The *Clostridium thermocellum* Cellulosome - the Paradigm of a Multienzyme Complex

Vladimir V. Zverlov[1,2], and Wolfgang H. Schwarz[2*]

[1]) Institute of Molecular Genetics, Russian Academy of Science, Kurchatov Sq., 123182 Moscow, Russia. [2]) Research Group Microbial Biotechnology, Technische Universitaet Muenchen, Am Hochanger 4, D-85350 Freising-Weihenstephan, Germany

The cellulosome is the large extracellular enzyme-complex of the anaerobic bacteria which is responsible for the highly efficient degradation of insoluble cellulose in lignocellulosic biomass. A great number of *Clostridium thermocellum* genes code for putative components of the cellulosome: the unfinished genomic sequence reveals more than 60 genes containing dockerin modules. However, the scaffoldin protein CipA has only nine dockerin binding sites (cohesins), which limits the number of components present in each individual cellulosome particle. Not all genes may be expressed: the components are not distributed equally in the multienzyme complex. Moreover: the *C. thermocellum* cellulosomes are not only composed of cellulases, but also of xylanases, pectinases, glycosidases and structural proteins.

By estimating the quantitative distribution of the proteins in a 2D-electrophoresis gel from a cellulosome preparation, the major components from cellulose-grown cells were identified. Major enzymatic components contain not only the already known endo-, processive endo- and exo-glucanases of the reducing- and non-reducing-end specificity type, but also hitherto unknown ß-glucanases, xylanases, xyloglucanases, and probably also at least one newly detected non-catalytic protein which are now under investigation. The quantitative distribution of the protein pattern present in cellulosomes is described and the major spots in the 2D-gel are identified.

ABUNDANCE OF LIGNOCELLULOSIC BIOMASS AND ENZYMATIC HYDROLYSIS

A huge amount of plant biomass is available in waste or crops: alone in Germany about 76 Mio metric tons per year of lignocellulosic biomass were collected but not useful otherwise in 2001. Much more cold be purposefully collected without additional planting efforts (37, 18). About 30 to 40 % of dry biomass is cellulose which can be hydrolyzed to glucose, the universal substrate for microbial fermentations. Complete hydrolysis can easily be performed at large industrial scale by treatment with sulfuric acid, or with a combination of heat and high pressure (TDH). However, a certain percentage of the sugars are lost due to unwanted chemical reactions. In contrast, enzymatic hydrolysis has a much higher yield of sugars per digested biomass, but is slow and expensive pretreatment is unavoidable. The cost effectiveness of the enzymatic process at a large scale is still to be shown (22). The search for more effective enzymes or enzyme systems goes on. Two methods to reach this goal are used: new genetically engineered enzymes are under development which promise a cost effective enzymatic hydrolysis in the near future. Alternatively new and more effective enzyme systems are screened from natural environments (30).

Native cellulose is a difficult substrate for enzymatic hydrolysis: it is partially crystalline and it is enwrapped with extremely heterogeneous hemicellulose (Fig. 1). This means that a great number of different enzymes is needed for digestion. It was observed that enzymes may support each other in a synergistic action. This synergism is not exerted by successive action: enzymes have to be present simultaneously; the greater the proximity of different enzyme components the greater the synergism (32).

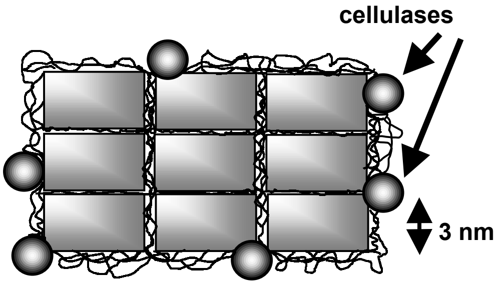

Fig. 1: Hypothetical structure of a cellulose fiber. The squares represent the cross-section of cellulose microfibrils, a number of which is glued together by hemicellulose and lignin (crooked lines). The approximate size of single cellulase enzymes is indicated by the circles.

