Overproduction, purification, crystallization and preliminary X-ray characterization of a novel carbohydrate-binding module of endoglucanase Cel5A from *Eubacterium cellulosolvens*

Ana S. Luís,[a] Victor D. Alves,[a] Maria J. Romão,[b] José A. M. Prates,[a] Carlos M. G. A. Fontes[a]* and Shabir Najmudin[a]*

[a]CIISA – Faculdade de Medicina Veterinária, Universidade Técnica de Lisboa, Avenida da Universidade Técnica, 1300-477 Lisboa, Portugal, and [b]REQUIMTE, Departamento de Química, FCT–UNL, 2829-516 Caparica, Portugal

Correspondence e-mail: cafontes@fmv.utl.pt, shabir@fmv.utl.pt

The anaerobic cellulolytic rumen bacterium *Eubacterium cellulosolvens* produces a large array of cellulases and hemicellulases that are responsible for the hydrolysis of plant cell-wall polysaccharides. One of these enzymes, endoglucanase Cel5A, comprises two tandemly repeated novel carbohydrate-binding modules (CBMs) and two catalytic domains belonging to glycoside hydrolase family 5 joined by flexible linker sequences. The novel CBM located at the N-terminus of the endoglucanase has been crystallized. The crystals belonged to the hexagonal space group $P6_122$ and contained a single molecule in the asymmetric unit. The structure of the L-selenomethionine derivative has been solved by a MAD experiment on crystals that diffracted to 1.75 Å resolution.

1. Introduction

In animals, the breakdown of plant feedstocks, particularly cellulose and other recalcitrant carbohydrates, occurs through fermentation by microbes in the rumen or the hindgut (Bryant *et al.*, 1958; Flint & Bayer, 2008; Prins *et al.*, 1972). Recalcitrant polysaccharides found in the plant cell wall pose numerous obstacles to hydrolysis through their development of complex chemical structures and also by becoming inaccessible to biocatalysts owing to the intricacy of their macromolecular structure. Thus, anaerobic bacteria and fungi in the rumen have developed a wide array of multi-modular cellulases and hemicellulases that act individually and as organized cellulosomes (an extremely complex and dynamic extracellular macromolecular assembly) for the hydrolysis of plant cell-wall polysaccharides to soluble sugars (for reviews, see Bayer *et al.*, 2004; Fontes & Gilbert, 2010). Glycoside hydrolases (GH) form one of the most widespread groups of enzymes which hydrolyze the glycosidic bond between polysaccharides, with more than 120 distinct families classified to date (see the CAZy website; http://www.cazy.org; Cantarel *et al.*, 2009). The *Eubacterium cellulosolvens* endoglucanase Cel5A (*Ec*Cel5A) has been sequenced and characterized (Yoda *et al.*, 2005). It is a 1148-amino-acid protein comprising a tandem repeat of an unclassified carbohydrate-binding module (here termed CBM-AL1/2) linked to a glycoside family 5 catalytic module (GH5) and a PT-rich module (PT). Thus, the architectural arrangement of Cel5A is N–CBM-AL1–GH5-1–PT1–CBM-AL2–GH5-2–PT2–C flanked by an N-terminal signal peptide and a C-terminal tail of unknown function. The two CBMs and GH5s show sequence identities of 73 and 94%, respectively, to each other. The enzyme has been shown to have activity towards a variety of cellulosic polysaccharides, including carboxymethyl cellulose, lichenan, acid-swollen cellulose and oat-spelt xylan (Yoda *et al.*, 2005). The catalytic domains of Cel5A display moderate (20–40%) sequence identity to other known GH5s (see CAZy database) and are thus expected to present the $(\beta/\alpha)_8$ TIM-barrel fold common to the GH-A clan (Henrissat, 1998). However, a *BLAST* search (Altschul *et al.*, 1990) showed that there are no structural homologues of CBM-AL1. In fact, to date only two other endoglucanases, one from *Cellulosilyticum ruminicola* (Cai *et al.*, 2010) and one from *Clostridium lentocellum* (Lucas *et al.*, 2010), have been found to possess this CBM module along with a GH5 catalytic
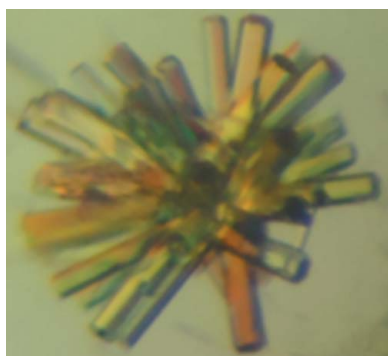
Table 1
Data-collection statistics.

Values in parentheses are for the lowest/highest resolution shells.

