Biochimica et Biophysica Acta 1804 (2010) 2054–2062

Contents lists available at ScienceDirect

Biochimica et Biophysica Acta

journal homepage: www.elsevier.com/locate/bbapap

Family 42 carbohydrate-binding modules display multiple arabinoxylan-binding interfaces presenting different ligand affinities ☆

Teresa Ribeiro [a,1], Teresa Santos-Silva [b,1], Victor D. Alves [a], Fernando M.V. Dias [a], Ana S. Luís [a], José A.M. Prates [a], Luís M.A. Ferreira [a], Maria J. Romão [b], Carlos M.G.A Fontes [a,*]

[a] CIISA-Faculdade de Medicina Veterinária, Pólo Universitário do Alto da Ajuda, Avenida da Universidade Técnica, 1300-477 Lisboa, Portugal
[b] REQUIMTE/CQFB, Departamento de Química, FCT-UNL, 2829-516 Caparica, Portugal

ARTICLE INFO

Article history:
Received 29 March 2010
Received in revised form 29 June 2010
Accepted 3 July 2010
Available online 14 July 2010

Keywords:
Protein:carbohydrate interactions
Carbohydrate-Binding Module (CBM)
β-trefoil fold
Clostridium thermocellum

ABSTRACT

Enzymes that degrade plant cell wall polysaccharides display a modular architecture comprising a catalytic domain bound to one or more non-catalytic carbohydrate-binding modules (CBMs). CBMs display considerable variation in primary structure and are grouped into 59 sequence-based families organized in the Carbohydrate-Active enZYme (CAZy) database. Here we report the crystal structure of *Ct*CBM42A together with the biochemical characterization of two other members of family 42 CBMs from *Clostridium thermocellum*. *Ct*CBM42A, *Ct*CBM42B and *Ct*CBM42C bind specifically to the arabinose side-chains of arabinoxylans and arabinan, suggesting that various cellulosomal components are targeted to these regions of the plant cell wall. The structure of *Ct*CBM42A displays a beta-trefoil fold, which comprises 3 sub-domains designated as α, β and γ. Each one of the three sub-domains presents a putative carbohydrate-binding pocket where an aspartate residue located in a central position dominates ligand recognition. Intriguingly, the γ sub-domain of *Ct*CBM42A is pivotal for arabinoxylan binding, while the concerted action of β and γ sub-domains of *Ct*CBM42B and *Ct*CBM42C is apparently required for ligand sequestration. Thus, this work reveals that the binding mechanism of CBM42 members is in contrast with that of homologous CBM13s where recognition of complex polysaccharides results from the cooperative action of three protein sub-domains presenting similar affinities.

1. Introduction

Plant cell wall polysaccharides represent the most abundant reservoir of organic carbon within the biosphere. Recycling of photosynthetically fixed carbon through the action of microbial plant cell wall hydrolases is, therefore, a fundamental biological process that has recently acquired considerable industrial importance [1]. Development of second generation bio-fuels derived from lignocellulosic biomass highlights the need to understand the biological processes that result in the production of soluble sugars from plant cell wall structural polysaccharides. It is well established that the complex and intricate nature of plant cell walls restricts the access of enzymes to their target substrates, primarily cellulose and hemicellulose. To overcome their limited accessibility to plant carbohydrates, microbial cellulases and hemicellulases have acquired complex molecular architectures generally comprising catalytic domains and non-catalytic carbohydrate-binding modules (CBMs). The primary role of CBMs is to target the appended catalytic module to the proximity of its substrate, thereby potentiating catalysis and reducing the accessibility constrains [2]. Carbohydrate modifying enzymes and their associated modules, which include CBMs, have been classified into sequence-based families in the CAZy database [3]. Currently there are more than 50 sequence-based families of CBMs (March 2010) which recognize a variety of microbial, plant and mammalian glycans. Based on the topology of the carbohydrate-binding site, which complements the conformation of the target ligand, CBMs have been classified into three types [2]. Thus, in type A modules, which interact with the flat surfaces of crystalline polysaccharides, the binding site comprises a planar hydrophobic platform that contains three exposed aromatic amino acids [4]. These CBMs show no significant affinity for soluble polysaccharides and the ligand specificity of CBM families that contain type A modules is, usually, invariant. In contrast, type B and type C CBMs recognize single carbohydrate chains either internally or at the termini, respectively, and present a ligand specificity that reflects the substrate specificity of the appended catalytic domain [5–7]. Structural studies revealed that type B and C CBMs accommodate their target ligands in clefts or pockets, respectively [2,8,9].