C. THERMOCELLUM IS A SUPERIOR CELLULOSE DEGRADER

Increased reaction temperature is favorable for biotechnological processes. It was intended to demonstrate the occurrence of thermophilic cellulolytic bacteria in various environments. Soil from grass lands and woods, or self heated compost from cellulosic plant or algae waste from southern Germany and northern France was incubated with various media at 60 and 65 °C. Anaerobic as well as aerobic conditions were used for enrichment with Whatman filter paper as sole carbohydrate. Invariably the anaerobic cultures degraded the filter paper completely with no visible fibers or pieces, whereas in the aerobic cultures the paper was only disintegrated into fibers (Fig. 2). Isolated from one of the enriched anaerobic cultures at 65 °C, the sequence of 10 from 10 PCR amplificates of the small subunit ribosomal DNA (16S rDNA) could be identified as closely related to *Clostridium thermocellum* (> 96 % sequence identity over 1450 bp). Similarly a complete degradation of filter paper was effected by *C. thermocellum* strains DSM 1237 (the type strain) and F7 in anaerobic flasks.

Surprisingly, none of the aerobic cultures did show efficient cellulose degradation at the high temperatures, although the operator of one of the composting plants has detected temperatures up to 70 °C in the aerated compost from which samples were taken. Consequently, the anaerobic bacterium *C. thermocellum* (and close relatives) is an ubiquitous and easily isolated anaerobic thermophilic bacterium which seems to play a major role in anaerobic cellulose degradation in nature.

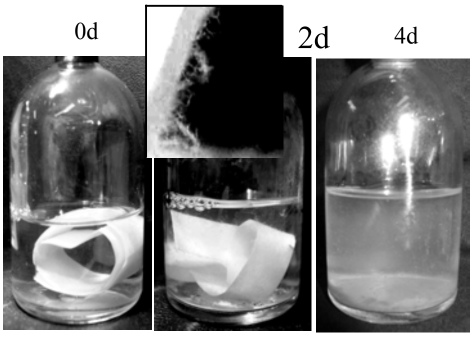

Fig. 2: Digestion of Whatman fiter paper No. 1 in anaerobic cultures. The rubber stoppered bottles containing reduced medium in N_2 atmosphere were inoculated with 1/100 vol. of a diluted soil or compost probe and incubated 0, 2 or 4 days at 60 °C. The insert shows the beginning of paper dissolution.

WHY IS BIOMASS SO STABLE?

Cellulose and hemicellulose are difficult to hydrolyze: although cellulose is chemically homogeneous, its physical structure is at least partially crystalline. This poses a number of problems for enzymes: only a small portion of the substrate is hydratized and exposed to the surface of the crystal in a way, the enzymes have access to it; the number of sites which can be attacked is very limited and different sites may need a different way of attack (Fig. 1). I.e. some enzymes degrade other parts of the substrate than others. Thus, in cellulose a very homogeneous chemistry is combined with a heterogeneous topography and a great number of enzymes is necessary to degrade them – but all enzymes cleave the same chemical bond. In hemicellulose the resistance to enzymatic degradation may rather come from the chemical heterogeneity: different sugars are linked with different chemical bonds. These polymers are often also derivatized with phenolic esters like feruylic acid. Here also a great number of different enzymes is necessary for hydrolysis, this time with activity for different chemical bonds.

Enzymes solve a part of the difficulties on cellulose with synergism. In a very much simplified manner one could say that endoglucanases may set a cut in a cellulose molecule on the surface of the substrate. An exo-glucanase widens the gap either from its reducing or non-reducing end and exposes another layer of the cellulose crystal. This layer could than in turn be opened by an endoglucanase cut and so on (Fig. 3). This picture addresses three important aspects of the synergism: endo- and exo-glucanases work together in degradation; exo- and other exo-glucanases work also together, e.g. in different directions or with a different specificity for cellulose topology; and all activities have to be present simultaneously, because their activity depends on repeated cycles of events involving all enzymes.

The simultaneous presence of many enzymes at a defined site on the substrate surface requires a high local concentration of all components. An organism can reach this goal by secreting a large amount of enzymes – the way aerobic bacteria and fungi do. Anaerobic bacteria are doomed to housekeeping due to the very limited energy supply by fermentation. They use a trick to get a high local enzyme concentration near the surface of the cell and produce enzyme complexes which contain all necessary enzymes, either as multifunctional enzymes with more than one catalytic module or as multienzyme complexes, the so-called cellulosomes (26).

