Volume 305, number 2, 91–96 FEBS 11195 June 1992

Fibronectin type III-like sequences and a new domain type in prokaryotic depolymerases with insoluble substrates

Christian K. Hansen

Instituto de Microbiología Bioquímica, Facultad de Biología, CSIC/Universidad de Salamanca, 37071 Salamanca, Spain

Received 13 April 1992; revised version received 11 May 1992

Fibronectin type III-like sequences are present in many proteins from higher eukaryotes and are involved in protein–protein interactions, heparin binding and cell adhesion. A nine-member family of bacterial sequences is shown to be significantly homologous to the type III-like sequences. All the sequences are contained in secreted depolymerases acting on complex, energy-rich insoluble substrates, in which they apparently do not participate in catalysis or substrate-binding, their exact function remaining unclear. Furthermore, a new family of sequences, present in some cellulases, is presented.

Amylase; Cellulase; Cellulose binding-domain; Chitinase; Fibronectin; Poly-3-hydroxybutyrate depolymerase

1. INTRODUCTION

Depolymerization of complex, insoluble, and sometimes crystalline natural polymers, such as cellulose, xylan, chitin and starch, often involves protein domains with catalytic and substrate-binding capabilities [1]. *Trichoderma reesei* [2], *Cellulomonas fimi* [3], *Bacillus lautus* [4,5], and some *Clostridia* [6–8] produce efficient and well-characterized cellulose- and xylan-degrading enzyme systems. While fungal, aerobic- and some anaerobic bacterial systems are believed to be composed of various non-associated polypeptides with catalytic and substrate-binding domains, *Cl. thermocellum* and *Cl. cellulovorans* produce a high molecular weight (2–4 MDa) cellulosome which consists of catalytic polypeptides joined to a large cellulose-binding protein [9,10].

By conventional primary amino acid sequence analysis and hydrophobic cluster analysis [11], nine families of catalytic domains (A–I) [1,12] and five families of cellulose-binding domains (I–V) have been identified (N.R. Gilkes, personal communication; Coutinho et al., paper submitted) which, in the non-associated enzyme systems, are joined by linker sequences [1]. In contrast, catalytic subunits of the cellulosome contain a 24 amino acid repeated sequence which binds to the central protein (CbpA) [13]. CbpA must in turn have cellulose-, subunit- and perhaps cell-binding functions. Shoseyov et al. [14] showed that CbpA of *Cl. cellulovorans* contains a family III cellulose-binding domain, a hydrophobic, probably subunit-binding domain type, and a hydrophilic domain, homologous to regions in endo-β-1,4-glucanase EG-Z of *Cl. stercorarium* [15].

It was recently reported that CenB of *Cell. fimi*, putative protein ORFX of *Cell. flavigena* [16], and chitinase A1 of *B. circulans* WL-12 [17] contain sequences homologous to type III repeats (Fn3's) in fibronectin, a eukaryotic extracellular matrix adhesion molecule [18]. Fn3's in fibronectin are probably involved in protein–protein interactions, heparin-binding and cell adhesion [18].

The presence of Fn3-like sequences and hydrophilic *Cl. cellulovorans* CbpA-like sequences in several hydrolytic enzymes acting on insoluble natural polymers is reported.

2. MATERIALS AND METHODS

The PCGENE software package (Genofit, Switzerland) was used on PC-AT compatible microcomputers for alignment of protein sequences (PALIGN and PCLUSTAL) and construction of phylogenetic trees (PCLUSTAL), while the FASTA software package was used for scanning data banks. PCOMPARE was used to calculate the alignment scores, using the MDM_{78} matrix [19]. HCAPLOT was used on Apple computers to present sequences for hydrophobic cluster analysis [11].

3. RESULTS

3.1. *Fn3's*

Comparison of the reported Fn3's of chitinase A1 of *B. circulans* WL-12 [17], CenB of *Cell. fimi* [16], and putative protein ORFX of *Cell. flavigena* [20] with the SWISS-PROT Protein Sequence Database release 20, using the FASTA program, revealed that an α-amylase–

Abbreviations: Fn3(s), fibronectin type III-like sequence(s).

Correspondence address: C.K. Hansen, c/o Per Linå Jørgensen, Novo Nordisk A/S, Novo Alle 1, DK-2880 Bagsværd, Denmark. Fax: (45) 4449 0070.

