Hydrophobic cluster analysis and secondary structure predictions revealed that major and minor structural subunits of K88-related adhesins of *Escherichia coli* share a common overall fold and differ structurally from other fimbrial subunits

Marie-Claire Méchin, Yolande Bertin, Jean-Pierre Girardeau*

Laboratoire de Microbiologie, INRA, Centre de Recherches de Clermont-Ferrand-Theix, 63122 Saint-Genes-Champanelle, France

Received 6 April 1995

Abstract The structural relatedness of K88-related major and minor subunits was deduced from their sequences by hydrophobic cluster analysis (HCA) and secondary structure predictions produced by the profile neural network prediction program (PHD) on multiple sequence alignments. Although the weak residue identity between major and minor subunits is evidence of a high evolutionary distance, an overall structural similarity was observed. In addition, clear amphipathic conformations were conserved in predicted secondary structure. On the basis of this predicted structural similarity, a schematic 2D model of ClpG subunit was developed.

Key words: Hydrophobic cluster analysis; Protein structure prediction; Structural similarity; Fimbrial subunit

1. Introduction

The K88 and F41 fimbrial adhesins produced by enterotoxigenic *Escherichia coli* are involved in host-specific adhesion and subsequent colonization and infection in piglets and calves respectively [1]. The CS31A antigen was described as a K88-related capsule-like surface protein surrounding the bacterial cells [2] and reported to promote adhesion to human Caco-2 cell line [3]. Analysis of the genetic organization of the K88 gene cluster has revealed that eight structural genes are involved in the biogenesis of K88 fimbriae. FaeC, FaeD and FaeE are minor components involved in the processing, transport and assembly of the composite fimbriae [4]. The K88 fimbriae is composed mainly of the repeating major subunits FaeG and of minor subunits FaeH, FaeI and FaeJ required for the initiation and formation of the K88 fibrillum [5]. As for the major subunit FaeG, the minor subunits FaeH, FaeI and FaeJ required association with the periplasmic chaperone FaeE to prevent aggregation and degradation by protease and to fold and assemble correctly [5]. F41 and CS31A share similar genetic and functional organization [6] and are mainly composed of repeating major subunits F41 and ClpG, respectively [7]. In the CS31A system, ClpH and ClpI are minor subunits closely related (90% identity) and homologous to FaeH and FaeI, respectively (unpublished results).

The significant sequence identity of ClpG with F41 and FaeG (24 and 43%, respectively) and occurence of five conserved hydrophobic amino acid clusters have revealed structural relatedness between the major subunits of CS31A, F41 and K88 surface antigens [7]. K88 minor subunits FaeH, FaeI and FaeJ share a much less significant sequence identity (15%) with the major subunit FaeG [5], but whether or not major and minor subunits support homology in structure was not known. We report here a comparison by hydrophobic cluster analysis (HCA) [8–10] and secondary structure predictions produced by the PHD program [11] of major (FaeG, ClpG, F41) and minor (FaeH, FaeI, FaeJ) subunits of the K88-related antigens. Our study shows structural relatedness between major and minor K88-related subunits and suggests that they differ structurally from type I or P-fimbrial subunits recently studied with the same strategy [12]. A schematic 2D model of the ClpG subunit was developed on the basis of secondary structure predictions.

2. Materials and methods

The protein sequences of major (FaeG, F41, ClpG) and minor (FaeH, FaeI, FaeJ) subunits belonging to the K88 fimbrial antigen family were analysed and compared through hydrophobic cluster analysis (HCA) using HCA-PLOT software from Doriane S.A. (France) running on a Macintosh computer. This method is able to detect similarities in the secondary folding of protein domains even if their sequence identity is low (≤10%). Briefly, HCA is a protein sequence comparison method based on helical representation of the sequences in which characteristics of the hydrophobic clusters, including their size and their shape, are compared, allowing alignment of the sequences. In related proteins, a precise sequence alignment can be deduced from the HCA plots by proceeding outward from the hydrophobic core of similar clusters toward cluster-linking regions where insertions/deletions can occur. For these HCA alignments, a numerical value (HCA score) can be calculated from clusters to assess the alignments [8,9]. HCA strategies and applications have been reviewed elsewhere [8–10].

Alignment and comparison of the sequences were performed with the PileUp algorithm [13] provided by the GCG package from the University of Wisconsin [14]. Deduced amino acid sequences were obtained from the National Biomedical Research Foundation (NBRF) data bases for F41 (S04072) and FaeG (S07202) subunits and from Gen Bank databases (Los Alamos, NM) for ClpG (M55389) and FaeH, FaeI and FaeJ subunits (Z11700).