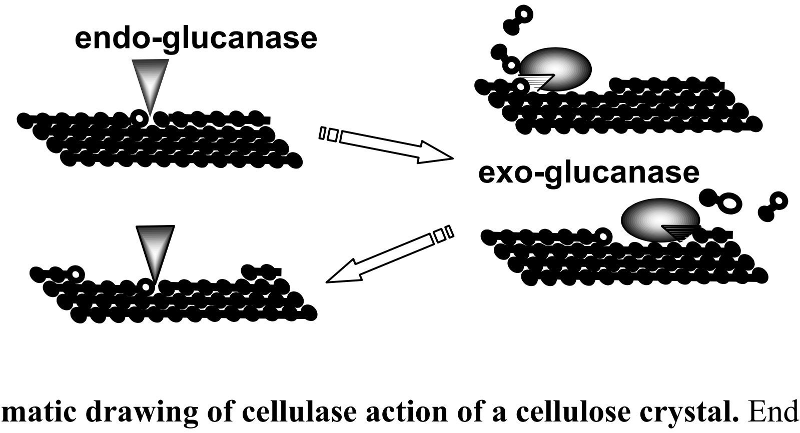

Fig. 3: Schematic drawing of cellulase action of a cellulose crystal. Endo-glucanases and exo-glucanases are represented by triangles and circles respectively. The small black circles are glucose molecules. Reducing ends of sugar chains are indicated by open circles.

THE CELLULOSOME, A MOST EFFICIENT DEPOLYMERIZATION MACHINERY

Evidence accumulated that the enzymatic cellulose degradation machinery of *C. thermocellum* is 50 to 100 times more efficient per gram of protein than the fungal enzymes, which are commercially exploited for cellulose hydrolysis (29). Even if the fungal enzyme system would be improved by a factor of ten, the *C. thermocellum* cellulosomes would be more efficient – and they were not optimized yet. However the production of the cellulosomes by the thermophilic host is presently very low. It could probably be improved by mutant selection or metabolic engineering. The high efficiency of the cellulase system counterbalances well the low energy efficiency of the anaerobic metabolism, which needs a high amount of glucose for little energy spent for the production of extracellular enzymes. In addition, these enzymes are located on the cell surface, minimizing diffusion of enzymes and hydrolysis products, and thus ensuring a high percentage of products taken up by the cells transport mechanism. This on the other hand is a disadvantage for technical enzyme production.

The Israeli group around Lamed and Bayer were the first to characterize the enzyme system of *C. thermocellum* as a huge extracellular enzyme complex which turned out to be arranged along the scaffoldin protein CipA (25,12). It can reach a mass of 1,6 MDa in some strains. This cellulosome is hold together by protein-protein interactions between the cohesin modules located on the scaffoldin and corresponding dockerin modules located mostly at the C-termini of the cellulosome components. The scaffoldin itself is docking to a cell wall anchoring protein (28). It also binds to the substrate with a very tightly binding carbohydrate binding module (CBM) and thus diminishes the necessity of the single enzymatic components to have a CBM by its own. The structure of the cellulosomes from *C. thermocellum* and other bacteria and the interaction between its components is described in recent reviews (4, 33, 35).

It is conspicuous that the production of cellulosomes is restricted to the bacterial family *Clostridiaceae* and the closely related *Lachnospiraceae* (anaerobic rumen bacteria) (34). All other reports of large extracellular enzyme complexes have so far no genetical data for support. The common gene structure, the similarity of the binding modules which hold the complexes together and the major composition of distinct enzyme families suggest heterologous gene transfer between the bacterial species. The very large enzyme complexes of the cellulosomes are an interesting model for the *in vitro* construction of bioengineered protein complexes (9).

GENOMIC LIBRARIES AND STRUCTURE OF CELLULOSOMAL PROTEINS

Libraries of genomic *C. thermocellum* DNA were screened for enzymatic activities known to be involved in biomass hydrolysis, such as ß-1,4-glucanase, cellobiohydrolase, and xylanase. A number of genes were identified with high redundancy (15). 25 of them contained a tandem repeat of a 24 amino acid peptide, called the dockerin module. It could be shown to bind strongly to the cohesin modules of the large scaffoldin CipA, which was identified by immunological screening (11). The *C. thermocellum* scaffoldin contains 9 cohesins type I, but carries itself a dockerin of type II with a different binding specificity to the cohesin of the cell wall anchoring protein OlpB, which has three repeats of a surface layer homologous module (SLH, reviewed in 3, 35). Substrate binding is mediated by a cellulose binding module of family CBM3a near the C-terminus of CipA. Only few of the enzymatic cellulosome components contain an own CBM which is commonly found in non-cellulosomal extracellular cellulases (table 1; Fig. 4).

