Proc. Natl. Acad. Sci. USA
Vol. 89, pp. 3483–3487, April 1992
Biochemistry

Primary sequence analysis of *Clostridium cellulovorans* cellulose binding protein A

(cellulase/cellulosome/*cbpA* gene/hydrophobic repeats)

ODED SHOSEYOV*, MASAHIRO TAKAGI, MARC A. GOLDSTEIN, AND ROY H. DOI†

Department of Biochemistry and Biophysics, University of California, Davis, CA 95616

Communicated by Paul K. Stumpf, December 24, 1991 (received for review August 7, 1991)

ABSTRACT **The *cbpA* gene for the *Clostridium cellulovorans* cellulose binding protein (CbpA), which is part of the multisubunit cellulase complex, has been cloned and sequenced. When *cbpA* was expressed in *Escherichia coli*, proteins capable of binding to crystalline cellulose and of interacting with anti-CbpA were observed. The *cbpA* gene consists of 5544 base pairs and encodes a protein containing 1848 amino acids with a molecular mass of 189,036 Da. The open reading frame is preceded by a Gram-positive-type ribosome binding site. A signal peptide sequence of 28 amino acids is present at its N terminus. The encoded protein is highly hydrophobic with extremely high levels of threonine and valine residues. There are two types of putative cellulose binding domains of ≈100 amino acids that are slightly hydrophilic and eight conserved, highly hydrophobic β-sheet regions of ≈140 amino acids. These latter hydrophobic regions may be the CbpA domains that interact with the different enzymatic subunits of the cellulase complex.**

The *Clostridium cellulovorans* cellulase is a complex enzyme consisting of several different polypeptides, including a large major cellulose binding protein (CbpA) with no apparent enzyme activity (1). This complex, which has been designated as the cellulosome in *Clostridium thermocellum* (2, 3), is capable of degrading crystalline cellulose, whereas the dissociated enzyme subunits with β-glucanase activity can hydrolyze only soluble substrates. Thus it appears that an association of the enzyme subunits with CbpA is necessary for true cellulase activity.

To elucidate the structure, function, and roles of the various subunits of the cellulosome, we have initiated a study of the enzymatic (4–6) and nonenzymatic (1) subunits of the *C. cellulovorans* cellulase complex. In this study we have cloned the gene for the largest cellulose binding protein in order to determine its base sequence and size and to derive the composition, biochemical properties, and putative functional domains of the protein that it encodes.‡ From homology studies with other cellulases and β-glucanases, we have found a remarkable set of five putative cellulose binding domains. Eight highly conserved hydrophobic sequences were found by a dot plot of CbpA.

The information derived from this study should allow us to define more precisely the functional domains of CbpA and the interactions that occur between CbpA and the enzymatic subunits.

MATERIALS AND METHODS

Library Construction. Chromosomal DNA from *C. cellulovorans* was purified as described (4–6) and digested with *Eco*RI restriction enzyme. This was ligated to λgt11 treated with *Eco*RI and calf intestinal phosphatase. The DNA was then packaged *in vitro* (Promega). *Escherichia coli* yt 1190 cells were transfected and plated on LB agar plates.

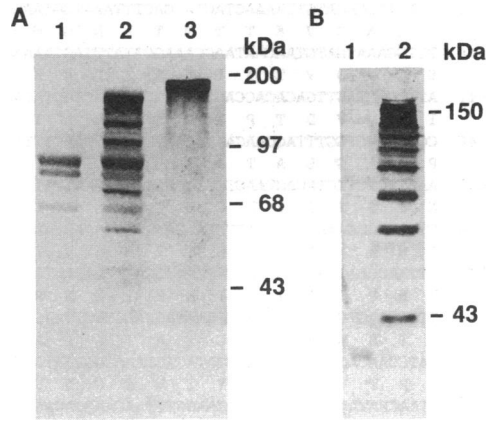

FIG. 1. Western blot analyses. Anti-P170 IgG was used to detect proteins after SDS/PAGE. (*A*) Native and cloned CbpA. Lane 1, extract from cells harboring plasmid pGEMEX-1 without insert as a control; lane 2, extract from cells harboring pCB1; lane 3, cellulase fraction from *C. cellulovorans* purified by the method reported (1). (*B*) Western blot analysis after cellulose binding assay. Lane 1, extract from pGEMEX-1-harboring cells; lane 2, extract from pCB1-harboring cells.

