JOURNAL OF BACTERIOLOGY, Oct. 2000, p. 5906–5910 Vol. 182, No. 20
0021-9193/00/$04.00+0

A Large Gene Cluster for the *Clostridium cellulovorans* Cellulosome

YUTAKA TAMARU,[1] SHUICHI KARITA,[2] ATEF IBRAHIM,[3] HELEN CHAN,[1] AND ROY H. DOI[1]*

Section of Molecular and Cellular Biology, University of California, Davis, California 95616[1]; Center for Molecular Biology and Genetics, Mie University, Tsu, Japan[2]; and Biotechnology and Genetic Engineering Institute, Menoufiya University, Sadat City, Egypt[3]

Received 6 March 2000/Accepted 14 July 2000

A large gene cluster for the *Clostridium cellulovorans* cellulosome has been cloned and sequenced upstream and downstream of the *cbpA* and *exgS* genes (C.-C. Liu and R. H. Doi, Gene 211:39–47, 1998). Gene walking revealed that the *engL* gene cluster (Y. Tamaru and R. H. Doi, J. Bacteriol. 182:244–247, 2000) was located downstream of the *cbpA-exgS* genes. Further DNA sequencing revealed that this cluster contains the genes for the scaffolding protein CbpA, the exoglucanase ExgS, several endoglucanases of family 9, the mannanase ManA, and the hydrophobic protein HbpA containing a surface layer homology domain and a hydrophobic (or cohesin) domain. The sequence of the clustered genes is *cbpA-exgS-engH-engK-hbpA-engL-manA-engM-engN* and is about 22 kb in length. The *engN* gene did not have a complete catalytic domain, indicating that *engN* is a truncated gene. This large gene cluster is flanked at the 5′ end by a putative noncellulosomal operon consisting of *nifV-orf1-sigX-regA* and at the 3′ end by noncellulosomal genes with homology to transposase (*trp*) and malate permease (*mle*). Since gene clusters for the cellulosome are also found in *C. cellulolyticum* and *C. josui*, they seem to be typical of mesophilic clostridia, indicating that the large gene clusters may arise from a common ancestor with some evolutionary modifications.

Clostridium cellulovorans (ATCC 35296) (19), an anaerobic, mesophilic, and spore-forming bacterium, produces extracellular polysaccharolytic multicomponent complexes called the cellulosome (1, 8), which has the ability to degrade cellulose, xylan, mannan, and pectin (19, 21). The *C. cellulovorans* cellulosome (3) consists of three major subunits, CbpA, P100, and P70, and several minor subunits (10, 16). We have previously cloned and sequenced several cellulosomal subunits, i.e., the scaffolding protein CbpA (18), the endoglucanases EngB (4, 17) and EngE (20), and the exoglucanase ExgS (9). More recently, we have completely sequenced the *engL* gene cluster, which consists of five different open reading frames (ORFs) containing a cellulosomal ManA-encoding sequence (21).

In a recent 16S rRNA gene analysis of polysaccharolytic clostridia, *C. cellulovorans* was classified in group I of the phylogenetic tree (13) while most cellulolytic clostridia, such as *C. cellulolyticum*, *C. josui*, *C. papyrosolvens*, and *C. thermocellum*, belonged to the same cluster (group III) (7). Although *C. cellulovorans* was located far from the other cellulolytic clostridia in the phylogenetic tree, the gene clusters of the *C. cellulovorans* cellulosome (22) seem similar to those of *C. cellulolyticum* (2) and *C. josui* (6, 7). Since a large gene cluster in *C. cellulolyticum* (*cipC-celF-celC-celG-celE-ORFX-celH-celJ-celK*) has recently been reported (2), such a gene cluster seems to be specific for mesophilic clostridia and did not occur in the thermophilic bacterium *C. thermocellum*. Furthermore, recent data obtained with *C. cellulovorans*, *C. cellulolyticum*, *C. josui*, and *C. acetobutylicum* revealed that all of these gene clusters begin with the scaffoldin gene, followed by a gene encoding a family 48 cellulase (2). It is of interest to determine the chromosomal organization of the genes of the cellulosome complex, since it may provide information concerning the number of genes, the transcriptional regulation, the coordinate expression, and the evolutionary relationship of the genes in the complex.

