The cellulosome concept as an efficient microbial strategy for the degradation of insoluble polysaccharides

Yuval Shoham, Raphael Lamed and Edward A. Bayer

The cellulosome is an extracellular supramolecular machine that can efficiently degrade crystalline cellulosic substrates and associated plant cell wall polysaccharides. The cellulosome arrangement can also promote adhesion to the insoluble substrate, thus providing individual microbial cells with a direct competitive advantage in the utilization of the soluble hydrolysis products.

Y. Shoham is in the Dept of Food Engineering and Biotechnology, Technion-Israel Institute of Technology, Haifa 32000, Israel; R. Lamed is in the Dept of Molecular Microbiology and Biotechnology, George S. Wise Faculty of Life Sciences, Tel Aviv University, Ramat Aviv 69978, Israel; and E.A. Bayer is in the Dept of Biological Chemistry, The Weizmann Institute of Science, Rehovot 76100, Israel.*
**tel: +972 4 829 3072,*
fax: +972 4 832 0742,
e-mail: yshoham@tx.technion.ac.il

In the early part of this century, various isolates of thermophilic clostridia were known to have cellulolytic activity. During the oil crisis of the late 1970s, research efforts in various laboratories were directed to the production of ethanol and other useful chemicals from renewable sources such as cellulose. *Clostridium thermocellum* and related species were chosen for mixed-culture cellulose fermentation in an attempt to develop a stable system for the production of ethanol from cellulose. During these studies, the effect of culture stirring was investigated, leading to the observation that *C. thermocellum* adheres strongly to cellulose before it is degraded.

On the basis of this initial observation, an attempt was made to identify the adherence factor linking the cells to the substrate[1]. By modifying an approach taken in earlier studies on oil-degrading bacteria[2], a mutant was obtained that was deficient in its ability to adsorb to cellulose. This mutant was isolated by an enrichment procedure involving repetitive cycles of growth on cellobiose and the selective removal of cellulose-adhering bacteria. Antibodies were then raised against all surface antigens of wild-type cells and rendered specific for the putative adherence factor by selective adsorption onto the mutant cells. The resultant antibody preparation was specific for a single surface antigen, termed the cellulose-binding factor (CBF; see Box 1 for a glossary of terms used).

Several lines of evidence suggested that the CBF was not a simple adherence factor. The CBF was found to contain 14 identifiable subunits and was produced in large quantities both on the cell surface and in the extracellular medium[3]. The near-identical composition of both forms strongly suggested that the high molecular weight CBF was not a non-specific protein aggregate but a discrete complex. The results of gel filtration, sedimentation velocity and electron microscopy showed that the CBF was a large entity (molecular mass in excess of 2 x 10^6 Da) of relatively uniform size (~18 nm diameter). The discovery of a potent cellulase activity tightly associated with the CBF led to it being renamed the 'cellulosome', to indicate its role in cellulose degradation[4].

During the past 15 years, the cellulosome from *C. thermocellum* and from other species has been studied in several laboratories around the world. These studies have combined many complementary approaches and have increased tremendously our understanding of cellulosome structure and function[5–9].

Multisubunit, multimodular cellulosome structure

In *C. thermocellum*, the cellulosome complex contains many different types of glycosyl hydrolases, including cellulases, hemicellulases and even carbohydrate esterases, all of which are bound to a major polypeptide called scaffoldin (also known as the cellulosome-integrating protein, CipA). The multiple roles of scaffoldin, namely the cellulose-binding and cell-anchoring functions, as well as its role in the organization of the enzyme subunits in the cellulosome complex, were recognized in the early stages of cellulosome research[10]. Similarly, early research also showed that scaffoldin promotes the activity of a cellulosomal enzyme subunit[11]. Scaffoldin contains many functional modules that dictate its various activities. These modules include a single cellulose-binding domain, or CBD, and nine very similar repeating domains, termed cohesins, which interact with the cellulosomal enzymes. The scaffoldin of the *C. thermocellum*

0966-842X/99/$ - see front matter © 1999 Elsevier Science. All rights reserved. PII: S0966-842X(99)01533-4