Plaque Screening. Plaques from the library plates were transferred onto Hybond membrane (Amersham). The membranes were washed and immunoscreened using anti-CbpA (anti-P170) IgG (1).

Western Blotting and Cellulose Binding Assay. The *Eco*RI fragment obtained from the λ clone was subcloned into pGEMEX-1 (Promega). Protein expression and detection were performed as described (7). A cellulose binding assay was done by the addition of Avicel to the cleared sonic extracts from cultures of plasmid-harboring *E. coli* JM109 (DE3) cells. Proteins binding to the Avicel were analyzed and purified by the method of Shoseyov and Doi (1).

DNA Sequencing. The 6.5-kilobase (kb) *Eco*RI fragment obtained from the λgt11 library was subcloned into the phagemid pBluescript II KS+/SK+ (Stratagene). Several sets of deletions were made from these subclones using an *Exo* III/mung bean nuclease kit (Stratagene). Single-stranded DNA (ssDNA) templates were prepared from phagemid-bearing cultures infected with the helper phage R408 (Stratagene). These were sequenced by the dideoxy method using Sequenase version 2 (United States Biochemical) following the manufacturer's protocol. Primers based on unique areas

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviation: ORF, open reading frame.

*Present address: Department of Horticulture, Hebrew University of Jerusalem, Rehovot, Israel 76100.
†To whom reprint requests should be addressed.

‡The sequence reported in this paper has been deposited in the GenBank data base (accession no. M73817).