All enzyme components of the cellulosome are composed of a catalytic and a dockerin module. Some have in addition a carbohydrate binding module or other non-catalytic modules (table 1). The CBM sometimes functions in addition to substrate binding in thermostabilization of the catalytic module. The non-catalytic modules also may modify the action mode of the catalytic module (reviewed in 35). For cellulases only a limited number of glycosyl hydrolase families is known: GH5, GH8 and GH9 for endo-, processive endo- and exo-glucanases, and GH48 for exo-glucanases (7). Examples for their structure are given in Fig. 4. Although family 48 enzymes are the most important exo-glucanases in cellulosomal systems, family 5 and family 9 enzymes are found as well to be exo- as endoglucanases. This was shown e.g. for the structural pairs CbhA/CelD or CelO/CelB (6, 13, 41, 42, 44).

In addition to cellulases (ß-1,4-glucanases) a number of other enzyme specificities have been identified: ß-1,3-1,4-glucanses, xylanases, mannanases and chitinases – not only in the cellulosomes of *C. thermocellum*, but also in that of mesophilic clostridia (table 1)(8).

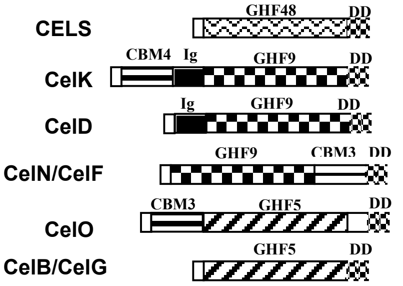

Fig. 4: Structure of selected enzymatic cellulosome components. Typical and repeatedly occurring module architectures are shown. The structure of the proteins is drawn schematically, beginning left with the N-terminus of the polypeptide. The bars are only approximately drawn to scale. GHF, glycosyl hydrolase family; DD, dockerin module; Ig, immunoglobulin-like fold; CBM, carbohydrate binding module.

IDENTIFICATION OF THE PROTEINS IN THE CELLULOSOME

The great number of genes and the limited number of docking sites on the scaffoldin rises the question, if on the average all gene products are expressed and present in the cellulosome. Earlier investigations made clear that there are a few major components which seem to dominate: the scaffoldin CipA, which must be present in each individual cellulosome particle, and the exo-glucanase Cel48S, of which more than 1 copy seems to be incorporated per particle. On the other hand, a chitinase ChiA protein is only present in each 20th cellulosomal particle (43).

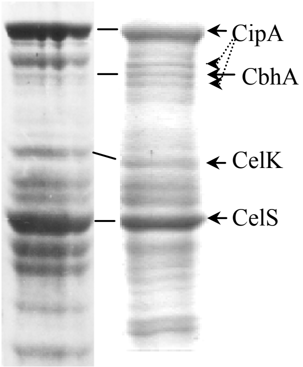

Fig. 5: SDS-PAGE of cellulosome preparations. A fresh (left lane) and a partially degraded cellulosome preparation (right lane) are shown. Bands identified by N-terminal sequencing in the partially degraded preparation (including multiple bands of CipA) and identical bands in the left lane are indicated.

To identify the major components, purified cellulosomes were denaturated by boiling in SDS/ß-mercaptoethanol and separated in denaturing SDS-PAGE. 14 protein bands, designated S1 to S14, were detected and the genes coding for most of them could be identified by several methods (Fig. 5)(25, 33). One method was the immunoblotting with specific antibodies against a recombinant protein devoid of the dockerin module which would give a reaction with all cellulosomal components (e.g. CbhA in 42; CipA in 11). Cross reactions between proteins having a highly homologous sequence can occur (e.g. CbhA/CelK; 42). Another problem was the instability of the cellulosomal components for protease degradation: the proteins have a modular structure and the modules are connected with flexible peptide linkers which are prone to protease attack. An example is shown in Figure 5: CipA fragments are frequently obtained if distinct protein bands were N-terminally sequenced. Particularly many CelA fragments were common in the low molecular weight region of the gel (unpublished data). Thus the number of proteins present in the cellulosome cannot be determined alone by counting protein bands in a gel.