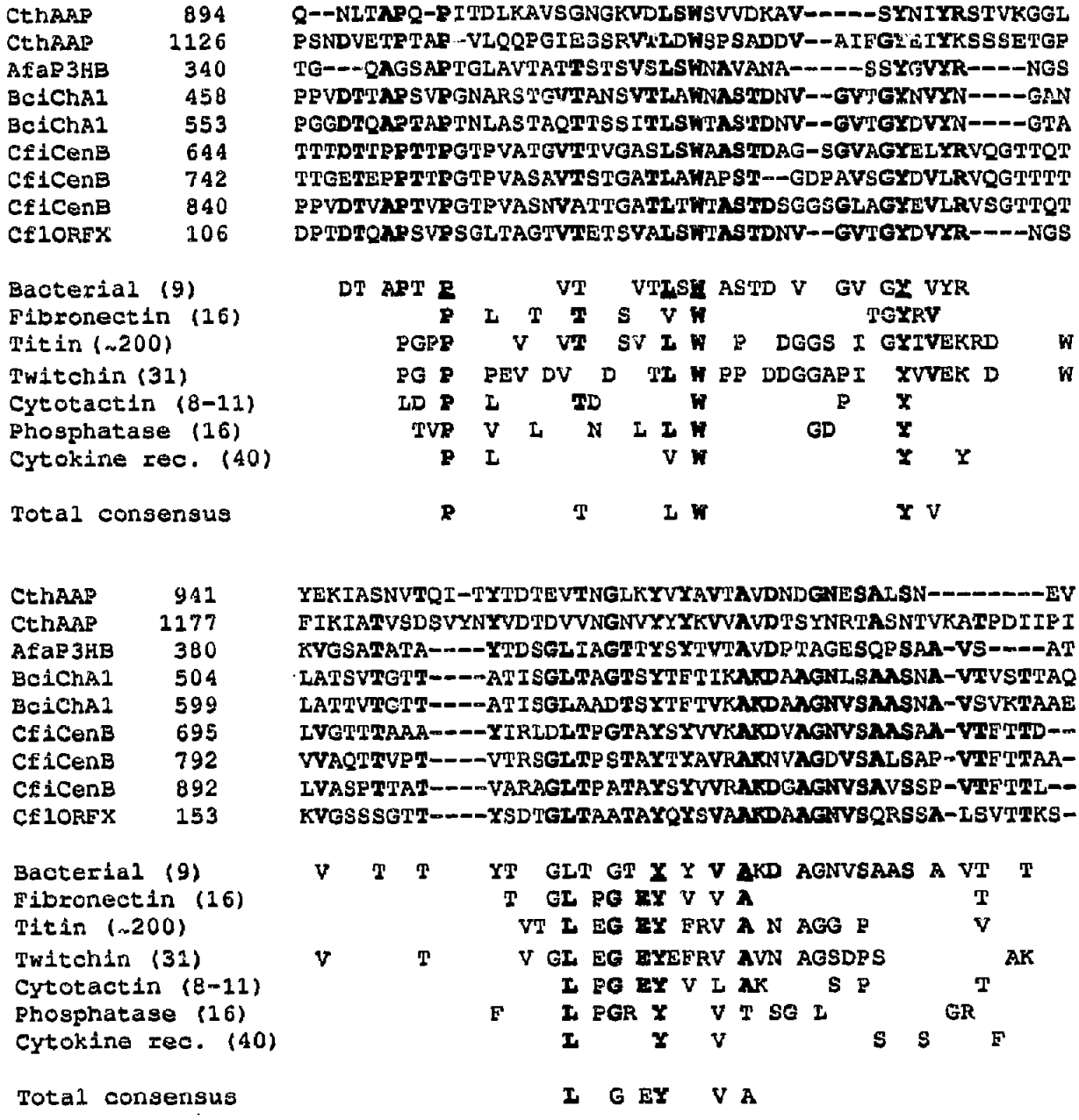

Fig. 1. Alignment of Fn3's. Homologous regions in *Cl. thermohydrosulphuricum* α-amylase–pullulanase (CthAAP) [21], *Al. faecalis* poly-3-hydroxybutyrate depolymerase (AfaP3HB) [22], *B. circulans* chitinase A1 (BciChA1) [17], *Cell. finni* CenB (CfiCenB) [16], and *Cell. flavigena* ORFX (CflORFX) [20] are shown in the upper part of the alignment. A consensus sequence for the bacterial sequences (Bacterial) was constructed with amino acids conserved in at least 5 sequences, with residues conserved in 8–9 sequences shown in bold, and completely conserved residues underlined. Amino acids in the individual sequences, which are identical to the consensus sequence, are also shown in bold. Partial consensus sequences of Fn3's in fibronectin [26], titin [23], twitchin [24], cytotactin [25], a protein-tyrosine phosphatase [27] and several cytokine receptors [28] are shown below. A total consensus sequence (bottom) was constructed by showing residues conserved in at least 4 of the individual consensus sequences, with completely conserved residues in bold. Residues in the individual consensus sequences identical to the total consensus are shown in bold. Numbers in parentheses refer to the number of sequences present in the protein or group of proteins.