In this paper, we describe the large gene cluster around the *cbpA* and *exgS* genes of *C. cellulovorans*. We also analyzed the amino acid sequences of the corresponding proteins and compared them with those of other proteins. Furthermore, this large gene cluster also codes for a small 25-kDa protein, hydrophobic protein A (HbpA), that showed homology with hydrophobic domains (HBDs or type I cohesins) in CbpA (18). The role of HbpA is still not understood, but it may function in a manner similar to that reported for OlpA of *C. thermocellum* (1) and ORFXp of *C. cellulolyticum* (11). The occurrence of this small HbpA may be widespread among mesophilic clostridia that produce cellulosomes.

Cloning and DNA sequencing of the gene cluster. The major gene cluster of the cellulosome consists of nine genes, as shown in Fig. 1. We have cloned and sequenced the *cbpA-exgS* gene cluster (9) and the *engL* gene cluster (pYI-1) harboring five different ORFs, i.e., *engK-hbpA-engL-manA-engM* (21). Since it was expected that the *engL* gene cluster might be located downstream of the *cbpA-exgS* gene cluster, we cloned the region between *exgS* and *engK* by gene walking. As shown in Fig. 1, the internal fragment between *exgS* and *engK* was amplified by PCR with two synthesized oligonucleotides, YT-12 (5′-CT GATATGAACGGTGATGGAAAAG-3′), corresponding to *exgS*, and YT-13 (5′-CCACCAGTTAATGTAGTTGGCA-3′), corresponding to *engK*. As a result, a 4.6-kb PCR fragment (pAI-1) was obtained and cloned into the pCR2.1 vector with a TA cloning kit (Invitrogen) and then sequenced (Fig. 1). The DNA sequence of the pAI-1 fragment contained the *engH* and *engK* genes. No potential transcription terminator was observed between *engH* and *engK*, while a large potential terminator (14) was seen after *engK*. This observation indicated that the *engH* and *engK* genes might be encoded by an operon. Likewise, since no repeat elements were observed between *cbpA* and *exgS* and between *hbpA* and *engL*, they appear to be encoded as operons; large transcriptional terminators were found between *exgS* and *engL*. There is a potential transcriptional terminator downstream of *manA*, indicating that *manA*

* Corresponding author. Mailing address: Section of Molecular and Cellular Biology, University of California, Davis, CA 95616. Phone: (530) 752-3191. Fax: (530) 752-3085. E-mail: rhdoi@ucdavis.edu.