CELLULOSOMAL PROTEINS SEPARATED BY 2D-GEL ELECTROPHORESIS

To reach a higher resolution of the proteins and to resolve common proteins besides rare ones, a 2D gel electrophoresis gel technique was developed, which involved denaturation of the proteins with urea and separation by pI (isolelectric focusing) followed by denaturing SDS-gel PAGE (46). The proteins separated well, most of them between pH 5 and 3 (data not shown). Protein smears due to overload could not be completely avoided in an attempt to identify also minor components.

Single spots were identified by MALDI-TOF and MALDI-TOF-TOF to avoid misinterpretation of spots due to protease action (method according to 36). However, esp. some of the minor spots could not be tested so far. Many of the major spots were identified against the library of sequenced genes. The most prominent proteins were CipA, CelS, CelA, CelK, XynC and XynZ. The locations for CbhA, CelG, CelN and the chitinase ChiA were also determined (46). However, the mass pattern of 3 major spots did not fit to the collection of known sequences of cellulosomal components. Only when the partial genomic sequence of *C. thermocellum* was included in the search (URL http://genome.ornl.gov/microbial/cthe/25jun02/cthe_contig9-15.html), three hitherto un-cloned open reading frames were identified.

The ORFs were amplified by PCR from strain *C. thermocellum* F7 genomic DNA. The resulting recombinant proteins were enzymatically active and could be purified. Preliminary biochemical data on the newly detected enzymes suggest that they code for an endo-glucanase (CelR), an endo-xyloglucanase (XghA) and an endo-xylanase (XynD) belonging to the glycosyl hydrolase families GH9, GH74 and GH10 respectively (Acc.no. AJ585348, AJ585344, AJ585345 resp.). They were among the most prominent spots in the gel when the cells were grown on cellulose.

The 2D-gels show a reproducible map of identified major protein spots and can now be used to look for differential expression of genes on various substrates or culture conditions. Preliminary data show differences in the expression of the cellulosomal genes (see also 16).

A GENOMIC ANALYSIS: MORE THAN 60 CELLULOSOME COMPONENTS?

Motivated by the detection of three new cellulosomal genes, the entire genomic sequence contigs were screened by a BLAST-search for reading frames containing dockerin sequences. This search resulted in a list of potential genes for cellulosome components which are shown in table 1. However, the genomic sequence is unfinished and not closed yet (10th of October, 2003); it contains a number of gaps, incompletely sequenced DNA stretches and overlaps. The screening revealed some reading frames connecting obviously unrelated modules and some sequences of already cloned genes were not found. Nevertheless the great number of 62 reading frames with dockerin modules is overwhelming and some of the presumed enzyme activities would never have been suspected and therefore have not been looked for in the genomic library screenings.

Besides the known genes containing catalytic modules of GH5, GH8, GH9 and GH48, few additional reading frames were found in the families GH5 and GH9. This may be due to the ease of screening for these enzymatic activities by conventional methods, and many research groups have undertaken extended screening programs for these ß-glucanases. However, it is evident that GH8 and GH48 are represented by only 1 gene each among the cellulosomal genes. Both are major components (Cel8A and Cel48S). ß-1,4-Glucanases of other families were not identified.

Surprising is the large number of hemicellulase components potentially present in the cellulosome (table 1). With XynD a 6th xylanase has been identified, 5 of which have been shown to be present in the isolated cellulosomes, 3 as major components. For complete hydrolysis of xylan, the presence of additional glycosidases has to be postulated, none of which was found so far by biochemical analysis or random cloning. But genes for ß-glucuronidase (GH2), and ß-xylosidase / α-arabinofuranosidase (GH39, GH43, GH54) are present and can now be investigated. In addition, genes for carbohydrate-esterases may also be expressed as components of the cellulosome which may split feruloyl residues and other esters of the hemicelluloses. Other hemicelluloses may be degraded by ß-1,3- and ß-1,6-glucanases (GH30, GH81), mannanase (GH26) or endo-ß-1,4-galactanase (GH53). A complete hydrolysis system for pectin including 4 families of pectate lyases may also be present (GH28, PL1, PL9, PL10, PL11).

Some cellulosome components contain more than one functionally related catalytic module like ß-xylosidase and α-arabinofuranosidase (GH54 + GH43), xylanase and carbohydrate-esterase (GH10 + CE1), ß-1,4-glucanase and mannanase (GH5 + GH26), or two different ß-1,4-glucanases (GH9 + GH44). A number of reading frames contains besides the dockerin module unknown modules with no obvious sequence homology in BLAST searches. They may be structural components or silent genes without function. However, one putative structural component has homology to the previously described CseP protein (#55 in table 1) for which a structural role has been expected (45).

