UNIVERSIDADE DE LISBOA
FACULDADE DE MEDICINA VETERINÁRIA

DIVERGENT CELLULOSOME ARCHITECTURE IN RUMEN BACTERIA: STRUCTURE AND FUNCTION STUDIES IN COHESIN-DOCKERIN COMPLEXES OF *RUMINOCOCCUS FLAVEFACIENS*

PEDRO MIGUEL BULE GOMES

Orientadores: Doutor Carlos Mendes Godinho de Andrade Fontes
Doutor Shabir Najmudin

Tese especialmente elaborada para obtenção do grau de Doutor em Ciências Veterinárias na Especialidade de Produção Animal

2017

UNIVERSIDADE DE LISBOA

FACULDADE DE MEDICINA VETERINÁRIA

DIVERGENT CELLULOSOME ARCHITECTURE IN RUMEN BACTERIA: STRUCTURE AND FUNCTION STUDIES IN COHESIN-DOCKERIN COMPLEXES OF *RUMINOCOCCUS FLAVEFACIENS*

PEDRO MIGUEL BULE GOMES

Orientadores: Professor Doutor Carlos Mendes Godinho de Andrade Fontes
Doutor Shabir Najmudin

Tese especialmente elaborada para obtenção do grau de Doutor em Ciências Veterinárias na Especialidade de Produção Animal

Júri:

Presidente: Doutor Rui Manuel de Vasconcelos e Horta Caldeira

Vogais:

- Doutor Harry J. Gilbert
- Doutor Gideon Davies
- Doutor Luís Manuel dos Anjos Ferreira
- Doutor José António Mestre Prates
- Doutor Carlos Mendes Godinho de Andrade Fontes

2017

Para a coautora das minhas aventuras

Joana Dias

Agradecimentos

Ao entrar na reta final de uma das fases mais desafiantes da minha vida, não poderia deixar de aproveitar esta oportunidade para agradecer a todos os que estiveram envolvidos no meu percurso como estudante de doutoramento. Embora as grandes conquistas científicas estejam invariavelmente associadas a um nome, o sucesso científico está longe de ser um esforço individual. Uma equipa que nos apoie, motive e estimule a persistir mesmo nos momentos de maior frustração, é essencial para alcançar os nossos objetivos. Nesse aspeto considero-me bastante afortunado pois o grupo de trabalho a que pertenci ao longo destes anos foi fundamental para o meu sucesso como investigador. Assim, gostaria de agradecer:

Ao meu orientador, Professor Doutor Carlos Fontes, que é o exemplo perfeito de um grande cientista e que irei para sempre ter como referência e modelo. Uma fonte inesgotável de genialidade, pragmatismo e perseverança tornam-no no líder perfeito do grupo de investigação. Muito obrigado pelo seu constante entusiamo, boa disposição, transmissão de conhecimento, amizade e também pela confiança que depositou no meu trabalho, o que tanto me motivou. Ao longo destes anos o Carlos foi um "chefe" que nunca me levou a olhar para o trabalho como uma obrigação mas sim como uma emocionante tarefa;

To my other supervisor, Dr. Shabir Najmudin, for all his teachings and for always pushing me to go further. He has given me the habit of always questioning everything, something that is at the very heart of science. His free and surprising adventurous spirit is inspiring and make him great company. Thank you, Shabir, for becoming a dear friend!;

Ao Professor Doutor Luís Ferreira pela sua incrível simpatia e sentido crítico. O seu entusiasmo e capacidade de transmitir conhecimento em forma de elogios são extraordinariamente motivadores;

Ao Professor Doutor José Prates pelos seus ensinamentos e pela natureza metódica do seu trabalho, a qual me inspirou a ser mais exigente e organizado;

Ao Professor Victor Alves pelos seus ensinamentos e pela excelente colaboração que mantivemos. A excelente forma com que os nossos resultados estão ilustrados devem-se maioritariamente à sua proficiência;

À Helena Santos, a alma do laboratório, pela sua incondicional disponibilidade para ajudar, pela qualidade do seu trabalho, organização e perseverança. Sem a presença da Lena jamais teríamos as condições necessárias para produzir boa ciência. Obrigado também pela sua simpatia e amizade;

À Virgínia Pires, a minha derradeira colega de laboratório, pela sua ajuda, pelos seus ensinamentos, pela sua capacidade de aturar a minha rabugice, por adicionar mais um ou dois protocolos à sua já enorme dose diária e me salvar do desespero da falta de tempo, pelos géis, placas e pipetas que lhe surripiei mas principalmente pela sua amizade;

À Joana Brás e à Teresa Ribeiro por me terem guiado na minha introdução ao mundo da ciência, pela vossa ajuda e simpatia. As vossas dicas preveniram uma série de eventos potencialmente catastróficos e com cheiro a queimado.

To Kate Cameron, my Coh-Doc buddy, for sharing with me the thrills of ITC runs, crystal fishing sessions and trips to the plate robot! Her sense of humor has kept me sane through the most tedious protocols. Thank you for your friendship, Kate!

To Immacolata Venditto for her help and friendship. Her efficiency and organization skills never stopped to amaze me.

À Dona Paula pela sua disponibilidade, bom humor e amizade. É um elemento fundamental do DPASA!

Ao grupo de Biologia Estrutural da Faculdade de Ciências e Tecnologia, com especial destaque para a Professora Ana Luísa Carvalho e a Cecília Bonifácio, pela sua ajuda, disponibilidade, paciência e ensinamentos. Foram fundamentais para a concretização do meu trabalho e também para a minha progressão enquanto investigador;

To Professor Harry Gilbert and all the support he has given me throughout my PhD, especially during my stay at Newcastle University's Institute for Cell and Molecular Biosciences. His input has been invaluable for my work.

To Professor Ed Bayer for being always ready to help and for being such an inspiring scientist with a wonderful personality. Our collaborations have been particularly fruitful!

To Professor Arun Goyal for his teachings, kind nature, hospitality and friendship. Also to all his students in IIT Guwahati who have made me feel at home during my stay at the institute.

To Professor Steven Smith for his fundamental help and contribution that guaranteed the quality of our work.

Ao grupo da NZYtech pela sua enorme ajuda. Com o vosso apoio consegui estabelecer a base sobre o qual todo o trabalho aqui apresentado foi construído.

À Susana Martins, Cristina Monteiro e Inês Viegas pela boa disposição e sentido de humor e por me terem apresentado ao fantástico mundo da parentalidade e todos os cremes a ele associados.

Num percurso repleto de altos e baixos, momentos de êxtase mas também de frustração, foi essencial partilhar essas experiências com aquelas pessoas que para mim são as mais importantes do mundo. Família e amigos são capazes de tornar um problema numa memória distante e uma pequena conquista num grande avanço científico. Mais uma vez, também neste aspeto, me considero uma pessoa sortuda. Quando estamos rodeados das pessoas que adoramos, amamos e admiramos, tudo se torna muito mais simples e é por isso que quero agradecer:

A todos os meus amigos pelos bons momentos que passamos, pela sua disponibilidade para ir beber um copo e ouvir as minhas queixas e por saber que irão estar sempre presentes. Nunca poderia de deixar de vos agradecer. Até porque vocês possuem demasiada informação sensível para que eu arrisque deixar-vos ofendidos... Em particular agradeço ao João Ferreira, Joana

Lobo, Rui Mascarenhas e Mafalda Coelho por me terem ajudado a manter a saúde mental ao longo destes anos.

À minha família, por todo o seu apoio e carinho, com destaque especial para os meus pais e irmã. Mãe e Pai, vocês fizeram de mim aquilo que eu sou, sempre me apoiaram nas minhas decisões e sempre me deram o vosso amor incondicional. Ensinaram-me que a vida não é uma série de problemas mas sim de soluções e que apenas podemos ser respeitados se respeitarmos os outros. São as duas melhores pessoas que eu conheço e não poderia ter tido melhores pais. Mana, obrigado por todas as parvoíces que partilhámos. Até fomos umas crianças felizes! Estou muito orgulhoso que também tu estejas a terminar uma etapa importante na tua vida.

À Joana Dias, a quem dedico esta tese, pela sua ajuda, apoio e amor incondicionais. Obrigado pelos teus conselhos, por me manteres focado e me inspirares, por me aturares nos dias mais rabugentos, por não te importares com jantares fora de hora, por aguentares as ausências mais prolongadas e por manteres a calma numa casa sem roupa lavada e comida no frigorífico. Tenho muita sorte em poder partilhar contigo, não só uma vida, mas também a paixão e o entusiasmo pela ciência: “That’s our thing” e é aquilo que vamos passar a quem aí vier. Obrigado meu Amor!

This work was supported by Fundação para a Ciência e a Tecnologia (Lisbon, Portugal) through grants PTDC/BIA-PRO/103980/2008, EXPL/BIA-MIC/1176/2012 and PTDC/BIA-MIC/5947/2014, by the European Union Seventh Framework Programme (FP7 2007-2013) under the WallTraC project (grant agreement No. 263916) and BioStruct-X (grant agreement No. 283570)

Resumo

Arquitetura celulossomal divergente em bactérias ruminais: estudos de estrutura e função em complexos coesina-doquerina do *Ruminococcus flavefaciens*

As interacções proteína-proteína desempenham um papel essencial em vários processos celulares, sendo exemplo disso a estruturação do celulosoma, um complexo bacteriano multienzimático altamente eficiente na degradação da celulose e hemicelulose. A montagem do celulosoma envolve interações de alta afinidade entre doquerinas do tipo I, presentes em enzimas, e os módulos coesina presentes em proteínas estruturais não catalíticas denominadas de escafoldinas. Adicionalmente, todo o complexo é ancorado à superfície bacteriana através da ligação de uma dockerina do tipo II, presente numa escafoldina, a coesinas ligadas à célula. Inicialmente, pensava-se que a arquitectura e organização dos celulosomas assentava exclusivamente em interacções coesina-doquerina do tipo I e II. Recentemente, foi sugerido que a microbiota ruminal contém bactérias produtoras de celulossoma com diferentes mecanismos de organização, envolvendo um terceiro tipo de complexos coesina-dockerina. O genoma da bactéria ruminal *Ruminococcus flavefaciens* FD-1, revelou um sistema celulossomal particularmente elaborado, montado a partir de uma biblioteca com mais de 200 componentes, através de complexos coesina-doquerina do tipo III. Estabelecer uma base estrutural para a especificidade exibida pelo crescente repertório de pares coesina-doquerina é não só fundamentalmente importante mas também essencial para o desenvolvimento de novas ferramentas com base no celulossoma. O presente trabalho teve como objetivo identificar a base estrutural para a especificidade coesina-doquerina do *R. flavefaciens*, permitindo descortinar os mecanismos por detrás da montagem dos celulosomas ruminais. Os dados obtidos revelaram uma colecção de interacções coesina-doquerina única, suportando a relevância funcional da classificação das doquerinas em grupos com base na homologia da sua estrutura primária. Mostraram ainda que o celulossoma do *R. flavefaciens* é montado através de um mecanismo envolvendo doquerinas com modo de ligação único mas não duplo. Isto contrasta com a maioria dos celulosomas descritos até à data, em que as doquerinas geralmente apresentam duas interfaces semelhantes de ligação à coesina, suportando um modo de ligação dupla. Tal é ilustrado pela estrutura de dois complexos coesina-doquerina do *Acetivibrio cellulolyticus*, envolvendo uma doquerina com modo de ligação dupla. Finalmente, esta informação estrutural foi usada para desenhar uma doquerina com dupla especificidade, mostranto a plasticidade da plataforma coesina-doquerina para o desenvolvimento de novas interações proteína:proteína.

Palavras-chave: celulossoma, coesina, doquerina, complexos proteína-proteína, *Ruminococcus flavefaciens*, *Acetivibrio cellulolyticus*, CAZYmes.

Abstract

Divergent cellulosome architecture in rumen bacteria: structure and function studies in cohesin-dockerin complexes of *Ruminococcus flavefaciens*

Protein-protein interactions play a vital role in many cellular processes as exemplified by the assembly of the cellulosome, a bacterial multi-enzyme complex that efficiently degrades cellulose and hemicellulose. Cellulosome assembly involves the high-affinity binding of type I enzyme-borne dockerins to repeated cohesin modules located on non-catalytic structural proteins termed scaffoldins. In addition, the complex is anchored into the bacterial surface through the binding of a scaffoldin type II dockerin to cell-bound cohesins. Initially, the architecture and organization of cellulosomes was thought to rely uniquely on type I and type II cohesin-dockerin interactions. It was recently suggested that cellulosomes from rumen bacteria are organized through different mechanisms involving a third type of cohesin-dockerin complexes. Thus, the genome of the major ruminal bacterium *Ruminococcus flavefaciens* FD-1 revealed a particularly elaborate cellulosome system that is assembled from a library of more than 200 different components through divergent cohesin-dockerin pairs. Providing structural insights for the specificity displayed by the increasing repertoire of cohesin-dockerin interaction is not only of fundamental importance but essential for the development of novel cellulosome based tools. The present work aimed to identify the molecular basis for the organization of *R. flavefaciens* cellulosome by dissecting the structural basis of cohesin-dockerin specificity in cellulosomes of rumen bacteria. The data revealed a collection of unique cohesin-dockerin interactions, supporting the functional relevance of dockerin classification in groups based on primary sequence similarity. In addition, *R. flavefaciens* cellulosome is assembled through a mechanism involving single but not dual-binding mode dockerins. This contrasts with the majority of the cellulosomes described to date where dockerins generally present two similar cohesin-binding interfaces, supporting a dual-binding mode. To illustrate this, the structures of two cohesin-dockerin complexes containing an *Acetivibrio cellulolyticus* dual-binding mode dockerin were solved. Finally, structural information was used to engineer a dockerin presenting a dual cohesin specificity, revealing the plasticity of the cohesin-dockerin platform to design novel protein-protein interactions.

Key-words: celulossome, cohesin, dockerin, protein-protein complexes, *Ruminococcus flavefaciens*, *Acetivibrio cellulolyticus*, CAZYmes.

This thesis was based on the following manuscripts:

Bule, P., Ruimy-Israeli, V., Cardoso, V., Bayer, E. A., Fontes, C. M. G. A., & Najmudin, S. (2014). Overexpression, crystallization and preliminary X-ray characterization of *Ruminococcus flavefaciens* scaffoldin C cohesin in complex with a dockerin from an uncharacterized CBM-containing protein. *Acta Crystallographica. Section F, Structural Biology Communications*, *70*(Pt 8), 1061–1064. https://doi.org/10.1107/S2053230X14012667

Bule, P., Correia, A., Cameron, K., Alves, V. D., Cardoso, V., Fontes, C. M. G. A., & Najmudin, S. (2014). Overexpression, purification, crystallization and preliminary X-ray characterization of the fourth scaffoldin A cohesin from *Acetivibrio cellulolyticus* in complex with a dockerin from a family 5 glycoside hydrolase. *Acta Crystallographica. Section F, Structural Biology Communications*, *70*(Pt 8), 1065–1067. https://doi.org/10.1107/S2053230X14013181

Bule, P., Alves, V. D., Leitão, A., Ferreira, L. M. A., Bayer, E. A., Smith, S. P., … Fontes, C. M. G. A. (2016). Single-binding mode integration of hemicellulose degrading enzymes via adaptor scaffoldins in *Ruminococcus flavefaciens* cellulosome. *Journal of Biological Chemistry*, jbc.M116.761643. https://doi.org/10.1074/jbc.M116.761643

Israeli-Ruimy, V., Bule, P., Jindou, S., Dassa, B., Moraïs, S., Borovok, I., … Bayer, E. A. (2017). Complexity of the *Ruminococcus flavefaciens* FD-1 cellulosome reflects an expansion of family-related protein-protein interactions. *Scientific Reports*, *7*, 42355. https://doi.org/10.1038/srep42355

Bule, P., Alves, V. D., Israeli-Ruimy, V., Carvalho, A. L., Ferreira, L. M. A., Smith, S. P., … Fontes, C. M. G. A. (2017). Assembly of *Ruminococcus flavefaciens* cellulosome revealed by structures of two cohesin-dockerin complexes. *Scientific Reports*, *7*(1). https://doi.org/10.1038/s41598-017-00919-w

Bule, P., Pires, V., Alves, V. D., Carvalho, A. L., Ferreira, L. M. A., Smith, S. P., Gilbert, H. J., Bayer E.A., Najmudin, S., Fontes, C. M. G. A. Assembly of primary scaffoldin to the cellulosome of *R. flavefaciens* involves a single binding mode dockerin
(manuscript submitted for publication)

Bule, P., Cameron, K., Ferreira, L. M. A., Smith, S. P., Gilbert, H. J., Bayer E.A., Najmudin, S., Fontes, C. M. G. A., Alves, V. D. Designing a dockerin with dual binding specificity based on the structure of the type I cohesin-dockerin complex of *Acetivibrio cellulolyticus* cellulosome
(manuscript submitted for publication)

Index

List of Figures

List of Tables

List of Abbreviations and Symbols

%	Percentage
Å	Angstrom
A. cellulolyticus	*Acetivibrio cellulolyticus*
A_{600}	Absorbance at 600 nanometers
Ala	Alanine (A)
Arg	Arginine (R)
Asn	Asparagine (N)
Asp	Aspartic acid (D)
P. cellulosolvens	*Pseudobacteroides cellulosolvens*
C. acetobutylicum	*Clostridium acetobutylicum*
C. cellulolyticum	*Clostridium cellulolyticum*
C. cellulovorans	*Clostridium cellulovorans*
CE	Carbohydrate esterase
Cel	Cellulase
Cip	Cellulosome integrating protein
C. josui	*Clostridium josui*
C. thermocellum	*Clostridium thermocellum*
Ca^{2+}	Calcium ion
$CaCl_2$	Calcium Chloride
CAZymes	Carbohydrate-active enzymes
CBM	Carbohydrate-binding module
Coh	Cohesin
Coh-Doc	Cohesin-dockerin complex
Cys	Cysteine (C)
Da	Dalton
DNA	Deoxyribonucleic acid
ΔG	Gibbs energy
ΔH	Change in Enthalpy of a system
Doc	Dockerin
E. coli	*Escherichia coli*
ESRF	European Synchrotron Radiation Facility
FPLC	Fast protein liquid chromatography

g	gram
GAX	Glucoronoarabinoxylans
gDNA	Genomic deoxyribonucleic acid
GH	Glycoside hydrolase
Gln	Glutamine (Q)
Glu	Glutamic acid (E)
Gly	Glycine (G)
h	hour
H_2O	Water molecule
HEPES	Hydroxyethyl piperazineethanesulfonic acid
His	Histidine (H)
His_6-tag	Six Histidines tag
Ig	Immunoglobulin
Ile	Isoleucine (I)
IMAC	Immobilised Metal Affinity Chromatography
IPTG	Isopropyl β-D-1-thiogalactopyranoside
ITC	Isothermal Titration Calorimetry
K	kelvin
K_A	Association constant
L	litre
LB	Luria Bertani
Leu	Leucine (L)
Lys	Lysine (K)
LPMO	Lytic Polyssacharide Monooxigenase
LRR	Leucine rich repeat
M	molar
Met	Methionine (M)
min	minute
mol	Molecular replacement
Mw	Molecular weight
NaCl	Sodium Chloride
NF	No Flap
NGE	Native Gel Electropheresis
Ni-NTA	Nickel-Nitrilotriacetilacid
NMR	Nuclear magnetic resonance

ºC	Degrees celcius
ORF	Open Reading Frame
PAGE	Polyacrylamide Gel Electrophoresis
PCR	Polymerase chain reaction
PD-10	Gel filtration columns GE Healthcare
PDB	Protein data bank
PEG	Polyethylene glycol
pH	Negative decimal logarithm of the hydrogen ion activity in a solution
PL	Polyssacharide lyase
Phe	Phenylalanine (F)
Pro	Proline (P)
R. champanellensis	*Ruminococcus champanellensis*
R. flavefaciens	*Ruminococcus flavefaciens*
Rpm	Rotation per minute
SAD	Single wavelength anomalous dispersion
SAXS	Small-angle X-ray scattering
Sca	Scaffoldin
SDS-PAGE	Sodium Dodecyl Sulfate-Polyacrylamide Gel Electrophoresis
SeMet	Seleno-Methionine
Ser	Serine (S)
Ser-Thr	Serine-Threonine pair
SLH	S-layer homology module
SLP	Surface layer protein
Thr	Threonine (T)
Tris	2-Amino-2-hydroxymethyl-propane-1,3-diol
Tyr	Tyrosine (Y)
TrxA	Thioredoxin A
Val	Valine (V)
UKN	Unknown
v/v	Volume per volume
w/w	Weight per weight
Wat	H_2O bridge
ELISA	Enzyme-linked immunosorbent assay
TCEP	Tris (2-carboxyethyl) phosphine hydrochloride
EM	Electron microscopy

HG	Homogalacturonan
RG-1	Rhamnogalacturonan -I
RGII	Rhamnogalacturonan -II
XGA	Xylogalacturonan
AGA	Apiogalacturonan
XyG	Xyloglucans
XDoc	X-module Dockerin
r.m.s.d.	Root mean square deviation
OD	Optical density
WT	Wild type
Xyn	Xylanase
SPL	Surface layer protein
°	Degree

Chapter 1

Bibliographic review and objectives

1.1. Introduction

For quite some time, society has been facing the challenge of replacing hydrocarbons as a primary energy source with cleaner, renewable alternatives. In recent years, a significant amount of resources has been applied to investigate the potential use of lignocellulosic biomass conversion to obtain fermentable sugars that could sustain the production of renewable fuels, such as ethanol. Plant cell walls, predominantly composed of cellulose and hemicellulose, are the most abundant source of biologically utilizable carbon on earth's surface. Photosynthetically fixed carbon is recycled by numerous microbial enzymes that hydrolyse cell wall polysaccharides, therefore playing a crucial role in the carbon cycle, while presenting a significant biotechnological potential. In general, aerobic microorganisms produce copious quantities of plant cell wall degrading enzymes that are secreted to the extracellular media and act individually in the hydrolysis of structural polysaccharides. The released products are then used as a carbon and energy source by the cells. By contrast, the energetic constraints posed by anaerobic ecosystems lead to the evolution of a remarkably highly efficient multi-enzyme complex, termed Cellulosome, which is attached to the microorganism and efficiently degrades a variety of plant cell wall polysaccharides. Anaerobic organisms generally have a lower capacity for protein synthesis and thus, the improved efficiency resulting from enzyme assembly, leads to a higher performance in lignocellulosic biomass degradation. The cellulosome of the Gram-positive thermophilic bacterium *Clostridium thermocellum* is the paradigm for the organization of enzymes into bacterial nanomachines but extensive genetic and genome sequencing studies have allowed the identification of several other species with cellulosomal systems. One of the most interesting cellulosomes produced by the rumen bacterium *Ruminoccoccus flavefaciens*, whose fiber-degrading capacity is predominant over the other ruminal microorganisms. Due to its elaborate architecture and potential industrial

applications it has been the focus of many studies. This chapter introduces and reviews the theme of this thesis. It begins with a general review of plant cell wall composition, with particular focus on cellulose and different polysaccharide constituents, followed by a description of the different mechanisms required for plant cell wall degradation. A short overview of the fiber degrading capabilities of the rumen follows, with special focus on fibrolytic bacteria. Subsequently, cellulosome complexity and functionality will be analysed according to the current knowledge on the structure and function of the different cellulosomal components. A detailed description of the mechanisms of cellulosome assembly will be provided, with a special focus on the cohesin-dockerin interaction, structure, specificity and plasticity. Finally, this chapter will finish with a short description of the current applications of the cellulosome system and with a clear identification of the main objectives of this project.

1.2. Plant cell wall

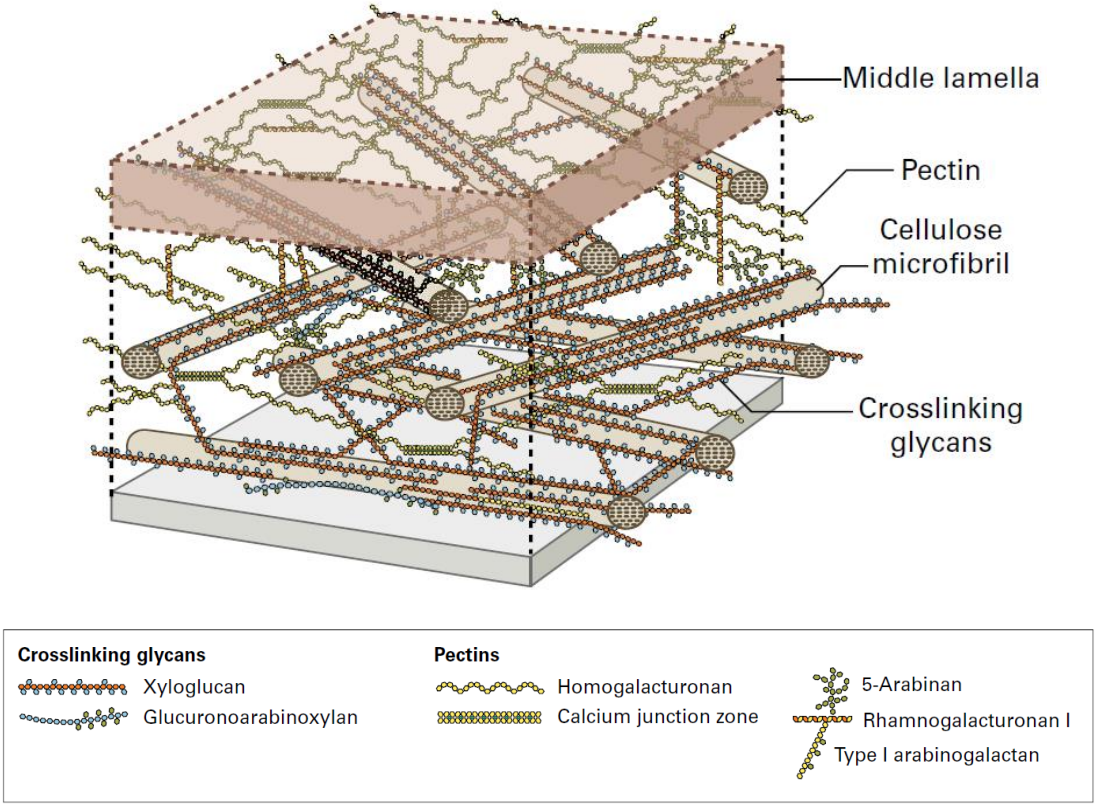
1.2.1. Plant cell wall polysaccharides

Due to the limited mobility of plants, the cell wall is essential for their survival. This highly complex macromolecule enables plants to withstand a variety of harsh environmental conditions and to survive attack by pathogens and herbivores (Caffall & Mohnen, 2009), while acting as a frontier between cells and the outside word. Plant cell walls are highly organized composites of many different polysaccharides, proteins and aromatic substances which contribute to support the plant and its ability to exist in diverse environmental niches (Caffall & Mohnen, 2009; Carpita, Ralph, & McCann, 2015). The cell wall has a well-adapted functional role in growth and development, defining the shape of a plant cell, contributing to structural integrity, cell adhesion and mediating the defence response. Plant cell walls limit the rate and direction of cell growth, exerting a profound influence on plant development and morphology (Carpita *et al.*, 2015), although maintaining a dynamic nature by changing throughout the life of the cell in response to growth, differentiation and environmental stimuli (Caffall & Mohnen, 2009; Scheller & Ulvskov, 2010). In plant cell walls, some structural molecules act as fibres, others as a crosslinked matrix much like the glass fibres and plastic matrix in fiberglass (Carpita *et al.*, 2015). Not all specialized functions of cell walls are structural. Some cell walls contain molecules that affect the pattern of cell development and mark a cell's position within the plant. In addition, plant cell walls contain signalling molecules that participate in cell–cell and wall–nucleus communication. Furthermore, fragments of cell wall polysaccharides may elicit secretion of defence molecules and the wall may become

impregnated with protein and lignin to armour it against invading fungal and bacterial pathogens. Cell wall surface molecules also allow plant cells to recognize their own kind in pollen-style interactions (Carpita *et al.*, 2015).

There is an abundance of research to date which has contributed to the elucidation of the structure and metabolic regulation of various cell wall components (Showalter, 1993). However, relatively little is known about their precise functions and intermolecular interactions, making this an area of scientific importance for extending our knowledge, as well as revealing and exploiting the uses of cell-wall polymers. However, understanding the extreme complexity, versatility and heterogeneity of plant cell walls presents vast technical challenges.

Figure 1.1 Simplified three-dimensional molecular model of the primary cell wall showing the molecular interactions between cellulose, cross-linked glycans (hemicellulose) and pectins.



The orthogonally arranged layers of cellulose microfibrils (brown) are hydrogen bonded with a network of cross-linking glycans (red and blue). This network coexists with a network of pectic polysaccharides (yellow). The cellulose-hemicellulose network provides tensile strength while the pectin network resists compression. The middle lamella is rich in pectin and cements adjacent cells together. Adapted from Carpita *et al.*, 2015.

Plants have two cell wall types with different functions and composition that are termed the primary and the secondary cell walls. Cell wall structure is continuously modified to accommodate the developmental stage and the environmental condition. The primary wall is

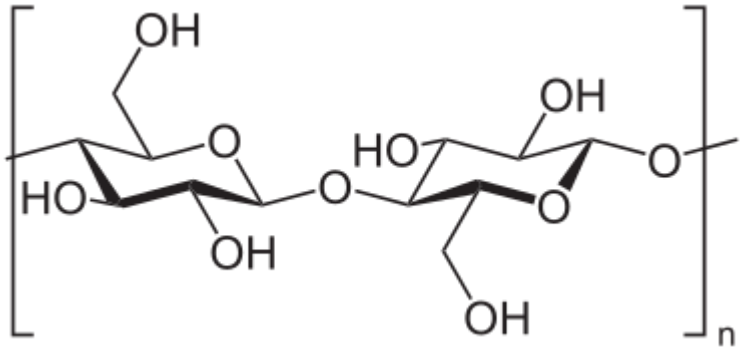
thought to contribute to wall structure integrity, cell adhesion and signal transduction. Primary cell walls (Figure 1.1) are generally composed of a thin, flexible and extensible layer formed while the cell is growing and is composed of a network of cellulose microfibrils, non-cellulosic polysaccharides and glycoproteins, which co-exist with a network of pectic polysaccharides such as homogalacturonan (HG), rhamnogalacturonan I (RG-I) and rhamnogalacturonan II (RG-II) that contribute to cell strength, cell adhesion, stomatal function and defence response (Popper *et al.*, 2011).

Many cells, particularly in higher plants, have a distinct secondary cell wall located between the primary cell wall and the plasma membrane, which is deposited at later stages of growth when cells assume their final stages of differentiation (Carpita *et al.*, 2015; Keegstra, 2010). Secondary cell walls provide additional protection and rigidity to the larger plant. It consists primarily of layered sheaths of cellulose, which are parallel within each layer, along with other polysaccharides, lignin and glycoproteins. Lignin, a phenolic polymer, cements and anchors the cellulose microfibrils among other matrix polysaccharides. The presence of lignin makes this wall less flexible and less permeable than the primary cell wall, stiffening the walls and thus preventing biochemical degradation and physical damage (Popper, 2008).

1.2.1.1. Cellulose

Cellulose is the most abundant plant polysaccharide, accounting for 15–30% of the dry mass of all primary cell walls and a much larger percentage of secondary walls (Brown, 2004; Carpita *et al.*, 2015). It is a chemically simple molecule that exists in the form of microfibrils, which are paracrystalline assemblies of several dozen β-1,4-linked D-glucan chains that interact with one another along their length via hydrogen bonds (Somerville, 2006). The β-1,4 link means that each glucose will be rotated 180º relatively to the adjacent molecule, making cellobiose (glucose dimer) the actual repeating unit in cellulose (Figure 1.2) (Béguin & Aubert, 1994).

Figure 1.2 Chemical structure of cellulose with the cellobiose unit highlighted between brackets

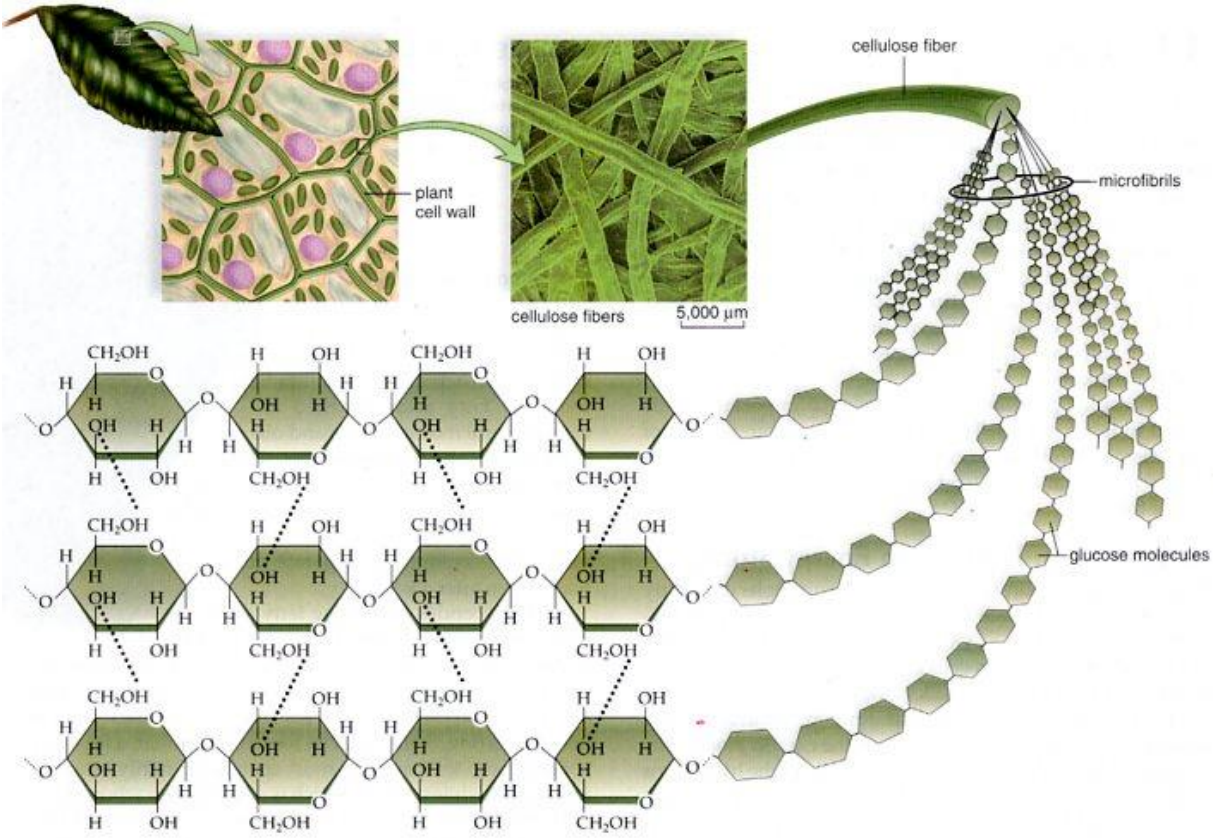


Each *β*-1,4-D-glucan chain may comprise several thousand units (≈ 2–3 μm long), but individual chains begin and end at different places within the microfibril to allow it to reach lengths of hundreds of micrometres and contain thousands of individual glucan chains. The microfibrils are then assembled into superstructures (Figure 1.3), such as cell walls, fibres, pellicles and so on (Bayer, Chanzy, Lamed, & Shoham, 1998).

Natural crystalline cellulose is named cellulose I, or native cellulose, and comprises the two forms I*α* and I*β*, in which these chains lie parallel (Jamal, Nurizzo, Boraston, & Davies, 2004). The I*α* form consists of a single-chain triclinic structure, whereas the I*β* form is monoclinic and is characterized by two parallel chains. The density and stability of the I*α* form was shown to be lower and its enzymatic or chemical reactivity is therefore higher (Bayer, Chanzy, *et al.*, 1998). Many non-natural forms of cellulose and crystalline arrays of cello-oligosaccharides form cellulose II in which the chains lie anti-parallel. This second most extensively studied form of the carbohydrate may be obtained from cellulose I by either of two processes: regeneration, which is the solubilisation of cellulose I in a solvent, followed by re-precipitation in water, or by mercerization, which is the process of swelling native fibres in concentrated sodium hydroxide, to yield cellulose II on removal of the swelling agent (O'Sullivan, 1997). In addition, many model sources of natural crystalline cellulose such as microcrystalline cellulose like Avicel™, bacterial microcrystalline cellulose, and tunicate cellulose appear to contain various proportions of unstructured cellulose, rather loosely termed "amorphous" cellulose. Enzymologists studying cellulose hydrolysis have long adopted a "binary" model of cellulose structure featuring "crystalline" and amorphous regions (Jamal *et al.*, 2004), although the amorphous regions still possess a degree of order (O'Sullivan, 1997). Cellulose is crystalline when molecules are tightly packed and is amorphous when they are loosely packed. The crystalline areas are more insoluble and inaccessible to enzymatic attack than the amorphous areas, making the hydrolysis of these regions more complex and difficult (Warren, 1996). Most of the "amorphous phase" of cellulose corresponds to chains that are located at the microfibril

surface, whereas crystalline components occupy the core (Bayer, Chanzy, *et al.*, 1998). With respect to cellulose biosynthesis, it is also known that it takes place at the non-reducing end of the growing chain (Bayer, Chanzy, *et al.*, 1998).

Figure 1.3 Arragements of fiber, microfibrils and cellulose chains in plant cell walls. Adapted from (Mader, Windelspecht, & Cognato, 2013)



1.2.1.2. Cross-linking glycans (Hemicellulose)

Crosslinking glycans are a class of highly branched polysaccharides that can hydrogen bond to cellulose microfibrils; they may coat microfibrils, but are also long enough to span the distance between microfibrils and linking them together to form a network (Carpita *et al.*, 2015). These crosslinks are responsible for the formation of a tough network, which is responsible for the mechanical strength of plant cell walls (Cooper & Hausman, 2009). Crosslinking glycans are often called "hemicelluloses," a widely-used term for all materials extracted from the cell wall with alkaline treatment, regardless of structure. Hemicelluloses are structurally homologous to cellulose but contain β-linked backbones decorated with a variety of sugars and acetyl groups, explaining why these polymers are not crystalline (Gilbert, 2010). The detailed structure of hemicelluloses and their abundance varies between different species and cell types. Xyloglucan, xylan, arabinoxylan, mannan and glucomannan are examples of these hemicellulosic

polysaccharides. Xyloglucan, xylan and arabinoxylan have a backbone composed of β-1,4 linked D-pyranosyl residues with O4 in equatorial orientation (Cosgrove, 2005).

Xyloglucan is the most abundant hemicellulosic polysaccharide in the plant cell wall of non-grasses and has a semirigid backbone of β-1,4-glucan that is decorated with two or three contiguous α-D-Xylose units linked to the O-6 position of the glucan units between unbranched glucan residues. Certain xylose units are substituted further with other monosaccharides, typically α-l-Arabinose or β-D-Galactose (Gal), to improve water solubility and the possible recognition by wall-modifying enzymes. In addition, Gal residues may be further substituted with α-l-Fucose (Carpita *et al.*, 2015; Cosgrove, 2005). Xylans are β-1,4 linked xylopyranose polymers that form twisted ribbons. Different xylans are variously substituted with acetyl, arabinofuranosyl and glucuronosyl residues (Warren, 1996). Xylans dominated with substitution with glucuronosyl residues are often known as glucuronoxylans (Scheller & Ulvskov, 2010). Arabinoxylan consists of a β-1,4 linked D-xylan backbone decorated with arabinose branches. Glucuronic acid and ferulic acid esters are other residues that can be attached to arabinoxylans, particularly in cereal grasses (Brett & Waldron, 1996). Mannan and glucomannan are also important components of plant biomass. Mannans, are relatively flexible and consist of a backbone of β-1,4 linked mannose residues, whereas glucomannan comprises a heterogeneous polymer of β-1,4 linked glucose and mannose sugars, randomly distributed. The backbone of both mannan and glucomannan can be decorated with α-1,6 linked galactosyl residues and thus these polysaccharides are often referred to as galactomannan and galactoglucomannan, respectively (Brett & Waldron, 1996).

1.2.1.3. Pectin

Pectins, the most soluble of the cell wall polysaccharides, comprise a mixture of heterogeneous, branched and highly hydrated polysaccharides rich in acidic sugars, such as glucuronic acid and galacturonic acid (Carpita *et al.*, 2015; Cosgrove, 1997). They include homogalacturonan (HG), xylogalacturonan (XGA), apiogalacturonan (AGA), rhamnogalacturonan I (RG-I) and rhamnogalacturonan II (RG-II). Pectins are structurally and functionally the most complex group of polysaccharides in plant cell walls. The intricate structure of pectic polysaccharides and the retention by plants of the large number of genes required to synthesise pectin, suggests their significant role in plant growth and development (Ridley, O'Neill, & Mohnen, 2001). Pectins perform many functions, such as determining wall porosity and providing charged surfaces that modulate wall pH and ion balance, regulating cell–cell adhesion at the middle lamella and serving as recognition molecules that alert plant cells to the presence of symbiotic

organisms, pathogens and insects. Particular cell wall enzymes may bind to the charged pectin network, constraining their activities to local regions of the wall. By limiting wall porosity, pectins may affect cell growth, thereby regulating access of wall-loosening enzymes to their glycan substrates. Pectins physical properties have been extensively explored in a number of technological and industrial applications such as a gelling and stabilizing polymer in diverse food and speciality products with positive effects on human health and multiple biomedical uses (Mohnen, 2008). HG, the simplest of these polymers, comprises a linear chain of α-1,4 D-galacturonic acid residues that can account for more than 60% of pectins in the plant cell wall. The backbone of HG is covalently linked to RG-I and RG-II, and is also hypothesised to be covalently cross-linked to the highly abundant hemicellulose xyloglucan (XG) *in muro* (Caffall & Mohnen, 2009). RG-I consists of alternative residues of α-1,4 D-galacturonic acid and α-1,2 L-rhamnose, decorated primarily with arabinan and galactan side chains. It has been suggested that RG-I functions as a scaffold to which other pectins, such as RG-II and HG are covalently attached as side chains (Somerville *et al.*, 2004). RG-II is structurally a complex pectic sub-domain of the plant cell wall, composed of more than 12 different sugars and 20 different linkages distributed in five side chains along an HG backbone. Although RG-II has long been described as highly conserved over plant evolution, recent studies have revealed variations in the structure of the polysaccharide (Buffetto *et al.*, 2014). RG-II plays a crucial role in the cell-wall integrity by enhancing cross-linking of the pectic network through formation of RG-II dimers binding two RG-II monomers via a borate di-ester bond. However, it is also the most complex of the plant polysaccharides and as with HG, more knowledge on its structural and functional roles are still required (Ndeh *et al.*, 2017; Pabst *et al.*, 2013).

1.2.1.4. Lignin

Although not a carbohydrate, lignin is very closely associated with plant cell wall polysaccharides. Lignin confers chemical and biological resistance to the cell wall, and mechanical strength to the plant. The term lignin does not refer to a single well-defined compound as it embraces a whole series of closely related polymers. Lignin originates from three derivatives of phenylpropane: p-coumaryl alcohol (H-lignin unit or “defensive lignin”), coniferyl alcohol (G-lignin unit) and sinapyl alcohol (S-lignin unit) which associate and form an amorphous polymer. The appearance of lignins coincides with the evolution of vascular plants. This compound is of particular importance because of its high resistance to chemical and enzymatic degradation. Physical incrustation of carbohydrates by lignin renders them inaccessible to glycoside hydrolases that would normally digest them. There is evidence that

strong chemical bonds exist between lignin and plant cell wall polysaccharides and proteins that prevents their enzymatic degradation (Cameron, 2015; Carpita *et al.*, 2015).

1.2.2. Plant cell wall hydrolysis

Plant cell wall polysaccharides, primarily cellulose and hemicelluloses, are a major reservoir of carbon and energy. Furthermore, the deconstruction of this complex macromolecule is of growing environmental and industrial significance as the demand for renewable bioenergy sources and substrates for the chemical industry increase (Himmel & Bayer, 2009). However, the chemical and physical complexity of plant cell walls result in an increased resistance to enzymatic degradation and only a restricted number of microorganisms have acquired the ability to deconstruct these structural carbohydrates (Fontes & Gilbert, 2010). Not surprisingly, the microbial degradation of plant cell walls is also a complex process in which an extensive battery of hydrolytic enzymes attacks an heterogeneous, insoluble and highly recalcitrant substrate. These enzymes are generally included in the so called Carbohydrate-Active enZYmes (CAZYmes). The requirement for a consortium of enzymes reflects the diversity and physical association of the polysaccharides within the plant cell wall, which demands that the catalytic entities to act in synergy to degrade this composite structure (Gilbert, 2007). For example, although only a single type of reaction, hydrolysis of β-1,4-glycosidic bonds, is required to convert cellulose to soluble products, degradation of this carbohydrate was shown to be complicated by the insolubility of the substrate and the inaccessibility of the glycosidic bonds, especially in the crystalline regions (Warren, 1996). Thus, the microbial degradation of polysaccharides entails diverse glycoside hydrolases with different specificities and modes of action. The spectrum of enzymes involved in plant cell wall degradation also includes polysaccharide lyases and carbohydrate esterases.

It is now well described that, at the molecular level, microorganisms can organize their plant cell wall CAZymes in two different systems. Thus, in aerobes, enzymes are secreted in copious amounts into the extracellular space and can act freely or associate into the outer membrane. Although these enzymes do not physically associate, they do display extensive biochemical synergy during plant cell wall hydrolysis. In addition, many of these biocatalysts possess a multi-modular structure composed of a catalytic module linked to one or more Carbohydrate-Binding Modules (CBMs), which improve enzyme efficiency by targeting the catalytic module to its target substrate (Gilbert, 2007; Gilbert, Stålbrand, & Brumer, 2008). Alternatively, the plant cell wall degrading enzymes in most anaerobic bacteria and fungi, associate into a large multi-enzyme complex (with a molecular weight higher than 3MDa), termed the Cellulosome,

which is usually anchored to the bacterial surface. The catalytic modules are integrated onto a non-catalytic scaffolding protein (scaffoldin) that may also contains a CBM, thus creating an intimate link between the cell and the substrate surface (Fontes & Gilbert, 2010). This integration is possible due to a strong interaction between the dockerin modules, appended to the enzymes, and the cohesin modules present in the scaffoldins. Scaffoldins also contain dockerin modules for binding to other scaffoldins. It is believed that the anaerobic environment imposes a greater selective pressure for the evolution of these highly efficient nanomachines (Bayer, Belaich, Shoham, & Lamed, 2004).

1.3. Rumen fibrolytic activity

Grasslands and savannas, covering about 20% of the earth's landscape, are a major source of nutrients for wild and domestic herbivores. In addition, annual forage crops are often the primary source of nutrients for domestic mammals. To maximize the value of these resources there is a continuing search for methods to improve the digestibility of both grasses and forage crops and the focus of these studies is the plant cell wall (Barrière, Guillet, Goffner, & Pichon, 2003). Ruminants make up a significant proportion of the domesticated animal species worldwide and, among farmed livestock, they are the best adapted to utilize the energy of plant cell walls (Hungate, 1966). For a long time it has been recognized that a complex community of fibrolytic microorganisms catalyses the degradation of fibre in the rumen. The three species of ruminal bacteria considered to be primarily responsible for plant cell wall biodegradation are *Fibrobacter succinogenes*, *Ruminococcus albus* and *Ruminococcus flavefaciens*. These species gain selective advantage in the rumen by optimizing cellulose hydrolysis (depolymerization) and efficient utilization of the hydrolysis products (cellodextrins) (Weimer, 1996). In addition, *Butyrivibrio fibrisolvens* is a highly xylanolytic Gram positive bacterium inhabiting the rumen, which has a central role in fiber digestion. *Prevotella* are not regarded as highly cellulolytic bacteria, but do produce a range of xylanases. A number of less well characterized cellulolytic bacteria, such as *Eubacterium cellulosolvens*, are also believed to have a significant role in fibre hydrolysis. In addition, anaerobic rumen fungi are considered critical to fibre digestion in the gastrointestinal tract of herbivores and one of the best-studied fibrolytic fungi is *Neocallimastix sp* (Chakrabarty, Demain, & Tiedje, 1997; Hobson & Stewart, 1997). There is also increasing evidence that the rumen protozoa may have the capacity to digest fibre (Devillard *et al.*, 2003). Rumen bacteria have been the subject of intensive studies over the past 60 years, and numerous studies have described the isolation and characterization of a variety of bacterial strains from various ruminant animals (Bryant, Small, Bouma, & Robinson, 1958; Hobson & Stewart,

1997). Out of the three main fibrolytic rumen bacteria, *F. succinogenes* is often the dominant species, although still representing a very small percentage of the total ruminal bacterial population (~0.1%) (Koike & Kobayashi, 2001). *F. succinogenes* activity owes much of their cellulolytic capacity to a strong binding to the surface of plant materials via adhesions which leads to extensive plant cell wall degradation. The fibrolytic enzymes of *F. succinogenes* are amongst the best studied within the rumen bacteria with at least 24 genes encoding endoglucanases and cellodextrinases already identified (Krause *et al.*, 2003) and 23 genes presumed to encode xylanases and other enzymes that hydrolyse non-cellulosic polysaccharides. A large range of glycoside hydrolases has also been isolated from several *R. albus* strains. A number of ORFs containing type I dockerins has been identified in *R. albus* supporting recent biochemical and genetic evidence that this species produces a cellulosome-like complex (Ohara, Noguchi, *et al.*, 2000; Ohara, Karita, Kimura, Sakka, & Ohmiya, 2000). In spite of that, only one cohesin sequence has been identified in *R. albus* genome which argues against the existence of an authentic cellulosome in this species (Artzi, Bayer, & Moraïs, 2016). In contrast, *R. flavefaciens* is the only rumen bacterium known to produce a defined cellulosome and, with more than 220 dockerin containing proteins and several cohesins identified in its proteome, it produces one of the most complex CAZyme rich multienzyme complexes known to date (Rincon *et al.*, 2010). *R. flavefaciens* is the second most abundant and probably most efficient ruminal fibrolytic bacterium. Recent genome sequencing analyses revealed that *R. flavefaciens* strain FD1 possesses at least 107 genes encoding glycoside hydrolases (Dassa *et al.*, 2014) suggesting that a large majority of the proteins assembled into the cellulosome have presently an unknown function.

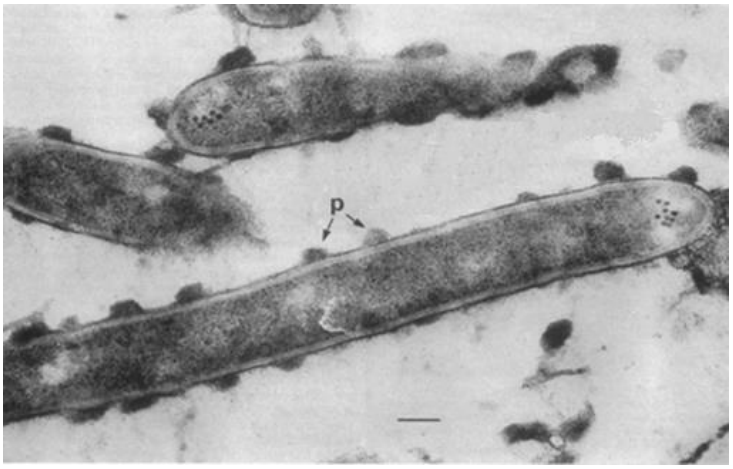
1.4. The Cellulosome

Anaerobic environments impose a greater selective pressure leading to the evolution of highly efficient machineries involved in the extracellular degradation of polymeric substrates. The energy levels in anaerobic bacteria limit the production of enzymes. Thus, to overcome this limitation, anaerobic bacteria have developed alternative strategies for degrading plant structural carbohydrates. The most remarkable one appears to be the organization of CAZYmes into cellulosomes – highly efficient, highly organized cell surface enzymatic systems that enable enzyme recycling and the direct assimilation of hydrolytic products (Bayer *et al.*, 2004). Cellulosomes are produced by a selected number of anaerobic bacteria that colonize different ecosystems, including forest and pasture soils, hot spring pools, sewage sludge, compost piles

and the microbiota of both vertebrates and invertebrates (Doi & Kosugi, 2004; Rosenberg, 2013).

In the early 1980s, the cellulosome complex was first described in the highly cellulolytic thermophilic anaerobe *Clostridium thermocellum* (Bayer, Kenig, & Lamed, 1983; Lamed, Setter, & Bayer, 1983). For many years it was known that bacteria and fungi produce many different types of cellulases that function collectively to promote an efficient degradation of cellulose. It started with biochemical studies of the cellulolytic activity of *C. thermocellum* that revealed the involvement of a large extracellular multi-component complex which is organized on the cell surface (Bayer, Shimon, Shoham, & Lamed, 1998). The presence of cellulosomes on the surface of *C. thermocellum* (Figure 1.4) was first visualized by immuno-cytochemical labelling and electron microscopy. The multienzyme complexes were found to be initially located in protuberances of the outermost layer of the cell envelope and to be subsequently released into the culture medium (Bayer & Lamed, 1986). After binding to cellulose, the cellulosome-containing protuberances elongate and form filamentous protractions which tether the bacterial cells to its substrate. Since all known sequences of cellulosomal polypeptides begin with a signal peptide, they are believed to be secreted individually through a general secretion pathway, therefore suggesting that cellulosome assembly takes place at the surface of the cells (Bégum & Lemaire, 1996).

Figure 1.4 Ultrastructure of *C. thermocellum* cell surface



Transmission Electron Microscopy (TEM) of cationized ferritin (CF)-labeled cellobiose-grown cells of *C. thermocellum* YS. Cells were grown on cellobiose. (p) nodulous protuberances which appear in large numbers over the entire cell surface. From (Bayer & Lamed, 1986)

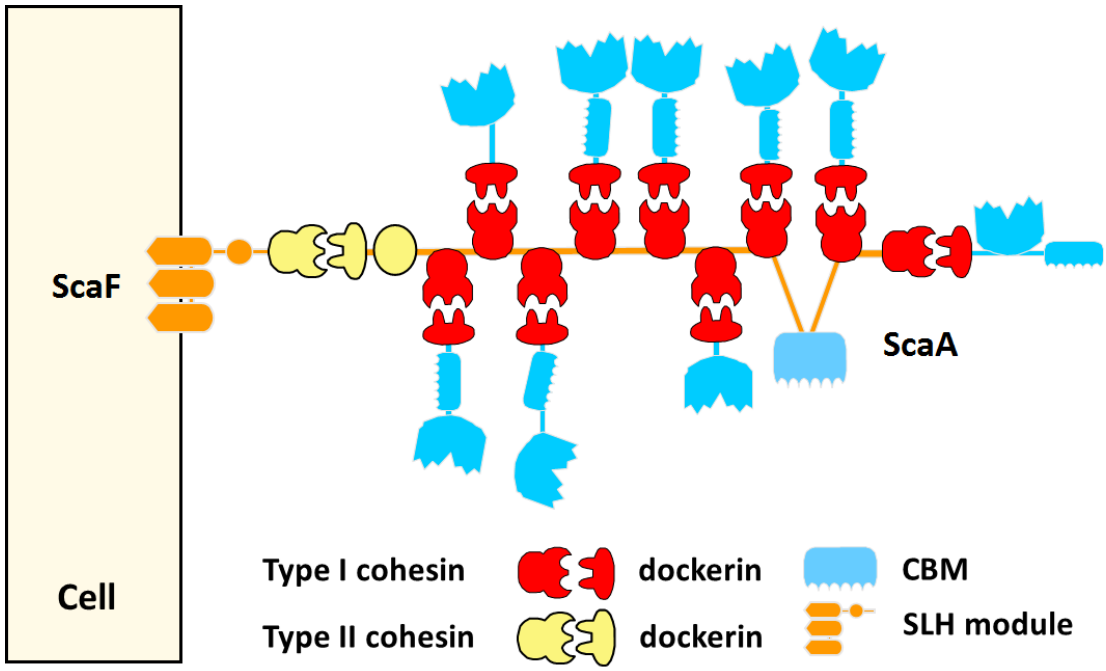
Sequencing of the genes that encode for cellulosomal proteins was fundamental to identify the most important components and the mechanisms of cellulosome assembly. Collaborative data

derived from molecular biology, bioinformatics, biochemistry and structural biology have provided a deep understanding of the molecular forces that enable cellulosome assembly and function (Bayer, Chanzy, *et al.*, 1998). These methods for the understanding of cellulosomal composition are continuously being employed and updated with advances in technology and are continuously broadening our knowledge and capabilities to understand the structure and arrangement of known and unknown cellulosomes from different anaerobic bacteria. For example, the genome of *R. flavefaciens* and *Ruminococcus champanellensis* were recently sequenced revealing that they possess unique cellulosomal architectures: *R. flavefaciens* has the most complex cellulosome described to date while *R. champanellensis* has the largest one ever identified (Ben David *et al.*, 2015; Moraïs *et al.*, 2016; Rincon *et al.*, 2010).

All cellulosomes that have been documented appear to have a similar molecular base for organization. They are composed of two main types of building block: dockerin-containing enzymes or other types of ancillary protein, and cohesin-containing structural proteins, which are termed scaffoldins. Cohesins and dockerins are complementary modules that bind tightly to each other (Artzi, Bayer, *et al.*, 2016). The binding specificity of different cohesin–dockerin pairs dictate the organization of the enzymes into the complex as well as its final architecture (Bayer, Morag, & Lamed, 1994; Doi & Kosugi, 2004). Scaffoldins can also contain a dockerin module for binding to other scaffoldins and a carbohydrate-binding module (CBM) for targeting the complex and its enzymes to appropriate sites on the plant cell wall substrate. Cellulosomes can be attached to the bacterial cell surface or can be released as cell-free cellulosomes (Hamberg *et al.*, 2014; Xu *et al.*, 2016).

C. thermocellum is the most widely studied cellulosome and has served as the archetypal example of these nanomachines and as a blueprint for cellulosome assembly. *C. thermocellum* cellulosomes are composed of a primary scaffoldin subunit, termed ScaA, which integrates enzymes through its nine highly conserved type I cohesins (Figure 1.5**)**. These enzymes or cellulosomal catalytic components contain type I dockerin modules, which bind specifically to the cohesin modules located in ScaA through very tight protein:protein interactions. The C-terminal region of ScaA contains a type II dockerin which interacts with a type II cohesin module located in proteins anchored to the bacterial peptidoglycan layer through an S-layer homology (SLH) module. Thus, type II cohesin-dockerin interactions tether the entire cellulosome to the bacterial cell surface (Bayer *et al.*, 2004; Brás *et al.*, 2016; Xu, Bayer, *et al.*, 2004).

