PROTEINS: Structure, Function, and Genetics 29:517–527 (1997)

Species-Specificity of the Cohesin-Dockerin Interaction Between *Clostridium thermocellum* and *Clostridium cellulolyticum:* Prediction of Specificity Determinants of the Dockerin Domain

Sandrine Pagès,[1] Anne Bélaïch,[1] Jean-Pierre Bélaïch,[1,2] Ely Morag,[3] Raphael Lamed,[4] Yuval Shoham,[5] and Edward A. Bayer[3]*

[1]*Bioénergétique et Ingéniérie des Protéines, Centre National de la Recherche Scientifique, IBSM-IFR1, Marseille, France*
[2]*Université de Provence, Marseille, France*
[3]*Department of Membrane Research and Biophysics, The Weizmann Institute of Science, Rehovot, Israel*
[4]*Department of Molecular Microbiology and Biotechnology, Tel Aviv University, Ramat Aviv, Israel*
[5]*Department of Food Engineering and Biotechnology, Technion–Israel Institute of Technology, Haifa, Israel*

ABSTRACT **The cross-species specificity of the cohesin–dockerin interaction, which defines the incorporation of the enzymatic subunits into the cellulosome complex, has been investigated. Cohesin-containing segments from the cellulosomes of two different species, *Clostridium thermocellum* and *Clostridium cellulolyticum,* were allowed to interact with cellulosomal (dockerin-containing) enzymes from each species. In both cases, the cohesin domain of one bacterium interacted with enzymes from its own cellulosome in a calcium-dependent manner, but the same cohesin failed to recognize enzymes from the other species. Thus, in the case of these two bacteria, the cohesin–dockerin interaction seems to be species-specific. Based on intra- and cross-species sequence comparisons among the different dockerins together with their known specificities, we tender a prediction as to the amino-acid residues critical to recognition of the cohesins. The suspected residues were narrowed down to only four, which comprise a repeated pair located within the calcium-binding motif of two duplicated sequences, characteristic of the dockerin domain. According to the proposed model, these four residues do not participate in the binding of calcium per se; instead, they appear to serve as recognition codes in promoting interaction with the cohesin surface. Proteins 29:517–527, 1997.**

Key words: cellulosome; cellulases; cohesin domain; scaffoldin subunit; EF-hand motif; molecular modeling

INTRODUCTION

Many anaerobic cellulolytic microorganisms produce high-molecular-weight, multienzyme complexes, called cellulosomes, which are highly active on crystalline cellulosic substrates.[1–6] The first cellulosome to be identified was that of *Clostridium thermocellum*, which contained about 14 different subunits and exhibited a molecular mass of about 2 MDa.[7] Cellulosomes are also produced by other cellulolytic clostridia, such as *C. cellulovorans*,[5] *C. cellulolyticum*,[8] and *C. papyrosolvens*.[9] In addition, cellulosome-like complexes are apparently found in a variety of anaerobic bacteria other than *Clostridium* spp., notably *Bacteroides cellulosolvens*[10] and *Acetivibrio cellulolyticus*[11] as well as rumen fungi such as *Neocallimastix frontalis* and *Piromyces* spp.[12]

The definitive characteristic of the clostridial cellulosomes is the presence of a large subunit (called the scaffolding protein or scaffoldin), which appears to be devoid of enzymatic activity and to which the catalytic subunits are attached.[2] The genes encoding

Abbreviations: CipA, the scaffoldin subunit from *C. thermocellum* (cellulase-integrating protein A); CbpA, the scaffoldin subunit from *C. cellulovorans* (cellulose-binding protein A); CipC, the scaffoldin subunit from *C. cellulolyticum;* CBD, cellulose-binding domain; $CelA_{2-DOC}$, recombinant, intact, cellulosomal endoglucanase CelA from *C. cellulolyticum,* which contains the CelA catalytic domain linked to the CelA dockerin domain; $CelA_3$, the same as $CelA_{2-DOC}$, but lacking the dockerin domain; CBD–Coh3-ct, recombinant probe from *C. thermocellum,* which contains the CBD and the cohesin-3 domain of the CipA scaffoldin subunit; CBD–HD1–Coh1-cc, recombinant probe from *C. cellulolyticum,* which contains the CBD, the first hydrophilic domain and the cohesin-1 domain of the CipC scaffoldin subunit

Contract grant sponsor: Centre National de la recherche scientifique; Contract grant sponsor: Universite of Provence; Contract grant sponsor: Région Provence-Alpes-Cotê d'Azur; Contract grant sponsor: Israel Science Foundation (Israel Academy of Sciences and Humanities, Jeruselum); Contract grant sponsor: European Commission (Biotechnology Programme); Contract grant number: PL962303
*Correspondence to: Dr. Edward A. Bayer, Department of Membrane Research and Biophysics, The Weizmann Institute of Science, Rehovot 76100, Israel.
E-mail: bfbayer@weizmann.weizmann.ac.il
Received 10 April 1997; Accepted 17 June 1997

three scaffolding proteins from three different cellulolytic bacterial species have either been totally or partially sequenced. These include CipA,[13] CbpA,[14] and CipC.[15,16] In all three cases, the deduced amino-acid sequence showed multiple copies of a hydrophobic domain of about 140 residues, which is now called the cohesin domain. Amino-acid sequence comparison of the various cohesins, both within a given species and among the three species, shows a strikingly high degree of similarity.

The cohesin domain interacts with a different type of noncatalytic domain, called the dockerin domain, a single copy of which is borne by each enzymatic subunit of the cellulosome.[17,18] The ~70-residue dockerin contains two duplicated sequences, each of about 22 amino acid residues. Like the cohesins, the sequences of the dockerins are also very similar.

It has been demonstrated for *C. thermocellum* and *C. cellulolyticum* that the incorporation of the enzymatic subunits into the cellulosome complex is accomplished via a specific interaction between the cohesin and the dockerin domains.[19] Our study and that of Yaron and colleagues,[20] found that the cohesin–dockerin interaction is calcium-dependent in both strains, and apparent equilibrium dissociation constants of the interaction have been reported.[8,21]

CipA also bears a single dockerin domain at its C terminus.[13] However, the CipA dockerin fails to recognize any of its own complement of cohesins,[21] rather, it binds specifically to a different class of cohesin (termed type-II cohesin) located on the cell surface.[22] This arrangement serves to anchor the CipA scaffoldin together with its set of enzyme subunits to the cell. The cellulosomes of *C. thermocellum* are thus implanted into a specialized type of cell-surface protuberancelike organelle.[23,24] By contrast, the mesophilic strains, *C. cellulolyticum* and *C. cellulovorans*, apparently lack a dockerin on their respective scaffoldin subunits, and it is yet unclear whether their cellulosomes are cell-surface components, and, if so, how they might be anchored thereto.

Several important issues concerning the cellulosome assemblage and its stoichiometry are still unclear. In this context, recent evidence has shown that the cohesins of both *C. thermocellum* and *C. cellulolyticum* recognize nearly all of the dockerins on their own enzymatic subunits,[8,16,20,25–27] which would seem to be consistent with the intrinsic similarity of their sequences. It thus seems that, within a given species, the incorporation of the cellulosomal enzymes into the complex is a nonselective process, and the catalytic subunits would appear to interact randomly along the CipA scaffolding protein.

