A Large Gene Cluster for the *Clostridium cellulovorans* Cellulosome

YUTAKA TAMARU,† SHUICHI KARITA,‡ ATEF IBRAHIM,¶ HELEN CHAN,† AND ROY H. DOI†*

Section of Molecular and Cellular Biology, University of California, Davis, California 95616;†
Center for Molecular Biology and Genetics, Mie University, Tsu, Japan;‡; and Biotechnology
and Genetic Engineering Institute, Menoufiya University, Sadat City, Egypt¶

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 mannannase 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-engN-hbpA-engM-engL-manA-engK-exgS-exgS 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 (imlA). 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 cellulosome (1) 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. thermocellum*, *C. cellulolyticum*, and *C. josui*, were classified in group II (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). How were the gene clusters prepared? EngX genes of *C. cellulovorans* have been cloned and sequenced upstream of the pAI-1 fragment (9) and the gene cluster (pYI-1) harboring five engX genes might be encoded by an operon.

**Cloning and DNA sequencing of the gene cluster.** The major gene family 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-engM-engN (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 engS and engK by gene walking. As shown in Fig. 1, the internal fragment between engS and engK was amplified by PCR with two synthesized oligonucleotides, YT-12 (5'-CT GATATGACGGTGATGGAAAAG-3') and between engK and engM (5'-CCACCATGTTAGGGTGCA-3'), corresponding to engS and yt-13 (5'-CCACCGTATTGATGAGCGG-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 engS and between hbpA and engL, they appear to be encoded as operons; large transcriptional terminators were found between engK 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.

http://jb.asm.org/ on September 7, 2017 by guest
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 HindIII or PstI 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 HindIII fragment) and pEngM53 (4.6-kb PstI 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 1.

### TABLE 1. Cellulosomal subunits of *C. cellulovorans*

<table>
<thead>
<tr>
<th>Gene product</th>
<th>Modular structure</th>
<th>No. of residues</th>
<th>Mol wt</th>
<th>Reference or source; GenBank accession no.</th>
</tr>
</thead>
<tbody>
<tr>
<td>EngE</td>
<td>(SLH)4-GH5-X-DS</td>
<td>1,030</td>
<td>111,796</td>
<td>20; AF105331</td>
</tr>
<tr>
<td>EngK</td>
<td>CBD1-ig-GH9-DS</td>
<td>892</td>
<td>97,024</td>
<td>This study; AF132735</td>
</tr>
<tr>
<td>EngM</td>
<td>CBD1-ig-GH9-DS</td>
<td>876</td>
<td>96,373</td>
<td>This study; AF132735</td>
</tr>
<tr>
<td>ExgS</td>
<td>GH48-DS</td>
<td>727</td>
<td>80,485</td>
<td>9; U34793</td>
</tr>
<tr>
<td>EngH</td>
<td>GH9-CBD111-DS</td>
<td>715</td>
<td>79,321</td>
<td>This study; U34793</td>
</tr>
<tr>
<td>EngL</td>
<td>GH9-DS</td>
<td>522</td>
<td>57,629</td>
<td>21; AF132735</td>
</tr>
<tr>
<td>EngB</td>
<td>GH5-DS</td>
<td>441</td>
<td>48,823</td>
<td>5; M37456</td>
</tr>
<tr>
<td>ManA</td>
<td>DS-GH5</td>
<td>425</td>
<td>47,156</td>
<td>21; AF132735</td>
</tr>
<tr>
<td>Chp</td>
<td>CBD-SLH-(HBD)2-SLH-(HBD)2-(SLH)2-HBD</td>
<td>1,848</td>
<td>189,149</td>
<td>18; M73817</td>
</tr>
<tr>
<td>HbpA</td>
<td>SLH-HBD</td>
<td>240</td>
<td>24,930</td>
<td>21; AF132735</td>
</tr>
</tbody>
</table>

*a* Catalytic modules are shown in boldface. Module abbreviations: CBDIV, 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.
1. We have previously characterized several cellulosomal subunits, i.e., CbpA (18), EngE (20), ExsS (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 NA1 is replaced by NAI. Since the cohesin-dockerin interaction in Clostridium species is a species-specific phenomenon (12), the C. cellulovorans NA1 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 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 the absence of an N-terminal cellulose-binding 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.
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