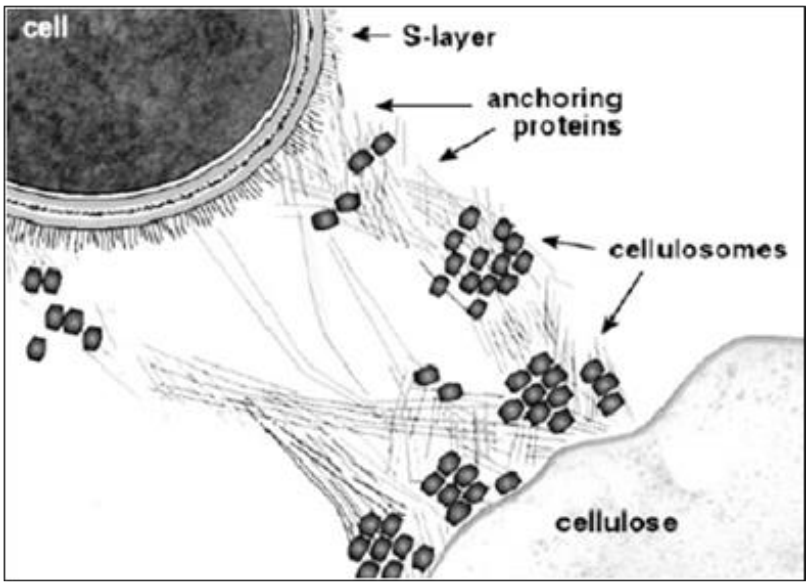
Figure 1.5 A simplistic representation of *C. thermocellum* cellulosome assembly



Schematic representation of *C. thermocellum* cellulosome. Dockerin containing enzymes bind selectively to any of the nine type I cohesins (1 to 9 from the N-terminal) of primary scaffoldin ScaA. The C-terminal X-dockerin dyad of the ScaA binds to the type II cohesin of the anchoring scaffoldin, ScaF, which is connected to the cell surface via an SLH module. The carbohydrate-binding module of the primary scaffoldin binds the cellulosome complex and attached cell to the cellulosic substrate. Adapted from (Brás *et al.*, 2016).

In spite of being structurally related, there is no cross-specificity between type I and type II cohesin-dockerin modules, ensuring an organized mechanism for cellulosome assembly and cell-surface attachment, respectively. ScaA also contains a family 3 carbohydrate binding module (CBM3) which interacts with crystalline cellulose and, therefore, plays a key role in targeting the cellulosome to its substrate, the plant cell wall (Bayer *et al.*, 2004; Gilbert, 2007). The physical association of proteins in cellulosomes is believed to potentiate the biochemical synergy between enzymes, suggesting that cellulosomes are more efficient at deconstructing plant structural polysaccharides when compared to the "free" enzyme systems produced by aerobic bacteria and fungi (Xu *et al.*, 2016) (Figure 1.6).

Figure 1.6 Schematic representation of polycellulosomes bound to cellulose cell surface.



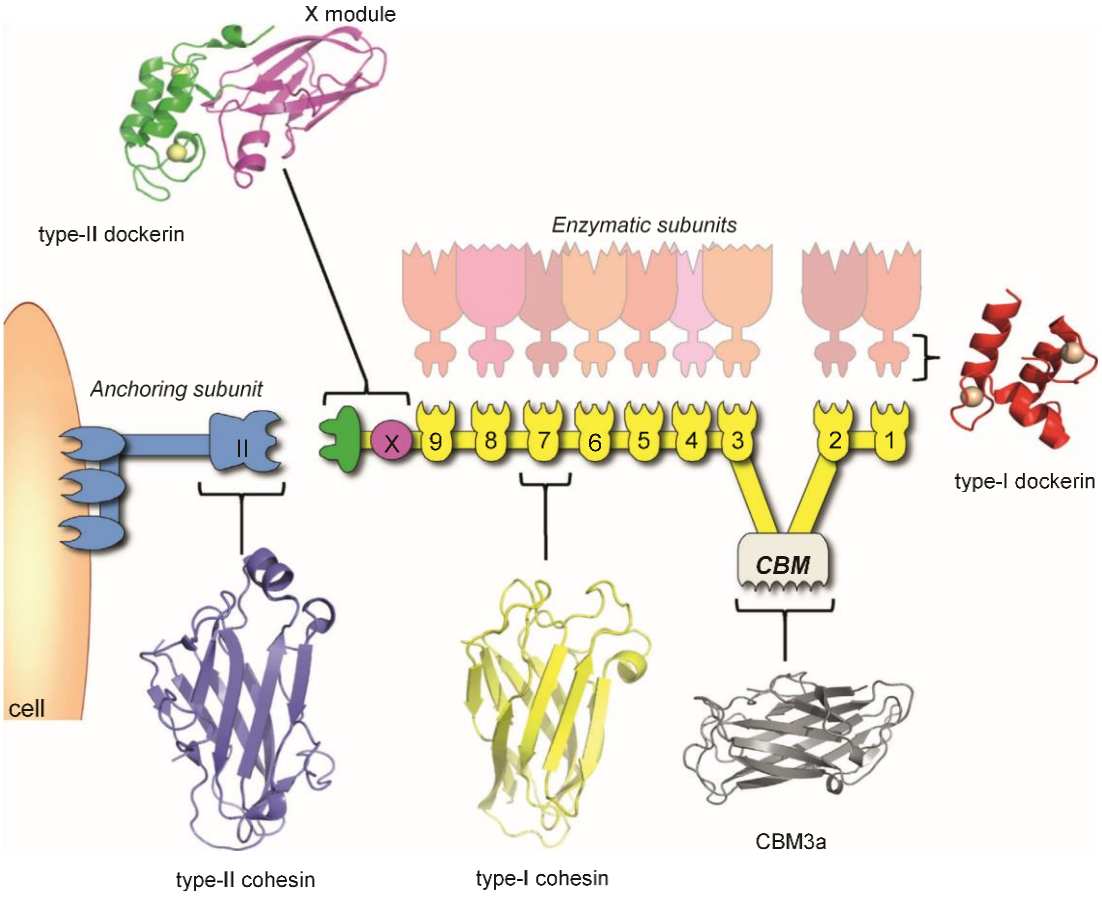
The cellulosome is mainly associated with the cellulose surface and connected to the cell via extended fibrous material. Adapted from (Bayer, Chanzy, *et al.*, 1998).

The synergistic effects result from the substrate targeting of the scaffoldin-born CBM, the proximity of the enzymes and also from the elimination of substrate inhibition due to the rapid uptake of released products (Goyal, Tsai, Madan, DaSilva, & Chen, 2011). *C. thermocellum* exhibits one of the highest known grow rates on cellulose. It has been reported that its cellulosome displays a specific activity against cellulose which is 50-fold higher than the CAZYme system secreted by the aerobic fungi *Tricoderma sp.* (Demain, Newcomb, & Wu, 2005). The genome sequences of several cellulosome producing bacteria are known and preliminary molecular and biochemical characterisation suggest an adaptation of the various cellulosome systems to the different ecological niches (Dassa *et al.*, 2012; Fontes & Gilbert, 2010). The composition of cellulosomes is dependent on the carbon source and other regulatory factors, and the diverse nature within given cellulosome systems has been investigated by transcriptomic and proteomic studies (Osiro *et al.*, 2017; Poudel *et al.*, 2017). There is still a lack of information concerning the molecular components of the cellulosomes of known and unknown organisms (Fontes & Gilbert, 2010) and the development of metagenomics will probably identify additional cellulosome-producing bacteria (Artzi, Bayer, *et al.*, 2016).

Structural biology has recently provided vital information concerning the function of various components of the cellulosome and a blue print for cellulosome assembly. As aforementioned, *C. thermocellum* serves as the archetype and many three-dimensional components of this system have been determined (Figure 1.7). These include several catalytic modules, the type I dockerin module from enzymatic subunits and various scaffoldin borne components, such as

type I cohesins, CBMs, X-modules, C-terminal type II dockerin and the type II cohesin module from anchoring scaffoldins (Smith & Bayer, 2013). This structural information provides fundamental insights into the individual components of the cellulosome. However, the analysis of individual cellulosomal components limits extrapolations concerning the molecular basis of cellulosome assembly and the arrangement of the nano-blocks within the context of the full length cellulosome.

Figure 1.7 Schematic representation of the basic modular architecture of *C. thermocellum*



Three-dimensional components of *C. thermocellum* adapted from (Smith & Bayer, 2013)

1.4.1. The Scaffoldin

The defining elements that distinguish the cellulosome from free enzyme systems are the cohesin containing multi-modular non-catalytic scaffoldins and the dockerin bearing enzymes. Three major types of scaffoldins have been identified by analysing different cellulosomes, namely primary, anchoring and adaptor scaffoldins. The primary scaffoldin is the most important and plays a central role in cellulosomal assembly. It is usually the most expressed

and contains multiple cohesin modules allowing simultaneous integration of several dockerin bearing enzymes (Artzi, Bayer, *et al.*, 2016). Anchoring scaffoldins allow tethering cellulosomes to the cell surface while adaptor scaffoldins are responsible for recruiting cellulosomal enzymes bearing divergent dockerins into the cellulosome. Since anaerobic fibrolytic microorganisms are found in various environmental niches, including the soil, wood chip piles, sewage and the rumen (Doi, Kosugi, Murashima, Tamaru, & Han, 2003), it is likely that these diverse environments have exerted an evolutionary pressure towards the development of complex cellulosomes ensuring a well-adapted system for the particular ecological conditions. In these bacteria expressing cell-surface cellulosomes, the scaffoldin gene is clustered together on the genome with one or more anchoring proteins. The genes for the various enzymes are distributed elsewhere on the genome either alone or in small clusters. In *C. thermocellum*, most anchoring proteins are encoded downstream of the ScaA-containing operon (Lemaire, Ohayon, Gounon, Fujino, & Béguin, 1995).

The first scaffoldin to be described belonged to *Clostridium cellulovorans* cellulosome. At the time, the cellulose-binding function of the cellulosome was recognized. However, the significance of the identified repeating elements (the cohesins) was not apparent (Shoseyov, Takagi, Goldstein, & Doi, 1992). It was not until the scaffoldin from *C. thermocellum* was sequenced, that the relationship between these elements, which were named cohesins, and the dockerins, located in the cellulosomal enzymes, was shown. Characterisation of *C. thermocellum* primary scaffoldin, recently renamed to ScaA, revealed an 1853 amino acid non-catalytic polypeptide containing nine highly conserved type I cohesins which were shown to recognise the type I dockerins from the catalytic subunits (Béguin & Aubert, 1994; Felix & Ljungdahl, 1993; Gerngross, Romaniec, Kobayashi, Huskisson, & Demain, 1993). Some mesophilic bacteria such as *Clostridium josui* and *Clostridium cellulolyticum* were also shown to contain scaffoldins similar to *C. cellulovorans* but, just like in the later, the mechanism by which their single primary scaffoldin is attached to the cell surface is unknown. In contrast, the cellulosome of *C. thermocellum* revealed to be somewhat more complex. *C. thermocellum* ScaA scaffoldin contains an additional C-terminal dockerin that is absent in the mesophilic cellulosomes. Since *C. thermocellum* ScaA C-terminal dockerin displays different cohesin specificity when compared with the type I dockerins located in the cellulosomal enzymes, it was termed a type II dockerin. Thus, it was shown that ScaA type II dockerin interacts specifically with type II cohesins of scaffoldins, which also contain a C-terminal S-layer homology (SLH) module. Biochemical evidence indicates that SLH modules bind to components of the cell envelope (Leibovitz & Béguin, 1996). Therefore, SLH module proteins that contain cohesins were called anchoring scaffoldins. Alternatively, in some species, the

anchoring scaffoldins can bind covalently to the cell surface through sortase (LPXTG) motifs (Rincon *et al.*, 2005). The interaction of primary scaffoldin, ScaA with the anchoring scaffoldins through the type II cohesin-dockerin complexes tethers the *C. thermocellum* cellulosome to the cell surface, leading to the hypothesis of the "enzyme-microbe synergy" or "cellulosome-cell synergy" (Hong *et al.*, 2014). Cellulosomes that assemble via a single primary scaffoldin are characteristic of most mesophilic *Clostridia,* such as *C. cellulolyticum*, *C. cellulovorans*, *C. josui* and *Clostridium acetobutylicum,* among others. These are considered the simplest cellulosomes (Kakiuchi *et al.*, 1998; Pagès *et al.*, 1999). The presence of a specialized dockerin on the primary scaffoldin that mediates cell surface attachment by interacting with the cohesin, or cohesins, of an anchoring scaffoldin is characteristic of highly structured cellulosomes (Bayer *et al.*, 2004). These contain several scaffoldins and many enzymes. In addition, the most complex cellulosomes contain adaptor scaffoldins that either connect two scaffoldins or a scaffoldin and an enzyme. These scaffoldins may have a regulatory role in determining the assembly and composition of a cellulosome complex, depending on the available substrate. Monovalent (single cohesin) adaptor scaffoldins can change the type of enzyme that is integrated into a cellulosome and can be regarded as a 'switch' that changes the cohesin specificity of the primary scaffoldin (Moraïs *et al.*, 2016; Rincon *et al.*, 2004). Depending on the substrate, different enzymes with different activities can thus be integrated into the cellulosomal complex. By contrast, polyvalent adaptor scaffoldins (containing several cohesins) can act as a platform for the expansion of the cellulosome complex and the integration of multiple enzymes (Dassa *et al.*, 2012; Xu *et al.*, 2003), thus enabling more efficient substrate hydrolysis. Distinct cohesin-dockerin specificities between primary scaffoldin complexes, cell anchoring complexes and adaptor scaffoldin complexes ensure a structured cellulosomal assembly.

The majority of cellulosomes that have been described to date are cell-anchored. However, recently, evidence for inherent cell-free scaffoldins was reported in *C. thermocellum*, *C. clariflavum* and *A. cellulolyticus*. The secretion of cellulosomes was verified experimentally for *C. thermocellum* and *C. clariflavum* (Artzi, Morag, Barak, Lamed, & Bayer, 2015; Xu *et al.*, 2016). The expression of cell-free versus cell-anchored cellulosomes has yet to be examined quantitatively. Cell-free cellulosomes are composed of different combinations of scaffoldins when compared with cell-anchored cellulosomes, suggesting that their expression is different.

In addition, in many cellulosome systems, the primary scaffoldin bears a single carbohydrate binding module (CBM). To date, all these scaffoldin-borne CBMs belong to the family-3 CBMs. This type of CBM binds strongly to the crystalline cellulose surface, which accounts for the primary targeting of the cellulosome to its substrate (Shimon *et al.*, 2000; Tormo *et al.*,

1996), and concentrates cellulolytic enzymes in their proximity, contributing further to the efficiency of cellulose degradation (Hong *et al.*, 2014).

1.4.2. The X-Module

N-terminal dockerins located in primary scaffoldins are usually associated with a module of unknown function that has been referred to as the X module. The exact function of this domain remains unknown. However, recent data suggest the involvement of the X-module in providing stability and enhanced solubility to the adjacent cellulosomal components (Adams, Webb, Spencer, & Smith, 2005; Kataeva *et al.*, 2004; Mosbah *et al.*, 2000; Schubot *et al.*, 2004). The crystal structure of a type II cohesin-X-Dockerin complex from *C. thermocellum* showed that the X-module and type II dockerin would allow the cellulosome to extend away from the bacterial envelope when in contact with the type II cohesin. It also suggests that the X-module may contribute to a greater cohesin-dockerin affinity due to the X-module-mediated stabilization of the type II Doc structure in solution combined with the hydrogen-bond contacts established between the X module and type II Coh (Adams, Pal, Jia, & Smith, 2006). The X-module –dockerin pairing has been described in other cellulosomes such as the ones of *Acetivibrio cellulolyticus*, *Pseudobacteroides cellulosolvens*, *Ruminococcus flavefaciens* and *Ruminococcus champanellensis* to name a few. The X-module found in *A. cellulolyticus* displays significant sequence similarity to that of the Ig-like module present in the ScaA C-terminal X-Dockerin of *C. thermocellum* (Xu, Barak, *et al.*, 2004). However, its function remains unknown and speculations about its structural stabilization role for increased affinity in cell surface attachment interactions remains experimentally unverified (Dassa *et al.*, 2012; Xu *et al.*, 2003). In *P. cellulosolvens*, an X-module that shows a high degree of similarity with other X proteins present in various bacteria, including *C. thermocellum*, is closely associated with an SLH-module (Xu, Barak, *et al.*, 2004). This could possibly suggest a greater level of stability with a reduced flexibility in the *P. cellulosolvens* cellulosome. Interestingly, unlike the X-module in the type II Coh-XDoc interaction of *C. thermocellum*, the X-module in the type IIIe CohE-XDoc complex from *R. flavefaciens* does not appear to contribute directly to the CohE-Doc binding surface. Rather, its elongated stalk-like conformation appears to serve as an extended spacer, which separates the cellulose-binding modules at the N-terminus of the primary scaffoldin and the bacterial cell wall (Salama-Alber *et al.*, 2013). Salama *et al*, (2013) suggest a role for these modules that would involve positioning of the catalytic modules away from the cell surface for optimal processing of the glycans of soluble or host-cell surface presented glycoconjugates.

1.4.3. Anchoring modules

Many gram-positive bacteria have a surface-layer protein (SLP) surrounding the exterior cell wall. This layer of proteins is attached to a secondary cell wall of polymers in the rigid cell wall layer. Several extracellular enzymes possess surface layer homology (SLH) domains that are homologous to regions of the SLP. It is believed that, like the SLP, SLH modules also attach to the secondary cell wall polymers binding these SLH-containing enzymes to the cell surface (Doi & Kosugi, 2004). SLH-domains are composed of about 50- to 60-amino acids segments, which are normally repeated threefold (Chauvaux, Matuschek, & Beguin, 1999). As the cellulosomes of several species like *C. thermocellum, A. cellulolyticus* or *P. cellulosolvens* are attached to SLH/cohesin-containing proteins through cohesin-dockerin interactions, it is believed that this interaction tethers them to the cells surface (Doi & Kosugi, 2004).

In some species, like *R. flavefaciens* or *R. champanellensis*, the anchoring scaffoldins do not possess an SLH module. Instead anchoring scaffoldins contain a C-terminal Gram-positive LPXTG-like motif, which is a site for proteolytic cleavage involved in the covalent binding of the scaffoldin to the bacterial cell wall via a sortase-mediated attachment mechanism (Salama-Alber *et al.,* 2013). Proteins for sortase-mediated cell wall anchoring contain several features that are essential for their localization and an N-terminal signal peptide directs these proteins to the secretory pathway. Three crucial features for cell wall anchoring are located at their carboxyl terminal, consisting of an LPXTG motif (leucine, proline, X, threonine and glycine, where X is any amino acid), a hydrophobic region and a tail of charged residues. These features are referred, collectively, as the cell wall sorting signal. During secretion, the hydrophobic domain and charged residues impede membrane translocation, allowing recognition of the LPXTG motif by the membrane-associated sortase enzyme. In a two-step transpeptidation reaction, sortase then cleaves the LPXTG motif between the threonine and glycine residues and covalently attaches the threonine to the amino group of the pentaglycine cell wall cross-bridge resulting in cell wall attached protein (Paterson & Mitchell, 2004).

1.4.4. Carbohydrate Binding Modules

Carbohydrate Binding Modules (CBMs) are non-catalytic domains that recognize different carbohydrates and are most commonly found in modular glycoside hydrolases and other carbohydrate modifying enzymes (Hammel *et al.,* 2005). Initially, these non-catalytic polysaccharide-recognizing modules were defined as Cellulose-Binding Domains (CBDs) since the first studied examples had crystalline cellulose as their primary ligand (Boraston,

Bolam, Gilbert, & Davies, 2004; Gilkes, Warren, Miller, & Kilburn, 1988). In order to reflect their diverse ligand specificity, these modules are now termed as Carbohydrate Binding Modules. Thus, CBMs that recognize crystalline cellulose, non-crystalline cellulose, chitin, β-1,3-glucans and β-1,3-1,4-mixed linkage glucans, xylan, mannan, galactan and starch have been described, while some CBMs display 'lectinlike' specificity and bind to a variety of cell-surface glycans (Boraston *et al.*, 2004). CBMs were also previously defined as a contiguous amino acid sequence, within a carbohydrate-active enzyme, with autonomous folding and skilled recognition for a specific carbohydrate motif (Gilbert, 1999). This definition is not entirely accurate though, as there are CBMs that are found isolated as independent proteins or that are part of cellulosomal scaffoldin proteins, responsible for targeting the cellulosome complex to the substrate (Boraston *et al.*, 2004; Fontes & Gilbert, 2010).

Like the catalytic modules of CAZYmes, CBMs are divided into families in the CAZy database (Lombard, Golaconda Ramulu, Drula, Coutinho, & Henrissat, 2014). There are currently 81 different CBM families registered on cazy.org, a number that has been consistently growing. The family classification of CBMs was introduced to aid in the identification of novel CBMs. In some cases, the family classification may allow predicting of the binding specificity while aiding in identifying functional residues and revealing evolutionary relationships (Gilbert, 1999). Within these families these modules have been shown to display three distinct specificities. Therefore, CBMs have been characterized into three types: type A CBMs which interact with crystalline polysaccharides, primarily cellulose, type B modules which bind to internal regions of single glycan chains, and type C CBMs that recognize small saccharides in the context of a complex carbohydrate (Fontes & Gilbert, 2010).

CBMs potentiate the function of associated CAZymes (Bolam *et al.*, 1998). Thus, CBMs play a key role in the deconstruction of complex insoluble composites by the appended catalytic modules and have three general effects with respect to the action of their cognate catalytic modules: a proximity effect, a targeting function and a disruptive function (Boraston *et al.*, 2004; Guillén, Sánchez, & Rodríguez-Sanoja, 2009). Previous studies showed that maintaining enzymes in the proximity of the insoluble substrate leads to a more rapid degradation of the recalcitrant polysaccharide. Therefore, the removal of CBMs from enzymes or from scaffoldins, dramatically reduces the enzymatic activity of the associated catalytic modules (Bolam *et al.*, 1998; Boraston, Kwan, Chiu, Warren, & Kilburn, 2003). However, the activity on soluble substrates is not frequently affected when CBMs are removed (Kleine & Liebl, 2006; Waeonukul *et al.*, 2009). Additionally, there are examples of CBMs that have become components of the substrate-binding sites of glycoside hydrolases, and that are pivotal to the substrate specificity and mode of action of the cognate enzymes. Hence, the efficient hydrolysis

of polysaccharide requires a dynamic interaction between CBMs and their substrates, where the catalytic domain is first positioned in the proximity of the substrate through the CBM. Then, the catalytic domain is able to hydrolyse the polysaccharide chains inserted in the active site. CBMs can be relocated to new regions on the ligand allowing a continuous hydrolysis of the substrate (Guillén, Sánchez, & Rodríguez-Sanoja, 2010).

Primary scaffoldins generally contain a family 3 CBM, such as the one present in ScaA of *C. thermocellum* (Fontes & Gilbert, 2010). CBM3 is a type A CBM and binds strongly to the crystalline surface of cellulose accounting for the primary targeting of the cellulosome to its substrate (Artzi, Bayer, *et al.*, 2016; Bayer *et al.*, 2004). The CBMs found on primary scaffoldins of *A. cellulolyticus* and *P. cellulosolvens* have also been classified as family 3 CBMs. However, they have been termed CBM3b. Even though they still bind strongly to crystalline cellulose, subtle differences at the primary and tertiary structure suggest that scaffoldin CBMs are more diverse than originally considered (Ding, Bayer, Steiner, Shoham, & Lamed, 1999, 2000). In contrast, the cellulosomal scaffoldins of *R. flavefaciens* do not contain such CBMs. Rather, this bacterium has an independent CttA scaffoldin, which possesses two putative CBMs and a dockerin that specifically recognizes a cohesin allocated on a anchoring scaffoldin and may thus serve to attach the bacterial cell to its substrate (Rincon *et al.*, 2010; Salama-Alber *et al.*, 2013).

Interestingly, a putative cell-free scaffoldin found in *A. cellulolyticus* cellulosomal system contains three cohesins and two family 2 CBMs (Dassa *et al.*, 2012). Comparable CBM2-containing scaffoldins were also observed in the genome of *C. clariflavum* (Artzi *et al.*, 2014). These are the only known examples of scaffoldins containing CBMs from a family other than CBM3. Family 2 CBMs are more commonly found associated with free, non-cellulosomal enzymes. They have been divided into two subfamilies, one of which binds to cellulose and the other binds to xylan (Simpson, Xie, Bolam, Gilbert, & Williamson, 2000). Family 2 CBMs are of interest, as they seem to be present only on cell-free scaffoldins and not on cell-anchored scaffoldins, which suggests that they have a different role than family 3 CBMs (Artzi, Bayer, *et al.*, 2016).

1.4.5. Catalytic components

Cellulosomal CAZYmes are modular enzymes containing, in addition to the dockerin domain, one or several catalytic modules and sometimes one or more CBMs (Fontes & Gilbert, 2010). The cellulosome was initially identified due to its ability to bind and degrade cellulose very effectively. Thus, numerous cellulases were first identified as part of cellulosomal complexes

(Lamed *et al.*, 1983). In addition to cellulases, other polysaccharide-degrading, carbohydrate-active cellulosomal enzymes were subsequently identified; most notably, xylanases, pectinases, carbohydrate esterases, mannanases and xyloglucanases (Artzi, Bayer, *et al.*, 2016).

CAZymes are particularly diverse and complex. They include glycoside hydrolases, carbohydrate esterases and polysaccharide lyases. These enzymes are broadly grouped according to their functionality, and are classified into families on the basis of their primary sequence homology (Cantarel *et al.*, 2009; Coutinho, Deleury, Davies, & Henrissat, 2003; Henrissat, 1998; Henrissat & Davies, 2000). The classification of glycoside hydrolases in families based on amino acid sequence similarities is useful as there is a direct relationship between sequences and folding. Hence, this classification better reflects the structural features of these enzymes than their corresponding substrate specificities. In addition, it helps to show the evolutionary relationships between carbohydrate active enzymes while providing a convenient tool to infer mechanistic information. The CAZy database provides a continuously updated list of the glycoside hydrolase families (Lombard *et al.*, 2014). As protein fold is better conserved than primary structure, some of the families can be further grouped into clans (Henrissat, 1998). Mechanistically, cellulose hydrolysis requires the cooperative action of at least 3 groups of enzymes: endo-(1,4)-*β*-D-glucanase, exo-(1,4)-*β*-D-glucanase and *β*-glucosidases. The exoglucanase acts on the ends of the cellulose chain and releases *β*-cellobiose as the major end product; endoglucanase randomly attacks the internal O-glycosidic bonds, resulting in glucan chains of different lengths; and the *β*-glucosidases act specifically on the *β*-cellobiose disaccharides producing glucose (Kuhad, Gupta, & Singh, 2011).

Interestingly, all known cellulosome-producing bacteria characteristically express large amounts of a single glycoside hydrolase 48 (GH48) exoglucanase, that is crucial for enzymatic activity (Artzi *et al.*, 2015; Morag, Halevy, Bayer, & Lamed, 1991; Ravachol *et al.*, 2015). Contrastingly, an extensive repertoire of family 9 glycoside hydrolases is generally secreted by these bacteria. Recently, large sets of GH9 enzymes from *R. champanellensis* and *C. cellulolyticum* were characterized. Independently of their modular organization, these enzymes were shown to exhibit different activities, distinct abilities to bind to cellulosic substrates and diverse synergies with the major Cel48A exoglucanase (Moraïs *et al.*, 2016; Ravachol, Borne, Tardif, de Philip, & Fierobe, 2014). This eventually suggests that enzyme diversity, especially of GH9 enzymes, reflects the structural diversity of cellulose and associated hemicellulose. Many other glycoside hydrolase families, such as GH5, GH10, GH11 and GH43, are also commonly found in cellulosome systems, providing bacteria with a powerful and diverse enzymatic apparatus for the effective hydrolysis of plant cell wall polysaccharides (Artzi, Bayer, *et al.*, 2016).

Polysaccharide lyases (PLs) are a group of enzymes which cleave the glycosidic bonds of uronic acid-containing polysaccharide chains via a *β*-elimination mechanism to generate an unsaturated hexenuronic acid residue and a new reducing end. These enzymes show a large variety of fold types (or classes), suggesting that polysaccharide lyases have been invented more than once during evolution from totally different scaffolds (Lombard *et al.*, 2010). Presently there are 24 families of PLs described.

Carbohydrate esterases generally remove ester based modifications present in mono-, oligo- and polysaccharides and thereby facilitate the action of GHs on complex polysaccharides. Since an ester is formed by an acid and an alcohol, at CAZy, two classes of substrates for carbohydrate esterases were considered: those in which the sugar plays the role of the "acid", such as pectin methyl esters and those in which the sugar behaves as the alcohol, such as in acetylated xylan. Presently 16 families are described in CAZy (Cantarel *et al.*, 2009).

CAZymes act synergistically to hydrolyse resistant plant-derived substrates. Synergism may be due to different modes of action towards the same substrate like when an endoglucanase hydrolyses the substrate, thereby producing additional chain ends that can be cleaved by an exoglucanase. It can also result from the hydrolysis of two tightly associated substrates, in which the action of one enzyme could make the concealed substrate accessible for the action of the second enzyme (eg. cellulases and xylanases). Also, the product inhibition effect over one enzyme could be decreased and its activity restored by another enzyme acting on said product (eg. *β*-glucosidases and cellulases) (Morag, Halevy, *et al.*, 1991).

In addition to CAZymes, other dockerin-containing proteins are present in cellulosomes, such as serpins (Kang, Barak, Lamed, Bayer, & Morrison, 2006), proteases (Levy-Assaraf *et al.*, 2013) and expansin-like proteins (Artzi, Morag, Shamshoum, & Bayer, 2016; Chen *et al.*, 2016). These enzymes have unique functions that are uncommon to cellulosomes, and their diverse roles may contribute to a range of physiological processes in bacteria, to the assembly and regulation of cellulosome components and/or indirectly to the degradation of biomass (Artzi, Bayer, *et al.*, 2016).

1.4.6. Linker regions

In general, enzymes that degrade plant cell wall polysaccharides display a modular architecture, which comprises one or more catalytic domains bound through flexible linker sequences to one or more non-catalytic modules. Previous studies have shown that modules in each cellulosomal subunit are interconnected by a variety of linker segments of different lengths and composition. Linkers are responsible for the connection of cohesin modules within the scaffoldin unit and

also to connect the dockerins with the catalytic subunits. Conformational changes in cellulosomal components, particularly the intermodular linker segments, have long been considered to play a physiological role in the efficient degradation of cellulose (Morag, Bayer, & Lamed, 1991). Although the structural and functional properties of individual cellulosomal modules are well documented, very little is known about the role of the linker peptide, and its structural properties are a matter of speculation (Noach *et al.*, 2009). Studies by small angle X-ray scattering have demonstrated that enzyme-borne linkers are not likely to be the main generators of such flexibility that may be required for plant cell wall degradation, once the enzyme is bound within the cellulosome. Alternatively, the linkers in the scaffoldin subunit probably possess the main flexibility role in cellulosome function. Thus, studies have shown that the intrinsic plasticity of linker segments allows a variety of dramatically different conformations of the scaffoldins (Hammel *et al.*, 2005). A more recent study suggests that the flexibility of the linkers connecting consecutive cohesins modules could control structural transitions and thus regulate substrate recognition and degradation. In addition, data comparing the efficacy of designer cellulosomes containing 3 different linker lengths between cohesins suggested that longer linkers seem to improve the hydrolytic efficacy (Vazana *et al.*, 2013).
Linker sequences, which connect the adjacent cohesin domains in *A. cellulolyticus* scaffoldin, are generally rich in prolines and threonines. Extended stretches of their sequences are remarkably similar, and reminiscent to those of the *C. thermocellum* scaffoldin subunit (Ding *et al.*, 1999). The high incidence of prolines suggests that linkers form extended configurations that physically separate the various modular domains (Ding *et al.*, 1999). In addition, proline-rich regions of proteins have been suggested to cause rapid and non-specific binding, which in the case of scaffoldins may promote intermodular and/or inter-subunit protein:protein interactions (Ding *et al.*, 1999). The numerous threonines would be suitable glycosylation sites, as demonstrated for the *C. thermocellum* and *P. cellulosolvens* scaffoldins (Gerwig *et al.*, 1993; Tomme, Warren, Gilkes, & Poole, 1995). Interestingly, linkers in *R. flavefaciens* cellulosomal enzymes are usually T-rich, but a significant number of GHs possess unusual asparagine-glutamine rich linkers previously only observed in free enzymes (Berg Miller *et al.*, 2009).

1.4.7. Cohesin-Dockerin

The cohesin–dockerin interaction is fundamental for the assembly of cellulosome complexes. This non-covalent protein:protein interaction has been shown to exhibit one of the strongest binding affinities known in nature (Gunnoo *et al.*, 2016; Stahl *et al.*, 2012; Valbuena *et al.*, 2009), being remarkably difficult to dissociate (Bhat, Goodenough, Bhat, & Owen, 1994).

Using single-molecule force spectroscopy, the force that is required to break the interaction between a cohesin–dockerin pair was estimated to be half of the force required to break a covalent bond (Schoeler *et al.*, 2015). Highly complex cellulosomes contain several interconnected components that articulate with each other through these strong interactions. Considering that the cellulosome is both tethered to the cell surface while also binding the substrate through its CBMs, it is likely that it will often be subjected to mechanical stress imparted by opposing forces. Hence, such a strong bond between the cohesin and dockerin modules is required to maintain the complex assembly and stability under adverse environmental conditions.

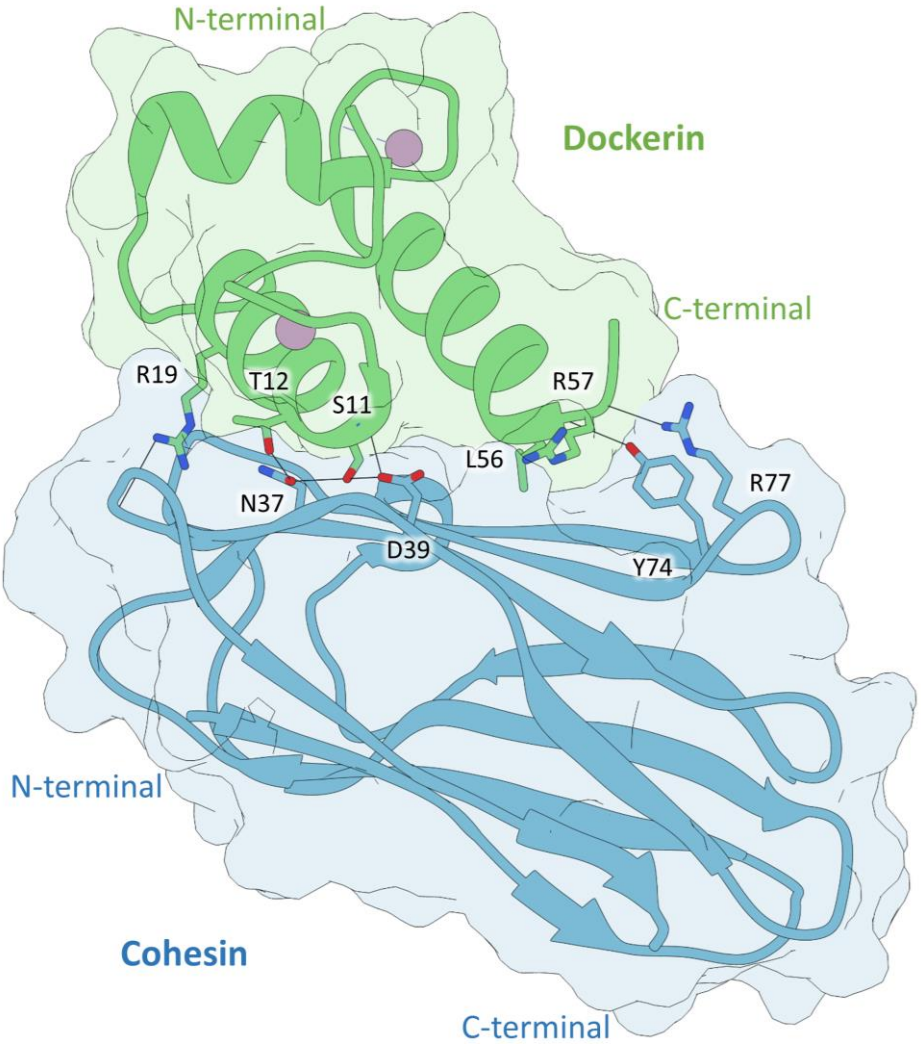
The organization and structural architecture of cellulosomes is orchestrated by the specificity of the different cohesin and dockerin modules. Based on primary structure analysis, three types of cohesin–dockerin interaction have been described. Type I interactions occur between dockerin-containing enzymes and cohesins of the primary scaffoldin. Type II interactions occur between two scaffoldins (usually primary and anchoring scaffoldins). Curiously, *Pseudobacteroides cellulosolvens* is the only known bacterium to have the opposite interaction pattern, with its enzymes containing type II dockerins and the scaffoldins containing type I dockerins (Xu, Bayer, *et al.*, 2004). Furthermore, Type III interactions are observed in ruminococcal cellulosomes and are distinct from the type I and II interactions from *Clostridium spp* (Dassa *et al.*, 2014; Karpol *et al.*, 2013). Dissection of cellulosomal components requires a closer analysis of the cohesin and dockerin modules and their individual properties as discussed below.

1.4.8. Types I interactions

The first cellulosomal components whose tridimensional structure was revealed were the type I cohesins from the scaffoldins of *C. thermocellum* and *C. cellulolyticum* (Shimon *et al.*, 1997; Spinelli *et al.*, 2000). The type I cohesin module is around 150-residues and is organized in a nine-stranded β-sandwich arranged in a jelly-roll topology with an elongated shape. The two sheets of the sandwich are composed of strands 8, 3, 6, 5 and 9, 1, 2, 7, 4, respectively, with β-strand 9 (C-terminus) and β-strand 1 (N terminus) running parallel to each other and the remaining running anti-parallel. The nine β -strands are assembled around an extensive aromatic core (Carvalho *et al.*, 2003; Shimon *et al.*, 1997). All three cohesin types interact with their dockerins through β-strands 5, 6, 3 and 8 of their β-sheets (Adams *et al.*, 2006; Carvalho *et al.*, 2003; Weinstein *et al.*, 2015).

Dockerins are usually present in a single copy at the C-terminus of cellulosomal enzymes. They consist of approximately 70 amino acids and contain 2 duplicated segments of around 22 residues, each of which comprises a distinctive Ca^{2+}-binding loop and α-helix. An NMR solution structure of the free *C. thermocellum* type I dockerin module from cellobiohydrolase Cel48S revealed that the first 12 residues of each duplicated segment resemble the calcium-binding loop of EF-hand motifs, in which the calcium-binding residues (aspartate and asparagine) and calcium coordination patterns are highly conserved (Fontes & Gilbert, 2010; Lytle, Volkman, Westler, Heckman, & Wu, 2001; Salamitou, Tokatlidis, Béguin, & Aubert, 1992; Tokatlidis, Salamitou, Béguin, Dhurjati, & Aubert, 1991). In this context, calcium dependence for dockerin function was demonstrated experimentally (Choi & Ljungdahl, 1996) and both duplicated segments were shown to be involved in cohesin recognition (Fierobe *et al.*, 1999). Additionally, the presence of the duplicated segments suggested that the structure of these modules may display a two-fold symmetry. The structure of a type I dockerin from *C. thermocellum* Xyn10B in complex with the second cohesin module of ScaA scaffoldin, obtained by (Carvalho *et al.*, 2003), confirmed this and provided the first insights into the mechanism by which cellulosomes are assembled (Figure 1.8). The dockerin module contains three α-helices, with helices 1 and 3 bearing the key conserved residues previously identified in the first and the second dockerin duplicated segments, respectively. Each duplicated segment displays remarkable structural conservation and also contributes an F-hand calcium-binding motif. Thus, two calcium ions are present in the dockerin within the two EF-hand loops. The three α -helices present a conformation defined by a loop-helix motif followed by a helix-loop-helix motif, connected by a six-residue segment. By revisiting the Cel48S NMR structure, a long-standing enigma has been recently resolved. It was believed that the dockerin undergoes conformational changes following cohesin binding. However, new evidence now favours an inherent cohesin-primed conformation of the dockerin without cohesin-induced alterations to its structure (Chen *et al.*, 2014). The structure of the type I complex illustrates that the cohesin interacts with its dockerin partner primarily along one face of its flattened β-barrel (β-strands 5, 6, 3 and 8). Although the dockerin presents remarkable internal symmetry, the detailed crystal structure of the first cohesin-dockerin complex revealed that the dockerin prefers binding to the cohesin through its second duplicated segment (helix 3) and only the C-terminal region of the helix 1 contributes to ligand recognition (Carvalho *et al.*, 2003, 2007). While hydrophobic forces dominate the cohesin-dockerin interface, the proteins also interact through hydrogen bonds in which a highly conserved Ser-Thr pair in helix 3 of the dockerin plays a central role in these polar interactions (Carvalho *et al.*, 2003; Gilbert, 2007).

Figure 1.8 Structure of the type I Coh-Doc complex of *Clostridium thermocellum*



The protein-protein complex formed between a cohesin molecule (blue) and a Ca^{2+}-bound dockerin (green). The most important residues involved in domain contacts are shown as stick models. The two Ca^{2+} ions are represented as pink spheres (Carvalho *et al.*, 2003).

In a subsequent study, Carvalho *et al.* (2007) revealed that the type I dockerin of *C. thermocellum* can bind to its cohesin partner through two distinct surfaces. By mutating the critical Ser-Thr pair located at the C-terminal helix, the binding was disrupted through this helix and reverse binding through the other helix was observed, thus revealing that the dockerin display a dual binding mode. Thus, substitution of the Ser-Thr pair of helix 3 with two alanine amino acids, lead to a 180º rotation of the dockerin with respect to the cohesin, with helix 1 assuming the position of helix 3, and the Ser-Thr pair in the first duplicated segment dominating the hydrogen bond network (Carvalho *et al.*, 2007). In essence, the equivalent residues in helix 1 of the mutant and helix 3 in the wild-type dockerin interact in the same manner with the cohesin module and so, an almost perfect overlapping of the two alternative binding interfaces

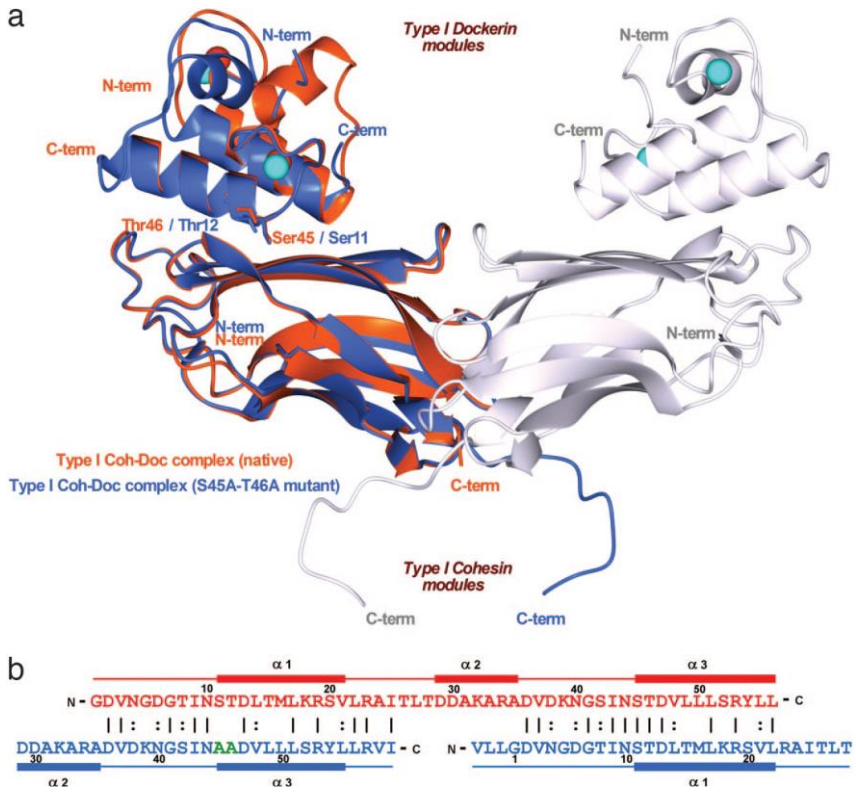
was observed (Figure 1.9). Karpol *et al* (2008) performed further truncation and mutation experiments to confirm the symmetry of the cohesin-dockerin interaction. It was found that the first calcium-binding loop can be deleted entirely, with almost full retention of binding to the cohesin. Likewise, significant deletion of the second repeated segment can be achieved, provided that its calcium-binding loop remains intact. In addition, mutations in one of the calcium-binding loops failed to disrupt cohesin recognition and binding, whereas a single mutation in both loops significantly reduced the affinity (Karpol, Barak, Lamed, Shoham, & Bayer, 2008). These results are in accordance to the structural data previously obtained.

Interestingly, data reported by Carvalho *et al.* (2007) revealed that the mutated (C-terminal Ser-Thr replaced by alanine) and wild type dockerins displayed equivalent affinities for the cohesin binding partner, suggesting that a dual binding operates in solution and most possibly *in vivo*. Thus, it is believed that the dual binding mode may be responsible for the introduction of required quaternary flexibility into the multi-enzyme complex and for the enhancement of the substrate targeting and synergistic interactions between complementary enzymes, particularly the exo- and endo-acting cellulases (Carvalho *et al.*, 2003; Gilbert, 2007). The stoichiometry of the binding of a variety of type I cohesin-dockerin complexes is consistently 1:1, which suggests that the two binding interfaces are not able to recognise their ligands simultaneously. Therefore, the dual binding mode may be responsible for reducing steric constrains that are likely to be imposed by assembling a large number of different catalytic and non-catalytic domains into a single cellulosome. Quaternary flexibility could be further provided by the proline-threonine rich linker sequences that join cohesins within scaffoldins. Indeed, probing cellulosome components by small angle X-ray scattering supports the proposal that the inter-module linkers in free enzymes are extended and flexible. The linker sequences joining the cohesin domains within the *C. thermocellum* scaffoldin are quite long, up to 35 residues, and thus the conformational freedom displayed by the scaffoldin protein may contribute to the synergy displayed by the enzymes within the cellulosomes (Hammel *et al.*, 2005; Hammel, Fierobe, Czjzek, Finet, & Receveur-Bréchot, 2004). Additionally, in order to optimise the synergy between specific enzymes, the efficiency of cellulosome function may require, temporarily, the switching of the enzymatic subunits from one cellulosome position to another. Since the cohesin-dockerin interaction is extremely tight, the existence of a second ligand binding surface in type I dockerins may facilitate the switching of the appended enzymes onto a different cellulosomal cohesin (Fontes & Gilbert, 2010). Béguin and colleagues performed site-directed mutagenesis and thermodynamic studies in different cohesin-dockerin complexes revealing that substitution of residues 11 and 12 (Ser-Thr pair) at one of the helices of *C. thermocellum* dockerin had no major impact on the cohesin-dockerin interaction (Miras,

Schaeffer, Béguin, & Alzari, 2002; Schaeffer *et al.*, 2002). Therefore, only the substitution of both serine-threonine motifs in helix-1 and helix-3 with bulky amino acids significantly reduces the affinity of the dockerin for its ligand (Carvalho *et al.*, 2007; Pinheiro *et al.*, 2012; Schaeffer *et al.*, 2002). These data are in accordance with the structures obtained for the Coh-Doc complexes. Similar observations in the type I interaction responsible for the binding of *A. cellulolyticus* adaptor scaffoldin ScaB to the anchoring scaffoldin ScaC were observed: only the mutation of both DocScaB key residue pairs (Ile-Asn) were able to disrupt binding (Cameron, Najmudin, *et al.*, 2015).

Recent transcriptomic, proteomic, and complimentary biochemical and structural studies have shown that type I cohesin modules are not exclusive to *C. thermocellum* cellulosome scaffoldins and the dual binding mode is not an entirely ubiquitous feature of the type I dockerin (Smith & Bayer, 2013). Four established *C. thermocellum* type-I dockerin modules, two associated with cellulases (Cel124A and Cel9D-Cel44A) and two others with proteins of unknown function (Cthe_0258 and Cthe_0918), contain sequential substitutions that do not allow a dual binding mode (Pinheiro *et al.*, 2009). The type-I dockerins from Cel124A and Cthe_0258 specifically bind the type-I cohesin module of the anchoring protein ScaG, while the type-I dockerin module from Cthe_918 similarly recognized the type-I cohesin modules from ScaA and ScaD. The structures of CohScaG-Cel124A and CohScaD-Cthe_0918 revealed that each of these dockerins display a single mode of binding with their cognate cohesin module; each being orientated 180º with respect to the other (Brás *et al.*, 2012). Thus, these data suggest that while the dual binding mode operates in dockerins that bind to the cellulosome, dockerins used to fix the appended enzymes to the bacterial cell-surface seem to display a single-binding mode.

Figure 1.9 The dual binding mode of *Clostridium thermocellum*'s Xyn10B dockerin.



A) Ribbon representation of the superposition of the type I Coh-Doc WT complex (in orange) with its S45A-T46A mutant complex (in blue). In the mutant complex, helix-1 (containing Ser-11 and Thr-12) dominates binding whereas, in the WT complex, helix-3 (containing Ser-45 and Thr-46) plays a key role in ligand recognition. Ser-11, Thr-12, Ser-45, and Thr-46, which interact with the cohesin module, are depicted as stick models and colored accordingly. The second molecule of the mutant complex is represented in light-gray ribbon. The Ca^{2+} ions are depicted as spheres and colored orange, in the case of the WT complex, and light blue, in the case of the mutant. The N- and C-terminal ends are labeled and colored accordingly. B) The structure based sequence alignment of the WT (in red) and S45A-T46A mutant (in blue) type I dockerins. Mutated residues, Ala-45 and Ala-46, are shown in green. Because of internal 2-fold symmetry of each dockerin module, the two structures overlap almost perfectly in their α1/ α3 regions. (Carvalho *et al.*, 2007).

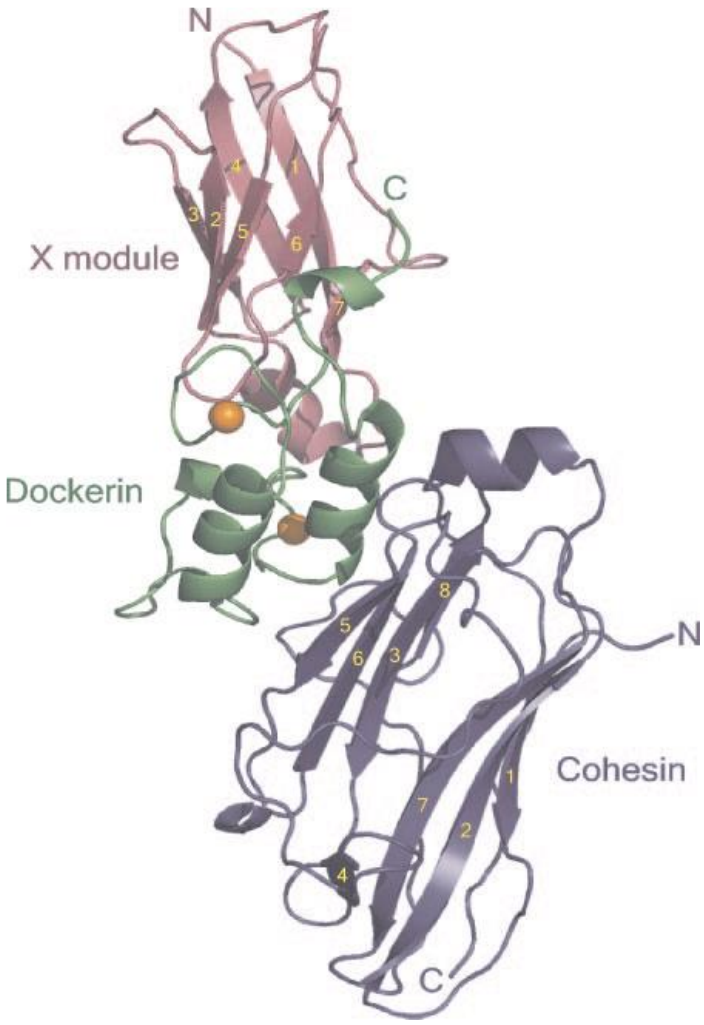
1.4.8.1. Type II interactions

Attachment of cellulosomes to the bacterial cell surface is a crucial mechanism for the optimal uptake of nutrients and consequently for the viability of the microbe. In *C. thermocellum*, type II dockerins tether the cellulosome to the peptidoglycan layer of the bacterial cell envelope through high-affinity interactions with type II cohesin modules located in cell-surface proteins ScaF, ScaB, ScaC. They can also bind to the cohesins from the extracellular scaffoldins ScaH and ScaE (Brás *et al.*, 2016; Fontes & Gilbert, 2010; Leibovitz & Béguin, 1996). The first type II cohesin crystal structure to be obtained was the type II cohesin of *P. cellulosolvens* of scaffoldin ScaA shortly followed by the structure of the type II cohesin from *C. thermocellum* anchoring protein ScaF (Carvalho *et al.*, 2005; Noach *et al.*, 2005). With the exception of a few

structural elements including the presence of an α-helix, between β-strand 6 and 7 and of two "β-flaps" interrupting β-strands 4 and 8, both structures have the same jelly-roll topology observed in type I cohesins. The sequences of these three secondary elements, as well as the rest of the structural elements, are more conserved between all type II cohesins than between type I cohesins (Carvalho *et al.*, 2005; Noach *et al.*, 2005). The crystal structure of the *C. thermocellum* ScaF type II cohesin in complex with the type II ScaA dockerin was obtained by Adams and colleagues (2006) (Figure 1.16). The type II cohesin also displays a typical jellyroll fold. Data indicated that the cohesin does not undergo significant conformational changes upon ligand binding (Adams *et al.*, 2006), a feature that is evident in type I cohesins from other microorganisms (Carvalho *et al.*, 2005; Noach *et al.*, 2003, 2005). It was shown that the type II dockerin displays a similar fold to its type I counterpart. However, type II dockerins closely interacts with a neighbouring module of unknown function, the previously mentioned X-module, which adopts an immunoglobulin-like fold. Unlike type I dockerins, in which cohesin recognition is dominated by only one of the dockerin helices, it was found that in type II dockerins both helices contact with the cohesin surface over their entire length. The interaction surfaces are significantly less charged, thus binding is predominantly hydrophobic. There is an extensive hydrogen-bonding network that involves residues from the X module, both dockerin helices and the β-strands 8-3-6-5 face of the cohesin module. Furthermore, the type II cohesin-dockerin complex reveals an intimate hydrophobic interface between the type II dockerin and the Ig-like X-module fold, giving the C-terminal region of the ScaA scaffoldin a rigid and elongated conformation. Besides interacting with the type II dockerin, the ScaA X-module also contributes to the different specificities displayed by the type I and the type II dockerin partners and might even contribute to structural stability and enhanced solubility of cellulosomal components.

Isothermal titration calorimetry (ITC) assays were performed in order to assess the binding affinity of the type II cohesin-X-dockerin interaction in solution. Titration of the X-dockerin into type II cohesin showed that these proteins bind with a 1:1 stoichiometry. However, it was impossible to determine an accurate affinity constant because this interaction has a very high affinity ($K_a > 10^{10}\ M^{-1}$) which exceeds the detection limits of this technique (Adams *et al.*, 2006). It was proposed that the increased affinity of the type II interaction is due to the X-module-mediated stabilisation of the type II dockerin structure in solution, combined with the hydrogen-bond contacts that exist directly between the X module and the type II cohesin. Thus, this crystal structure has extended our understanding of the extraordinary diversity in specificities displayed by type I and type II cohesin-dockerin protein partners.

Figure 1.10 Structure of the type II Cohesin-Xdockerin complex (ScaFCoh-ScaAXDoc).



Ribbon representation of the type II cohesin-dockerin complex with the cohesin module in blue, the dockerin in green and the X module in magenta. The *β*-strands of the X-module and the type II cohesin are numbered in yellow. The N- and C-termini are labelled accordingly and the Ca^{2+} ions are depicted as orange spheres (Adams *et al.*, 2006).

Structural and biochemical data revealing the dual binding mode, has dramatically affected the way that cohesin-dockerin interactions are perceived. Dockerins displaying a dual binding mode contain a near perfect 22-residue repeat, unlike that of dockerins exhibiting a single binding mode. With this in mind Noach *et al.* (2010) have hypothesised that the type II dockerin of *A. cellulolyticus* may in fact display a dual binding mode, due to its near identical segment repeat in contrast to the type II dockerin of *C. thermocellum* (Figure 1.17). Their attempts to crystallize the type II cohesin-dockerin complex were however, unsuccessful. The authors reasoned that the apparent symmetry of the type II dockerin module of this bacterium may lead to a dual binding mode which would in turn result in formation of heterogeneous complexes hindering the crystallisation process.

Table 1.1 Symmetrical and asymmetrical dockerin sequences

		***Ct* type II dockerin (Asymmetric)**	***Ac* type II dockerin (Symmetric)**
Seg 1		DIVKDNSINLLDVAEVIRCFNA	GGTQDGAINLEDILEICKAFNT
Seg 2		DINRNGAINMQDIMIVHKHFGA	DLNRDGAISLEDVMIVAKHFNK
		** ::.:**: *: * : *.*	. .:****.***:: :.* **.

Sequence alignment of the two 22 segment repeats of the asymmetric *C. thermocellum* type II dockerin (from ScaA) and the symmetrical type II ScaA dockerin of *A. cellulolyticus*. Identical residues are indicated with asterisks. Adapted from (Noach *et al.*, 2010).

In conclusion, although cohesin and dockerin modules have been classed by different types, it is becoming more and more apparent that this classification is only relevant in terms of phylogenetic similarities and in fact it could be argued that each of the modules from a given species should be viewed and characterised individually given its intrinsic functional properties. Despite initial evidence that the type I and type II interactions reflect dual-binding and single-binding modes, respectively, it is now clear that the mode of binding is not strictly indicative of the modular type (Nash, Smith, Fontes, & Bayer, 2016). It is rather the conserved or divergent nature of the recognition residues in the two repeated segments that determines the binding mode of a given dockerin.

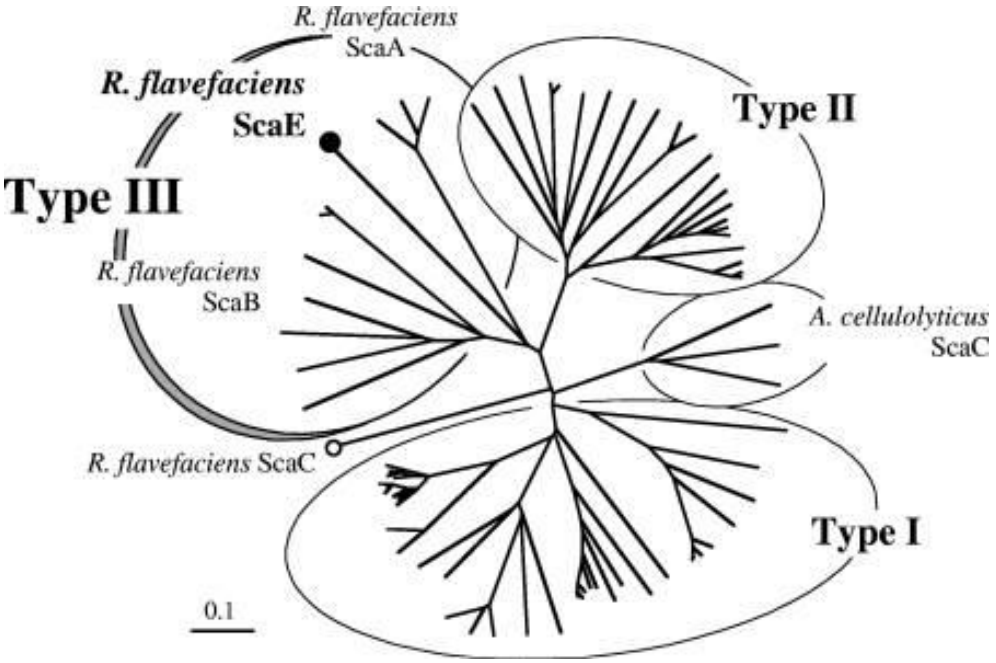
1.4.8.2. Type III interactions

The cohesin, dockerin and X-modules of *Ruminococcae* cellulosomal components were found to be divergent in sequence from previously known type I and type II cellulosomal modules, and their Coh-Doc interactions were therefore collectively designated type III based on their respective phylogenetic trees (Rincon *et al.*, 2003, 2004, 2005, 2007). The cohesin from the *R. flavefaciens* adaptor scaffoldin ScaC is more divergent from the other ruminococcal cohesins and seems to be more closely related to type I cohesins from other species (Figure 1.11). This was confirmed by determining the structure of CohScaC in complex with a group 3 dockerin from strain FD-1 (Bule *et al.*, 2016). This data will be discussed in chapter 3 of this thesis.