Due to the intraspecies sequence similarity among the various cohesins and dockerins, it was of interest to determine whether cohesins from cellulosomes of one species of bacterium will recognize dockerins from another. In this communication, we show conclusively that the cohesin–dockerin interaction between *C. thermocellum* and *C. cellulolyticum* cellulosomes is a species-specific phenomenon. Based on cross-species sequence comparison, we predict that four amino-acid residues on the dockerin domain serve as recognition codes, critical to the observed binding specificity.

MATERIALS AND METHODS

Bacterial Strains and Vectors

Genomic DNA was prepared from *C. cellulolyticum* ATCC 35319 and *C. thermocellum* YS as described previously.[8,20] *Escherichia coli* BL21(DE3) was used as the host for the *C. cellulolyticum* expression vector pETCip1 (pET22b(+) derivative plasmid; Novagen, Abingdon, UK).[8] *E. coli* strain XL-1 Blue (Stratagene; La Jolla, CA) was used as the host for the expression vector pCBD3[20] from *C. thermocellum* [pTrc99A derivative plasmid; Pharmacia, Uppsala, Sweden].

Dockerin Sequences

The sequences for the dockerins were obtained from their respective cellulosome component (when available, Swiss-Prot accession number included parenthetically). From *C. thermocellum:* cellulosome subunits CelA (P04955); CelB (P04956); CelD (P04954); CelE (P10477); CelF (P26224); CelG (Q05332); CelH (P16218); CelJ[28]; CelS (P38686); LicB (P29716); XynY (P51584); XynZ (P10478); and the scaffoldin CipA (Q06851). From *C. cellulolyticum:* CelA (P17901); CelC (P37699); CelD (P25472); CelF (P37698); CelG (P37700). The cellulosomal cellulases CelE, CelH, and CelJ from *C. cellulolyticum* were recently sequenced (A. Bélaïch et al., unpublished results).

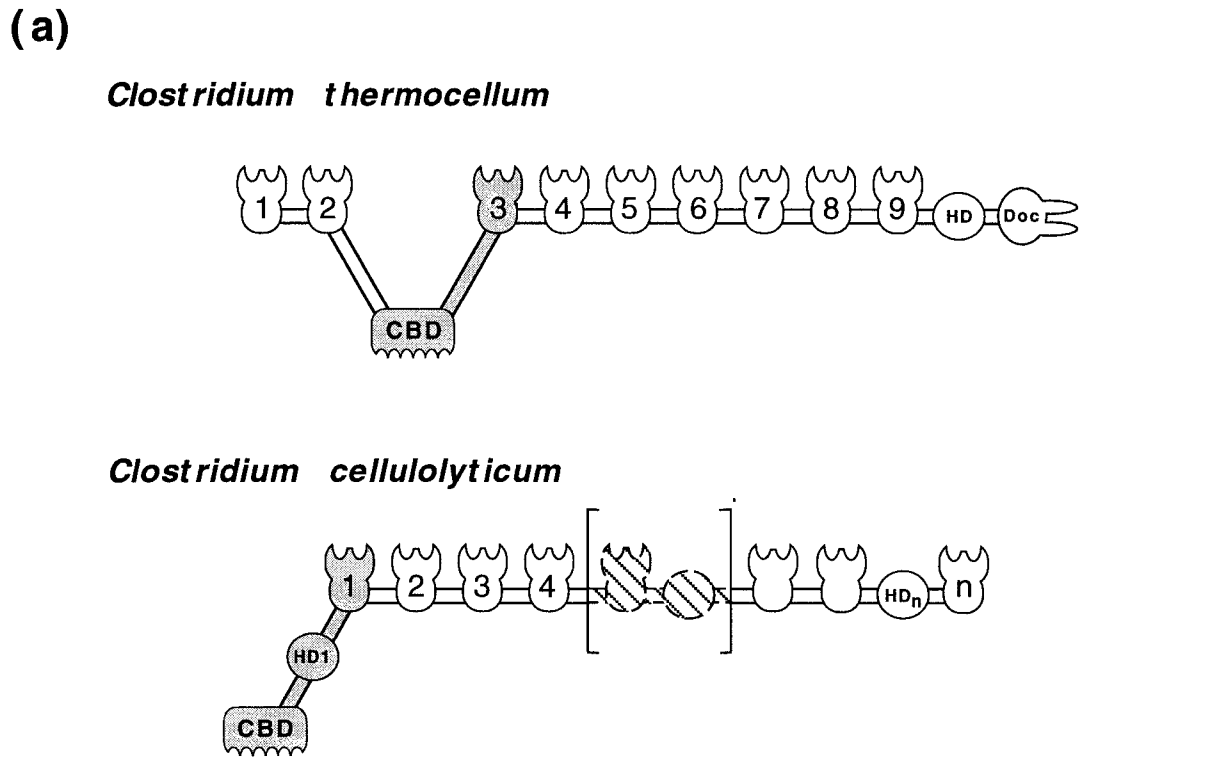
Cellulosome Purification

The purification of the cellulosome of *C. cellulolyticum*, grown on Mn 300 cellulose (7.5 g/l) (Fluka Chemie AG, Buchs, Switzerland), was carried out as described by Gal et al.[16] The cellulosome of *C. thermocellum* was purified from culture broth using the affinity digestion procedure.[29]

Expression and Purification of Recombinant Proteins

The nature of the two cohesin-containing probes used in these studies is shown schematically in Figure 1.

Plasmid DNA ($pETCipC_1$) was used to transform *E. coli* BL21(DE3) containing an inducible T7 polymerase. The general protocol used for expression of the *C. cellulolyticum* construct (CBD–HD1–Cohl-cc; originally termed $miniCipC_1$) has been described previously.[8] The pTrc99A derivative plasmid containing the trc promotor (induced with 1–5 mM of IPTG) was used to transform *E. coli* XL1-Blue. The transformed cells were grown on LB medium containing



(b)

		β1 β2 β3 β4 β5
Coh3-ct	562	IKIKVDTVNAKPGDTVNIPVRFSGIPS-KGIANCDFVYSYDPNVLEIIEIKPGELIVDPNPDKSFDTAVYP
Coh1-cc	288	LKVTVGTANGKPGDTVTVPVTFADVAKMKNVGTCNFYLGYDASLLEVVSVDAGPIVKN--AAVNFSSSASN
		β6 β7 β8 β9
Coh3-ct	632	DRKIIVFLFAEDSGTGAYAITKDGVFATIVAKVKSGAPNGLSVIKFVEVGGFANNDLVEQ-RTQFFDGGVNV
Coh1-cc	425	G--TISFLFLDNT-ITDELITADGVFANIKFKLKSVTAKTTTPVTFKDGGAFGDGTMSKIASVTKTNGSVTI