Data set	Peak	Inflection point	Remote
Wavelength (Å)	0.9793	0.97934	0.9769
Unit-cell parameters			
$a = b$ (Å)	47.86	48.04	48.21
c (Å)	191.60	191.97	192.34
Resolution limits (Å)	63.87–1.75 (63.87–5.53/1.84–1.75)	63.99–1.75 (63.99–5.53/1.84–1.75)	64.11–1.75 (64.11–5.53/1.84–1.75)
No. of observations	76770 (2783/6703)	77755 (2577/7493)	76501 (2588/6260)
No. of unique observations	13971 (558/1889)	14089 (526/1959)	14094 (530/1836)
Multiplicity	5.5 (5.0/3.5)	5.5 (4.9/3.8)	5.4 (4.9/3.4)
Completeness (%)	99.1 (97.6/94.8)	99.4 (93.2/98.2)	98.5 (92.9/91.6)
$\langle I/\sigma(I)\rangle$	10.4 (22.5/1.7)	9.6 (19.9/1.7)	15.1 (1.3)
R_{merge}†	6.8 (3.7/31.8)	7.5 (3.7/32.0)	7.1 (3.8/31.4)

† $R_{\mathrm{merge}} = \sum_{hkl} \sum_i |I_i(hkl) - \langle I(hkl)\rangle| / \sum_{hkl} \sum_i I_i(hkl)$, where $I_i(hkl)$ is the intensity of the measurement of reflection hkl and $\langle I(hkl)\rangle$ is the mean value of $I_i(hkl)$ for all i measurements.

module. In order to gain insight into the structural properties that govern ligand recognition by this novel CBM, we aimed to determine the crystal structure of *E. cellulosolvens* CBM-AL1. In the present communication, we describe the overproduction, purification, crystallization and preliminary X-ray analysis of the *Ec*Cel5A N-terminal CBM-AL1 module.

2. Materials and methods

2.1. Protein production and purification

*Ec*Cel5A is a modular enzyme containing an N-terminal CBM followed by a GH5 catalytic domain. The two domains are duplicated in tandem and the enzyme contains an additional C-terminal domain of unknown function. The gene encoding the N-terminal CBM of *Ec*Cel5A (residues 37–170) was synthesized (NZYTech Ltd, Portugal) with codon usage optimized for expression in *Escherichia coli*. The synthesized gene contained engineered *Nhe*I and *Xho*I recognition sequences at the 5′ and 3′ ends, respectively, which were used for subsequent subcloning into pET-28a (Novagen), generating pAL1 which encodes CBM-AL1. CBM-AL1 contains an N-terminal His$_6$ tag. *E. coli* Tuner DE3 cells harbouring pAL1 were cultured in Luria–Bertani broth at 310 K to mid-exponential phase ($A_{600\,\mathrm{nm}} = 0.6$) and recombinant protein overproduction was induced

by the addition of 0.2 m*M* isopropyl β-D-1-thiogalactopyranoside and incubation for a further 16 h at 292 K. The His$_6$-tagged recombinant protein was purified from cell-free extracts by immobilized metal-ion affinity chromatography (IMAC) as described previously (Najmudin *et al.*, 2005). Purified CBM-AL1 was buffer-exchanged into 50 m*M* HEPES–HCl pH 7.5, 200 m*M* NaCl, 5 m*M* CaCl$_2$ and then subjected to gel filtration using a HiLoad 16/60 Superdex 75 column (GE Healthcare) at a flow rate of 1 ml min^{-1}. Preparation of *E. coli* to generate selenomethionylated CBM-AL1 was performed as described in Carvalho *et al.* (2004) and the protein was purified using the same procedures as employed for the native CBM. Purified CBM-AL1 (Fig. 1) was concentrated using an Amicon 10 kDa molecular-mass centrifugal concentrator and washed three times with 5 m*M* DTT (for the SeMet protein) or water (for native CBM).

2.2. Crystallization

Crystallization conditions were screened by the hanging-drop vapour-diffusion method using Crystal Screen, Crystal Screen 2 and PEG/Ion Screen (Hampton Research). Drops consisting of 1 µl 80 mg ml^{-1} CBM-AL1 solution and 1 µl reservoir solution were prepared at 292 K. Crystals (maximum dimension of ∼50 µm) grew
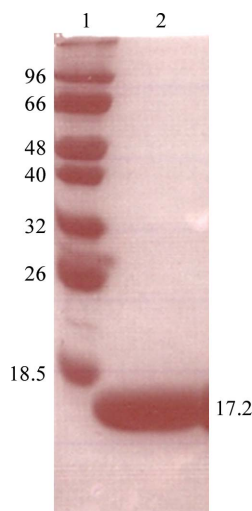