Type III interactions can also be of either single-binding or dual-binding mode. In fact, the ruminococcal type III cohesins and dockerins are highly diverse and possess different specificities within the same species (Israeli-Ruimy *et al.*, 2017). The type III dockerin of the CttA protein from *R. flavefaciens* contains two additional helices, but interacts with the ScaE cohesin in a manner similar to that of the type I interaction. The specialized atypical type III

dockerins that have extra helices are rare and contain three unusual sequence inserts that act as structural buttresses to support the extended stalk-like neighbouring X module. The latter module probably maintains the parent protein at a fixed distance from the cell surface and thus requires the additional physical reinforcement that is provided by the inserts (Salama-Alber *et al.*, 2013). Type III cohesin of *R. flavefaciens* ScaE has a very similar topology to that of type I cohesins and interacts with its dockerin module through *β*-strands 5, 6, 3 and 8, similar to that of the type I interaction. It contains two '*β*-flaps' between *β*-strands 4 and 8, similar to those of type II cohesins, but also has a prominent 13-residue *α*-helix that is enveloped by an extensive amino-terminal loop that is not found in other cohesin types (Salama-Alber *et al.*, 2013). There is still, however, a paucity of data regarding type III cohesin-dockerin complexes. The diversity of type III modules has made it clear that there is more to Coh-Doc interactions than single vs dual binding mode. They have created a new opportunity to build on current information about this extremely strong and specific protein:protein interaction that can have major applications in a plethora of different fields.

Figure 1.11 Phylogenetic relationship of *R. flavefaciens 17* and type I and II cohesin domains.



ScaE-Coh maps as a type III cohesin, separated from the other type III cohesins of ScaA and ScaB. In contrast, the *R. flavefaciens* ScaC cohesin maps on a separate branch, closer to *A. cellulolyticus* ScaC and the type I cohesins of other species (Adapted from (Rincon *et al.*, 2005).

1.4.8.3. Specificity

Structure-function studies in a variety of cohesin-dockerin systems suggest that, within the same species, there is no cross-specificity between enzyme recruiting and cellulosome-cell anchoring cohesin-dockerin partners, providing an effective mechanism for cellulosome assembly and cell-surface attachment (Miras *et al.*, 2002; Schaeffer *et al.*, 2002). In addition, it is now well recognized that there is no cross-species specificity among the most popular cohesin dockerin pairs within different species. In addition, the sequence duplication observed in type I dockerins from a variety of organisms, beyond *C. thermocellum*, indicates that the dual binding mode may be replicated in the majority of other microbial cellulosomes (Table 1.2). Analysis of the aligned dockerin sequences suggests a correlation with the essential Ser-Thr pair (positions 11 and 12) found in both duplicated segments, which very likely represent specificity determinants of the interaction in *C. thermocellum*. In the first two species that were examined, *C. thermocellum* and *C. cellulolyticum*, these positions were essentially preserved within each species but divergent between them (Pagès *et al.*, 1997). The same interspecies divergence exists between the cohesin-dockerin interaction of *C. thermocellum* and *C. josui* (Jindou *et al.*, 2004). Thus, mutagenesis studies by Pages *et al*, (1997) showed that the type I dockerin found in *C. cellulolyticum* cellulosomal enzymes do not interact with the cohesins found in *C. thermocellum* scaffoldins and vice-versa. In general, residues at positions 11 and 12 of *C. cellulolyticum* dockerin were changed to the equivalent amino acids in *C. thermocellum* dockerins and vice-versa, and the resultant variants recognised cohesins from both clostridia (Pagès *et al.*, 1997). Later studies suggested that besides residues at positions 11 and 12, residues at positions 18, 19 and 23 of the dockerin are also involved in species-specific ligand recognition (Mechaly *et al.*, 2001). Hence, the very tight interaction between cohesins and dockerins is generally species specific, although there is a considerable similarity in sequence and structure between cohesins and dockerins from different species (Bayer *et al.*, 2004). Altogether, these data suggest that cellulosomal enzyme sharing is not an evolutionary driver in different cellulolytic organisms. However, it may be possible that microorganisms inhabiting the same ecological niche, in extreme circumstances, benefit from the sharing of cellulosomal enzymes. An example that supports this is the considerable sequence homology that results in cross-species specificity observed in *C. cellulolyticum* and *C. josui* dockerins (Fontes & Gilbert, 2010; Jindou *et al.*, 2004). In contrast to the type I interaction, type II cohesin-dockerin complexes have a relatively extensive cross-species plasticity. For example, the type II cohesin of the *C. thermocellum* anchoring scaffoldin ScaF binds not only to the *C. thermocellum* ScaA type II dockerin, but also to both *P. cellulosolvens* and *A. cellulolyticus* type II cohesins. Additionally, a type II dockerin of *A. cellulolyticus* binds both *A. cellulolyticus* and *C.*

thermocellum type II cohesins (Haimovitz *et al.*, 2008). The biological relevance, if any, of the promiscuity of the type II cohesin-dockerin interaction remains unknown. In conclusion, as more cellulosome-producing bacteria and their cohesins and dockerins are sequenced, our views on the sequence characteristics of these modules have changed. Although cohesin–dockerin specificity of a given type is usually preserved within a species, there are exceptions. In addition, restricted specificity between species is not always observed. Broad interspecies recognition is common among the simple cellulosome systems of mesophilic clostridia and some cross-species overlap has been observed with type II interactions (Haimovitz *et al.*, 2008).

Table 1.2 Suspected recognition residues of different dockerin domains derived from cellulosomal components of different species.

Organism	Protein(s)	Duplicated segment 1					Duplicated segment 2				
		Positions					Positions				
		11	12	18	19	23	11'	12'	18'	19'	23'
C. thermocellum	Enzymes[a]	S	T	K	R	K,R,G	S	T	K,H,S	R,K	K,R
	CipA	L	L	I	R	A	M	Q	H	K	A
A. cellulolyticus	Cel9B	S	I	R	K	G	S	L	R	Q	G
	ScaA	L	E	C	K	T	L	E	A	K	K
	ScaB	I	N	R	D	G	I	N	R	D	G
B. cellulosolvens	Cel48A	M	A	A	Q	K	M	A	A	Q	K
	ScaA	S	D	R	Q	G	S	D	R	Q	G
R. flavefaciens	Enzymes (EndB-like)[a]	L	A	M	Q	N	G,N,A,E	D	Q	K,R,E,L	R,K,H,G
	CesA	S	F	R	K	N	V	A	Q	S	G
	XynE	S	L	R	R	K	V	A	K	R	G
	ScaA	V	A	N	R	D	D	K	D	P	K
	ScaC	L	A	M	Q	N	G	D	Q	K	T
C. cellulolyticum	Enzymes[a]	A	L,I	K	K	G	A	L,I	K	K	S,G
C. josui	Enzymes[a]	A	L,I	K	K	N,T	A	I	K	K,V	A,N
C. cellulovorans	Enzymes[a]	A,S	L,I	K	K	D,N,S,G	A	I	K	K	S,G,A
C. acetobutylicum	Enzymes[a]	G	R	R,T	K,Q	G	G	R	I,R	K,Q	G,N,S

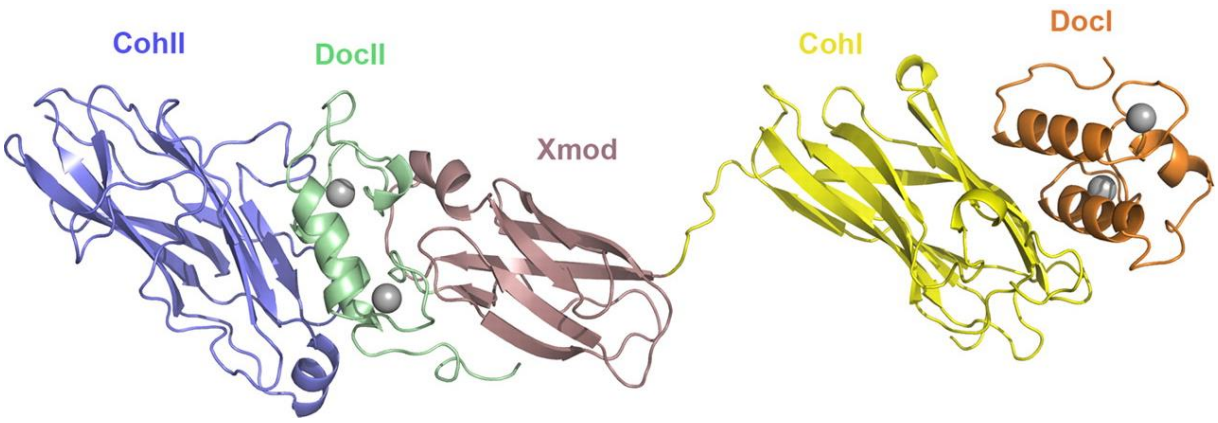
Scaffoldin-borne dockerins are highlighted in grey. [a] Consensus residues represent the dominant amino acids that appear in the designated position from the indicated group of cellulosomal enzymes. Adapted from (Bayer *et al.*, 2004).

1.4.9. Quaternary Structural organization

Whereas the high-resolution structures of several cellulosomal components have been elucidated, the structural organisation of the complete cellulosome remains poorly understood. The success at determining the structures of individual cohesins and dockerins and their

combined complexes has generated ambitious attempts to crystallize larger portions of cellulosomal components. However, these efforts have proved problematic, and only isolated crystal structures of larger cellulosome fragments have been described (Currie *et al.*, 2012). Initial studies using electron microscopy indicated that polycellulosome organelles are located on the cell surface and appear as extended protuberances in the presence of a cellulose substrate (Bayer & Lamed, 1986). Small-angle X-ray scattering studies showed that the conformational flexibility provided by the linker regions between type I cohesin modules of the scaffoldin allow for optimal positioning of the enzymatic subunits onto the substrate. However, the linker regions present between the dockerin modules and the catalytic core of the enzymatic cellulosomal components were proposed to be predominantly rigid (Hammel *et al.*, 2005, 2004). Additionally, the crystal structure of the type II cohesin-dockerin complex showed an unexpected extensive modular interface between the type II dockerin and its neighbouring X module, which revealed that the C-terminal region of the ScaA scaffoldin has an elongated topology (Adams *et al.*, 2006). The most extensive crystal structure determined by x-ray includes three different proteins comprising five separate modules of the *C. thermocellum* cellulosome complex (Currie *et al.*, 2012). It contains the type I Doc of Cel9D bound to the C-terminal trimodular fragment of the ScaA scaffoldin (the ninth cohesin I, $CohI_9$, connected by a linker to the X-dockerin II) which in turn is bound to the ScaF type II cohesin (Figure 1.12). The structure reveals an elongated topology with a flexible 13-residue linker connecting the ninth type I cohesin module and the X module. Elevated temperature factors suggest that the linker is highly dynamic. This flexibility could allow the ninth type I cohesin to explore a larger conformational space, providing closer proximity with the type II X-dockerin-cohesin region. Four molecules of the DocI·$CohI_9$–X-DocII·CohII ternary complex were found in the asymmetric unit of the crystal structure. Alignment of the X-DocII·CohII region from the four molecules of the complex, as well as the $CohI_9$–X-DocII·CohII complex, reveals slightly different orientations of the DocI·$CohI_9$ region. Together, these studies suggest the existence of several possible conformations of the linker sequences when bound to the neighbouring cohesin modules, indicating that structural changes in the linker regions may contribute to modulate the overall conformation of the cellulosome (Adams, Currie, *et al.*, 2010).

Figure 1.12 Crystal structure of the DocI·CohI9–X-DocII·CohII ternary cellulosomal complex.



The backbone ribbon representation depicts ScaF CohII in blue, the ScaA DocII in green, X module in rose, $CohI_9$ in yellow, and the Cel9D DocI in orange. Calcium ions are shown as gray spheres (Adapted from Currie *et al.*, 2012).

(García-Alvarez *et al.*, 2011)), used cryo-electron microscopy and revealed that a large fragment of the cellulosome presents a very compact conformation in solution. The three-dimensional structure of a *C. thermocellum* mini-cellulosome that comprises three consecutive cohesin ScaA modules (third, fourth and fifth) bound to three Cel8A cellulases, through their native dockerin modules, was solved (García-Alvarez *et al.*, 2011). Unlike what was observed by SAXS experiments (Hammel *et al.*, 2005), the structure showed that the linker regions between cohesin modules exhibited a restricted flexibility. This compact conformation is thought to be a result of the stabilisation of specific contacts between cohesin modules by the linkers. Further work is required in order to obtain novel insights into the higher-order scaffoldin interactions which are behind cellulosome modular architecture and polycellulosome formation. Reports on the deconstruction of the cellulosome to its various components are extremely helpful providing a structural basis for previous biochemical studies. However, this approach is limited regarding the molecular basis of cellulosome assembly or for the specific arrangement of the various modules in terms of the full-length multi-protein nanomachine (Smith & Bayer, 2013).

1.4.10. Cellulosome diversity

When cellulosomes were discovered it was initially thought that cellulosome-producing bacteria would be prevalent in nature. However, it has become increasingly apparent that cellulosomes are specialized and rare, although essential for degradation of recalcitrant polysaccharides derived from plant cell walls in lignocellulosic ecosystems (Artzi, Bayer, *et al.*, 2016). Nevertheless, with the increased availability of genomic and metagenomic

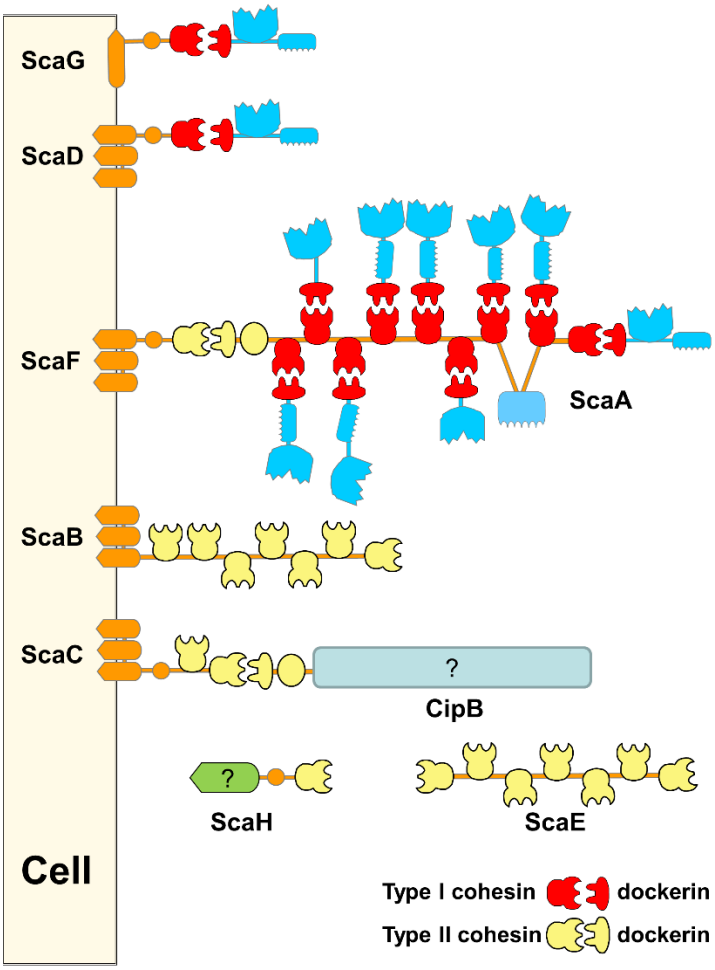
information cellulosome diversity is expanding and is likely to still increase. It is apparent that scaffoldins have the potential to be exceptionally varied in size, structural organisation and nature of their modular components (Ding *et al.*, 2000). Thus, two major types of cellulosomes have been described to date, depending on the presence of exclusive primary scaffoldins or both primary and anchoring scaffoldins. Future descriptions of new microbial scaffoldins will further contribute to our knowledge concerning the similarity and diversity among the cellulosome systems in nature (Ding *et al.*, 2000). In order to demonstrate the potential of the cellulosomal system to adopt several distinct elaborate architectures, the cellulosomes of four bacterial species are described below.

1.4.10.1. *Clostridium thermocellum*

As the archetypal cellulosome, the organization of *C. thermocellum* multi-enzyme complex has already been partially described. *C. thermocellum* contains multiple anchoring scaffoldins, that are connected to the cell surface via SLH modules. The cellulosomes of *C. thermocellum* were shown to be located at the cell surface in the early stages of growth (Leibovitz, Ohayon, Gounon, & Béguin, 1997). They have a central non-catalytic primary scaffoldin, which has been termed ScaA (Figure 1.13). ScaA harbours nine type I cohesin modules separated by linker regions of varying lengths which bind the various type I dockerin containing subunits, most of which consist of CAZymes (Smith & Bayer, 2013). The more recently described CipB contains a signal peptide followed by four modules of unknown function, including a region containing 19 repeats of a 41-residue motif with three highly conserved replicated cysteine residues (Brás *et al.*, 2016). Both CipB and ScaA contain a C-terminal divergent type II X-dockerin module that recognise type II cohesins located in one of three anchoring proteins named ScaB, ScaC and ScaF, that carry seven, two and one cohesin(s), respectively, through a highly specific and extremely tight interaction (Lemaire, Miras, Gounon, & Béguin, 1998). In all cases studied so far, SLH repeats are found in these proteins and biochemical evidence indicates that they bind to components of the cell envelope (Leibovitz & Béguin, 1996). Recent genomic sequencing efforts identified 72 dockerin-containing proteins, suggesting the potential for related cellulosomal subunits encoded in the genome (Raman *et al.*, 2009). Genomic, transcriptomic and proteomic analysis confirmed this assumption and revealed that *C. thermocellum* has two additional type II cohesin-containing anchoring proteins, ScaH and ScaE, the latter of which is believed to be exclusively extracellular, as there is no SLH domain present and it comprises seven type II cohesin modules, allowing the potential to form polycellulosomal structures that may contain up to 63 catalytic subunits (Fontes & Gilbert, 2010; Pinheiro *et al.*, 2009; Raman

et al., 2009). Additionally, type I cohesin modules from *C. thermocellum* were identified in two cell surface proteins, ScaG and ScaD, suggesting that cellulosomal enzymes can also adhere directly, and individually, onto the bacterial cell envelope (Fontes & Gilbert, 2010; Salamitou *et al.*, 1994). However,(Pinheiro *et al.*, 2009), proved that Xylanase 10B-like dockerins, which are the most common in *C. thermocellum*, seem to display a much higher affinity for ScaA cohesins than to ScaG, the dominant type I cohesin-containing cell surface protein (Raman *et al.*, 2009). It was therefore suggested that cellulosomal enzymes may transiently interact with the bacterium's cell surface by binding to ScaG, before they are assembled into the multi-enzyme complexes (Pinheiro *et al.*, 2009).

Figure 1.13 Organization of *C. thermocellum* cellulosome.



The *C. thermocellum* scaffoldin (ScaA) contains nine type I cohesins and thus organizes a multiprotein complex with nine enzymes (blue), through the binding to type I dockerins. ScaA also contains a cellulose-specific family 3 CBM (dark blue). The C-terminal type II dockerin domain of ScaA binds specifically type II cohesin domains found in cell-surface proteins ScaF, ScaB, and ScaC (orange) or in the extracellular ScaE and ScaH. Cellulosomal enzymes may adhere directly to the bacterium cell surface by binding the single type I cohesin domains found in ScaD and ScaG. The linkers joining the

modules in the scaffoldin and catalytic subunits are shown as orange and blue lines, respectively. Adapted from (Fontes & Gilbert, 2010).

1.4.10.2. *Acetivibrio cellulolyticus*

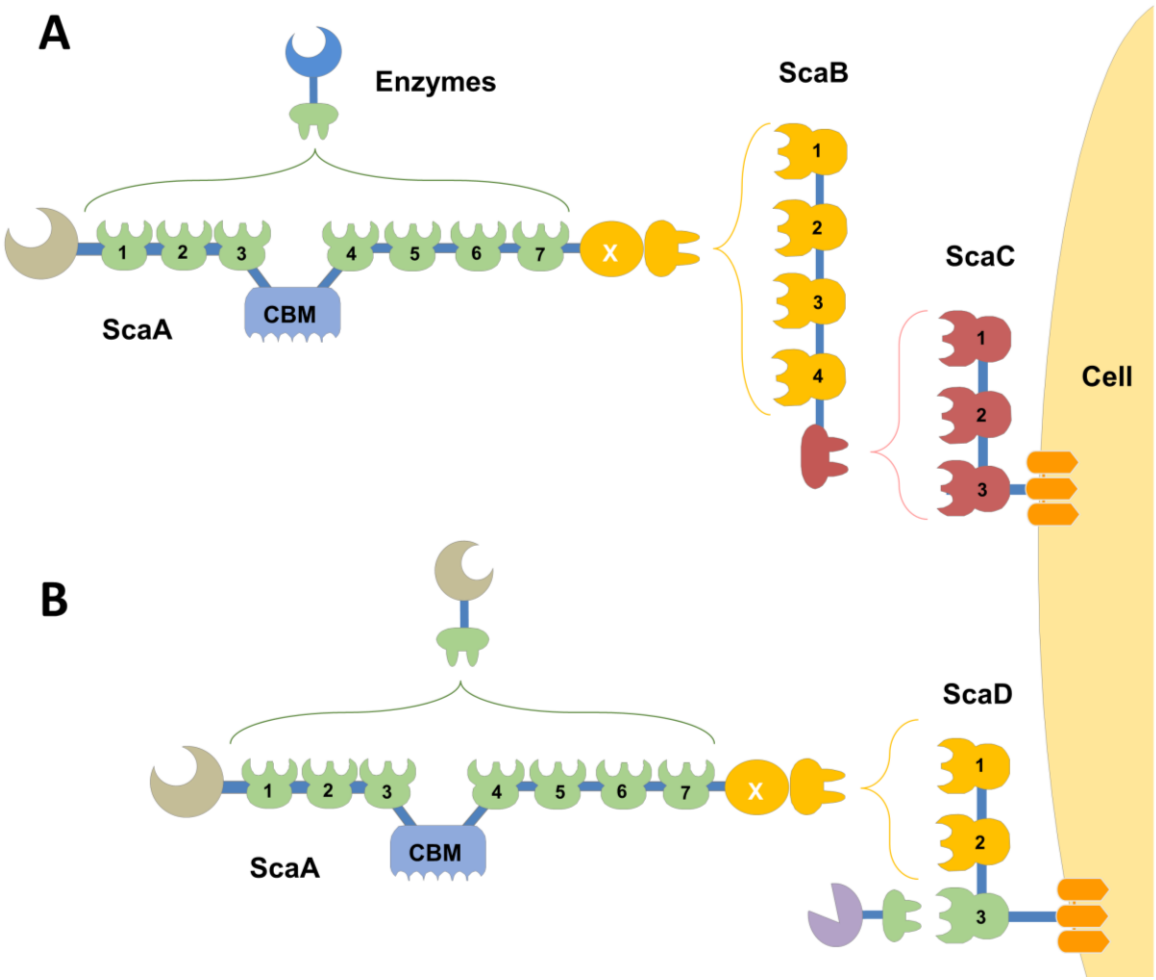
Scanning electron microscopy allowed Lamed *et al*, (1987) to identify protuberance-like structures on the cell surface of many cellulolytic micro-organisms, including the mesophilic, anaerobic bacterium *Acetivibrio cellulolyticus*. In the early 80's *A. cellulolyticus*, which inhabits sewage sludge, was shown to contain a cellulolytic enzyme system capable of hydrolysing a range of cellulosic materials as efficiently if not better than commercially available enzyme preparations (Khan, 1980). The identification of recognizable sequences, mainly through the sequencing of genes encoding cohesin and dockerin domains, along with previous biochemical studies supported the idea that cellulosomes occur in *A. cellulolyticus* (Bayer, Chanzy, *et al.*, 1998).

A. cellulolyticus primary scaffoldin ScaA (originally termed CipV), was sequenced more than ten years ago, and comparative sequence analysis of its functional modules with those of earlier sequenced scaffoldins provided insight into the structural arrangement and phylogeny of this family of microbial proteins (Ding *et al.*, 1999). The ScaA from *A. cellulolyticus* shares the main traits found in the primary ScaA scaffolding of *C. thermocellum*, containing an internal CBM, bordered by seven type I cohesin domains, a single X-module and a divergent C-terminal type II dockerin. Interestingly, a family 9 glycoside hydrolase (GH9) sequence was identified as being part of the polypeptide chain; this was surprising as until then catalytic modules had only been seen in free enzymes or non-scaffoldin cellulosomal subunits (Ding *et al.*, 1999). Further genomic sequencing downstream of the ScaA scaffoldin locus revealed the gene encoding two more scaffoldin protein: ScaB and ScaC. ScaB was found to contain four type II cohesin modules, which interact with the C-terminal type II dockerins of the ScaA, and a divergent C-terminal type I dockerin which in turn interacts with the type I cohesin modules found on the ScaC scaffoldin. ScaB essentially plays the role of an adaptor protein, which mediates the interaction between ScaA (and its incorporated enzymes) and ScaC. This ScaB scaffoldin was the first example of an adaptor protein. In turn ScaC, acts as an anchoring scaffoldin by virtue of its C-terminal SLH module (Xu *et al.*, 2003).

Later, Xu *et al.* (2004) completed the sequence of a successive gene in the ScaA gene locus of *A. cellulolyticus* that was termed ScaD. ScaD contains two different types of cohesins, type I and type II, therefore exhibiting two divergent dockerin-binding specificities. The consequence of this molecular arrangement is that ScaD can integrate two primary scaffoldins via its resident type II cohesins and, additionally, a single dockerin-containing enzyme via the type I cohesin. Like ScaC,

ScaD was also found to contain C-terminal segments encoding an SLH module making it an anchoring protein. Since ScaA can harbour eight enzymes, the ScaD-anchored cellulosome system of *A. cellulolyticus* has the potential to incorporate up to 17 enzymes, in addition to the 96 enzymes that can be assembled through the ScaC-anchored system. (Figure 1.14) (Xu, Bayer, *et al.*, 2004).

Figure 1.14 Schematic representation of the *A. cellulolyticus* cellulosomal components

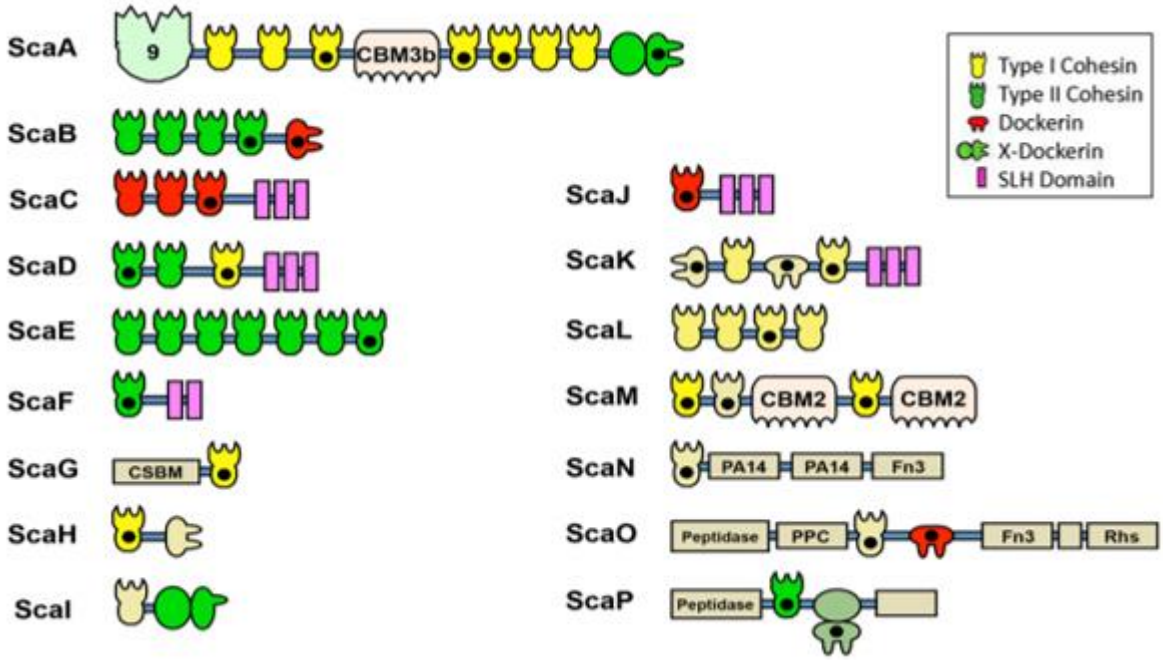


In A) Dockerin-containing enzymes are incorporated into the ScaA scaffoldin through interaction with the seven ScaA cohesins (light green). ScaB plays the role of an adaptor protein that mediates between the ScaA dockerin (yellow) and the cohesins of the anchoring scaffoldin (red) ScaC. The entire complex appears to be cell associated via the resident SLH module of ScaC (orange). ScaA contains also a CBM (blue) and a GH9 (light brown) catalytic module. In B) An additional mechanism of cellulosome attachment; ScaA is bound to the type II cohesins of ScaD (yellow), which can also accept a single enzyme via its third type I cohesin (light green). The SLH module of ScaD serves to anchor the alternative complex to the cell surface. Adapted from Bayer *et al.*, 2008.

The complete sequencing of the *A. cellulolyticus* CD2 genome has enabled the identification and analysis of numerous additional cellulosomal components, gene regulatory elements and cell anchoring modules, revealing a much more elaborate and sophisticated cellulosome system than originally proposed (Dassa *et al.*, 2012). Analysis of the genome sequence uncovered 41

putative cohesin modules which are distributed among 16 scaffoldins (including the four genes of the Sca cluster ScaA, ScaB, ScaC and ScaD), some of which have both cohesins and dockerins in the same polypeptide chain (Figure 1.15).

Figure 1.15 Modular architecture of the array of scaffoldins identified in the *A. cellulolyticus* CD2 genome. Adapted from *(Dassa et al., 2012)*.



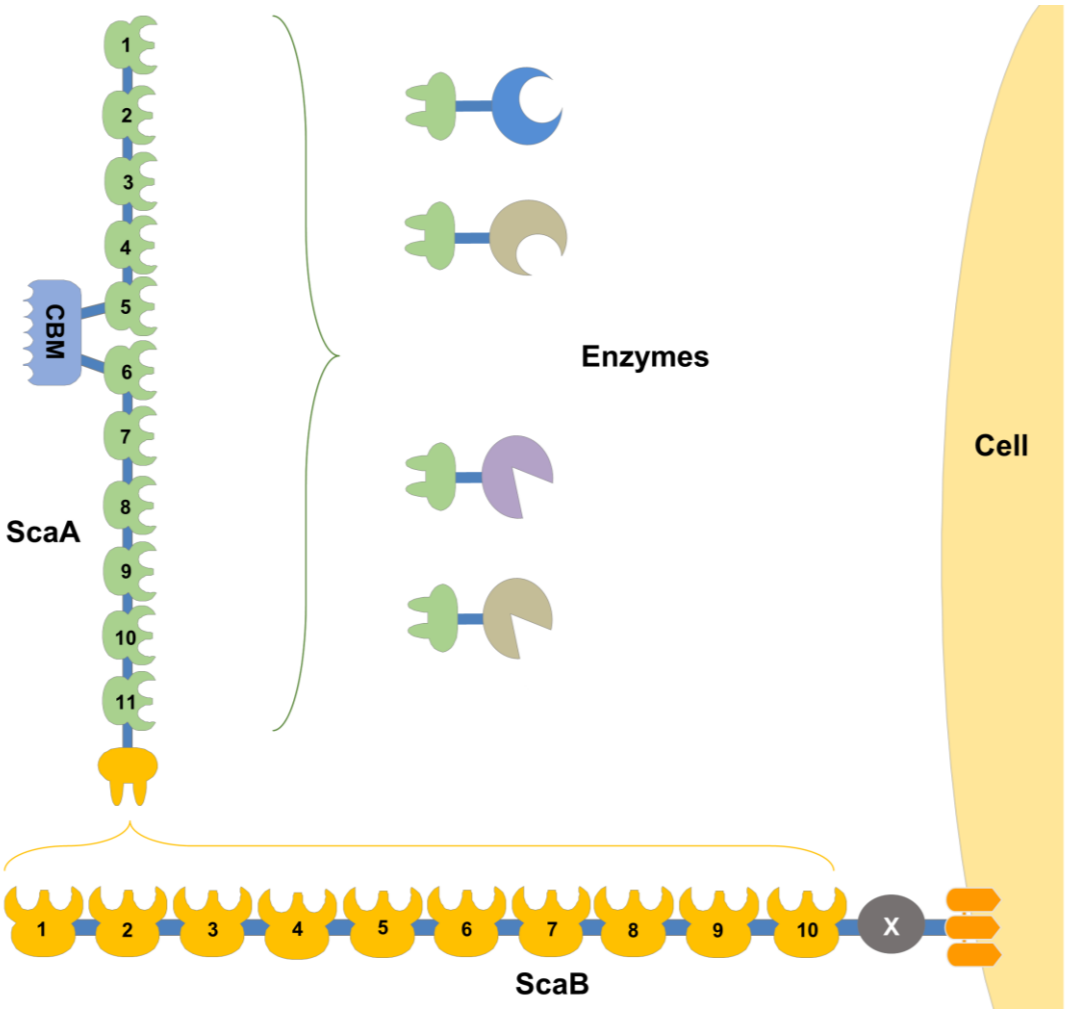
All of the scaffoldins discovered, except for ScaI, appear to contain a signal peptide, suggesting that all cellulosomal components are secreted. Like *C. thermocellum*, *Acetivibrio* cohesins have been classed in two types: type I, of which there are 26 members, and type II, with 15 members. (Cao & Yin, 2014; Dassa *et al.*, 2012). However, the cohesin type does not necessarily indicate its binding specificity to a given dockerin. More recent re-mining of this cellulosome-producing genome has revealed a further 16 putative cohesins in 15 proteins, falling typically within the type I and type II clusters, although phylogenetically they have longer branches to previously published cohesins (Cao & Yin, 2014). The genome of *A. cellulolyticus* has proved to be particularly enriched with dockerin-containing genes, and 143 genes that contain putative dockerin modules were identified. Therefore, the *A. cellulolyticus* contains twice the number of dockerins as any other Clostridial bacteria, with *Ruminococcus flavefaciens* FD-1 being the only currently known genome to contain more dockerin-containing genes (Berg Miller *et al.*, 2009; Dassa *et al.*, 2012). As with the cohesin modules, a further 67 new dockerin domains in 64 proteins were also identified from the *A. cellulolyticus* genome, four of which coexist with carbohydrate-active enzymes and 15 which associate with cohesins or SLH modules (Cao & Yin, 2014). Therefore, it is reasonable to assume that the complete and exact architecture of the

A. cellulolyticus is still to be finalised and more work is required to correctly define and assign the different cellulosomal components both structurally and functionally.

1.4.10.3. *Pseudobacteroides cellulosolvens*

Phylogenetic relationships between cellulolytic bacteria do not necessarily reflect the characteristics of their respective cellulase systems (Ding *et al.*, 2000). Analysis of the enzymes secreted by anaerobic organisms which do not belong to the genus *Clostridium* suggest significant differences in the mechanisms of cellulosome assembly. One such example is the cellulosome produced by the mesophilic anaerobic bacterium *Bacteroides cellulosolvens*, recently reclassified as *Pseudobacteroides cellulosolvens* (Horino, Fujita, & Tonouchi, 2014). This bacterium was shown to exhibit cellulosome-like complexes in both the cell-associated and the extracellular fractions. However, the existence of a multi-enzyme complex was not confirmed until the early 90's. The biochemical evidence in favour of a cellulosome included the presence of cell surface protuberance-like structures, the interaction with a αGal-specific lectin that specifically recognises the S1 subunit of the cellulosome and the cross-reactivity with an anti-cellulosome-specific antibody preparations of *C. thermocellum* (Lamed, Morag (Morgenstern), Mor-Yosef, & Bayer, 1991). Furthermore, extensive structural analysis showed that *P. cellulosolvens* cell surface organelles are extremely similar to the cellulosome-associated oligo-sugars from *C. thermocellum* (Gerwig *et al.*, 1993). These preliminary observations provide evidence for the existence of cellulosomes in *P. cellulosolvens*. However, it was not until the beginning of 2000 that Ding *et al.* (2000) pursued a genetic program to expand on earlier biochemical findings and confirmed that *P. cellulosolvens* produces a defined cellulosome. Their work uncovered a simple yet enlarged cellulosome (Figure 1.16) which comprises two large scaffoldin components similar in arrangement to *C. thermocellum*. Interestingly, sequence analysis indicated that the types of cohesins found on the primary and anchoring scaffoldins appeared to be reversed, as described above. The primary scaffoldin harboured type II cohesins and the anchoring scaffoldin comprised type I cohesins (Ding *et al.*, 2000; Xu, Bayer, *et al.*, 2004).

Figure 1.16The cellulosome system of *Pseudobacteroides cellulosolvens*



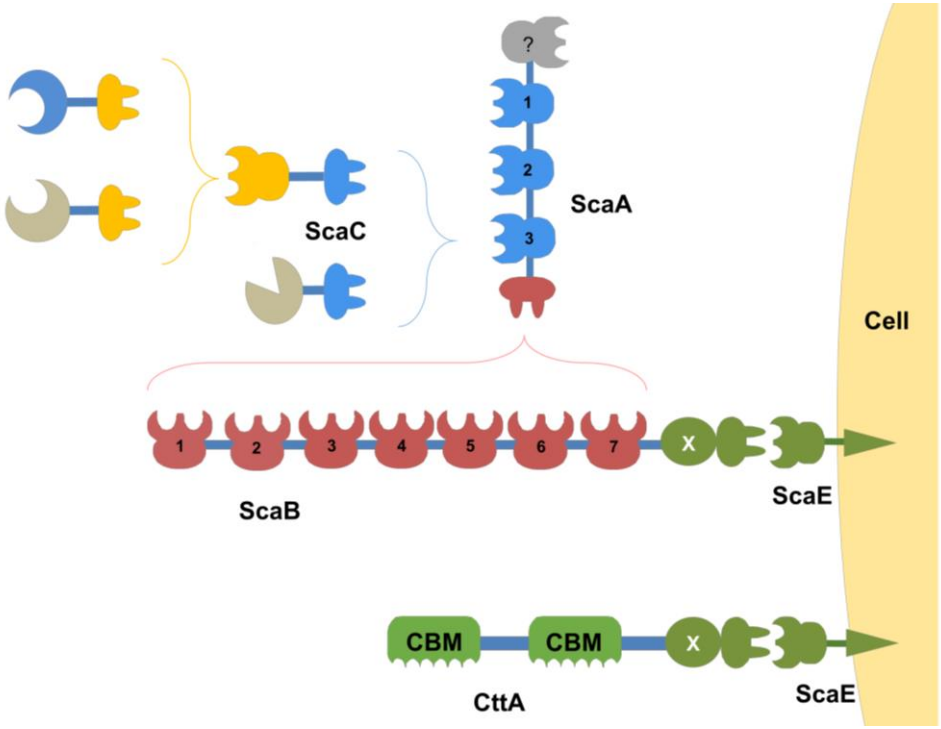
The cellulosome of *P. cellulosolvens* comprises a primary scaffoldin, named ScaA, which contains 11 type II cohesins (light green) with a C-terminal type I dockerin (yellow), and an anchoring scaffoldin, named ScaB, which bears 10 type I cohesins (yellow). This cellulosome also contains a CBM to direct the complex to its substrate and the whole system is tethered to the bacterial cell wall via ScaB's SLH-module (orange).

The number of cohesins on these scaffoldins is very high, with a 10 cohesin anchoring scaffoldin and an 11 cohesin primary scaffoldin. Therefore, a theoretical capacity to incorporate 110 different cellulosomal subunits in a single complex is expressed by *P. cellulosolvens* cellulosome (Bayer, Lamed, White, & Flint, 2008). Additionally the C-terminal dockerin domain of the primary scaffoldin lacks the X-module which is usually associated to type II dockerins located in primary scaffoldins (Ding *et al.*, 2000). These characteristics are largely different to previously discovered cellulosomes, strengthening the idea that nature and ecology play an important part in the evolution of exquisite and original cellulosomes.

1.4.10.4. *Ruminococcus flavefaciens*

To date the most elaborate cellulosome discovered is that from the ruminal bacteria *R. flavefaciens*. Initial studies aiming at clarifying the organization of ruminal cellulosomes were carried out on *R. flavefaciens* strain 17 **(**Figure 1.17). ScaA and ScaB were the first scaffoldins to be identified (Ding *et al.*, 2001). ScaA contains 3 cohesins recognized by enzyme associated dockerins and a C-terminal dockerin that binds to one of the seven ScaB cohesins. ScaB was initially thought to be the anchoring scaffoldin (Rincon *et al.*, 2003), but later it was discovered that it contains a C-terminal dockerin associated to an X-module that binds to another scaffoldin, ScaE, which is the actual anchoring protein (Salama-Alber *et al.*, 2013). This anchoring apparatus is unique as ScaE contains a non-catalytic module covalently attached to the bacterial cell surface through a sortase-mediated mechanism, in place of the more common SLH module (Bayer *et al.*, 2008). Another unique feature is the presence of a monovalent adaptor scaffoldin, ScaC. This scaffoldin binds to ScaA through its C-terminal dockerins and contains a single cohesin, different from the ScaA ones. This allows increasing the repertoire of enzymes that can be integrated into the cellulosome (Rincon *et al.*, 2004). It is possible that this adaptor protein may regulate the assembly of the different enzymatic units onto the complex under different environmental conditions (Rincon *et al.*, 2004; Salama-Alber *et al.*, 2013). In addition, the ruminal cellulosome does not appear to contain the CBM module which has been a key characteristic of all of the scaffoldins from clostridia. However, identification of a fifth anchoring scaffoldin termed CttA revealed the presence of two putative CBMs which may satisfy the role of presenting the cell-attached complex in close proximity to its insoluble substrate (Rincon *et al.*, 2007). Initial characterization of *R. flavefaciens* 17 cohesins suggested sequence and structural differences to the previously described type I and type II cohesin-dockerin complexes, resulting in the cohesin-dockerin interactions of *R. flavefaciens* being classified as type III modules (Ding *et al.*, 2001).

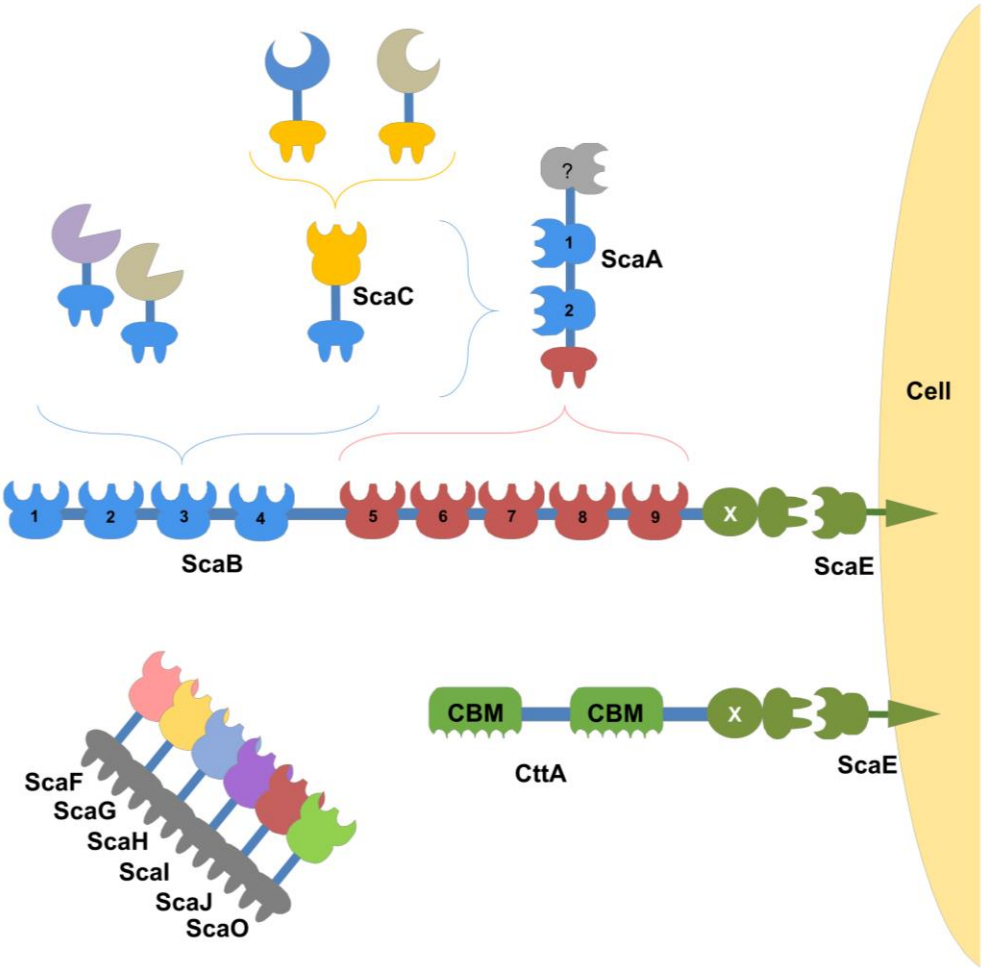
Figure 1.17 Schematic overview of the cellulosome system in *Ruminococcus flavefaciens* strain 17



The cellulosome is characterised by at least four different cohesin–dockerin specificities. The conserved X-Dockerin dyad of ScaB and CttA is bound to the anchoring ScaE cohesin. The seven ScaB cohesins interact with the ScaA dockerin, thereby increasing the number of components that are incorporated into the *R. flavefaciens* cellulosome. The ScaA cohesins bind directly to a group of dockerin-containing enzymes (Cel44A-like). Alternatively, they bind to the ScaC dockerin, whose divergent cohesin recognizes and incorporates into the cellulosome a different set of dockerin-containing enzymes and other components (Bayer *et al.*, 2008).

Recently the genome sequence of *R. flavefaciens* FD-1 strain revealed the most intricate and potentially versatile cellulosomal complex described so far (Figure 1.18). Like strain 17 it possesses scaffoldins ScaA, ScaB, ScaC, ScaE and the CBM bearing CttA, although some of the cellulosomal components have a slightly different architecture. ScaB contains nine cohesins presenting two different specificities: four recognise the dockerins of the catalytic subunits and five bind to ScaA. This primary scaffolding protein, ScaA, is capable of binding a group of dockerin-containing enzymes to its two cohesin domains and as a result, amplifies the number of enzymes in the *R. flavefaciens* cellulosome. ScaC binds to both ScaA and ScaB via its dockerin and, like in strain 17, possesses a single divergent cohesin domain that was show to bind to a different group of dockerins.

Figure 1.18 The complexity of *R. flavefaciens* strain FD-1 cellulosome

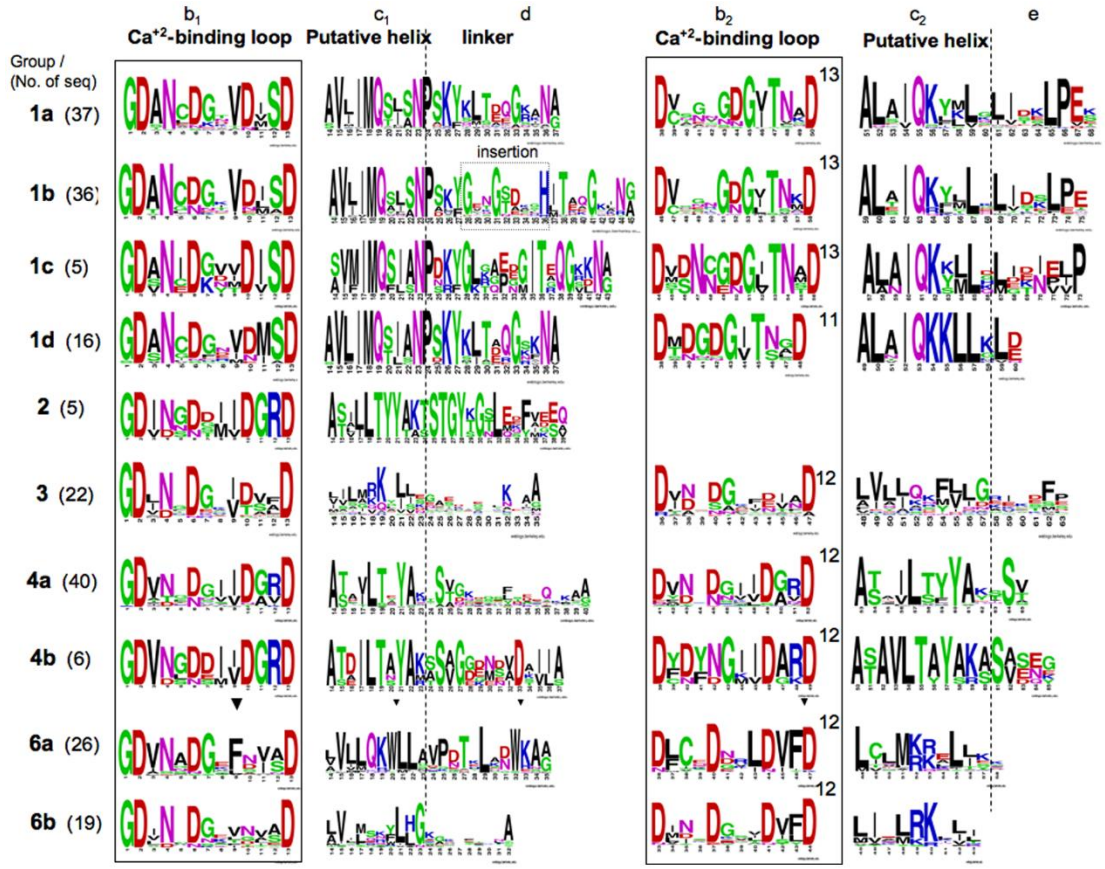


The single cell-surface scaffoldin, ScaE, may bind CttA, which carries two CBMs that mediate the primary anchorage to the plant cell wall or to ScaB. ScaB contains cohesins with two different specificities. One type (red) exclusively interacts with the adaptor scaffoldin ScaA. The other type of ScaB cohesins (blue) binds cellulosomal enzymes or ScaC. In addition, ScaA contains two cohesins that present a similar specificity to the second set of cohesins of ScaB. Like ScaA, ScaC is an adaptor scaffoldin that recognizes a different set of dockerin-containing proteins. Other adaptor scaffoldins, presenting a similar structure to ScaC, but displaying a yet unknown specificity, exist in *R. flavefaciens*. Adapted from Bayer *et al.*, 2008.

Several other monovalent adaptor scaffoldins, similar to ScaC, were also identified in *R. flavefaciens* genome, raising the total number of confirmed cohesins to 19. In addition, *R. flavefaciens* FD-1 strain encodes more than 220 dockerin-containing proteins, the largest number to be recorded in any given cellulosome so far (Bayer *et al.*, 2008; Rincon *et al.*, 2010). Unlike the dockerins of *C. thermocellum* and *C. cellulolyticum* cellulosomes, where the great majority are very similar in their sequences, the dockerin sequences of the *R. flavefaciens* FD-1 cellulosome can be divided into more distinctive groups based on sequence similarity (Rincon *et al.*, 2010) (Figure 1.19). Thus, 6 dockerin groups have been defined in *R. flavefaciens*: group 1 is the largest with 96 members divided into 4 subgroups; group 2 consists of truncated

dockerins possessing only the first repeat; group 3 has 22 members mainly appended to hemicellulases; groups 4 and 6 are each subdivided into 2 subgroups and group 5 has the ScaA dockerin as its single member.

Figure 1.19 Conservation patterns of different dockerin groups from *R. flavefaciens* FD-1.



The 222 dockerins were clustered into groups according to their conserved sequence features, and their sequence logo is presented. Segments along the dockerin modules (b–e at top) are labelled according to (Pagès *et al.*, 1997). The length of the second repeat is marked for each group. (Adapted from (Rincon *et al.*, 2010)

The patterns observed in the dockerin-containing proteins provide another level of complexity to the *R. flavefaciens* FD-1 cellulosome. Apart from a large number of modules with unknown function, catalytic modules (glycoside hydrolases, polysaccharide lyases, carbohydrate esterases and associated CBMs) were particularly associated with only a few select groups (1a, 1b, 3 and 6). Attempts to understand this complexity included inspection of the levels of gene expression, which mainly revealed that multi-modular proteins were mostly up-regulated in cells grown on cellulose versus growth on cellobiose (Rincon *et al.*, 2010). However, it remains ambiguous to simply equate dockerin clusters with their specificities. This question can only be answered through careful and extensive functional studies on the interactions between purified modules, and through the structural analysis of the determinants of binding specificity. Previous

studies revealed that, within *R. flavefaciens*, cellulosome structural organisation varies between different strains, reflecting the complexity of the rumen ecosystem and the diversity of the lignocellulosic substrate (Jindou *et al.*, 2006, 2008). For example, differences in cellulosomal organisation between strains 17 and FD-1 reflect their sampling in different animals, originated in different geographical locations and collected in different time frames (Salama-Alber *et al.*, 2013).

1.4.11. Biotechnological and potential applications for cellulosomes

Providing a structural insight for the specificity displayed by the increasing repertoire of cohesin-dockerin pairs and exploring the dynamic structural features of the scaffoldin subunit is essential for the development of cellulosome based/inspired tools (Smith & Bayer, 2013). Artificial multi-enzyme complexes that mimic cellulosomes were proposed over two decades ago (Bayer *et al.*, 1994), and since, they have been produced extensively, both *in vitro* and *in vivo*. Minicellulosomes and designer cellulosomes have been used both as tools for the study of cellulosome action and as potential replacements for, or extensions of, native cellulosomes for nanobiotechnological applications, notably for the production of biofuels from cellulosic biomass (Fierobe *et al.*, 2005; Moraïs *et al.*, 2012; Stern, Moraïs, Lamed, & Bayer, 2016). In this case, naturally evolved nanomachines could be used as a blueprint for the design, construction and exploitation of tailor-made catalytic multi-protein complexes with precise functions (Fontes & Gilbert, 2010). The hydrolysis of cellulose remains a major limiting factor for the efficient utilisation of lignocellulosic materials (Matano, Hasunuma, & Kondo, 2012). The activities of multiple enzymes, including endoglucanase, exoglucanase, and β-glucosidase, are required to release soluble sugars from cellulose, (Hasunuma *et al.*, 2013) therefore making the use of cellulosomal enzymes an ideal solution. Cellulosomes integrating fungal and bacterial enzymes from non-aggregating systems, displaying particular promise in biomass saccharification, can be generated to improve hydrolytic activities. To broaden cellulosome diversity and increase substrate degradation, these 'external' enzymes, such as β-glucosidases, lytic polysaccharide monooxygenases (LPMOs) or expansins, have been incorporated into designer cellulosomes. This incorporation complemented the complex with novel enzymatic activities that generally resulted in an enhancement of overall activity (Arfi, Shamshoum, Rogachev, Peleg, & Bayer, 2014; Chen *et al.*, 2016; Gefen, Anbar, Morag, Lamed, & Bayer, 2012). Additionally, more than a decade ago it was estimated that the sale of industrial enzymes would reach a market value of approximately 1.6 billion dollars, of which cellulases and associated enzymes represented a significant amount. The potential of cellulosomes and their

association with a vast array of different cellulases and hemicellulases and extreme habitat variability make them an exceptional tool for the bioenergy field (Karmakar & Ray, 2011). The value of the proximity effect in cellulosomes has also been proven transferrable to other novel platforms. By drawing inspiration from cellulosome architecture other structures with an increased number of enzymes in a single complex were designed. These include self-assembled 12-enzyme and 18-enzyme complexes (Mitsuzawa *et al.*, 2009; Moraïs *et al.*, 2010), cellulases that are covalently bound to nanospheres (Blanchette, Lacayo, Fischer, Hwang, & Thelen, 2012), cellulases that are bound to streptavidin and inorganic particles (Kim *et al.*, 2012), and cellulases that are bound to a DNA scaffold (Mori *et al.*, 2013).

1.4.11.1. Other applications

It is well established that inclusion of microbial cellulases and hemicellulases in wheat, barley and rye-based diets for simple-stomach animals, such as broilers, improves the efficiency of feed utilisation, enhances growth and contributes for a better use of low cost feed ingredients (Bedford, 2000). Previous research on cellulosomes and 'designer' cellulosomes has shown that cellulosomal cellulases act together in an enhanced synergistic manner in the degradation of cellulosic substrates. Thus, it is possible to integrate the current knowledge on the mechanisms of cellulosome assembly and CBM function to produce more efficient biocatalysts for feed supplementation (Costa *et al.*, 2014). Due to the modular nature of the cellulosome, its components have been proposed for use in many biotechnological applications (Bayer *et al.*, 2004), especially together with other affinity systems (such as protein A, antibodies and lectins). The potential of employing the cohesin-dockerin interaction to support the development of innovative techniques with various purposes has attracted the attention of many groups. These include immunoassays and blotting, microarray technology, drug delivery, localization and cytochemistry, isolation and immobilization, affinity chromatography and cell separation (Bayer *et al.*, 1994). The high-fidelity, high-affinity cohesin–dockerin interaction could be used as a partner in other affinity-based applications, such as those that involve avidin biotin (Wilchek, Bayer, & Livnah, 2006). A high sensitivity and selectivity self-assembling biosensor based on the cohesin–dockerin interaction was even recently developed (Hyeon, Kang, & Han, 2014). The near-irreversible cohesin–dockerin binding (Mahalingeswara Bhat & Wood, 1992; Mori, 1992), is a major limitation in some of these techniques. Nevertheless, engineering of the *C. thermocellum* dockerin allowed decreasing its affinity for the cohesin, which enabled its use as an affinity tag for protein purification (Demishtein, Karpol, Barak, Lamed, & Bayer, 2010; Sakka *et al.*, 2011). Karpol *et al.* (2009) proposed a protein affinity tag based on the cohesin-

dockerin interaction combined with the binding of a CBM to cellulose matrices. The affinity purification system consisted of a recombinant *C. thermocellum* scaffoldin fragment that included a CBM and adjacent cohesin, such that the cohesin bound to a mutated dockerin and its host protein and the CBM bound the cellulose column. Effectively, the mutated dockerin retained high levels of affinity for its complementary cohesin, yet enabled complete dissociation of the dockerin from the CBM-cohesin affinity column upon purification (Karpol *et al.*, 2009). Later, Demishtein *et al* (2010) designed a new protein affinity tag, in which specific residues of the *C. thermocellum* Cel48S dockerin were mutated, so that the binding affinity for its cohesin partner was reduced. Besides proving a very efficient and robust alternative system for affinity chromatography, the affinity tag was also shown to have little effect on the properties of the proteins tested, including enzymes. Furthermore, the relatively inexpensive costs of cellulose-based affinity columns together with their reusable nature and high capacity makes this system very attractive for affinity protein purification (Demishtein *et al.*, 2010). The utilisation of enzymes and cellulosomes are also being considered as valuable alternatives for the usage of agro-wastes and organic pollutants as a renewable resource, reducing he consequent environmental pollution (Bayer, Lamed, & Himmel, 2007; Karmakar & Ray, 2011). Nevertheless, only a few of the many possible research applications have been explored and, as our knowledge of these multi-enzyme complexes increases, so does the potential for future innovation.