Fig. 1. Comparative organization of the cohesin domains within the scaffoldin subunits of *C. cellulolyticum* and *C. thermocellum.* **a:** Scaffoldin subunits from the two species. In each case, the shaded portion of the scaffoldin subunit (designated CBD–Coh3-ct and CBD–HD1–Coh1-cc) indicates the contents of the cohesin-containing probe used for interaction with desired dockerin-containing components. Shown are the positions of the cohesin domains (numbered) relative to the CBD of the designated species. The scaffoldin of *C. cellulolyticum* also bears at least two hydrophilic domains (HD), whereas that of *C. thermocellum* has only one such domain. The *C. thermocellum* scaffoldin also harbors a single N-terminal dockerin domain (Doc). The scheme of the scaffoldin from *C. thermocellum* is based on its sequence, determined by Gerngross et al.,[13] and the putative scheme for the *C. cellulolyticum* scaffoldin is based on the nearly completed sequence (A. Bélaïch et al., unpublished results). Confirmed portions of the latter include the CBD, the first and last HD, and the first four and last three cohesin domains. The sequence of an ~1-kb section (crosshatched and bracketed) is still pending; this section could conceivably accommodate two extra cohesin domains, or, alternatively, an additional cohesin domain plus one or two HDs. **b:** Structure-based, amino acid alignment of the *C. cellulolyticum* and *C. thermocellum* cohesin domains used in this study. Identical residues are indicated by vertical lines and similar residues by dots. Only the sequences of the respective cohesin domain is shown (i.e., without the CBD, adjacent linker sequences and hydrophilic domain). The position of the β strands is based on the recently reported structure of cohesin-2 from CipA.[36] Coh3-ct, CipA cohesin domain 3 from *C. thermocellum;* Coh1-cc, CipC cohesin domain 1 from *C. cellulolyticum.*

ampicillin. For the *C. thermocellum* probe, the recombinant CBD–Coh3-ct product was expressed as described by Yaron et al.[20] Both probes were purified by using an affinity-chromatographic procedure, by virtue of the inherent cellulose-binding properties of the CBD present in the recombinant proteins.

The expression and purification of $CelA_2$–DOC and $CelA_3$ were performed as described by Fierobe et al.[30]

Recognition of Cellulosomal Subunits Using Biotinylated Proteins

Biotinylation of CBD–HD1–Coh1-cc and CBD–Coh3 via their lysine residues was performed using biotinyl *N*-hydroxysuccinimide ester, as described by Bayer and Wilchek[31] (biotin labeling kit; Boehringer-Mannheim, Germany). The cellulosomal subunits, and the endoglucanases $CelA_2$–DOC and $CelA_3$ were separated by SDS-PAGE or native-PAGE, transferred onto BA83 nitrocellulose membranes (Schleicher & Shuell, Dassel, Germany), and treated as described by Pagès et al.[8] Blots were incubated in the presence of 15 mM $CaCl_2$ for 1 hour at 23°C before introduction of the desired biotinylated probe.[20] Biotinylated proteins were detected using streptavidin-conjugated peroxidase (Biotin/Streptavidin Western blotting kit; Boehringer-Mannheim).[32] In indicated experiments, calcium was omitted and blots were treated in the presence of EDTA (5 mM EDTA dissolved in 50 mM Tris and 150 mM sodium chloride, pH 7.5, containing 2% bovine serum albumin). As described previously,[8,20,33] a variety of controls was routinely instituted to guarantee specificity of the interaction. In all cases, protein transfer was monitored by staining the nitrocellulose membrane with Ponceau Red.

Secondary Structure Prediction

Alignment of the dockerin sequences and secondary structure prediction was carried out according to the EMBL network systems: MAXHOM multiple sequence alignment and PHD–Profile fed neural network systems from HeiDelberg.[34,35]

RESULTS

Interaction of Recombinant Cohesins With Cellulosomal Subunits

Blots containing cellulosomal components from either *C. thermocellum* or *C. cellulolyticum* were probed with recombinant segments of the scaffoldin subunit from the two species. For this purpose, a construct termed CBD–Coh3-ct was used, which contains the cellulose-binding domain and the cohesin-3 domain of scaffoldin subunit CipA (Fig. 1A). Likewise, a second construct termed CBD–HD1–Coh1-cc was used, which contains the CBD, the first hydrophilic domain and the cohesin-1 domain of the CipC scaffoldin. The recombinant cohesin-containing proteins were biotinylated, and blots were developed using a streptavidin peroxidase conjugate.

Samples of the *C. thermocellum* cellulosome preparation were separated into subunits by SDS-PAGE and transferred onto nitrocellulose membranes (Fig. 2A,B). As described previously,[20] the CBD–Coh3-ct construct interacts with all cellulosomal subunits in *C. thermocellum* between S3 and S13 (inclusive) (Fig. 2A), and the recognition pattern of the CBD–Coh3-ct product is similar for the three cellulosomal concentrations tested. It was previously reported that the interaction between the recombinant cohesin domains and the cellulosomal subunit is, in the case of *C. thermocellum*, enhanced by calcium.[20] Therefore all binding experiments reported in this paper were performed in 15 mM $CaCl_2$.

A second blot of the SDS-PAGE-treated *C. thermocellum* cellulosome was probed with the recombinant *C. cellulolyticum* cohesin (Fig. 2B). No binding was detected between CBD–HD1–Coh1-cc and any of the *C. thermocellum* cellulosomal subunits.

In another series of experiments, *C. cellulolyticum* cellulosomal subunits and purified, recombinant endoglucanases $CelA_2$–DOC (a 50-kDa construct, which includes the C-terminal dockerin domain) and $CelA_3$ (44-kDa, the same enzyme, but lacking the dockerin domain) were separated by SDS-PAGE and blotted onto nitrocellulose (Fig. 2C,D). In one experiment, blots were allowed to interact with biotinylated CBD–HD1–Coh1-cc (Fig. 2C) and stained with streptavidin peroxidase. The results corroborated previous reports[16] that the cohesin-containing construct interacts specifically with most or all of the cellulosomal enzymatic subunits from *C. cellulolyticum* (Fig. 2C, lane 1), including the two major subunits CelE (94 kDa) and CelF (80.6 kDa). This interaction[8] is clearly dockerin-dependent, since only $CelA_2$–DOC interacts with the biotinylated probe (Fig. 2C, lane 3), whereas $CelA_3$ (Fig. 2C, lane 2) fails to interact.

In a second experiment, blots containing the *C. cellulolyticum* samples were incubated with biotinylated CBD–Coh3-ct (from *C. thermocellum*). In this case, no binding was detected (Fig. 2D); the probe failed to interact with the cellulosome subunits or with the two endoglucanases ($CelA_2$–DOC and $CelA_3$).

Collectively, these results indicate that the formation of the cohesin–dockerin complex is species-specific.

Interaction of Recombinant Cohesins With Undenatured Endoglucanases

To determine whether the presence of detergent in SDS-PAGE may have interfered subsequently with the cross-species cohesin–dockerin interaction, the binding between the two CelA preparations from *C. cellulolyticum* and the recombinant cohesins were analyzed under nondenaturing conditions (Fig. 3A,B). The endoglucanases (Fig. 3, lane 2) were separated