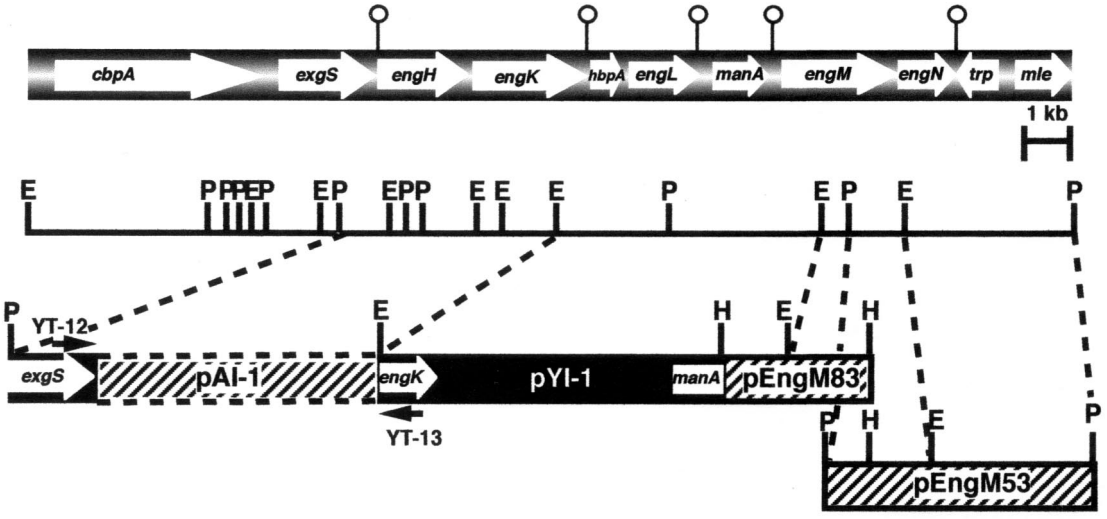

FIG. 1. Restriction enzyme map of a cellulosomal gene cluster. The genes coding for CbpA, ExgS, EngH, EngK, HbpA, EngL, ManA, EngM, and EngN are shown at the top. The pin-like marks indicate palindromes. E, H, and P indicate *Eco*RI, *Hin*dIII, and *Pst*I restriction sites, respectively.

is a monocistronic gene. In fact, ManA production is repressed by cellobiose (21) while the three major cellulosome subunits are expressed in the presence of cellobiose (10). Thus, it will be of extreme interest to study the regulation of expression of these putative operons. One might expect coordinated expression of the operons for the enzymatic subunits with the *cbpA-exgS* operon.

To obtain the complete *engM* gene, Southern hybridization analysis with a partial *engM* fragment of pYI-1 as a probe was carried out. Either *Hin*dIII or *Pst*I digestion of *C. cellulovorans* chromosomal DNA gave a 3.3- or 4.6-kb fragment which was associated with the probe (data not shown). As a result of screening by colony hybridization with the same probe, we cloned two kinds of plasmids that were named pEngM83 (3.3-kb *Hin*dIII fragment) and pEngM53 (4.6-kb *Pst*I fragment), respectively (Fig. 1). The DNA sequence of these fragments contained four ORFs. The first ORF coded for EngM; the second ORF, named *engN*, encoded only the N-terminal amino acid sequence of family 9 cellulases. The last two ORFs coded for proteins that were homologous to transposase (*trn*) and malate permease (*mle*), respectively (Fig. 1), and these two genes flanked the cellulosome gene cluster at the 3′ end. On the other hand, the gene cluster was flanked at the 5′ end by the noncellulosomal gene cluster *nifV-orf1-sigX-regA* (S. Karita and R. H. Doi, unpublished data; 18). There are three cellulosomal genes that are unlinked to the major gene cluster and unlinked to each other, i.e., *engB* (17), *engE* (20), and *engY-pelA* (22).

The *engN* gene is an anomaly, since the coding sequence, which has been checked several times in all three reading frames, indicated that EngN does not have a complete catalytic domain. Repeated sequencing experiments indicate strongly that *engN* is a truncated gene. Furthermore, no duplicated sequence (DS) is present in the coding sequence. The cloned *engN* gene also does not express any endoglucanase activity in *Escherichia coli*, while the other enzymatic genes are expressed in *E. coli* as active enzymes. Since *engN* is flanked by *engM* and the transposase gene (Y. Tamaru and R. H. Doi, unpublished data), there does not appear to have been some accidental deletion during cloning.

Amino acid sequences encoded by the gene cluster. The cellulosomal subunits of *C. cellulovorans* are summarized in Table

TABLE 1. Cellulosomal subunits of *C. cellulovorans*

Gene product	Modular structure[a]	No. of residues[b]	Mol wt[b,c]	Reference or source; GenBank accession no.
EngE	$(SLH)_3$-**GH5**-X-DS	1,030	111,796	20; AF105331
EngK	CBD_{IV}-Ig-**GH9**-DS	892	97,024	This study; AF132735
EngM	CBD_{IV}-Ig-**GH9**-DS	876	96,373	This study; AF132735
ExgS	**GH48**-DS	727	80,485	9; U34793
EngH	**GH9**-CBD_{III}-DS	715	79,321	This study; U34793
EngL	**GH9**-DS	522	57,629	21; AF132735
EngB	**GH5**-DS	441	48,823	5; M37456
ManA	DS-**GH5**	425	47,156	21; AF132735
Cbp	CBD-SLH-$(HBD)_2$-SLH-$(HBD)_6$-$(SLH)_2$-HBD	1,848	189,149	18; M73817
HbpA	SLH-HBD	240	24,930	21; AF132735

[a] Catalytic modules are shown in boldface. Module abbreviations: CBD_{IV}, family IV cellulose-binding domain; **GH9**, family 9 glycosyl hydrolase; Ig, immunoglobulin-like domain; X, unknown domain.
[b] Includes signal sequence.
[c] Molecular weights were determined from the peptide sequences.