1.5. Objectives

This work aims to clarify a series of unresolved questions concerning the structure, function and relevance of novel cohesin-dockerin interactions belonging to the cellulosome complexes of *Ruminococcus flavefaciens* and *Acetivibrio cellulolyticus*. Due to the relatively scarce data about the more recently described type III cohesin-dockerin complexes, the bulk of the work was focused in identifying and thoroughly characterizing the interaction of novel type III complexes from *Ruminococcus flavefaciens.* Objectively, the main goals of this project can be summarized as follows:

- To identify novel cohesin-dockerin interactions and the multiple Coh-Doc specificities present in *R. flavefaciens* cellulosome, in order to fully decipher the mechanism of cellulosome organization in this ruminal bacterium (**Chapter 2**).
- To determine the structural basis for the novel cohesin-dockerin interaction responsible for enzyme recruitment into *Ruminococcus flavefaciens* cellulosome via an adaptor scaffoldin (**Chapter 3**).

- To determine the structural basis for the novel type III cohesin-dockerin interaction responsible for the direct recruitment of enzymes into *Ruminococcus flavefaciens* through the binding to the primary scaffoldin (**Chapter 4**).
- To determine the structural basis for the novel type III cohesin-dockerin interaction responsible for the articulation between the primary and the primary/adaptor scaffoldin from *Ruminococcus flavefaciens* cellulosome (**Chapter 5**).
- To determine the structure of a novel type I cohesin-dockerin complexes of *A. cellulolyticus* in order to investigate the molecular mechanisms of cross-specificity in type I complexes involved in enzyme recruiting and cell wall anchoring (**Chapter 6**).

1.5.1. Thesis outline

To properly address and discuss the above-mentioned objectives, this thesis has been divided into seven chapters. The first Chapter consists of a detailed state of the art review. Several concepts concerning plant cell wall composition, degradation, complexity and functionality are addressed. This is followed by an in-depth discussion of cellulosomes, including their diversity of components encompassing cohesin-dockerin complexes, other non-catalytic and catalytic modules and general architectures. A brief discussion about the current and potential future biotechnological applications of cellulosomes is also included. Chapters two to four are adapted from scientific papers already published in international peer reviewed journals. Chapters five and six are also based on scientific manuscripts which are currently in preparation for submission. Finally, the last Chapter integrates the results presented in each of the previous Chapters with a complete discussion and conclusion.

With the exception of the ELISA and Microarray protocols described in chapters 2 and 4, all of the work described troughout this thesis has been entirely developed by the student. The student also contributed in the elaboration of all manuscripts from which the several chapters were adapted from.

Chapter 2

The organization of *Ruminococcus flavefaciens* cellulosome

Complexity of the *Ruminococcus flavefaciens* FD-1 cellulosome reflects an expansion of family-related protein-protein interactions

Vered Israeli-Ruimy[1,*], Pedro Bule[2,*], Sadanari Jindou[3], Bareket Dassa[1], Sarah Moraïs[1], Ilya Borovok[3], Yoav Barak[1,4], Michal Slutzki[1], Yuval Hamberg[1], Vânia Cardoso[2], Victor D. Alves[2], Shabir Najmudin[2], Bryan A. White[5], Harry J. Flint[6], Harry J. Gilbert[7], Raphael Lamed[3], Carlos M.G.A. Fontes[2] and Edward A. Bayer[1**]

[1]Department of Biomolecular Sciences, The Weizmann Institute of Science, Rehovot, Israel. [2]CIISA – Faculdade de Medicina Veterinária, Universidade de Lisboa, Avenida da Universidade Técnica, 1300-477 Lisboa, Portugal. [3]Department of Molecular Microbiology and Biotechnology, Tel Aviv University, Ramat Aviv, Israel. [4]Chemical Research Support, The Weizmann Institute of Science, Rehovot, Israel. [5]Department of Animal Sciences, Institute for Genomic Biology, University of Illinois at Urbana–Champaign, Champaign, IL, USA. [6]Microbiology Group, Rowett Institute of Nutrition and Health, University of Aberdeen, Foresterhill, Aberdeen, Scotland, UK. [7]Institute for Cell and Molecular Biosciences, Newcastle University, The Medical School, Newcastle upon Tyne NE2 4HH, UK. ***Equal contribution.*** ** ***Corresponding author.***

Adapted from *Scientific Reports*, 2017 Feb 10;7:42355 (Israeli-Ruimy *et al.*, 2017)

Abstract

Protein-protein interactions play a vital role in cellular processes as exemplified by assembly of the intricate multi-enzyme cellulosome complex. Cellulosomes are assembled by selective high-affinity binding of enzyme-borne dockerin modules to repeated cohesin modules of structural proteins termed scaffoldins. Recent sequencing of the fiber-degrading *Ruminococcus*

flavefaciens FD-1 genome revealed a particularly elaborate cellulosome system. In total, 223 dockerin-bearing ORFs potentially involved in cellulosome assembly and a variety of multi-modular scaffoldins were identified, and the dockerins were classified into six major groups. Here, extensive screening employing three complementary medium- to high-throughput platforms was used to characterize the different cohesin-dockerin specificities. The platforms included (i) cellulose-coated microarray assay, (ii) enzyme-linked immunosorbent assay (ELISA) and (iii) *in-vivo* co-expression and screening in *Escherichia coli*. The data revealed a collection of unique cohesin-dockerin interactions and support the functional relevance of dockerin classification into groups. In contrast to observations reported previously, a dual-binding mode is involved in cellulosome cell-surface attachment, whereas single-binding interactions operate for cellulosome integration of enzymes. This sui generis cellulosome model enhances our understanding of the mechanisms governing the remarkable ability of *R. flavefaciens* to degrade carbohydrates in the bovine rumen and provides a basis for constructing efficient nano-machines applied to biological processes.

2.1. Introduction

Cellulose degradation has long been focus of many studies in the fields of renewable energy and waste management (Bayer *et al.*, 2007; Himmel *et al.*, 2007; Himmel & Bayer, 2009; Ragauskas *et al.*, 2006; Schubert, 2006). Cellulose is the most abundant naturally occurring organic material, yet its recalcitrant nature renders it largely unavailable for extensive biodegradation (Meng & Ragauskas, 2014; O'Sullivan, 1997). Herbivores feed on plants as a sole carbon source. The rumen is a highly populated and competitive ecological niche, where a complex and diversified repertoire of microbial enzymatic systems participate in deconstruction of recalcitrant carbohydrates through molecular mechanisms which remain poorly understood (Flint & Bayer, 2008; Flint, Bayer, Rincon, Lamed, & White, 2008; White, Lamed, Bayer, & Flint, 2014). An enormous concentration of archaea, protozoa, fungi and bacteria colonize the rumen. Although only a small fraction of these microbes are directly engaged in fiber degradation, they all benefit from the metabolic by-products. Dominant rumen species identified as primary degraders of crystalline forms of polysaccharides are fibrolytic bacteria, namely *Fibrobacter succinogenes*, *Ruminococcus flavefaciens* and *Ruminococcus albus* (Flint *et al.*, 2008; Hespell, Akin, & Dehority, 1997).

R. flavefaciens is a Gram-positive, anaerobic bacterium of the Firmicutes phylum. It is the only known bacterium in the rumen shown to possess a definitive cellulosome, i.e., a discrete multi-enzyme complex specialized in the breakdown of cellulose and associated plant cell-wall polysaccharides (Aurilia *et al.*, 2000; Ding *et al.*, 2001; Kirby, Aurilia, McCrae, Martin, & Flint,

1998). The cellulosome complex carries three fundamental features. Firstly, cellulosome assembly results from the incorporation of cellulosomal enzymes, e.g. glycoside hydrolases (GH), carbohydrate esterases (CE), and polysaccharide lyases (PL), into structural scaffoldin subunits through high-affinity interactions between cohesin and dockerin modules. Cohesins are modular components of scaffoldins, whereas dockerins are borne by individual cellulosomal enzymes that are integrated into the complex through interaction with the cohesins (Bayer *et al.*, 2004, 1994; Shoham, Lamed, & Bayer, 1999; Tokatlidis, Dhurjati, & Béguin, 1993). Secondly, cellulosomes are anchored to the cell-surface through a mechanism, which may take place either covalently through enzymatic mediation or non-covalently via a specialized module (Navarre & Schneewind, 1994; Rincon *et al.*, 2005; Zhao *et al.*, 2006). Thirdly, a non-catalytic substrate (carbohydrate)-binding module (CBM) attaches the entire complex to cellulose (Poole *et al.*, 1992; Rincon *et al.*, 2007; Shoseyov *et al.*, 1992). Cellulosomes thus present a complex functional machinery of great environmental flexibility and adaptation, gained by the many possible arrangements of its modular components, as dictated by the deployment of different cohesin-dockerin pairs.

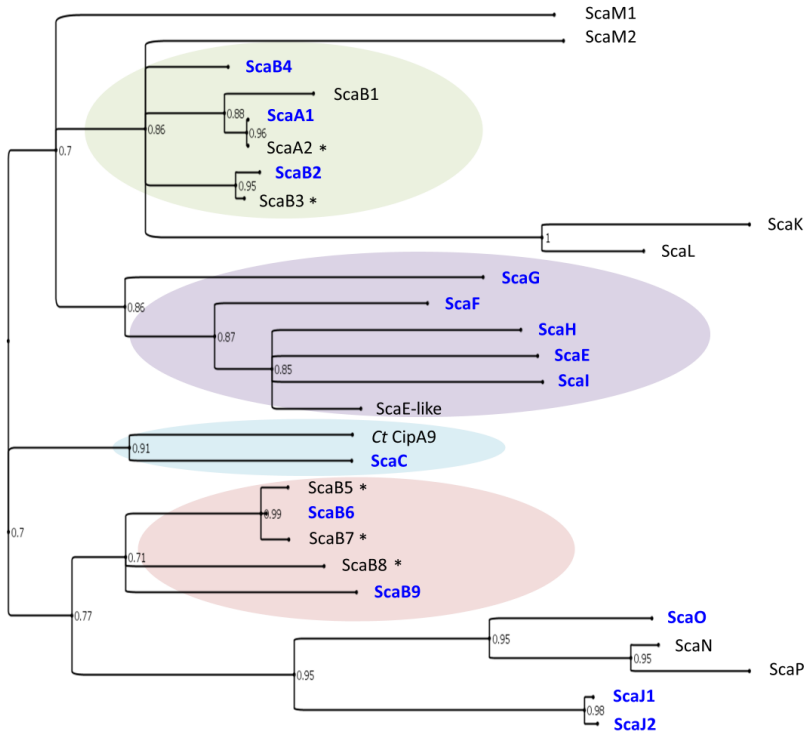
The profile of *R. flavefaciens* presents a multiplicity of rumen strains, both similar to and phylogenetically distinct from previously discovered strains (Brulc *et al.*, 2011; Jindou *et al.*, 2008; Krause, Bunch, Smith, & McSweeney, 1999). All members of this species have been shown to possess a scaffoldin-encoding sca gene cluster, and thus appear to synthesize a cellulosome. The locus encodes scaffoldins ScaC, ScaA, ScaB and ScaE, as well as a CttA protein, believed to include two consecutive carbohydrate-binding modules (CBMs) (Jindou *et al.*, 2008). *R. flavefaciens* strains have in common an enzyme-integrating subunit, ScaB, which carries a C-terminal X module-dockerin (XDoc) dyad that in turn recognizes the single cohesin of the surface-anchored scaffoldin, ScaE (Dassa *et al.*, 2014; Jindou *et al.*, 2006). ScaE is covalently linked to the bacterial envelope via an LPXTG motif, mediated by the enzyme sortase; thus the entire multi-enzymatic cellulosome assembly is bound to the bacterial cell surface (Rincon *et al.*, 2005). In addition, the ScaE cohesin also binds the CttA protein, which, like ScaB, carries a C-terminal XDoc dyad and would thus promote substrate targeting and bacterial adhesion via its CBM modules, thereby initiating deconstruction of the cellulosic substrate. Moreover, the XDoc modules of CttA and ScaB include three unique insertions within their structure, recently proposed to mechanically support the bulky complex and its anchoring to the cell via ScaE (Rincon *et al.*, 2007; Salama-Alber *et al.*, 2013; Venditto *et al.*, 2015).

The main difference among the various *R. flavefaciens* strains is the number and types of cohesins borne by the main ScaB subunit and their specificity(ies) towards cognate dockerins.

In strain FD-1, ScaB harbors nine cohesins, four of which (cohesins 1-4) are similar in sequence to the two ScaA cohesins, whereas the others (cohesins 5-9) bind to the unique ScaA dockerin. Previous studies have demonstrated variation in scaffoldin recognition by different classes of enzymes in *R. flavefaciens*. Some enzymes bind directly to ScaA and ScaA-like cohesins on ScaB, whereas others bind via the intermediary ScaC cohesin (Rincon *et al.*, 2004), which acts as a selective "adaptor" scaffoldin that alters enzymatic composition of the cellulosome. These divergent interactions and their significance towards cellulosome organization are presumably governed by the sequence and consequent specificity of the enzyme-borne dockerin.

In the past, cohesins were distinguished into three types: I, II and III, based on phylogeny of the primary sequences. Likewise dockerins that interacted with these cohesins were regarded as the same type. The cohesins and dockerins of *R. flavefaciens*, belong to type III albeit with considerable internal diversity (Figure 2.1). Curiously, the ScaC cohesin of *R. flavefaciens* maps onto a divergent phylogenetic branch, closer to those of the clostridial type-I cohesins (Figure 2.1). Only a single enzyme-borne dockerin, CE3B, a family 3 carbohydrate esterase, had been shown previously to bind to the ScaC cohesin, whereas the general binding specificity and range of proteins it serves to integrate remains obscure (Jindou *et al.*, 2006).

Figure 2.1 Phylogenetic tree of the *R. flavefaciens* FD-1 cohesins.



Cohesins B1-B4 are located together in the tree (mint green), consistent with reports in the literature, i.e., closer to one another and to ScaA cohesins than to cohesins B5-9 (pink). Cohesins selected for the microarrays assay are shown in blue font. *C. thermocellum* ScaA cohesin 9 (CtScaA9) was used as a marker to represent type I cohesins. Note that the cohesin borne by the ScaC adaptor scaffoldin is

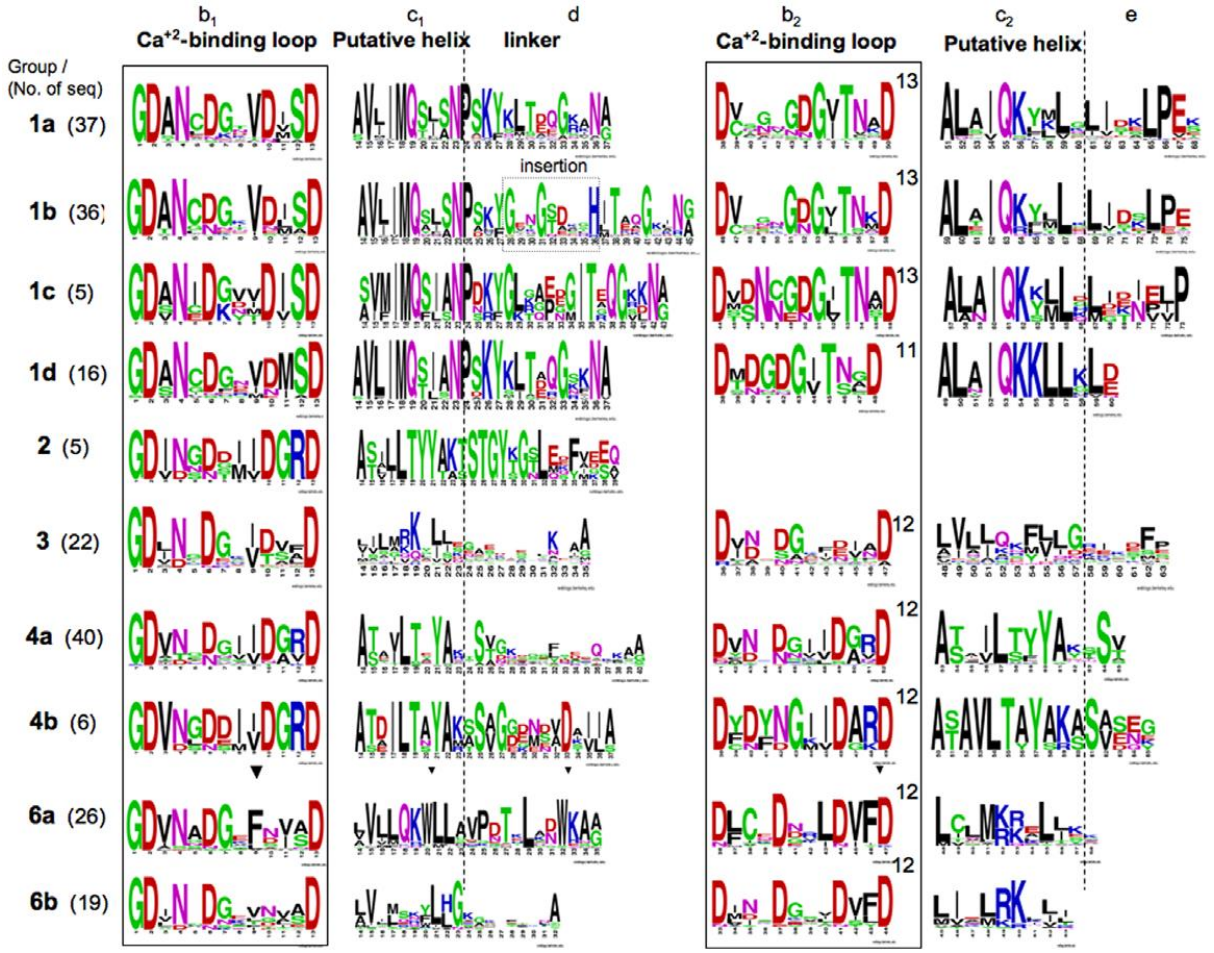
associated with the type I cohesins (powder blue) and thus diverges from the type III *R. flavefaciens* cohesins. Another cluster of cohesins is marked in lavender. Asterisks (*) indicate cohesins tested in both complementary ELISA and non-denaturing PAGE studies. The tree was generated using PhyML software (http://www.atgc-montpellier.fr/phyml) and processed using FigTree v1.4.2 (http://tree.bio.ed.ac.uk/software/figtree). Bootstrap threshold of 0.7 is presented.

A draft genome of *R. flavefaciens* strain FD-1 has been published, revealing 223 dockerin-containing ORFs (Dassa *et al.*, 2014; Rincon *et al.*, 2010). This is triple the number of cellulosomal components observed for clostridial species, rendering the *R. flavefaciens* cellulosome the most intricate described to date. The bacterium comprises an abundant repertoire of catalytic and CBM modules frequently organized in multi-modular protein architectures (Berg Miller *et al.*, 2009). The presence of numerous genes encoding for highly complex multi-modular hemicellulases is particularly striking. Nevertheless, many of the dockerin-bearing parent proteins appear to be unrelated to traditional cellulosome activities, with predicted functions, such as serpins, peptidases, LRR (leucine-rich repeats) proteins and transglutaminases.

The dockerin sequences of *R. flavefaciens* FD-1 exhibit great sequence diversity that ranges between 20%-98% homology. This has led to their categorization into six distinct major groups and eleven sub-groups, based on sequence conservation patterns, secondary structural elements and postulated Ca^{2+}-binding and cohesin-recognition residues (Rincon *et al.*, 2010). Each group exhibits unique and recognizable features, such as the presence of an atypical number of conserved residues in the second repeat. Some dockerins resemble known dockerins (groups 3 and 6) and some are exclusive to *R. flavefaciens* FD-1 (groups 1-2). The conservation pattern of the group classification of the *R. flavefaciens* dockerins from (Rincon *et al.*, 2010) is available in Figure 2.2.

Nonetheless, the functional significance of dockerin classification into these different groups remains unclear. It was thus uncertain whether the dockerin grouping reflected variation in ligand (cohesin) specificity or stability factors within the context of their parent proteins. To clarify these issues, the present report describes a combined experimental approach to investigate cellulosome configuration in *R. flavefaciens* strain FD-1.

Figure 2.2 Conservation patterns of different dockerin groups from *R. flavefaciens* FD-1



The 222 dockerins were clustered into groups by (Rincon *et al.*, 2010), according to their conserved sequence features, and their sequence logo is presented. The length of the second repeat is marked for each group.

2.2. Experimental Procedures

2.2.1. Protein microarrays

PCR primers were designed to amplify different dockerin- and cohesin-containing genes from the gDNA of *R. flavefaciens* strain FD-1. A full list of primers is available in Table S2.1 (Annexes). Constructs were prepared by standard molecular techniques. Briefly, dockerin inserts were cloned into the pET9d plasmid, supplemented with an N-terminal xylanase T-6 module, derived from *Geobacillus stearothermophilus*, and His-tag (Handelsman *et al.*, 2004). Cohesins were cloned into the pET28a plasmid, supplemented with an N-terminal family-3a carbohydrate-binding module (CBM3a) from ScaA of *C. thermocellum* (Barak *et al.*, 2005). PCR reactions were conducted with Phusion DNA polymerase and DNA restriction reactions with Fermentas Fast Digest enzymes (ThermoFisher Scientific, Waltham, MA, USA). Preparation of xylanase-fused X-dockerins (XynDocs) of ScaB and CttA, and ScaE CBM-fused cohesin (CBM-Coh) were described earlier (Barak *et al.*, 2005; Handelsman *et al.*, 2004).

Plasmid DNA was extracted using a QIAprep Spin Miniprep Kit (Qiagen GmbH, Hilden, Germany). DNA integrity was confirmed by sequencing. *E. coli* strain BL-21(λDE3) pLysS cells were used to over-express XynDoc and CBM-Coh fusion proteins as described (Barak *et al.*, 2005; Haimovitz *et al.*, 2008). XynDocs were incubated at 16 ºC for 16 h; CBM-Cohs at 37ºC for 3 h, post induction. To normalize protein levels, whole-cell extracts of over-expressed CBM-Coh and XynDoc fusion proteins were examined on SDS-PAGE gels (12 %), using ScaB2 CBM-Coh and ScaC XynDoc as standards, respectively. Rabbit *α*-Xyn T6 and *α*-CBM antibodies were produced as described earlier (Morag *et al.*, 1995) and labeled with Cy3 and Cy5 mono-reactive dyes, respectively (GE Healthcare Bio-Sciences AB Uppsala, Sweden). Conjugates were dialyzed against Tris-buffered saline (137 mM NaCl, 2.7 mM KCl, 25 mM Tris pH 7.4; TBS). The procedure to evaluate the Coh-Doc interactions upon cellulose-coated microarrays was followed as documented (Hamberg *et al.*, 2014) with the following modification: Cohesin crude extracts were diluted 3-fold in TBS and printed in quintuplicate on cellulose slides. Nonspecific binding events were assessed using an unrelated cohesin (ScaA-Coh3 of *C. thermocellum*), and crude extracts of transformed *E. coli* cells harboring an empty pET28a-CBM3a vector as negative controls. Fluorescent signal intensities were measured using ImageJ (http://imagej.nih.gov/ij). Following assignment of Cy3/Cy5 ratios, interactions were normalized according to a control XynCBM Cy3/Cy5 ratio of 1.

2.2.2. Enzyme-linked immunosorbent assay (ELISA).

ELISA was conducted using XynDoc/CBM-Coh fusion protein pairs to evaluate cohesin-dockerin interactions as described (Barak *et al.*, 2005). DNA isolation/cloning and protein expression/purification were as above.

2.2.3. *In-vivo* screening of cohesin-dockerin interactions

Genes encoding 45 representative *R. flavefaciens* dockerins (Table 2.1) were cloned using the Gateway recombination cloning technology (ThermoFisher Scientific). Sequences were amplified by PCR using *R. flavefaciens* FD-1 genomic DNA as template and primers with engineered ends that allow site-specific recombination without need for restriction enzymes (Table S2.2 Annexes). Amplified genes were inserted into pDONR201 entry vector and subsequently into two distinct protein expression destination vectors, pDest17 and pETG-20A, according to the manufacturer's protocol. In both expression vectors the genes are under T7 promoter control. pDest17 allowed fusion of N-terminal His-tags onto dockerins, while pETG-20A allowed fusion of N-terminal thioredoxin A with internal His-tag for increased stability/solubility. Genes encoding ten diverse cohesins were cloned into Novagen pCDFDuet vector (Merck Millipore, Darmstadt, Germany) using traditional restriction-enzyme methods.

Cohesin sequences were isolated via PCR using genomic DNA as template. Primers incorporated 5'-NcoI and NheI restriction sites and 3'-XhoI (Table S2.3 Annexes). Digesting PCR products with NcoI and XhoI allowed cloning in the pCDFDuet vector, whereby engineered recombinant proteins contained no His-tag. Digesting with NheI and XhoI allowed cloning in pET21a to produce cohesins with C-terminal His-tags. Genes were sequenced in both directions to confirm that no mutations had occurred during amplification.

To detect novel cohesin-dockerin complexes, the cohesins and dockerins were co-expressed *in vivo* and co-purified by IMAC using the His-tagged dockerin. Initially, *E. coli* BL21 (DE3) cells were transformed with pCDFDuet poly-histidine tag lacking cohesins. Cohesin-harboring *E. coli* strains were made competent following conventional protocols. Each *E. coli* pCDFDuet-cohesin was retransformed with pDest17 and pETG20-A derivatives encoding the 45 dockerins. The two plasmids, pCDFDuet and either pDest17 or pETG20-A, have compatible origins of replication and independent antibiotic selection, leading to a total of 720 different recombinant *E. coli* strains expressing 720 cohesin-dockerin combinations. Cells transformed with the two plasmids were used to inoculate 5 ml of LB media with Ampicillin and Streptomycin in deep-well plates and grown to OD600 of approximately 0.4-0.6. Expression was induced by 1 mM of IPTG, and cells were grown an additional 16 h at 19ºC before harvesting. Cell pellets were then re-suspended in 1 ml lysozyme-containing buffer (8.4 mM 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid (HEPES), pH 7.5, 10 mM imidazole, 167 mM NaCl, 0.83 mM $CaCl_2$ and 0.25 mg/mL lysozyme) and kept at -80ºC for 1 h. IMAC was performed in 96-well plates using a manifold vacuum system (Merck Millipore, Darmstadt, Germany). Purified samples were then subjected to 14% SDS-PAGE, and visualization of either two (cohesin + His tagged dockerin complex), one (His-tagged dockerin) or no bands (no expression) was annotated.

For the non-denaturing PAGE assays, cohesins and dockerins were expressed and purified independently. pET21a plasmid derivatives encoding cohesins were used to transform BL21 (DE3) cells. The cells were used to inoculate 200-ml LB media with Ampicillin and grown to OD600 of approximately 0.4-0.6. Expression was induced by IPTG as above. Cells were harvested and kept at -20º C for 1 h, lysed by ultra-sonication in 50 mM HEPES buffer, pH 7.5, containing 1-M NaCl, 5 mM $CaCl_2$, 10 mM imidazole. Protein purification was performed through IMAC in 1-ml His GraviTrap gravity flow columns (GE Healthcare, Little Chalfont, Buckinghamshire, UK). Dockerins were expressed in deep-well plates and purified with the manifold vacuum system. Each cohesin was incubated with each dockerin (25 μM of each module in elution buffer) for 30 min. Samples of the isolated dockerins, cohesins and respective mixtures were examined by non-denaturing PAGE.

2.2.4. Isothermal titration calorimetry (ITC)

Affinity and thermodynamics of representative cohesin-dockerin interactions was evaluated by ITC. Recombinant cohesins (pET21a vector; containing a His-tag) and dockerins (pETG20A vector; containing an N-terminal thioredoxin-His tag) were produced separately and purified by IMAC. Proteins were buffer exchanged by PD-10 Sephadex G-25M gel filtration (GE Healthcare) columns into 50-mM Na-HEPES buffer, pH 7.5, containing 2 mM $CaCl_2$ and 0.5 mM tris(2-carboxyethyl)phosphine (TCEP). Briefly, thioredoxin-fused dockerins (20-30 μM) were stirred at 307 rpm in the reaction cell, injected with 10-μL aliquots of 80-180 μM cohesin solution at 220-s intervals. Titrations were performed at 308.16 K. Integrated heat effects, after correction for heats of dilution, were analyzed by nonlinear regression using a single-site model (Microcal ORIGIN version 7.0, Microcal Software, Northampton, MA). The fitted data yielded the association constant (K_A) and enthalpy of binding (ΔH). Other thermodynamic parameters were calculated by the standard thermodynamic equation: $\Delta RTlnKA=\Delta G=\Delta H-T\Delta S$.

2.2.5. Alanine-scanning assay

The two-fold symmetry observed in some group-4 dockerin sequences renders them similar to those of type I, previously shown to exhibit dual-binding mode (Carvalho *et al.*, 2007; Pinheiro *et al.*, 2008). Consequently, putative group-4-dockerin recognition residues of cysteine peptidase (ZP_06142181) and ScaH (ZP_06142361) were chosen for alanine-scanning, based on sequence similarity to ScaB and CttA XDocs in their presumed cohesin-recognition residues (Figure 2.3). To substitute two amino acids simultaneously, overlap-extension PCR was conducted (Mechaly *et al.*, 2001). Thus, in two sequential PCR reactions exploiting two sets of primers (Table S2.4 Annexes), double Ala mutations were introduced into the first Ca^{2+}-binding loop and the third helix of the dockerins instead of the original Gly-Arg residues (positions 10-11 and 48-49). The resultant three variants included a mutant carrying Ala-Ala in positions 11-12 of the first dockerin repeat (mutant 1), a mutant carrying Ala-Ala in positions 48-49 (mutant 2) of the second dockerin repeat and a variant harboring both sites of mutations (double mutant). Mutated XynDocs were expressed and purified by IMAC. Interaction between each XynDoc and CBM-CohE was evaluated by ELISA and further processed as described (Barak *et al.*, 2005).

Figure 2.3 Alignment of the dockerins belonging to groups 4 and 2.

Group-4 dockerins exhibit an atypical two-fold symmetry that resembles modules of type I rather than type III, prevalent in *R. flavefaciens*. The dominant cohesin-recognition residues at positions 10-11 and 17-18 of the two repeats are Gly/Ala-Arg/Ile/Val and Leu-Thr/Ser, respectively. Interestingly, the three dockerins of group-2 (marked in red), comprising the first Ca^{2+}-binding loop-helix motif alone, are remarkably conserved with respect to the canonical 1st helix-loop segment of group-4 dockerins. The alignment was performed in Clustal Omega (http://www.ebi.ac.uk/Tools/msa/clustalo/). Dockerins are numbered according to Table 2.1. Note the insertions of group-4 ScaB and CttA XDocs that are absent in other members of the group. These insertions have been shown to form supporting buttresses that interact with the upstream X-module of the above-mentioned scaffoldins (Rincon *et al.*, 2007; Salama-Alber *et al.*, 2013). Residues involved in Ca^{2+}-binding are designated in cyan while residues involved in cohesin recognition are highlighted in yellow.

2.3. Results

2.3.1. Selection of representative cohesin and dockerin modules

Past studies have predicted 223 genomically encoded dockerin-bearing proteins in *R. flavefaciens* (Rincon *et al.*, 2010). Taken together with the 29 predicted cohesin modules (Dassa *et al.*, 2014), a theoretical matrix of 6467 potential cohesin-dockerin interactions was generated. In this work, we accumulated data using three complementary experimental platforms to identify interacting cohesin and dockerin pairs that may shape cellulosomal architecture and enzyme composition in *R. flavefaciens* FD-1. Dockerin modules were selected to represent the previously established bioinformatic sequence diversity. Table 2.1, provides a list of the 77 dockerins selected for recombinant production and subsequent testing within the different experimental platforms. The selected dockerins originated from all of the different groups and subgroups (Rincon *et al.*, 2010) as designated in Table 2.1. The nature of the parent protein was also considered in dockerin selection. Thus, some dockerins belong to proteins bearing typical plant cell wall-degrading catalytic modules (e.g., various GH and CE families) while others are part of proteins containing structural or functional components (e.g., CBM, predicted cohesin-bearing scaffoldins, serpins and LRR motifs). In addition, dockerins belonging to proteins whose expression was upregulated by growth on cellulose were also targeted (Berg Miller *et al.*, 2009) (e.g., Doc 11-13, Table 2.1).

Table 2.1 Summary of interacting *R. flavefaciens* FD-1 cohesin and dockerin modules depicted by the various strategies used in this work: Cellulose-coated microarrays, ELISA, and in-vivo screening followed by non-denaturing PAGE.

	Accession no.	Group No.	Cohesin / Architecture of parental-enzyme	A1	B2	B3	B4	B5-B9	C	E	G	H
1	ZP_06141990	1a	UNK-**Doc**	+		+		-	-	-		-
2	ZP_06142678		GH9-CBM3-**Doc**	+	+	+	+	-	-	-	-	-
3	ZP_06143384		GH44-**Doc**	+		+		-	-	-	-	-
4	ZP_06143935		LRR-**Doc**	+		+		-	-	-		-
5	ZP_06144449		UNK-CE12-CBM13-**Doc**-CBM35-CE12	+		+		-	-	-		-
6	ZP_06145345		UNK-**Doc**	+		+		-	-	-		-
7	ZP_06145412		LRR-**Doc**	+		+		-	-	-		-
8	ZP_06145411		GH5-**Doc**	+		+		-	-	-	-	-
9	ZP_06145755		GH5-**Doc**	+		+		-	-	-	-	-
10	ZP_06144897		UNK-**Doc**	+		+		-	-	-	-	-
11	ZP_06142769		GH11-CBM22-GH10-**Doc**-CBM22-CE4	+	+		-	-	-	-	-	-
12	ZP_06142857		GH11-CBM22-**Doc**-GH11-CE3	+	+		-	-	-	-	-	-
13	ZP_06142983		UNK-CE12-CBM13-**Doc**-CBM35-CE12	+	+		+	-	-	-	-	-
14	ZP_06145360		GH48-**Doc**	+	+		-	-	-	-	-	-
15	ZP_06144535		Coh-**Doc** (ScaO)	+	+		-	-	-	-	-	-
16	ZP_06145505		Coh-**Doc** (ScaM)	+	+		-	-	-	-	-	-
17	ZP_06141671	1b	CBM-GH9-**Doc**	+		+		-	-	-		-
18	ZP_06144353		LRR-**Doc**	+		+		-	-	-		-
19	CAK18894		Coh-**Doc** (ScaC)	*	*	*	*	-	-	-	-	-
20	ZP_06141810		UNK-**Doc**	+		+		-	-	-	-	-
21	ZP_06142866		GH9-UNK(CBM?)-UNK(CBM?)-**Doc**	+	+		+	-	-	-	-	-
22	ZP_06145705		GH43-UNK-CBM13-CBM13-**Doc**	+	+		+	-	-	-	-	-
23	ZP_06142105	1c	UNK-LamGL(CBM?)-**Doc**	+	+		+	-	-	-	-	-
24	ZP_06142374	1d	UNK- **Doc**	+		+		-	-	-		-
25	ZP_06144548		UNK- **Doc** -UNK	+		+		-	-	-		-
26	ZP_06145497		Coh-Coh- **Doc** (ScaJ)	+	+	+	-	-	-	-	-	-
27	ZP_06144651	2	LRR- **Doc**	-				-		+		+
28	ZP_06143271		UNK- **Doc** -UNK	-				-		+		+
29	ZP_06143424	3	PL-CBM- **Doc**	-				-	+	-		-
30	ZP_06145446		CBM4-GH10-CBM9- **Doc**	-		-		-	+	-	-	-
31	ZP_06143878		CE-CBM- **Doc** -UNK (known as "CE3B")	-		-		-	*	-	-	-
32	ZP_06141916		GH43-X19-CBM22- **Doc** -CE1	-	-	-	-	-	+	-	-	-
33	ZP_06143260		GH53-CE- **Doc**	-		-		-	+	-	-	-
34	ZP_06142964		UNK- **Doc**						+			
35	ZP_06144896		GH11-UNK- **Doc**	-	-		-	-	+	-	-	-
36	CAK18896	4a	Coh-Coh-Coh-Coh-Coh-Coh-Coh-Coh-Coh-**X-Doc** (ScaB)	-	-	-	-	-	-	*	-+	-+
37	CAK18897		CBM-CBM- **X-Doc** (CttA)	-	-	-	-	-	-	*	+	+
38	ZP_06142651		UNK- **Doc**	-		-		-	-	+		
39	ZP_06142361		Coh- **Doc** (ScaH)	-	-	-	-	-	-	+	+	+
40	ZP_06144588		Coh- **Doc** (ScaF)	-	-	-	-	-	-	+	-+	-+
41	ZP_06142181		Peptidase-UNK- **Doc**	-	-		-	-	-	+	-	-
42	ZP_06143695		UNK- **Doc**	-				-	-	+		+
43	ZP_06145744		LRR-Coh- **Doc** (ScaI)	-	-		-	-	-	+	-	-
44	CAK18895	5	UNK-Coh-Coh- **Doc** (ScaA)	-	-	-	-	*	-	-	-	-
45	ZP_06142459	6a	"zincins"- **Doc** -UNK	-				-	+	-		-
46	ZP_06144432		UNK- **Doc**	-				-	+	-		-
47	ZP_06145118		GH18- **Doc**	-				-	+	-		-
48	ZP_06142855		UNK-PL- **Doc**	-		-		-	+	-	-	-
49	ZP_06143179		UNK-PL- **Doc**	-		-		-	+	-	-	-
50	ZP_06143476		UNK(LbetaH-LamGL)- **Doc**	-	-		-	-	+	-	-	-
51	ZP_06142906	6b	**Doc** -Serpin	-	-		-	-	+	-	-	-
52	ZP_06144185		UNK-LRR-Cysteine proteinase- **Doc**	-				-	+	-		
53	ZP_06143078		GH5-CBM32-CBM32- **Doc**	-	-		-	-	+	-	-	-

Accession numbers, architecture of the dockerin-bearing parent proteins and group classification (see also Figure 2.2) are designated. The dockerin module is marked in boldface for each ORF. Dockerins 1-16, 17-22, 23, 24-26, 27-28, 29-35, 36-43, 44, 45-50, 51-53 represent dockerin groups: 1a, 1b, 1c, 1d, 2, 3, 4a, 5, 6a and 6b, respectively. Twenty-four dockerins that were cloned and expressed but did not exhibit any interaction are available in Table S2.5 (Annexes). Glycoside hydrolase families 5, 9, 44 and 48 are putative cellulases and families 10, 11 and 43 are putative xylanases. Key to symbols in the Table: + Novel interactions discovered in the present study. * Previously reported interactions. − Interactions examined but found to be negative. Untested pairs by the designated methods were left blank.

While most dockerins are located at the C-terminus of their host protein, a few are at the N-terminus or in the middle of the polypeptide chain (e.g., Doc 11-13, 36, 50 and 55, Table 2.1). The dockerin of the family 48 GH was also included (Doc 14, Table 2.1), since this enzyme represents a major contributing component of every cellulosome system thus far described.

A collection of 19 cohesin modules was selected from the eight previously identified scaffoldins of the bacterium, including ScaA cohesins 1 and 2 (ScaA1-2), ScaB cohesins 2 to 9 (ScaB2-9), and the single cohesins in ScaC, ScaE, ScaF, ScaG, ScaH and ScaI (based on bioinformatic data, cohesins ScaB1-4 are highly similar (Dassa *et al.*, 2014); cohesins B2-B4 were thus selected and included as representatives. Cohesin B1 was not included. Additionally, three putative cohesin modules were selected: ScaJ cohesins 1-2 (ScaJ1-2) and ScaO, whose sequence divererge from canonical cohesins (Figure 2.1). The sequences of 19 selected cohesins are typical of type-III cohesins (Alber *et al.*, 2009; Dassa *et al.*, 2014; Salama-Alber *et al.*, 2012), except for ScaC, which is more related to the type-I cohesins (dendrogram in Figure 2.1). Nevertheless, sequence variations exist among the type-III cohesins. Therefore, the selected modules were chosen from different branches of the dendrogram. Putative cohesins, deemed too divergent from classic cohesins (namely ScaK, ScaL, ScaM1, ScaM2, ScaN and ScaP), were not selected for biochemical analysis.

2.3.2. Identification of novel cohesin-dockerin interactions in *R. flavefaciens*

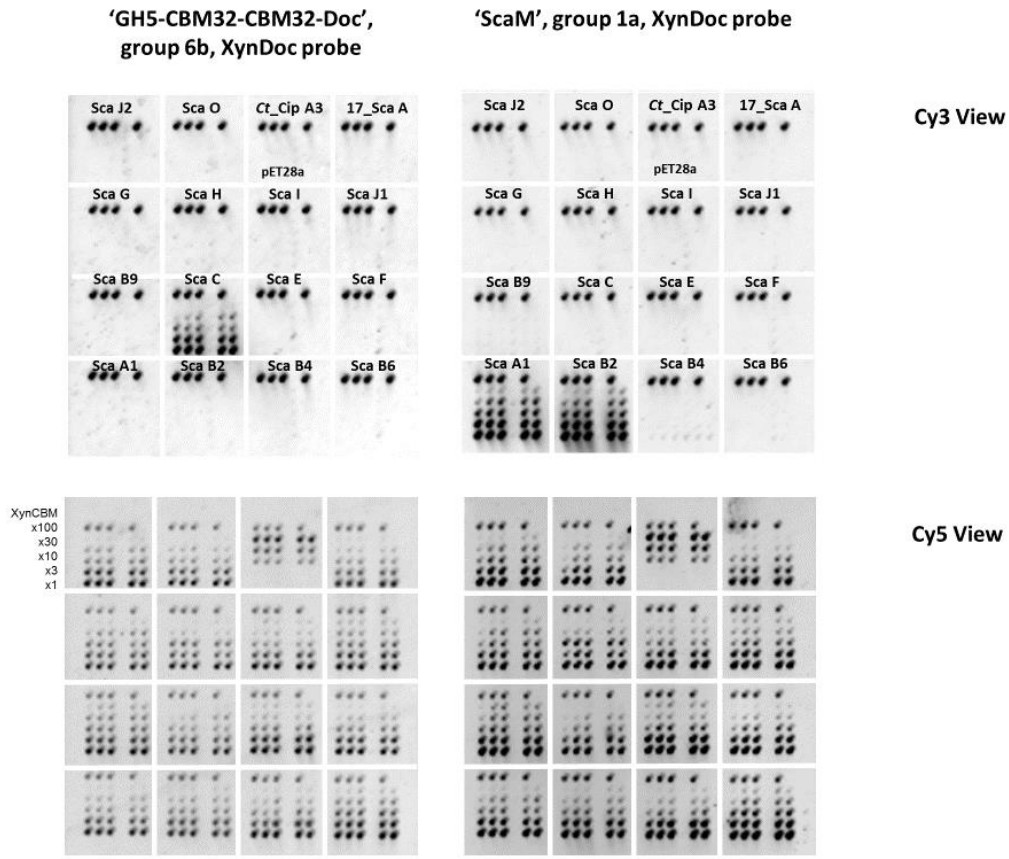
Unraveling the selective pattern of cohesin-dockerin binding within the *R. flavefaciens* cellulosome was achieved by employing three different approaches to detect protein-protein interactions. The three strategies are complementary and comprise cellulose-coated microarray, affinity-based ELISA assay, and *in-vivo* screening of co-expressed cohesin and dockerin modules with subsequent in-vitro validation by non-denaturing PAGE.

Microarray. Recombinant xylanase-fused dockerins (XynDocs) were interacted with CBM-fused cohesins (CBM-Coh). The latter allowed selective attachment to cellulose-coated slides (Slutzki *et al.*, 2012). The methodology was streamlined by applying crude cell extracts containing both CBM-Coh and XynDoc (Hamberg *et al.*, 2014), thereby facilitating analysis of large numbers of candidate modules.

In Figure 2.4, the data are presented for a series of representative CBM-Cohs applied to a cellulose-coated slide, subsequently interacted with a XynDoc probe (14 interactions tested per slide). The microarray technology was used to examine 14 *R. flavefaciens* cohesins (Figure 2.1) and 32 dockerins (Table 2.1), yielding 448 possible interactions. Figure 2.5 shows representative interactions for different dockerin-containing scaffoldins and enzymes (in many cases, multi-functional). The data are shown as bar graphs taking into account non-specific

background binding (Lytle, Myers, Kruus, & Wu, 1996). All reported binding levels were significantly above background. Note cohesin recognition trends delineate the different dockerin groups. Internal dockerins and N-terminal dockerins were as active as C-terminal dockerins. Curiously, most dockerins originating from LRR-containing parent proteins of the different groups did not interact with tested cohesins.

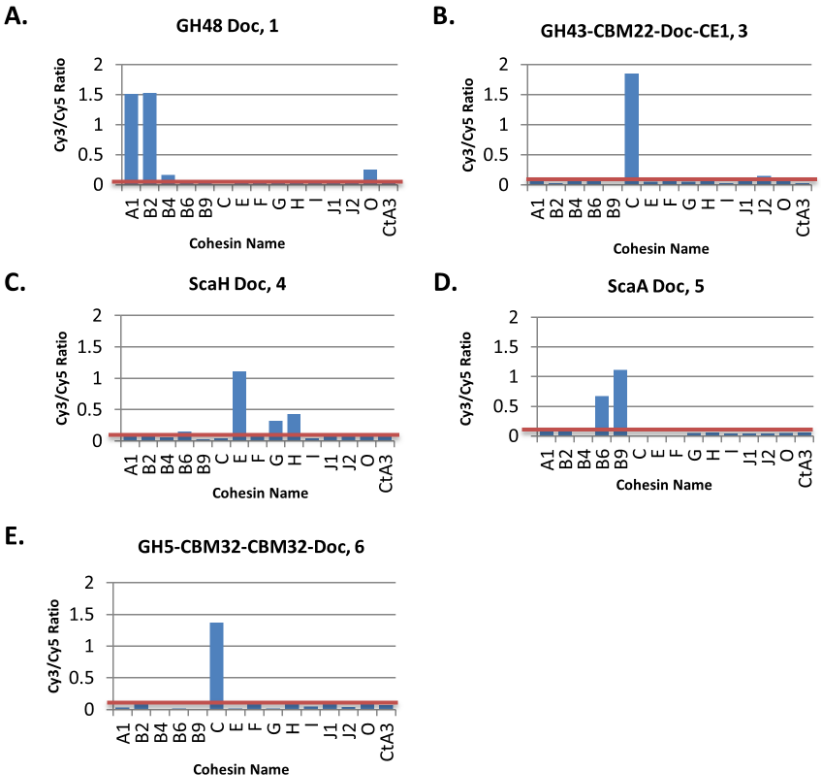
Figure 2.4 Representative cellulose-coated protein microarray screening, using crude cell extracts of both dockerin- and cohesin-fused proteins



XynDoc extracts derived from ScaM and a GH5 enzyme are shown as examples as probes against crude extracts of different CBM-cohesins, applied onto a cellulose-coated glass slide.
Upper panel: Cy3-derivatized anti-Xyn antibody labeling revealed strong interaction of the group-6b GH5-borne dockerin and the ScaC cohesin (left), whereas the group-1a ScaM dockerin (right) interacted with ScaA cohesin 1 (A1) and ScaB cohesin 2 (B2). *C. thermocellum* ScaA cohesin 3 (Ct_Cip A3) and the crude bacterial extract (transformed *E. coli* BL21 with an empty plasmid (pET28a) were used as negative controls. ScaA cohesin 3 of *R. flavefaciens* strain 17 (17_ScaA) was used to examine whether cross-strain interaction occurs.
Lower panel: Cy5-derivatized anti-CBM antibody labeling observed for all of the printed protein spots on the microarray. The intensity of each spot is in linear correlation with the amount of CBM-Coh present. The array is divided into subarrays, each containing a different CBM-Coh sample. The top row of each subarray includes a XynCBM positive control, below which are serial dilutions by a factor of 3

of the crude cell extracts. Each CBM-Coh was printed in quintuplicate for each dilution. The scheme of all printed microarray samples is shown at the bottom left.

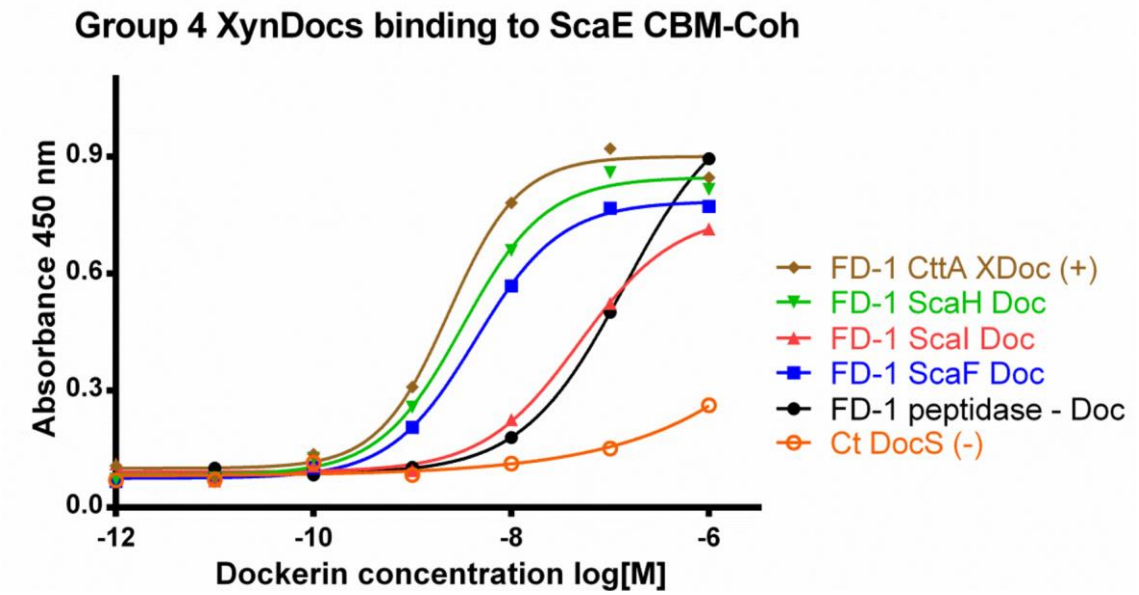
Figure 2.5 Quantification of representative interacting cohesin-dockerin pairs from *R. flavefaciens* strain FD-1 on cellulose-coated microarrays.



Each bar graph represents interactions of a designated dockerin probe vs. 14 different cohesins (abscissa: ScaA1, ScaB2, ScaB4, etc.) and *C. thermocellum* ScaA-CohA3 (CtA3) as a control. (A) Group-1 dockerins, represented by ZP_06145360 (GH48 Doc). (B) Group-3 dockerins, represented by ZP_06141916 (GH43-CBM22-Doc-CE1). (C) Group-4 dockerins, represented by ZP_06142361 (ScaH-Doc). (D) The lone group-5 dockerin, ScaA-Doc (CAK18895). (E) Group-6 dockerins, represented by ZP_06143078 (GH5-CBM32-CBM32-Doc). See Table 2.1 for complete summary of the cohesin-dockerin interactions investigated in this work.

ELISA. The interaction of various *R. flavefaciens* recombinant XynDocs (Table 2.1) with CBM-Cohs was also tested using an ELISA approach. The binding of group-4 dockerins (i.e., ScaF, ScaH and ScaI dockerins, as well as peptidase-Doc) to ScaE, indicates that these components attach to the bacterial cell surface (Figure 2.6). Several of these interactions displayed only weak binding using cellulose microarrays, yet IC50 indicate high-affinity binding (in the nano-molar range) of CttA XDoc, ScaH and ScaF, and an order-of-magnitude less for ScaI and peptidase-Doc. Based on these results, we concluded that such apparent low-affinity interactions, as revealed by the cellulose microarrays, should be regarded as possible positive hits, requiring further confirmation by complementary approaches.

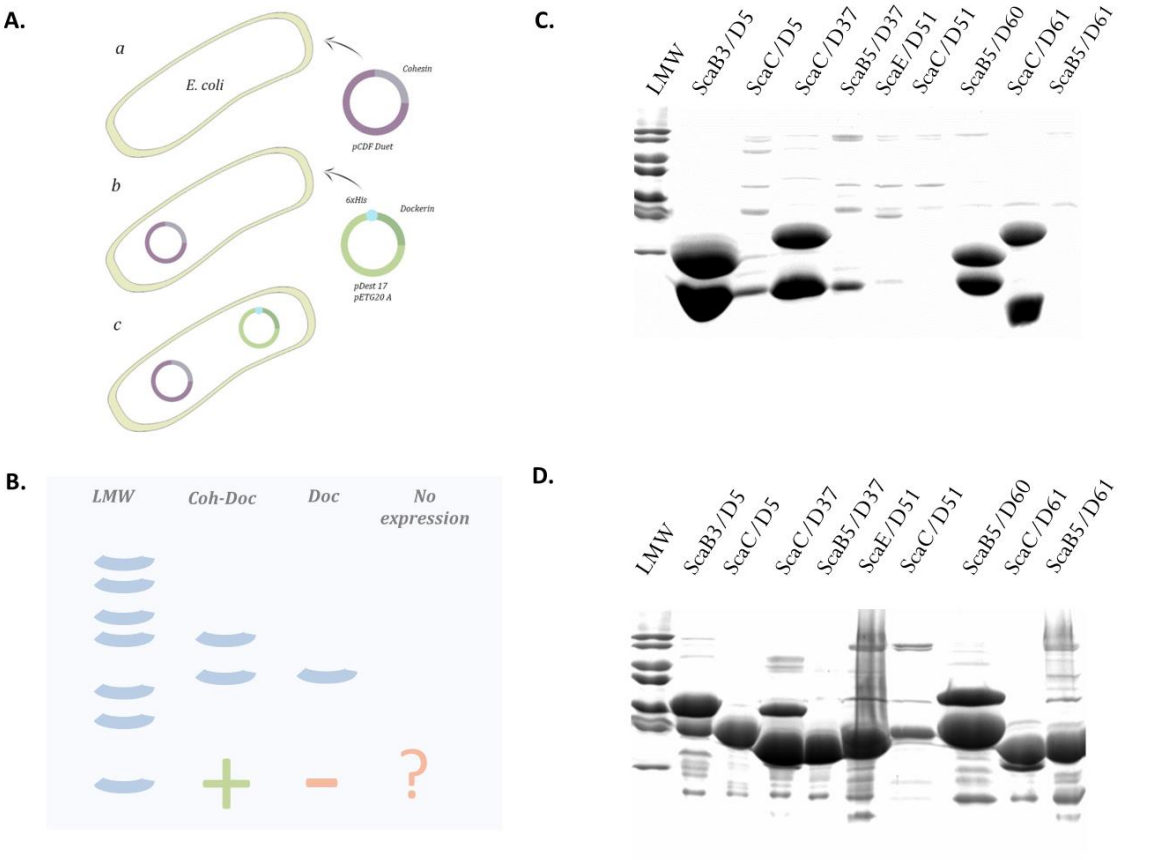
Figure 2.6 Binding of group-4 dockerins to ScaE cohesin probed by an ELISA assay.



XynDocs of CttA XDoc, ScaH, ScaF, peptidase-Doc and ScaI were purified on Ni-NTA columns and interacted with the ScaE cohesin. *C. thermocellum* DocS was chosen as a negative control. From the IC50 values it is clear that these dockerins indeed bind the ScaE cohesin, even though only weak interaction was observed using the cellulose microarray approach. Surprisingly, the interaction is relatively strong when compared with other known protein-protein interactions.

***In-vivo* co-expression.** Dockerins are small unstable protein modules prone, to degradation when expressed in *E. coli*. However, recombinant dockerins are stabilized when bound to their counterpart cohesin. Thus, we devised a third complementary approach to identify novel cohesin-dockerin interactions within the *R. flavefaciens* cellulosome. Genes encoding different cohesin/dockerin partners were isolated and cloned into two compatible vectors for co-expression in *E. coli*. Recombinant dockerins contained an engineered N-terminal His tag. Immobilized metal-ion affinity chromatography (IMAC) was used to purify the recombinant dockerins together with the cohesins, upon binding between the two modules. Thus, protein complex formation was analyzed through SDS-PAGE by detecting the presence of a recombinant cohesin (Figure 2.7A,B and Figure 2.8). For these experiments 10 cohesins (Figure 2.1) and 45 dockerins (Table 2.1) were selected.

Figure 2.7 Identification of cohesin-dockerin complexes following recombinant in-vivo co-expression

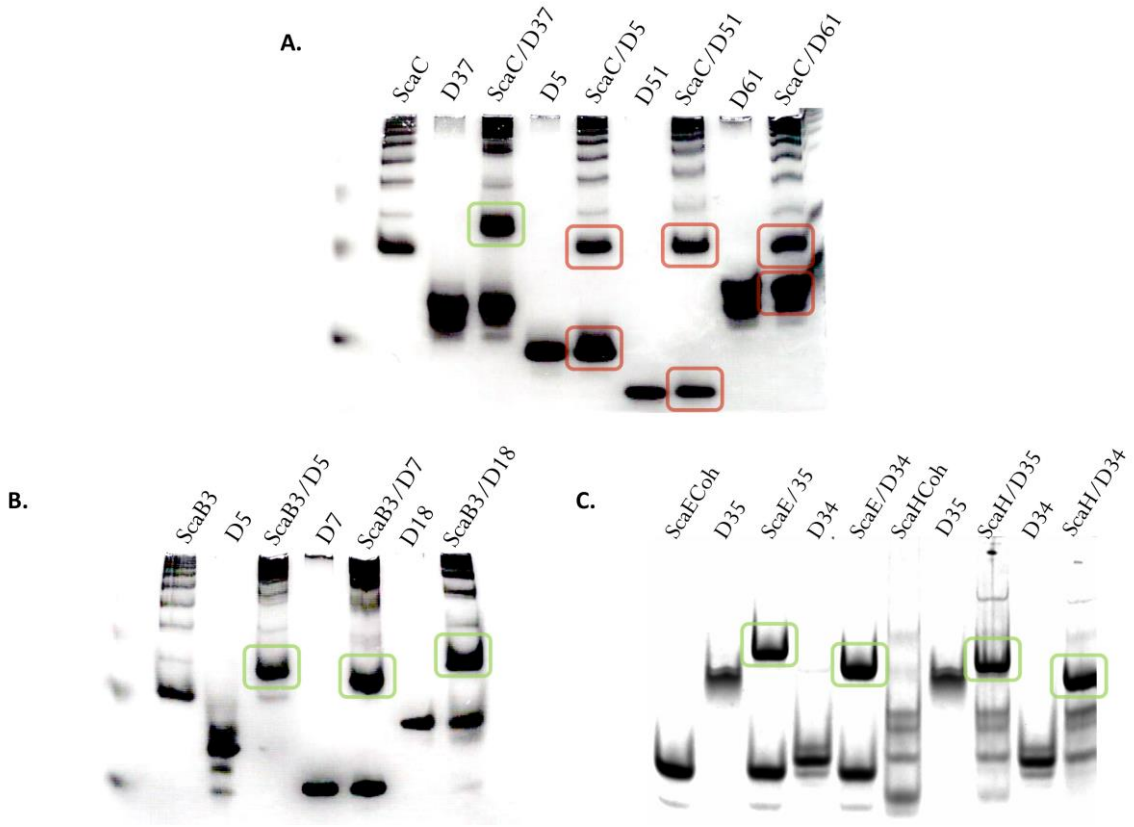


(A) Schematic depiction of the recombinant in-vivo co-expression strategy. Cohesin-encoding genes were inserted into the pCDFDuet plasmid that was used to transform *E. coli* BL21(DE3) competent cells. Cells were made competent again and re-transformed with 45 Dockerins previously inserted into pDest17 (His-tag) and pETG20A (TrxA-His-tag). A total of 720 different clones (8 cohesins x 45 dockerins x 2 vectors) were obtained and used for co-expression. (B) Schematic illustration of the expected results. After purification by IMAC, in-vivo complex formation was evaluated by loading the purified samples onto SDS-PAGE gels. Since only the dockerins possessed a His tag, identification of complex formation was determined by the appearance of two bands in the gel, corresponding to the His-tagged dockerin and the bound cohesin. A single band corresponded to the isolated dockerin alone. The absence of bands indicated that the dockerin was either insoluble or did not express. (C). Representative experiment showing SDS-PAGE of selected samples: Two bands indicating in-vivo complex formation are clearly evident in the cases of ScaB3/D5 (group 1), ScaC/D37 (group 3), ScaB5/D60 (ScaADoc) and ScaC/D61 (group 6). Dockerin stability is greatly improved when bound to the cohesin as indicated by the difference in band intensity between bound and unbound dockerins. (D) Duplication of the experiment with TrxA-fused dockerins was carried out to eliminate false negatives due to low dockerin expression or insolubility. See Table 2.1 for complete summary of cohesin-dockerin interactions.