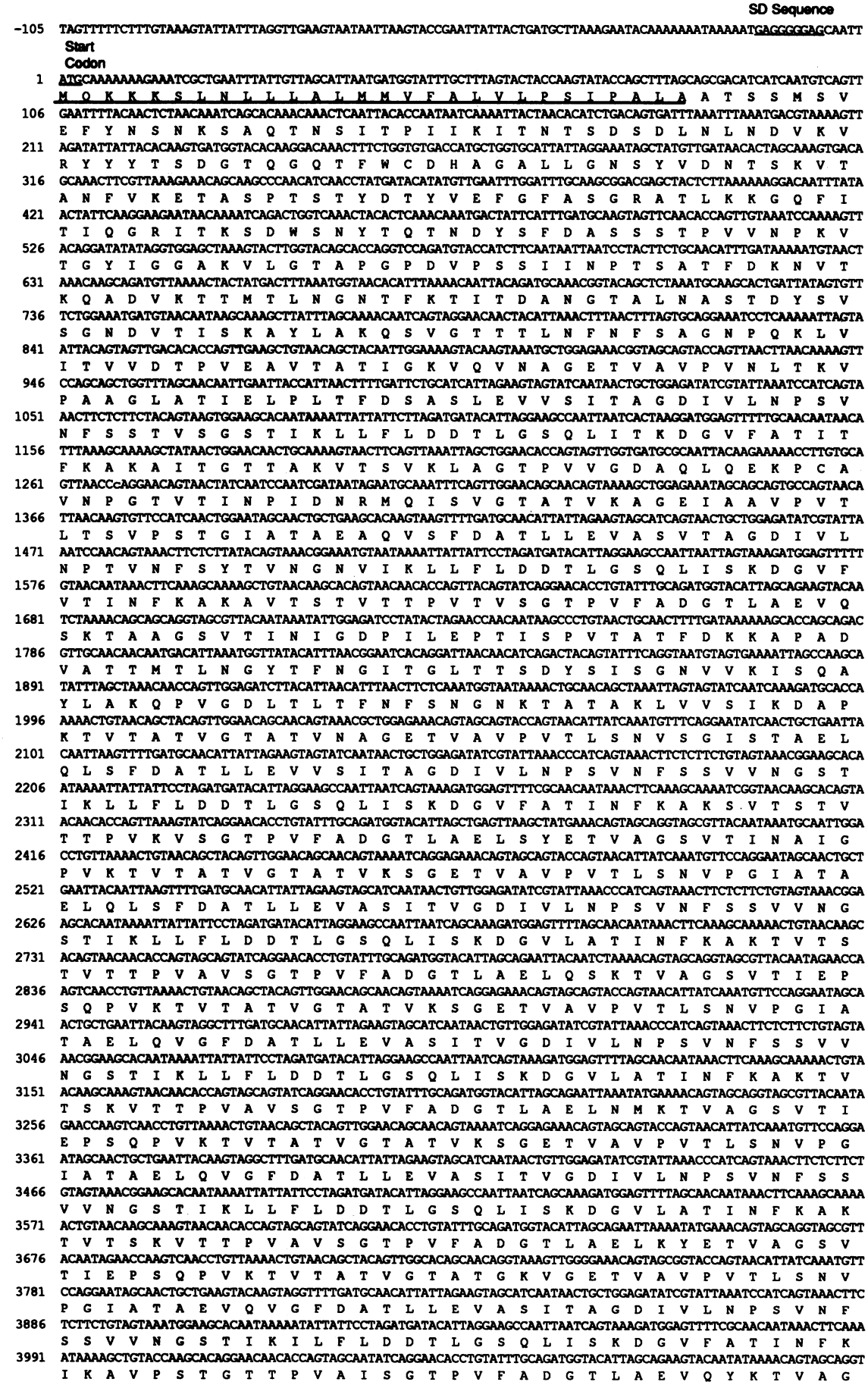

FIG. 2. *(Figure continues on the opposite page.)*

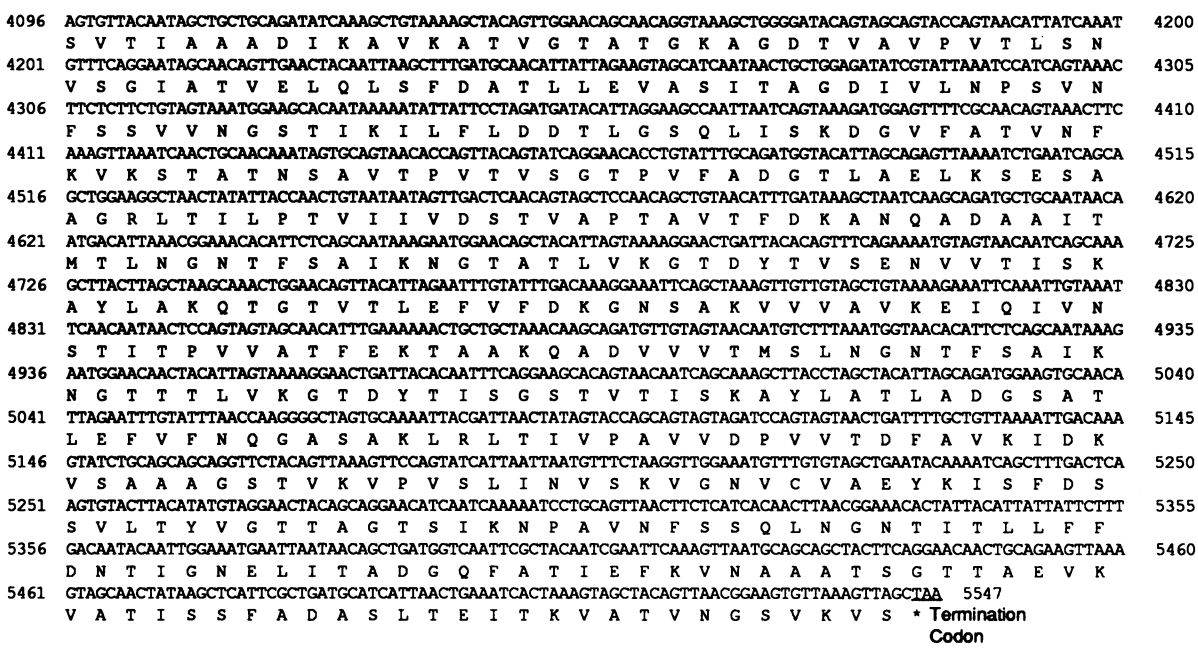

FIG. 2. DNA and deduced amino acid sequence of CbpA. The putative Shine–Dalgarno sequence as well as the start and stop codons are underlined. The predicted signal peptide is also underlined. The processing site was predicted based upon statistical data reported previously (11).

of the sequence were also used. Both strands were sequenced. Inverse polymerase chain reaction (PCR) was used to rescue a fragment of DNA beyond one end of the λgt11 clone (8). Clones from two independent reactions were sequenced by the above methods. Additionally, asymmetric PCR was used to generate ssDNA sequencing templates as an added confirmation of the sequence.