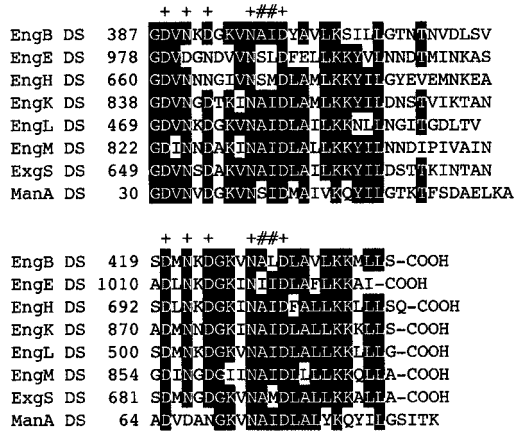

FIG. 2. Alignment of the DSs of cellulosomal subunits of *C. cellulovorans*. Amino acids which are conserved in at least five of the eight sequences are highlighted. Identical amino acid residues are highlighted. Pluses indicate amino acid residues involved in calcium binding. Residues suspected of serving as selectivity determinants are indicated by pound signs.

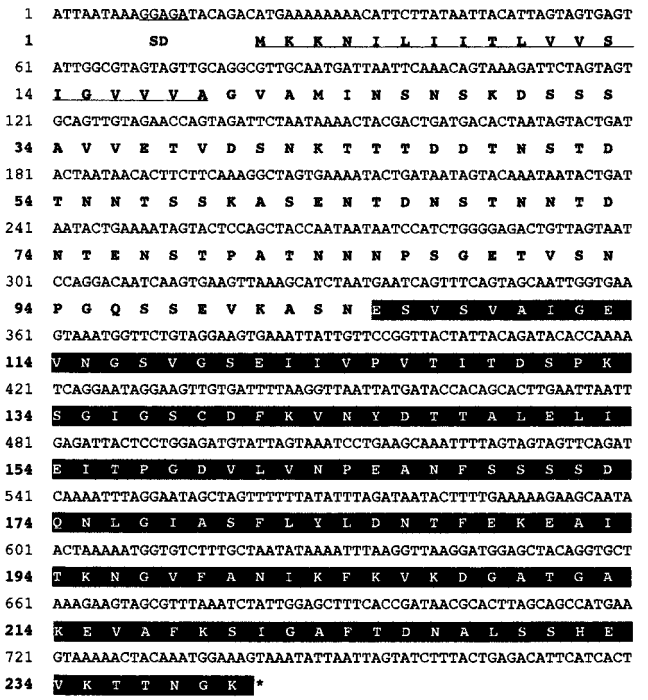

FIG. 3. Nucleotide and deduced amino acid sequences of *hbpA* and HbpA, respectively. The Shine-Dalgarno (SD) and signal peptide sequences are underlined. The stop codon is indicated by an asterisk. The amino acids of the HBD are highlighted.

1. We have previously characterized several cellulosomal subunits, i.e., CbpA (18), EngE (20), ExgS (9), EngB (4, 5), and ManA (21). Four family 9 cellulases, i.e., EngH, EngK, EngL, and EngM, have been found in the gene cluster. EngK and EngM belong to subfamily E1 in family 9, while EngH and EngL belong to subfamily E2 in family 9. Also, except for EngL, family 9 cellulases in the gene cluster contain a cellulose-binding domain (CBD). EngH contains a family IIIc CBD, while EngK and EngM have a family IV CBD.