pullulanase from *Cl. thermohydrosulphuricum* [21] and a poly-3-hydroxybutyrate depolymerase from *Alcaligenes faecalis* [22] also contain such motifs, repeated twice in the former.

Alignment of the bacterial Fn3's shows conservation of 6 amino acid residues, of which 3 are aromatic (Fig. 1). The latter and a proline are also conserved in Fn3's of the eukaryotic proteins: titin [23], twitchin [24], cytotactin [25], fibronectin [26], a protein-tyrosine phosphatase [27] and several cytokine receptors [28].

Pairwise comparison of the bacterial sequences with PALIGN showed that they have 19.2–71.6% identical,

and 33.7–83.2% identical plus similar amino acids (Table I). The homology of these sequences and Fn3's in fibronectin was assessed with PCOMPARE, which assigns significance to homologous sequence-pairs with SD values larger than 3. Each of the bacterial sequences was compared to the other bacterial Fn3's and to the 16 repeats in fibronectin. SD values of 3.98–22.58 were obtained when the bacterial sequences were compared, clearly demonstrating the significance of their similarity (Table I). In addition, each bacterial sequence scored SD values above 3 when compared to at least 6 of the 16 Fn3's in fibronectin. The phylogenetic tree of Fn3's clearly shows distinct prokaryotic and eukaryotic sequence clusters (Fig. 2).

3.2. *Hydrophilic* Cl. cellulovorans *CbpA-like sequences*

By comparing the hydrophilic repeats of *Cl. cellulovorans* CbpA [14], and domains B and B' of *Cl. stercorarium* EG-Z [15] with the protein sequence databank, endo-β-1,4-glucanase EG-B of *B. lautus* [4] was shown to contain a homologous domain, surrounded by putative linker sequences (SSTASS and SSTTGTTSS) (Fig. 3) [1]. Alignment of the sequences shows that 8 amino acids are generally conserved. Pairwise comparison of the sequences showed 22.1–63.2% identical, and 38.3–83.3% identical plus similar amino acids (Table II). SD values larger than 3 (3.53–17.78) were obtained in all pairwise evaluations.

Program HCAPLOT was used to present region B of EG-Z, hydrophilic repeat 1 of CbpA and the homologous region of EG-B for hydrophobic cluster analysis (Fig. 4). Similar cluster shapes (hydrophobic regions) (S1, S2 and S3) are observed around the conserved asparagine-glycine, aspartic acid-tyrosine and phenylalanine residues, respectively, emphasizing their probable importance for the conserved residue micro-environment.

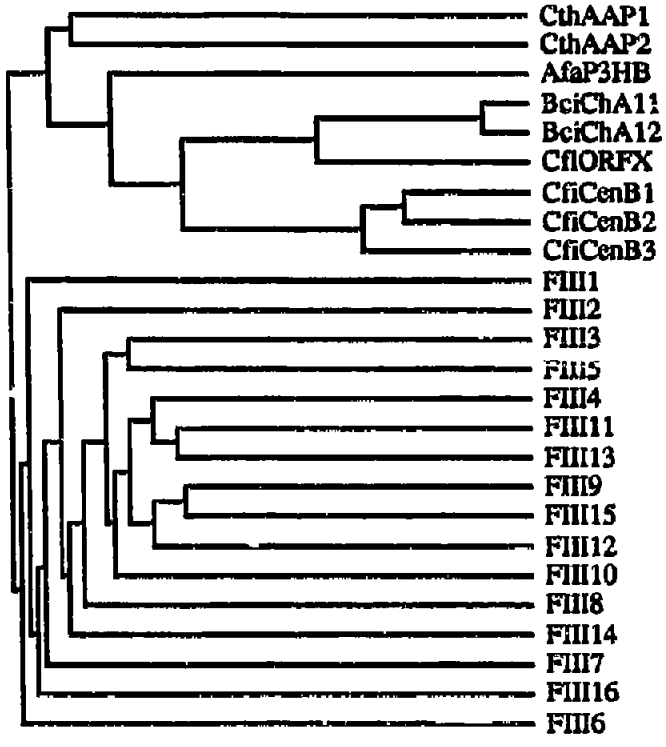

Fig. 2. Phylogenetic tree for the Fn3's of bacterial enzymes and fibronectin. Repeats 1–16 from fibronectin are designated FIII1, FIII2, etc. Bacterial sequences are designated as in Fig. 1, repeats from the same protein carrying a suffix to indicate their position in the protein (e.g. CfiCenB1, CfiCenB2, CfiCenB3, etc.).