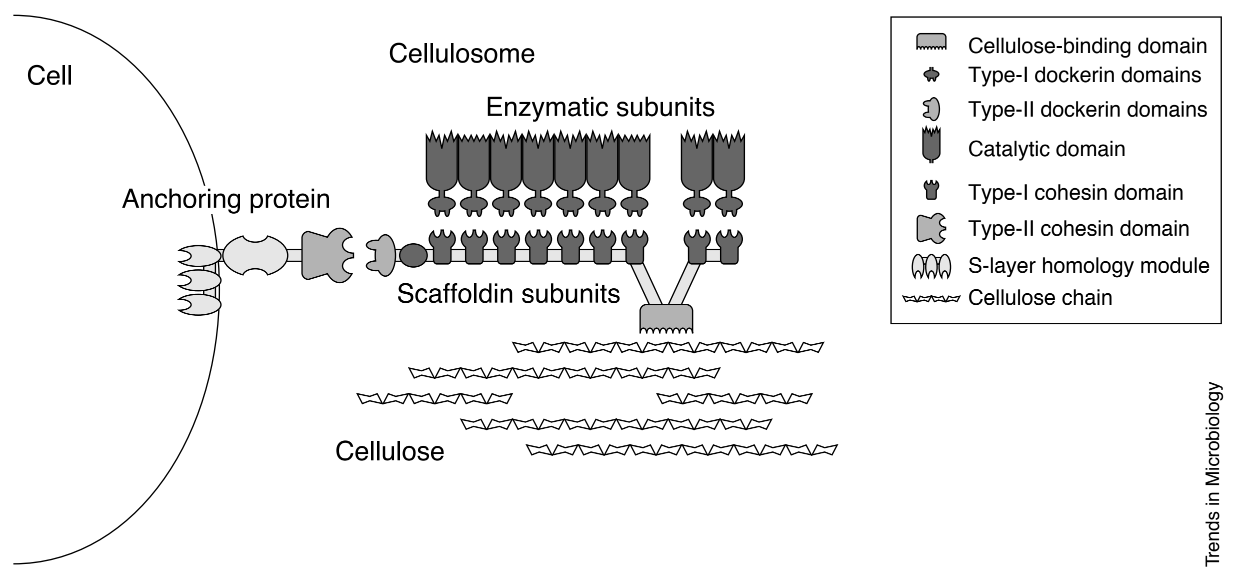


Fig. 1. Simplified schematic view of the interaction between the *Clostridium thermocellum* cellulosome and its substrate, and its connection to the cell surface via an associated anchoring protein.

cellulosome has an additional domain that allows it to attach to the cell surface.

The cellulosomal enzymes are also modular in nature. In addition to a definitive catalytic module, they all possess an additional domain, called a dockerin, that binds tightly with the cohesins of the scaffoldin. The cohesin–dockerin interaction therefore governs the assembly of the complex, while the interaction of the complex with cellulose is mediated by the scaffoldin-borne CBD. The three-dimensional structures of the CBD (Ref. 12) and of two cohesin domains[13,14] from the *C. thermocellum* scaffoldin have been solved. The CBD and cohesin domains have a similar type of fold, but their functional components are clearly different. A schematic view of the cellulosome and its interaction with cellulose and the cell surface is presented in Fig. 1.

The high molecular weight scaffoldin of *C. thermocellum* is highly glycosylated[15] and antigenic. As scaffoldins from different cellulosome-producing species are inherently crossreactive[16], these properties might serve as a tool for identification of new scaffoldins. The sequences of four complete cellulosomal scaffoldin genes have been published[17–20]. All of the known scaffoldins contain the same type of CBD and cohesins, although their number and internal arrangement differ.

Enzymes galore

The enzymes associated with the *C. thermocellum* cellulosome are also relatively large proteins, ranging in molecular mass from 40 to 180 kDa. Each enzyme contains one or more catalytic modules and a single dockerin domain that mediates its interaction with the scaffoldin cohesins. The ~70-residue homologous dockerin domains include a conserved duplicated sequence that resembles the EF-hand motif, which is a conserved helix–loop–helix motif specific for calcium binding, found in proteins such as troponin C and calmodulin[21]. Although the structure of the dockerin domain has yet to be determined, its homology with the EF-hand motif suggests a similar fold, particularly with respect to the calcium-binding loop. In addition, correlation analysis among dockerins of distinct specificities has allowed the identification of putative recognition determinants in the dockerin sequence[22].

The catalytic modules can be grouped, according to sequence similarity and/or general fold, into known glycosyl-hydrolase families and clans[23]. To date, genes encoding 18 different dockerin-containing enzymes have been cloned and sequenced from *C. thermocellum* (Table 1). As expected, many of the enzymes are classical cellulases, including both endo- and exo-acting β-glucanases, enzymes that sever the cellulose chain internally or at one of the ends, respectively. The enzymes most powerful in their action on crystalline substrates appear to be the ‘processive’ cellulases, which cleave the cellulose chain sequentially. Examples of such enzymes in the *C. thermocellum* cellulosome are CelS, CbhA, CelK and CelF.

The CelS subunit appears to be the main catalytic component of the cellulosome. This intriguing processive enzyme is a member of the Family-48 glycosyl hydrolases, and exhibits exocellulolytic, and some endocellulolytic, activity[24,25]. Many of the properties of the intact cellulosome are reflected in those of CelS (Ref. 24). The crystal structure of a related cellulosomal Family-48 enzyme from *C. cellulolyticum* has recently been solved[26]. Another cellulosomal subunit, CelF (Ref. 27), is a Family-9 glycosyl hydrolase that contains a special type of CBD fused to its catalytic site. This type of CBD does not bind to crystalline cellulose *per se*, but appears to bind instead to a single cellulose chain, presumably directing the carbohydrate chain to the active site. The three-dimensional structure of such an enzyme, cellulase E4 from *Theromonospora fusca*, has recently been described[28].