Computer Analysis. The DNA sequence was entered into the DNA STRIDER program (Cedex, France). This program created a map of the restriction sites as well as a plot of potential open reading frames (ORFs). The program also used the deduced amino acid sequence to produce a hydropathy profile by the method of Kyte and Doolittle (9). The sequence data were then transferred to a VMS mainframe computer for analysis by the GCG package (Genetics Computer Group, Madison, WI). Repeats and homologies with other amino acid sequences were detected by dot plot analysis.

RESULTS AND DISCUSSION

Cloning of *cbpA*. In a previous study (1) we characterized a cellulose binding protein with a molecular mass of ≈170 kDa. To isolate the gene for this protein (CbpA), a λgt11 gene bank of *C. cellulovorans* was screened with anti-CbpA antiserum.

Several positive plaques from the immunoscreened plates were selected for analysis. λgt11 inserts of two *Eco*RI fragments, 5.9 kb and 6.5 kb, were obtained. These inserts had similar restriction maps. A Southern blot was performed to determine if either or both of these fragments were present in the *C. cellulovorans* chromosome. A 6.5-kb chromosomal fragment was detected (data not shown). The 6.5-kb insert was chosen for further study because the smaller 5.9-kb fragment may have undergone an internal deletion.

Gene Expression. The 6.5-kb *Eco*RI fragment was cloned into the T7 expression vector pGEMEX-1, and the resulting plasmid was designated pCB1. This recombinant plasmid was transformed into *E. coli* JM109 (DE3), which contained the T7 phage RNA polymerase gene. Extracts of plasmid-harboring cell cultures were subjected to Western blotting analysis and a cellulose binding assay, as shown in Fig. 1. The antibody, anti-P170 IgG, interacted with proteins produced by cells harboring pCB1 (Fig. 1*A*). Some background interactions with control *E. coli* proteins were detected (Fig. 1*A*, lane 1), but none of these proteins exhibited cellulose binding ability (Fig. 1*B*, lane 1). The proteins produced in *E. coli* (pCB1) were able to bind to crystalline cellulose (Fig. 1*B*), but no binding was seen from proteins of the control cells. The largest protein band corresponded to a mass of >150 kDa, and the smallest corresponded to a mass of ≈43 kDa. These smaller proteins were probably the result of partial proteolysis of CbpA in the *E. coli* host. Native CbpA appeared as a single band of ≈170 kDa. The difference between the apparent size of the large bands may be the result of glycosylation of CbpA in *C. cellulovorans*. It has been found that native CbpA contains about 10% (wt/wt) carbohydrate as determined by phenol/sulfuric acid analysis.

DNA Sequencing. The nucleotide sequence of the 6.5-kb *Eco*RI fragment was determined. Two distinctive ORFs were found. The first coded for a protein, RegA, as described (10). The second ORF was preceded by a typical Shine–Dalgarno sequence and ribosome binding site and contained a putative signal peptide 28 amino acids long at the N-terminal region of the ORF (11). However, the 6.5-kb fragment did not contain a termination codon for this ORF. As the expected molecular mass of the coded polypeptide (184 kDa) was very similar to that found on SDS gels (170 kDa), it was predicted that the termination codon would be found nearby in the adjacent sequence of the chromosome. Inverse PCR can be used to walk small distances along genomic DNA (8). Southern blot analysis was done to create a restriction map of the flanking region. *Hin*dIII cleavage of genomic DNA produced a 900-base-pair (bp) fragment that contained ≈500 bases beyond the end of the *Eco*RI segment. The 900-bp fragment was used for inverse PCR rescue of the flanking sequence. A PCR product containing 461 bases of flanking DNA was obtained. Sequencing of this product showed that it contained a continuation of the ORF and a translation termination codon at a location 135 bases downstream from the *Eco*RI site. The complete nucleotide sequence for *cbpA* is given in Fig. 2. The complete ORF was 5544 bp long and encoded a protein, of 1848 amino acids, with a calculated molecular mass of 189