The secondary structure and solvent accessibility predictions were produced by the profile neural network prediction programs PHDsec [11] and PHDacc [15] respectively, which use information from multiple sequence alignments. The two programs were provided by the software facilities 'PredictProtein' of EMBL Heidelberg.

3. Results and discussion

3.1. Hydrophobic cluster analysis of major and minor subunits

The segmentation centered on hydrophobic clusters was clearly conserved between the major subunits ClpG, FaeG and

*Corresponding author. Fax: (33) 73 62 45 81.

SSDI 0014-5793(95)00417-3

F41 (Fig. 1). Segments S5, S6, S7, S9 and S13, which correspond to the five conserved proline-associated hydrophobic clusters (P1 to P5) previously identified [7], served as major anchors for the alignment. Topological similarities between hydrophobic clusters, which mainly correspond to the core-forming part of regular secondary structures [9,10], and groups of identical (or similar) amino acid residues were recognized. Motifs KWxWxxGxG (segment S2), IPxIx(F/M)xxY(D/E)G (S6) and striking amphipathic consensus included in segments S1, S6, S8, S9 and S13 were present in the three major subunits (Fig. 2). The validity of the proposed alignment can be evaluated by the cluster matching score. The highest HCA score (80%) was shared by ClpG and K88 subunits (Table 1) and provided good evidence of a common folding since an HCA score ≥80% is commonly found in protein with superimposable 3D structures (r.m.s. deviation between polypeptide chain less than