The presence of DSs (or dockerins), each sequence consisting of about 22 amino acids, is one of the tell-tale signs of a cellulase enzyme belonging to the cellulosome. The cellulosomal gene products are all characterized by the presence of a DS, usually at the C terminus of the protein, although the DS of ManA is located at its N terminus (Fig. 2). Although a DAL or DAI motif is conserved in the DSs from *C. cellulolyticum* and *C. josui* and an NST motif is conserved in those from *C. thermocellum* (7), this motif of *C. cellulovorans* is replaced by NAI. Since the cohesin-dockerin interaction in *Clostridium* species is a species-specific phenomenon (12), the *C. cellulovorans* NAI motif may be essential as a recognition code for binding specificity. Furthermore, the linkage of the DS to the catalytic domain may have a special structure since, almost invariably, when these enzyme subunits are expressed in *E. coli*, a protease in *E. coli* cleaves off the DS and leaves a still-active catalytic domain. This suggests strongly that a protease-accessible structure is present between the catalytic domain and DS domains of *C. cellulovorans* cellulosomal enzymes.

DNA sequence of *hbpA* and domain structure of HbpA. Figure 3 shows the complete nucleotide sequence of the *hbpA* structural gene along with its flanking regions. The *hbpA* gene consists of 720 nucleotides encoding a protein of 240 amino acids with a predicted molecular weight of 24,930. The putative initiation codon (ATG) is preceded by a spacing of 7 bp and by a typical ribosome-binding sequence, AGGAG, which is homologous to the consensus Shine-Dalgarno sequence. Downstream of the TAA translation termination codon, a transcription terminator was not observed, suggesting that *hbpA* and *engL* are in an operon.

The N-terminal amino acid sequence of HbpA exhibits a typical signal peptide and consensus sequence (Val-X-Ala) (23), where the predicted cleavage site is located between positions 19 (Ala) and 20 (Gly) (Fig. 3). The N-terminal region of HbpA (residues 20 to 104) contains a surface layer homology (SLH) domain which shows homology with S-layer proteins from *Mycoplasma hyorhinis* (18.5% identity and 84.5% similarity among 103 amino acids; accession no. P29228) and *Plasmodium reichenowi* (26.5% identity; 91.6% similarity among 83 amino acids; accession no. Z30339) (Fig. 4A). The SLH sequences vary among different surface layer proteins but can be recognized as SLH domains by a few conserved identical amino acids (15).

Also, the N terminus of HbpA has several potential O-glycosylation sites. Since it does not contain a DS, HbpA most likely does not bind to CbpA and is not part of the cellulosome. The C-terminal region (residues 105 to 240) shows 32 to 37% identity with HBDs of CbpA (18) (Fig. 4B), while this region has about the same identity with type I cohesins of other *Clostridium* species (data not shown). Furthermore, the whole HbpA sequence reveals 29.6% identity and 86.2% similarity to *C. cellulolyticum* ORFXp (11) (Fig. 5). The presence of the N-terminal SLH domain suggests that HbpA is a cell surface-bound protein with some function in cellulosome assembly, as postulated previously for a similar protein, ORFXp, from *C. cellulolyticum* (11). It was postulated that the cohesin in ORFXp acts as a temporary binding station for cellulosomal enzymes that are destined for CipA during the assembly of the cellulosome (11). A significant difference between *C. cellulolyticum* ORFXp and *C. cellulovorans* HbpA is the absence of an SLH domain in ORFXp. The presence of the glycosylation sites suggests that HbpA can be glycosylated, while ORFXp is highly glycosylated (11). Thus, the occurrence of this small, hydrophobic protein may be widespread among mesophilic clostridia that produce cellulosomes.