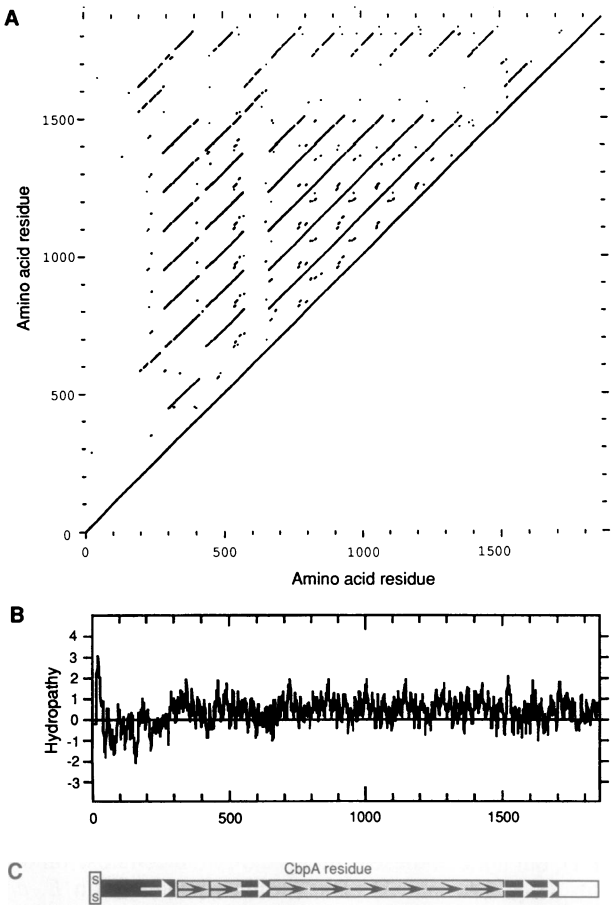

FIG. 3. Primary structure analysis of CbpA. (*A*) Dot matrix plot of the CbpA sequence to detect repeated areas (comparison window = 30, stringency = 15). (*B*) Hydropathy profile by the method of Kyte and Doolittle (9). Higher numbers on the *y* axis represent greater hydrophobicity. (*C*) Schematic diagram of CbpA. The white box with § represents the signal sequence. Black regions indicate areas showing homology to reported cellulose binding proteins, with white arrows indicating hydrophilic repeats. Shaded regions with black arrows represent hydrophobic repeats. A C-terminal region of undetermined function is represented by a white box. The same scale is used in *A–C*.

kDa. These facts and the data shown above strongly indicated that the cloned gene encoded the cellulose binding protein, CbpA. The complete amino acid sequence of CbpA deduced from the nucleotide sequence is shown in Fig. 2. N-terminal sequencing of native CbpA was attempted, in order to confirm that our ORF was the *cbpA* gene, but was unsuccessful, due to the blockage of the N terminus of native CbpA or the extensive glycosylation in the native host mentioned above.

Primary Structure Analysis. The amino acid sequence derived from the ORF was analyzed and it was found that the amino acid composition of CbpA exhibited three cysteine residues (two in the N-terminal region and one near the C terminus) and was extremely threonine and valine rich (14.4 mol% and 13.4 mol%, respectively). Dot plot analysis of the amino acid sequence (Fig. 3*A*) revealed several interesting features. Two distinct sequences were repeated. One of these was almost 100 amino acids in length and was found four times (Fig. 3*C*, white arrows), whereas the other was longer (about 140 amino acids) and was present eight times (Fig. 3*C*, dark arrows). The hydropathy profile of the protein indicated that CbpA was very hydrophobic, as shown in Fig. 3*B*. There was a typical hydrophobic signal sequence at the N terminus of the protein, which was followed by a more hydrophilic region, 300 amino acids long, containing the first 100-amino acid repeat. These shorter repeats occur among the most hydrophilic regions of the protein, whereas the longer 140-amino acid repeats formed the most hydrophobic sections. Each of these longer repeats contained a hydrophobic spike in the center of the sequence. As six 140-amino acid repeats were clustered together in the center of the primary sequence, this formed a core region of the protein of very high average hydrophobicity. A much less homologous, truncated copy of the 140-amino acid repeat was found near the C terminus of the protein.