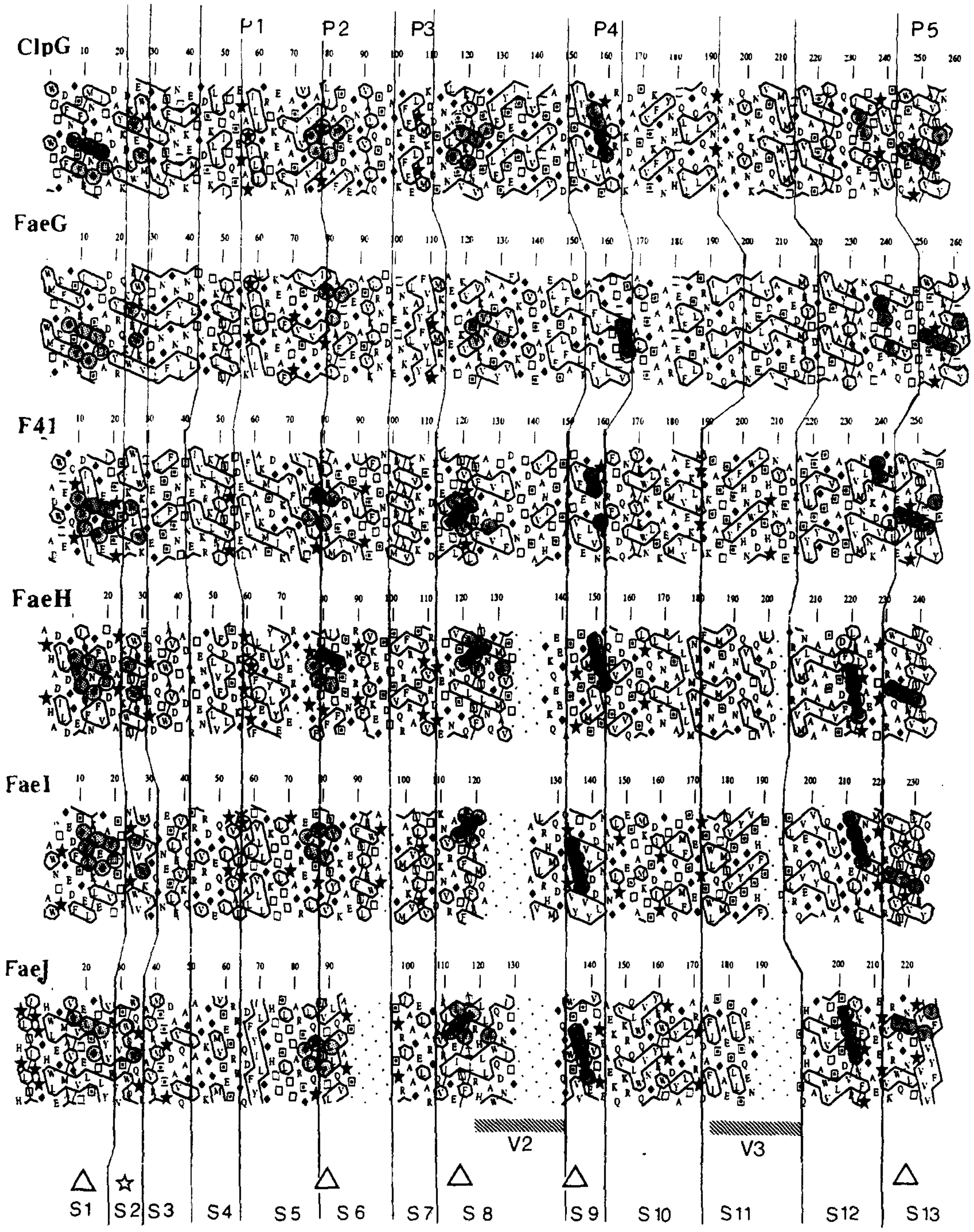

Fig. 1. Alignment of HCA plots of major subunits (FaeG, F41, ClpG) and minor subunits of K88 (FaeH, FaeI, FaeJ). In these plots, the amino acid sequence of the protein is written on a duplicated helical net and the clusters formed by contiguous hydrophobic residues are drawn. The different clusters are delineated by vertical lined which, when joined, show the correspondence between the major segments (S1 to S13). All subunits possess the WxW motif (☆) within the segment S2. Particular conserved features including the five consensus sequences (Fig. 2) are indicated (△) and conserved amino acid are shaded. The one-letter amino acid code is used except for proline (☆), glycine (◇), threonine (□) and serine (⊡).

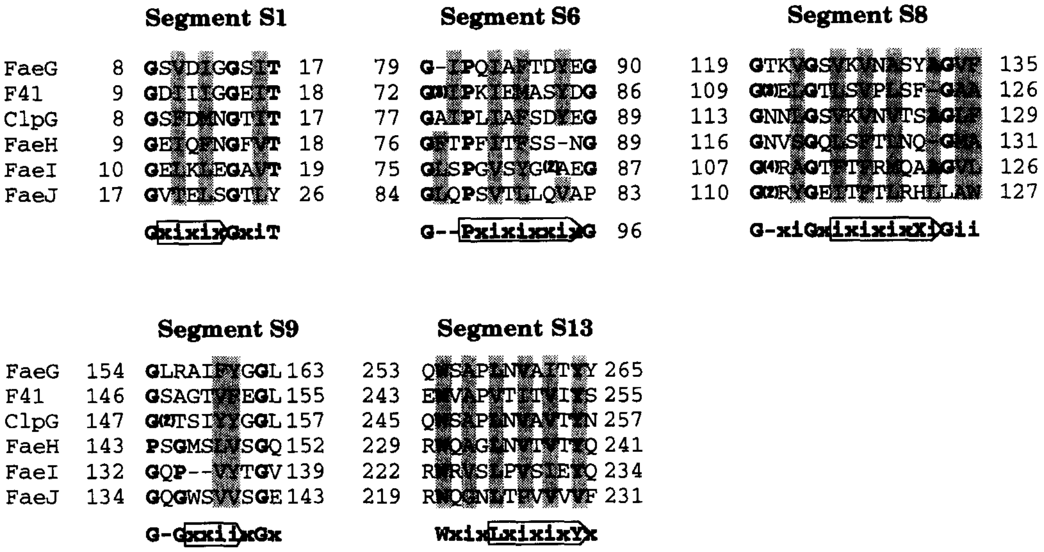
Segment S1