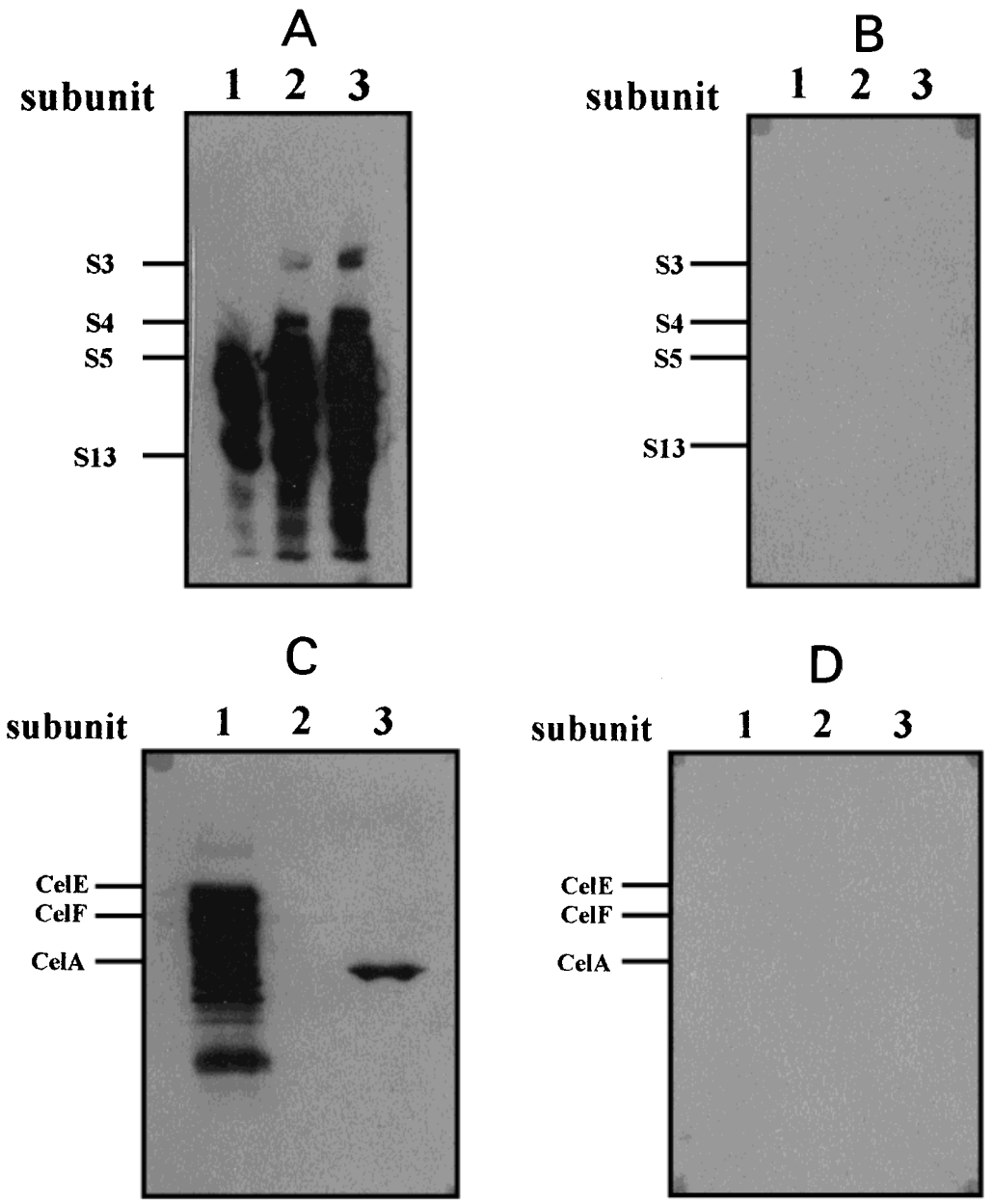


Fig. 2. Intra- and cross-species recognition of SDS-PAGE-separated cellulosomal components by recombinant cohesin domains. SDS-PAGE-separated samples were blotted onto nitrocellulose and probed with the desired biotinylated cohesin construct (10 μg per sample). Binding experiments were performed in the presence of 15 mM $CaCl_2$. Blots were developed with streptavidin peroxidase complex. **Gels A and B:** Cellulosome preparations from *C. thermocellum; lanes 1–3:* different amounts of sample applied (0.2, 0.7, and 2 μg, respectively). **Gels C and D:** Samples from *C. cellulolyticum; lane 1:* a sample (5 μg) of the *C. cellulolyticum* cellulosome; *lane 2:* a 3.5-μg sample of endoglucanase $CelA_3$ (which lacks a dockerin domain); and *lane 3:* an equivalent sample of endoglucanase $CelA_{2-DOC}$ (which contains a dockerin domain). **Gels A and D:** Affinity blotting using CBD–Coh3-ct (from *C. thermocellum*). **Gels B and C:** Affinity blotting using CBD–HD1–Coh1-cc (from *C. cellulolyticum*). The standard positions for various cellulosomal enzyme subunits from each of the two bacterial strains are indicated.

by native-PAGE and blotted onto nitrocellulose membranes. The latter were probed with the recombinant cohesin-containing proteins from both *C. cellulolyticum* and *C. thermocellum*. Similar to the results of the SDS-PAGE-treated cellulosomal subunits, CBD–HD1–Coh1-cc interacted exclusively with endoglucanase $CelA_{2-DOC}$ (Fig. 3A), while CBD–Coh3-ct failed to react with either $CelA_{2-DOC}$ or $CelA_3$, that is, the interaction is species-dependent.

Calcium Dependence of the Cohesin-Dockerin Interaction in *C. cellulolyticum*

It was demonstrated in a previous study,[20] that the interaction between CBD–Coh3-ct and the cellu-

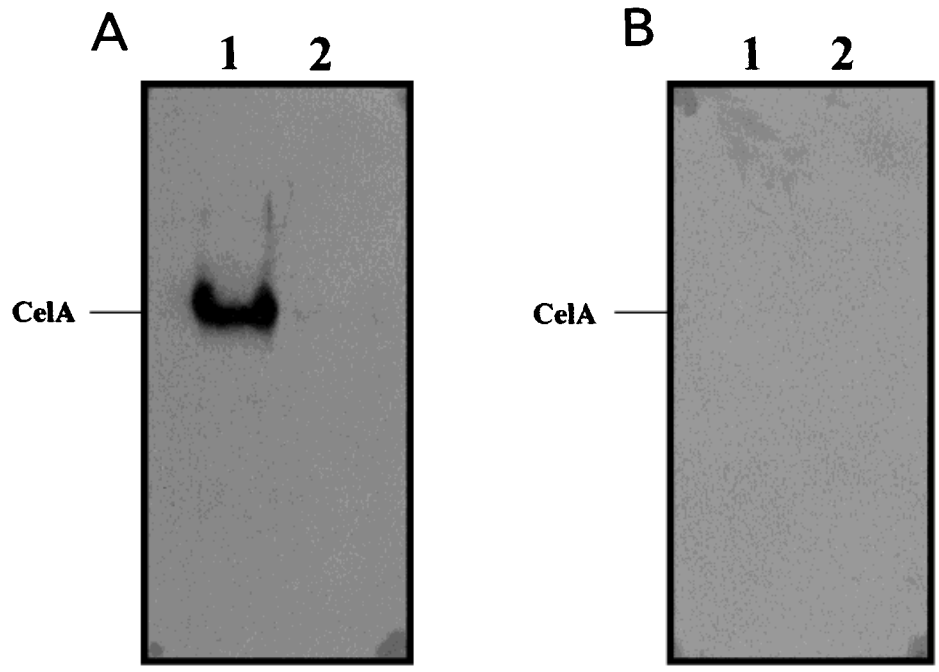


Fig. 3. Interaction between recombinant cohesin domains and purified endoglucanases from *C. cellulolyticum,* following native-PAGE. Dockerin-containing endoglucanase $CelA_{2-DOC}$ and the truncated "dockerin-less" $CelA_3$ (*lanes 1* and *2,* respectively; 3.5 µg per sample) were subjected to native-PAGE (in the absence of SDS) and blotted onto nitrocellulose membranes. Samples were probed using biotinylated CBD–HD1–Coh1-cc (**gel A**) and CBD–Coh3-ct (**gel B**) as described in the legend to Figure 2. The known position for endoglucanase CelA is designated.

losomal subunits of *C. thermocellum* is calcium-dependent. We therefore examined whether the same conditions used for the *C. thermocellum* system would reveal a similar calcium dependence for the *C. cellulolyticum* system. For this purpose, the purified dockerin-containing endoglucanase ($CelA_{2-DOC}$) was subjected to SDS-PAGE, blotted, allowed to interact with biotinylated CBD–HD1–Coh1-cc in the presence and absence of EDTA. The results (not shown) demonstrated unequivocally that the binding between the cohesin-1 and the dockerin of endoglucanase $CelA_{2-DOC}$ is calcium-dependent. In the presence of EDTA, no label could be observed, whereas in the absence of EDTA, a clear band was evident.