Predicted cellulosomal genes with interesting architecture or hitherto not detected probable activities are presently under investigation. Not all of these genes will be expressed or even incorporated into the cellulosome. And some will be expressed possibly only under certain induction conditions, i.e. by growth on certain substrates. Some genes will be poorly expressed and not identified with the 2D gel-electrophoresis. But the number of cellulosomal components seems to be by far greater than estimated so far. However, only a few enzyme components will play a major role in the degradation of crystalline cellulose or of defined biomass substrates. These components have to be identified by biochemistry and gene technology.

The data presented here are a good basis for doing functional genomics with *C. thermocellum*. The methods will be applied for identifying the protein components present in the cellulosome under different conditions of growth on a range of carbohydrate substrate, the number of which is limited due to growth restrictions of the *Clostridium thermocellum strain*, e.g. on xylan or glucose. Probably not all reading frames identified will produce cellulosome components in vivo. But the major components produced on crystalline cellulose as substrate are already identified.

REFERENCES

1) Ahsan, M. M. et al., J. Bacteriol. **178**, 5732-5740 (1996).
2) Arai, T. et al., Appl. Microbiol. Biotechnol. **57**, 660-666 (2001).
3) Bayer, E. A. et al., in M. Claeyssens, et al. (eds.), Carbohydratases from *Trichoderma reesei* and other microorganisms, p. 39-65. The Royal Society of Chemistry, London (1998).
4) Bayer, E. A., et al., in Dworkin, M. et al. (eds.), "The Prokaryotes: An Evolving Electronic Resource for the Microbiological Community", 3rd edition. Springer-Verlag, New York. (2000).
5) Béguin, P. et al., J. Bacteriol. **162**, 102-105 (1985).
6) Chauvaux, S. et al., J. Biol. Chem. **267**, 4472-4478 (1992).
7) Coutinho, P.M. & Henrissat, B., URL: http://afmb.cnrs-mrs.fr/~cazy/CAZY/index.html (1999).
8) Doi, R. H. et al., J. Bacteriol. **185**, 5907-14 (2003).
9) Fierobe, H. P. et al., J. Biol. Chem. **276**, 21257-21261 (2001).
10) Fontes, C. M. G. A. et al., Biochem. J. **307**, 151-158 (1995).
11) Fujino, T. et al., FEMS Microbiol. Let. **94**, 165-170 (1992).
12) Gerngross, U. T. et al., Molec. Microbiol. **8**, 325-334 (1993).
13) Grépinet, O., & Béguin, P., Nuc. Ac. Res. **14**, 1791-1799 (1986).
14) Grepinet, O. et al., J. Bacteriol. **170**, 4582-4588 (1988).
15) Guglielmi, G., & Béguin, P., FEMS Microbiol. Let. **161**, 209-215 (1998).
16) Halliwell, G. et al., Process Biochem. **30**, 243-250 (1995).
17) Halstead, J. R. et al., Microbiol. **145**, 3101-3108 (1999).
18) Hartmann, H. & Kaltschmitt, M. (Eds.): Biomasse als erneuerbarer Energieträger. Schriftenreihe „Nachwachsende Rohstoffe", Vol. 3. Landwirtschaftsverlag, Münster (2002).
19) Hayashi, H. et al., J. Bacteriol. **179**, 4246-4253 (1997).
20) Hayashi, H. et al., Appl. Microbiol. Biotechnol. **51,** 348-357 (1999).
21) Hazlewood, G. P. et al., Enz. Microb. Technol. **12**, 656-662 (1990).
22) Himmel, M. E. et al., Curr. Opin. Biotechnol. **10**, 358-364 (1999).
23) Joliff, G. et al., *Nuc. Acids Res*. **14**, 8605-8613 (1986).
24) Kurokawa, J. et al., Appl. Microbiol. Biotechnol. **59**, 455-461 (2002).
25) Lamed , R. et al., J. Bacteriol. **156**, 828-836 (1983).
26) Lamed, R. et al., J. Bacteriol. **169**, 3792-3800 (1987).
27) Lemaire, M., & Béguin, P., J. Bacteriol. **175**, 3353-3360 (1992)
28) Lemaire, M. et al., J. Bacteriol. **177**, 2451-2459 (1995).
29) Lynd, L. et al., Microbiol. Molec. Biol. Rev. **66**, 506-577 (2002).
30) Montoya, D. et al., J. Ind. Microbiol. Biotechnol. **27**, 329-335 (2001).
31) Navarro, A. et al., Res. Microbiol. **142**, 927-936 (1991).
32) Riedel, K., & Bronnenmeier, K., Molec. Microbiol. **28**, 767-775 (1998).
33) Schwarz, W. H., Appl. Microbiol. Biotechnol. **56**, 634-649 (2001).
34) Schwarz, W. H., URL http://www.wzw.tum.de/mbiotec/cellmo.htm (2003).
35) Schwarz, W. H. et al., Adv. Appl. Microbiol. (in press).
36) Shevchenko, A. et al., Anal. Chem. **68**, 850-858 (1996).
37) van Wyk, J. P. H., Trends Biotechnol. **19**, 172-177 (2001).
38) Wang, W. K. et al., J. Bacteriol. **175**, 1293-1302 (1993).
39) Yaguee, E. et al., Gene **89**, 61-67 (1990).
40) Zverlov, V. V. et al., Biotechnol. Let. **16**, 29-34 (1994).
41) Zverlov, V. V. et al., J. Bacteriol. **180,** 3091-3099 (1998).
42) Zverlov, V. V. et al., Appl. Microbiol. Biotechnol. **51**, 852-859 (1999).