Segment S6
Segment S8
FaeG 8 GSVDIGGSIT 17 79 G-IPQIAFTDYEG 90 119 GTKVGSVKVNASYAGVF 135
F41 9 GDIIIGGEIT 18 72 G(3)IPKIEMASYDG 86 109 G(3)ELGTLSVPLSF-GAA 126
ClpG 8 GSFDMNGTIT 17 77 GAIPLIAFSDYEG 89 113 GNNLGSVKVNVTSAGLF 129
FaeH 9 GEIQFNGFVT 18 76 GFTPFITFSS-NG 89 116 GNVSGQLSFTLNQ-GAA 131
FaeI 10 GELKLEGAVT 19 75 GLSPGVSYG(2)AEG 87 107 G(4)RAGTFITRMQAAGVL 126
FaeJ 17 GVTELSGTLY 26 84 GLQPSVTLLQVAP 83 110 G(2)RYGEITPTLRHLLAW 127
GxixixGxiT G--PxixixxixG 96 G-xiGxixixixXiGii
Segment S9
Segment S13
FaeG 154 GLRAIFYGGL 163 253 QWSAPLNVAITYY 265
F41 146 GSAGTVFEGL 155 243 EWVAPVTIIVIYS 255
ClpG 147 G(2)TSIYYGGL 157 245 QWSAPLNVAVTYN 257
FaeH 143 PSGMSLVSGQ 152 229 RWQAGLNVTVTYQ 241
FaeI 132 GQP--VYTGV 139 222 RWRVSLPVSIEYQ 234
FaeJ 134 GQGWSVVSGE 143 219 RWQGNLTPVVVVF 231
G-GxxiixGx WxixLxixixYx