DISCUSSION

The results of this study demonstrate that the formation of cohesin-dockerin complexes between the two anaerobic, cellulosome-producing bacteria, *C. thermocellum* and *C. cellulolyticum,* is species-specific, that is, a cohesin of one species will only interact with dockerins of the same species. Numerous sequences of dockerin and cohesin domains from *C. thermocellum, C. cellulolyticum,* and *C. cellulovorans* are currently known. Phylogenetic trees based on these sequences[15] revealed that both types of domain are generally separated into two clusters, thus distinguishing the thermophilic *C. thermocellum* from the mesophilic *C. cellulolyticum* and *C. cellulovorans.*

The crystal structure of one of the cohesin domains from the CipA scaffoldin subunit of *C. thermocellum* has recently been solved.[36,37] The domain forms a 9-stranded β sandwich with a jellyroll topology, surprisingly similar to the fold displayed by the CBD derived from the same subunit.[38] It is clear from sequence homology that the closely related cohesins from the three confirmed species of cellulosome-producing bacteria would share the same structural fold. Indeed, the sequences of the two cohesin domains used in this study (cohesin-3 of the *C. thermocellum* cellulosome and cohesin-1 of *C. cellulolyticum*) are very similar, displaying about 50% identical or similar residues (Fig. 1B). Based on surface residues among the known cohesins that exhibit both intraspecies conservation and cross-species dissimilarity, a preferred dockerin-binding face was proposed,[36] which could include residues on β strand 5, the 4/5, 5/6, and 6/7 loops.

On the other hand, the structure of the dockerin domain is not yet known. Nevertheless, we may have some clues as to its possible structure by analyzing the known dockerin sequences in clostridia. For such an analysis, we can subdivide the dockerin sequence into seven different parts (Fig. 4). All clostridial dockerins thus far described contain a duplicated

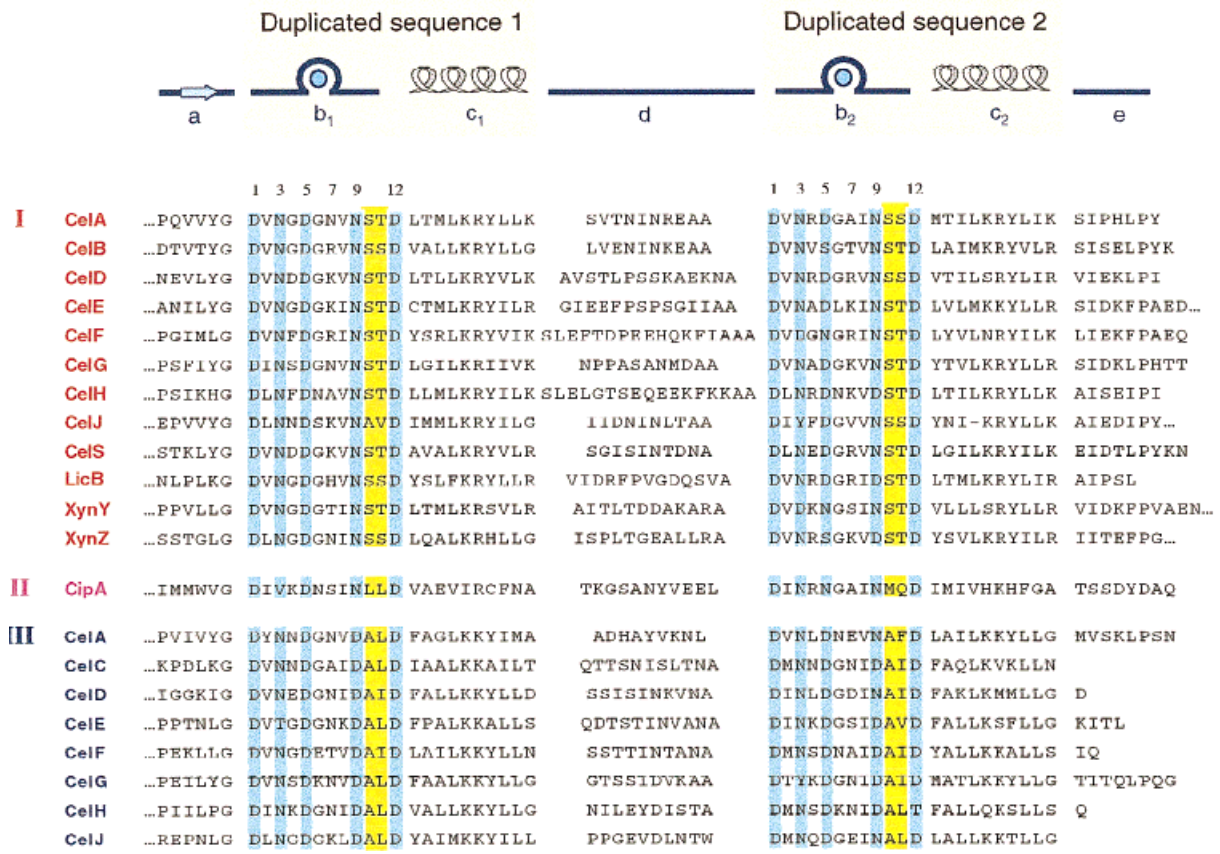


Fig. 4. Amino acid alignment of the known dockerin domains from *C. thermocellum* and *C. cellulolyticum*. **Group I:** dockerins of cellulosomal enzymes from *C. thermocellum*. **Group II:** dockerin of scaffoldin subunit from *C. thermocellum*. **Group III:** dockerins of cellulosomal enzymes from *C. cellulolyticum*. Dockerin sequences for CelE, CelH and CelJ from *C. cellulolyticum* were determined recently (A. Bélaïch et al., unpublished results). The sequences are partitioned into seven segments (designated *a* through *e*), and the 22-residue duplicated sequences are indicated. The residues of the calcium-binding motifs (b_1 and b_2), in which the side chains purportedly react with the calcium ion, are shaded in *blue*. The residues (positions *10* and *11* of the calcium motifs), suspected of serving as selectivity determinants, are indicated in *yellow*.

sequence of about 22 amino acid residues, the first 12 of which (Fig. 4, segments b_1 and b_2) are homologous to a known structure, that is, the calcium-binding loop in the EF-hand motif.[39,40] It is assumed that this 12-residue sequence in the dockerin structure would adopt the same fold when bound to calcium (see schematic diagrams depicted in Fig. 2.13 of Ref. 40).

The potential calcium-binding characteristic of this motif in the duplicated sequence of a cellulosomal enzyme was first noted by Chauvaux et al.[41] In this regard, it has been demonstrated that the cohesin–dockerin interaction of *C. thermocellum* is in fact calcium-dependent.[20] This property probably reflects the latter motif of the dockerin, since the cohesins themselves apparently lack calcium.[36] Moreover, in nearly all of the dockerins, all of the designated calcium-binding residues (notably the aspartic acid and asparagine or serine at positions 1, 3, 5, 9, and 12) are conserved in both duplicated sequences, the only exceptions being position 3 in segment b_1 of the CipA dockerin, and the same position in segment b_2 of the CelJ (*C. thermocellum*) and CelG (*C. cellulolyticum*) dockerins. It is not known whether these substitutions and the nonglycine residues at position 6[40] would serve to hinder (i.e., lower the affinity) or block entirely the proposed calcium-binding property of these segments.