The cellulosomal enzymes are not all cellulases, but include classic xylanases from Families 10 and 11, a Family-26 mannanase, a Family-16 lichenase, and even a Family-18 chitinase. Several of the enzyme subunits carry more than one catalytic module in the same polypeptide, as already discussed. Notably, some of the xylanase subunits also contain carbohydrate esterases able to hydrolyse acetyl or feruloyl groups from the main hemicellulose backbone. Interestingly, *C. thermocellum* can use only cellulose and its degradation products. Hence, the wealth of non-cellulolytic enzymes in the cellulosome apparently allows the removal or detachment of plant cell wall polymers – hemicellulose and lignin – that are in close contact with cellulose.

Table 1. **Cellulosomal subunits of *Clostridium thermocellum*** [a]

Gene product	Description and modular structure[b,c]	No. of residues[d]	Mol. mass (Da)[e]	GenBank Accession No.	Ref.
CipA	Scaffoldin 2(Coh$_{\mathrm{I}}$)–CBD$_{\mathrm{IIa}}$–7(Coh$_{\mathrm{I}}$)–X2–Doc$_{\mathrm{II}}$	1853	196 902	L08665	18
CelJ	Cellulase J X–Ig–**GH9**–**GH44**–Doc$_{\mathrm{I}}$–X	1601	178 382	D83704	50
CbhA	Exoglucanase CBD$_{\mathrm{IV}}$–Ig–**GH9**-2(X1)–CBD$_{\mathrm{III}}$–Doc$_{\mathrm{I}}$	1230	138 078	X80993	51
XynY	Xylanase Y X6–**GH10**–X6–Doc$_{\mathrm{I}}$–**FAE**	1077	119 672	X83269	52
CelH	Endoglucanase H **GH26**–**GH5**–CBD$_{\mathrm{XI}}$–Doc$_{\mathrm{I}}$	900	102 415	M31903	53
CelK	Cellulase K CBD$_{\mathrm{IV}}$–Ig–**GH9**–Doc$_{\mathrm{I}}$	895	100 712	AF039030	NA
XynZ	Xylanase Z **FAE**–CBD$_{\mathrm{VI}}$–Doc$_{\mathrm{I}}$–CBD$_{\mathrm{VI}}$–**GH10**	837	92 262	M22624	54
CelE	Endoglucanase E **GH5**–Doc$_{\mathrm{I}}$–**AXE**	814	90 244	M22759	55
CelS	Cellulase S **GH48**–Doc$_{\mathrm{I}}$	741	83 558	L06942	25
CelF	Endoglucanase F **GH9**–CBD$_{\mathrm{IIIc}}$–Doc$_{\mathrm{I}}$	739	82 088	X60545	27
XynA, XynU	Xylanase A or U **GH11**–CBD$_{\mathrm{VI}}$–Doc$_{\mathrm{I}}$–**NodB**	683	74 511	AB010958 AF047761	NA
CelD	Endoglucanase D Ig–**GH9**–Doc$_{\mathrm{I}}$	649	72 441	X04584	56
XynC	Xylanase C X6–**GH10**–Doc$_{\mathrm{I}}$	619	69 517	D84188	57
CelB	Endoglucanase B **GH5**–Doc$_{\mathrm{I}}$	563	63 929	X03592	58
CelG	Endoglucanase G **GH5**–Doc$_{\mathrm{I}}$	566	63 199	X69390	59
ChiA	Chitinase A **GH18**–Doc$_{\mathrm{I}}$	482	55 028	Z68924	NA
CelA	Endoglucanase A **GH8**–Doc$_{\mathrm{I}}$	477	52 594	K03088	58
XynB, XynV	Xylanase B or V **GH11**–CBD$_{\mathrm{VI}}$–Doc$_{\mathrm{I}}$	457	49 833	AB010958 AF047761	NA
LicB	Lichenase B or Laminarinase 1 **GH16**–Doc$_{\mathrm{I}}$	334	37 897	X63355	60

[a] Modified from Ref. 8.

[b] Abbreviations: AXE, acetyl xylan esterase; CBD$_{\mathrm{III}}$, CBD$_{\mathrm{IIIa}}$ etc., cellulose-binding domain (Families III, IIIa, etc.); Coh$_{\mathrm{I}}$, type-I cohesin domain; Doc$_{\mathrm{I}}$, type-I dockerin domain; Doc$_{\mathrm{II}}$, type-II dockerin; FAE, ferulic acid esterase; GH, glycosyl hydrolase; Ig, immunoglobulin-like domain; Mol. mass, molecular mass; NA, not available; NodB, enzyme activity similar to AXE, but unrelated in sequence; X, other modules or linking segments of unknown function.

[c] Catalytic modules are shown in bold.

[d] Includes signal sequence.