Fig. 2. Consensus sequences included in the hydrophobic clusters used to anchor the HCA alignment. Conserved residues are shown in bold and hydrophobics are shaded. Open arrows below the sequence indicate the conserved amphiphatic β strands predicted from the PHD program (i = hydrophobic, x = any residue).

1,5 Å) [9,16]. The HCA score of 70% shared by ClpG and F41 subunits correspond to value found in proteins having the same overall fold but with significant structural divergences [16].

Although a weak sequence identity (15%) was observed between the minor subunits FaeH and FaeI, the conserved segmentation and the significant HCA score (71%) indicate that these subunits were structurally related. To improve the HCA alignment between FaeI and FaeJ, two gaps were introduced in segments S6 and S11 of the FaeJ sequence. Given this arrangement, the HCA segmentation between FaeH, FaeI and FaeJ was conserved and the HCA score reached 70% (Table 1). Taking into account these two insertions in the FaeJ sequence, amphipathic consensus included in segments S1, S6, S8, S9 and S13 were found in all minor subunits (Fig. 2). In addition, two conserved motifs (WxW in S2 and PxRW(R/Q)xxL in S13) were recognized within sequences of minor subunits. Thus, the topological similarities between hydrophobic clusters, the significant HCA scores and the presence of consensus in the same segment suggest that the minor subunits FaeH, FaeI and FaeJ have evolved from a common ancestral gene and have probably conserved an overall fold. Similarities extend to ClpH and ClpI in CS31A system (not shown).

Because of the high sequence divergence, classical sequence alignment between major and minor subunits have not been successfully achieved. By contrast, simultaneous comparisons of HCA plots of major and minor subunits revealed a similar distribution of hydrophobic clusters over the entire lenght of the sequence. Moreover, the introduction of one gap (9 residues) in segment S8 of all minor subunits (FaeH, FaeI and FaeJ) improved the alignment between both protein families (Fig. 1). A first assessment of the significance of this alignment was made by calculation of the pairwise HCA homology score

Table 1
Comparison scores between major and minor subunits of the K88 related fimbrial antigens

Sequences	ClpG (257)	F41 (264)	FaeG (254)	FaeH (242)	FaeI (233)	FaeJ (229)
ClpG		70 (24) **24**	80 (46) **48**	68 (21) **24**	68 (16) **20**	59 (14) **17**
F41			70 (22) **22**	70 (18) **20**	71 (20) **20**	61 (13) **14**
FaeG				68 (18) **20**	70 (19) **20**	60 (13) **15**
FaeH					71 (15) **18**	69 (16) **17**
FaeI						70 (17) **21**

For each entry, the sequence identity rate (%) is given in patentheses from PileUp [13] or in bold from HCA base alignment, and the overall HCA score is given above. The HCA score was manually calculated, as described in Törrönen et al. [16]. Random sequences would score around 37 ± 6% [9,10]. Sequence length for each subunit is given in parentheses.