The second halves of the duplicated sequences (Fig. 4, segments c_1 and c_2) appear to form α helices (Table I), according to PHD secondary structure prediction (PHDsec).[34,35] These helices would be analogous to the F helix of the EF-hand motif. The two duplicated sequences of the dockerin domain are connected by a sequence of variable length (Fig. 4, segment d) in the different dockerins, which could, perhaps, form a structured loop, according to PHDsec analysis. Finally, the dockerins bear connecting stretches (segments a and e) on the N and C termini of the duplicated sequences. A short β strand is predicted for part of segment a. The evidence thus suggests that, structurally, the dockerin domain may not be strictly equivlent to the EF-hand motif, but

Table I. Predicted Secondary Structure for Dockerin Domains*

	....,....1....,....2....,....3....,....4....,....5....,....6....,..
AA	PQVVYGDVNGDGNVNSTDLTMLKRYLLKSVTNINREAADVNRDGAINSSDMTILKRYLIKSIPHLPY
PHD sec	EEEEE HHHHHHHHHHHHHH HHHHHHHHHHHHHHHHH
Re1 sec	97888532798975657799999999883077763345538899656265999999999999972499
prH sec	00000000000001167789999999885411011221110100001577899999999999975200
prE sec	02888764100011000000000000000000011322122000012000000000000000000000
prL sec	97111235898986721100000000113488875556658899767322000000000000014699
Subset	LLEEEE..LLLLLLLHHHHHHHHHHHHH..LLLL...LL.LLLLLLL.HHHHHHHHHHHHHH..LL

Abbreviations: H, helix; E, extended (sheet); blank, other (loop); Rel, Reliability index of prediction (0–9); pr, "probability" for assigning designated secondary structure (scaled to the nearest integer, 0–9); Subset, a subset of the prediction, for all residues with an expected average accuracy >82% (tables in header), where L signifies loop and "." indicates lack of prediction (Rel <5).
*According to PHD: Profile network prediction HeiDelberg—PHDsec. Protein predicted length: 67 AA residues, based upon MAXHOM multiple sequence alignment of 20 different dockerin sequences from *C. thermocellum* and *C. cellulolyticum*.

would represent a variation thereof. The duplicated stretches would each contain a calcium-binding loop and an F helix equivalent, but the equivalent of an E helix* may be missing (Table I) and may be replaced by alternative structural elements in segments a and d. The exact nature of the structural elements of the dockerin domain remains to be resolved by x-ray crystallography and NMR.

As stated above and shown previously,[8,16,20,25,27] the cohesin–dockerin interaction appears to be relatively nonselective among components of the same species. Yet, between species, the interaction seems to be specific, despite the high degree of sequence similarity (Fig. 4, groups I and III). Moreover, the dockerin of the scaffoldin subunit (CipA) in *C. thermocellum* (group II) fails to recognize any of the cohesins on the same subunit. These data and knowledge of the various dockerin sequences allow us to speculate on the specificity determinants of the dockerin domain.

Strict cross-species conservation of residues among the dockerins at a given position could signify either a critical contribution to the overall structural fold or a general participation of these residues in the binding interaction. On the other hand, strict intraspecies conservation of residues coupled with cross-species variance would indicate a role in the biorecognition of the cohesin domain. Thus, in analyzing the sequences from the three groups of dockerin domains, we can establish the following criteria based on the known specificity characteristics: (i) *intraspecies* comparison of residues should be conserved (i.e., identity or similarity†) at a given position in dockerins from group I or group III (Fig. 4), (ii) *cross-species* comparison between the invariant residues of group I and group III should show a high degree of dissimilarity, and (iii) residues of the CipA (scaffoldin) dockerin from *C. thermocellum* (group II) should differ from those of either of the two other groups. It is also clear that the side chains of residues known to participate directly in calcium-binding would presumably be involved in this function and would not contribute directly to the specificity properties of the dockerin domain.

Using these criteria, residues of the N- and C-terminal segments (a and e) can immediately be excluded, since they fail to fulfill criteria i and ii. Regarding segment d, although intra- and cross-species relationships among the different dockerins are evident, it appears that this segment would not confer specificity toward the cohesin-dockerin interaction, if based solely on the varying length of the observed sequences.

We now address the status of the duplicated sequences. In reviewing the positions of the residues in segments c_1 and c_2 (the predicted α helices), once again criteria i, ii, and/or iii are not maintained. Regarding the calcium-binding segments (b_1 and b_2), the contribution of residues 1, 3, 5, 9, and 12 can be eliminated, since they are involved in calcium binding. In addition, criterion ii is not fulfilled. Most of the other residues in this segment also fail to satisfy one or more of the criteria, *except* residues in positions 10 and 11. These four positions (i.e., the two copies of positions 10 and 11 in the duplicated sequences) reveal both intraspecies conservation and cross-species diversity. The CipA dockerin is also different than the other two groups at these positions.

We therefore propose these four amino acid residues of the dockerin domain as primary candidates for conferring specificity towards the cohesin–dockerin interaction. The apparent location of these four residues is also of interest. As part of the calcium-binding motif, but not directly involved in calcium binding, the four residues would apparently be displayed correctly *when calcium is bound*. This would also explain the observed calcium dependence and interference in the cohesin-dockerin interaction by EDTA. In singling out these four residues as

*E-helix equivalents would not be part of the duplicated sequences, but would comprise N-terminal flanking regions in segments a and d (see Fig. 4).

†For the purposes of this article, similar residues are defined as follows: (A, G), (D, E), (D, N), (E, Q), (I, L, M, V), (K, R), (S,T).

recognition determinants, we distinguish their function from other dockerin residues which could also participate in the *binding* of the cohesins and which could be conserved among the different species.

In a recent study by Choi and Ljungdahl,[42] synthetic branched duplicated sequences were prepared and subjected to interaction with cellulosomal subunits. It was concluded by the authors that only the first duplicated sequence (which contains segments b_1 and c_1) is capable of binding calcium and interacting with the cohesins, whereas the second sequence (i.e., segments b_2 and c_2) fails to do either. The authors presented a model whereby calcium is assumed to participate *directly* in the binding between the cohesin domain and the first duplicated sequence of the dockerin domain. The model for the cohesin-dockerin interaction as presented in the present work would thus contradict the model presented by Choi and Ljungdahl. Close inspection of the synthetic sequences used for that work reveals that the first duplicated sequence contains 3 extra residues located in segment a, whereas the second sequence lacks analogous flanking residues and commences precisely at the beginning of segment b_2 (i.e., the initial aspartate residue). The extra flanking residues might be important for inducing the correct fold of the calcium binding motif. This difference in the two synthetic derivatives might indeed explain the experimentally observed differences in binding of both calcium and the cellulosomal enzyme components by the synthetic sequences.[42]

In reviewing the various dockerin sequences from *C. thermocellum*, it is interesting to note that criterion i does not hold for the CelJ dockerin (see Fig. 4). In this case, positions 10 and 11 (Ala-Val) of segment b_1 are clearly different from those (Ser-Ser) of the CelJ segment b_2 and from those (Ser-Ser/Thr) of the other dockerins from this species. However, CelJ was recently identified as the S2 subunit of the cellulosome,[28] and, in an independent study,[20] this particular subunit failed to interact with recombinant cohesins of the scaffoldin subunit. The CelJ dockerin thus provides an internal inconsistency which provides further proof to the rule. CelJ may therefore be a special type of cellulosomal subunit; the specificity characteristics of its dockerin domain and its possible mode of incorporation into the complex have yet to be determined.