Initially, the capacity of recombinant *E. coli* strains to produce all 10 cohesins was evaluated. Two cohesins, from ScaG and ScaI, were insoluble when expressed under various conditions. Therefore, the *in vivo* expression studies were performed with the eight cohesins that expressed at detectable levels. Recombinant *E. coli* strains expressing the soluble cohesins were rendered

competent and retransformed with 45 plasmids encoding dockerins. Since dockerins were expressed with either a single His-tag (in pDest17) or a thioredoxin fusion partner for increased solubility (pET20G), in total 720 interactions were tested (8 cohesins x 45 dockerins x 2 vectors). Analysis of the 720 recombinant strains, transformed with the cohesin- and dockerin-containing plasmids (exemplified in Figure 2.7 C,D) revealed that the capacity of *E. coli* to produce dockerins was severely impaired in the absence of a fusion protein (Figure 2.7 C). However, dockerin yield was significantly higher when a co-purified cohesin band was observed, confirming that binding to cohesin stabilizes dockerin structure leading to significant levels of protein production (Figure 2.7 D). Both co-expression experiments, using unfused and fused dockerins, generally revealed identical cohesin-dockerin specificity patterns. However, in some cases the size of the dockerin-fused protein was similar to that of the cohesin, making binding difficult to detect. Thus, the interaction of cohesin and dockerin pairs was validated by independent production of the two proteins in *E. coli*, using the TrxA-His fused dockerin derivative and His-tag fused cohesins. Following purification by IMAC, cohesin and dockerin modules were incubated to promote complexation, which allowed clarification of the cohesin-dockerin interactions.

Figure 2.8 Confirmation of *in vivo* co-expression data by non-denaturing PAGE.



The first lane of gels A and B were loaded with the cohesin (ScaC in A and ScaB3 in B). Adjacent lanes were loaded with a test dockerin and with both cohesin and dockerin modules together after 60-min

incubation at equimolar concentrations. Dockerins are numbered according to Table 2.1. The appearance of a band with a different migration pattern (green highlights) in lanes containing the complex represents a positive result (e.g. ScaC/D37), while a negative result (e.g. ScaC/D5) is given by the appearance of only the individual dockerin and cohesin bands (red highlights). ScaC interacts with group-3 dockerin D37 but not with groups-1, -4 or -6 dockerins D5, D51 or D61. ScaB3 binds to group-1 dockerins D5, D7 and D18. Gel C shows the binding of both group-2 dockerins to ScaE and ScaH.

2.3.3. Novel cohesin-dockerin specificities reveal the overall architecture of the *R. flavefaciens* cellulosome

Data concerning the novel cohesin-dockerin specificities observed in *R. flavefaciens* cellulosomes, as evaluated by the three different platforms described above, are summarized in Table 2.1. In general, 5 major patterns of selectivity between cohesins and dockerins were observed, as follows:

- A broad range of group-1 dockerins recognized ScaA cohesins 1-2 and ScaB cohesins 2-4. Many of the dockerins in this group are components of enzymes, bearing catalytic motifs crucial for carbohydrate-degradation such as GHs in families 5, 9, 10, 11, 26, 43 and 48, which include the major cellulases and some hemicellulases; CEs from families 1, 3, 4 and 12) and CBMs. Some dockerins originate from established and putative cohesin-containing proteins, including ScaC, ScaE-like scaffoldin (ZP_06142991), ScaJ, ScaO, ScaM (Table 2.1).
- Both group-2 dockerins recognized the cohesins of ScaE and ScaH, as revealed by *in-vivo* co-expression and isothermal titration calorimetry **(**ITC) (see below).
- Dockerins of groups 3 and 6, exclusively recognized the same binding partner, the ScaC adaptor cohesin. Prior to the present work, only the dockerin of the enzyme CE3B (Table 2.1, Doc 31) was demonstrated to bind the ScaC cohesin (Jindou *et al.*, 2006). This dockerin was included as a member of the group-3 dockerins (Rincon *et al.*, 2004, 2010). Our study broads the range of possible interactions between the ScaC cohesin and dockerins belonging to groups 3 and 6. In this regard, the fact that the ScaC cohesin and dockerins of groups 3 and 6 share high sequence similarity with type I, and not type III, modules is of note (Rincon *et al.*, 2010) (Figure 2.1). This type of dockerin is almost exclusively a component of hemicellulases (GH families 5, 10, 11, 16, 24, 26, 43, 53 and 97), associated CEs, and some PLs.
- Similar to the group-2 dockerins, group-4 dockerins (notably those of CttA, ScaB, ScaF, ScaH, ScaI and peptidase-Doc) recognized the ScaE cohesin. Moreover, very weak binding of the CttA-XDoc and ScaH-Doc to cohesin H and the standalone cohesin G was observed in cellulose microarrays. The binding of group-4 dockerins to cohesins G

and H was further supported by ELISA data, which provided evidence for ScaB-XDoc and ScaF-Doc as binding partners for these cohesins. Using the *in-vivo* screening approach, ScaH-Doc and another dockerin of a parent protein (ZP_06143271) of unknown function (UNK) were found to recognize cohesin H in addition to cohesin E. Interestingly, ScaH-Doc recognized its own cohesin. The ScaB and CttA dockerins were expressed with their adjacent upstream X-modules to ensure their functionality, as discussed previously (Alber *et al.*, 2009; Rincon *et al.*, 2005, 2007; Salama-Alber *et al.*, 2012). As mentioned above, group-4 dockerins have a symmetrical sequence, as reflected by their two Ca^{2+}-binding repeats, an apparent peculiarity for type III dockerin modules (Rincon *et al.*, 2010). Further analysis of a possible dual-binding mode of group-4 dockerins by alanine scanning assay coupled with ELISA is detailed below.

- The unique ScaA dockerin is the only member of group 5. It was found to bind cohesins 5 through 9 on the ScaB scaffoldin, as formerly reported (Karpol *et al.*, 2013; Rincon *et al.*, 2003; Slutzki *et al.*, 2013).

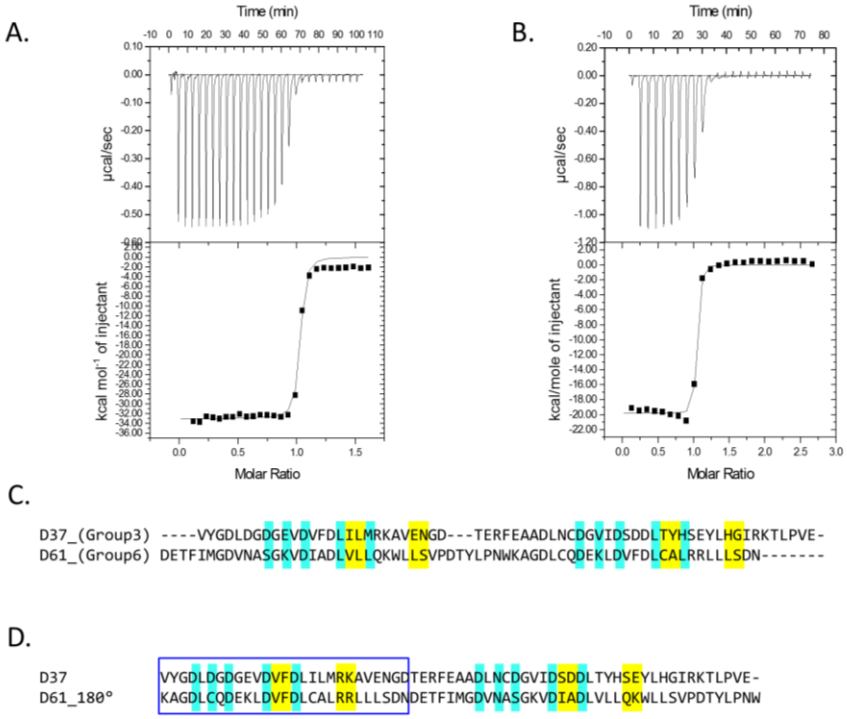
2.3.4. Probing the specificities of groups-2 & -4 dockerins and groups-3 & -6 dockerins by ITC

The data presented above suggest that dockerins of groups 3 and 6 bind exclusively to the ScaC cohesin. The interaction between representative members of groups-3 and -6 dockerins and ScaC cohesin was evaluated by ITC at 35°C, the temperature of the *R. flavefaciens* microbial niche. The data (Figure 2.9, Table 2.2) reveal macromolecular association of high affinity (K_a 10^8 M^{-1}; stoichiometry of approximately 1:1). The sequences of these two dockerin groups indicate an asymmetric distribution of predicted recognition residues, suggesting a single-binding mode. When the two dockerins are aligned after swapping the C- and N-terminal halves of the group-6 dockerin, the identity at the putative cohesin-interacting region increases (Figure 2.9D). A similar twofold alternative specificity mechanism was recently observed for cohesin-dockerin recognition in another ruminococcal species (Moraïs *et al.*, 2016).

Group-2 dockerins resemble truncated versions of group-4 modules (Rincon *et al.*, 2010). ITC using representative members of groups-2 and -4 dockerins was performed to quantify the affinity of both interactions. Data, presented in Figure 2.11 and Table 2.2, suggest a lower affinity constant (K_a of 10^6-10^7 M^{-1}) compared with groups-3 and -6 dockerins. Alignments of groups-2 and -4 dockerins suggest that group-2 dockerins are highly homologous to the C-terminus of group-4 proteins (Figure 2.3 and Figure 2.10C). ITC experiments also confirmed the affinity of group-2 dockerins to the ScaH cohesin (data not shown), although the interaction

was too tight to accurately determine the K_a using this method. As described for other cohesin-dockerin pairs the interactions described here between *R. flavefaciens* cohesin-dockerin pairs are both enthalpically and entropically unfavourable (Carvalho *et al.*, 2007; Pinheiro *et al.*, 2008).

Figure 2.9 Binding of group-3 and group-6 dockerins to ScaC cohesin evaluated by ITC.



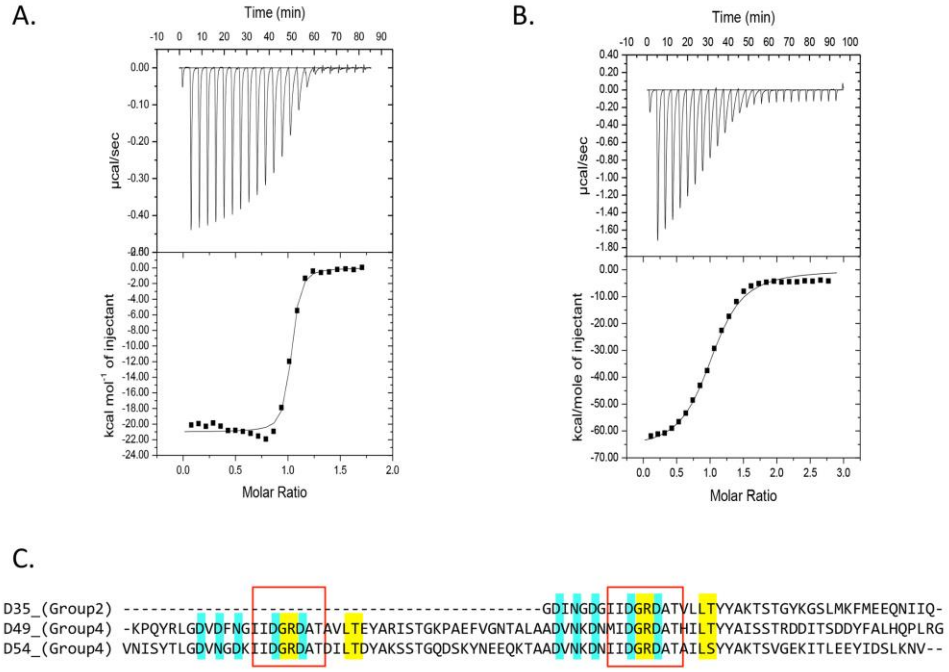
The dockerins are numbered according to Table 2.1. Representative titrations are displayed in panel (A), ScaC Coh and dockerin 37 (D37), and (B), ScaC Coh and dockerin 61 (D61). The upper part of each panel shows the raw heats of binding, whereas the lower parts comprise the integrated heats after correction for heat dilution. The curve represents the best fit to a single-site binding model. (D) Alignment of dockerin D37 (group 3) with D61 (group 6) and of dockerin D37 with D61_180° (a mutated version of D61 in which the C-terminal half was switched with the N-terminal half). Note the similarity in the cohesin-recognition residues in the aligned first repeat (blue box, yellow highlight). Residues involved in Ca^{2+}-binding are colored in cyan while putative residues involved in cohesin recognition are highlighted in yellow.

Table 2.2 Thermodynamics of novel cohesin-dockerin interactions identified in *R. flavefaciens* cellulosome as evaluated by ITC.

Interaction	K_a M^{-1}	ΔG^o kcal m	ΔH kcal mol^{-1}	$T\Delta S^o$ kcal mol^{-1}
ScaCCoh/D32 (Group 3)	2.69E8 ± 2.52E7	-11.85	-36.33 ± 0.055	-24,48
ScaCCoh/D53 (Group 6)	3.54E8 ± 1.37E7	-12.05	-19.79 ± 0.183	-7.73
ScaECoh/D30 (Group 2)	6.17E7 ± 3.13E6	-10.99	-23.00 ± 0.520	-12.01
ScaECoh/ScaHDoc (Group 4)	1.56E6 ± 1.61E5	-5,90	-64.11 ± 1.181	-58.21

Thermodynamic parameters were determined at 308.16 K.

Figure 2.10 Binding of group-2 and group-4 dockerins to ScaE evaluated by ITC



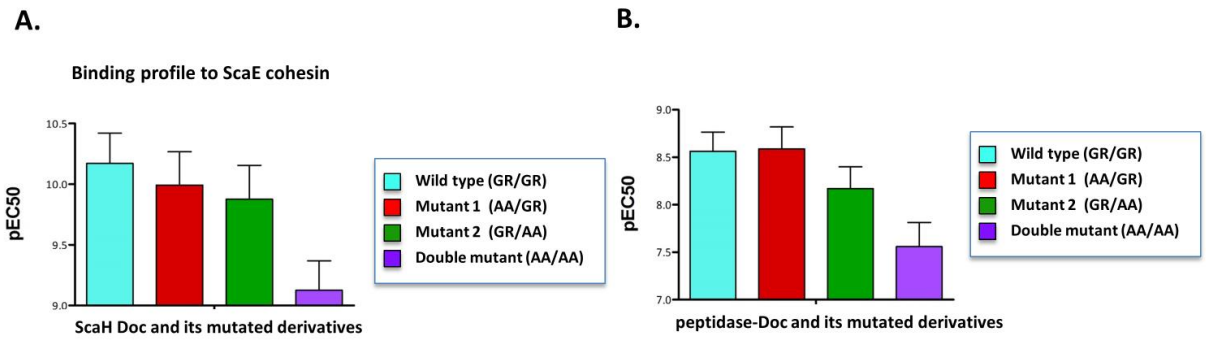
The dockerins are numbered according to Table 2.1. Representative titrations are displayed in panel (A), ScaE Coh and dockerin 35 (D35), and panel (B), ScaE Coh and dockerin 49 (D49). The upper part of each panel shows the raw heats of binding, whereas the lower parts comprise the integrated heats after correction for heat dilution. The curve represents the best fit to a single-site binding model. (C) Alignment of dockerin D35 (group 2) with two group-4 dockerins, D49 (ScaHDoc) and D54. The conservation of the postulated cohesin recognition site is highlighted with a red box. Residues involved in Ca+2-binding are colored in cyan while putative residues involved in cohesin recognition are highlighted in yellow.

2.3.5. Dual-binding mode in group-4 type III dockerins

Data presented here suggest that group-4 dockerins associate to the bacterial cell envelope via recognition of the anchoring ScaE cohesin, without an upstream X-module and internal insertions (Rincon *et al.*, 2005, 2007; Salama-Alber *et al.*, 2013). Furthermore, these *R. flavefaciens* dockerins are generally distinctive within the realm of the type-III modules for their unique symmetrical nature. Alignment of these dockerins together with the XDocs of ScaB and CttA (Figure 2.3) revealed that several of them, notably peptidase-Doc (ZP_06142181) and ScaH-Doc (ZP_06142361), exhibit similar Gly-Arg residues at postulated cohesin-recognition sites (Levy-Assaraf *et al.*, 2013; Rincon *et al.*, 2007). Interestingly, the dockerins of ScaB and CttA also possess duplicated Gly-Arg residues in both of their purported recognition sites, but the overall symmetry is disrupted by the characteristic extended insertions. Dockerins that exhibit symmetrical sequences have been shown in other bacterial species to possess two identical binding sites (i.e., dual-binding mode), thought to promote conformational flexibility to facilitate integration of enzymes into the cellulosomal complex and/or to overcome steric interactions which may interfere with the action of cellulosomal enzymes with the substrate

(Carvalho *et al.*, 2007; Pinheiro *et al.*, 2008). To investigate such a role in *R. flavefaciens* strain FD-1, mutants of the above-designated symmetrical group-4 dockerins, containing Ala-Ala substitutions for the Gly-Arg dyad in one or both of the putative repeated recognition sites. From the extrapolated pEC50 values (Figure 2.11), binding to the counterpart cohesin of ScaE was only impaired in the double mutant. Binding, however, was not completely eliminated due to apparent involvement of additional interacting residues. These results clearly indicate a dual-binding mode for the symmetrical group-4 dockerins.

Figure 2.11 Dual-binding mode in the symmetrical group-4 dockerins.

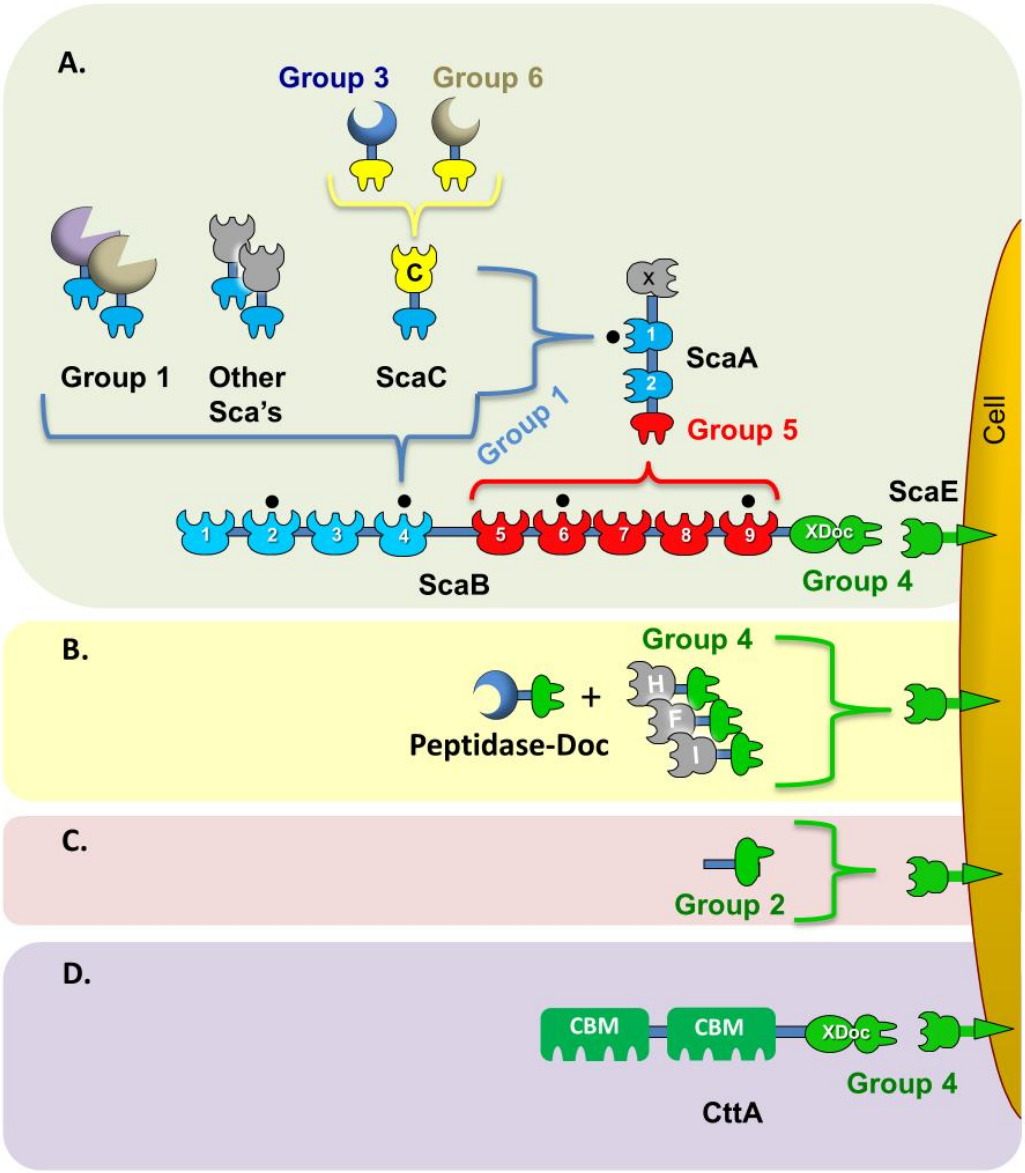


(A) ScaH Doc (ZP_06142361) and (B) peptidase-Doc (ZP_06142181). Alanine mutations were inserted at the major putative cohesin-recognition residues: positions G11/R12 and/or G50/R51, representing mutations in the first or second repeated segment of the dockerins, or the double mutant. Binding ability of the wild-type and mutants to the ScaE cohesin was examined by ELISA, and pEC50 values were determined as described previously (Barak *et al.*, 2005).

2.4. Discussion

The complexity of the *R. flavefaciens* FD-1 cellulosome system is reflected by its numerous secreted fiber-degrading dockerin-containing enzyme and non-enzymatic subunits and encoded scaffoldins, which can potentially generate innumerable configurations of cellulosome assemblies (Berg Miller *et al.*, 2009; Dassa *et al.*, 2014; Rincon *et al.*, 2010). Using three experimental approaches to screen for cohesin-dockerin interactions, we accumulated evidence for several novel interactions between type III cohesins and their cognate dockerins belonging to heterogeneous groups. The results present recognition preference between the different cohesins and dockerins groups in this ruminal bacterium. They provide a snapshot of the molecular organization of the intricate *R. flavefaciens* cellulosome system, thus enabling routes of elaborate assembly of these multienzyme complexes, a model of which is proposed in Figure 2.12.

Figure 2.12 Current model of cellulosome assembly in *R. flavefaciens* strain FD-1.



The scheme is color-coded to highlight the four subgroups of cohesin-dockerin specificities: Dockerins and cognate cohesin counterparts of the different groups are marked in light blue (Group-1 dockerins), yellow (Groups 3 and 6), green (Groups 2 and 4) and red (Group 5), respectively. The interacting partner(s) of cohesin modules marked gray, are yet to be discovered (and consequently yet to be confirmed as bona fide cohesins). (A) Cellulosomal proteins. (B) Cell wall-attached proteins. (C) Short (half) dockerins of group 2. (D) CttA subunit, purportedly mediating substrate attachment (Rincon *et al.*, 2007).

The data correlate well with previous bioinformatic observations that *R. flavefaciens* dockerins exhibit exclusive sequence features allowing their classification into six distinct groups (Rincon *et al.*, 2010). The second-order classification of the dockerin groups into eleven subgroups was found to be functionally redundant, since cohesin recognition among the various subgroups did not segregate with this subgroup classification. The subgrouping of these dockerin sequences

may infer structural variations that reflect the stability of interaction with the cohesin or secondary interactions with the parent protein.

(Borne, Bayer, Pagès, Perret, & Fierobe, 2013) have recently demonstrated that, despite the general lack of interspecies cohesin-dockerin specificity, cellulosomes are not necessarily assembled in solution at random. The same study argued that enzyme binding to a cohesin will directly influence subsequent incorporation of other enzymes by mechanisms other than steric hindrance. These results support previous coarse-grain molecular modeling studies by (Bomble *et al.*, 2011) Moreover, preferential integration may also be related to inter-cohesin linker length (Vazana *et al.*, 2013).

Group-1 dockerins comprise the majority of the encoded dockerins in the *R. flavefaciens* genome (96 ORFs) and mainly include multi-functional catalytic modules, such as numerous GHs, CEs, PLs and CBMs (Berg Miller *et al.*, 2009; Dassa *et al.*, 2014). The data presented here support previous claims (Jindou *et al.*, 2006) that Group-1 dockerins, whose sequence profile is exclusive to *R. flavefaciens*, preferentially bind cohesins ScaA1-2 and ScaB1-4.

Dockerins of groups 3 and 6 (mainly originating from hemicellulases) preferentially bound to the ScaC adaptor cohesin (Table 2.1). The common recognition profile suggests that enzymes associated with these dockerins might functionally interact. Interestingly, the putative recognition residues of these two dockerin groups are largely reversed, reminiscent of a similar phenomenon recently described for groups-3 and 4 dockerins of the human isolate *Ruminococcus champanellensis* (Berg Miller *et al.*, 2009; Dassa *et al.*, 2014; Moraïs *et al.*, 2016). Significantly, the ScaC cohesin is similar to type I cohesins of other cellulosome-producing bacteria, as opposed to the majority of type III cohesins in this bacterium.

Intriguingly, growth of *R. flavefaciens* strain 17 on xylan was shown to upregulate dockerin-containing enzymes that interact with the ScaC cohesin versus cultures grown on microcrystalline cellulose (Rincon *et al.*, 2004). Moreover, the same study showed that components from cultures cultivated on xylan are enriched with very high-molecular-weight dockerin-bearing components that interact strongly with the ScaC Coh (and also with that of ScaA). In this context, high-molecular-weight multifunctional xylanases and carbohydrate esterases are produced by the various strains of *R. flavefaciens* (Dassa *et al.*, 2014; Rincon *et al.*, 2010). The combined evidence suggests that ScaC may be involved in a regulatory mechanism that governs preferential expression of enzymes that act on hemicelluloses.

A serpin-associated group-6 dockerin was also observed (Rincon *et al.*, 2010). The serpin in this context may play a role in protecting the enormous cellulosome assembly from inadvertent extra-cellular proteolytic cleavage (O Cuív, Gupta, Goswami, & Morrison, 2013; Steenbakkers *et al.*, 2008). Such serpins also exist in other cellulosomal systems, such as those of *C.*

thermocellum and *R. albus* (Irving *et al.*, 2002; Kang *et al.*, 2006). Other putative roles could be regulatory in nature, since serpins are involved in cascade control processes or spatial confinement of developing signals (Hashimoto, Kim, Weiss, Miller, & Morisato, 2003).

Previously, the ScaE cohesin had only been reported to interact with three proteins that share an X module-dockerin dyad: ScaB, CttA and a putative cysteine peptidase (Levy-Assaraf *et al.*, 2013; Rincon *et al.*, 2005). The well-characterized interactions of ScaB and CttA link the entire cellulosome machinery to the bacterial envelope and mediating substrate recognition and cell adhesion (Jindou *et al.*, 2006; Rincon *et al.*, 2005, 2007). Single-molecule force spectroscopy revealed one of the strongest bimolecular protein-protein interactions yet reported for this type of interaction (Schoeler *et al.*, 2014). The dockerins possess three unique insertion regions that are absent in other dockerins. Recently, the crystal structure of the CttA-XDoc complex with ScaE was solved (Salama-Alber *et al.*, 2013; Venditto *et al.*, 2015), indicating that the insertions serve to reinforce the stalk like structure of the X module. Another form of X module was found to be involved in *C. thermocellum* type II interactions (Adams, Pal, *et al.*, 2005). These modules are believed to contribute to the solubility, conformational state, structural and thermal stability and spatial flexibility of the cohesin-dockerin pair.

The dual-binding mode is proposed to decrease steric constraints imposed when multiple enzymes are integrated into a single scaffoldin unit, resulting, in some cases, in a bias towards cellulosome integration. Some *C. thermocellum* enzymes harbour unique type I dockerins, which are directed to the cell surface and appear to interact via a single-binding mode, since their pivotal cohesin-recognition residues at positions 11 and 12 of one of the dockerin-binding interfaces were atypical (Brás *et al.*, 2012). It was suggested that cellulosomal enzymes with dual-binding-mode dockerins may transiently interact with the bacterial cell surface before they are assembled into the multi-enzyme complexes. This mechanism would ensure retention by the bacterium even if cohesins are saturated. In addition, single-binding-mode dockerins recruit appended enzymes specifically to the cell surface. It is possible that synergism between cell surface-bound enzymes and cellulosomal enzymes may contribute to efficient hydrolysis of structural carbohydrates. Curiously, dockerin members of group 4 display internal symmetry of the two calcium-binding repeats, a phenomenon usually common to the majority of type I dockerins, but not prevalent in type III dockerins.

To summarize, this study has verified four major cohesin-dockerin recognition specificities in the cellulosome assembly of *R. flavefaciens* strain FD-1. Our findings provide an answer to the fundamental question whether bioinformatic classification of the 223 dockerin modules into groups with distinct sequence characteristics reflects binding specificity (Rincon *et al.*, 2010). The data provided herein revealed the most complex and diverse cellulosome described to date.

Not only does *R. flavefaciens* form the largest enzymatic consortium thus far identified, it also comprises the largest number of different cohesin-dockerin interactions observed in a single bacterium. This study demonstrates how a set of complimentary medium to high-throughput techniques can be applied to address functionally relevant questions concerning the activity of highly efficient nano-machines. We provide the basis for future exploration of novel cohesin-dockerin interactions in the field of nano-biotechnology, whereby recombinant chimeric scaffoldin constructs, harboring cohesins of different selective specificities, allow precise incorporation of matching dockerins attached to selected enzyme hybrids, thus promoting synergistic action of all biological processes that benefit from enzyme/protein proximity.

Chapter 3

Ruminococcus flavefaciens Coh-Doc complexes involving dockerins from groups 3 & 6

Single-binding mode integration of hemicellulose degrading enzymes via adaptor scaffoldins in *Ruminococcus flavefaciens* cellulosome

Pedro Bule[a], Victor D. Alves[a], André Leitão[a], Luís M.A. Ferreira[a], Edward A. Bayer[b], Steven P. Smith[c], Harry J. Gilbert[d], Shabir Najmudin[a] and Carlos M.G.A. Fontes[a,1]

[a] CIISA – Faculdade de Medicina Veterinária, ULisboa, Pólo Universitário do Alto da Ajuda, Avenida da Universidade Técnica, 1300-477 Lisboa, Portugal. [b] Department of Biomolecular Sciences, The Weizmann Institute of Science, Rehovot 76100 Israel. [c] Department of Biomedical and Molecular Sciences, Queen's University, Kingston, ON K7L 3N6, Canada. [d] Institute for Cell and Molecular Biosciences, Newcastle University, The Medical School, Newcastle upon Tyne NE2 4HH, United Kingdom. ***[1]Corresponding author***

Adpated from *Journal of Biologycal Chemistry* 2016 Dec 23;291(52):26658-26669 (Bule *et al.*, 2016) and *Acta Crystallographica Section F Structural Biology Communications* 2014 Aug;70(Pt 8):1065-7 (Bule, Ruimy-Israeli, *et al.*, 2014)

Abstract

The assembly of one of Nature´s most elaborate multi-enzyme complexes, the cellulosome, results from the binding of enzyme-borne dockerins to reiterated cohesin domains located in a non-catalytic primary scaffoldin. Generally, dockerins present two similar cohesin binding interfaces that support a dual binding mode. The dynamic integration of enzymes in cellulosomes, afforded by the dual binding mode, is believed to incorporate additional flexibility in highly populated multi-enzyme complexes. *Ruminococcus flavefaciens*, the

primary degrader of plant structural carbohydrates in the rumen of mammals, uses a portfolio of more than 220 different enzymes to assemble the most intricate cellulosome known to date. A sequence-based analysis organized *R. flavefaciens* dockerins into six groups. Strikingly, a subset of *R. flavefaciens* cellulosomal enzymes, comprising dockerins of groups 3 and 6, were shown to be indirectly incorporated into primary scaffoldins, via an adaptor scaffoldin termed ScaC. Here we report the crystal structure of a group 3 *R. flavefaciens* dockerin, Doc3, in complex with ScaC cohesin. Doc3 is unusual as it presents a large cohesin-interacting surface that lacks the structural symmetry required to support a dual binding mode. In addition, dockerins of groups 3 and 6, which bind exclusively to ScaC cohesin, display a conserved mechanism of protein recognition that is similar to Doc3. Group 3 and 6 dockerins are predominantly appended to hemicellulose degrading enzymes. Thus, single binding mode dockerins interacting with adaptor scaffoldins exemplify an evolutionary pathway developed by *R. flavefaciens* to recruit hemicellulases to the sophisticated cellulosomes acting on the gastro intestinal tract of mammals.

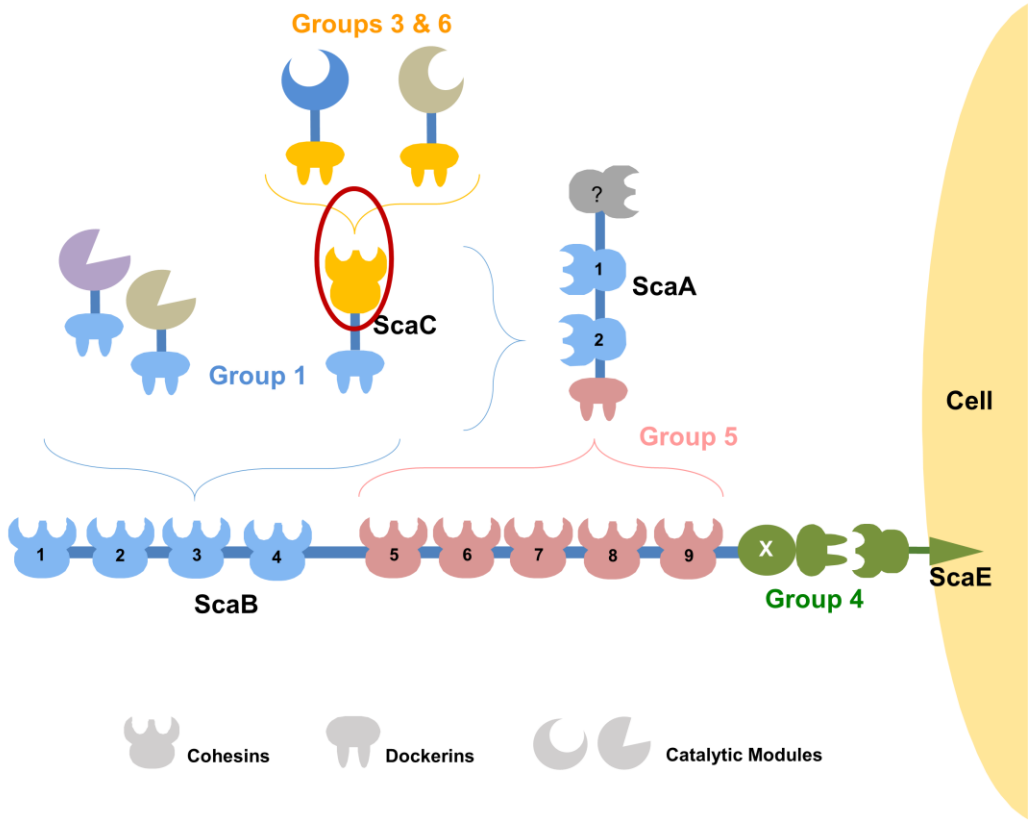
3.1. Introduction

Plant cell wall polysaccharides, primarily cellulose and hemicellulose, are the most abundant organic molecules produced in Nature, thus constituting a major reservoir of carbon and energy (Burton & Fincher, 2014). The intricate organization of structural carbohydrates in plant cell walls and their inherent heterogeneity pose significant constraints to polysaccharide degradation, which usually requires a wide array of catalytic activities acting cooperatively (Bayer *et al.*, 2007; Gilbert, 2010). In highly competitive anaerobic environments, such as the rumen of mammals, enzymatic systems that recycle the carbon stored in plant cell walls are organized in high molecular mass multi-enzyme complexes termed cellulosomes (Bayer *et al.*, 2004; Fontes & Gilbert, 2010). Molecular integration of microbial biocatalysts into these extremely elaborate nanomachines results from the binding of enzyme-borne dockerin modules (Doc) to reiterated cohesin domains (Coh) located in large non-catalytic scaffoldins, a mechanism that promotes enzyme synergy and stability. In addition, recruitment of cellulosomes to the bacterial cell surface via divergent Coh-Doc interactions allows the immediate uptake of released sugars, which are used by microbes as an energy source.

Ruminococcus flavefaciens is a Gram-positive, anaerobic bacterium of the Firmicutes phylum and the only species in the rumen that has been shown to possess a definitive cellulosome. With over 220 Doc-containing proteins, *R. flavefaciens* strain FD-1 has potentially the most complex cellulosome described to date (Figure 3.1) (Berg Miller *et al.*, 2009). Based on primary structures identity, *R. flavefaciens* Docs have been organized into six major groups (Rincon *et*

al., 2010). Recently, classification of *Ruminococcus* Docs into groups was shown to be functionality relevant as members of the same Doc group present similar Coh specificities (Ruimy, 2013). The major player in the organization of the *R. flavefaciens* FD-1 cellulosome is scaffoldin B (ScaB), which, in combination with SacA, can bind up to 14 of the 96 group 1 Doc containing proteins. These modular proteins possess catalytic modules with different activities, including glycoside hydrolases, carbohydrate esterases, polysaccharide lyases, carbohydrate-binding modules and also domains with currently unknown function (Figure 3.1). The binding of the C-terminal group 4 Doc located in ScaB to the Coh of ScaE, a cell-bound anchoring scaffoldin, provides the molecular mechanism to tether *R. flavefaciens* cellulosome to the bacterial cell surface (Rincon *et al.*, 2005). Unique to the *R. flavefaciens* FD-1 cellulosome is the presence of the adaptor scaffoldin ScaC, which contains a group 1 doc and thus can interact with either ScaA or ScaB (Rincon *et al.*, 2004). Sca C also contains a single Coh that is capable of interacting with Group 3 and 6 Docs. The ScaC adaptor scaffoldins may thus modulate integration of alternative types of enzymes into the cellulosome when this is functionally relevant.

Figure 3.1 Group-specific interactions that contribute to cellulosome assembly in *R. flavefaciens* strain FD-1.



The scheme is color-coded to highlight the four subgroups of Coh-Doc specificities: Docs and cognate Coh counterparts of the different groups are marked in light blue (Group 1 Docs), yellow (Groups 3 and 6), green (Groups 2 and 4) and red (Group 5), respectively. Group 2 Docs are truncated derivatives of

group 4 and are not represented in the figure for simplification. The red oval marks the complex of the Group 3 interaction, whose structure is reported here.

Structural studies on Coh-Doc complexes from *Clostridium thermocellum* (12–13), *C. cellulolyticum* (Pinheiro *et al.*, 2008) and *Acetivibrio cellulolyticus* (Cameron, Najmudin, *et al.*, 2015), revealed that the observed primary structure duplication in Docs appended to cellulosomal enzymes supports a dual-binding mode with their target protein partners. This consists in the Doc's ability to bind the Coh in two different orientations, 180º opposite to each other. The dual binding mode is believed to confer additional flexibility to the macromolecular organization of cellulosomes. Primary structure analysis revealed that *R. flavefaciens* Group 3 and 6 Docs, although appended to enzymes, do not seem to possess the internal sequence symmetry found in other enzyme associated Docs that is required to support the dual binding mode. Here, we report the structure of the protein complex between ScaC Coh and a group 3 Doc from *R. flavefaciens* FD-1. A comprehensive biochemical analysis guided by structural information confirmed that group 3 and 6 Docs present a single Coh binding interface. Since Docs of groups 3 and 6 are appended, essentially, to hemicellulases the data suggest that *R. flavefaciens* FD-1 has evolved an original molecular mechanism, using single binding mode Docs that exclusively interact to adaptor scaffoldins, to recruit this subset of highly important plant cell wall degrading enzymes to the cellulosome.

3.2. Experimental procedures

3.2.1. Gene synthesis and DNA cloning

Docs are inherently unstable when produced in *Escherichia coli*. To promote Doc stability, *R. flavefaciens* FD-1 Doc3 of protein ZP_06143424 (residues 888-952) was co-expressed *in vivo* with CohScaC. The immediate binding of Doc3 to CohScaC confers the necessary Doc stabilization. The genes encoding the two proteins were designed with a codon usage optimized to maximize expression in *E. coli*, synthesized *in vitro* (NZYTech Ltd, Lisbon, Portugal) and cloned into pET28a (Merck Millipore, Germany) under the control of separate T7 promoters. The Doc3-encoding gene was positioned at the 5' end and the CohScaC-encoding gene at the 3' end of the artificial DNA. A T7 terminator sequence (to terminate transcription of the Doc gene) and a T7 promoter sequence (to control transcription of the Coh gene) were incorporated between the sequences of the two genes. This construct contained specifically tailored NheI and NcoI recognition sites at the 5' end and XhoI and SalI at the 3' end to allow subcloning the nucleic acid into pET-28a (Merck Millipore, Germany) such that the sequence encoding a six-residue His tag could be introduced either at the N-terminus of the Doc (through digestion with

NheI and SalI, incorporating the additional sequence MGSSHHHHHHSSGLVPRGSHMAS at the N-terminus of the Doc3) or at the C-terminus of the CohScaC (by cutting with NcoI and XhoI, which incorporates the additional sequence LEHHHHHH at the C-terminus of the Coh). Thus, as a result of this strategy two pET28a plasmid derivatives were produced: pET28DtC with the engineered tag at the Doc and pET28DCt where the engineered tag is attached to the Coh. The two plasmids were used to express *Rf*CohScaC-Doc3 complexes in *E. coli*. Recombinant Doc3 and CohScaC primary structures are presented in Table 3.1.

Table 3.1 Recombinant protein sequences of CohScaC, Doc 3 and mutant variants produced for the interaction studies.

Dockerin	Protein Sequence
CohScaC	AGETVQISASNAEAKAGDQFEVKVSLADVPSTGIQGIDFAVTYDNTVVTIDKITVGEIADTKAASSDQTASLLPTF DVSIQNSEGYSSVIWSTAVEDSSYWISKDGVLCTITGTVSSNAKPGAESPIKLEA**VKRETYVGSGTDNSS**ISAGYS ANDKAVKYTVKATNGKISVPSAEV
CohScaC No Flap	MAGETVQISASNAEAKAGDQFEVKVSLADVPSTGIQGIDFAVTYDNTVVTIDKITVGEIADTKAASSDQTASLLPT FDVSIQNSEGYSSVIWSTAVEDSSYWISKDGVLCTITGTVSSNAKPGAESPIKLEAISAGYSANDKAVKYTVKATN GKISVPSAEV
Doc 3 WT	VYGDLDGDGEVDVFDLILMRKAVENGDTERFEAADLNCDGVIDSDDLTYHSEYLHGIRKTLPVE
Doc 3 F902A	VYGDLDGDGEVDV**A**DLILMRKAVENGDTERFEAADLNCDGVIDSDDLTYHSEYLHGIRKTLPVE
Doc 3 R908A	VYGDLDGDGEVDVFDLILM**A**KAVENGDTERFEAADLNCDGVIDSDDLTYHSEYLHGIRKTLPVE
Doc 3 H943A	VYGDLDGDGEVDVFDLILMRKAVENGDTERFEAADLNCDGVIDSDDLTYHSEYL**A**GIRKTLPVE
Doc 3 F902A/R908A	VYGDLDGDGEVDV**A**DLILM**A**KAVENGDTERFEAADLNCDGVIDSDDLTYHSEYLHGIRKTLPVE
Doc 3 F902A/H943A	VYGDLDGDGEVDV**A**DLILMRKAVENGDTERFEAADLNCDGVIDSDDLTYHSEYL**A**GIRKTLPVE
Doc 3 R908A/H943A	VYGDLDGDGEVDVFDLILM**A**KAVENGDTERFEAADLNCDGVIDSDDLTYHSEYL**A**GIRKTLPVE
Doc 3 F902A/R908A/H943A	VYGDLDGDGEVDV**A**DLILM**A**KAVENGDTERFEAADLNCDGVIDSDDLTYHSEYL**A**GIRKTLPVE

The underlined fragment represents the fraction missing in the No Flap variant of CohScaC.

To produce recombinant Cohs and Docs individually, an ELISA-based system designed to probe Coh-Doc affinities that requires fusion with xylanase or carbohydrate-binding modules (CBMs) was selected, as it allows production of highly stable and functional Coh and Doc derivatives (Barak *et al.*, 2005). Thus, sequences encoding Doc3 and CohScaC were amplified from *R. flavefaciens* FD-1 genomic DNA by PCR, using NZYProof polymerase (NZYTech Ltd, Portugal) and the primers shown in Table S3.1 (Annexes). After gel purification the Doc3 encoding amplicon was inserted into a Xylanase-Doc cassette in pET9d plasmid after digestion with KpnI and BamHI and ligation with T4-ligase. The resulting expressed product constitutes a His-tagged Doc3 fused to xylanase T-6 from *Geobacillus stearothermophilus* at the N-terminus of the polyhistidine tag (XynDoc3). The CohScaC encoding gene was cloned into a CBM-Coh cassette in pET28a after digestion with BamHI and XhoI restriction enzymes. This resulted in a His-tagged CohScaC recombinant derivative fused to a CBM3a from the *Clostridium thermocellum* scaffoldin ScaA (CBMCohScaC) (Handelsman *et al.*, 2004).

To identify the Doc residues that modulate Coh recognition, several XynDoc3 protein derivatives were produced using site directed mutagenesis. PCR amplification of the Doc

containing plasmid using the primers presented in Table S3.1 (Annexes), allowed the production of seven Doc3 protein derivatives, namely F902A, R908A, H943A, F902A/R908A, F902A/H943A, R908A/H943A and F902A/R908A/H943A. Each of the newly generated gene sequence was fully sequenced to confirm that only the desired mutation accumulated in the nucleic acid.

In order to remove the 15-residue *β*-flap present in *β*-strand 8, an overlapping PCR protocol was carried using the plasmid encoding CBMCohScaC as template. The two gene regions on each side of the 15 residue coding sequence (5' fragment and 3' fragment) were amplified in two separate reactions using the primers shown in Table S3.1 (Annexes). This resulted in the 3' end of the amplified 5' fragment being complementary to the 5' end of the amplified 3' fragment. The two fragments were then mixed at equimolar concentrations (0.15 pmol) and used as the template for a third PCR reaction using the forward primer from the first reaction and the reverse primer from the second. The resulting product was cloned back into pET21a by cutting with NheI/XhoI restriction enzymes and sequenced to confirm the integrity of the recombinant gene. The concentrations of each fragment were estimated in a NanoDrop 2000c spectrophotometer (Thermo Scientific, USA).

The genes encoding several Group 3 and 6 Docs were cloned using the Gateway recombination cloning technology (Thermo Scientific, USA). Sequences were amplified by PCR using *R. flavefaciens* FD-1 genomic DNA as template and using primers with engineered ends that allow site-specific recombination without the need for restriction enzymes (Table S3.2 Annexes). Amplified genes were inserted into pDONR201 entry vector and from there into the protein expression destination vector pETG-20A, according to the manufacturer's protocol (Thermo Scientific, USA). The genes are under the control of a T7 promoter. pETG-20A allowed the fusion of an N-terminal thioredoxin A (McCoy & La Ville, 2001) and an internal His tag to the recombinant Doc to promote protein stability and solubility.

3.2.2. Expression and purification of recombinant proteins

Preliminary expression screens revealed that when the polyhistidine tag was located at the Doc N-terminal end in *Rf*CohScaC-Doc3 complexes, the expression levels of both Coh and Doc were higher. Tagging the Coh resulted in the accumulation of large levels of unbound Coh in the purification product suggesting that Coh was expressed at higher levels than Docs. Consequently, the plasmid pET28DtC was used to transform *E. coli* BL21 (DE3) cells in order to produce *Rf*CohScaC-Doc3 complex in large quantities. Transformed *E. coli* were grown at 37ºC to an OD_{600} of 0.5. Recombinant protein expression was induced by the addition of 1 mM isopropyl *β*-D-1-thiogalactopyranoside followed by incubation at 19ºC for 16 hours. Cells were

harvested by 15 min centrifugation at 5000 x g and resuspendend in 20 mL of IMAC binding buffer (50 mM HEPES, pH 7.5, 10 mM imidazole, 1 M NaCl, 5 mM $CaCl_2$). Cells were then disrupted by sonication and the cell free supernatant recovered by 30 min centrifugation at 15,000 x g. After loading the soluble fraction into a HisTrap^{TM} nickel charged sepharose column (GE Healthcare, UK), initial purification was carried out by IMAC in an FPLC system (GE Healthcare, UK) using conventional protocols with a 35 mM imidazole wash and a 35-300 mM imidazole gradient. The buffer of all recovered fractions containing the purified Coh–Doc complex was exchanged into 50 mM HEPES, pH 7.5, containing 200 mM NaCl, 5 mM $CaCl_2$ using a PD-10 Sephadex G-25M gel-filtration column (Amersham Pharmacia Biosciences, UK). A further purification step by gel-filtration chromatography was performed by loading the samples onto a HiLoad 16/60 Superdex 75 (GE Healthcare, UK) at a flow rate of 1 ml min^{-1}. Fractions containing the purified complex were then concentrated with Amicon Ultra-15 centrifugal devices with a 10 kDa cutoff membrane (Millipore, USA) and washed three times with molecular biology grade water (Sigma) containing 0.5 mM $CaCl_2$. The protein concentration was estimated in a NanoDrop 2000c spectrophotometer (Thermo Scientific, USA) using a molar extinction coefficient (ε) of 26 025 M^{-1} cm^{-1}. The final protein concentration was adjusted to 81 $mg.mL^{-1}$ in molecular biology grade water containing 0.5 mM $CaCl_2$. The purity and molecular mass of the recombinant complex was confirmed by 14 % (w/v) SDS–PAGE.

Group 3 and 6 Docs, CBMCohScaC, XynDoc3 and respective protein derivatives used in ITC and native PAGE experiments were expressed as described above and purified with IMAC using nickel charged sepharose His GraviTrap gravity-flow columns (GE Healthcare, UK). After IMAC, the recombinant Coh and Docs were buffer exchanged to 50 mM HEPES pH 7.5, 0.5 mM $CaCl_2$ and 0.5 mM TCEP using PD-10 Sephadex G-25M gel filtration columns (GE Healthcare, UK).

3.2.3. Nondenaturing gel electrophoresis (NGE)

For the NGE experiments each of the XynDoc3 variants, at a concentration of 30 μM, was incubated in the presence and absence of 30 μM ScaCCoh for 30 min at room temperature and separated on a 10 % native polyacrilamide gel. Electrophoresis was carried out at room temperature. The gels were stained with Comassie Blue. Complex formation was detected by the presence of an additional band displaying a lower electrophoretic mobility than the individual modules.

3.2.4. Isothermal titration calorimetry

All ITC experiments were carried out at 308 K. The purified XynDoc3 variants and CohScaC were diluted to the required concentrations and filtered using a 0.45 µm syringe filter (PALL). During titrations the Doc constructs were stirred at 307 revolutions/min in the reaction cell and titrated with 28 successive 10 µL injections of CohScaC at 220 s intervals. Integrated heat effects, after correction for heats of dilution, were analyzed by nonlinear regression using a single-site model (Microcal ORIGIN version 7.0, Microcal Software, USA). The fitted data yielded the association constant (K_A) and the enthalpy of binding (ΔH). Other thermodynamic parameters were calculated using the standard thermodynamic equation: $\Delta RTlnK_A=\Delta G=\Delta H-T\Delta S$.

3.2.5. Crystallization, structural determination and refinement

The crystallization conditions were set up using the sitting-drop vapor-diffusion method with an Oryx8 robotic nanodrop dispensing system (Douglas Instruments, UK; (Bule, Ruimy-Israeli, *et al.*, 2014)). The commercial kits Crystal Screen, Crystal Screen 2, PEG/Ion and PEG/Ion 2 (Hampton Research, California, USA), JCSG+ HT96 (Molecular Dimensions,UK) and an in-house screen (80 factorial) were used for the screening. Precisely 0.7 µl drops of 40 and 81 mg ml^{-1} *Rf*CohScaC-Doc3 were mixed with 0.7 µl reservoir solution at room temperature per well containing 50 µl of the crystallization solution. The resulting plates were then stored at 292 K. Crystal formation was observed in 35 conditions after a period of approximately 30 days (maximum dimensions ~100 x 20 x 20 µm). All the crystals were obtained from the initial screens. These crystals were cryoprotected with mother solution containing 20–30 % glycerol or with 100 % Paratone-N (Hampton Research, USA) and flash-cooled in liquid nitrogen.

Data were collected on beamline I04 at the Diamond Light Source, Harwell, England using a PILATUS 6M detector (Dectris Ltd) from crystals cooled to 100 K using a Cryostream (Oxford Cryosystems Ltd). A systematic grid search was carried out on all of these crystals to select the best diffracting part of each crystal. EDNA (Winter & McAuley, 2011) and iMosflm (Battye, Kontogiannis, Johnson, Powell, & Leslie, 2011) were used for strategy calculation during data collection. All data sets were processed using the Fast_dp and xia2 (Winter, 2010) packages, which use the programs XDS (Kabsch, 2010), POINTLESS and SCALA (Evans, 2006) from the CCP4 suite (Winn *et al.*, 2011). Data-collection statistics are given in Table 3.2.

The best diffracting crystals were the ones formed in condition JCSG+ 2.33 [0.1 M potassium thiocyanate, 30 % (w/v) PEG 2000 MME] and diffracted to a resolution of 2.16 Å and belonged to the orthorhombic spacegroup $P2_12_12_1$. BALBES was used to carry out molecular replacement (Long, Vagin, Young, & Murshudov, 2008). The best solution was found using the type I Coh–

Doc complex from *C. thermocellum* (PDB entry 2CCL; (Carvalho *et al.*, 2007), the Coh of which displayed a sequence identity of 30.0 % and 31.7 % for its Doc, with an R factor and Rfree of 24.45 % and 30.58 %, respectively, and a Q-factor of 0.506 after REFMAC5 (Murshudov *et al.*, 2011) at the end of the BALBES run. Two copies of the heterodimer *Rf*CohScaC-Doc3 complex are present in the asymmetric unit. This model was adjusted and refined using REFMAC5 and PDB REDO (Joosten, Long, Murshudov, & Perrakis, 2014) interspersed with model adjustment in COOT to give the final model (Protein Data Bank code 5LXV, Table 3.2). The final round of refinement was performed using the TLS/restrained refinement procedure using each module as a single group. The root mean square deviation of bond lengths, bond angles, torsion angles and other indicators were continuously monitored using validation tools in COOT and MOLPROBITY. A summary of
the refinement statistics is shown in Table 3.2.

Table 3.2 X-ray crystallography data collection and refinement statistics for RfCohScaC-Doc3.

Data collection	
Beamline	*IO4-1, Diamond*
Space Group	*$P2_12_12_1$*
Wavelength (Å)	*0.920*
Unit-cell parameters:	
a, b c (Å)	*59.59, 66.73, 109.59*
α, β , γ (º)	*90, 90, 90*
Vm≠ ($Å^3$ Da^{-1})	*2.04*
Solvent Content (%)	*40*
Resolution limits (Å)	*57.00 - 2.40 (2.49 - 2.40)*
No. of observations	*68111 (7819)*
No. of unique observations	*17195 (1942)*
Multiplicity	*4.0 (4.0)*
Completeness (%)	*97.6 (99.3)*
<I/σ(I)>	*51.7 (5.0)*
$CC_{1/2}$†	*0.985 (0.936)*
Wilson B-factor	*27.51*
Rmerge ‡	*0.078 (0.328)*
Structure refinement	
R-work § , R-free ¥	*0.2170, 0.2600*
No. of Non-H atoms	*3658 (A: 1837 B: 1821)*
macromolecules	*3542 (A: 1771 B: 1771)*
ligands	*4 (A:2 B:2)*
water	*112 (A:64 B:48)*
Protein residues	*475 (A: 237 B: 238)*
RMS(bonds)	*0.009*
RMS(angles)	*0.95*
Ramachandran favored (%)	*97*
Ramachandran outliers (%)	*0*
Clash score	*0.57*
Average B-factor	*36.30*
macromolecules	*36.50*
ligands	*31.00*
solvent	*30.20*
PDB accession code	5LXV

Values in parenthesis are for the highest resolution shell.# Matthews coefficient (Matthews, 1968). † $CC_{1/2}$ = the correlation between intensities from random half-dataset (Diederichs & Karplus, 2013) ‡ $R_{merge} = \Sigma_{hkl} \Sigma_i |I_i(hkl) - \langle I(hkl) \rangle| / \Sigma_{hkl} \Sigma_i I_i(hkl)$, where $I_i(hkl)$ is the *i*th intensity measurement of reflection *hkl*, including symmetry-related reflections and $\langle I(hkl) \rangle$ is its average. §$R_{work} = \Sigma_{hkl} ||F_{obs}| - |F_{calc}|| / \Sigma_{hkl} |F_{obs}|$. ¥$R_{free}$ as R_{work}, but summed over a 5% test set of reflections.