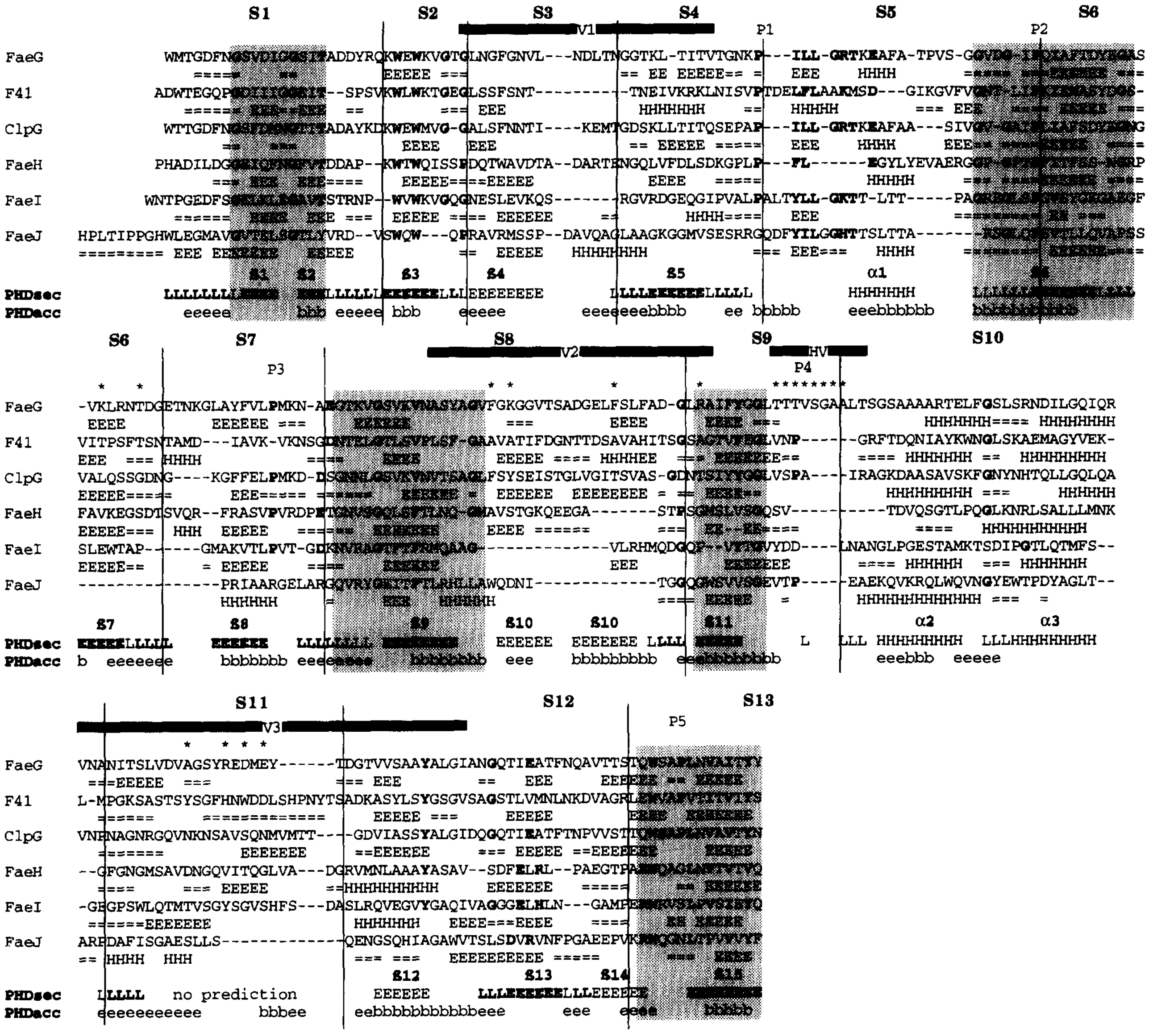

Fig. 3. Alignment of the major and minor subunits of the K88-related fimbrial antigens, based on HCA alignment reported in Fig. 1. Vertical lines delineate segments (S1 to S13) as on HCA plots. Conserved residues are shown in bold and the five consensus sequences shown in Fig. 2 are shaded. The conserved prolines associated with hydrophobic clusters are noted (P1 to P5). The previously identified hypervariable region (HV) on FaeG [19,20] and variable regions (V1, V2, V3) on ClpG [7] are indicated by solid bars. Residues involved in the receptor binding site of FaeG [4,5] are indicated (*). Individual secondary structure predictions produced by PHDsec program on single sequence are reported: β strand (E), α helice (H), loop (=). The position of β strands (β1 to β15), α-helices (α1 to α3) and loops (L) predicted on ClpG with the PHDsec program using the multiple sequence alignments (FaeG, ClpG, F41, FaeH and FaeI), were shown below the alignments. Secondary structure with high prediction potential are in bold. Residues predicted to be exposed (e) or buried (b) are indicated following predictions obtained on multiple sequence alignments with the PHDacc program. Only solvent accessibility predicted with high potential are reported.