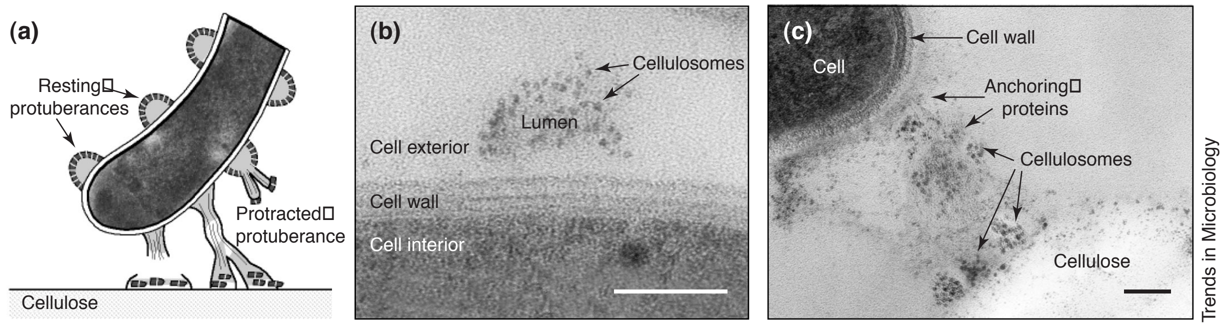
[e] Calculated values are from the peptide sequence.

The *C. thermocellum* cellulosome is cell bound

The arrangement of cellulosomes on the cell surface of *C. thermocellum* was visualized in early research using immunocytochemical labelling and electron microscopy[3,29–31]. The complex is arranged on the cell surface as polycellulosomal protuberance-like organelles (Fig. 2). These protuberances comprise multiple copies of the cellulosome, associated with an interior matrix that contains fibrous material[8]. The protuberances are associated with the cell surface, at intervals, on a layer of exocellular anionic material[30,32]. Upon binding to cellulose, these organelles undergo a dramatic conformational change to form elongated fibres between the substrate and the cell surface. These fibres might direct the soluble products from the insoluble substrate to the cell permeases. The attachment of the cellulosome to the cell surface is mediated by a unique type of cohesin–dockerin interaction. The carboxy-terminus of scaffoldin contains a type-II dockerin that fails to bind to its own type-I cohesins but instead interacts with complementary type-II cohesins of cell-surface anchoring proteins[33,34]. These anchoring proteins also contain an SLH (S-layer homology)[35] module, believed to be associated with the cell surface of Gram-positive bacteria. Thus, the SLH module interacts with the cell surface, and the type-II cohesin, in turn, interacts with scaffoldin via its type-II dockerin, thereby incorporating the cellulosome into the cell surface.

Cellulosome assembly and regulation

Very little is known about cellulosome assembly and what controls the exact composition of each individual complex. All of the cellulosomal components are secreted outside the cell and possess typical leader peptides, which are cleaved during the export process. The complex is assembled extracellularly, probably



Trends in Microbiology

Fig. 2. Ultrastructure of the *Clostridium thermocellum* cell surface. **(a)** Diagrammatic representation of a typical cell bound to cellulose. **(b)** Transmission electron micrograph of a resting polycellulosomal protuberance. **(c)** Transmission electron micrograph of a protracted polycellulosomal protuberance. The cellulosome is mainly associated with the cellulose surface and connected to the cell via extended fibrous material, believed to comprise the anchoring proteins. Scale bars = 100 nm.

in close contact with the cell surface. The number of known dockerin-bearing enzymes in *C. thermocellum* is at least double the number of cohesins in the scaffoldin subunit. A unique interaction between specific cohesin–dockerin pairs is therefore unlikely. In fact, biochemical evidence indicates that the interaction among the cohesins and dockerins within a given species is non specific[36,37]. A possible consequence of this phenomenon is that the composition of the cellulosome is regulated by the relative amounts of the available dockerin-containing polypeptides, which are incorporated randomly into the complex. Individual cellulosome complexes would therefore differ in their exact content and distribution of subunits[38].

The heterogeneous nature of the cellulosome probably affects its overall structure. The flexibility of the many glycosylated linkers, which interconnect the various domains in the scaffoldin and the cellulosomal enzymes, allows multiple degrees of freedom; for this reason, it is unlikely that a precise crystal structure of the entire complex will be forthcoming.

Box 1. Glossary

Cellulose-binding domain (CBD): Domain that mediates the interaction of the cellulosome and its enzyme components with the substrate.
Cellulosomal enzymes: Multimodular enzymes that contain a definitive dockerin domain and one or more catalytic modules.
Cellulosome: A discrete, multienzymatic complex that degrades crystalline cellulosic substrates efficiently.
Cellulosome signature sequences: The presence of dockerin- and/or cohesin-like sequences in a protein.
Cohesin: A functional domain on one molecule that selectively binds to a dockerin domain on another, thereby causing the tenacious association of the two.
Dockerin: The molecular counterpart of the cohesin domain.
Scaffoldin: The cellulosome subunit that integrates the other (enzymatic) subunits into the complex.
Type-I cohesin–dockerin interaction: The interaction between the cohesins on scaffoldin with the dockerins of the enzymatic subunit.
Type-II cohesin–dockerin interaction: The interaction between the carboxy-terminal dockerin of scaffoldin with the cohesin domain(s) of specialized cell-surface anchoring proteins.