4. DISCUSSION

Fn3's were until recently only reported in proteins from higher eukaryotes such as the (i) muscle proteins, i.e. titin, twitchin, smooth muscle myosin light chain kinase, C-protein and 86K protein [23], (ii) cell adhesion proteins, i.e. neural adhesion protein L1 [29], fi-

Table I

Pairwise alignment of bacterial fibronectin type III-like sequences

Domain	CthAAP1	CthAAP2	AfaP3HB	BciChA11	BciChA12	CflORFX	CfiCenB1	CfiCenB2	CfiCenB3
CthAAP1		31.5/46.1	34.1/43.5	25.8/41.5	23.6/39.3	21.3/38.2	24.7/33.7	23.6/34.8	23.6/38.0
CthAAP2	7.17		24.7/40.0	22.1/40.0	20.0/42.2	24.5/41.5	21.4/41.8	23.5/40.8	19.2/37.4
AfaP3HB	9.19	8.01		36.5/54.1	43.5/60.0	50.0/70.6	38.8/50.6	29.4/48.2	28.2/48.2
BciChA11	3.98	7.26	7.11		71.6/83.2	52.1/68.1	46.3/66.3	41.1/57.9	45.3/66.4
BciChA12	4.90	9.70	10.46	22.27		53.2/73.4	45.3/64.2	44.2/61.0	42.1/63.2
CflORFX	5.98	9.58	14.20	20.06	16.03		45.7/60.6	38.3/57.4	48.9/62.7
CfiCenB1	8.06	6.48	8.21	13.26	10.31	12.19		60.2/72.4	65.3/78.6
CfiCenB2	4.93	6.56	6.09	12.04	13.27	13.51	22.58		61.2/76.5
CfiCenB3	5.12	4.54	5.99	15.57	14.03	13.78	21.19	18.17	
Fibronectin reps. 1–16	11/4/1	12/3/1	15/1/0	10/4/2	7/7/2	6/9/1	7/4/5	11/4/1	11/4/1

The upper triangle shows percentage identities/identities + similarities. The lower triangle shows significance scores for comparisons of the sequences. The scores, in SD units, were calculated from 50 random runs with the PCOMPARE program using the MDM_{78} matrix [19]. A bias of 60 and a gap penalty of 60 were used to detect the relationships. The bottom line indicates the number of repeats in fibronectin scoring above/between 2 and 3/below 2 SD units when compared to the respective bacterial sequences.

```
CceCbpA-1   213-   ADVKTTMTLNGNTFKTITDANGTALNASTDYSVSGNDVTISKAY
CceCbpA-2   594-   ADVATTMTLNGYTFNGITGL------TTSDYSISGNVVKISQAY
CceCbpA-3  1535-   ADAAITMTLNGNTFSAIKN-GTATLVKGTDYTVSENVVTISKAY
CceCbpA-4  1628-   ADVVVTMSLNGNTFSAIKN-GTTTLVKGTDYTISGSTVTISKAY
CstEGZ-B    667-   ENIQVVMNLNGNTLNGIKY-GNTYLREGTDYTVSGDTVTILKSF
CstEGZ-B'   760-   RDVKVKLVFNGNTLLAVKK-DGEALVLGRDYSIDGDEVTIFREY
BlaEGB      385-   KDTSVQLNLNGNTLTSLS-VNGTTLKSGTDYTLNSSRLTFKASQ

Consensus          ADV VTM LNGNTF  I     T L  GTDYT S   VTISKAY


CceCbpA-1   257-   LAKQSV------GTTTLNFNFSAGN---PQKLV--ITVVDTPVEA
CceCbpA-2   632-   LAKQPV------GDLTLTFNFSNGNKTATAKLV--VSIKDAP-KT
CceCbpA-3  1578-   LAL-QT------GTVTLEFVFDKGNSAKVVVAVKEIQIVNSTITP
CceCbpA-4  1671-   LATLAD------GSATLEFVFNQGASAKLRLT------IVPAVVDP
CstEGZ-B    710-   LNSFDT------STVQLIFDFSAGRDPVLTVNIIDT--------TT
CstEGZ-B'   803-   LADQPV------GRVTLTFDFDRGTDPVLTINITDSRQVETGVIQ
BlaEGB      428-   LTKLTSLGKLGVNATIVTKFNRGADWKFNVVLYNTPKLSSTTGT

Consensus          LA          G  TL F E  G              V
```