3.3. Results and Discussion

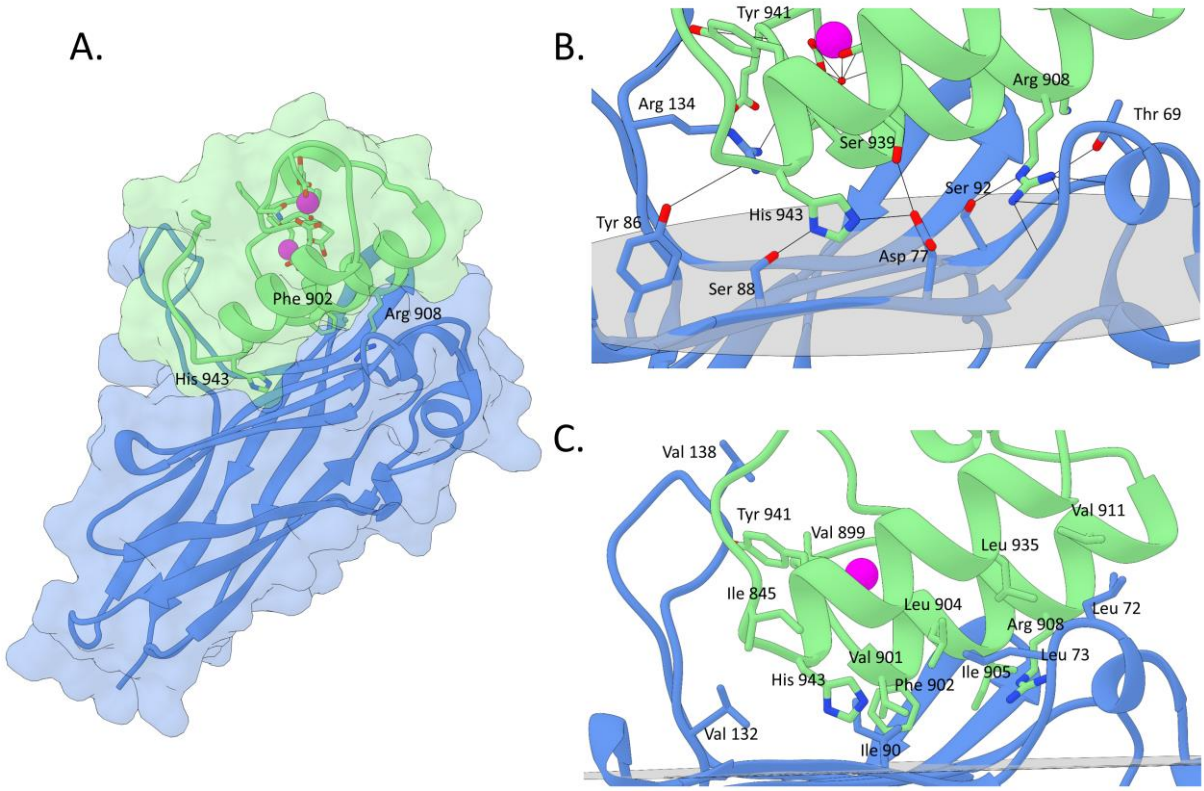
3.3.1. Expression and crystallization of a novel *R. flavefaciens* Coh-Doc complex

In a previous study (Israeli-Ruimy *et al.*, 2017), *R. flavefaciens* group 3 and 6 Docs were shown to bind specifically to the Coh of adaptor scaffoldin ScaC. Out of the 21 Docs selected for those studies, the Doc of protein WP_009985128 displayed the highest levels of expression. WP_009985128 contains a 730 residue long N-terminal X141 module of unknown function. X141 was expressed individually and its capacity to degrade a range of substrates was evaluated. The data revealed that X141 is unable to attack structural polysaccharides including pectins (data not shown). In addition, WP_009985128 also contains an internal family 6 carbohydrate-binding module (CBM6) and a C-terminal group 3 Doc (defined henceforth as Doc3). To gain insights into the molecular mechanisms of cellulosome assembly involving adaptor scaffoldins, the complex combining ScaC Coh (CohScaC) and Doc3 was expressed at high levels, purified and crystallized. Established strategies for the production and purification of Coh-Doc complexes, which involve the heterologous co-expression of both proteins in *E. coli* (Carvalho *et al.*, 2003), were employed, which allowed generating high quality crystals of the *Rf*CohScaC-Doc3 complex.

3.3.2. Structure of the *R. flavefaciens* CohScaC-Doc3complex

The crystal structure of *Rf*CohScaC-Doc3 complex was solved by molecular replacement using the structure of *C. thermocellum* (PDB code 2CCL; (Carvalho *et al.*, 2007)) type-I complex as a search model. The *Rf*CohScaC-Doc3 structure included 2 molecules of the heterodimer in asymmetric unit, as well as 118 water molecules, with each Doc coordinating two calcium ions. The dimer resulted from interactions established between two CohScaC modules. Thus, Nε2 of molecule A CohScaC's Gln-6 interacts with Oε1 of molecule B CohScaC's Gln-19 while Nζ of molecule A CohScaC's Lys-163 hydrogen bonds Oδ1 of molecule B CohScaC's Thr-114. The biological relevance of these crystallographic interactions, if any, is presently unclear. The *Rf*CohScaC-Doc3 complex displayed an elongated shape with overall dimensions of 30 × 35 × 60 Å and includes residues 3–174 from CohScaC and residues 889–953 of Doc3 from *R. flavefaciens* FD-1 (Figure 3.2). Crystal parameters and data collection statistics are summarized in Table 3.2.

Figure 3.2 Structure and Coh-Doc interface in the *R. flavefaciens* CohScaC–Doc3 complex.



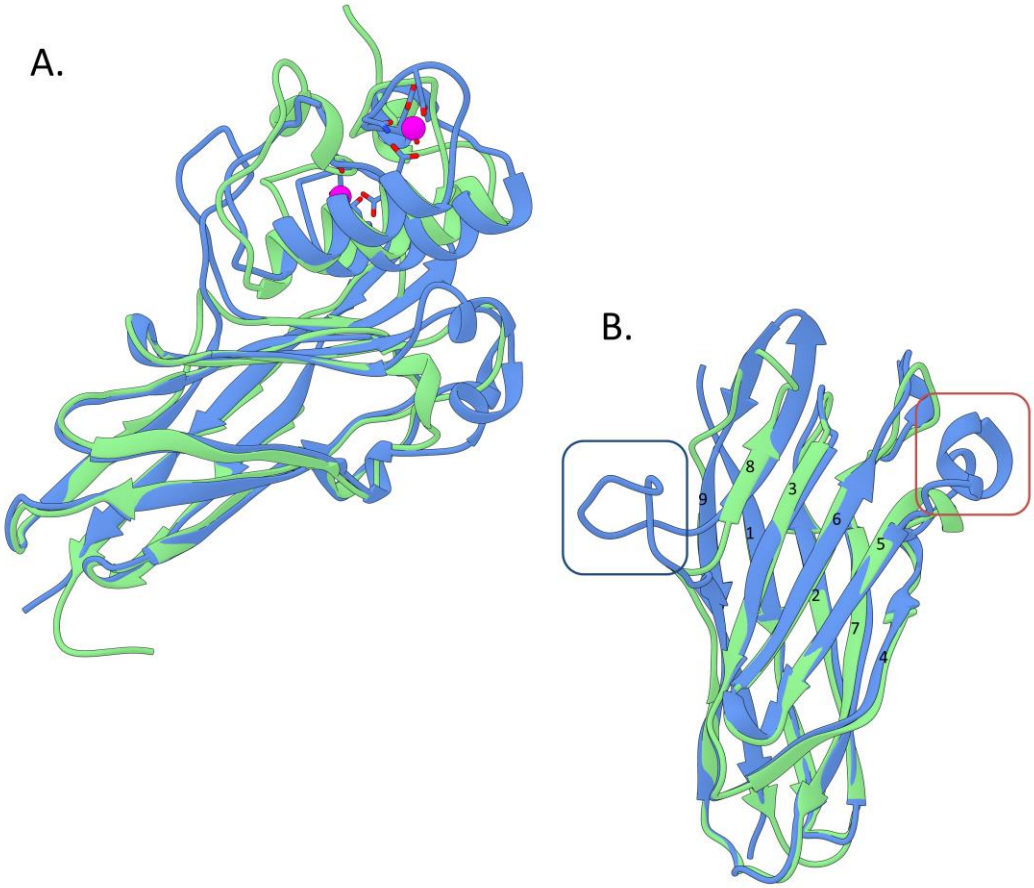
A. Structure of CohScaC-Doc3 complex with the Doc in green and the Coh in blue. Phe-902, Arg-908 and His-943 that dominate Coh recognition are labelled and shown as stick configuration. Ca^{2+} ions are depicted as purple spheres. *B.* Polar interactions at the complex interface. *C.* Hydrophobic interactions at the complex interface. The most important residues in both types of interaction are shown as sticks. The transparent grey disk in *B* and *C* marks the plane defined by the 8-3-6-5 *β*-sheet, where the *β*-strands form a distinctive Doc interacting plateau.

3.3.3. Structure of ScaC Coh

R. flavefaciens FD-1 CohScaC in complex with its cognate Doc3 displayed an elliptical structure comprising nine *β*-strands arranged in two *β*-sheets that form a *β*-barrel with the classic "jelly roll" topology (Figure 3.2). The two sheets are formed by *β*-strands 9, 1, 2, 7 and 4 on the non-interacting face and *β*-strands 8, 3, 6 and 5 on the Doc3 contacting face. All *β*-strands are antiparallel except for 1 and 9 which are parallel to each other and complete the jelly roll topology. Unusually, *β*-strand 8 is disrupted by a 17 residue long *β*-flap that extending from Ala-131 to Ile-147. Furthermore, two *α*-helices are present between *β*-strands 4 and 5 and *β*-strands 6 and 7. Although these motifs are somewhat similar to those observed previously for the type-II Cohs from *C. thermocellum*, *Bacteroides cellulosolvens* and *A. cellulolyticus* (PDB codes 3BM3, 1TYJ and 1QZN: SSM z-scores of 1.5, 6.3 and 7.2), structural similarity using SSM (Krissinel & Henrick, 2004) revealed that the closest functionally relevant structural

homologue of CohScaC was the type-I Coh from *A. cellulolyticus* ScaC (PDB code 4UYP for ScaCCoh-ScaBDoc complex) with a z-score of 11.4, a root mean square deviation (r.m.s.d) of 1.24 Å, over 136 aligned residues out of a possible 146, and a total sequence identity of 30%. However, the α-helix connecting β-strands 4 and 5 is longer in *R. flavefaciens* CohScaC and the *Acetivibrio* homologue lacks the large insertion identified in β-strand 8 of CohScaC (Figure 3.3). *C. thermocellum* ScaA Coh is the second closest structural homologue to CohScaC (PDB code 2ccl for ScaA Coh complexed with a Doc) with a z-score of 11, an r.m.s.d of 1.39 Å over 130 aligned residues out of a possible 149 and a total sequence identity of 27%. CohScaC also shows homology with Cohs of *C. cellulolyticum*, *C. perfringens*, *B. cellulosolvens*, and the Coh from *R. flavefaciens* ScaE (r.m.s.d >1.8 Å; sequence identity <20%). Secondary structure comparison of CohScaC with representative members of other Cohs with different specificities revealed the distinctive features of the *R. flavefaciens* ScaC Coh to be the well-defined α-helix connecting β-strands 4 and 5 and the β-flap disrupting β-strand 8 (Figures 3.3*B* & 3.4).

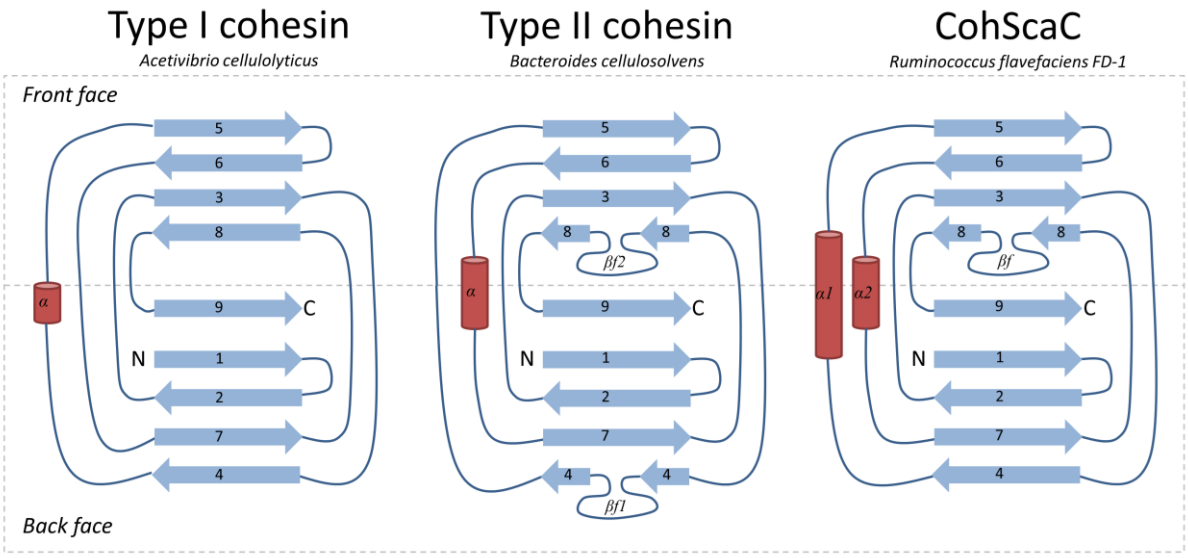
Figure 3.3 Overlay of the *R. flavefaciens* CohScaC–Doc3 complex with the *A. cellulolyticus* type-I Coh-Doc complex



A. An overlay of CohScaC–Doc3 (depicted in blue) with the *Ac*ScaCCoh3-ScaBDoc type-I complex from *A. cellulolyticus* (depicted in green, PDB code 4UYP), with the Docs rotated 180° relative to each other, showing the high degree of overall similarity. *B.* Overlay of both Cohs isolated from the

complexes and rotated approximately 90º down and right relatively to *A*, with the Doc interacting plateau in the first plane. This view highlights the main differences between the two Cohs which consist in the large *β*-flap extension that interrupts *β*-strand 8 (dark-blue box) and the well-defined *α*-helix connecting *β*-strands 4 and 5 (red box). These two structural elements together with the loop formed by the distal part of *β*-strand 8 and the proximal section of *β*-strand 9 that is tilted towards the Doc, form a claw like interaction interface.

Figure 3.4 Topology diagram of CohScaC compared with previously described type-I and type-II Cohs.



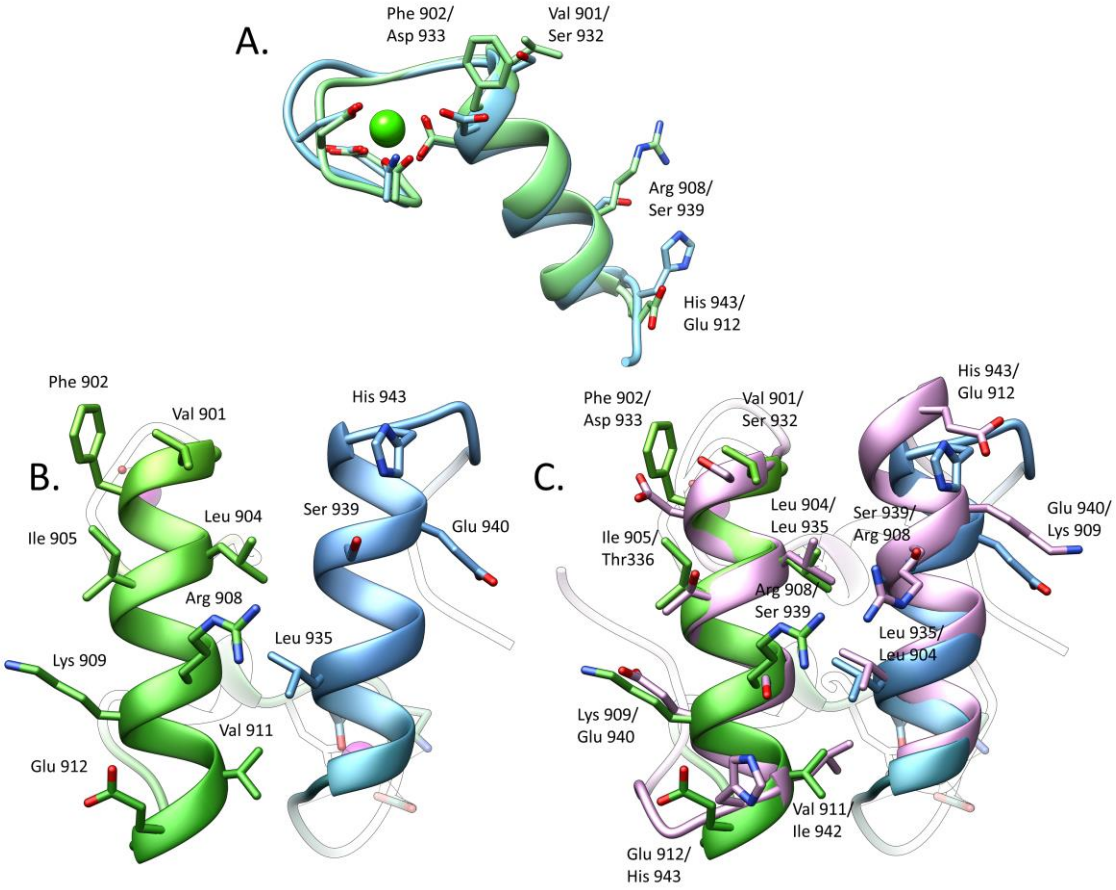
The CohScaC module (right) forms the classical nine-stranded *β*-sandwich with jelly-roll topology, which is essentially analogous to that of the type-I and type-II Coh modules: *Ac*ScaCCoh3 (left, PDB code 4UYP) and *Bc*ScaACoh11 (middle, PDB code 1TYJ), respectively. Like the type II example, CohScaC presents a *β*-flap extension that interrupts *β*-strand 8 and a *α*-helix between *β*-strands 6 and 7. In addition, it also possesses an *α*-helix connecting *β*-strands 4 and 5, much like the *A. cellulolyticus* type I but much better defined. The exuberant nature of this *α*-helix together with the pronounced *β*-flap are the two distinctive features of the *R. flavefaciens* ScaC Coh.

3.3.3.1. Structure of *R. flavefaciens* FD-1 group 3 dockerin (Doc3)

Within the complex, Doc3 comprises two *α*-helices arranged in an antiparallel orientation extending from Val-901 to Asn-913 (helix-1) and Ser-932 to His-943 (helix-3), while the loop connecting these structural elements contains a four residue *α*-helix (helix-2) extending from Phe-919 to Ala-922 (Figure 3.2). The overall tertiary structure of Doc3 is very similar to enzyme Docs from *C. thermocellum* (r.m.s.d of 0.9 Å) and *A. cellulolyticus* (r.m.s.d of 1.4 Å), which display a dual binding mode. Doc3 contains two Ca^{2+} ions coordinated by several amino-acid residues, similar to the canonical EF-hand loop motif. Both of the Ca^{2+} ions have an *n, n+2, n+4, n+11* plus a water molecule pattern of coordination. Thus, the Ca^{2+} ion located at the N-terminus is coordinated by the side chains of Asp-892, Asp-894, Asp-896 and Asp-903 (both the Oδ1 and Oδ2), the latter belonging to *α*-helix 1. The octahedral geometry of the coordination is completed by the main chain carbonyl of Glu-908 and one water molecule. The

second Ca^{2+} site stabilizes the loop connecting α-helices 2 and 3 and is coordinated by the side chains of Asp-923, Asn-925, Asp-927 and Asp-934 (both the Oδ1 and Oδ2) as well as the carbonyl from Val-929 and a water molecule. A structural overlay of the two duplicated sequences observed in Doc3, indicated they are structurally similar with an r.m.s.d of 0.8 Å for all main-chain atoms (Figure 4*A*).

Figure 3.5 Significant differences between the two Coh binding interfaces do not allow the dual binding mode of type-I Doc from *R. flavefaciens*.



A. An overlay of the two Doc repeats observed in Doc3 showing that the structures are similar (r.m.s.d of 0.82 Å) for the main chain atoms but have considerable differences in the side-chains. *B.* The two interacting helices of Doc3, helix 1 (bright green) and helix 3 (dark green), with the most important Coh recognition residues displayed as sticks. *C.* Comparison of the two putative binding surfaces by overlaying Doc3 with a version of itself rotated by 180º (in pink) and shows a lack of conservation in the key contacting residues. Lack of internal symmetry in Doc3 and the involvement of the two helices in Coh recognition suggest that Doc3 displays a single Coh-binding platform.

3.3.4. *Rf*CohScaC-Doc3 complex interface

Doc3 interacted with the 8-3-6-5 sheet of the ScaCCoh β-sandwich, which presents a predominantly flat surface. However, the C-terminus of β-strand 8 is elevated in relation to the 8-3-6-5 plane which enables the N-terminus of β-strand 9 to interact with Doc3. The β-flap on

one side of the CohSacC 8-3-6-5 sheet and the α-helix, between β-strands 4 and 5, on the other side generate the appropriate topology at the surface of the Coh to accommodate Doc3. A large network of polar (Table 3.3) and hydrophobic interactions (Table 3.4) were identified at the complex interface.

Table 3.3 Main polar contacts between CohScaC and Doc3.

Doc3					CohScaC		
	Atom	Residue	Residue #		Atom	Residue	Residue #
Hydrogen Bonds							
	N	Val	899	< >	O	Thr	136
H1	NH1	Arg	908	< >	OG1	Thr	69
H1	NH1	Arg	908	< >	O	Leu	73
H1	NH2	Arg	908	< >	O	Leu	73
H1	NH2	Arg	908	< >	O	Pro	74
H1	NE	Arg	908	< >	OG	Ser	92
H1	NZ	Lys	909	< >	O	Ala	94
H3	OG	Ser	939	< >	OD2	Asp	77
H3	OH	Tyr	941	< >	O	Glu	135
H3	O	Leu	942	< >	NH1	Arg	134
H3	NE2	His	943	< >	OD2	Asp	77
H3	O	His	943	< >	OH	Tyr	86
H3	ND1	His	943	< >	OG	Ser	88
Salt Bridges							
	OD2	Asp	896	< >	NZ	Lys	159
	OE1	Glu	898	< >	NZ	Lys	159
	OD2	Asp	900	< >	NZ	Lys	159
H1	NZ	Lys	909	< >	OE2	Glu	96
H1	NZ	Lys	909	< >	OD1	Asp	155
H3	NE2	His	943	< >	OD1	Asp	77

Table was made using the PDBePISA server and the contacts were further verified manually with Coot. Some of the Doc residues are marked as belonging either to helix 1 (H1) or to helix 3 (H3) interfaces.

Their total number is greater than that observed in any related clostridial Coh-Doc complex that involves the recruitment of enzymes into clostridial cellulosomes [defined as type I doc-Coh pairs (Brás *et al.*, 2012; Cameron, Weinstein, *et al.*, 2015)]. In these dual binding mode Docs the C-terminal region of one of the helices interacts with the Coh, while the entire length of the second interacting helix binds to the protein ligand. Doc binding can switch and, as a result of a 180º rotation of the Doc on the Coh surface, the Doc helix with the previous lower number of contacts can dominate Coh recognition, supporting the well described dual binding mode. In contrast, in the *Rf*CohScaC-Doc3 complex the two Doc3 helices (helix 1 and helix 3) make similar contributions to CohScaC recognition (Table 3.3 and Table 3.4).

Table 3.4 Main hydrophobic contacts between CohScaA and Doc3

Doc3					CohScaC
		Residue	Residue #		Residues
		Asp	894	< >	Ala 157
		Asp	896	< >	Lys 159
		Gly	897	< >	Tyr 137
		Glu	898	< >	Thr 136, Lys 159
		Val	899	< >	Arg 134, Thr 136
		Asp	900	< >	Arg 134, Lys 159
H1		Val	901	< >	Ile 90, Ser 148
H1		Phe	902	< >	Ala 149, Gly 150, Tyr 151, Ala 157, Lys 159
H1		Leu	904	< >	Ser 92
H1		Ile	905	< >	Gln 35, Gly 36, Thr 93
H1		Leu	906	< >	Asp 155
H1		Arg	908	< >	Thr 69, Leu 73, Pro 74, Ser 92, Ala94
H1		Lys	909	< >	Ala 94, Glu 96, Asp 155
H1		Val	911	< >	Leu 72
H1		Glu	912	< >	Gln 68, Ala 94
	H3	Leu	935	< >	Leu 73
	H3	Ser	939	< >	Leu 73, Asp 77
	H3	Tyr	941	< >	Arg 134, Glu 135, Tyr 137, Val 138
	H3	Leu	942	< >	Ile 90, Val 132, Arg 134
	H3	His	943	< >	Asp 77, Ser 79, Tyr 86, Ser 88, Ile 90, Val 132
		Gly	944	< >	Tyr 86
		Ile	945	< >	Ser 79, Gln 81
		Leu	949	< >	Val 138

Table was made using the PDBePISA server. Some of the Doc residues are marked as belonging either to helix 1 (H1) or to helix 3 (H3) interfaces.

The elevation of the α-helix located between β-strands 4 and 5 of CohScaC over the plane of the protein interacting surface allows the entire Coh surface to be in closer proximity to both Doc α-helices. This observation together with the lack in symmetry of the binding residues, which is described below, suggests that, in contrast to what was previously observed in several Coh-Doc complexes involving enzyme recruitment, Doc3 presents a single binding mode.

The interactions between α-helix-1 of Doc3 and CohScaC are dominated by Val-901, Phe-902, Ile-905, Arg-908 and Lys-909 of Doc3, and Gln-35 and Ser-92 of CohScaC (Figure 3.2). The side chains of the Val-901/Phe-902 pair, occupying positions 11 and 12 of Doc3 that were previously suggest to modulate specificity in type-I interactions (Cameron, Weinstein, *et al.*, 2015), lie in the hydrophobic pocket formed by CohScaC residues Gly-36 and Gly-150. The hydrophobic character of α-helix-1 interaction is reinforced by the interaction of Ile-905 with CohScaC Gln-35. The more distal α-helix-1 Arg-908 and Lys-909 pair contributes to the hydrogen-bond network with CohScaC by contacting residues Leu-72, Ser-92 and Glu-96, while the aliphatic side chains of these residues make comprehensive hydrophobic contacts with CohScaC Ala-94. In addition, the Nη1 of Lys-909 contributes two important salt bridges with Oδ2 of Glu-96 and Oδ1 of Asp-155 of the CohScaC. In α-helix-3 the contacts are dominated by the important salt bridge established between Nε2 of His-943 and Oδ1 of

CohScaC Asp-77. His-943 also establishes important hydrogen bonds with Tyr-86 and Ser-88 of the Coh. In addition, the side chains of Leu-935 and Leu-942 make non-polar contacts with CohScaC amino acid residues Leu-73 and Ile-90, respectively.

One of the notable features of CohScaC is the presence of an extensive loop disrupting *β*-strand 8. Residues located at this loop make a significant number of contacts with Doc3. Thus, Tyr-941 located in *α*-helix-3 of Doc3, is hydrogen bonded to the carbonyl group of CohScaC loop residue Glu-135, while Leu-942 makes a polar contact with CohScaC Arg-134. Furthermore, Val-899 located at the N-terminus of Doc3 forms a hydrogen bond with CohScaC Thr-136. Additional van der Waals interactions are established between CohScaC loop residues Val-132 and Val-138 with Doc3 Gly-944 and Leu-949. Strikingly, residues located at the C-terminus of CohScaC *β*-strand 8 and the N-terminus of *β*-strand 9 make important contributions for Doc3 recognition. The twisted conformation of these two *β*-strands provides a platform that binds Doc3 amino acids located at the N-terminal loop. Hence, the Nη1 of Lys-159 located in *β*-strand 9 makes three important hydrogen bonds with Asp-896, Glu-898 and Asp-900, which are Doc3 residues participating in the coordination of the calcium ion of the first Doc repeat. In addition, *β*-strand 9 Ala-157 and the aliphatic chain of Lys-159 provide an important hydrophobic environment to accommodate the Doc3 side chain of Phe-902. Collectively, these observations suggest an extensive interface in *Rf*CohScaC-Doc3 complex not previously observed in type-I Coh-Doc interaction.

3.3.5. Doc3 presents a single Coh-binding interface

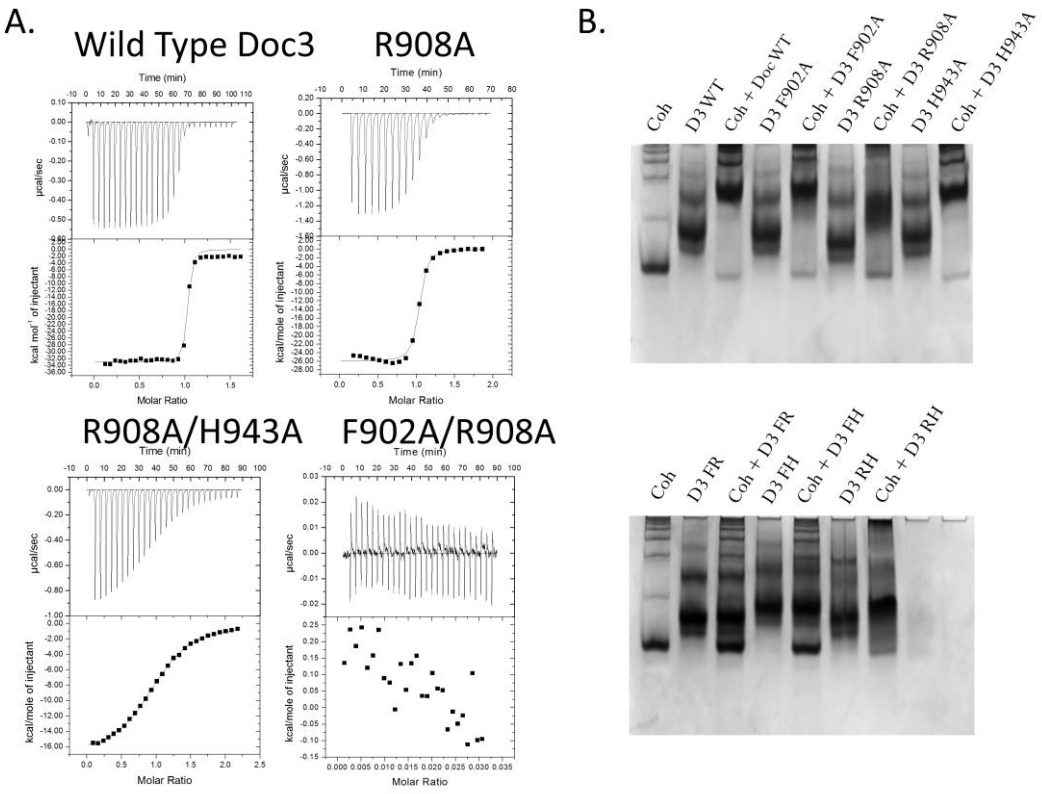
The binding thermodynamics of Doc3 to CohScaC were assessed by isothermal titration calorimetry (ITC) at 308 K, consistent with the approximate temperature of rumen. The data, presented in Table 3.5 and exemplified in Figure 3.6, revealed a macromolecular association with a stoichiometry of 1:1 and a K_a of ~ 10^7 M^{-1}, an affinity similar to other type-I interactions. It is noteworthy that the apparent hydrophobic nature of the CohScaC-Doc interaction is associated with an enthalpy driven interaction, a property previously observed in other Coh-Doc complexes. The importance of Doc3 Phe-902, Arg-908 and His-943 for CohScaC recognition was also probed by ITC. The data (Table 3.5, Figure 3.6) revealed that alanine substitutions of residues Phe-902 and His-943 had no effect in the affinity of Doc3 for its Coh partner. In contrast, the R908A Doc3 derivative displayed a 10-fold lower affinity for the CohScaC (Table 3.5, Figure 3.6).

Table 3.5 Thermodynamics of interaction between wild type CohScaC and wild type and mutant variants of Doc3.

Dockerin	$K_a\ M^{-1}$	ΔG^o *kcal mol*$^{-1}$	ΔH *kcal mol*$^{-1}$	$T\Delta S^o$ *kcal mol*$^{-1}$
Doc3 WT	2.69E8 ± 2.52E7	-11.89	-36.33 ± 0.055	-24.48
Doc3 F902A	1.10E8 ± 1.20E7	-11.415	-22.20 ± 0.087	-10.79
Doc3 R908A	3.07E7 ± 5.53E6	-10.55	-25.95 ± 0.238	-15.40
Doc3 H943A	1.58E8 ± 1.21E7	-11.54	-25.98 ± 0.062	-14.44
Doc3 F902A/R908A	*Nb*	*Nb*	*Nb*	*Nb*
Doc3 F902A/H943A	*Nb*	*Nb*	*Nb*	*Nb*
Doc3 R908A/H943A	6.86E5 ± 1.75E4	-8.24	-17.11 ± 0.086	-8.87
Doc3 F902A/R908A/H943A	*Nb*	*Nb*	*Nb*	*Nb*
Doc3 WT vs CohScaC NF	1.07E8 ± 1.45E7	-11.31	-16.12 ± 0.064	-4.80

The last row refers to the interaction between wild type (WT) Doc3 and CohScaC without the flap insertion (NF). All Thermodynamic parameters were determined at 308 K.

Figure 3.6 Determination of Doc3 Phe-902, Arg-908 and His-943 importance for CohScaC recognition.



A. Representative binding isotherms of the interaction between CohScaC/Doc3, CohScaC/Doc3 R908A mutant, CohScaC/Doc3 R908A/H943A double mutant and CohScaC/Doc3 F902A/R908A double mutant. The upper part of each panel shows the raw heats of binding, whereas the lower parts comprise the integrated heats after correction for heat of dilution. The curve represents the best fit to a single-site binding model. The corresponding thermodynamic parameters are shown in Table 3. *B.* Nondentauring gel electrophoretic analysis of CohScaC-Doc3 interaction. The first lane both gels were loaded with the cohesin (Coh). Adjacent lanes were loaded with the Doc (D3) and with both Coh and Doc modules together after 60-min incubation at equimolar concentrations. The appearance of a band with a different migration pattern in lanes containing the complex represents a positive result (e.g. D3 WT), while a negative result (e.g. D3 FR) is given by the appearance of only the individual Doc and Coh bands. A

faint Coh band is seen even in the lanes were there is complex formation which results from excess Coh probably due to not all the Doc in solution being active.

The fact that single amino acid substitutions at the Coh-binding surface of Doc3 had a marginal or no effect on Coh recognition may be accounted for by at least two explanations: (1) Doc3 displays a dual binding mode typical of other Docs and mutation of a single residue has no effect in affinity as it leads to a 180º rotation of the Doc and the presence of the mutated residue is compensated by its 2-fold symmetry related counterpart; or (2) Doc3 presents a single CohScaC binding platform so extensive that single substitutions have marginal effects on affinity. To distinguish between these two possibilities, we probed the internal symmetry of Doc3 by overlaying its structure with the 2-fold related derivative using the Matchmaker procedure from Chimera (27), which showed an r.m.s.d of 0.36 Å for 115 atoms (Figure 3.5*A,B*). The superposition highlights the lack of conservation in the contacting residues when the two putative Coh-binding surfaces were compared (Figure 3.5*C*). For example, the key Val-901/Phe-902 pair located at position 11 and 12 of the first repeat is replaced by Ser-932 and Asp-933, while His-943, that dominates the hydrogen bond network with the Coh at the C-terminal α-helix, superposes with a Glu-912 (Figure 3.5*C*). The lack of internal symmetry in Doc3 and the involvement of α-helices 1 and 3 in Coh recognition confirm that Doc3 displays a single Coh-binding platform. Thus, the importance of Phe-902, Arg-908 and His-943 in binding of CohScaC was investigated by probing the capacity of double and triple mutant derivatives to recognize the CohScaC. The data (Table 3.3) suggests that although *Rf*CohScaC-Doc3 complex presents an extensive protein-protein interface, Doc3 Phe-902, Arg-908 and His-943 dominate Coh recognition, as replacement of these residues by Ala in double and triple mutants significantly diminishes or abrogates binding. It is noteworthy that ΔG values for the single mutant interactions do not vary much from the ones observed for the wild type interaction. On the other hand double and triple mutant interactions have a significant decrease in ΔG values. This means the $\Delta\Delta G$ values between wild-type and multiple mutant interactions is greater than the $\Delta\Delta G$ obtained by adding the several $\Delta\Delta G$ values between the wild-type and the several single-mutant interactions. One explanation might be that the reduced affinity may also result from conformational changes in addition to the lack of contacts made by the mutated side-chains. Further support for the single binding-mode involving several interacting residues is provided by the observation that removal of the CohScaC loop that interrupts β-strand 8, which makes several contacts with Doc3, had no influence in affinity (Table 3.3).

3.3.6. *R. flavefaciens* FD-1 Group 3 and Group 6 Docs present a non-dynamic binding mode to CohScaC

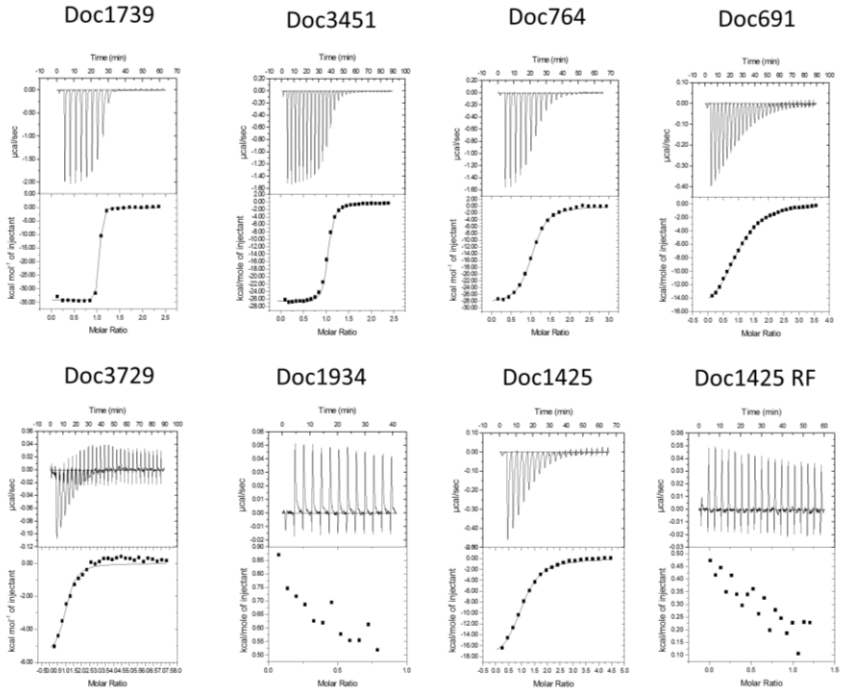
Recent data suggests that *R. flavefaciens* FD-1 Group 3 and 6 Docs display tight specificity for CohScaC (Israeli-Ruimy *et al.*, 2017). Prevalence of xylan (GH10, GH11 and GH43), and pectin (PL11, CE1, CE3 and CE15) degrading catalytic modules (and associated carbohydrate binding CBM22 and CBM6 modules) in *R. flavefaciens* FD-1 cellulosomal proteins containing Group 3 Docs suggests that this subset of enzymes is particularly suited to deconstruct hemicellulose and pectin (Figure 3.7). In addition, Group 6 Docs are appended to a broader range of enzymes that include GH5, GH26, GH43, GH44, GH97, PL1, PL11, CE1, CE3 and CE4. The structure of *Rf*CohScaC-Doc3 complex provides an opportunity to identify the residues that modulate ligand specificity within these two Doc groups.

Figure 3.7 Modular architecture of group 3 dockerin bearing proteins.

Prevalence of GH10, GH11, GH43s (in red), PLs (in blue) and CEs (in orange), in addition to CBM6 and CBM22 (in green), in *R. flavefaciens* FD-1 cellulosomal proteins containing group 3 Docs, suggest that this subset of enzymes is particularly suited to deconstruct hemicellulose and pectin. Doc modules are displayed in yellow.

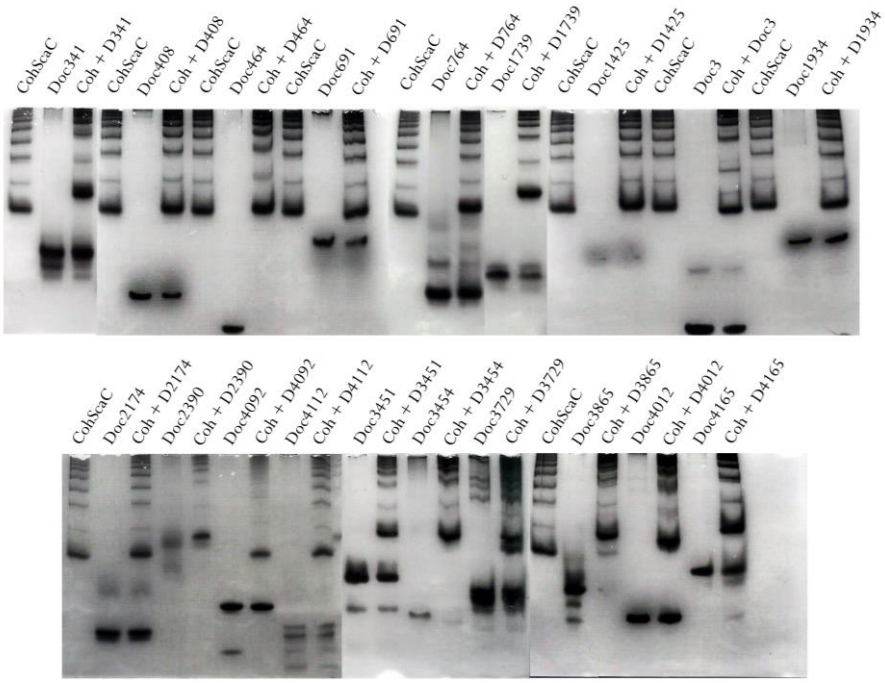
All twenty *R. flavefaciens* group 3 Docs were expressed and purified. From a total of 19 recombinant Docs (one of the Docs was insoluble when produced in *E. coli*), 16 were shown to bind to CohScaC using ITC (Figure 3.8) and nondenaturing gel electrophoresis (Figure 3.9).

Figure 3.8 Binding affinity of group 3 Docs to CohScaC determined by ITC.



These examples show the different degrees of affinity among the several group members. The bottom right binding isotherm results from the experiment using the F46A/R52A mutant of Doc1425 confirming the hypothesis of an opposite orientation binding mode.

Figure 3.9 Binding affinity of group 3 dockerins to CohScaC determined by NGE.



The lanes marked Coh were loaded with the Coh. Adjacent lanes were loaded with the Docs and with both Coh and Doc modules together after 60-min incubation at equimolar concentrations. The

appearance of a band with a different migration pattern in lanes containing the complex represents a positive result (e.g. Doc3865), while a negative result (e.g. Doc1934) is given by the appearance of only the individual Doc and Coh bands. Some results were difficult to interpret and therefore inconclusive. The ITC experiments (Table 3.6 and Figure 3.8) were able to clear up any doubts left from the NGE experiments.

Strikingly, the data (Table 3.6 and Figure 3.6) revealed that the affinities of the Group 3 Docs ranged from K_a $<10^5$ to 10^8 M^{-1}. Inspection of the alignment of the *R. flavefaciens* FD-1 Group 3 Docs revealed that the three members which did not bind CohScaC lack the three residues that were shown to dominate CohScaC recognition (Figure 3.10). Moreover, in two cases (Doc381 and Doc1425) the Docs would appear to recognize CohScaC in the reverse orientation relative to Doc3 since the three residues involved in Coh recognition (i.e., Phe-902, Arg-908 and His-943) are identified in the opposite helices to those of Doc3. To verify this possibility, Phe and Arg residues observed in α-helix 3 of Doc1425 were mutated to Ala and the affinity of the mutant Doc for CohScaC determined. The data, presented in Table 3.6, indicated that Phe-46Ala/Arg-52Ala mutant displays no affinity for CohScaC, suggesting that α-helix-3 should occupy the position of Doc3 α-helix-1 during Coh recognition. Variation in CohScaC affinities may be explained by replacement of at least one of the three residues important for Coh recognition by a non-conserved homologue. For example, Doc3729 and Doc3865, which display the lowest affinities for CohScaC, have His replaced by a Ser and Phe substituted by a Met (Figure 3.10).

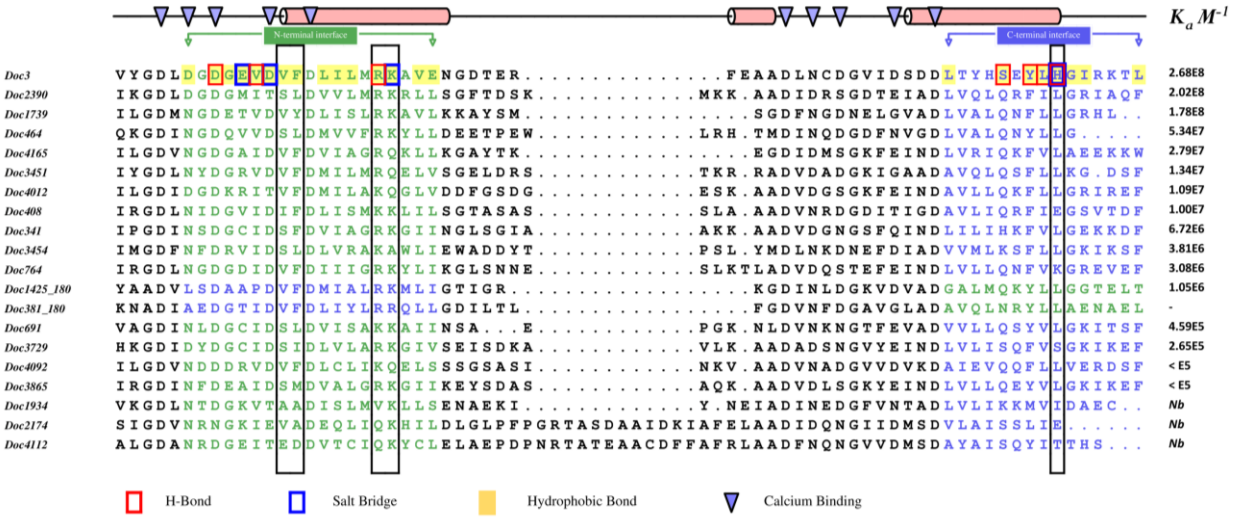
Table 3.6 Thermodynamics of interaction between wild type CohScaC and Docs from groups 3 and 6.

Group 3 Dockerin	K_a M^{-1}	ΔG^o kcal mol^{-1}	ΔH kcal mol^{-1}	$T\Delta S^o$ kcal mol^{-1}
Doc3	2.69E8 ± 2.52E7	-11.89	-36.33 ± 0.055	-24.48
Doc2390	2.02E8 ± 2.65E7	-11.76	-23.17 ± 0.05	-11.40
Doc1739	1.79E8 ± 3.89E7	-11.81	-34.31 ± 0.17	-22.49
Doc464	5.34E7 ± 1.90E7	-10.88	-18.96 ± 0.30	-8.07
Doc4165	2.79E7 ± 5.79E6	-10.52	-44.11 ± 0.57	-33.58
Doc3451	1.34E7 ± 6.99E5	-10.05	-26.94 ± 0.09	-16.88
Doc4012	1.09E7 ± 7.00E5	-9.93	-39.70 ± 0.16	-29.76
Doc408	1.00E7 ± 2.08E6	-9.86	-26.66 ± 0.51	-16.79
Doc341	6.72E6 ± 1.22E6	-9.64	-35.71 ± 0.61	-26.06
Doc3454	3.81E6 ± 7.43E5	-9.53	-31.10 ± 0.73	-21.57
Doc764	3.08E6 ± 1.92E5	-9.16	-28.82 ± 0.24	-19.66
Doc1425	1.05E6 ± 9.02E4	-8.50	-19.47 ± 0.50	-10.97
Doc691	4.59E5 ± 2.06E4	-7.97	-16.97 ± 0.20	-8.99
Doc4092	2.65E5 ± 3.92E4	-7.63	-17.03 ± 1.2	-9.398
Doc3729	*< E5*	-	-	-
Doc3865	*< E5*	-	-	-
Doc1934	*Nb**	-	-	-
Doc2174	*Nb**	-	-	-
Doc4112	*Nb**	-	-	-
Doc1425 F46A /R52A	*Nb**	-	-	-
Group 6 Dockerin	K_a M^{-1}	ΔG^o kcal mol^{-1}	ΔH kcal mol^{-1}	$T\Delta S^o$ kcal mol^{-1}
Doc903	*Id†*	*Id†*	*Id†*	*Id†*
Doc1965	1.13E7 ± 1.02E6	-9.49	-35.68 ± 0.28	-26.19

Doc1369	1.07E7 ± 4.25E5	-9.90	-23.77 ± 0.07	-13.86
Doc1804	*< E5*	-	-	-
Doc2712	1.33E5 ± 6.59E3	-7.10	-84.14 ± 8.76	-77.03

The last row of group 3 refers to the interaction between the F46A /R52A mutant variant of Doc 1425 and CohScaC. All Thermodynamic parameters were determined at 308 K. *† Impossible to determine:* Binding too strong to accurately calculate the thermodynamic parameters. * *No binding*: the corresponding Doc did not bind CohScaC

Figure 3.10 Alignment of Group 3 dockerins.

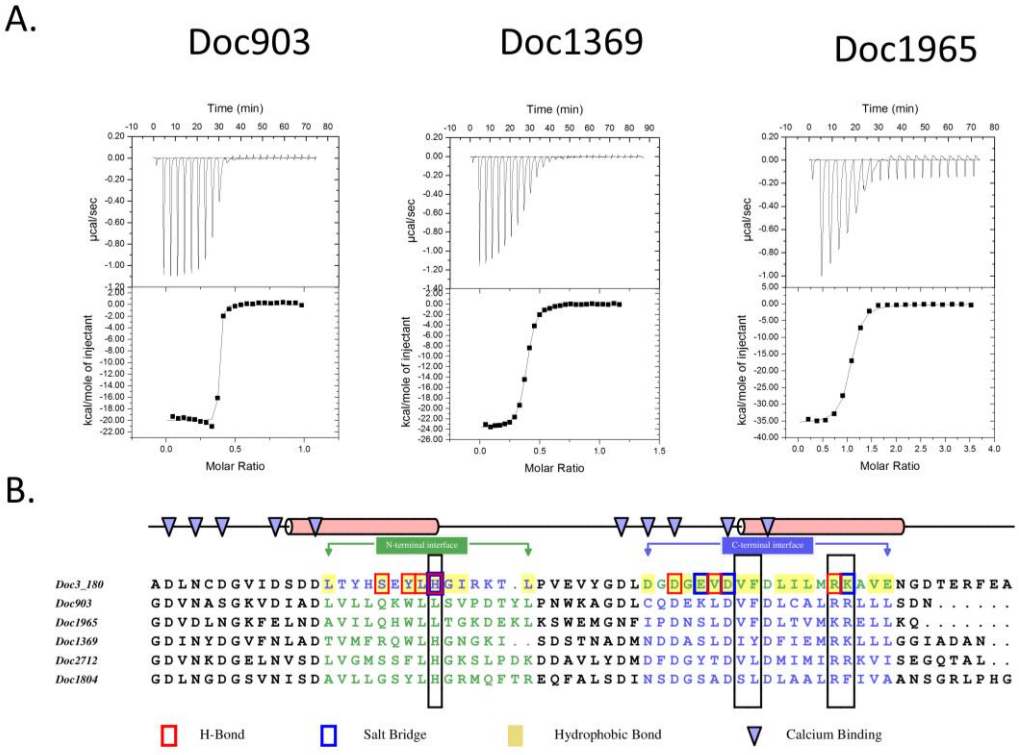


All group 3 Docs were aligned using Clustal Omega multiple sequence alignment software and are organized according to their affinity to CohScaC, from the highest Ka value to the lowest, as determined by ITC. Docs 381 and 1425 were aligned in opposite orientation relative to the other members by switching the N-terminal half with the C-terminal half of the sequence. This resulted in the N-terminal interface (blue residues) of Docs 381 and 1425 being perfectly aligned with the C-terminal interface (green residues) of the remaining group 3 members and vice-versa, supporting the theory that they will bind the Coh in an opposite orientation. The top line matches the protein secondary structure (red cylinders) to the primary structures, as observed in the Doc3 structure and also points to the calcium coordinating residues (blue triangles). All residues involved in the Doc3 interaction with CohScaC are highlighted according to the colour code displayed at the bottom. Conservation of key residues for Coh recognition along the group is highlighted with black boxes. To some extent, this conservation pattern seems to correlate to the CohScaC affinity profile of the group.

Similarly to what was observed for Group 3 Docs, the majority of Group 6 Doc-containing enzymes were previously annotated as hemicellulases (Rincon *et al.*, 2010). To understand why Group 3 and 6 Docs have similar Coh specificities, six representative members of *R. flavefaciens* FD-1 Group 6 Docs (out of a total of 45) were produced recombinantly and their affinities for CohScaC probed through ITC. The thermodynamic data, displayed in Table 3.6 as well as the binding thermograms presented in Figure 3.11*A*, revealed that Group 6 Docs also display significant differences in affinity for CohScaC. For example, the affinity of Doc903

was beyond the detectable limit of the calorimeter suggesting a $K_a > 10^9$ M^{-1} while the other five Docs displayed affinities that were either too low to be quantified (Doc1804) or had a K_a that ranged from 10^5-10^7 M^{-1}. To rationalize these observations the Group 6 Docs were aligned with Doc3 (Figure 3.11*B*). Residues required for CohScaC recognition are observed in the opposite helices when compared with Doc3, suggesting that in Group 6 Docs α-helix 3 interacts with CohScaC similar to α-helix 1 of Doc 3. In addition, insertion of a Ser residue at position 11 that is usually occupied by a hydrophobic amino acid may abrogate binding of Doc1804 to CohScaC.

Figure 3.11 Binding affinity of group 6 Docs to CohScaC determined by ITC.



Representative binding isotherms are displayed in *A*, CohScaC/Doc903, CohScaC/Doc1369 and CohScaC/Doc1965. The upper part of each panel shows the raw heats of binding, whereas the lower parts comprise the integrated heats after correction for heat of dilution. The curve represents the best fit to a single-site binding model. The corresponding thermodynamic parameters are shown in Table 3.6. *B.* Alignment of tested Group 6 Docs with a version of Doc3 in which the C-terminal half, was switched with the N-terminal half (Doc3_180), resulting in the N-terminal interface (blue residues) of Doc3 being perfectly aligned with the C-terminal interface (green residues) of the Group 6 members and vice-versa, supporting the theory that they will bind the Coh in opposite orientation. Residues involved in Ca^{2+}-binding are pointed out by the blue triangles at the top. All residues involved in the Doc3 interaction with CohScaC are highlighted according to the colour code displayed at the bottom. Conservation of key residues for Coh recognition is highlighted with black boxes.

3.4. Conclusions

In nature, Coh-Doc interactions are essential for cellulosome assembly by providing a molecular base for the integration of microbial enzymes onto a primary scaffoldin. Enzyme-containing Docs present a dual binding mode resulting from the presence of two identical Coh-binding faces. Data presented here reports a notable exception to this general rule by analyzing the incorporation of cellulosomal enzymes into *R. flavefaciens* cellulosomes through adaptor scaffoldins such as ScaC. Previously, *R. flavefaciens* Group 3 and Group 6 Docs were shown to specifically recognize the single Coh of scaffolding ScaC. Here, the structure of a Group 3 Doc, Doc3, in complex with CohScaC, revealed the presence of a single Coh-binding interface that involves both Doc helices. These observations contrast with the dual binding mode mechanism previously identified in Docs used by the majority of cellulosome producing bacteria, such as *C. thermocellum*, *A. cellulolyticus* and *C. cellulolyticum*, to recruit cellulosomal enzymes into primary scaffoldins. Lack of internal symmetry in group 3 and 6 *R. flavefaciens* Docs generated an unconventional single protein-binding interface that specifically interacts with the Coh of ScaC adaptor scaffoldin. Notably, group 3 and 6 Docs were found to be predominantly appended to hemicellulases, suggesting that *R. flavefaciens* has evolved an original mechanism to recruit this subset of enzymes that are critical to plant cell wall degradation to the cellulosome. Thus, instead of binding a significant array of hemicellulases directly to primary scaffoldins, enzymes affixed with group 3 and 6 Docs are mobilized to the highly intricate bacterial nanomachines produced by *R. flavefaciens* to degrade recalcitrant polysaccharides *via* an adaptor scaffoldin. This observation suggests that hemicellulases may either act freely during carbohydrate hydrolysis, in the absence of ScaC, or be recruited to cellulosomes once ScaC adaptor scaffoldin is expressed. This hypothesis is currently under investigation and may indicate that *R. flavefaciens* has developed highly elaborate mechanisms to fine tune plant cell wall degradation.

Chapter 4

Ruminococcus flavefaciens Coh-Doc complexes involving group 1 dockerins

Assembly of *Ruminococcus flavefaciens* cellulosome revealed by structures of two cohesin-dockerin complexes

Pedro Bule[a], Victor D. Alves[a], Vered Israeli-Ruimy[b], Ana Luísa Carvalho[c], Luís M.A. Ferreira[a], Steven P. Smith[d], Harry J. Gilbert[e], Shabir Najmudin[a], Edward A. Bayer[b] and Carlos M.G.A. Fontes[a,1]

[a] *CIISA – Faculdade de Medicina Veterinária, ULisboa, Pólo Universitário do Alto da Ajuda, Avenida da Universidade Técnica, 1300-477 Lisboa, Portugal.* [b] *Department of Biomolecular Sciences, The Weizmann Institute of Science, Rehovot 76100 Israel.* [c] *UCIBIO-REQUIMTE, Departamento de Química, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, 2829-516 Caparica, Portugal.* [d] *Department of Biomedical and Molecular Sciences, Queen's University, Kingston, ON K7L 3N6, Canada.* [e] *Institute for Cell and Molecular Biosciences, Newcastle University, The Medical School, Newcastle upon Tyne NE2 4HH, United Kingdom.* [1]***Corresponding author***

Adapted from *Scientific Reports* 2017 Apr 7;7(1):759. (Bule *et al.*, 2017)

Abstract

Cellulosomes are sophisticated multi-enzymatic nanomachines produced by anaerobes to effectively deconstruct plant structural carbohydrates. Cellulosome assembly involves the binding of enzyme-borne dockerins (Doc) to repeated cohesin (Coh) modules located in a non-catalytic scaffoldin. Docs appended to cellulosomal enzymes generally present two similar Coh-binding interfaces supporting a dual-binding mode, which may confer increased positional adjustment of the different complex components. *Ruminococcus flavefaciens'* cellulosome is assembled from a repertoire of 223 Doc-containing proteins classified into 6 groups. Recent studies revealed that Docs of groups 3 and 6 are recruited to the cellulosome *via* a single-binding

mode mechanism with an adaptor scaffoldin. To investigate the extent to which the single-binding mode contributes to the assembly of *R. flavefaciens* cellulosome, the structures of two group 1 Docs bound to Cohs of primary (ScaA) and adaptor (ScaB) scaffoldins were solved. The data revealed that group 1 Docs display a conserved mechanism of Coh recognition involving a single-binding mode. Therefore, in contrast to all cellulosomes described to date, the assembly of *R. flavefaciens* cellulosome involves single but not dual-binding mode Docs. Thus, this work reveals a novel mechanism of cellulosome assembly and challenges the ubiquitous implication of the dual-binding mode in the acquisition of cellulosome flexibility.