Early observations on the cellulosome indicated that the complex might assume different forms. Cellulosomes isolated at early stages of growth appeared compact, whereas during the later stages of cultivation they take on a more relaxed conformation[31]. It is tempting to speculate that the cellulosomal structure could also be influenced by the structure of the substrate it degrades. For example, cellulosic substrates with high hemicellulose content may induce formation of cellulosomes rich in hemicellulolytic enzymes. There are some indications that the cellulosome structure changes upon adsorption to cellulose[39], and models incorporating the spacing between the catalytic groups have been proposed[31].

The expression of many cellulosomal genes in *C. thermocellum* appears to be constitutive and does not involve induction by oligosaccharides derived from cellulose[10]. The highest expression seems to be achieved during carbon-source limitation, presumably by a mechanism analogous to catabolite repression. Little is known about the relative expression of the various cellulosomal genes that, for the most part, are monocistronic and scattered throughout the chromosome of *C. thermocellum*[40]. In contrast, many of the cellulosomal genes in *Clostridium cellulolyticum* are part of a large chromosomal cluster[41].

In *C. thermocellum*, growth on different substrates appears to alter the relative content of the enzymes within the complex[10]. The clearest example of this phenomenon is the amplification of the Family-48 enzyme CelS in the cellulosome during growth of the bacterium on cellulose instead of cellobiose. Transcriptional analysis of the *celA*, *celD* and *celF* genes[42] indicates that the level of transcripts is highest in the early part of the stationary phase, and the transcription starts from two different sites resembling the *Bacillus subtilis* σ^A- and σ^D-like promoters. More research into the regulation of enzyme expression is necessary, not only for *C. thermocellum* but also for other cellulosome-producing bacteria.

Why cellulosomes?

The complex enzymology associated with the degradation of insoluble cellulosic substrates makes it difficult to assess whether the arrangement of plant cell wall degrading enzymes into a cellulosome complex has advantages over free enzyme systems (e.g. that of *Trichoderma reesei*). An early report[43], which compared the extracellular cellulase activity of *C. thermocellum* with that of *T. reesei,* indicated that much less total protein from *C. thermocellum* was required to completely solubilize the crystalline cellulose substrate. Indeed, a recent study showed that the *C. thermocellum* cellulosome is particularly efficient at solubilizing cellulosic substrates of the highest-known crystalline content[44]. This suggests that the specific activity of the cellulosome for such substrates is higher than that of free enzyme systems. It is clear that the organization of enzymes into a cellulosome 'concentrates' them, and perhaps positions them in a suitable orientation both with respect to each other and to the cellulosic substrate, thereby facilitating stronger synergism among the catalytic units. Because of the overall length of the scaffoldin subunit, the cellulosomal enzymes might also have a relatively high degree of flexibility while still attached to the crystalline cellulose, compared with free cellulases, each of which harbours its own CBD.

However, the arrangement of enzymes into a cellulosome could also offer advantages in other respects. In *C. thermocellum*, for example, the cellulosome is attached to the cell surface, localizing the complement of enzymes at the interface between the cell and the insoluble substrate. As it is impossible to maintain equivalent rates of cellulose hydrolysis and cellobiose uptake into cells, hydrolysis is controlled tightly by feedback inhibition. Hence, with the proximity of the cellulosome to the cell, cellobiose would not simply accumulate and dissipate away from the cells, but would be maintained at appropriate concentrations for most efficient use by the cell.

The cellulosome-mediated attachment of the cells to cellulose provides yet another elegant solution to the problem of cell-density-dependent growth[45]. When microorganisms attempt to use high molecular weight polymers, they are forced to produce extracellular enzymes. Free enzymes are soluble and can diffuse away from the cell. Consequently, at very low cell densities, the concentrations of the soluble products might be too low to support growth. However, when the hydrolytic process occurs at the cell–substrate interface, growth on polymers can be initiated by even a single cell, because an adequate concentration of product is maintained. Cellulosomes are found mainly in anaerobic systems where metabolic economy is crucial, and this might indicate that cellulolytic complexes provide a more efficient way to solubilize cellulose.

Conclusions

Despite the overwhelming evidence in favour of the cellulosome concept as a major paradigm for microbial cellulose degradation, many questions still remain unanswered. One important area of heightened research activity is the investigation of the presence of cellulosomes in different bacteria and even fungi. Until recently, the presence of cellulosomes has been confirmed at the genetic level in only four clostridial species. However, numerous cellulosome-related signature sequences have now been described in many other cellulolytic microorganisms (Table 2). The presence of sequences consistent with dockerins and cohesins is considered to be indicative of cellulosomes, and these discoveries support the original biochemical evidence[16] that prompted the notion that cellulosomes are widely distributed among cellulolytic microorganisms. Most of the new publications have reported dockerin-containing enzymes, although a few new scaffoldins (containing type-I cohesins) have also been described. The list of microorganisms in Table 2 reveals that cellulosomes are not limited to anaerobic clostridia, but include anaerobic fungi and even an aerobic bacterium.