Fig. 3. Alignment of regions homologous to *Cl. cellulovorans* CbpA hydrophilic repeats (CceCbpA1-CceCbpA4) [14]. Homologous regions in *Cl. stercorarium* EG-Z (CstEGZB and CstEGZB') [15] and *B. lautus* EG-B (BlaEGB) [4] are also shown. The consensus sequence was constructed with residues conserved in at least 4 sequences. Amino acids conserved in 6–7 sequences are shown in bold, while completely conserved residues are underlined. In the individual sequences, amino acids identical to the consensus sequence are also shown in bold.

bronectin [18] and cytotactin [25], and (iii) a family of cytokine receptors [28]. In many of these proteins, Fn3's are known to interact with other proteins, heparin and the cell, although the molecular basis of the interactions in most cases is unclear.

The function of Fn3's in the bacterial proteins is unknown, although they would be expected to be non-essential for the catalytic and substrate-binding activities since they are found in enzymes with activity against different substrates, namely cellulose, starch, chitin and poly-3-hydroxybutyrate. For example, CenB-activity on soluble substrate and qualitative cellulose-binding activity has been shown to be independent of the presence of the Fn3's [16]. Likewise, in chitinase A1, elimination of the Fn3 and a region of unknown function (perhaps a chitin-binding domain), did not abolish activity against insoluble chitin, although some difference was observed and attributed to loss of the chitin-binding ability [17]. Similarly, the poly-3-hydroxybutyrate depolymerase was shown to lose substrate-binding capability and activity against insoluble substrates upon treatment with trypsin, an event accompanied by deletion of approximately 60 amino acids, probably from the C-terminus and not including the Fn3 [30] (Fig. 5). The role of Fn3's in bacterial proteins, therefore, remains unclear, although functions such as binding to other components of the hydrolytic system, the substrate or the cell can be imagined.

The role of hydrophilic regions in CbpA, EG-Z and EG-B is also unclear, since none of the regions have been characterized individually. However, an EG-Z peptide composed of domains C', B, B' and C has been shown to bind to cellulose [15], although it remains unclear if this binding is mediated by C-sequences alone

Table II

Pairwise alignments of CbpA hydrophilic-like sequences

Domain	CceCbpA-1	CceCbpA-2	CceCbpA-3	CceCbpA-4	CstEGZ-B	CstEGZ-B'	BlaEGB
CceCbpA-1		63.2/69.1	50.0/59.7	51.2/67.2	47.9/54.8	43.8/52.0	28.8/49.3
CceCbpA-2	10.80		55.9/64.7	51.5/60.3	36.8/47.1	41.2/45.6	22.1/38.3
CceCbpA-3	11.70	13.40		72.2/83.3	48.6/55.5	43.1/55.6	31.9/51.3
CceCbpA-4	15.24	10.42	17.42		48.0/57.3	41.3/50.6	36.0/50.7
CstEGZ-B	15.02	9.91	12.47	17.78		42.7/52.0	29.3/42.6
CstEGZ-B'	11.98	11.84	14.20	12.38	12.21		29.3/44.0
BlaEGB	6.28	3.53	7.34	6.22	6.94	6.65	

The upper right triangle shows percentage identities/identities + similarities. The lower left triangle shows significance scores for the comparisons of the sequences, calculated as in Table I. A bias of 60 and a gap penalty of 50 were used to detect the relationships.