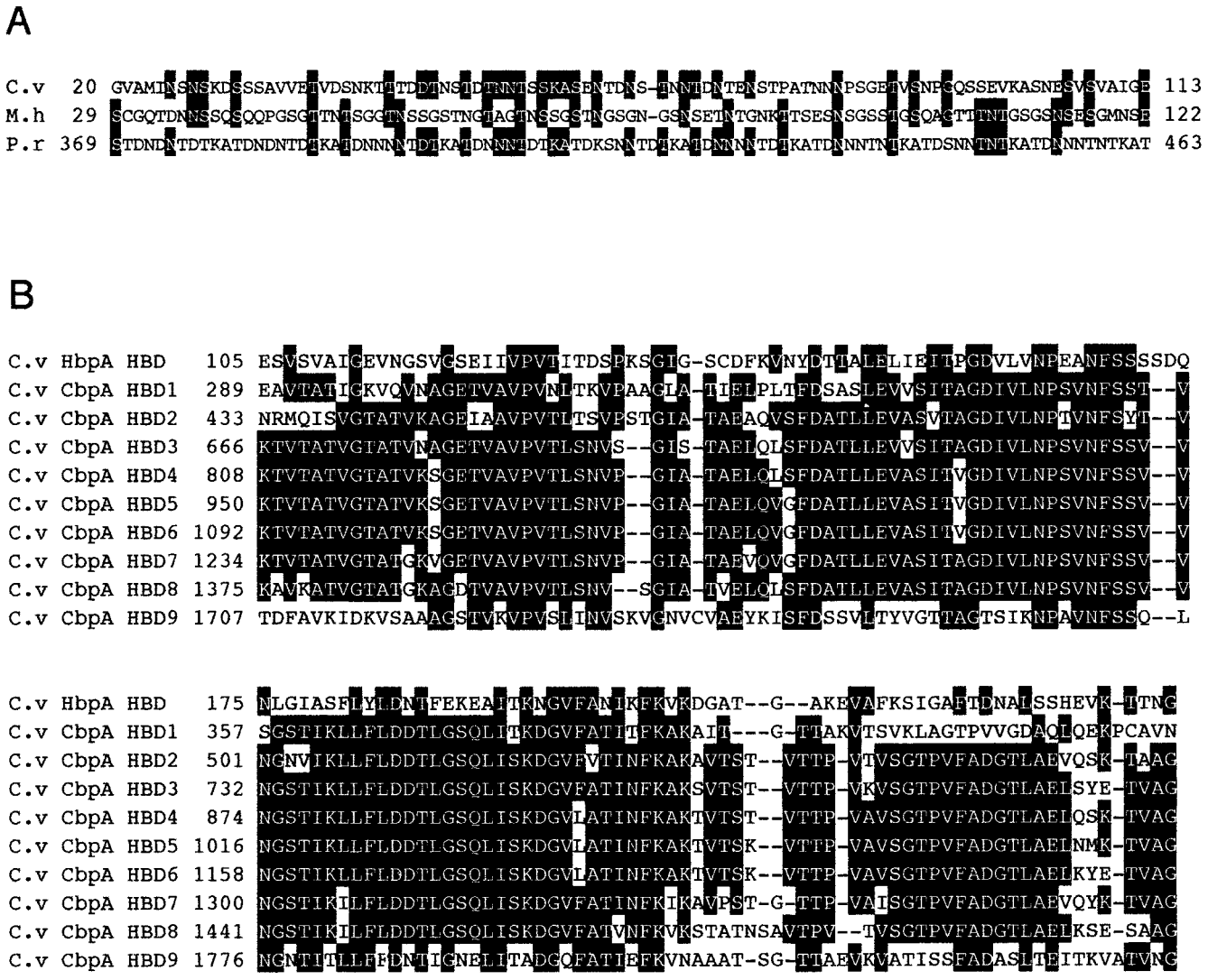

FIG. 4. (A) Alignment of the N-terminal region of HbpA from *C. cellulovorans* (C.v) with the corresponding proteins from *M. hyorhinis* (M.h) and *P. reichenowi* (P.r). (B) Alignment of the C-terminal region of HbpA with HBDs of CbpA from *C. cellulovorans* (C.v). Identical amino acids are highlighted. Gaps left to improve the alignment are indicated by dashes. The numbers refer to amino acid residues at the start of the respective lines; all sequences are numbered from Met-1 of the peptide.

```
C.v HbpA    1 MKKNILIITLVVSIGVVVAGVAMINSNSKDSSSAVVETVDSNKTTTDDTNSTDTNNTSSK
                                                    ....*........*.*
C.c ORFXp   1                   MKKVLLAILIIVCATALVIFGVNHFTSDSETSTSTSTISSTASP

C.v HbpA   61 ASENTDNSTNNTDNTENSTPATNNNPSGETVSNPGQSSEVKASNESVSVAIGEVNGSVGS
              .* ....*.........*.. ......... ...  ... *.........**.*.*..**
C.c ORFXp  45 TSPSASASVSTSSQSKSSDSKSAKTSAAKDSKDTKSNPKDKTPGGEAEISIGKVSGATGS

C.v HbpA  121 EIIVPVTITDSPKSGIGSCDFKVNYDTTALELIEITPGDVLVNPEANFSSSSDQNLGIAS
              ....**.... **.**** .*...***.***..*...*... . ..**. .  .. *..*
C.c ORFXp 105 TVTIPVKLNNLPKKGIGSFNFNIKYDTDALEVVEVKSGEIFGSNNSNFDYTVIDTTGLVS

C.v HbpA  181 FLYLDNTFEKEAITKNGVFANIKFKVKDGA-TGAKEVAFKSIGAFTDNALSSHEVKTTNG
              *** ... .*.*.** **..**.**.**.* .*. ...  . ***.*..*.. .   *.*
C.c ORFXp 165 FLYTSSNSGKDAVTKPGVITNITFKIKDNAKKGSIKISQGTSGAFGDTELKKINPVFTEG
```