Homology with Other Sequences. Many endoglucanases can digest soluble forms of cellulose but not the crystalline form. We have shown that a principal step in the breakdown of crystalline cellulose is the attachment of the enzyme to the crystalline substrate (1). We have proposed that CbpA, which does not show any enzymatic properties of its own, can coordinate the digestion of crystalline cellulose by interacting

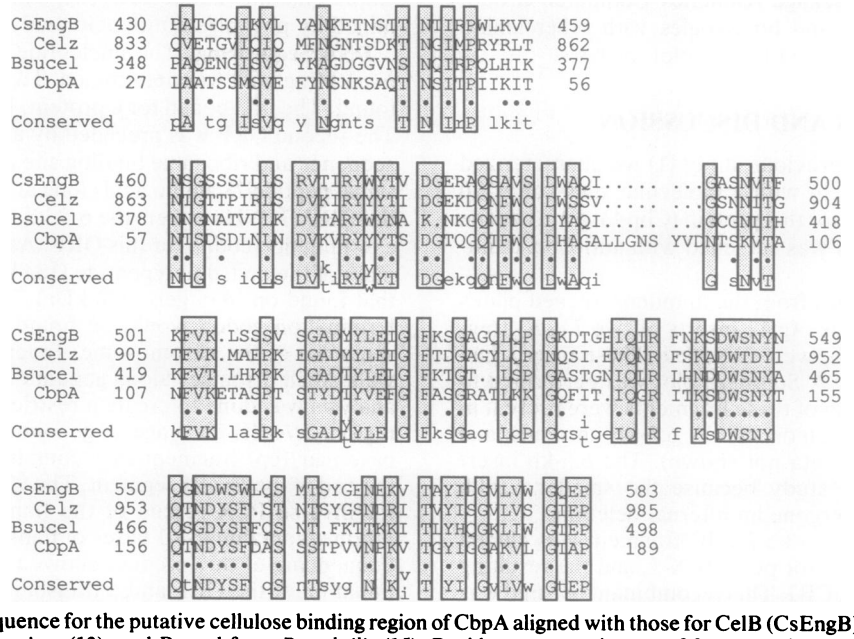

FIG. 4. Amino acid sequence for the putative cellulose binding region of CbpA aligned with those for CelB (CsEngB) from *C. saccharolyticum* (14), CelZ from *C. stercorarium* (13), and Bsucel from *B. subtilis* (15). Residues present in two of four protein sequences are represented by lowercase letters. Regions of higher conservation are enclosed by shaded boxes and are represented by uppercase letters. ●, Positions where CbpA matches the consensus.

with the enzymatic subunits (endoglucanases) and bringing them into proximity with the fibrous substrate. Here we suggest that CbpA has at least two types of domains capable of carrying out these functions: (*i*) a cellulose binding domain and (*ii*) a region of interaction with endoglucanases. Homology searches were done to determine whether CbpA exhibited domains homologous to those reported for various bacterial and fungal cellulases. Although no significant homology was found between CbpA and published fungal and bacterial cellulose binding domain consensus sequences as reviewed by Gilkes *et al.* (12), strong homologies were detected between CbpA and nonenzymatic domains of Avicelase I (CelZ) from *Clostridium stercorarium* (13), CelB from *Caldocellum saccharolyticum* (14), and Bsucel from *Bacillus subtilis* (15). These enzymes contained a sequence related to the first 160 amino acids at the N terminus of mature CbpA (Fig. 4). This region exhibited approximately 40% identity and 55% similarity when compared to the three cellulases mentioned above. Further, the hydrophilic repeat present four times in CbpA (Fig. 3*C*, white arrows) was homologous to a repeated (two occurrences) portion of CelZ. These regions of CelZ have been shown to bind to crystalline cellulose, such as Avicel (13). These homologies to reported cellulose binding domains indicate that CbpA contains one or more regions capable of interacting with crystalline cellulose. This is consistent with the result that even small (≈43 kDa) proteolysis products exhibited binding to Avicel (Fig. 1*B*).