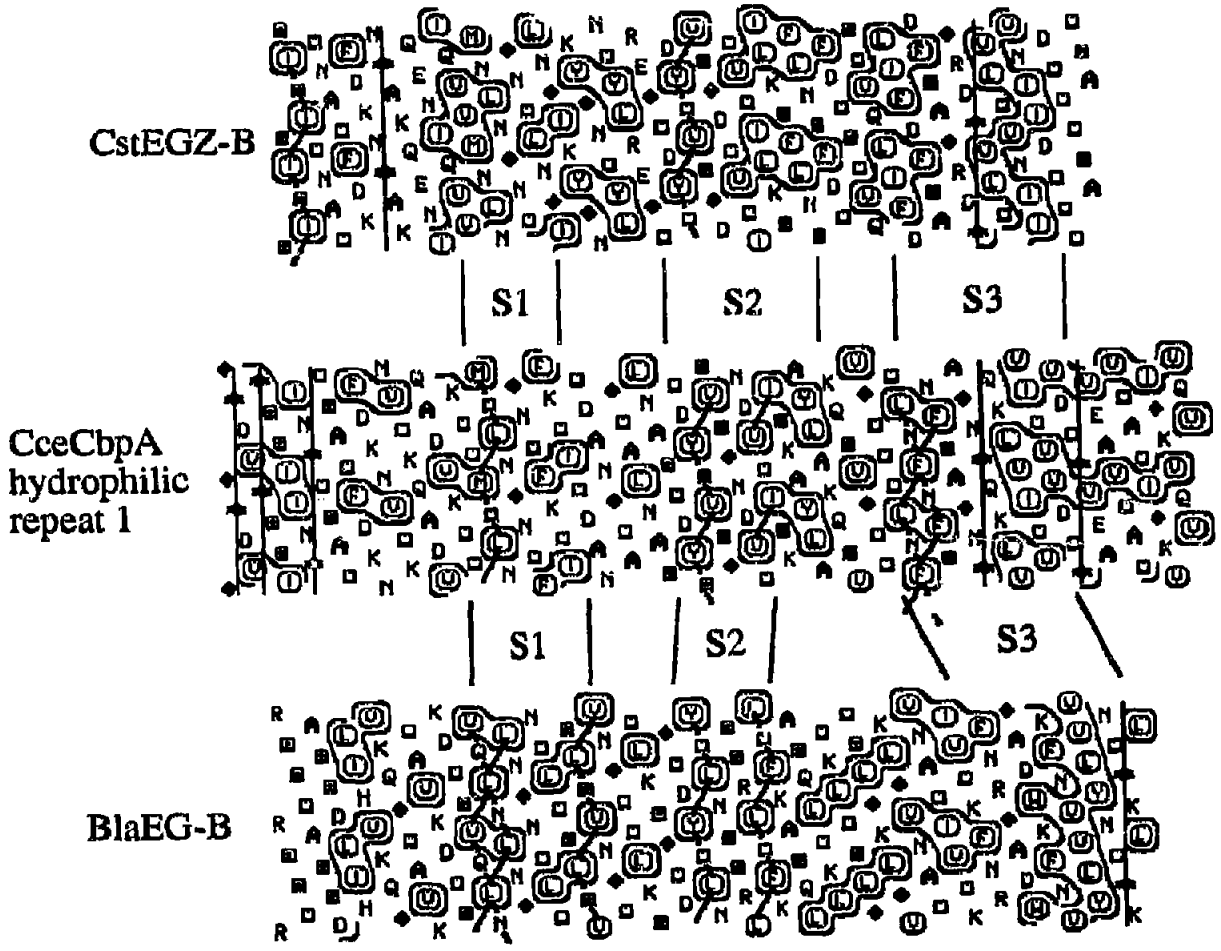

Fig. 4. Alignment of *Cl. cellulovorans* CbpA hydrophilic domain-like sequences by hydrophobic cluster analysis [11]. Sequences were presented with the HCAPLOT program and aligned by eye. Conserved regions S1, S2 and S3 are shown. Abbreviations are as in Fig. 3. Open square, threonine; filled square, serine; star, proline; diamond, glycine.

(cellulose-binding domains) or is enhanced by the B-sequences. CbpA, as a whole, has been proposed to contain the substrate-, subunit-, and perhaps cell-binding functions of the *Cl. cellulovorans* 'cellulosome' [14].

It shall be interesting indeed to determine the roles played by these regions in bacterial enzymes and perhaps contribute in the investigation of Fn3's present in so many eukaryotic proteins.

Acknowledgements: Drs. Neil R. Gilkes and Roi H. Doi are thanked for making data available prior to publication. Dr. Ramon Santamaria, in whose laboratory this work was carried out, Miss Maj-Britt Kaltoft, and Drs. Pierre Béguin, Peter Carlsson, Bernard Henrissat and Per L. Jørgensen are thanked for helpful discussions. The financial support from General Directorate XII of the Commission of European Communities is greatly appreciated.