The three-dimensional structure of most CBMs conforms to a β-sandwich fold in which a single ligand-binding site lies in a cleft located on the concave surface of the protein [2]. Ligand plasticity in

☆ Data deposition: Coordinates and observed structure factor amplitudes for the CBM42A have been deposited in the Protein Data Bank with the PDB ID code 3KMV.

* Corresponding author. Tel.: +351 213652876; fax: +351 213652889.
E-mail address: cafontes@fmv.utl.pt (C.M.G.A. Fontes).

[1] Equal contribution.

1570-9639/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.bbapap.2010.07.006

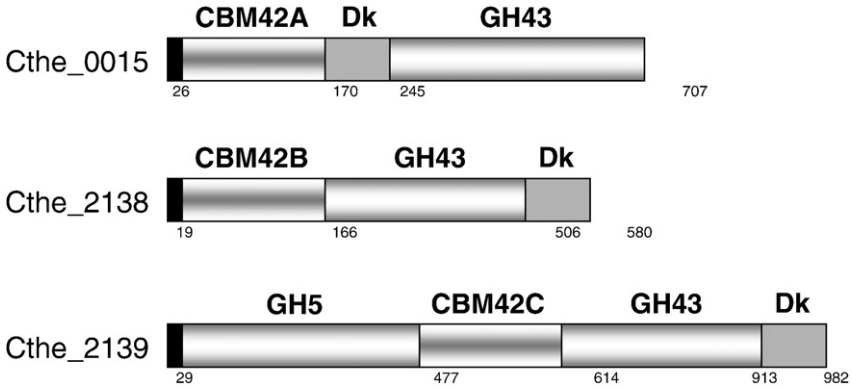

Fig. 1. Modular architecture of *C. thermocellum* enzymes containing CBM42 modules. Abbreviations: Dk, type I dockerin domain; GH5, family 5 glycoside hydrolase domain; GH43, family 43 glycoside hydrolase domain; CBM42, family 42 carbohydrate-binding domain.

CBMs built on a β-sandwich platform usually results from subtle variations at the binding interface that confer capacity to accommodate heterogeneity in the composition and linkage of the sugar backbone *per se* or in the branches that may decorate the carbohydrate polymers [10,11]. In contrast, a variety of CBMs have evolved the capacity to recognize their target ligands at multiple binding sites, as exemplified by members of families CBM13 and CBM42, which assume a β-trefoil fold [12,13]. These modules, which are typical type C CBMs, show a sequential 3-fold internal repeat of ~ 45 amino acid residues comprising three sub-domains, denoted as α, β and γ, each one containing a discrete ligand binding site. Although CBM13 and CBM42 are built on a similar scaffold, the ligand-binding sites of the two structurally related families display different topologies and locations within the protein. However, the three type C binding interfaces of CBM13 and CBM42 have a pocket-like topology, which is particularly suited to recognize small sugars [14,15]. Thus, the fungal family 42 CBM (*Ak*CBM42) of *Aspergillus kawachii* arabinofuranosidase GH54 [12,15], termed *Ak*Abf54 was shown to bind arabinose side-chains of arabinoxylans [12]. Asp435 and Asp488, located in *Ak*CBM42 binding pockets β and γ, respectively, form two pivotal hydrogen bonds with the O-2 and O-3 atoms of an arabinose molecule captured in complex with the CBM and thus play a key role in ligand recognition [12]. In addition, *Ak*CBM42 His416 (pocket β) and His463 (pocket γ) form an additional hydrogen bond with the O-5 atom of the arabinose moiety. Furthermore, an aromatic stacking effect is accomplished by Tyr417–Phe419–Tyr456 triad in pocket β and Tyr464–Tyr359 in pocket γ. The pocket of *Ak*CBM42 sub-domain α was, apparently, non-functional and it was suggested that this is due to the replacement of an aspartate by a glutamate at position 387. Members of family CBM42 were shown to bind specifically to the arabinose side-chain of arabinoxylans, while not interacting directly with the xylan backbone individually [12,16]. In contrast, members of CBM13 found in xylanases were shown to display a higher degree of affinity and specificity for insoluble xylan [13,17]; subtle variations in the ligand-binding sites allow the α, β and γ sub-domains to bind cooperatively three different xylan strands of the insoluble macromolecule. Both CBM13s found in xylanases and CBM42s located in arabinofuranosidases were shown to promote the activity of the appended catalytic domains against insoluble xylans [12,16].