Figure 1
A Coomassie Brilliant Blue-stained 14% PAGE gel evaluation of protein purity. Lane 1, molecular-mass markers (kDa); lane 2, CBM-AL1.
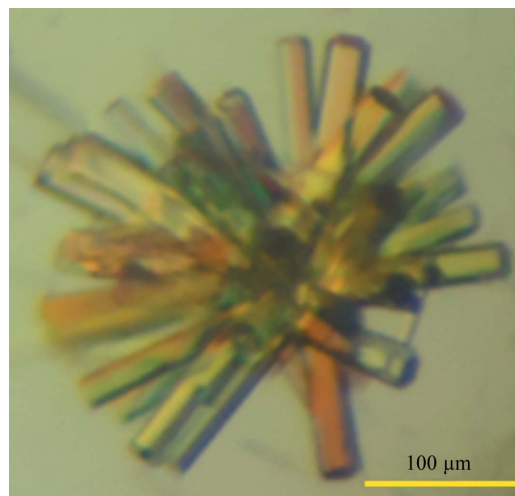

Figure 2
Crystals of CBM-AL1 obtained by hanging-drop vapour diffusion in the presence of 0.23 *M* ammonium sulfate, 0.1 *M* sodium acetate pH 4.6, 26%(*m/v*) PEG 2K MME. The largest crystals are approximately 120 × 30 × 30 µm in size. Crystals were prised apart by gentle prodding with a needle before choosing the best one for data collection.
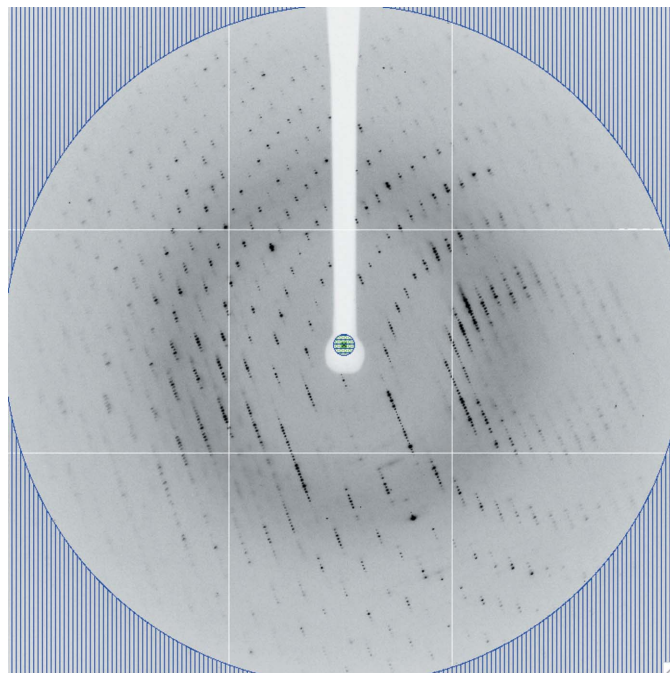
Figure 3
Representative diffraction pattern of an SeMet CBM-AL1 crystal (the outer circle corresponds to 1.75 Å resolution).

within 3 d in the following two conditions: (i) 0.2 M Li$_2$SO$_4$, 25% polyethylene glycol monomethyl ether 2000 (PEG 2K MME) and (ii) 0.2 M ammonium sulfate, 0.1 M sodium acetate pH 4.6, 30%(m/v) PEG 2K MME. Crystals of L-selenomethionine-containing protein were also obtained by vapour diffusion using the hanging-drop method with equal volumes (1 µl) of protein solution (80 mg ml^{-1} in 5 mM DTT) and reservoir solution from a fine screen based around the above two successful conditions for the native crystals. The best crystals (maximum dimension of ∼100 µm) grew over a few days in 0.20–0.23 M ammonium sulfate, 0.1 M sodium acetate pH 4.6, 26%(m/v) PEG 2K MME in clusters (Fig. 2). The crystals were cryocooled in liquid nitrogen after soaking in cryoprotectant [30%(v/v) glycerol added to the crystallization buffer] for a few seconds.

2.3. Data collection and processing

MAD data were collected for selenomethionylated CBM-AL1 on beamline ID23-1 at the ESRF (Grenoble, France) using a Quantum 315r charge-coupled device detector (ADSC, USA) from a crystal cooled to 100 K using a Cryostream (Oxford Cryosystems Ltd.) (Table 1 and Fig. 3). All data sets were processed using the programs *iMOSFLM* (Leslie, 1992) and *SCALA* (Evans, 2006) from the *CCP*4 suite (Collaborative Computational Project, Number 4, 1994). The crystals belonged to the hexagonal system with either the $P6_122$ or $P6_522$ space-group enantiomorph as determined by *POINTLESS* (Evans, 2006). The Matthews coefficient (V_M = 2.11 Å^3 Da^{-1}) suggested the presence of one molecule in the asymmetric unit and a solvent content of 42% (Matthews, 1968). The SeMet crystal diffracted to a resolution of 1.75 Å. Data were collected at 0.97934 Å (inflection point, f' = −9.72, f'' = 3.43), 0.97930 Å (peak, f' = −8.10,