FIG. 5. Alignment of *C. cellulovorans* (C.v) HbpA with *C. cellulolyticum* (C.c) ORFXp. The gap left to improve the alignment is indicated by a dash. Identical and similar amino acid residues are indicated by asterisks and dots, respectively. The numbers refer to amino acid residues at the start of the respective lines; all sequences are numbered from Met-1 of the peptide.

Nucleotide sequence accession numbers. The nucleotide sequence data reported in this paper have been submitted to GenBank under accession no. U34793 and AF132735.

This research was supported in part by grant DE-FG03-92ER20069 from the U.S. Department of Energy.

REFERENCES

1. **Bayer, E. A., L. J. W. Shimon, Y. Shoham, and R. Lamed.** 1998. Cellulosomes—structure and ultrastructure. J. Struct. Biol. **124:**221–234.
2. **Belaich, J. P., A. Belaich, H. P. Fierobe, L. Gal, C. Gaudin, S. Pages, C. Reverbel-Leroy, and C. Tardif.** 1999. The cellulolytic system of *Clostridium cellulolyticum*, p. 479–487. *In* K. Ohmiya, K. Hayashi, K. Sakka, Y. Kobayashi, S. Karita, and T. Kimura (ed.), Genetics, biochemistry and ecology of cellulose degradation. Uni Publishers, Tokyo, Japan.
3. **Doi, R. H., J.-S. Park, C.-C. Liu, L. M. Malburg, Y. Tamaru, A. Ichi-Ishi, and A. Ibrahim.** 1998. Cellulosome and noncellulosomal cellulases of *Clostridium cellulovorans*. Extremophiles **2:**53–60.
4. **Foong, F., T. Hamamoto, O. Shoseyov, and R. H. Doi.** 1991. Nucleotide sequence and characteristics of endoglucanase gene *engB* from *Clostridium cellulovorans*. J. Gen. Microbiol. **137:**1729–1736.
5. **Foong, F. C.-F., and R. H. Doi.** 1992. Characterization and comparison of *Clostridium cellulovorans* endoglucanases-xylanases EngB and EngD hyper-

