (6) and are frequently isolated from septicemic animals and from some cases of human septicemia (5). Although the CS31A antigen was highly expressed in the intestinal lumen of animals, the role of this fimbrial protein in virulence remains to be determined. In any event, CS31A belongs to a group of fimbrillar proteins with an interesting evolutionary history, which have conserved a compatible mechanism of pilin processing and assembly but which possibly possess an interchangeable host-fimbria function relationship.

ACKNOWLEDGMENT

This work was supported by Eclair program grant AGREE-008 from the European Economic Community.

REFERENCES

1. **Anderson, D. G., and S. L. Moseley.** 1988. *Escherichia coli* F41 adhesin: genetic organization, nucleotide sequence, and homology with the K88 determinant. J. Bacteriol. **170**:4890–4896.
2. **Birnboim, H. C., and J. Doly.** 1979. A rapid alkaline extraction procedure for screening recombinant plasmid DNA. Nucleic Acids Res. **7**:1515–1523.
3. **Bowie, J. U., J. F. Reidhaar-Olson, W. L. Lim, and R. T. Sauer.** 1990. Deciphering in the message in protein sequences: tolerance to amino acid substitutions. Sciences **247**:1306–1310.
4. **Casey, T. A., S. L. Moseley, and H. W. Moon.** 1990. Characterization of bovine septicemic, bovine diarrheal, and human enteroinvasive *Escherichia coli* that hybridize with K88 and F41 accessory gene probes but do not express these adhesins. Microb. Pathog. **8**:383–392.
5. **Cherifi, A., M. Contrepois, P. Picard, P. Goullet, J. De Rycke, J. M. Fairbrother, and J. Barnouin.** 1990. Factors and marker of virulence in *Escherichia coli* from human septicemia. FEMS Microbiol. Lett. **70**:279–284.
6. **Contrepois, M., J. M. Fairbrother, Y. K. Kaura, and J. P. Girardeau.** 1989. Prevalence of CS31A and F16S surface antigens in *Escherichia coli* isolates from animals in France, Canada and India. FEMS Microbiol. Lett. **59**:319–324.
7. **Dayhoff, M. O., W. C. Barker, and L. T. Hunt.** 1983. Establishing homologies in protein sequences. Methods Enzymol. **91**: 524–545.
8. **Dykes, C. W., I. J. Halliday, M. J. Read, A. N. Hobden, and S. Harfard.** 1985. Nucleotide sequence of four variants of the K88 gene of porcine enterotoxigenic *Escherichia coli*. Infect. Immun. **50**:279–283.
9. **Ebina, Y., and A. Nabajawa.** 1983. Cyclic AMP-dependent initiation and rho-dependant termination of colicin E1 gene transcription. J. Biol. Chem. **258**:7072–7088.
10. **Gaastra, W., and F. K. de Graaf.** 1982. Host-specific fimbrial adhesins of noninvasive enterotoxigenic *Escherichia coli* strains. Microbiol. Rev. **46**:129–161.
11. **George, D. G., W. C. Barker, and L. T. Hunt.** 1986. The protein identification resource (PIR). Nucleic Acids Res. **14**:11–15.
12. **Gibrat, J. F., J. Garnier, and B. Robson.** 1987. Further developments of proteins secondary structure prediction using information theory. J. Mol. Biol. **198**:425–443.
13. **Girardeau, J. P., M. Der Vartanian, J. L. Ollier, and M. Contrepois.** 1988. CS31A, a new K88-related fimbrial antigen on bovine enterotoxigenic and septicemic *Escherichia coli* strains. Infect. Immun. **56**:2180–2188.
14. **Guinée, P. A. M., and W. H. Jansen.** 1979. Behavior of *Escherichia coli* K antigens K88ab, K88ac, and K88ad in immunoelectrophoresis, double diffusion, and hemagglutination. Infect. Immun. **23**:700–705.
15. **Hamers, A. M., J. P. Herman, G. A. Willshaw, J. G. Kusters, B. A. M. van der Zeijst, and W. Gaastra.** 1989. The nucleotide sequence of the first two genes of the CFA/I fimbrial operon of human enterotoxigenic *Escherichia coli*. Microb. Pathog. **6**:297–309.
16. **Hanahan, H.** 1983. Studies on transformation of *Escherichia coli* with plasmids. J. Mol. Biol. **166**:557–580.
17. **Heuzenroeder, M. W., T. R. Elliot, C. J. Thomas, R. Halter, and P. A. Manning.** 1990. A new fimbrial type (PCF09) on enterotoxigenic *Escherichia coli* 09:H-LT+ isolated from case of infant diarrhea in central Australia. FEMS Microbiol. Lett. **66**:55–60.
18. **Hopp, T. P., and K. R. Woods.** 1981. Prediction of protein antigenic determinants from amino acid sequences. Proc. Natl. Acad. Sci. USA **78**:3824–3828.
19. **Hoschützky, H., W. Nimmich, F. Lottspeich, and K. Jann.** 1989. Isolation and characterization of the non-fimbrial adhesin NFA-4 from uropathogenic *Escherichia coli* O7:K98:H6. Microb. Pathog. **6**:351–359.
20. **Humphreys, G. O., G. A. Willshaw, and E. S. Anderson.** 1975. A simple method for the preparation of large quantities of pure plasmid DNA. Biochim. Biophys. Acta **383**:457–463.
21. **Ikemura, T.** 1981. Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for synonymous codon choice that is optimal for the *E. coli* translational system. J. Mol. Biol. **151**:389–409.
22. **Jacobs, A. A. C., J. Venema, R. Leeven, H. van Pelt-Heerschap, and F. K. de Graaf.** 1987. Inhibition of adhesive activity of K88 fibrillae by peptides derived from the K88 adhesin. J. Bacteriol. **169**:735–741.
23. **Josephsen, J., F. Hansen, F. K. de Graaf, and W. Gaastra.** 1984. The nucleotide sequence of the protein subunit of the K88ac fimbriae of porcine enterotoxigenic *Escherichia coli*. FEMS Microbiol. Lett. **25**:301–306.
24. **Klemm, P.** 1985. Fimbrial adhesins of *Escherichia coli*. Rev. Infect. Dis. **7**:321–340.
25. **Levine, M. M., J. B. Kaper, R. E. Black, and M. L. Clements.** 1983. New knowledge of pathogenesis of bacterial enteric infections as applied to vaccine development. Microbiol. Rev. **47**:510–550.
26. **Lipman, D. J., and W. R. Pearson.** 1985. Rapid and sensitive protein similarity searches. Sciences **227**:1435–1440.
27. **Maniatis, T., E. F. Fristch, and J. Sambrook.** 1982. Molecular cloning: a laboratory manual. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.
28. **Martin, C., C. Boeuf, and F. Bousquet.** *Escherichia coli* CS31A fimbriae: molecular cloning, expression, and homology with the