Recently, a new type of scaffoldin from *Acetivibrio cellulolyticus* was identified and sequenced (Ref. 46;

Table 2. Evidence for cellulosomes in cellulolytic microorganisms[a]

Organism	Cellulosome signature sequence(s)		Refs
	Protein	Domain[b]	
Anaerobic bacteria			
Clostridium thermocellum	Scaffoldin Surface-anchoring proteins Enzymes	Coh$_{\text{I}}$+CBD+Doc$_{\text{II}}$ Coh$_{\text{II}}$ Doc$_{\text{I}}$	7,18
Clostridium cellulovorans, Clostridium cellulolyticum, Clostridium josui	Scaffoldin Enzymes	Coh$_{\text{I}}$+CBD Doc$_{\text{I}}$	17,19,20
Acetivibrio cellulolyticus	Scaffoldin and surface-anchoring protein	Coh$_{\text{I}}$+CBD+Doc$_{\text{II}}$ Coh$_{\text{II}}$	46
Bacteroides cellulosolvens	Scaffoldin or surface anchoring protein	Coh$_{\text{II}}$+CBD	46
Ruminococcus albus, Ruminococcus flavefaciens	Enzymes	Doc$_{\text{I}}$	61–63
Aerobic bacteria			
Vibrio sp.	Enzyme	Fungal-type dockerin	64
Anaerobic fungi			
Neocallimastix, Piromyces, Orpinomyces	Enzymes	Fungal dockerins	65,66

[a] Modified from Ref. 9.

[b] Abbreviations: CBD, cellulose-binding domain; Coh$_{\text{I}}$, type-I cohesin domain; Coh$_{\text{II}}$, type-II cohesin domain; Doc$_{\text{I}}$, type-I dockerin domain; Doc$_{\text{II}}$, type-II dockerin domain.

Questions for future research

- What is the advantage of the arrangement of cellulolytic and related enzymes into a cellulosome?
- What is the overall structure of the cellulosome complex? What are the dynamics of its action on crystalline substrates?
- How, and where, is the cellulosome assembled?
- How is the expression of cellulosomal genes regulated?
- What is the structure of the cohesin–dockerin complex?
- What are the roles and structures of the accessory domains?
- How diverse are the above phenomena in different microorganisms?

S-Y. Ding, unpublished). Surprisingly, this scaffoldin contains, at its amino-terminus, an enzymatic module homologous to the Family-9 glycosyl hydrolases. In this particular case, the scaffoldin can presumably function as an enzyme, and the presence of multiple cohesins indicates that a full complement of other enzymes is integrated into the complex. Clearly, more scaffoldin sequences are required from different types of organisms to provide further insight into the diversity of cellulosome structure and function in nature.

Finally, the organization of enzymatic components into functionally efficient macromolecular complexes is rapidly becoming a popular subject of scientific research[47]. The terms proteosome, spliceosome, degradosome and signalosome are now well-established. However, the cellulosome remains a paradigm which might prove to be applicable to the degradation of other natural polymeric substrates, such as xylan, chitin, pectin, starch and proteins. Indeed, xylanosomes and amylosomes have already been reported[48,49]. In the future, the cellulosome and the scaffoldin subunit could serve as conceptual templates for the production of tailor-made macromolecular machines, which could be used, for example, in the degradation of unnatural polymers, such as nylon, polyesters and even plastics.

Acknowledgements
This article is dedicated to Professor Arny Demain, in honour of his pioneering contributions to the cellulase field. Our work is supported by grants from the Israel Science Foundation (administered by the Israel Academy of Sciences and Humanities, Jerusalem), and the European Commission (Biotechnology Programme, BIO4-97-2303). Support was also provided by the Technion Otto Meyerhof Centre for Biotechnology, established by the Minerva Foundation, Germany.