The proposed involvement of positions 10 and 11 of the two calcium-binding dockerin motifs leads to an additional prediction: the amino acids at these positions in *C. cellulolyticum* bear close resemblance to those of the known dockerin sequence from another mesophilic species (from *C. cellulovorans*).[43] This would infer that the dockerins and cohesins from these two mesophilic clostridial cellulosomes may cross-react, and that none of them would interact with their counterparts from *C. thermocellum*. The interspecies cross-specificities of *C. cellulovorans* are currently unknown and will be examined in future studies, when the relevant dockerin and cohesin probes become available.

Duplicated calcium-binding motifs are also present in the sequences of various ruminococcal cellulases, thus suggesting the existence of ruminococcal dockerin domains and hence cellulosomes. In the few sequences currently available,[44,45] positions 10 and 11 of the motif appear to exhibit characteristic species specificity, which is different than those of the clostridial dockerins.

A recent study of fungal enzymes[46] has also revealed 40-residue duplicated sequences reminiscent of the dockerins, which bound to separate polypeptide components from two different fungal species but failed to recognize bacterial components. It is thus presumed that such fungi indeed produce a cellulosomelike complex in a manner similar to bacteria and that the specificity of the docking function is conserved among their cellulosomal enzymes. Interestingly, 12-residue stretches, which conform to the calcium-binding motif, can also be discerned, commencing at position 24 within the 40-residue duplicated fungal sequences (see Ref. 46). Position 33 of the latter sequence (i.e., the equivalent of position 10 of the bacterial calcium-binding motif) is invariably asparagine; position 34 (the position-11 equivalent) is either asparagine, aspartic acid, or glycine. This may indicate a very strict requirement for the characteristic recognition residue in position 10.

The data presented here suggest an approach for interconverting the specificity determinants of the *C. thermocellum* and *C. cellulolyticum* dockerin domains by site-directed mutagenesis. In modifying the suspected recognition codes, we hope to reverse the inherent specificities, such that the dockerin of one species will no longer recognize its own cohesins but will recognize those of the second species. Molecular modeling procedures may also provide additional structural evidence concerning the array of amino acid residues on both the dockerin and cohesin domains which may be responsible for the binding specificity and strong interaction. In the final analysis, the structures of the dockerin and the cohesin–dockerin complex await solution by crystallography and/or NMR techniques.

ACKNOWLEDGMENTS

We thank H.-P. Fierobe for helpful discussions, L. Gal for preparation of the *C. cellulolyticum* cellulosome, and Sima Yaron for preparation of the *C. thermocellum* cellulosome and CBD-Coh3 construct. This research was supported by grants from the Centre National de la recherche Scientifique, the Université de Provence and the Région Provence-Alpes-Côte d'Azur and from the Israel Science Foundation (administered by the Israel Academy of Sciences and Humanities, Jerusalem), and a contract

from the European Commission (Biotechnology Programme, Proposal PL962303). Technical support was provided by the Technion Otto Meyerhof Biotechnology Laboratories established by the Minerva Foundation, Germany.