K88 determinants. Microb. Pathog., in press.
29. **Mooi, F. R., I. Claassen, D. Bakker, H. Kuipers, and F. K. de Graaf.** 1986. Regulation and structure of an Escherichia coli gene coding for an membrane protein involved in export of K88ab fimbrial subunits. Nucleic Acids Res. **14**:2443–2457.
30. **Mooi, F. R., and F. K. de Graaf.** 1985. Molecular biology of fimbriae of enterotoxigenic *Escherichia coli*. Curr. Top. Microbiol. Immunol. **118**:119–138.
31. **Mooi, F. R., N. Harms, D. Bakker, and F. K. de Graaf.** 1987. Organization and expression of genes involved in the production of the K88ab antigen. Infect. Immun. **32**:1155–1163.
32. **Mooi, F. R., C. Wanters, A. Wijfjes, and F. K. de Graaf.** 1982. Construction and characterization of mutants impaired in the biosynthesis of the K88ab antigen. J. Bacteriol. **150**:512–521.
33. **Moseley, S. L., G. Dougan, R. A. Schneider, and H. W. Moon.** 1986. Cloning of chromosomal DNA encoding the F41 adhesin of enterotoxigenic *Escherichia coli* and genetic homology between adhesins F41 and K88. J. Bacteriol. **167**:799–804.
34. **Orskov, I., A. Birch-Andersen, J. P. Duguid, J. Stenderup, and F. Orskov.** 1985. An adhesive protein capsule of *Escherichia coli*. Infect. Immun. **47**:191–200.
35. **Sanger, F., S. Nicklen, and A. R. Coulson.** 1977. DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. USA **74**:5463–5467.
36. **Simons, B. L., P. Rathman, C. R. Malij, B. Oudega, and F. K. de Graaf.** 1990. The penultimate tyrosine residue of the K99 fibrillar subunit is essential for stability of the protein and its interaction with the periplasmic carrier protein. FEMS Microbiol. Lett. **6**:107–112.
37. **Zuker, M., and P. Stiegler.** 1981. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. Nucleic Acids Res. **9**:133–148.