expressed in *Escherichia coli*. J. Bacteriol. **174:**1403–1409.
6. **Fujino, T., S. Karita, and K. Ohmiya.** 1993. Nucleotide sequence of the *celB* gene encoding endo-1,4-β-glucanase-2, ORF1 and ORF2 forming a putative cellulase gene cluster of *Clostridium josui*. J. Ferment. Bioeng. **76:**243–250.
7. **Kakiuchi, M., A. Isui, K. Suzuki, T. Fujino, E. Fujino, T. Kimura, S. Karita, K. Sakka, and K. Ohmiya.** 1998. Cloning and DNA sequencing of the genes encoding *Clostridium josui* scaffolding protein CipA and cellulase CelD and identification of their gene products as major components of the cellulosome. J. Bacteriol. **180:**4303–4308.
8. **Lamed, R., and E. A. Bayer.** 1988. The cellulosome concept: exocellular and extracellular enzyme factor centers for efficient binding and cellulolysis, p. 101–116. *In* J.-P. Aubert, P. Béguin, and J. Millet (ed.), Biochemistry and genetics of cellulose degradation. Academic Press, Inc., San Diego, Calif.
9. **Liu, C.-C., and R. H. Doi.** 1998. Properties of *exgS*, a gene for a major subunit of the *Clostridium cellulovorans* cellulosome. Gene **211:**39–47.
10. **Matano, Y., J.-S. Park, M. A. Goldstein, and R. H. Doi.** 1994. Cellulose promotes extracellular assembly of *Clostridium cellulovorans* cellulosomes. J. Bacteriol. **176:**6952–6956.
11. **Pagès, S., A. Bélaïch, H.-P. Fierobe, C. Tardif, C. Gaudin, and J.-P. Bélaïch.** 1999. Sequence analysis of scaffolding protein CipC and ORFXp, a new cohesin-containing protein in *Clostridium cellulolyticum*: comparison of various cohesin domains and subcellular localization of ORFXp. J. Bacteriol. **181:**1801–1810.
12. **Pagès, S., A. Bélaïch, J.-P. Bélaïch, E. Morag, R. Lamed, Y. Shoham, and E. A. Bayer.** 1997. Species-specificity of the cohesin-dockerin interaction between *Clostridium thermocellum* and *Clostridium cellulolyticum*: prediction of specificity determinants of the dockerin domain. Protein **29:**517–527.
13. **Rainey, F. A., and E. Stackebrandt.** 1993. 16 S rDNA analysis reveals phylogenetic diversity among the polysaccharolytic clostridia. FEMS Microbiol. Lett. **113:**125–128.
14. **Rosenberg, M., and D. Court.** 1979. Regulatory sequences involved in the promotion and termination of RNA transcription. Annu. Rev. Genet. **13:**319–353.
15. **Sára, M., and U. B. Sleytr.** 2000. S-layer proteins. J. Bacteriol. **182:**859–868.
16. **Shoseyov, O., and R. H. Doi.** 1990. Essential 170 kDa subunit for degradation of crystalline cellulose of *Clostridium cellulovorans* cellulase. Proc. Natl. Acad. Sci. USA **87:**2192–2195.
17. **Shoseyov, O., T. Hamamono, F. Foong, and R. H. Doi.** 1990. Cloning of *Clostridium cellulovorans* endo-1,4-β-glucanase genes. Biochem. Biophys. Res. Commun. **169:**667–672.
18. **Shoseyov, O., M. Takagi, M. Goldstein, and R. H. Doi.** 1992. Primary sequence analysis of *Clostridium cellulovorans* cellulose binding protein A (CbpA). Proc. Natl. Acad. Sci. USA **89:**3483–3487.
19. **Sleat, R., R. A. Mah, and R. Robinson.** 1984. Isolation and characterization of an anaerobic, cellulolytic bacterium, *Clostridium cellulovorans* sp. nov. Appl. Environ. Microbiol. **48:**88–93.
20. **Tamaru, Y., and R. H. Doi.** 1999. Three surface layer homology domains at the N terminus of the *Clostridium cellulovorans* major cellulosomal subunit EngE. J. Bacteriol. **181:**3270–3276.
21. **Tamaru, Y., and R. H. Doi.** 2000. The *engL* gene cluster of *Clostridium cellulovorans* contains a gene for cellulosomal ManA. J. Bacteriol. **182:**244–247.
22. **Tamaru, Y., C.-C. Liu, L. Malburg, and R. H. Doi.** 1999. The *Clostridium cellulovorans* cellulosome and non-cellulosomal cellulases, p. 488–494. *In* K. Ohmiya, K. Hayashi, K. Sakka, Y. Kobayashi, S. Karita, and T. Kimura (ed.), Genetics, biochemistry and ecology of cellulose degradation. Uni Publishers, Tokyo, Japan.
23. **von Heijne, G.** 1985. Signal sequences: the limits of variation. J. Mol. Biol. **184:**99–105.