REFERENCES

1. Lamed, R., Bayer, E.A. The cellulosome concept: Exocellular/extracellular enzyme reactor centers for efficient binding and cellulolysis. In: "Biochemistry and Genetics of Cellulose Degradation." Aubert, J.-P., Beguin, P., Millet, J. (eds.). London: Academic Press, 1988:101–116.
2. Bayer, E.A., Morag, E., Lamed, R. The cellulosome: A treasure-trove for biotechnology. Trends Biotechnol. 12:378–386, 1994.
3. Bayer, E.A., Morag, E., Shoham, Y., Tormo, J., Lamed, R. The cellulosome: A cell-surface organelle for the adhesion to and degradation of cellulose. In: "Bacterial Adhesion: Molecular and Ecological Diversity." Fletcher, M. (ed.). New York: Wiley-Liss, Inc., 1996:155–182.
4. Béguin, P., Lemaire, M. The cellulosome: an exocellular, multiprotein complex specialized in cellulose degradation. Crit. Rev. Biochem. Mol. Biol. 31:201–236, 1996.
5. Doi, R.H., Goldstein, M., Hashida, S., Park, J.S., Takagi, M. The *Clostridium cellulovorans* cellulosome. Crit. Rev. Microbiol. 20:87–93, 1994.
6. Felix, C.R., Ljungdahl, L.G. The cellulosome: The exocellular organelle of Clostridium. Annu. Rev. Microbiol. 47:791–819, 1993.
7. Lamed, R., Setter, E., Kenig, R., Bayer, E.A. The cellulosome: A discrete cell surface organelle of *Clostridium thermocellum* which exhibits separate antigenic, cellulose-binding and various cellulolytic activities. Biotechnol. Bioeng. Symp. 13:163–181, 1983.
8. Pagès, S., Belaich, A., Tardif, C., Reverbel-Leroy, C., Gaudin, C., Belaich, J.-P. Interaction between the endoglucanase CelA and the scaffolding protein CipC of the *Clostridium cellulolyticum* cellulosome. J. Bacteriol. 178:2279–2286, 1996.
9. Pohlschröder, M., Leschine, S.B., Canale-Parola, E. Multicomplex cellulase-xylanase system of *Clostridium papyrosolvens* C7. J. Bacteriol. 176:70–76, 1994.
10. Lamed, R., Morag (Morgenstern), E., Mor-Yosef, O., Bayer, E.A. Cellulosome-like entities in *Bacteroides cellulosolvens*. Curr. Microbiol. 22:27–33, 1991.
11. Lamed, E., Naimark, J., Morgenstern, E., Bayer, E.A. Specialized cell surface structures in cellulolytic bacteria. J. Bacteriol. 169:3792–3800, 1987.
12. Wilson, C., Wood, C.M. The anaerobic fungus *Neocallimastix frontalis:* Isolation and properties of a cellulosome-type enzyme fraction with the capacity to solubilize hydrogen-bond-ordered cellulose. Appl. Microbiol. Biotechnol. 37:125–129, 1992.
13. Gerngross, U.T., Romaniec, M.P.M., Kobayashi, T., Huskisson, N.S., Demain, A.L. Sequencing of a *Clostridium thermocellum* gene (cipA) encoding the cellulosomal S_L-protein reveals an unusual degree of internal homology. Mol. Microbiol. 8:325–334, 1993.
14. Shoseyov, O., Takagi, M., Goldstein, M.A., Doi, R.H. Primary sequence analysis of *Clostridium cellulovorans* cellulose binding protein A. Proc. Natl. Acad. Sci. U.S.A. 89:3483–3487, 1992.
15. Belaich, J.-P., Tardif, C., Belaich, A., Gaudin, C. The cellulolytic system of *Clostridium cellulolyticum*. J. Biotechnol. in press, 1997.
16. Gal, L., Pagès, S., Gaudin, C., Belaich, A., Reverbel-Leroy, C., Tardif, C., Belaich, J.-P. Characterization of the cellulolytic complex (cellulosome) produced by *Clostridium cellulolyticum*. Appl. Environ. Microbiol. 63:903–904, 1997.
17. Tokatlidis, K., Salamitou, S., Béguin, P., Dhurjati, P., Aubert, J.-P. Interaction of the duplicated segment carried by *Clostridium thermocellum* cellulases with cellulosome components. FEBS Lett. 291:185–188, 1991.
18. Salamitou, S., Tokatlidis, K., Béguin, P., Aubert, J.-P. Involvement of separate domains of the cellulosomal protein S1 of *Clostridium thermocellum* in binding to cellulose and in anchoring of catalytic subunits to the cellulosome. FEBS Lett. 304:89–92, 1992.
19. Tokatlidis, K., Dhurjati, P., Béguin, P. Properties conferred on *Clostridium thermocellum* endoglucanase CelC by grafting the duplicated segment of endoglucanase CelD. Protein Eng. 6:947–952, 1993.
20. Yaron, S., Morag, E., Bayer, E.A., Lamed, R., Shoham, Y. Expression, purification and subunit-binding properties of cohesins 2 and 3 of the *Clostridium thermocellum* cellulosome. FEBS Lett. 360:121–124, 1995.
21. Salamitou, S., Raynaud, O., Lemaire, M., Coughlan, M., Béguin, P., Aubert, J.-P. Recognition specificity of the duplicated segments present in *Clostridium thermocellum* endoglucanase CelD and in the cellulosome-integrating protein CipA. J. Bacteriol. 176:2822–2827, 1994.
22. Leibovitz, E., Béguin, P. A new type of cohesin domain that specifically binds the dockerin domain of the *Clostridium thermocellum* cellulosome-integrating protein CipA. J. Bacteriol. 178:3077–3084, 1996.
23. Bayer, E.A., Setter, E., Lamed, R. Organization and distribution of the cellulosome in *Clostridium thermocellum*. J. Bacteriol. 163:552–559, 1985.
24. Bayer, E.A., Lamed, R. Ultrastructure of the cell surface cellulosome of *Clostridium thermocellum* and its interaction with cellulose. J. Bacteriol. 167:828–836, 1986.
25. Kruus, K., Lua, A.C., Demain, A.L., Wu, J.H.D. The anchorage function of CipA (CelL), a scaffolding protein of the *Clostridium thermocellum* cellulosome. Proc. Natl. Acad. Sci. U.S.A. 92:9254–9258, 1995.
26. Lytle, B., Myers, C., Kruus, K., Wu, J.H.D. Interactions of the CelS binding ligand with various receptor domains of the *Clostridium thermocellum* cellulosomal scaffolding protein, CipA. J. Bacteriol. 178:1200–1203, 1996.
27. Reverbel-Leroy, C., Pagés, S., Belaich, A., Belaich, J.-P., Tardif, C. The processive endocellulase CelF, a major component of the *Clostridium cellulolyticum* cellulosome: Purification and characterization of the recombinant form. J. Bacteriol. 179:46–52, 1997.
28. Ahsan, M.M., Kimura, T., Karita, S., Sakka, K., Ohmiya, K. Cloning, DNA sequencing, and expression of the gene encoding *Clostridium thermocellum* cellulase CelJ, the largest catalytic component of the cellulosome. J. Bacteriol. 178:5732–5740, 1996.
29. Morag, E., Bayer, E.A., Lamed, R. Affinity digestion for the near-total recovery of purified cellulosome from *Clostridium thermocellum*. Enzyme Microb. Technol. 14:289–292, 1992.
30. Fierobe, H.-P., Gaudin, C., Belaich, A., Loutfi, M., Faure, F., Bagnara, C., Baty, D., Belaich, J.-P. Characterization of endoglucanase A from *Clostridium cellulolyticum*. J. Bacteriol. 173:7956–7962, 1991.
31. Bayer, E.A., Wilchek, M. Protein biotinylation. Methods Enzymol. 184:138–160, 1990.
32. Bayer, E.A., Ben-Hur, H., Wilchek, M. Analysis of proteins and glycoproteins on blots. Methods Enzymol. 184:415–427, 1990.
33. Reverbel-Leroy, C., Belaich, A., Bernadac, A., Gaudin, C., Belaich, J.-P., Tardif, C. Molecular study and overexpression of the *Clostridium cellulolyticum celF* cellulase gene in *Escherichia coli*. Microbiology 142:1013–1023, 1996.
34. Rost, B., Sander, C. Prediction of protein secondary structure at better than 70% accuracy. J. Mol. Biol. 232:584–599, 1993.
35. Rost, B., Sander, C. Combining evolutionary information and neural networks to predict protein secondary structure. Proteins 19:55–72, 1994.
36. Shimon, L.J.W., Bayer, E.A., Morag, E., Lamed, R., Yaron, S., Shoham, Y., Frolow, F. The crystal structure at 2.15 Å resolution of a cohesin domain of the cellulosome from *Clostridium thermocellum*. Structure 5:381–390, 1997.
37. Shimon, L.J.W., Frolow, F., Yaron, S., Lamed, R., Morag, E., Bayer, E.A., Shoham, Y. Crystallization and preliminary X-ray analysis of a cohesin domain of the cellulosome from

Clostridium thermocellum. Acta Crystallogr. D 53:114–115, 1997.
38. Tormo, J., Lamed, R., Chirino, A.J., Morag, E., Bayer, E.A., Shoham, Y., Steitz, T.A. Crystal structure of a bacterial family-III cellulose-binding domain: A general mechanism for attachment to cellulose. EMBO J. 15:5739–5751, 1996.
39. Herzberg, O., James, M.N. Refined crystal structure of troponin C from turkey skeletal muscle at 2.0 Å resolution. J. Mol. Biol. 203:761–779, 1988.
40. Brändén, C.I., Tooze, J. "Introduction to Protein Structure." New York: Garland, 1991.
41. Chauvaux, S., Béguin, P., Aubert, J.-P., Bhat, K.M., Gow, L.A., Wood, T.M., Bairoch, A. Calcium-binding affinity and calcium-enhanced activity of *Clostridium thermocellum* endoglucanase D. Biochem. J. 265:261–265, 1990.
42. Choi, S.K., Ljungdahl, L.G. Structural role of calcium for the organization of the cellulosome of *Clostridium thermocellum*. Biochemistry 35:4906–4910, 1996.
43. Foong, F., Hamamoto, T., Shoseyov, O., Doi, R.H. Nucleotide sequence and characteristics of endoglucanase gene *eng*B from *Clostridium cellulovorans*. J. Gen. Microbiol. 137:1729–1736, 1991.
44. Flint, H.J., Martin, J., McPherson, C.A., Daniel, A.S., Zhang, J.X. A bifunctional enzyme, with separate xylanase and β(1,3-1,4)-glucanase domains, encoded by the xynD gene of *Ruminococcus flavefaciens*. J. Bacteriol. 175:2943–2951, 1993.
45. Zhang, J.X., Martin, J., Flint, H.J. Identification of noncatalytic conserved regions in xylanases encoded by the xynB and xynD genes of the cellulolytic rumen anaerobe *Ruminococcus flavefaciens*. Mol. Gen. Genet. 245:260–264, 1994.
46. Fanutti, C., Ponyi, T., Black, G.W., Hazlewood, G.P., Gilbert, H.J. The conserved noncatalytic 40-residue sequence in cellulases and hemicellulases from anaerobic fungi functions as a protein docking domain. J. Biol. Chem. 270:29314–29322, 1995.