f'' = 5.38) and 0.97560 Å (remote). The crystal was fragile and suffered radiation damage after peak-data collection. Thus, the structure was solved by the SAS protocol of *Auto-Rickshaw* using only the peak data with the EMBL Hamburg automated crystal structure determination platform (Panjikar *et al.*, 2005). The input diffraction data were processed and converted for use in *Auto-Rickshaw* using programs from the *CCP*4 suite (Collaborative Computational Project, Number 4, 1994). F_A values were calculated using the program *SHELXC* (Sheldrick, 2008). Based on an initial analysis of the data, the maximum resolution for substructure determination and initial phase calculation was set to 1.75 Å. Both of the heavy atoms requested were found using the program *SHELXD* (Sheldrick, 2008) and *SHELXE* gave better statistics for the $P6_122$ space group. 94% of the model was built using the program *ARP/wARP* (Perrakis *et al.*, 1999; Morris *et al.*, 2004).

Structure completion and analysis are ongoing.

References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). *J. Mol. Biol.* **215**, 403–410.
Bayer, E. A., Belaich, J. P., Shoham, Y. & Lamed, R. (2004). *Annu. Rev. Microbiol.* **58**, 521–554.
Bryant, M. P., Small, N., Bouma, C. & Robinson, I. M. (1958). *J. Bacteriol.* **76**, 529–537.
Cai, S., Li, J., Hu, F. Z., Zhang, K., Luo, Y., Janto, B., Boissy, R., Ehrlich, G. & Dong, X. (2010). *Appl. Environ. Microbiol.* **76**, 3818–3824.
Cantarel, B. L., Coutinho, P. M., Rancurel, C., Bernard, T., Lombard, V. & Henrissat, B. (2009). *Nucleic Acids Res.* **37**, D233–D238.
Carvalho, A. L., Goyal, A., Prates, J. A., Bolam, D. N., Gilbert, H. J., Pires, V. M., Ferreira, L. M., Planas, A., Romão, M. J. & Fontes, C. M. (2004). *J. Biol. Chem.* **279**, 34785–34793.
Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* D**50**, 760–763.
Evans, P. (2006). *Acta Cryst.* D**62**, 72–82.
Flint, H. J. & Bayer, E. A. (2008). *Ann. N. Y. Acad. Sci.* **1125**, 280–288.
Fontes, C. M. & Gilbert, H. J. (2010). *Annu. Rev. Biochem.* **79**, 655–681.
Henrissat, B. (1998). *Biochem. Soc. Trans.* **2**, 153–156.
Leslie, A. G. W. (1992). *Jnt CCP4/ESF–EACBM Newsl. Protein Crystallogr.* **26**.
Lucas, S., Copeland, A., Lapidus, A., Cheng, J.-F., Bruce, D., Goodwin, L., Pitluck, S., Land, M. L., Hauser, L., Currie, C. & Woyke, T. (2010). US DOE Joint Genome Institute Project, ID 4086514. http://www.jgi.doe.gov.
Matthews, B. W. (1968). *J. Mol. Biol.* **33**, 491–497.
Morris, R. J., Zwart, P. H., Cohen, S., Fernandez, F. J., Kakaris, M., Kirillova, O., Vonrhein, C., Perrakis, A. & Lamzin, V. S. (2004). *J. Synchrotron Rad.* **11**, 56–59.
Najmudin, S., Guerreiro, C. I. P. D., Ferreira, L. M. A., Romão, M. J. C., Fontes, C. M. G. A. & Prates, J. A. M. (2005). *Acta Cryst.* F**61**, 1043–1045.
Panjikar, S., Parthasarathy, V., Lamzin, V. S., Weiss, M. S. & Tucker, P. A. (2005). *Acta Cryst.* D**61**, 449–457.
Perrakis, A., Morris, R. & Lamzin, V. S. (1999). *Nature Struct. Biol.* **6**, 458–463.
Prins, R. A., Van Vugt, F., Hungate, R. E. & Van Vorstenbosch, C. J. (1972). *Antonie Van Leeuwenhoek*, **38**, 153–161.
Sheldrick, G. M. (2008). *Acta Cryst.* A**64**, 112–122.
Yoda, K., Toyoda, A., Mukoyama, Y., Nakamura, Y. & Minato, H. (2005). *J. Appl. Environ. Microbiol.* **71**, 5787–5793.