43) Zverlov, V. V. et al., Appl. Environ. Microbiol. **68**, 3176-3179 (2002).
44) Zverlov, V. V. et al., Microbiol. **148**, 247-255 (2002).
45) Zverlov, V. V. et al., Microbiol. **149**, 515-524 (2003).
46) Zverlov, V. V., in preparation.

ACKNOWLEDGMENTS

This work was supported by grants from the Deutsche Forschungsgemeinschaft DFG and the Leonhard-Lorenz-Foundation to WHS, and from the A.-v.-Humboldt Foundation to VVZ. We are very grateful to W. L. Staudenbauer for countless stimulating discussions. D. Minde provided the data on paper hydrolysis under elevated temperature conditions. C. Arevalo is thanked for valuable support with the 2D-gel electrophoresis and J. Kellermann (MPI for Biochemistry, Martinsried) for the performing all the MALDI-TOF determinations.

TABLE 1: List of cellulosomal components in the genome of *C. thermocellum*

Hypothetical reading frames and cloned genes in the unfinished genome sequence. Only reading frames containing dockerin modules, a Shine-Dalgarno sequence and stop codon, and a recognizable module composition are listed. The protein (gene) designation and enzymatic activity or function (if known), and the ORF number from *http://genome.ornl.gov/microbial/cthe/* (Cthe) are given (as of October 2003). If no Cthe number is indicated, the gene was cloned but is missing in the genomic sequence. The components are sorted according to their GH family or their putative function, as obvious from the catalytic module family. Components with more than one catalytic module or unknown modules are listed at the end.

	Gene*	Reading frame / function	Structure **	Reference localisation
		Structural component		
1.	CipA +	scaffoldin, Cthe1933-1930	2(Coh1)-CBM3a-7(Coh1)-X2-Doc2	11, 42
		GH2		
2.		Cthe1580	GH2-CBM6-Doc1	
		GH5		
3.	CelO	cellobiohydrolase, Cthe1674	CBM3b-GH5-Doc1	44
4.		Cthe1575	GH5-CBM6-Fn3–Doc1	
5.	CelB	endoglucanase, Cthe0374	GH5-Doc1	13
6.	CelG +	endoglucanase, Cthe0885	GH5-Doc1	27
7.		Cthe0444	GH5-Doc1	
		GH8		
8.	CelA +	endoglucanase, Cthe0722	GH8-Doc1	5; 46
		GH9		
9.	CbhA +	cellobiohydrolase	CBM4-Ig-GH9-2(Fn3)-CBD3b-Doc1	41
10.	CelK +	cellobiohydrolase, Cthe2598	CBM4-Ig-GH9-Doc1	41
11.	CelD	endoglucanase, Cthe0968	Ig-GH9-Doc1	23
12.		Cthe1953	GH9-CBM3c-CBM3b-Doc1	
13.		Cthe0850	GH9-CBM3c-CBM3b- Doc1	