(Table 1). When each of the minor subunits was compared with the major subunits, a significant HCA score (70%) was observed for FaeH and FaeI, but a limited score (60%) was obtained for FaeJ. A second assessment of alignment was made on the basis of topological similarities between hydrophobic clusters and occurence of consensus sequences in segments S1, S6, S8, S9 and S13 (Fig. 2). Segment S8, which is characterized by a distinctive 'Zig-Zag' hydrophobic cluster present in all subunits, was one of the major anchors for the alignment. This segment contained the motif iGxixF/VxixxiG (i = hydrophobic residue, x = any residue) which is compatible with a conserved amphipatic β strand [9,10]. Amphipathicity in secondary structural protein regions are critical characteristics for stability, folding and function, and it would expected to find a preferential conservation in ordered amphipathic conformations. The occurrence of the five amphipathic consensus sequences, use as major anchors for the HCA segmentation, are in agreement with this finding and are consistent with the clear conservation of amphipathicity during divergent evolution of homologous proteins recently reported by Pascarella and Argos [17].

3.2. Secondary structure prediction on multiple sequence alignments and schematic 2D model of ClpG subunit

Introducing gaps and arranging sequences according to coin-

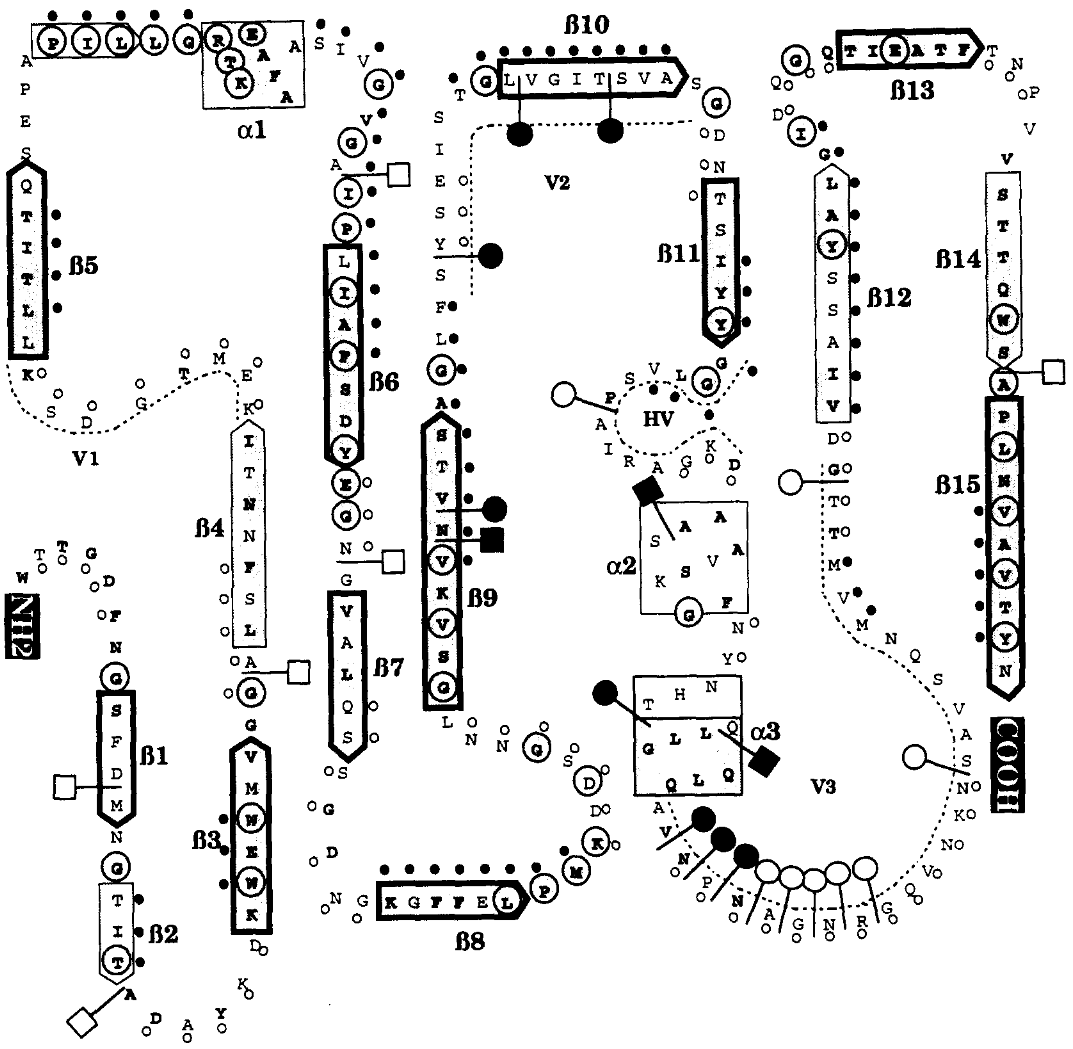

Fig. 4. Schematic diagram of the predicted secondary structure of the ClpG subunit. The amino acid sequence is given as a single letter code. Conserved residues between ClpG and FaeG are shown in bold and conserved regions are shaded. The topologically conserved residues in all subunit sequences are circled. The previously identified hypervariable region on FaeG [HV] and variable regions on ClpG (V1, V2, V3) are indicated (dashed line). Open arrows correspond to the predicted β strands (β to β15) and the predicted α-helices are shown as boxes (α1 to α3) with strong predictions indicated by heavy outline. Residues predicted to be exposed (○) or (●) are reported following predictions of solvent accessibility produce by PHDacc. The position at which insertion did not interfere in biogenesis of K88 (□–) or CS31A (○–) and the position at which insertion greatly interfered with K88 (■–) or CS31A (●–) biogenesis are as previously reported [21–24].

ciding hydrophobic clusters showed that secondary structure predictions of major and minor subunits according to the method of Rost and Sander are markedly alike (Fig. 3). Except for the FaeJ subunit, which is much more divergent, all the subunit sequences could be aligned in a similar secondary structure prediction pattern. Hence, the five consensus sequences resulting from the HCA comparison (Fig. 2) were consistent with the highly predicted β strands, B1, B6, B9, B11 and B15 (Fig. 3). Regions predicted as coil and turn connecting β strands were also remarkably alike in position. Loops, coils and turns were preponderantly predicted in regions which coincide with positions where alignments based on HCA required gaps (<10 residues). This is in agreement with analysis of insertions/deletions (indels) characteristics in protein structures [17,18] indicating that during the evolutionary divergence of related proteins all the elements (loops, coils and turns) that connect secondary structural units are targets for insertions and deletions, which rarely (3%) encroach upon helices and strands. Regions differing markedly in their predicted secondary structures and in which greater gaps are required (≥ 12 residues) coincide with the variable (V2, V3) and hypervariable (HV) regions [7,19] which contain in FaeG, the amino acid residues involved in the receptor binding site [20].

Although the weak sequence identity (<20%) suggest a considerable evolutionary distance between minor and major subunits, the conservation of secondary structural features (amphipathic β strands B3, B6, B9, B11, B15) could reflect common conformational (or functional) constraints in all subunit members of the K88-related family. Thus, similarities between all subunits extend mainly to their N-terminal regions which are thought to be involved in subunit-subunit interactions [20,21] and to their C-terminal regions, which are probably involved in periplasmic chaperonine binding through a conserved amphipatic β-strand [22].

In a recent study [12], we showed that the C-terminal domain of adhesin (minor protein) of type I and P-fimbriae, fold as a pilin subunit (major protein) probably to serve as an assembly module required for presentation of adhesin in the pilus. Compared to type I pilus, K88 fimbriae differ mainly in the functional and structural organization of the composite fiber [5,20,22]. This study shows that K88 fimbriae also differ from type I and P-fimbriae in the folding of the structural subunits. Unlike the major and minor subunits of type I pili system which differ structurally (adhesin appears as a duplicate pilin), the major and minor subunits of K88 system appear with a similar overall fold.