REFERENCES

[1] Gilkes, N.R., Henrissat, B., Kilburn, D.G., Miller Jr., R.C. and Warren, R.A.J. (1991) Microbiol. Rev. 55, 303–315.
[2] Mandels, M., Sternberg, D. and Andreoti, A.E. (1975) in: Proceedings of Symposium on Enzymatic Hydrolysis of Cellulose (Bailey, M., Enari, T.M. and Linko, M. eds.) pp. 81–109, Aulanko, Finland.
[3] Gilkes, N.R., Langsford, M.L., Kilburn, D.G., Miller Jr., R.C. and Warren, R.A.J. (1984) J. Biol. Chem. 259, 10455–10459.
[4] Jørgensen, P.L. and Hansen, C.K. (1990) Gene 93, 55–60.
[5] Hansen, C.K. (1992) Ph.D. Thesis, The Technical University of Denmark, Copenhagen.
[6] Weimer, P.J. and Zeikus, J.G. (1977) Appl. Environ. Microbiol. 33, 289–297.
[7] Bronnenmeier, K. and Staudenbauer, W.L. (1988) Appl. Microbiol. Biotechnol. 27, 432–436.
[8] Sleat, R., Mah, R.A. and Robinson, R. (1984) Appl. Environ. Microbiol. 48, 88–93.
[9] Lamed, R. and Bayer, E.A. (1988) Adv. Appl. Microbiol. 33, 1–46.
[10] Shoseyov, O. and Doi, R.H. (1990) Proc. Natl. Acad. Sci. USA 87, 2192–2195.
[11] Gaboriaud, C., Bissery, V., Benchetritit, T. and Mornon, J.-P. (1987) FEBS Lett. 224, 149–155.
[12] Henrissat, B., Claeyssens, M., Tomme, P., Lemesle, L. and Mornon, J.-P. (1989) Gene 81, 83–95.
[13] Tokatlidis, K., Salamitou, S., Béguin, P., Djurjati, P. and Aubert, J.-P. (1991) FEBS Lett. 291, 185–188.
[14] Shoseyov, O., Takagi, M., Goldstein, M.A. and Doi, R.H. (1992) Proc. Natl. Acad. Sci. USA (in press).