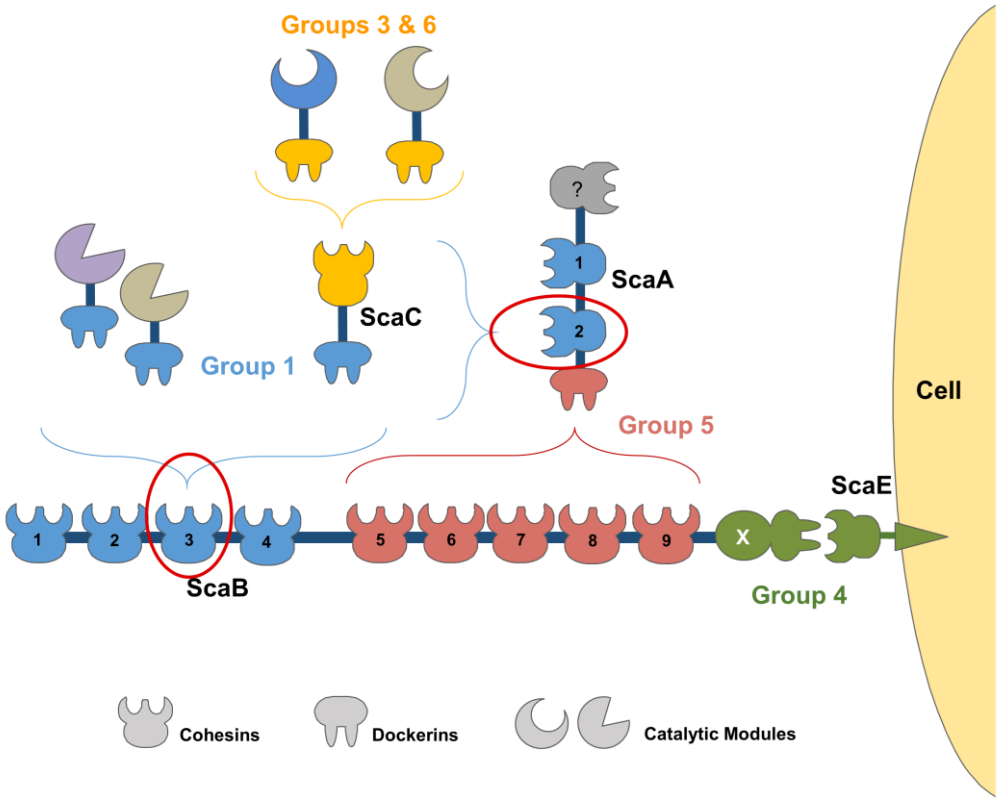
4.1. Introduction

The cellulosome is one of the most intricate nanomachines Nature has evolved. Cellulosomes combine an extensive repertoire of enzymes, including glycoside hydrolases, pectate lyases and carbohydrate esterases, into a large multi-enzyme complex (molecular mass >3 MDa) that efficiently deconstructs especially recalcitrant plant structural carbohydrates, such as cellulose and hemicellulose. Highly ordered protein-protein interactions are critical to a large array of cellular and biological processes. Thus, cellulosome assembly results from the binding of enzyme borne dockerin modules (Docs) to cohesin modules (Cohs) located in macro-molecular scaffolds (scaffoldins). Integration of enzymes into the cellulosome is believed to enhance the synergistic interactions between enzymes with complementary activities while promoting enzyme stability (Bayer *et al.*, 2004; Fontes & Gilbert, 2010). This process is critical to the cycling of carbon between microbes, herbivores and plants. In addition, cellulases and hemicellulases are now used in several biotechnology-based industries, such as the bio-conversion of plant biomass into renewable fuels and the development of specific molecules with biomedical applications (Bayer *et al.*, 2007, 1994; Demain *et al.*, 2005).

The rumen, which essentially constitutes a large fermentation chamber in the gastrointestinal tract of ruminant mammals, is a highly competitive ecological niche colonized by symbiotic microbes that have specialized in the hydrolysis of recalcitrant carbohydrates. *Ruminococcus flavefaciens*, a Gram-positive anaerobic bacterium of the Firmicutes phylum, is one of the major cellulolytic ruminal bacteria and the only species in this microbial ecosystem that has been shown to possess a definitive cellulosome (Ding *et al.*, 2001). Intriguingly, the rumen houses numerous subspecies of this bacterium, each with a similar set of scaffoldins but with its own spectrum of dockerin-bearing proteins (enzymes) and cellulosome architecture (Dassa *et al.*, 2014; Jindou *et al.*, 2008). The genome sequence of *R. flavefaciens* strain FD-1 revealed the presence of 223 dockerin-containing proteins (154 of which were identified as carbohydrate-

active enzymes) (Dassa *et al.*, 2014), indicating that this bacterial nanomachine is the most complex cellulosome described to date (Berg Miller *et al.*, 2009) (Figure 4.1). *R. flavefaciens* Docs have been organized into six groups based on primary structure homology (Rincon *et al.*, 2010). This classification was recently found to be functionally relevant (Israeli-Ruimy *et al.*, 2017), with the binding of group 1 Docs to the Cohs of scaffoldins ScaA and ScaB providing the major mechanism for cellulosome assembly in *R. flavefaciens*. The 96 group 1 Docs have been classified in four subgroups (a to d) although the functional significance of this subdivision remains unclear. The cellulosome is tethered to the surface of *R. flavefaciens* through the binding of the group 4 Doc of ScaB to the Coh of the cell surface protein ScaE. A variety of other proteins were found to contain Docs that specifically interact with cell surface Cohs rather than to the cellulosomal Cohs. These Docs were classified into group 4 and group 2. Finally, hemicellulases containing group 3 or 6 Docs bind to the adaptor scaffoldin ScaC, whose group 1 Doc locks onto the Cohs in ScaA or ScaB Cohs 1-4 (Bule *et al.*, 2016; Rincon *et al.*, 2004). The ScaA Doc is the only member of group 5 and binds exclusively to ScaB Cohs 5-9. Figure 4.1 provides an overview of the organization of *R. flavefaciens* cellulosome.

Figure 4.1 Group-specific interactions that contribute to the major cellulosome assembly in *R. flavefaciens* strain FD-1.



The scheme is color-coded to highlight the four subgroups of cohesin-dockerin specificities: Dockerins and cognate cohesin counterparts of the different groups are marked in blue (Group 1 dockerins), yellow (Groups 3 and 6), green (Groups 2 and 4) and red (Group 5), respectively. Group 2 dockerins are

truncated derivatives of group 4 and are not represented in the figure for simplification. The red ovals mark the complexes of the Group 1 interactions, whose structures are reported here.

In all clostridial cellulosomal systems described to date, such as *Clostridium thermocellum* (Carvalho *et al.*, 2003, 2007), *C. cellulolyticum* (Pinheiro *et al.*, 2008) and *Acetivibrio cellulolyticus* (Cameron, Najmudin, *et al.*, 2015), Docs interact with their cognate Cohs through a dual-binding mode. Thus, these Docs possess the ability to bind the cognate Coh in two different orientations, by rotating ~180º with respect to its protein ligand, resulting in two different Coh-Doc conformations. The dual-binding mode results from the characteristic internal symmetry of the Doc sequence and is believed to confer additional flexibility to the macromolecular organization of cellulosomes. Recent structure/function studies, unexpectedly, showed that groups 3 and 6 *R. flavefaciens* Docs display a single-binding mode for their target Cohs. Intriguingly, the sequence of group 1 Docs, do not seem to possess the internal symmetry required to support the dual-binding mode. This suggests that group 1 Docs may bind to their target Cohs through a single-binding mode. To test this hypothesis, we determined the X-ray crystal structure of two *R. flavefaciens* group 1 Docs, Doc1a and Doc1b, in complex with a ScaB (CohScaB3) and a ScaA Coh, respectively. These structures together with comprehensive biochemical analyses suggest that integration of a large repertoire of enzymes into the *R. flavefaciens* cellulosome operates through a single-binding mode.

4.2. Experimental procedures

4.2.1. Gene synthesis and DNA cloning

Dockerins are inherently unstable when produced in *Escherichia coli*. To promote dockerin stability, *R. flavefaciens* FD-1 group 1 dockerins from protein WP_009986495 (residues 577-649) and protein WP_009982745 (residues 783-862), termed Doc1a and Doc1b, were co-expressed *in vivo* with ScaB cohesin 3 (CohScaB3) and ScaA cohesin (CohScaA), respectively. The immediate binding of the expressed dockerins to the expressed cohesins confers the necessary dockerin stabilization. The genes encoding the proteins were designed considering the optimization of codon usage to maximize expression in *E. coli*, synthesized *in vitro* (NZYTech Ltd, Lisbon, Portugal) and cloned into pET28a (Merck Millipore, Germany) under the control of separate T7 promoters. The dockerin-encoding genes were positioned at the 5' end and the cohesin-encoding genes at the 3' end of the artificial DNA. A T7 terminator sequence (to terminate transcription of the dockerin gene) and a T7 promoter sequence (to control transcription of the cohesin gene) were incorporated between the sequences of the two genes. This construct contained *Nhe*I and *Nco*I recognition sites at the 5' end and *Xho*I and *Sal*I at the 3' end specifically tailored to allow subcloning into pET-28a (Merck Millipore, Germany), such that the sequence encoding a six-residue His tag could be introduced either at

the N-terminus of the dockerin (through digestion with *Nhe*I and *Sal*I, incorporating the additional sequence MGSSHHHHHHSSGLVPRGSHMAS at the N-terminus of the polypeptide) or at the C-terminus of the cohesin (by cutting with *Nco*I and *Xho*I, which incorporates the additional sequence LEHHHHHH at the C-terminus of the polypeptide). Thus, as a result of this strategy, two pET28a plasmid derivatives were produced for each Coh-Doc pair: one leading to the expression of dockerin with an engineered hexa-histidine tag and a second derivative where the engineered tag is attached to the cohesin. The plasmids were used to express *Rf*CohScaA-Doc1a and *Rf*CohScaB3-Doc1b protein complexes in *E. coli*. Recombinant Doc1a, Doc1b, CohScaA and CohScaB3 primary sequences are presented in Table 4.1.

Table 4.1 Recombinant protein sequences of *Rf*Doc1a, *Rf*Doc1b, *Rf*CohScaA, RfCohScaB3 and mutant variants of these proteins produced for the interaction studies.

Dockerin	Protein Sequence
Doc1a	EAVQKFPGDANCDGIVDISDAVLIMQTMANPSKYQMTDKGRINADVTGNSDGVTVLDAQFIQSYCLGLVELPPVEYVNVTKQPVEPA
Doc1a I39A	EAVQKFPGDANCDGIVD**A**SDAVLIMQTMANPSKYQMTDKGRINADVTGNSDGVTVLDAQFIQSYCLGLVELPPVEYVNVTKQPVEPA
Doc1a S40A	EAVQKFPGDANCDGIVDI**A**DAVLIMQTMANPSKYQMTDKGRINADVTGNSDGVTVLDAQFIQSYCLGLVELPPVEYVNVTKQPVEPA
Doc1a V43A	EAVQKFPGDANCDGIVDISDA**A**LIMQTMANPSKYQMTDKGRINADVTGNSDGVTVLDAQFIQSYCLGLVELPPVEYVNVTKQPVEPA
Doc1a Q47A	EAVQKFPGDANCDGIVDISDAVLIM**A**TMANPSKYQMTDKGRINADVTGNSDGVTVLDAQFIQSYCLGLVELPPVEYVNVTKQPVEPA
Doc1a K54A	EAVQKFPGDANCDGIVDISDAVLIMQTMANPS**A**YQMTDKGRINADVTGNSDGVTVLDAQFIQSYCLGLVELPPVEYVNVTKQPVEPA
Doc1a Q83A	EAVQKFPGDANCDGIVDISDAVLIMQTMANPSKYQMTDKGRINADVTGNSDGVTVLDAQFI**A**SYCLGLVELPPVEYVNVTKQPVEPA
Doc1a L87A	EAVQKFPGDANCDGIVDISDAVLIMQTMANPSKYQMTDKGRINADVTGNSDGVTVLDAQFIQSYC**A**GLVELPPVEYVNVTKQPVEPA
Doc1a I39A + V43A	EAVQKFPGDANCDGIVD**A**SDA**A**LIMQTMANPSKYQMTDKGRINADVTGNSDGVTVLDAQFIQSYCLGLVELPPVEYVNVTKQPVEPA
Doc1a V43A + Q47A	EAVQKFPGDANCDGIVDISDA**A**LIM**A**TMANPSKYQMTDKGRINADVTGNSDGVTVLDAQFIQSYCLGLVELPPVEYVNVTKQPVEPA
Doc1b	NVTLWGDANCDGIVDISDAVIIMQSLSNPSKFDRNGNDEHHITAQGELNGDVNENGNGITNADALAIQKYLLNLIGNLPE

Cohesin	Protein Sequence
CohScaB3 WT	MPVANADVVFDFQNYTAKAGDEVTVDVLVDSKNK PISAMDVKFKVDSPLTIEEIDKESLAFNTTVMTNMAILGANFKSLDDKGEPLVPKDGAAVFTLYVNVPANTPDGTYYVGFNGKNEVHKSNDGS QFTVASKNGAITVGTPNEEG
CohScaB3 A38Q	...PIS**Q**MDVKFKVDSPLTIEEIDKESLAFNTTVMTNMAILGANFKSLDDKGEPLVPKDGAAVFTLYVNVPANTPDGTYYVGFNGKNEVHKSNDGS...
CohScaB3 N68A	...PISAMDVKFKVDSPLTIEEIDKESLAFNTTVMT**A**MAILGANFKSLDDKGEPLVPKDGAAVFTLYVNVPANTPDGTYYVGFNGKNEVHKSNDGS...
CohScaB3 N75A	...PISAMDVKFKVDSPLTIEEIDKESLAFNTTVMTNMAILGA**A**FKSLDDKGEPLVPKDGAAVFTLYVNVPANTPDGTYYVGFNGKNEVHKSNDGS...
CohScaB3 K77A	...PISAMDVKFKVDSPLTIEEIDKESLAFNTTVMTNMAILGANF**A**SLDDKGEPLVPKDGAAVFTLYVNVPANTPDGTYYVGFNGKNEVHKSNDGS...
CohScaB3 L79A	...PISAMDVKFKVDSPLTIEEIDKESLAFNTTVMTNMAILGANFKS**A**DDKGEPLVPKDGAAVFTLYVNVPANTPDGTYYVGFNGKNEVHKSNDGS...
CohScaB3 E84A	...PISAMDVKFKVDSPLTIEEIDKESLAFNTTVMTNMAILGANFKSLDDKG**A**PLVPKDGAAVFTLYVNVPANTPDGTYYVGFNGKNEVHKSNDGS...
CohScaB3 H121A	...PISAMDVKFKVDSPLTIEEIDKESLAFNTTVMTNMAILGANFKSLDDKGEPLVPKDGAAVFTLYVNVPANTPDGTYYVGFNGKNEV**A**KSNDGS...
CohScaB3 N124A	...PISAMDVKFKVDSPLTIEEIDKESLAFNTTVMTNMAILGANFKSLDDKGEPLVPKDGAAVFTLYVNVPANTPDGTYYVGFNGKNEVHKS**A**DGS...
CohScaB3 N75A + E84A	...PISAMDVKFKVDSPLTIEEIDKESLAFNTTVMTNMAILGA**A**FKSLDDKG**A**PLVPKDGAAVFTLYVNVPANTPDGTYYVGFNGKNEVHKSNDGS...
CohScaB3 N75A + H121A	...PISAMDVKFKVDSPLTIEEIDKESLAFNTTVMTNMAILGA**A**FKSLDDKGEPLVPKDGAAVFTLYVNVPANTPDGTYYVGFNGKNEV**A**KSNDGS...
CohScaB3 N75A + N124A	...PISAMDVKFKVDSPLTIEEIDKESLAFNTTVMTNMAILGA**A**FKSLDDKGEPLVPKDGAAVFTLYVNVPANTPDGTYYVGFNGKNEVHKS**A**DGS...
CohScaB3 E84A + H121A	...PISAMDVKFKVDSPLTIEEIDKESLAFNTTVMTNMAILGANFKSLDDKG**A**PLVPKDGAAVFTLYVNVPANTPDGTYYVGFNGKNEV**A**KSNDGS...
CohScaB3 E84A + N124A	...PISAMDVKFKVDSPLTIEEIDKESLAFNTTVMTNMAILGANFKSLDDKG**A**PLVPKDGAAVFTLYVNVPANTPDGTYYVGFNGKNEVHKS**A**DGS...
CohScaB3 H121A + N124A	...PISAMDVKFKVDSPLTIEEIDKESLAFNTTVMTNMAILGANFKSLDDKGEPLVPKDGAAVFTLYVNVPANTPDGTYYVGFNGKNEV**A**KS**A**DGS...
CohScaA	MQPVANADVIFDFGNYEAKAGEEVQVDVTVDSKNKAISAMDVVFAIDSPLTIDEIDKESLAFKTTAMTNIAILGANFKSLDDKGEPLVPTKDPVFTLYV TVPATTPDGVYNVGFGNKCEVHKSNDGSKYSSTAINGKIKVGNPVDDP

The mutated residues are highlighted in black.

To produce the recombinant cohesins and dockerins individually, two distinct cloning methods were used. Digesting the previously described cohesin-tagged version of the pET28 derivatives with BglII allowed removal of the dockerin sequence. Plasmid integrity was reconstituted by re-ligating. This strategy allowed producing two novel pET28a derivatives encoding recombinant cohesins CohScaA and CohScaB3 containing C-terminal hexa-histidine tags. Dockerin-encoding genes were cloned into the pHTP2 vector (NZYtech, Lisbon, Portugal) using NZYEasy Cloning & Expression System (NZYtech, Lisbon, Portugal), following the manufacturer's protocol. Dockerin genes were isolated by PCR using *R. flavefaciens* FD-1 genomic DNA as a template and the primers shown in Table S4.1 (Annexes). Recombinant dockerins encoded by the pHTP2 derivatives contained an N-terminal thioredoxin A and an internal hexa-histidine tag for increased protein stability and solubility. Sequences of all plasmids produced were confirmed by Sanger sequencing.

To identify the residues that modulate Coh-Doc specificity, several Doc1a and CohScaB3 protein derivatives were produced by site-directed mutagenesis of the pHTP2 and pET28a derivatives encoding the two genes. Site-directed mutagenesis was performed by PCR amplification using the primers presented in Table S4.1 (Annexes), which allowed the production of nine Doc1a protein derivatives (I39A, S40A, V43A, Q47A, K54A, Q83A, L87A, I39A + V43A, V43A + Q47A) and fourteen CohScaB3 protein derivatives (A38Q, N68A, N75A, K77A, L79A, E84A, H121A, N124A, E84A + H121A, N75A + H121A, N75A + N124A, N75A + E84A, E84A + N124A, H121A + N124A). Each of the newly generated gene sequences was fully sequenced to confirm that only the desired mutation accumulated in the nucleic acid.

For the cellulose microarray experiments, a system designed to fuse the Docs with a xylanase and the Cohs to a carbohydrate-binding module was selected. This allows production of highly stable and functional Cohs that can be immobilized in a cellulose-coated glass slide and Docs that can be recognized by an α-xylanase antibody (Haimovitz *et al.*, 2008). Thus, sequences encoding the various cohesins and selected group 1 Docs were amplified from *R. flavefaciens* FD-1 genomic DNA by PCR, using NZYProof polymerase (NZYTech Ltd., Portugal) and the primers shown in supplemental Table S4.2 (Annexes). After gel purification, the Doc-encoding amplicons were inserted into a xylanase-Doc cassette in the pET9d plasmid after digestion with *Kpn*I and *Bam*HI and ligation with T4 ligase. The resulting expressed products consist of His-tagged Docs fused to xylanase T-6 from *Geobacillus stearothermophilus* at the N terminus of the polyhistidine tag (XynDoc). The Coh-encoding genes were cloned into a CBM-Coh cassette in pET28a after digestion with *BamH*I and *Xho*I restriction enzymes. This resulted in His-

tagged Coh recombinant derivatives fused to a CBM3a from the *C. thermocellum* scaffoldin ScaA (CBMCoh) (Barak *et al.*, 2005; Handelsman *et al.*, 2004).

4.2.2. Expression and Purification of Recombinant proteins

Preliminary expression screens revealed that when the hexa-histidine tag was located at the dockerin N-terminal end of both *Rf*CohScaB3-Doc1a and *Rf*CohScaA-Doc1b complexes, the expression levels of both cohesin and dockerin were higher. Tagging the cohesin resulted in the accumulation of large levels of unbound cohesin in the purification product suggesting that cohesin was expressed at higher levels than dockerins or that untagged dockerin was less stable. Thus, pET28a derivatives encoding the protein complexes formed using the tagged dockerin were subsequently used to transform *E. coli* BL21 (DE3) cells in order to produce *Rf*CohScaB3-Doc1a and *Rf*CohScaA-Doc1b protein complexes in large quantities. Recombinant *E. coli* were grown at 37ºC to an OD_{600} of 0.5. Recombinant protein expression was induced by the addition of 1 mM isopropyl *β*-D-1-thiogalactopyranoside followed by incubation at 19ºC for 16 hours. Cells were harvested by 15 min centrifugation at 5000 x *g* and resuspendend in 20 mL of immobilized-metal affinity chromatography (IMAC) binding buffer (50 mM HEPES, pH 7.5, 10 mM imidazole, 1 M NaCl, 5 mM $CaCl_2$). Cells were then disrupted by sonication and the cell-free supernatant recovered by 30 min centrifugation at 15,000 x *g*. After loading the soluble fraction into a HisTrap^{TM} nickel-charged Sepharose column (GE Healthcare, UK), initial purification was carried out by IMAC in a FPLC system (GE Healthcare, UK) using conventional protocols with a 35 mM imidazole wash and a 35-300 mM imidazole elution gradient. Fractions containing the purified cohesin–dockerin complexes were buffer exchanged into 50 mM HEPES, pH 7.5, containing 200 mM NaCl, 5 mM $CaCl_2$ using a PD-10 Sephadex G-25M gel-filtration column (Amersham Pharmacia Biosciences, UK). A further purification step by gel-filtration chromatography was performed by loading the Coh-Doc complexes onto a HiLoad 16/60 Superdex 75 (GE Healthcare, UK) at a flow rate of 1 ml min^{-1}. Fractions containing the purified complexes were then concentrated with Amicon Ultra-15 centrifugal devices with a 10-kDa cutoff membrane (Millipore, USA) and washed three times with molecular biology grade water (Sigma) containing 0.5 mM $CaCl_2$. The protein concentration was estimated in a NanoDrop 2000c spectrophotometer (Thermo Scientific, USA) using a molar extinction coefficient (ε) of 9 075 M^{-1} cm^{-1} for *Rf*CohScaB3-Doc1a and 13 075 M^{-1} cm^{-1} for *Rf*CohScaA-Doc1b. The final protein concentrations were adjusted to 40 $mg.mL^{-1}$ for the *Rf*CohScaB3-Doc1a complex and 27 $mg.mL^{-1}$ for *Rf*CohScaA-Doc1b, and stored in molecular biology grade water containing 0.5 mM $CaCl_2$. The purity and molecular mass of the recombinant complexes were confirmed by 14 % (w/v) SDS–PAGE. A similar protocol was

used to produce *Rf*CohScaB3 used in the crystallization trials and its seleno-methionine derivative, except that in the latter the protein was expressed in the methionine auxotroph B834 strain of *E. coli*, using the growth conditions described by Ramakrishnan *et al.* 1993 (Ramakrishnan, Finch, Graziano, Lee, & Sweet, 1993), and a reducing agent was added to all the buffers: 5 mM of 2-mercaptoethanol in affinity-chromatography buffers, 5 mM DTT in size-exclusion chromatography buffer and 1 mM TCEP in storage buffer. The final protein concentrations were adjusted to 47 $mg.mL^{-1}$.

Group 1 dockerins and *R. flavefaciens* cohesins and their respective mutant derivatives used in native PAGE and ITC experiments were expressed as described before and purified with IMAC using nickel-charged Sepharose His GraviTrap gravity-flow columns (GE Healthcare, UK). After IMAC, the recombinant cohesin and dockerins were buffer exchanged to 50 mM HEPES pH 7.5, 0.5 mM $CaCl_2$ and 0.5 mM TCEP using PD-10 Sephadex G-25M gel filtration columns (GE Healthcare, UK).

4.2.3. Nondenaturing gel electrophoresis (NGE)

For the NGE experiments, each Doc variant (30 μM) was incubated in the presence and absence of 30 μM Coh for 30 min at room temperature and separated on a 10 % native (lacking SDS) polyacrylamide gel. Electrophoresis was carried out at room temperature. The gels were stained with Coomassie Blue. Complex formation was detected by the presence of an additional band, usually displaying a lower electrophoretic mobility than that of the individual modules.

4.2.4. Isothermal Titration Calorimetry

All ITC experiments were carried out at 308 K. The purified Doc and Coh variants were diluted to the required concentrations and filtered using a 0.45 μm syringe filter (PALL). During titrations the dockerin constructs were stirred at 307 revolutions/min in the reaction cell and titrated with 28 successive 10 μL injections of cohesin at 220 s intervals. Integrated heat effects, after correction for heats of dilution, were analyzed by nonlinear regression using a single-site model (Microcal ORIGIN version 7.0, Microcal Software, USA). The fitted data yielded the association constant (K_A) and the enthalpy of binding (ΔH). Other thermodynamic parameters were calculated using the standard thermodynamic equation: $\Delta RTlnK_A=\Delta G=\Delta H-T\Delta S$.

4.2.5. Cellulose microarray

The cellulose microarray approach was conducted using the XynDoc/CBM-Coh fusion protein pairs, in order to evaluate cohesin-dockerin interactions by refining the method described in

Barak *et al.*(Haimovitz *et al.*, 2008) DNA isolation and cloning were performed as described above. The strong selective binding of the CBM to the cellulose-coated slides was used as an intrinsic purification step so that cohesins were thus applied to the glass slides as crude extracts. The dockerins were purified as described above.

Rabbit anti-XynT6 primary antibody was conjugated with fluorescent Cy3 dye and rabbit anti-CBM primary antibody with fluorescent Cy5 dye, in order to assess signal intensity and normalize with the amount of protein, respectively. Xyn-CBM fusion protein was designed, cloned and expressed in the form of crude extract, as a positive control for the Cy3- and Cy5-conjugated antibodies. For biological positive controls, pre-established interactions were included in the setup. To eliminate the possibility of any of *E. coli*'s background components generating a false signal, BL-21 were transformed with an empty pET28a vector, which lacks a CBM or a cohesin module. The cellulose-coated glass slides were printed with crude extracts of this negative control that were subjected to the same treatment and storage conditions.

Although protein amounts were validated on SDS-PAGE gels prior to screening, there was still printing variation resulting from the use of a hand arrayer. It was therefore necessary to estimate the ratio between the Cy3 signal intensity, which indicates the presence of XynDoc, and the Cy5 signal intensity, which stands for the amount of CBM-Coh that is present in the area of a specific spot. This was done with 'Array Vision Evaluation 8.0' software. Raw data were further processed in Excel to generate bar graphs.

4.2.6. X-ray crystallography, Structural Determination and Refinement

Crystallization conditions were set up using the sitting-drop vapor diffusion method with a robotic nanodrop dispensing system Oryx8 (Douglas Instruments, UK). Commercial kits Crystal Screen, Crystal Screen 2, PEG Ion Screen I and II from Hampton Research (California, USA), JCSG+ HT96 (Molecular Dimensions, UK) and an in-house screen (80 factorial) were used for the screening. For *Rf*CohScaB3, 1.0 μL drops of 22 and 47 $mg \cdot mL^{-1}$ of protein were mixed with 1.0 μL of reservoir solution at room temperature per well containing 50 μL of the crystallization solution. The same procedure was used for *Rf*CohScaA-Doc1b and *Rf*CohScaB3-Doc1a with protein drops at concentrations of 40 and 20 $mg \cdot mL^{-1}$ and 27 and 13.5 $mg \cdot mL^{-1}$, respectively. The resulting plates were then stored at 292 K.

Crystal formation from the initial screens was observed in the following 2 different conditions with the C-terminal tagged native *Rf*CohScaB3: 0.2 M lithium sulfate, 0.1 M sodium acetate pH 4.5, 30% w/v PEG 8000; and 0.17 M ammonium sulfate, 25.5% w/v PEG 4000, 15% v/v glycerol. SeMet-derivative plates were immediately set up for structure determination, should

molecular replacement methods fail. For the SeMet-*Rf*CohScaB3 an optimization screen was set up around the range 0.1-0.5 M lithium sulfate, 0.1 M sodium acetate pH 4.5, 10-32% w/v PEG 8000 for the first condition; and 0.1-0.5 M ammonium sulfate, 10-32% w/v PEG 4000, 15% v/v glycerol for the second. The glycerol concentration was maintained at 15%, which acted as a cryoprotectant. Diffracting crystals were obtained in 12 of the 96 wells of the second optimization screen. These crystals grew to a maximum dimension of ~500 x 80 x 80^3 μm, within two weeks. In addition, diffracting N-tagged *Rf*CohScaB3-Doc1a crystals were obtained in a 0.2 M ammonium nitrate and 20% w/v PEG 3550 solution while *Rf*ScaSCoh-Doc1b crystalized in a 0.2 M calcium acetate, 0.1 M sodium cacodylate trihydrate pH 6.5 and 18% PEG 8000 solution. All crystals were cryoprotected with mother solution containing 15–30 % glycerol or with 100 % Paratone-N (Hampton Research, USA) and flash-cooled in liquid nitrogen.

4.2.7. Data collection, processing, structure determination and refinement

Data for the SeMet *Rf*CohScaB3 derivatives were collected on beamline ID23-2 at the European Synchrotron Radiation Facility (ESRF), Grenoble, France. 360° of data were collected with a $\Delta\varphi$ of 0.1° and an exposure of 0.04 sec. The data were collected at the wavelength of 0.8726 Å for a single-wavelength anomalous diffraction experiment. The crystal was cooled to 100 K using a gaseous nitrogen cryostream (Oxford Cryosystems) and data collected using the CCD MARMOSAIC 225 detector. The data sets were processed using iMOSFLM (Battye *et al.*, 2011) or XDS (Kabsch, 2010) and AIMLESS (Evans, 2006) from the CCP4 suite (Collaborative Computational Project, Number 4, 1994 (Winn *et al.*, 2011)). Data collection statistics are given in Table 4.2. The crystals belong to the tetragonal space group ($P4_12_12$), with a single molecule in the asymmetric unit, a solvent content of ~51 % and a Matthews coefficient of ~2.49 $Å^3$ Da^{-1} (Matthews, 1968). The SeMet-*Rf*CohScaB3 structure was determined by single wavelength anomalous dispersion experiment with AUTOSOL (Terwilliger *et al.*, 2009) from the PHENIX suite (Adams, Afonine, *et al.*, 2010)). AUTOBUILD was used for building the initial structure (Terwilliger *et al.*, 2008). Refmac5 (Murshudov *et al.*, 2011) interspersed with model adjustment in COOT (Emsley & Cowtan, 2004) were used for structure refinement and rebuilding. PDB_REDO was used in the penultimate round of refinement for validation purposes (Joosten *et al.*, 2014). The root mean square deviation of bond lengths, bond angles, torsion angles and other indicators were continuously monitored using validation tools in COOT and MOLPROBITY. Final coordinates and structure factors were deposited in PDB under accession codes 5AOZ and R5AOZSF, respectively.

Data for the Coh-Doc complexes were collected on beamline I04-1 at the Diamond Light Source, Harwell, England (*Rf*CohScaB3-Doc1a) and at the ESRF beamline ID-23, Grenoble, France (*Rf*CohScaA-Doc1b) using a PILATUS 6M detector (Dectris Ltd). Data collection and processing was done as described above. Data collection statistics are given in Table 4.2. The best diffracting *Rf*CohScaB3-Doc1a crystals diffracted to a resolution of 1.26 Å and belonged to the orthorhombic space group $P2_12_12_1$ with a single cohesin-dockerin complex in the asymmetric unit, a solvent content of ~43 % and a Matthews coefficient of ~2.15 $Å^3$ Da^{-1}. PHASER (McCoy *et al.*, 2007) was used to carry out molecular replacement using *Rf*CohScaB3 (5AOZ) and BUCCANEER (Cowtan, 2006) helped building the initial dockerin model. Refinement and model rebuilding were carried out as described for *Rf*CohScaB3. The final round of refinement was performed using the TLS/restrained refinement procedure using each module as a single group. The best diffracting *Rf*CohScaA-Doc1b crystals diffracted to 1.70 Å and belonged to the orthorhombic spacegroup $P2_12_12_1$ with a single cohesin-dockerin complex in the asymmetric unit, a solvent content of ~47 % and a Matthews coefficient of ~2.33 $Å^3$ Da^{-1}. PHASER was used to carry out molecular replacement using the *Rf*CohScaB3-Doc1a model. Refinement occurred has described for *Rf*CohScaB3-Doc1a. A summary of the refinement statistics is shown in Table 4.2. Molecular representation figures were prepared with UCSF Chimera (Pettersen *et al.*, 2004). Final coordinates and structure factors were deposited in PDB under accession codes 5M2O and SF5M2O for *Rf*CohScaB3-Doc1a, and 5M2S and SF5M2S for *Rf*CohScaA-Doc1b, respectively.

Table 4.2 X-ray crystallography data collection and refinement statistics for *Rf*CohScaB3, *Rf*CohScaB3-Doc1a and *Rf*CohScaA-Doc1b.

Dataset	*Rf*CohScaB3	*Rf*CohScaB3-Doc1a	*Rf*CohScaA-Doc1b
Data Colection			
Beamline	*ESRF ID23-2*	*Diamond I04-1*	*ESRF-ID23*
Space Group	$P4_12_12$	$P2_12_12_1$	$P12_11$
Wavelength (Å)	*0.8726*	*0.920*	*0.873*
Unit-cell parameters			
a, b c (Å)	*60.427, 60.427, 86.509*	*42.77, 63.51, 84.48*	*45.61, 64.49, 47.67*
α, β , γ (º)	*90, 90, 90*	*90, 90, 90*	*90, 116.72, 90*
V_m# (Å3 Da-1)	*2.36*	*2.15*	*2.33*
Solvent Content (%)	*48.01*	*42.94*	*47.27*
Resolution limits (Å)	*49.54 – 1.14 (1.18 – 1.14)*	*20.27 – 1.26 (1.305– 1.26)*	*42.58 – 1.7 (1.761 – 1.7)*
No. of observations	*606740 (55700)*	*460418 (38386)*	*112328 (7417)*
No. of unique observations	*58923 (5791)*	*62519 (6069)*	*26481 (2322)*
Multiplicity	*10.3 (9.6)*	*7.4 (6.3)*	*4.2 (3.2)*
Completeness (%)	*99.91 (99.27)*	*99.6 (98.09)*	*97.38 (86.13)*
<I/σ(I)>	*18.21 (1.73)*	*5.74 (2.56)*	*9.33 (4.34)*
CC1/2†	*0.999 (0.582)*	*0.976 (0.783)*	*0.991 (0.845)*
Wilson B-factor	*11.66*	*6.76*	*8.48*
Rmerge ‡	*0.073 (1.327)*	*0.2322 (0.5811)*	*0.1134 (0.2575)*
Strtucture Refinement			
R-work § , R-free ¥	*0.1184, 0.1424*	*0.1318, 0.1535*	*0.1313, 0.1592*
No. of Non-H atoms	*1331*	*1947*	*2041*
Macromolecules	*1100*	*1622*	*1695*
Ligands	*6*	*2*	*21*
Water	*225*	*323*	*325*
Protein residues	*141*	*211*	*220*
RMS(bonds)	*0.016*	*0.0178*	*0.019*
RMS(angles)	*1.75*	*1.780*	*1.87*
Ramachandran favored (%)	*95*	*97.2*	*98*
Ramachandran outliers (%)	*0*	*0*	*0*
Clash score	*2.71*	*1.24*	*3.2*
Average B-factor	*17.60*	*10*	*12.50*
macromolecules	*14.80*	*7.5*	*9.80*
ligands	*16.30*	*4.4*	*25.60*
solvent	*31.00*	*22.90*	*25.70*
PDB accession code	*5AOZ*	*5M2O*	*5M2S*

Values in parenthesis are for the highest resolution shell. # Matthews coefficient (Matthews, 1968). † $CC_{1/2}$ = the correlation between intensities from random half-dataset (Diederichs & Karplus, 2013) ‡ $R_{merge} = \Sigma_{hkl} \Sigma_i |I_i(hkl) - \langle I(hkl)\rangle| / \Sigma_{hkl} \Sigma_i I_i(hkl)$, where $I_i(hkl)$ is the *i*th intensity measurement of reflection *hkl*, including symmetry-related reflections and $\langle I(hkl)\rangle$ is its average. §$R_{work} = \Sigma_{hkl}||F_{obs}|-|F_{calc}||/\Sigma_{hkl}|F_{obs}|$. ¥$R_{free}$ as R_{work}, but summed over a 5% test set of reflections.

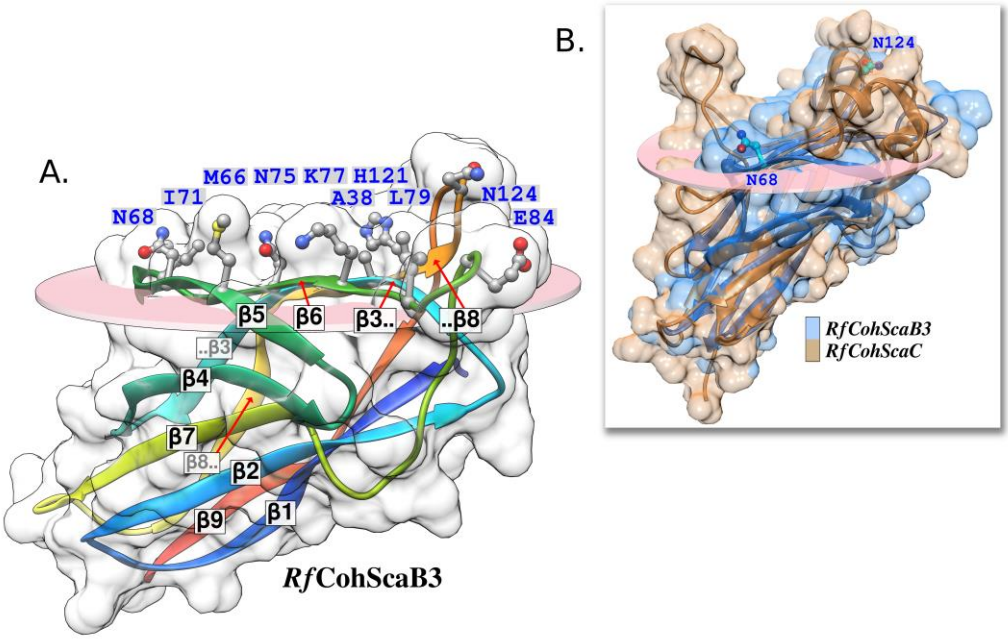
4.3. Results and Discussion

4.3.1. Structure of *R. flavefaciens* ScaB cohesin 3 (*Rf*CohScaB3)

In an initial attempt to understand the structural determinants of Coh-Doc specificity that orchestrate the correct assembly of *R. flavefaciens* cellulosome, the structure of the third Coh of ScaB, termed *Rf*CohScaB3, was solved by SAD phasing. Crystals belong to space group $P4_12_12$ with unit cell dimensions of $a = b = 60.43$ Å, $c = 86.51$ Å. Final data and structure-

quality statistics are shown in Table 4.2. *Rf*CohScaB3 displays an elliptical structure with nine β-strands, which form two β-sheets aligned in an elongated *β*-barrel that displays a classical "jelly-roll fold" (Figure 4.2A). The two sheets comprise *β*-strands 9, 1, 2, 7, 4 on one face and *β*-strands 8, 3, 6, 5 on the other face. Strands 1 and 9 align parallel to each other, thus completing the jelly-roll, while the other *β*-strands are antiparallel. Structural similarity search using the PDBeFold server (http://www.ebi.ac.uk/msd-srv/ssm/) revealed that the closest, functionally relevant, structural homologs of *Rf*CohScaB3 are Cohs that bind Docs appended to enzymes, although levels of sequence similarity were relatively low. They include the Cohs from *C. thermocellum* ScaA (PDB code 1AOH; z score of 6.4 and root mean square deviation (r.m.s.d) of 2.3 Å over 126 aligned residues), *Pseudobacteroides cellulosolvens* ScaB (PDB code 4UMS; z-score of 6.6 and r.m.s.d of 1.97 Å over 120 aligned residues), *C. cellulolyticum* ScaA (PDB code 2VN5; z-score of 6.8 and r.m.s.d of 2.3 Å over 124 aligned residues) and *R. flavefaciens* ScaC cohesin in complex with *Rf*Doc3 (PDB code 5LXV; z-score of 6.9 and r.m.s.d of 2.1 Å over 124 aligned residues). Major differences between the Coh structures were observed at *β*-sheet 8-3-6-5, which constitutes the protein-interacting interface (Figure 4.3). In particular, the ligand binding interfaces of *Rf*CohScaB3 and *Rf*CohScaC are dramatically different explaining differences in specificity as will be described below (Figure 4.2B). These observations suggest that *Rf*CohScaB3 displays a unique mechanism of dockerin recognition not described in other Coh-Doc complexes.

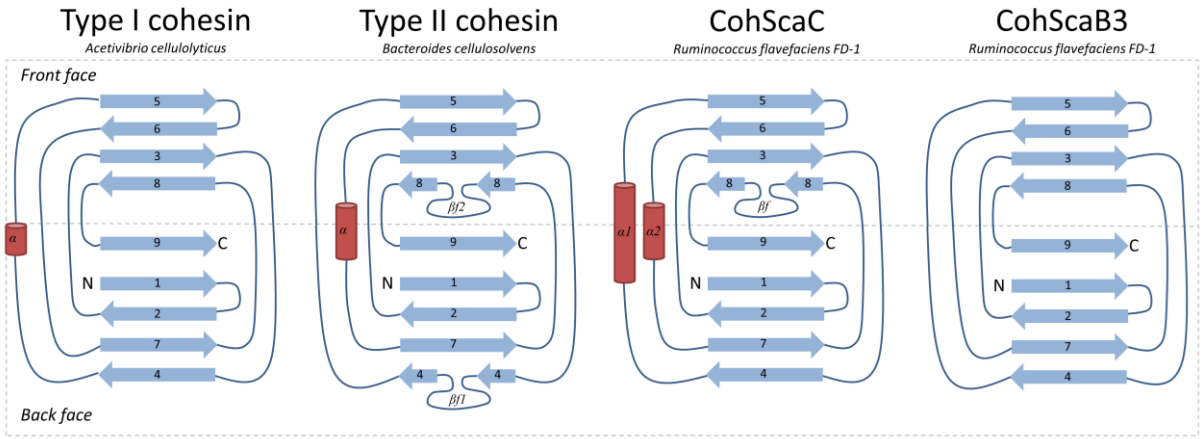
Figure 4.2 Structure of *Rf*CohScaB3.



A. The structure of CohScaB3 is represented in color ramped style from the blue N-terminus to the red C-terminus. Below the transparent molecular surface, the most important residues for dockerin

interaction are shown in ball&stick representation, above the pink oval disk that marks the plane defined by the 8-3-6-5 β-sheets. Each of the 9 β-strands is labeled. *B.* Overlay of *Rf*CohScaB3 with *Rf*CohScaC, with the blue and tan colored transparent molecular surface, respectively, revealing the secondary structure and the major differences, particularly at the dockerin-interacting plateau highlighted above the same oval pink plane representation. *Rf*CohScaB3 cyan-colored residues N-68 and N-124 were left on panel B. as orientation reference points relative to panel A.

Figure 4.3 Topology diagram of *Rf*CohScaB3 compared with previously described cohesins and *Rf*CohScaC.



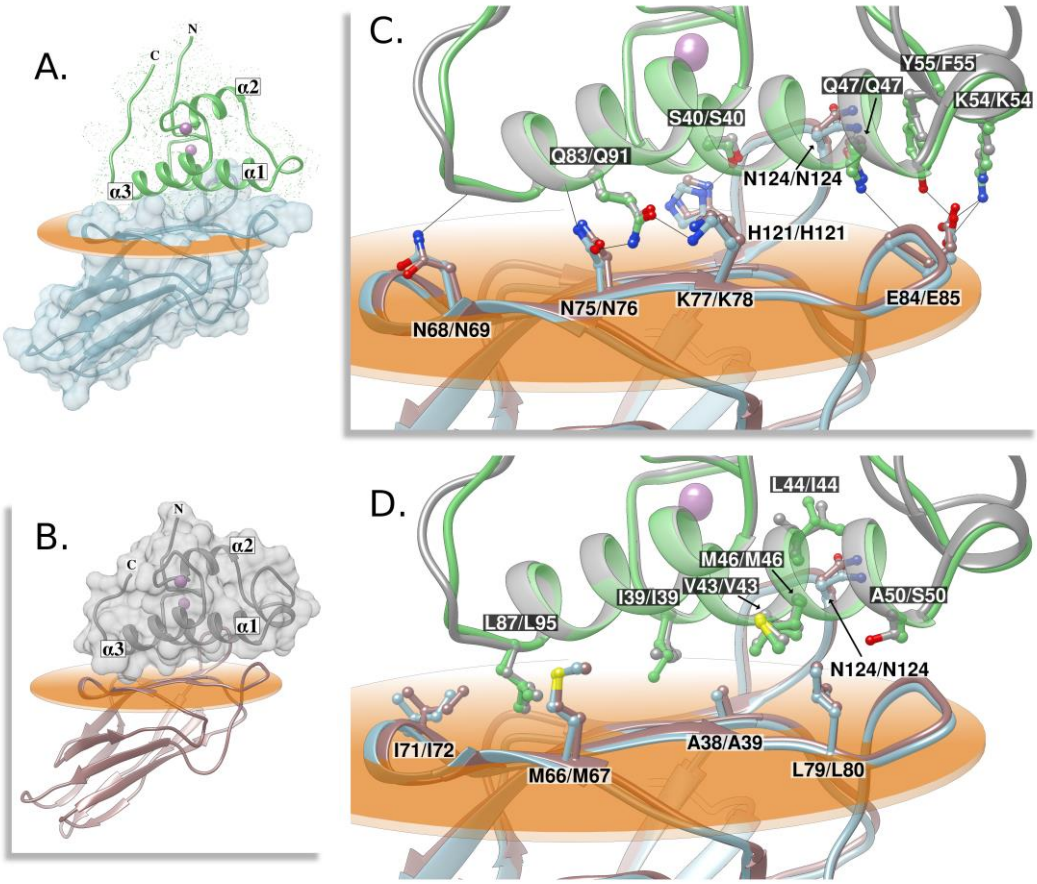
The CohScaB3 module (last) forms the classical nine-stranded β-sandwich with jelly-roll topology, which is essentially analogous to that of the cohesin modules *Ac*ScaCCoh3 (first, PDB code 4UYP) and *Bc*ScaACoh11 (second, PDB code 1TYJ), respectively. Unlike *Ac*ScaCCoh3 and *Bc*ScaACoh11, *Rf*CohScaC (third, PDB code 5LXV) does not possess any β-flap extensions interrupting β-strands or α-helices between β-strands.

4.3.2. Structure of novel *R. flavefaciens* Coh-Doc complexes

In a previous study (Israeli-Ruimy *et al.*, 2017), ScaB Cohs 1 to 4 and ScaA Cohs were shown to bind specifically to group 1 Docs. In those studies, highly stable complexes were formed between *Rf*CohScaB3 and a group 1a Doc, *Rf*Doc1a, and between *Rf*CohScaA and a group 1b Doc, *Rf*Doc1b. *Rf*Doc1a is a component of a family 12 carbohydrate esterase, and *Rf*Doc1b is the C-terminal component of a family 9 glycoside hydrolase. To gain insight into the molecular mechanisms of cellulosome assembly the X-ray crystal structures of *R. flavefaciens* ScaA and ScaB Cohs in complex with group 1b and 1a Docs, defined as *Rf*CohScaA-Doc1b and *Rf*CohScaB3-Doc1a, respectively, were determined. The structure of *Rf*CohScaB3-Doc1a was solved by molecular replacement using the *Rf*CohScaB3 structure, described above, as the search model. The *Rf*CohScaB3-Doc1a structure includes a single copy of the heterodimer in the asymmetric unit, as well as 323 water molecules, with *Rf*Doc1a coordinating two calcium ions. The complex displays an elongated shape with overall dimensions of 40 × 35 × 66 Å and includes residues 5 – 141 of *Rf*CohScaB3 and residues 23 – 96 of *Rf*Doc1a from *R. flavefaciens*

FD-1 (Figure 4.4A). The structure of *Rf*CohScaA-Doc1b was also solved by molecular replacement using *Rf*CohScaB3-Doc1a as the search model. Like *Rf*CohScaB3-Doc1a it includes a single copy of the heterodimer in the asymmetric unit, 325 water molecules and 2 calcium ions coordinated by the Doc. *Rf*CohScaA-Doc1b is virtually identical to *Rf*CohScaB3-Doc1a and includes residues 3 – 143 from *Rf*CohScaA and residues 24 – 102 from *Rf*Doc1b (Figure 4.4B). Crystal parameters for the structure of the two protein complexes and data collection statistics are summarized in Table 4.2. In both Coh-Doc complexes the group 1 Docs bind the 8-3-5-6 sheet of the *Rf*CohScaB3 and *Rf*CohScaA β-sandwiches, which present a predominantly flat surface. Significantly, the structures of the *Rf*CohScaB3-Doc1a and *Rf*CohScaA-Doc1b complexes were found to be very similar to each other, with an average r.m.s.d of 0.6 Å for the two chains (Figure 4.4C,D). This reflects the high degree of primary structure identity (72.7% for the Cohs and 42.2% for the Docs) shown by the two complementary protein modules.

Figure 4.4 Structure and cohesin-dockerin interface of *Rf*CohScaB3-Doc1a and *Rf*CohScaA-Doc1b.



A. Structure of *Rf*CohScaB3-Doc1a complex with the dockerin in green and the cohesin in light blue. The dockerin N- and C- terminus and the α-helices are labeled, and a dotted molecular surface representation is shown. The cohesin blue molecular surface is represented. *B.* Structure of *Rf*CohScaA-Doc1b complex with the dockerin in gray and the cohesin in brown, using a similar layout as in panel

A. but showing instead the transparent gray molecular surface of the dockerin. *C.* Overlay of both complexes showing the main polar interactions at the Coh-Doc interface. D. Overlay of both complexes showing the main hydrophobic interactions at the Coh-Doc interface. In panels C. and D. the most important residues involved in Coh-Doc recognition are depicted as ball&stick configuration, with a dark background label for the Doc residues and a light background label for the Coh residues, using the Doc1a/Doc1b and CohScaB3/CohScaA numbering. Solid black lines mark hydrogen-bonds interactions. Ca^{2+} ions are depicted as purple spheres. In all panels, the transparent orange disk marks the plane defined by the 8-3-6-5 β-sheet, where the β-strands form a distinctive dockerin-interacting plateau.

4.3.3. Structures of *Rf*CohScaB3 and *Rf*CohScaA in complex with their cognate Docs

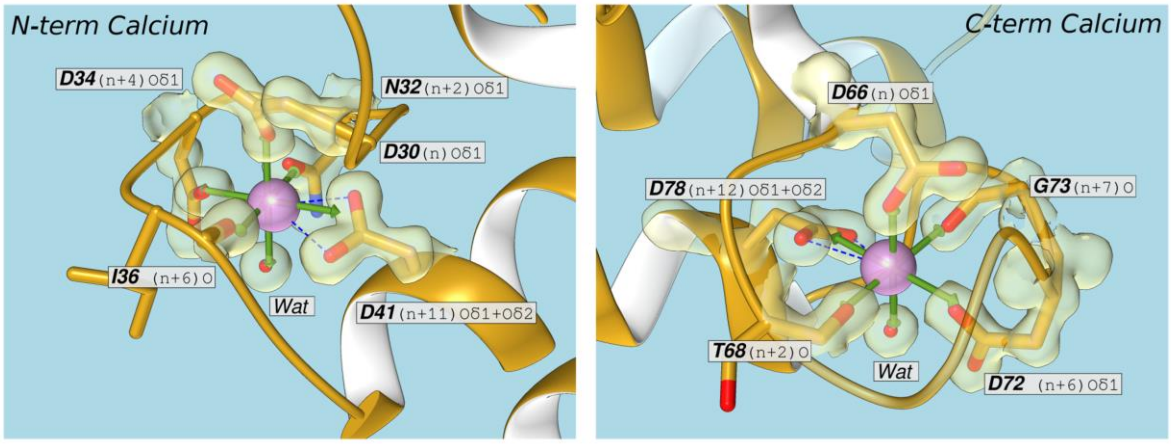
The structures of *R. flavefaciens Rf*CohScaB3 and *Rf*CohScaA Cohs in complex with *Rf*Doc1a and *Rf*Doc1b, respectively display striking structural similarities presenting a r.m.s.d of 0.45 Å over 136 main chain carbon atoms. As proposed above, the Doc-interacting β-sandwich face comprised β-strands 8, 3, 6 and 5 (Figure 4.3). No α-helices were identified in *Rf*CohScaB3 and *Rf*CohScaA Cohs (Figure 4.4A, B; Figure 4.3), and they thus lack the distinctive α-helix connecting β-strands 4 and 5 in other bacterial Cohs as well as the large β-flap disrupting β-strand 8, previously observed in the *R. flavefaciens* ScaC group 3 Coh ((Bule *et al.*, 2016); Figure 4.3). The structure of *Rf*CohScaB3, whether unbound or in complex with *Rf*Doc1a, was essentially identical (r.m.s.d ~ 0.37 Å). Thus, similar to previous descriptions (Carvalho *et al.*, 2007; Pinheiro *et al.*, 2008), Cohs appear to be highly stable modules that do not undergo significant conformational changes upon binding to their Doc ligands.

4.3.4. Structures of *Rf*Doc1a and *Rf*Doc1b in complex with their cognate Cohs

The structures of *Rf*Doc1a and *Rf*Doc1b in complex with *Rf*CohScaB3 and *Rf*CohScaA Cohs, respectively, comprise two α-helices arranged in antiparallel orientation extending from residues (using *Rf*Doc1a/*Rf*Doc1b numbering) Ile-39/Ile-39 to Tyr-55/Phe-55 (helix-1) and Val-76/Asn-84 to Leu-89/Leu-97 (helix-3). The two loops connecting these structural elements, in *Rf*Doc1a and *Rf*Doc1b, contain a seven-residue α-helix (helix-2) extending from Asp-59/Ala-67 to Ala-65/Gly-73, respectively (Figure 4.4A, B). The tertiary structures of *Rf*Doc1a and *Rf*Doc1b adopt a similar fold with an r.m.s.d of 0.9 Å over 68 main chain carbon atoms. Major structural differences between *Rf*Doc1a and *Rf*Doc1b Docs involve the loop extending from helix-1 and helix-2, which is longer in *Rf*Doc1b reflecting the previously identified longer linker region connecting the two duplicated repeats of group 1b Docs (Rincon *et al.*, 2010). The overall tertiary structure of *Rf*Doc1a and *Rf*Doc1b is very similar to the enzyme-borne Docs from *C. thermocellum* (r.m.s.d of ~1.4 Å, over 64 residues), *A. cellulolyticus* (r.m.s.d of ~1.8

Å, over 67 residues), and *R. flavefaciens* group 3 Doc (Doc3) that binds the ScaC Coh (r.m.s.d of 1.82 Å, over 59 residues). Both *Rf*Doc1a and *Rf*Doc1b contain two Ca^{2+} ions coordinated by several amino-acid residues, similar to the canonical EF-hand loop motif described in all other Docs(Kretsinger & Nockolds, 1973). The Ca^{2+} bound to the N-terminal repeat has a typical *n, n+2, n+4, n+11*, plus a water molecule, pattern of coordination (Figure 4.5). In contrast, the second Ca^{2+}-binding region has an atypical coordination arrangement of *n, n+6, n+12* plus a water molecule (Figure 4.5).

Figure 4.5 Dockerin *Rf*Doc1a calcium octahedral coordination.



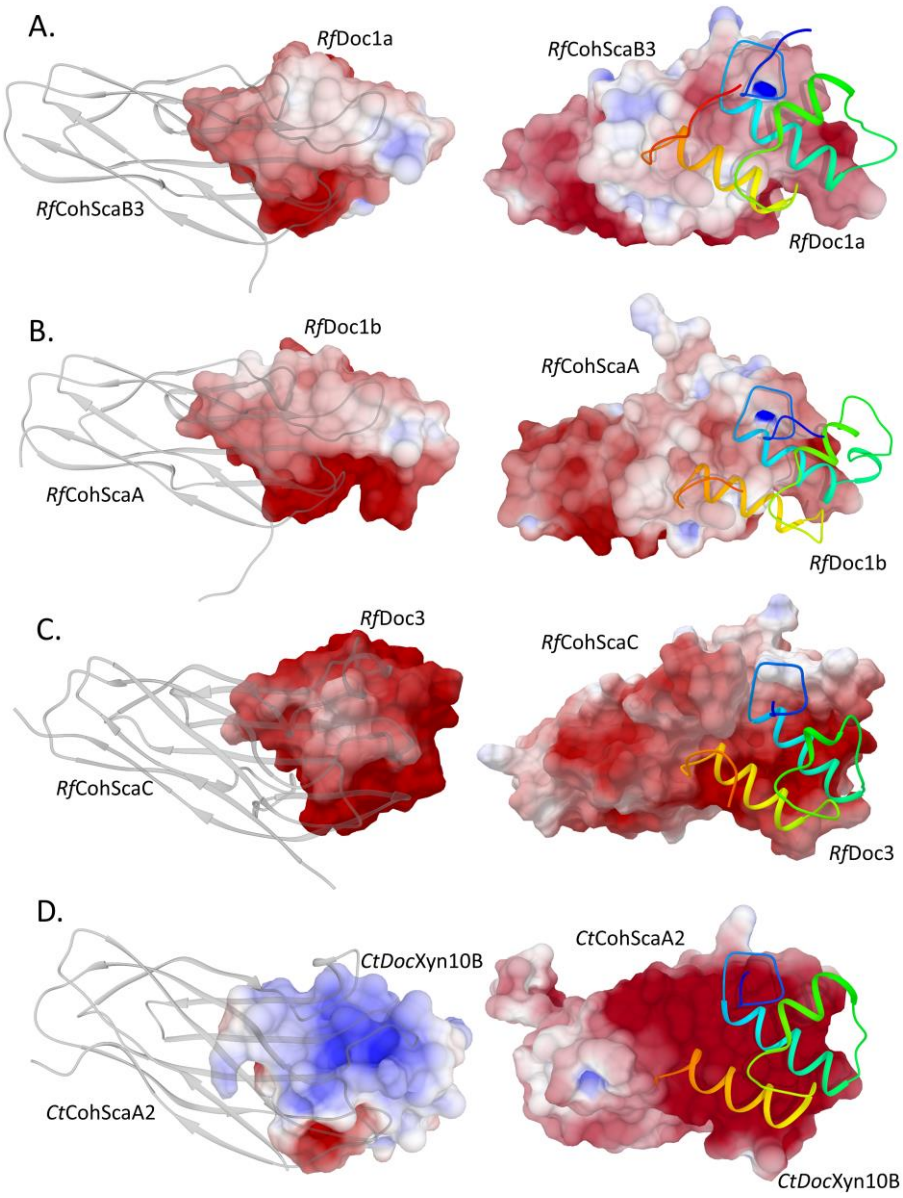
The left and right panels show a representation of the *Rf*Doc1a N- and C-terminal Ca^{2+} ions, respectively. In both panels the secondary structure ribbon representation of *Rf*Doc1a highlights the amino-acid residues (in stick representation) involved in the metal coordination, surrounded by a transparent light yellow representation of the Refmac5 maximum-likelihood σ_A–weighted $2F_o-F_c$ electron density map contoured at 1σ (0.46 electrons/A^3). The labels show the *Rf*Doc1a residue and coordination position numbers and also the atoms involved. Both calcium ions are depicted as purple spheres and are overlaid with an idealized octahedral geometry representation (green arrows). A single water molecule (Wat) completes the coordination sphere. The bidentate nature of the Asp-34 and Asp-78 coordination is highlighted with blue dashed lines.