References

1 Bayer, E.A., Kenig, R. and Lamed, R. (1983) *J. Bacteriol.* 156, 818–827
2 Bayer, E.A., Rosenberg, E. and Gutnick, D. (1981) *J. Gen. Microbiol.* 127, 295–300
3 Lamed, R., Setter, E. and Bayer, E.A. (1983) *J. Bacteriol.* 156, 828–836
4 Lamed, R. *et al.* (1983) *Biotechnol. Bioeng. Symp.* 13, 163–181
5 Lamed, R. and Bayer, E.A. (1988) *Advan. Appl. Microbiol.* 33, 1–46
6 Felix, C.R. and Ljungdahl, L.G. (1993) *Annu. Rev. Microbiol.* 47, 791–819
7 Béguin, P. and Lemaire, M. (1996) *Crit. Rev. Biochem. Mol. Biol.* 31, 201–236
8 Bayer, E.A. *et al.* (1998) *J. Struct. Biol.* 124, 221–234
9 Bayer, E.A. *et al.* (1998) *Curr. Opin. Struct. Biol.* 8, 548–557
10 Bayer, E.A., Setter, E. and Lamed, R. (1985) *J. Bacteriol.* 163, 552–559
11 Wu, J.H.D., Orme-Johnson, W.H. and Demain, A.L. (1988) *Biochemistry* 27, 1703–1709
12 Tormo, J. *et al.* (1996) *EMBO J.* 15, 5739–5751
13 Shimon, L.J.W. *et al.* (1997) *Structure* 5, 381–390
14 Tavares, G.A., Béguin, P. and Alzari, P.M. (1997) *J. Mol. Biol.* 273, 701–713
15 Gerwig, G. *et al.* (1989) *J. Biol. Chem.* 264, 1027–1035
16 Lamed, R. *et al.* (1987) *J. Bacteriol.* 169, 3792–3800
17 Shoseyov, O. *et al.* (1992) *Proc. Natl. Acad. Sci. U. S. A.* 89, 3483–3487
18 Gerngross, U.T. *et al.* (1993) *Mol. Microbiol.* 8, 325–334
19 Kakiuchi, M. *et al.* (1998) *J. Bacteriol.* 180, 4303–4308
20 Pagès, S. *et al.* (1999) *J. Bacteriol.* 181, 1801–1810
21 Chauvaux, S. *et al.* (1990) *Biochem. J.* 265, 261–265
22 Pagès, S. *et al.* (1997) *Proteins* 29, 517–527
23 Henrissat, B. and Davies, G. (1997) *Curr. Opin. Struct. Biol.* 7, 637–644
24 Morag, E. *et al.* (1991) *J. Bacteriol.* 173, 4155–4162
25 Wang, W.K., Kruus, K. and Wu, J.H.D. (1993) *J. Bacteriol.* 175, 1293–1302
26 Parsiegla, G. *et al.* (1998) *EMBO J.* 17, 5551–5562
27 Navarro, A. *et al.* (1991) *Res. Microbiol.* 142, 927–936
28 Sakon, J. *et al.* (1997) *Nat. Struct. Biol.* 4, 810–818
29 Coughlan, M.P. *et al.* (1985) *Biochem. Biophys. Res. Commun.* 130, 904–909
30 Bayer, E.A. and Lamed, R. (1986) *J. Bacteriol.* 167, 828–836
31 Mayer, F. *et al.* (1987) *Appl. Environ. Microbiol.* 53, 2785–2792
32 Lemaire, M. *et al.* (1998) *Microbiology* 144, 211–217
33 Salamitou, S. *et al.* (1994) *J. Bacteriol.* 176, 2822–2827
34 Leibovitz, E. and Béguin, P. (1996) *J. Bacteriol.* 178, 3077–3084
35 Sleytr, U.B. and Beveridge, T.J. (1999) *Trends Microbiol.* 7, 253–260
36 Yaron, S. *et al.* (1995) *FEBS Lett.* 360, 121–124
37 Lytle, B. *et al.* (1996) *J. Bacteriol.* 178, 1200–1203
38 Pohlschröder, M., Leschine, S.B. and Canale-Parola, E. (1994) *J. Bacteriol.* 176, 70–76
39 Morag, E., Bayer, E.A. and Lamed, R. (1992) *Appl. Biochem. Biotechnol.* 33, 205–217
40 Guglielmi, G. and Béguin, P. (1998) *FEMS Microbiol. Lett.* 161, 209–215
41 Bagnara-Tardif, C. *et al.* (1992) *Gene* 119, 17–28
42 Mishra, S., Béguin, P. and Aubert, J. (1991) *J. Bacteriol.* 173, 80–85
43 Johnson, E.A. *et al.* (1982) *Appl. Environ. Microbiol.* 43, 1125–1132
44 Boisset, C. *et al.* (1999) *Biochem. J.* 340, 829–835
45 Rosenberg, E., Keller, K.H. and Dworkin, M. (1977) *J. Bacteriol.* 129, 770–777
46 Bayer, E.A. *et al.* (1999) in *Genetics, Biochemistry and Ecology of Lignocellulose Degradation* (Ohmiya, K. and Hayashi, K., eds), pp. 428–436, Uni Publishers, Tokyo
47 Alberts, B. (1998) *Cell* 92, 291–294
48 Sunna, A. and Antranikian, G. (1997) *Crit. Rev. Biotechnol.* 17, 39–67
49 Lamed, R. *et al.* (1988) in *Proceedings of the 8th International Biotechnological Symposium* (Vol. 1) (Durand, G., ed.), pp. 371–383, Société Française de Microbiologie, Paris
50 Ahsan, M.M. *et al.* (1996) *J. Bacteriol.* 178, 5732–5740
51 Zverlov, V.V. *et al.* (1998) *J. Bacteriol.* 180, 3091–3099
52 Fontes, C.M. *et al.* (1995) *Biochem. J.* 307, 151–158
53 Yagüe, E., Béguin, P. and Aubert, J-P. (1990) *Gene* 89, 61–67
54 Grépinet, O., Chebrou, M-C. and Béguin, P. (1988) *J. Bacteriol.* 170, 4582–4588
55 Hall, J. *et al.* (1988) *Gene* 69, 29–38
56 Joliff, G., Béguin, P. and Aubert, J-P. (1986) *Nucleic Acids Res.*