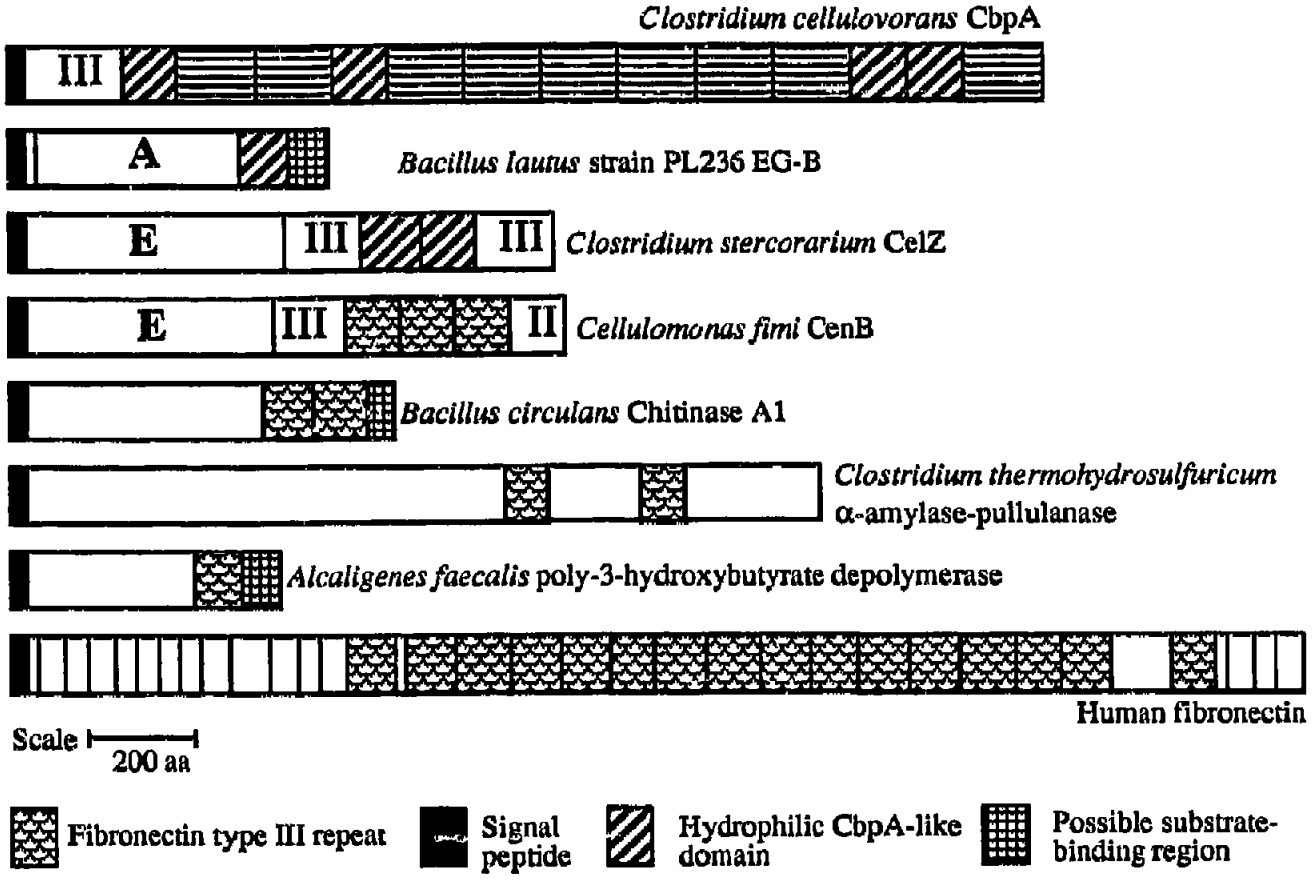

Fig. 5. Domain organization in the bacterial sequences aligned in Figs. 1 and 3, and fibronectin. Letters indicate membership of a β-glycanase catalytic domain family [1], and roman numerals indicate membership of a cellulose-binding domain family (N.R. Gilkes, personal communication; Coutinho, et al., paper submitted). Other codes are explained in the figure.

[15] Jauris, S., Rücknagel, K.P., Schwarz, W.H., Kratzsch, P., Bronnenmeier, K. and Staudenbauer, W.L. (1990) Mol. Gen. Genet. 223, 258–267.
[16] Meinke, A., Gilkes, N.R., Kilburn, D.G., Miller Jr., R.C. and Warren, R.A.J. (1991) J. Bacteriol. 173, 7126–7135.
[17] Watanabe, T., Suzuki, K., Oyanagi, W., Ohnishi, K. and Tanaka, H. (1990) J. Biol. Chem. 265, 15659–15665.
[18] Ruoslahti, E. (1988) Annu. Rev. Biochem. 57, 375–413.
[19] Dayhoff, M.O. (1978) in: Atlas of Protein Sequence and Structure, vol. 5, suppl. 3 (Dayhoff, M.O. ed.) pp. 1–8, National Biomedical Foundation, Washington, DC.
[20] Al-Tawheed, A.R. (1988) M.Sc. Thesis, Trinity College, Dublin.
[21] Melasniemi, H., Paloheimo, M. and Hemiö, L. (1990) J. Gen. Microbiol. 136, 447–454.
[22] Saito, S., Suzuki, K., Yamamoto, J., Fukui, T., Miwa, K., Tomita, K., Nakanishi, S., Odani, S., Suzuki, J.-I. and Ishikawa, K. (1989) J. Bacteriol. 171, 184–189.
[23] Labelt, S., Barlow, D.P., Gautel, M., Gibson, T., Holt, J., Hsieh, C.-L., Francke, U., Leonard, K., Wardale, J., Whiting, A. and Trinick, J. (1990) Nature 345, 273–276.
[24] Benian, G.M., Kiff, J.E., Neckelmann, N., Moerman, D.G. and Waterston, R.H. (1989) Nature 342, 45–50.
[25] Jones, F.S., Hoffmann, S., Cunningham, B.A. and Edelman, G.M. (1989) Proc. Natl. Acad. Sci. USA 86, 1905–1909.
[26] Kornblihtt, A.R., Umezawa, K., Vibe-Pedersen, K. and Baralle, F.E. (1985) EMBO J. 4, 1755–1759.
[27] Krueger, N.X., Streuli, M. and Saito, H. (1990) EMBO J. 9, 3241–3252.
[28] Bazan, J.F. (1990) Proc. Natl. Acad. Sci. USA 87, 6934–6938.
[29] Moos, M., Tacke, R., Scherer, H., Teplow, D., Früh, K. and Schner, M. (1988) Nature 334, 701–703.
[30] Fukui, T., Narikawa, T., Miwa, K., Shirakura, Y., Saito, T. and Tomita, K. (1988) Biochim. Biophys. Acta 952, 164–171.