No other significant homology between CbpA and other proteins has been detected. However, repeats of as yet unknown function have been found in endoglucanases associated with the cellulosome (ref. 16; F. Foong and R.H.D., unpublished data). It has been reported that these repeats are not required for the catalytic activity of the enzymes (16). It is possible that these regions may interact with the hydrophobic repeats of CbpA. Secondary structural predictions, by the method of Chou and Fasman (17), indicate that there is one major turn in one of the hydrophilic regions, in the range of residues 600–650. The rest of the protein was predominantly β sheets (data not shown). This type of structure is consistent with the high threonine and valine content in the hydrophobic repeats of CbpA. The presence of three cysteine residues, two near the N terminus and one near the C end, indicates the possibility of disulfide bonds curling the molecule, bringing the N and C termini together at the tertiary structural level. This would have the effect of clustering the putative cellulose binding domains together in three-dimensional space. Further work is necessary to confirm that our putative cellulose binding domains do interact as expected with the cellulose substrate and to provide a functional analysis of the hydrophobic repeats.

We thank Janet F. Aoyama for her excellent technical assistance. This research was supported in part by Department of Energy Grant DE-FG03-87ER13705 (to R.H.D.). O.S. was supported by Fellowship SI-0057-87 from the U.S.-Israel Binational Agricultural Research & Development Fund (BARD). M.A.G. was supported partially by a Jastro–Shields Research Scholarship.

1. Shoseyov, O. & Doi, R. H. (1990) *Proc. Natl. Acad. Sci. USA* **87**, 2192–2195.
2. Lamed, R., Setter, E. & Bayer, A. E. (1983) *J. Bacteriol.* **156**, 827–836.
3. Mayer, F., Coughlan, M. P., Mori, Y. & Ljungdahl, L. G. (1987) *Appl. Environ. Microbiol.* **53**, 2785–2792.
4. Shoseyov, O., Hamamoto, T., Foong, F. & Doi, R. H. (1990) *Biochem. Biophys. Res. Commun.* **169**, 667–672.
5. Hamamoto, T., Shoseyov, O., Foong, F. & Doi, R. H. (1990) *FEMS Microbiol. Lett.* **72**, 285–288.
6. Foong, F., Hamamoto, T., Shoseyov, O. & Doi, R. H. (1991) *J. Gen. Microbiol.* **137**, 1729–1736.
7. Chang, B. Y. & Doi, R. H. (1990) *J. Bacteriol.* **172**, 3257–3263.
8. Innis, M. A., Gelfand, D. H., Sninsky, J. J. & White, T. J. (1990) in *PCR Protocols* (Academic, San Diego), pp. 219–227.
9. Kyte, J. & Doolittle, R. F. (1984) *J. Mol. Biol.* **157**, 105–132.
10. Shoseyov, O., Goldstein, M. A., Foong, F., Hamamoto, T. & Doi, R. H. (1991) *Nucleic Acids Res.* **19**, 1710.
11. Watson, M. E. E. (1984) *Nucleic Acids Res.* **12**, 5145–5164.
12. Gilkes, N. R., Henrissat, B., Kilburn, D. G., Miller, R. C., Jr., & Warren, R. A. J. (1991) *Microbiol. Rev.* **55**, 303–315.
13. Jauris, S., Rücknagel, K. P., Schwartz, W. H., Kratzsch, P., Bronnenmeier, K. & Staudenbauer, W. L. (1990) *Mol. Gen. Genet.* **223**, 258–267.
14. Saul, D. J., Williams, L. C., Love, D. R., Chamley, L. W. & Bergquist, P. L. (1989) *Nucleic Acids Res.* **17**, 439.
15. MacKay, R. M., Lo, A., Willick, G., Zucker, M., Daird, S., Dove, M., Moranelli, F. & Seligy, V. (1986) *Nucleic Acids Res.* **14**, 9159–9170.
16. Hall, J., Hazelwood, G. P., Barker, P. J. & Gilbert, H. J. (1988) *Gene* **69**, 29–38.
17. Chou, P. Y. & Fasman, G. D. (1978) *Annu. Rev. Biochem.* **47**, 251–276.