Clostridium thermocellum produces a remarkably complex functional nanomachine, termed the cellulosome, which efficiently degrades plant cell wall polysaccharides [18,19]. Cellulosome assembly results from the interaction of type I dockerin domains, present in cellulosomal cellulases and hemicellulases, and the cohesin domains of a large non-catalytic integrating protein which acts as a molecular scaffold, termed CipA [20,21]. CipA contains a family 3 CBM that binds crystalline cellulose, thus anchoring the enzyme complex onto the plant cell wall [22]. In addition, most cellulosomal enzymes also contain CBMs that bind a variety of carbohydrates, allowing the individual catalytic units to interact with their specific target substrates [18]. *C. thermocellum* proteome presents 72 polypeptides containing type I dockerins, which allowed assigning those proteins as cellulosomal components [23]. Inspection of the primary sequence of those enzymes revealed that proteins with accession numbers Cthe_0015, Cthe_2138 and Cthe_2139 contain family 42 CBMs. The three enzymes are putative GH43 arabinofuranosidases, although their precise role in the function of the associated enzymes remains unknown (see Fig. 1 for molecular architecture of the three proteins). Here we report the structural and biochemical characterization of *C. thermocellum* cellulosomal CBM42s. The structure of one of these proteins was solved and was used to perform a mutagenesis study on the ligand specificity of the various CBM42. The data suggest that cellulosomal CBM42s display a restricted

Table 1
Data collection and refinement statistics.

Crystal		
Space group		$P3_221$
Unit cell parameters (Å) $a=b$, i		106.37, 237.56
Mathews parameter (Å^3/Da)		2.93
Data collection statistics		
X-ray source		ESRF, ID29
Wavelength (Å)		0.976
No. of observed reflections		1414710
No. of unique reflections		144547 (20870)
Resolution limits (Å)		45.22–1.80 (1.90–1.8)
Completeness (%)		99.9 (99.8)
Redundancy		9.8 (9.2)
Multiplicity		9.8 (9.2)
Average $I/\sigma(I)$		19.8 (4.9)
Rsym (%)		0.089 (0.381)
Refinement statistics		
Resolution limits (Å)		45.22–1.80
R-factor (%) (No. of reflections)		0.169 (144547)
R-free (%) (No. of reflections)		0.197 (7239)
No. protein residues in the asymmetric unit		1103
No. water molecules in the asymmetric unit		864
No. atoms in the asymmetric unit		10307
rmsd bond length (Å)		0.013
rmsd bond angles (°)		1.416
Average temperature factor (Å^2)	Protein main chain atoms	7.1
	Protein side chain atoms	9.0
	Water molecules	15.7
	Ca^{2+}	29.6
Ramachandran plot	Residues in most favoured regions (%)	87.6
	Residues in additionally allowed regions (%)	11.6
	Residues in generously allowed regions (%)	0.8
Overall G-factor		0.01