14, 8605–8613
57 Hayashi, H. *et al.* (1997) *J. Bacteriol.* 179, 4246–4253
58 Cornet, P. *et al.* (1983) *Biotechnology* 1, 589–594
59 Lemaire, M. and Béguin, P. (1993) *J. Bacteriol.* 175, 3353–3360
60 Zverlov, V.V. *et al.* (1994) *Biotechnol. Lett.* 16, 29–34
61 Kirby, J. *et al.* (1997) *FEMS Microbiol. Lett.* 149, 213–219
62 Ohmiya, K. *et al.* (1997) *Biotechnol. Genet. Eng. Rev.* 14, 365–414
63 Karita, S., Sakka, K. and Ohmiya, K. (1997) in *Rumen Microbes and Digestive Physiology in Ruminants* (Onodera, R. *et al.*, eds), pp. 47–57, Japan Sci. Soc. Press
64 Tamaru, Y. *et al.* (1997) *J. Ferment. Bioeng.* 83, 201–205
65 Fanutti, C. *et al.* (1995) *J. Biol. Chem.* 270, 29314–29322
66 Li, X., Chen, H. and Ljungdahl, L. (1997) *Appl. Environ. Microbiol.* 63, 4721–4728

Variation and evolution of the citric-acid cycle: a genomic perspective

Martijn A. Huynen, Thomas Dandekar and Peer Bork

The presence of genes encoding enzymes involved in the citric-acid cycle has been studied in 19 completely sequenced genomes. In the majority of species, the cycle appears to be incomplete or absent. Several distinct, incomplete cycles reflect adaptations to different environments. Their distribution over the phylogenetic tree hints at precursors in the evolution of the citric-acid cycle.

M.A. Huynen, T. Dandekar and P. Bork are in the European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany, and in the Max-Delbrück-Centre for Molecular Medicine, 13122 Berlin-Buch, Germany;*
M.A. Huynen is also in the Bioinformatics Group, Utrecht University, Padualaan 8, 3584 CH Utrecht, The Netherlands.
**tel: +49 6221 387372,*
fax: +49 6221 387517,
e-mail: huynen@embl-heidelberg.de

Completely sequenced genomes have provided a new way of analysing the biochemical pathways in a species: using the presence of genes encoding the enzymes that catalyse its reactions[1,2]. By studying the variation in metabolic pathways and the way that they are encoded in a rapidly growing set of sequenced genomes, we can elucidate their evolution. Here, we present an investigation of the presence and absence of genes, in prokaryotes and yeast, that code for the enzymes involved in the citric-acid cycle (CAC), including variations such as the reductive CAC and the branched citric-acid pathway, the glyoxylate shunt, and in the reactions connecting the CAC to pyruvate and phosphoenolpyruvate.

Our analysis has combined a thorough examination of sequence data, which included improving the annotation of genes in the GenBank genome database, with an analysis of the biochemical data on the compared species. We examined the genomes of unicellular organisms published to date, including those of four Archaea, 14 Bacteria and one Eukaryote. For an overview of the published genomes, including references, see `http://www.tigr.org/tdb/mdb/mdb.html.`

Variability of the pathway

The genes involved in the CAC and its connections to pyruvate and phosphoenolpyruvate in the various genomes are indicated in Table 1 and a graphical display of the reaction steps for which genes can be found in the selected genomes is given in Fig. 1. The first striking feature in most of the genomes is the incompleteness of the CAC. Only the four largest genomes, those of *Escherichia coli*, *Bacillus subtilis*, *Mycobacterium tuberculosis* and *Saccharomyces cerevisiae*, and the small genome of *Rickettsia prowazekii*, encode the genes for a complete CAC. In the other genomes, the cycle has gaps or is completely absent. In these incomplete cycles, the genes that are present generally code for reactions that are connected to each other, suggesting there are functional connections between the genes. In incomplete cycles, the last part of the oxidative cycle (steps 6–8 in Fig. 1a), leading from succinate to oxaloactetate, is the most highly conserved, whereas the initial steps (steps 1–3), from acetyl CoA to 2-ketoglutarate, show the least conservation.

When interpreting the role of incomplete CACs, it is important to realize that, as well as the oxidation of acetyl CoA, the CAC also plays a role in the generation of intermediates for anabolic pathways. Specifically, 2-ketoglutarate (between steps 3 and 4), oxaloacetate (between steps 8 and 1) and succinyl CoA (between steps 5 and 6) are starting points for the synthesis of glutamate, aspartate and porphyrin, respectively. The autotrophic species that are missing a small part of the CAC are still able to generate 2-ketoglutarate, oxaloacetate and succinyl CoA from

0966-842X/99/$ - see front matter © 1999 Elsevier Science. All rights reserved. PII: S0966-842X(99)01539-5