14.	CelN +	endoglucanase, Cthe1222	GH9-CBM3c-Doc1	46
15.	CelR +	endoglucanase, Cthe1837	GH9-CBM3c–Doc1	46
16.	CelQ +	endoglucanase, Cthe0300	GH9-CBM3c-Doc1	2
17.	CelF	endoglucanase, Cthe0382	GH9-CBM3c-Doc1	31
18.		Cthe1308	GH9-CBM3c–Doc1	
19.		Cthe0727	GH9-Doc1	
20.	CelT +	endoglucanase	GH9-Doc1	24
		Xylanases		
21.	XynD +	xylanase, Cthe0688	CBM22-GH10–Doc1	46
22.	XynC +	xylanase, Cthe0626	CBM22-GH10-Doc1	19
23.	XynA, XynU +	xylanase, Cthe1161	GH11-CBM4-Doc1-NodB	20
24.	XynB, XynV +	xylanase	GH11-CBM4-Doc1	19
		Other hemicellulases		
25.	LicB +	lichenase	GH16-Doc1	40
26.	ChiA +	chitinase	GH18-Doc1	43
27.	ManA +	mannanase, Cthe0533	CBM-GH26-Doc1	17
28.		Cthe2142	GH26-Doc1	
29.		Cthe1127	GH30-CBM6-Doc1	
30.		Cthe2333	GH53-Doc1	
31.		Cthe0269	GH81-Doc1	
		Putative glycosidases		
32.		Cthe1665	GH39-2(CBM6)-Doc1	
33.		Cthe1579	GH43-CBM6-Doc1	
34.		Cthe0268	GH43-CBM13-Doc1	
35.		Cthe0484	GH43-2(CBM6)-Doc1	
		GH48		
36.	CelS +	exoglucanase, Cthe0939	GH48-Doc1	38
		Xyloglucanhydrolase		
37.	XghA +	xyloglucanase, Cthe2335	GH74-CBM2-Doc1	46
		Putative carbohydrate esterases		
38.		Cthe0066	Fn3-CE12-Doc1-CBM6-CE12	
39.		Cthe1577	CE1-CBM6-Doc1	
		Putative pectinases		
40.		Cthe2008	GH28-Doc1	
41.		Cthe2236	PL1-Doc1-CBM6	
42.		Cthe1810	<-Doc1-CBM6-PL9	
43.		Cthe2234	PL10-UN-Doc1	
44.		Cthe0702	Doc1-CBM6-PL11	
		Multifunctional components		
45.	CelJ +	cellulase, Cthe0301	X-Ig-GH9-GH44-Doc1-X	1
46.	CelH	endoglucanase, Cthe0837	GH26-GH5-CBD9-Doc1	39
47.		Cthe1667	GH30-GH54-GH43-Doc1	
48.		Cthe1211	GH54-Doc1-GH43	
49.		Cthe1666	GH54-GH43-Doc1	
50.	XynZ +	xylanase, Cthe1691	CE1-CBM6-Doc1-GH10	14

51.	XynY	xylanase, Cthe2036	CBM22-GH10-CBM22-Doc1-CE1	10
52.	CelE +	endoglucanase, Cthe0940, Cthe2702, Cthe2514	GH5-Doc1-CE2	21
		Putative protease inhibitors		
53.		Cthe1412	Fn3-Doc1-serpin	
54.		Cthe1413	DOC1-SERPIN	
		Components with unknown function		
55.		Cthe0694	2(UN)-UN-UN(CelP 550-870)-Doc1	
56.		Cthe1578	UN-CBM6-Doc1	
57.	CseP +	Cthe1223	UN-Doc1	45
58.		Cthe1474	Doc1-UN	
59.		Cthe0287	UN1-UN2-Doc1	
60.		Cthe0416	Doc1-UN	
61.		Cthe0073	UN-Doc1	
62.		Cthe0649	UN-Doc1	

*A “+” in the reference column indicates, that the component was shown to be present in the cellulosome.
** Module classification according to (7), URL:

http://afmb.cnrs-mrs.fr/CAZY/: Coh, cohesin module; Doc, dockerin module; CBM, carbohydrate binding module; X, hydrophobic module; GH, glycosyl hydrolase family; Fn3, fibronectin III module; Ig, immunoglobulin like fold; NodB, acetylxylan-esterase NodB type; CE, carbohydrate esterase; PL, pectin lyase; UN, unknown module; serpin, serine-protease inhibitor homologue.