From the secondary structure and solvent accessibility predictions produced by the PHD program using information from the multiple sequence alignments, we developed a schematic 2D model of ClpG subunit (Fig. 4). In our model, the ClpG subunit shows a high percentage of β-sheet content (40%) mixed with α-helix (10%) and random-coil conformations (50%). As shown in figure 4, there is generally good agreement between conserved hydrophobic residues and predicted buried β-strand as well as between predicted surface loops and sequence variability. However, the variable region V2 form exception since residues from this region were predicted to be buried and fold with a β-conformation. This finding was consistent with mutagenesis experiments showing that insertions or deletions in V2, greatly interfered with CS31A biogenesis [23]. The crucial positions of glycine and proline residues topologically conserved in all subunits are highlighted by the 2D representation. In subunit members of the K88-related family, which do not possess any cysteine residue, conserved glycine and proline residues are thought to participate in maintaining the overall structure of the subunit. The 2D model closely matches the results of site-directed mutagenesis, and agrees with the analysis of indels characteristics [18]. Indeed, the positions of the permissive insertions of linker [21,24] or foreign epitopes [23] were mostly located in predicted turns and loops, wheareas insertions that greatly interfere with K88 or CS31A biogenesis intruded in predicted α-helices and β-strands. Further mutagenesis experiments and immunostructural analysis of subunits of the K88-related family will help to assess the accuracy of the predicted 2D model of the ClpG subunit and to elucidate the structure-function relationship of the subunit protein.

Acknowledgements: Authors thank Stéphanie Dutilloy for her helpfull secretarial assistance.

References

[1] Klemm, P. (1985) Rev. Infect. Dis. 7, 321–340.
[2] Girardeau, J.P., Der Vartanian, M., Ollier, J.L. and Contrepois, M. (1988) Infect. Immunol. 56, 2180–2188.
[3] Jallat, C., Darfeuille-Michaud A., Girardeau, J.P., Rich, C. and Joly, B. (1994) Infect. Immunol. 62, 2865–2873.
[4] Bakker, D., Vader, C.E.M., Roosendaal, B., Mooï, F.R., Oudega, B. and De Graaf, F.K. (1991) Mol. Microbiol. 5, 875–886.
[5] Bakker, D., Willemsen, P.T.J., Willems, R.H., Huisman, T.T., Mooï, F.R., Oudega, B., Stegehuis, F. and De Graaf, F.K. (1992) J. Bacteriol. 174, 6350–6358.
[6] Martin, C., Boeuf, C. and Bousquet, F. (1991) Microb. Pathogen. 10, 429–442.
[7] Girardeau, J.P., Bertin, Y., Martin, C., Der Vartanian, M. and Boeuf, C. (1991) J. Bacteriol. 173, 7673–7683.
[8] Gaboriaud, C., Bissery, V., Benchetrit, T. and Mornon, J.P. (1987) FEBS Lett. 224, 149–155.
[9] Lemesle-Varloot, L., Henrissat, B., Gaboriaud, C., Bissery, V., Morgat, A. and Mornon, J.P. (1990) Biochimie 72, 555–574.11.
[10] Woodcock, S., Mornon, J.P. and Henrissat, B. (1992) Protein Eng. 5, 629–635.
[11] Rost, B. and Sander C. (1993) Proc. Natl. Acad. Sci. USA 90, 7558–7562.
[12] Girardeau, J.P. and Bertin, Y. (1995) FEBS Lett. 357, 103–108.
[13] Higgins, D.G. and Sharp, P.M. (1989) CABIOS 5, 151–153.
[14] Devereux, J., Haeberli, P. and Smithies, O (1984) Nucleic Acids Res. 12, 387–395.
[15] Rost, B. and Sander, C. (1994) Proteins 19, 55–72.
[16] Törrönen, A., Kubicek, C.P. and Henrissat, B. (1993) FEBS Lett. 321, 135–139.
[17] Pascarella, S. and Argos, P. (1994) Protein Eng. 7, 185–193.
[18] Pascarella, S. and Argos, P. (1992) J. Mol. Biol. 224, 461–471.
[19] Josephsen, J., Hansen, F., de Graaf, F.K. and Gaastra, W. (1984) FEMS Microbiol. Lett. 25, 301–306.
[20] Bakker, D., Willemsen, P.T.J., Simons, L.H., van Zijderveld, F.G. and De Graaf, F.K (1992) Mol Microbiol. 6, 247–255.
[21] Pedersen, P.A. (1991) Mol. Microbiol. 5, 1073–1080.
[22] Kuehn, M.J., Ogg, D.J., Kihlberg, J., Slonim, L.N., Flemmer, K., Bergfors, T. and Hultgren, S.J. (1993) Science 262, 1234–1241.
[23] Bousquet, F., Martin, C., Girardeau, J.P., Méchin, M.C., Der Vartanian, M., Laude, H. and Contrepois, M. (1994) Infect. Immunol. 62, 2553–2561.
[24] Der Vartanian, M., Méchin, M.C., Jaffeux, B., Bertin, Y., Felix, I. and Gaillard-Martinie, B. (1994) Gene 148, 23–32.