4.3.5. *Rf*CohScaB3-Doc1a and *Rf*CohScaA-Doc1b complex interfaces

*Rf*Doc1a and *Rf*Doc1b helices 1 and 3 make various contacts with the surface of 8-3-6-5 *β*-sheets of *Rf*CohScaB3 and *Rf*CohScaA, respectively (Figure 4.4C,D). Although the Coh-interacting platform is predominantly flat, the loop connecting *β*-strands 8 and 9 is elevated in relation to the 8-3-6-5 plane, thus remaining in close proximity to the N-terminus of helix-1 in the Doc structure. A slight elevation is also observed in the loop connecting *β*-strands 6 and 7, leading to a closer interaction with the C-terminus of helix-1. This means that the entire length of helix-1 of *Rf*Doc1a and *Rf*Doc1b interacts with the Coh surface, while helix-3 binds the Coh platform predominantly by the C-terminus. This contrasts with the interface of the recently described *R. flavefaciens Rf*CohScaC-Doc3 complex where the two Doc3 helices (helix 1 and

helix 3) make similar contributions to CohScaC recognition (Bule *et al.*, 2016). In *Rf*CohScaC-Doc3, CohScaC's α-helix located between β-strands 4 and 5, which is absent in Rf*CohScaB3 and* Rf*CohScaA*, is elevated in relation to the 8-3-6-5 plane allowing the entire Coh surface to be in closer proximity to both Doc α-helices. The surface electrostatic potential calculated for *Rf*CohScaB3-Doc1a and *Rf*CohScaA-Doc1b complexes reveal that the Coh- and Doc-interacting faces are predominantly uncharged (Figure 4.6). This is in contrast with *C. thermocellum* Coh-Doc complexes where a predominantly positive-charged Doc binds a negatively charged Coh, while the *Rf*CohScaC-Doc3 complex interface has an intermediate charge (Figure 4.6).

Figure 4.6 Electrostatic surface potential for the Coh-Doc interface.



In each panel the right images show the cohesin binding plateau with the bound dockerin partner in N- to C-terminus rainbow color-ramped style on top, while the left images depict the same complex after a

180º rotation along axis "X", thus allowing the view of the molecular dockerin-binding surface, below a transparent view of the secondary structure of the Coh partner. *A.* *Rf*CohScaB3 and *Rf*Doc1a (PDB code 5M2O). *B.* *Rf*CohScaA and *Rf*Doc1b (PDB code 5M2S). *C.* *Rf*CohScaC and *Rf*Doc3 (PDB code 5LXV) and *D.* *Ct*CohScaA2 and *Ct*DocXyn10B (PDB code 2CCL). The figure was prepared with UCSF Chimera using APBS (Adaptive Poisson-Boltzmann Solver) and the electrostatic potential was contoured from -6 (red) to +6 (blue) (arbitrary Chimera units).

A large network of polar (Table 4.3) and hydrophobic interactions (Table 4.4) were identified at the *Rf*CohScaB3-Doc1a and *Rf*CohScaA-Doc1b complex interfaces (Figure 4.4C,D). Although a few differences were observed, the contacts are highly conserved between the two complexes (Figure 4.4C,D).

Table 4.3 . Main polar contacts between *Rf*CohScaB3 and *Rf*Doc1a and *Rf*ScaACoh and *Rf*Doc1b.

Hydrogen Bonds								
***Rf*Doc1a**					***Rf*CohScaB3**			
		Atom	**Residue**	**Residue #**		**Atom**	**Residue**	**Residue #**
		ND2	ASN	32	< >	O	ASN	124
H1		OG	SER	40	< >	ND1	HIS	121
H1		OG	SER	40	< >	ND1	HIS	121
H1		NE2	GLN	47	< >	O	GLY	83
H1		OE1	GLN	47	< >	ND2	ASN	124
		NZ	LYS	54	< >	OE2	GLU	84
		OH	TYR	55	< >	OE2	GLU	84
	H3	NE2	GLN	83	< >	OD1	ASN	75
	H3	O	GLN	83	< >	ND2	ASN	75
	H3	OE1	GLN	83	< >	NZ	LYS	77
	H3	O	CYS	86	< >	NZ	LYS	117
	H3	O	LEU	87	< >	ND2	ASN	68
***Rf*Doc1b**					***Rf*CohScaA**			
		Atom	**Residue**	**Residue #**		**Atom**	**Residue**	**Residue #**
		ND2	ASN	32	< >	O	ASN	124
H1		OG	SER	40	< >	ND1	HIS	121
H1		OG	SER	40	< >	ND1	HIS	121
H1		OE1	GLN	47	< >	ND2	ASN	124
H1		NE2	GLN	47	< >	O	GLY	84
		NZ	LYS	54	< >	OE2	GLU	85
		O	HIS	63	< >	ND2	ASN	124
	H3	OE1	GLN	91	< >	NZ	LYS	78
	H3	NE2	GLN	91	< >	OD2	ASN	76
	H3	O	LEU	95	< >	ND2	ASN	69

Table was made using the PDBePISA server and the contacts were further verified manually with Coot. Some of the dockerin residues are marked as belonging either to helix 1 (H1) or to helix 3 (H3) interfaces.

The interactions between α-helix-1 of the Docs and the *R. flavefaciens* Cohs are dominated by Ile-39, Ser-40, Val-43, Met-46, Gln-47 and Lys-54 of *Rf*Doc1a and *Rf*Doc1b and His-121/His-121, Ala-38/Ala-39, Leu-79/Leu-80 and Glu-84/Glu-85 of *Rf*CohScaB3/*Rf*CohScaA Cohs, respectively (Figure 4.4C,D). The side chains of the Ile-39/Val-43, at positions 11 and 15 of *Rf*Doc1a and *Rf*Doc1b, dominate the hydrophobic recognition of the Coh by contacting with

the hydrophobic platform of the Coh created by Ala-38/Ala-39 and Leu-79/Leu-80 in *Rf*CohScaB3/*Rf*CohScaA, respectively. The highly hydrophobic character of α-helix-1 interaction is reinforced by the contacts established by Leu-44/Ile-44, Met-46/Met-46 and Ala-50/Ser-50 of *Rf*Doc1a/*Rf*Doc1b with *Rf*CohScaB3/*Rf*CohScaA Leu-79/Leu-80 and the aliphatic region of Asn-124 side-chain. The hydrogen bond network established by α-helix 1 is dominated by the interaction of Ser-40, Gln-47 and Lys-54 with His-121/His-121, Asn-124/Asn-124 and Glu-84/Glu-85 of *Rf*CohScaB3/*Rf*CohScaA, respectively. The two Docs are less conserved at the C-terminus of helix-1 and this generates differences in the interaction with the Coh. Thus, *Rf*Doc1a establishes an extra hydrogen bond between Tyr-55 Oη (Phe-55 in *Rf*Doc1b) and Oδ2 of *Rf*CohScaB3 Glu-84. In addition, the longer loop connecting helices 1 and 2 in *Rf*Doc1b allows the carbonyl of His-63 to form a hydrogen bond with Asn-124 Nδ2 of *Rf*CohScaA. In α-helix-3 the contacts are dominated by the important salt bridges established between Nε2 and Oε1 of Gln-83/Gln-91 of *Rf*Doc1a/*Rf*Doc1b with Oδ1 of Asn-75/Asn-76 and Nζ of Lys-77/Lys-78 of *Rf*CohScaB3/*Rf*CohScaA Cohs. In addition, the side chains of Leu-87/Leu-95 of *Rf*Doc1a/*Rf*Doc1b occupy the hydrophobic pocket created by Gly-73/Gly-74, Ile-71/Ile-72 and the aliphatic portion of Met-66/Met-67 of *Rf*CohScaB3/*Rf*CohScaA Cohs. The closer proximity of the two protein partners at the C-terminus of helix-3 in *Rf*CohScaB3-Doc1a protein complex allows the formation of two extra hydrogen bonds between *Rf*Doc1a and *Rf*CohScaB3 that are absent in *Rf*CohScaA-Doc1b.

Table 4.4 Main hydrophobic contacts between *Rf*CohScaB3 and *Rf*Doc1a and *Rf*CohScaA and *Rf*Doc1b.

	*Rf*Doc1a			*Rf*CohScaB3
	Residue	**Residue #**		**Residues**
	ASN	32	< >	ASN124, ASP125, GLY126
	ASP	34	< >	ASP125, GLY126
	ASP	38	< >	HIS121
H1	ILE	39	< >	ALA38, MET39, PHE76, HIS121
H1	SER	40	< >	HIS121, SER123, ASN 124, GLY126
H1	VAL	43	< >	SER37, ALA38, SER123
H1	MET	46	< >	LYS77, LEU79
H1	GLN	47	< >	GLY83, ASN124
H1	ALA	50	< >	ASP81, LYS82, GLY83
	ASN	51	< >	LYS82
	LYS	54	< >	GLU84
	TYR	55	< >	GLU84, ASN124
H3	GLN	80	< >	MET66
H3	GLN	83	< >	MET66, ASN75, PHE76, LYS77
H3	SER	84	< >	MET66
H3	CYS	86	< >	ASP40, ASN75, LYS117
H3	LEU	87	< >	MET66, ASN68, ILE71, GLY73, ALA74, ASN75
	LEU	89	< >	ASN68
	*Rf*Doc1b			*Rf*CohScaA
	Residue	**Residue #**		**Residues**
	ASN	32	< >	ASN124, ASP125, GLY126
	ASP	34	< >	ASP125, GLY126
	ASP	38	< >	HIS121
H1	ILE	39	< >	ALA39, PHE77, HIS121,
H1	SER	40	< >	HIS121, ASN124, GLY126, SER 123, ASP125
H1	VAL	43	< >	SER123, ASN124, SER38, ALA39
H1	ILE	44	< >	ASN124
H1	MET	46	< >	LEU80, LYS 78,
H1	GLN	47	< >	GLY84, ASN124
H1	SER	50	< >	ASP82, LYS83, GLY84
H1	ASN	51	< >	LYS83
	LYS	54	< >	GLU85
	PHE	55	< >	GLU85
	HIS	63	< >	ASN124
H3	LEU	88	< >	MET67
H3	GLN	91	< >	ASN76, MET67, LYS78
H3	LYS	92	< >	MET67
H3	LEU	94	< >	LYS117, ASP41, ASN76
H3	LEU	95	< >	ILE72, ASN69, ASN76, MET67, GLY74, ALA75
H3	ASN	96	< >	ASN69
H3	LEU	97	< >	THR68, ASN69

Table was made using the PDBePISA server. Some of the dockerin residues are marked as belonging either to helix 1 (H1) or to helix 3 (H3) interfaces.

4.3.6. *Rf*Doc1a and *Rf*Doc1b present a single Coh-binding interface

The binding thermodynamics of *Rf*Doc1a and *Rf*Doc1b to *Rf*CohScaB3 and *Rf*CohScaA were assessed by isothermal titration calorimetry (ITC) at 308 K, consistent with the approximate temperature of the rumen. The data, presented in Table 4.5 and exemplified in Figure 4.7, revealed a macromolecular association with a 1:1 stoichiometry and a K_a of ~ 10^7-10^8 M^{-1}, an affinity similar to other Coh-Doc interactions. Binding was driven by changes in enthalpy with the reduction in entropy having a negative impact on affinity.

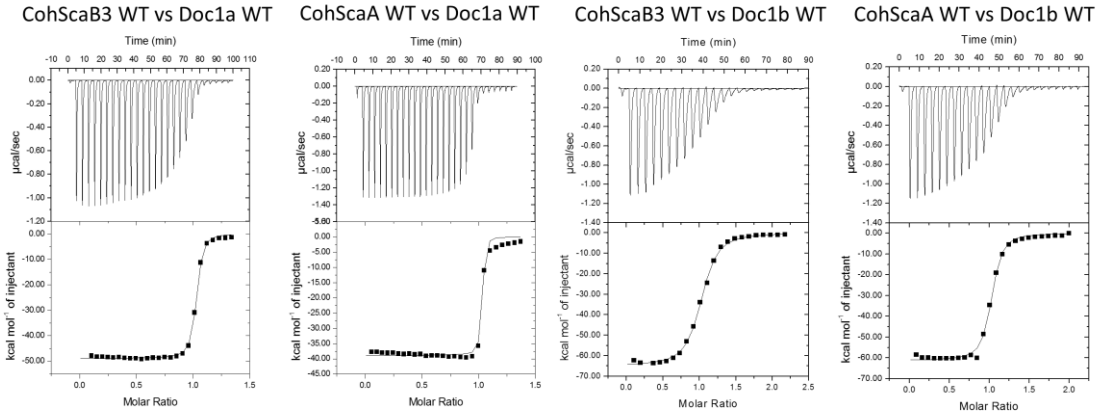
Table 4.5 Thermodynamics of the several interactions tested by ITC. All Thermodynamic parameters were determined at 308 K.

Cohesin	*Dockerin*	*K_a M^{-1}*	*ΔG^o kcal mol^{-1}*	*ΔH kcal mol^{-1}*	*$T\Delta S^o$ kcal mol^{-1}*	*N*
CohScaA	**Doc1bWT**	2.67E7 ± 3.78E6	-10.37	-61.19 ± 0.50	-50.82	1
	Doc1aWT	5.03E8 ± 2.36E8	-12.28	-38.92 ± 0.33	-26.64	1
CohScaB3 WT	**Doc1b WT**	1.03E7±7.63E5	-9.80	-64.94 ± 0.44	-55.13	1
	Doc1aWT	1.18E8 ± 2.00E7	-11,23	-50.68 ± 0.29	-39.44	1
	Doc1a I39A	3.86E6 ± 7.66E4	-9.18	-57.26 ± 0.12	-48.07	1
	Doc1a S40A	1.09E8 ± 1.17E7	-11.47	-46.60 ± 0.15	-35.12	1
	Doc1a V43A	1.91E6 ± 3.02E4	-8.94	-52.08 ± 0.14	-43.14	1
	Doc1a Q47A	1.71E7 ± 6.89E5	-10.05	-48.27 ± 0.12	-38.21	1
	Doc1a K54A	1.96E7 ± 1.28E6	-10.42	-59.73 ± 0.25	-49.30	1
	Doc1a Q83A	1.54E8 ± 1.03E7	-11.71	-39.45 ± 0.83	-27.73	1
	Doc1a L87A	2.81E7 ± 1.61E6	-10.53	-59.84 ± 0.19	-49.30	1
	Doc1a I39A + V43A	*Nb**	*Nb**	*Nb**	*Nb**	*Nb**
	Doc1a V43A + Q47A	4.36E5 ± 9.66E3	-7.81	-48.79 ± 0.39	-40.98	1
CohScaB3 A38Q		*Nb**	*Nb**	*Nb**	*Nb**	*Nb**
CohScaB3 N68A		1.10E8 ± 1.20E7	-11.48	-58.63 ± 0.22	-47.15	1
CohScaB3 N75A		9.09E7 ± 8.96E6	-11.10	-52.70 ± 0.17	-41.60	1
CohScaB3 K77A		4.23E8 ± 7.13E7	-12.12	-57.42 ± 0.23	-45,30	1
CohScaB3 L79A		7.59E6 ± 3.73E5	-9.71	-51.93 ± 0.20	-42.21	1
CohScaB3 E84A		3.62E7 ± 3.45E6	-10.77	-55.45 ± 0.26	-44.68	1
CohScaB3 H121A		2.66E7 ± 1.74E6	-10.41	-52.63 ± 0.18	-42.27	1
CohScaB3 N124A	**Doc1aWT**	5.54E7 ± 4.68E6	-10.89	-52.80 ± 0.19	-41.90	1
CohScaB3 E84A + H121A		1.65E6 ± 2.50E5	-8.62	-66.86 ± 1.49	-58.24	1
CohScaB3 N75A + H121A		2.49E6 ± 6.66E4	-8.93	-50.84 ± 0.18	-41.90	1
CohScaB3 N75A + N124A		1.86E6 ± 3.59E5	-8.85	-56.31 ± 1.56	-47.45	1
CohScaB3 N75A + E84A		2.08E7 ± 1.29E6	-10.47	-51.76 ± 0.18	-41.29	1
CohScaB3 E84A + H121A		1.65E6 ± 2.50E5	-8.62	-66.86 ± 1.49	-58.24	1
CohScaB3 E84A + N124A		1.53E7 ± 9.83E5	-10.05	-55.04 ± 0.23	-44.99	1
CohScaB3 H121A + N124A		2.28E6 ± 9.39E4	-8.84	-46.44 ± 0.24	-37.59	1

**Nb* - No binding

.

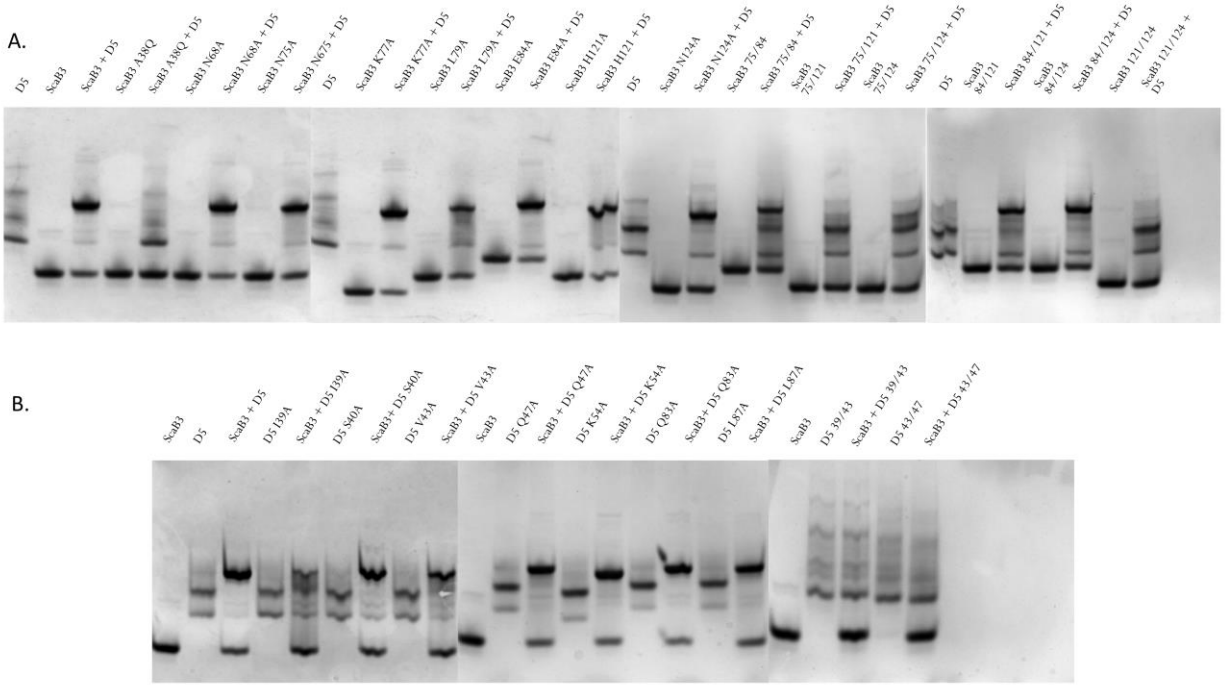
Figure 4.7 Binding affinity of wild-type *Rf*Doc1a and 1b to both *Rf*CohScaB3 and *Rf*CohScaA determined by ITC.



Binding isotherms for *A. Rf*CohScaB3 vs *Rf*Doc1a, *B. Rf*CohScaA vs *Rf*Doc1a, *C. Rf*CohScaB3 vs *Rf*Doc1b and *D. Rf*CohScaA vs *Rf*Doc1b are displayed. The upper part of each panel shows the raw heats of binding, whereas the lower parts comprise the integrated heats after correction for heat of dilution. The curve represents the best fit to a single-site binding model. The corresponding thermodynamic parameters are shown in Table 4.5.

The importance of *Rf*Doc1a and *Rf*CohScaB3 residues for Coh-Doc recognition was initially probed through non-denaturing gel electrophoresis (NGE) (Figure 4.8) and then extensively explored through ITC.

Figure 4.8 Binding affinity of CohScaB3 and Doc1a mutant derivatives to their wild-type partners, determined by non-dentaturing gel electrophoresis (NGE).

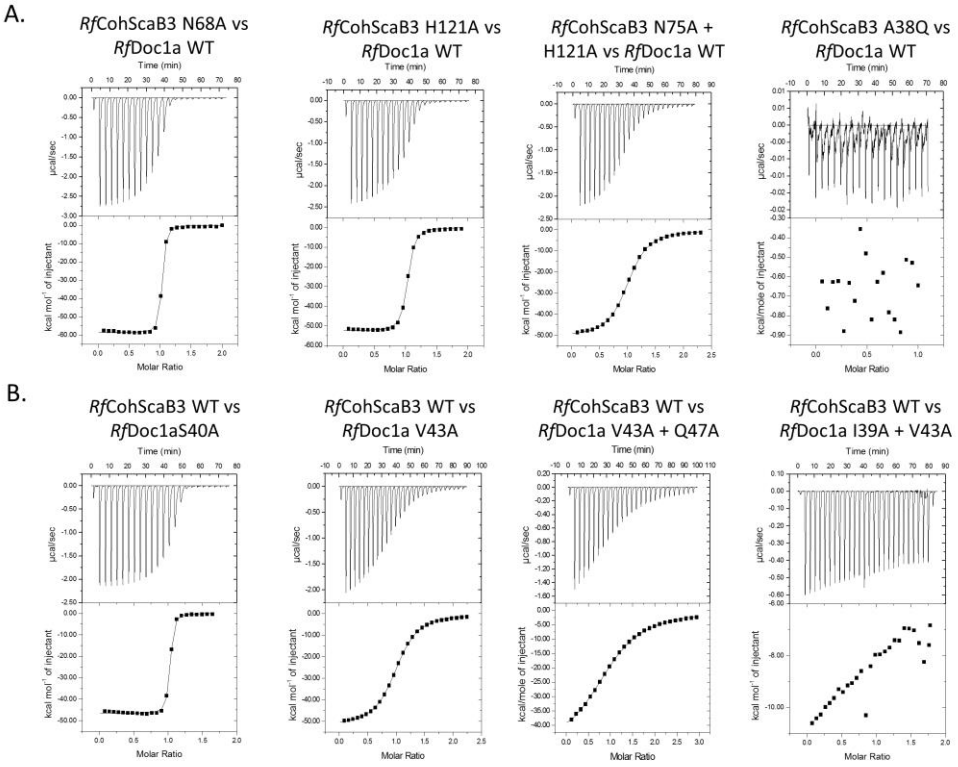


A. The lanes marked D5 were loaded with the dockerin alone. Adjacent lanes were loaded with the cohesin mutant derivatives and with both cohesin and dockerin modules after 60-min incubation at equimolar concentrations. *B.* The lanes marked ScaB3 were loaded with the cohesin alone. Adjacent lanes were loaded with the dockerin mutant derivatives and with both cohesin and dockerin modules after 60-min incubation at equimolar concentrations. The appearance of a band with a different migration pattern in lanes containing the complex represents a positive result (e.g. ScaB3), while a negative result (e.g. ScaB3 A38Q) is given by the appearance of only the individual dockerin and cohesin bands.

The data (Table 4.5, Figure 4.9) revealed that alanine substitutions of *Rf*Doc1a residues Ile-39 and Val-43 resulted in ~100-fold reduction in affinity of the Doc for the *Rf*CohScaB3. Complete abolition of Coh recognition resulted from the substitution of these two non-polar residues simultaneously (Table 4.5, Figure 4.9). The alanine substitution of *Rf*Doc1a residues that participate in the hydrogen bond network with the Coh (namely Ser-40, Gln-47, Lys-54 and Gln-83) had little impact on affinity (Table 4.5, Figure 4.9). In addition, a significant reduction in the affinity of *Rf*CohScaB3 for *Rf*Doc1a was observed following the substitution of Ala-38 with Gln and Leu-79 with Ala, the two Coh residues that create the hydrophobic environment at the *Rf*CohScaB3 platform that binds to *Rf*Doc1a. Again, the data suggest that *Rf*CohScaB3

residues that hydrogen bond with *Rf*Doc1a play a relatively small role in the binding; even when double mutants were generated the reduction in affinity was never higher than ~ 100 fold. Overall the data suggest that the residues that mostly influence *Rf*CohScaB3-Doc1a interaction are Ile-39 and Val-43 at helix-1 of *Rf*Doc1a and Ala-38 and Leu-79 located at the flat surface of *Rf*CohScaB3 8-3-6-5 β-sheet. Thus it seems that hydrophobic interactions play a major role in *Rf*CohScaB3-Doc1a assembly.

Figure 4.9 Determination of the contribution of key residues of *Rf*Doc1a and *Rf*CohScaB3 for the Coh-Doc interaction.

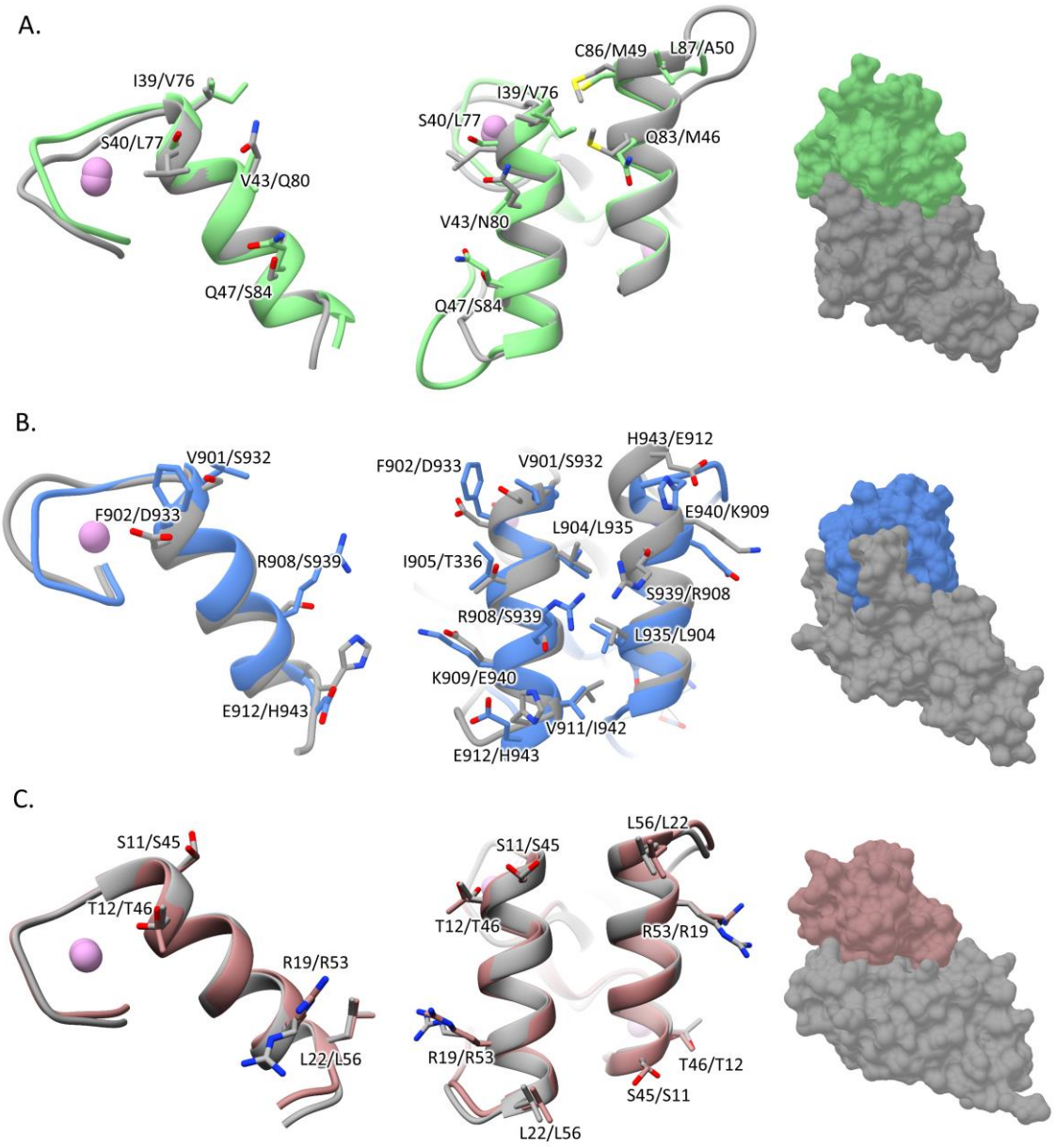


A. Representative binding isotherms of the interactions between the wild-type *Rf*Doc1a and several cohesin mutants. *B.* Representative binding isotherms of the interactions between the wild-type *Rf*CohScaB3 and several dockerin mutants. The isotherms are arranged according to loss of function, from no loss to complete loss. The upper part of each panel shows the raw heats of binding, whereas the lower parts comprise the integrated heats after correction for heat of dilution. The curve represents the best fit to a single-site binding model. The corresponding thermodynamic parameters are shown in Table 4.5.

The observation that the Ile-39Ala/Val-43Ala Doc mutant did not bind to its target Coh suggests that *R. flavefaciens* group 1 Docs present a single-binding mode, in contrast to previous observations for the majority of Docs appended to enzymes in other organisms. When Docs present a dual-binding mode, mutation of a single or two closely positioned residues usually has no effect on affinity, as the other (duplicated) binding site is functional and can be accessed

by its target Coh through a 180º rotation of the Doc. Inspection of the *Rf*CohScaB3-Doc1a structure revealed that the symmetry-related residues to Ile-39 and Val-43 (amino acids that occupy the equivalent position to Ile-39 and Val-43 when the Doc has been rotated 180º) are, respectively, Val-76 and Gln-80. While the side chain of Val-76 and Ile-39 are compatible, the bulky polar side chain of Gln-80 would be incompatible with the hydrophobic pocket in the cognate Coh that interacts with Val-43. Recent data revealed that both group 3 and group 6 *R. flavefaciens* Docs display a single-binding mode with the ScaC Coh. The internal symmetry of *R. flavefaciens* group 1 and group 3 Docs when compared with the well-described dual-binding mode of enzyme Docs from *C. thermocellum* was therefore probed by overlaying the various structures with their 2-fold related derivatives using the Matchmaker procedure from Chimera (Pettersen *et al.*, 2004). The superposition, displayed in Figure 4.10, highlights the lack of conservation in the contacting residues when the group 1 and group 3 Docs were overlayed with their 180-rotated versions. In addition to the previously mentioned changes in group 1 Docs, Ser-40 is replaced by the non-polar Leu-77 while the critical Gln-47 is replaced by Ser-84 (Figure 4.10). The lack of internal symmetry is also observed in the group 3 Docs, where both α-helices 1 and 3 are involved in Coh recognition. These data, together with the extensive mutagenesis analyses presented here, suggest that group 1 Docs display a single Coh-binding platform. In contrast, the superposition of *C. thermocellum* enzyme Docs revealed a well-defined internal symmetry with conservation of the Coh-interacting residues when the Doc is rotated by 180º, a property that supports a dual-binding mode (Figure 4.10).

Figure 4.10 Significant differences between the two cohesin-binding interfaces do not allow the dual-binding mode of dockerins from *R. flavefaciens*.

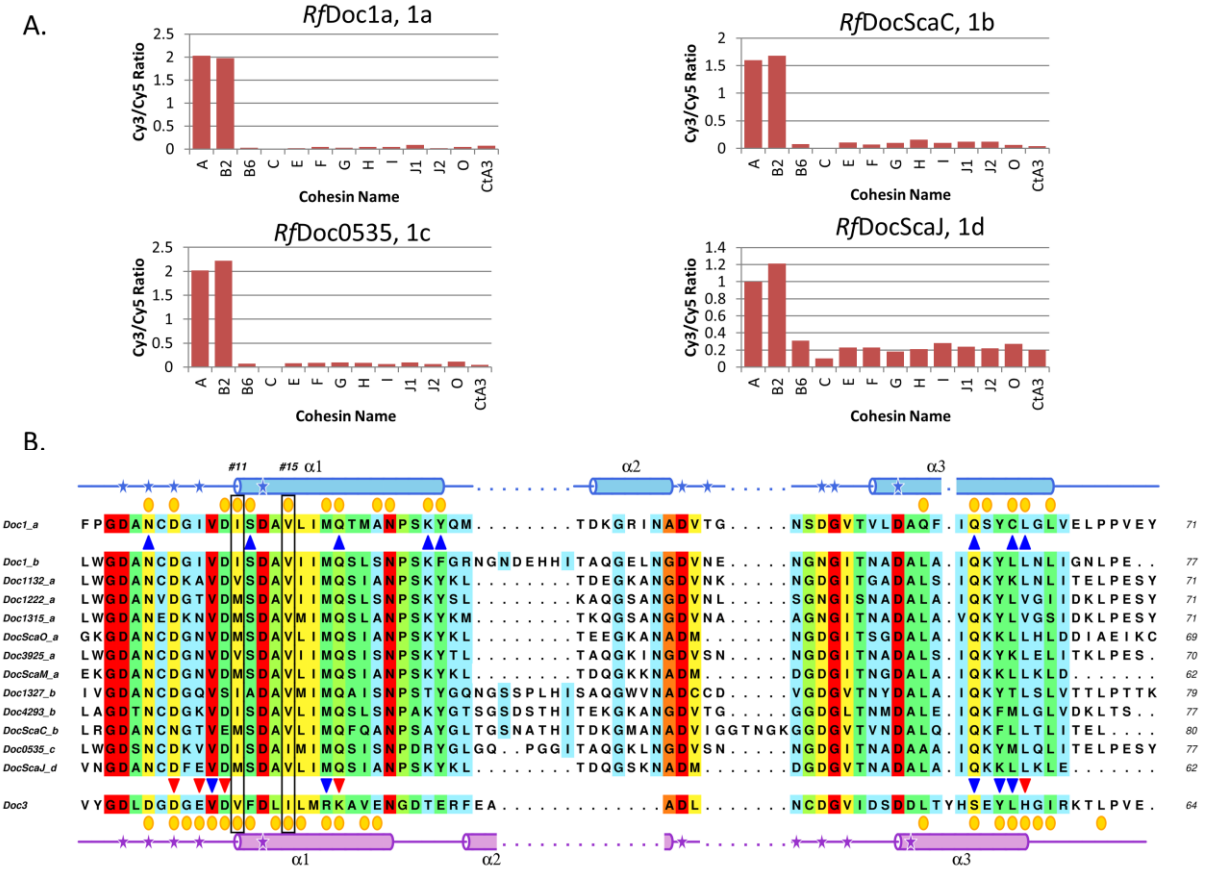


A. R. flavefaciens Group1 Doc. B. *R. flavefaciens* Group3 Doc. C. *C. thermocellum* Doc. The first image of each panel shows an overlay of the N-terminal and C-terminal dockerin repeats. In all cases it is apparent that both repeats are similar at the main-chain atoms but only the *C. thermocellum* Doc (*C*) shows conservation in the side chains, allowing the dual-binding mode. The middle image of each panel shows a comparison of the two putative binding surfaces by overlaying the dockerins with a version of themselves rotated by 180º (in grey) and shows a lack of conservation in the key contacting residues in both *R. flavefaciens* dockerins (*A,B)*. Contrary to the *C. thermocellum* Doc (*A*), lack of internal symmetry in Doc1a and Doc3 and the involvement of the two helices in cohesin recognition suggest that they display a single cohesin-binding platform. The final image of each panel shows the molecular surface of the several complexes, with the cohesin in grey and the dockerin in green (*Rf*Doc1a), blue (*Rf*Doc3) or pink (*C. thermocellum* Doc).

4.3.7. *R. flavefaciens* FD-1 Group 1 Docs have a functional conservation

The 96 group 1 Docs identified in the proteome of *R. flavefaciens* FD-1 were previously organized in 4 subgroups, termed 1a to 1d (Rincon *et al.*, 2010). *Rf*Doc1a and *Rf*Doc1b belong to group 1a (37 members) and group 1b (36 members), respectively, the most represented group 1 Docs. It was previously observed that group 1b Docs contain the longest linker region between the two Ca^{2+} repeats, although the functional significance of this remains obscure (Rincon *et al.*, 2010). Recent data suggest that *R. flavefaciens* group 1 Docs display tight specificity for ScaA (Coh 1 and 2) and ScaB (Coh 1 to 4) Cohs. However, it remains unknown if the sub-classification of *R. flavefaciens* group 1 Docs has a functional significance. Thus, representative members of all *R. flavefaciens* Doc subgroups were expressed and purified. The capacity of the Docs to bind a range of representative Cohs from *R. flavefaciens* proteome was probed using a previously described cellulose microarray assay method (Haimovitz *et al.*, 2008). The data, presented in Figure 4.11 and Figure S4.1 (Annexes), revealed that all twelve Docs presented a similar binding specificity; all group 1 Docs bind tightly to CohScaA1 and CohScaB2, while not interacting with the other Cohs analyzed, including a Coh from *A. cellulolyticus* used as control. The primary sequences of all 13 Docs were aligned with those of group 3 Docs (Figure 4.11). Initial inspection of the aligned sequences confirms, as described above, that group 1 Docs present a single-binding mode, due to a lack of internal symmetry (Figures 4.10, 4.11). With some exceptions, strong conservation was observed in the most important residues involved in Coh recognition, namely Ile-39, Val-43, Gln-47 in helix-1 and Gln-83 and Leu-87 in helix-3 (*Rf*Doc1a residue numbering). There are, however, a few substitutions at the Ile-39 position, but these are all to non-polar residues such as Val and Met, suggesting functional conservation at this position. Taken together, the data suggest that the subgrouping of *R. flavefaciens* has no functional implications.

Figure 4.11 Coh-binding range and multiple sequence alignment of *R. flavefaciens* group 1 dockerins.

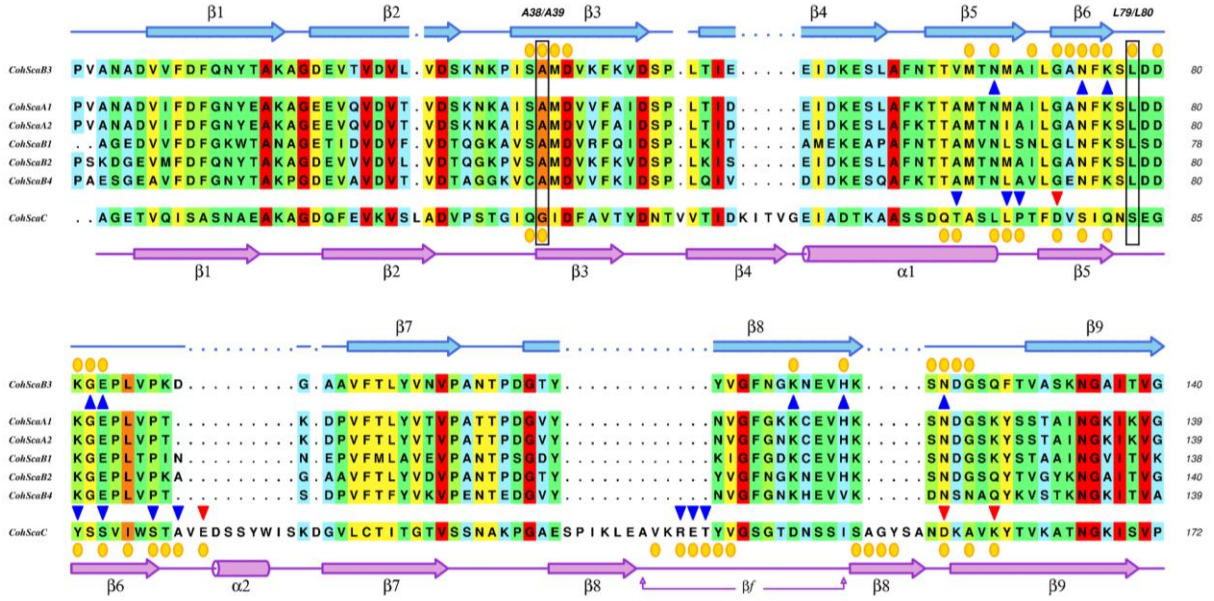


A. Results of Coh-Doc interactions using a cellulose microarray assay with XynDoc/CBM-Coh fusion protein pairs. Each bar graph represents the recognition profile of one dockerin from a different group 1 subgroup and 12 cohesins. The bar values correspond to the ratio between the measured Cy3 and Cy5 signals. Intensity values were calculated by Array Vision Evaluation 8.0 software and all data processing was made in Excel. B. Multiple sequence alignment of *R. flavefaciens* group 1 Docs and group 3 Doc (Doc3). The primary sequence background is colored according to the ALSCRIPT Calcons convention, implemented in ALINE (Bond & Schüttelkopf, 2009): red, identical residues; orange to blue, lowering color-ramped scale of conservation. Above and below the alignment lies a cartoon representation of the secondary structure of Doc1a (blue color) and Doc3 (purple color), respectively (Coh-Doc complexes PDB codes: 5M2O and 5LXV, respectively). Also for these two Docs, the residues involved in molecular interactions with the Coh partner are represented as follows: blue triangle for hydrogen bonds, red triangle for salt bridges and yellow circles for hydrophobic contacts. Critical residues for *Rf*Doc1a/*Rf*Doc1b Coh-binding are marked with a black box, highlighting the #11 and #15 positions.

Recent studies suggest that within the *R. flavefaciens* proteome six Cohs, CohScaA1 and CohScaA2 and CohScaB1-4 (Figure 4.1), are able to bind the 96 group 1 Docs that recruit cellulosomal enzymes to the multi-enzyme complex (Israeli-Ruimy *et al.*, 2017). Residues at *Rf*CohScaB3 and *Rf*CohScaA Cohs which make direct contacts with the Doc domains, as shown in the *Rf*CohScaB3-Doc1a and *Rf*CohScaA-Doc1b structures, are mostly conserved in the four other Cohs of *R. flavefaciens* ScaA and ScaB scaffoldins (Figure 4.12). Changes that might

disturb the Coh-Doc interaction are observed in CohScaB4, with the replacement of the conserved Ser-37 by a Cys (*β*-strand 3) and the highly conserved His-121 (*β*-strand 8) by a Val. The His121Val substitution would remove the hydrogen bond partner for Doc Ser-40. However, this may be compensate by the Gly126Asn change observed in the loop connecting *β*-strands 8 and 9 of CohScaB4, which can form the required hydrogen-bonding partner for Doc Ser-40. Thus, overall conservation in the residues involved in cellulosome assembly suggests that CohScaA1, CohScaA2 and CohScaB1-4 of *R. flavefaciens* will be unable to discriminate between the different group 1 Docs appended to cellulosomal enzymes. In contrast, comparison of the structure of the group 1 Coh-Doc complexes with that of the group 3 *Rf*CohScaC-Doc3 complex explains why the ScaA and ScaB Cohs cannot bind group 3 or 6 Docs, while conversely ScaC Coh is unable to recognize group 1 Docs. Other differences besides the presence of the important loop interrupting *β*-strand 8 in ScaC Coh, include the presence of the bulky hydrophobic side chain (usually Phe) of group 3 and 6 Docs at the critical Ser-40 position of group 1 Docs, which would make steric clashes with group 1 Cohs. Conversely, Ser-40 in group 1 Docs would not make productive interactions with the hydrophobic pocket in the ScaC Coh that is occupied by Phe side-chain in group 3 Docs.

Figure 4.12 Multiple sequence alignment of *R. flavefaciens* ScaA, ScaB and ScaC cohesins.



The primary sequence background is colored according to the ALSCRIPT Calcons convention, implemented in ALINE (Bond & Schüttelkopf, 2009): red, identical residues; orange to blue, lowering color-ramped scale of conservation. Above and below the alignment lies a cartoon representation of the secondary structure of CohScaB3 (blue color) and CohScaC (purple color), respectively, with the *β*-strands numbering. Also for these two Cohs, the residues involved in molecular interactions with the Doc partner (Coh-Doc complexes PDB codes: 5M2O and 5LXV, respectively) are represented as follows: blue triangle for hydrogen bonds, red triangle for salt bridges and yellow circles for

hydrophobic contacts. Critical residues for *Rf*CohScaB3/*Rf*CohScaA Doc-binding are marked with a black box and labelled on the top.

4.4. Conclusions

Previous structure-function studies of the cellulosomes of *C. thermocellum* (Carvalho *et al.*, 2003, 2007) and *C. cellulolyticum* (Pinheiro *et al.*, 2008) revealed that Docs used to recruit the microbial enzymes to these highly intricate multi-enzyme complexes display a dual-binding mode. In addition, recent reports revealed that the attachment of cellulosomes to the *P. cellulosolvens* (Cameron, Weinstein, *et al.*, 2015) and *A. cellulolyticus* cell surface is also mediated by Docs that display a dual-binding mode (Cameron, Najmudin, *et al.*, 2015) (Brás *et al.*, 2016). The structure of dual-binding mode Docs presents a 2-fold internal symmetry that allow binding to the Coh partner in two 180º-related alternate positions. The fact that Docs, in general, possess two different Coh-interacting platforms displaying identical specificities suggests that the dual-binding mode could contribute to enhance the conformational flexibility of the quaternary architecture of the highly populated multi-enzyme complex. This was supported by the observation that non-cellulosomal Docs that recruit single enzymes directly to the cell surface of *C. thermocellum* present a single-binding mode (Brás *et al.*, 2012). In addition, the Coh-Doc interaction used by *C. perfringens* to assemble a two-protein toxin, which is thus also not related to cellulosome assembly, was also shown to display a single-binding mode (Adams, Gregg, Bayer, Boraston, & Smith, 2008). In contrast, a recent analysis of the *R. flavefaciens* cellulosome describes a new system in which this is not observed (Bule *et al.*, 2016). In this bacterium, a large repertoire of hemicellulases is appended to group 3 and 6 Docs, which specifically bind to the Coh of the adaptor scaffoldin ScaC. ScaC contains a group 1 Doc, similar to *Rf*Doc1a and *Rf*Doc1b, which interacts with ScaB and ScaA Cohs. Notably, the structure of a *R. flavefaciens* group 3 Doc, Doc3, in complex with CohScaC, revealed the presence of a single Coh-binding interface that involves both Doc helices (Bule *et al.*, 2016). Here, we extended these studies to establish if Docs displaying a single-binding mode mechanism is a generic feature of enzyme recruitment into the *R. flavefaciens* cellulosome. The data revealed that, similar to previously reported group 3 and 6 Docs, lack of internal symmetry in group 1 *R. flavefaciens* Docs generated an unconventional single protein-binding interface. This property might be widespread among all the 96 group 1 Docs, suggesting that assembly of *R. flavefaciens* cellulosome involves, uniquely, single-binding mode Docs. The data presented in this report questions the widely held hypothesis that the dual-binding mode mechanism provides the conformational flexibility required to degrade plant cell walls in which the topology of these composite structures varies between plants and during the degradative

process. We propose that the dual-binding mode mechanism has evolved to enable rotation of the Docs in cellulosomes with a limited scaffoldin repertoire, a requirement to minimize steric clashes between the enzyme components thus increasing the number of enzyme combinations that can populate these protein complexes. The complexity of the *R. flavefaciens* cellulosome primary and adaptor scaffoldins reduces the steric constraints imposed by enzyme assembly obviating the need for Docs to display a dual-binding mode.

Chapter 5

Ruminococcus flavefaciens Coh-Doc complex involving the dockerin of ScaA

Assembly of primary scaffoldin to the cellulosome of *R. flavefaciens* involves a single binding mode dockerin

Pedro Bule[a], Virgínia Pires[a], Victor D. Alves[a], Ana Luísa Carvalho[b], Luís M.A. Ferreira[a], Steven P. Smith[c], Harry J. Gilbert[d], Edward A. Bayer[e], Shabir Najmudin[a] and Carlos M.G.A. Fontes[a,f,1]

*[a] CIISA – Faculdade de Medicina Veterinária, ULisboa, Pólo Universitário do Alto da Ajuda, Avenida da Universidade Técnica, 1300-477 Lisboa, Portugal. [b] UCIBIO-REQUIMTE, Departamento de Química, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, 2829-516 Caparica, Portugal. [c] Department of Biomedical and Molecular Sciences, Queen's University, Kingston, ON K7L 3N6, Canada. [d] Institute for Cell and Molecular Biosciences, Newcastle University, The Medical School, Newcastle upon Tyne NE2 4HH, United Kingdom. [e] Department of Biomolecular Sciences, The Weizmann Institute of Science, Rehovot 76100 Israel. [f] NZYTech genes & enzymes, Estrada do Paço do Lumiar, 1649-038 Lisboa, Portugal. [1]**Corresponding author***

Adapted from a manuscript in preparation

Abstract

Cellulosomes are highly sophisticated molecular nanomachines that play a major role in the deconstruction of cellulose and hemicellulose in the rumen of mammallian herbivores. The primary mechanism of cellulosome assembly arises from the binding of enzyme-borne Dockerin (Doc) domains to repeated cohesin (Coh) modules located in a non-catalytic primary scaffoldin. In some cases, as exemplified by the cellulosome of the major cellulolytic ruminal bacterium *Ruminococcus flavefaciens*, primary scaffoldins bind to an adaptor scaffoldin that

further interacts with the cell surface providing a mechanism for the amplification of cellulosome complexity. We have elucidated the structure of the Doc of *R. flavefaciens* FD-1 primary scaffoldin ScaA bound to Coh 5 of adaptor scaffoldin ScaB. The *Rf*CohScaB5-DocScaA complex has an elliptical architecture similar to others previously described in a variety of ecological niches. ScaA Doc presents a single binding mode which is similar to the ones described for the other two specificities that contribute to cellulosome assembly in *R. flavefaciens*. This contrasts with the majority of cellulosomes described to date where Docs generally present two similar Coh-binding interfaces supporting a dual-binding mode. Thus, *R. flavefaciens* cellulosome is assembled through an original mechanism involving single, but not dual-binding mode Docs. Whether these single-binding mode Coh-Doc interactions observed in ruminal cellulosomes represent an adaptation to the singular properties revealed by the rumen of mammals remains to be elucidated.

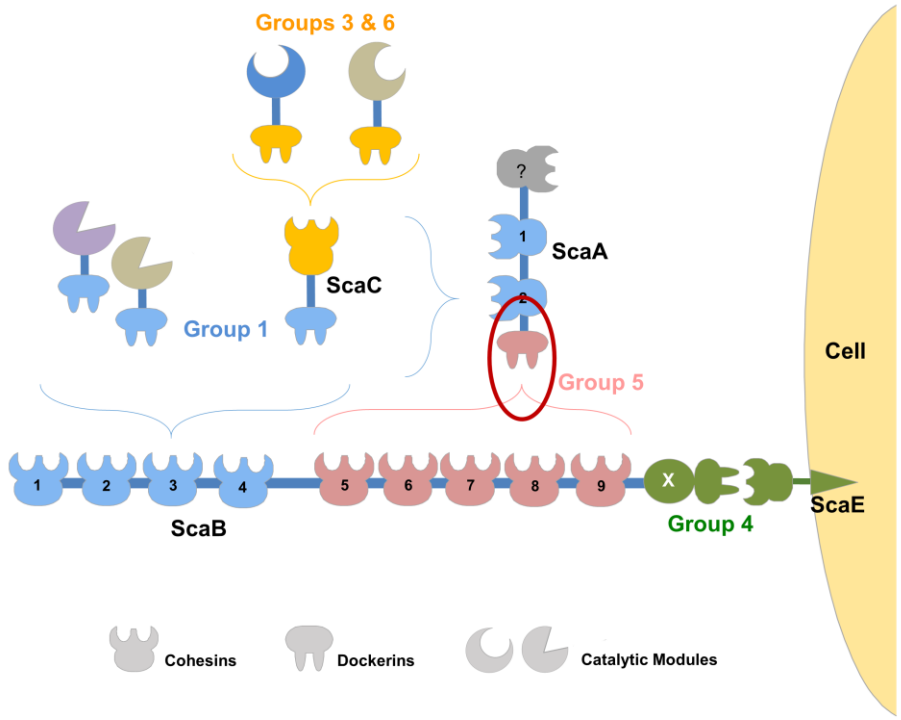
5.1. Introduction

The cellulosome is a highly intricate molecular nanomachine produced by anaerobic microorganisms to efficiently deconstruct complex plant cell wall polyssacharides, such as cellulose and hemicellulose. It consists of a multi-protein complex with several independent enzymatic components arranged around a molecular scaffold, termed scaffoldin. Cellulosomes combine an extensive repertoire of enzymes, including glycoside hydrolases, pectate lyases and carbohydrate esterases. Integration of these enzymes into the cellulosome is believed to enhance the synergistic interactions between enzymes with complementary activities while promoting enzyme stability (Bayer *et al.*, 2004; Fontes & Gilbert, 2010). This process allows anaerobic microorganisms to gain a critical advantage when extracting energy in highly competitive ecological niches and is critical to the recycling of carbon between microbes, herbivores and plants. Furthermore, cellulases and hemicellulases have captured the attention of several biotechnology-based industries due to their potential application for the bio-conversion of plant biomass into renewable fuels and the development of molecules with biomedical application (Bayer *et al.*, 2007, 1994; Demain *et al.*, 2005). Protein:protein interactions established between dockerin (Doc) modules, located in the cellulosomal enzymes, and cohesin (Coh) domains of the molecular scaffoldin are the primary force for cellulosomal assembly.

Previously, extensive structural and biochemical characterization of type I and type II cohesin-dockerin (Coh-Doc) complexes revealed the molecular determinants of cellulosome assembly in *Clostridium thermocellum*, *Clostridium cellulolyticus, Pseudobacteroides cellulosolvens and Acetivibrio cellulolyticus*, species that colonize different ecological niches(Cameron,

Weinstein, *et al.*, 2015; Carvalho *et al.*, 2003; Noach *et al.*, 2005; Pinheiro *et al.*, 2009). In general, Coh-Doc complexes involved in cellulosome assembly are of type I, while type II Coh-Doc interactions recruit cellulosomes into the cell surface (Gefen *et al.*, 2012; Moraïs *et al.*, 2012; Stahl *et al.*, 2012). In contrast, the relevance of cellulosomes to fiber digestion in the rumen remains mostly unexplored. The rumen, which essentially constitutes a large fermentation chamber in the gastrointestinal tract of ruminant mammals, is a highly competitive ecological niche colonized by symbiotic microbes that have specialized in the hydrolysis of recalcitrant carbohydrates. So far, *Ruminococcus flavefaciens*, a Gram-positive anaerobic bacterium of the Firmicutes phylum, is the only species in this microbial ecosystem that has been shown to possess a definitive cellulosome (Ding *et al.*, 2001). Intriguingly, the rumen houses numerous subspecies of this bacterium, all with a similar set of scaffoldins but each with its own array of dockerin-bearing proteins (enzymes) and cellulosome architecture (Dassa *et al.*, 2014; Jindou *et al.*, 2008). The genome sequence of *R. flavefaciens* strain FD-1 revealed the presence of 223 dockerin-containing proteins (154 of which were identified as carbohydrate-active enzymes) (Dassa *et al.*, 2014), revealing the most complex cellulosome described to date (Berg Miller *et al.*, 2009) (Figure 5.1). *R. flavefaciens* Docs have been organized into six groups based on primary structure homology (Rincon *et al.*, 2010). This classification was recently found to be functionally relevant (Israeli-Ruimy *et al.*, 2017), with different Doc groups displaying different binding specificities. Thus, the 96 group 1 Docs of *R. flavefaciens* FD-1 bind to the two cohesins of ScaA and cohesins 1 to 4 of ScaB. Hemicellulases, contain group 3 or 6 Docs that specifically bind to adaptor scaffoldin ScaC, whose group 1 Doc locks onto the Cohs of ScaA or Cohs 1-4 of ScaB (Bule *et al.*, 2016; Rincon *et al.*, 2004). The cellulosome is tethered to the surface of *R. flavefaciens* through the binding of group 4 Doc of ScaB to the Coh of cell surface protein ScaE. A variety of other proteins were found to contain Docs that specifically interact with cell surface Cohs rather than to the cellulosomal Cohs. These Docs were classified into groups 4 and 2. Group 2 Docs are functional truncated derivatives of group 4 Docs. Finally, ScaA Doc, which is the sole member of group 5, binds exclusively to ScaB Cohs 5-9. This interaction has a central role in cellulosomal assembly as it allows the binding of up to 5 ScaA primary scaffoldins to ScaB and, as such, up to 14 enzymes to a single cellulosome (Figure 5.1).

Figure 5.1 Cellulosome of *R. flavefaciens* strain FD-1 displaying the different group-specific Coh-Doc interactions involved in the multi-enzyme complex assembling.



The scheme is color-coded to highlight the four subgroups of cohesin-dockerin specificities: Dockerins and cognate cohesin counterparts of the different groups are marked in blue (Group 1 dockerins), yellow (Groups 3 and 6), green (Groups 2 and 4) and red (Group 5), respectively. Group 2 dockerins are truncated derivatives of group 4 and are not represented in the figure for simplification. The red oval marks the complex between DocScaA and CohScaB, which structure is reported this Chapter.

Initial studies of *R. flavefaciens* Coh and Doc modules suggested that these sequences diverge from the previously described type I and type II modules and were, therefore, collectively classified as of type III (Rincon *et al.*, 2003, 2004, 2005, 2007). Until recently, only a single crystal structure of a type III Coh-Doc complex had been reported, comprising the X-module associated group 4 Doc of ScaB bound to ScaE Coh. This structural divergence from previously described type I and II Coh-Doc complexes is pronounced, especially on the dockerin side. Thus, ScaB's XDoc has 5 α-helices instead of the traditional 3 and three inserts that act as structural buttresses that reinforce the stalk like conformation of the X module (Salama-Alber *et al.*, 2013). In the previous Chapters of this thesis three more structures of *R. flavefaciens* Coh-Doc complexes have been described: ScaC Coh bound to a group 3 Doc, a ScaA Coh bound to a group 1b Doc and a ScaB Coh bound to a group 1a Doc. While the first of the three is very similar to the previously described type I complexes, Coh-Doc complexes involving *R. flavefaciens* group 1 Docs do not bear much homology with any other complexes described to date. Although these three complexes are responsible for the integration of enzymes into the

primary scaffoldins, either directly or through an adaptor scaffoldin, none possess a dual binding mode as observed in other cellulosomes (Bule *et al.*, 2016, 2017).

Here, we report the crystal structure of the *R. flavefaciens* strain FD-1 Coh-Doc complex established between ScaA Doc and the fifth cohesin of ScaB (*Rf*CohScaB-DocScaA). ScaADoc exhibits an atypical Ca^{2+} binding site due to several sequence alterations and a 12 residue insert in the midst of the Ca^{2+} coordination loop. A comprehensive biochemical analysis of CohScaB-DocScaA interaction informed by the structural data suggests a non-dynamic single-binding mode. Thus, in contrast to the other known cellulosomes, this work supports the view that in *R. flavefaciens* cellulosome protein assembly is the result of exclusively non-dynamic Coh-Doc interactions.

5.2. Experimental procedures

5.2.1. Gene synthesis and DNA cloning

Docs are inherently uns5 when produced in *Escherichia coli*. To promote stability, *R. flavefaciens* FD-1 DocScaA (WP_009986657.1 residues 648-730) was co-expressed *in vivo* with CohScaB5 (WP_009986658.1 residues 737-880). The immediate binding of DocScaA to CohScaB5 is believed to confer immediate stabilization of the Doc structure. The genes encoding the two proteins were designed with a codon usage optimized to maximize expression in *E. coli*, synthesized *in vitro* (NZYTech Ltd, Lisbon, Portugal) and cloned into pET28a (Merck Millipore, Germany) under the control of separate T7 promoters. The DocScaA-encoding gene was positioned at the 5' end and the CohScaB5-encoding gene at the 3' end in the synthetic DNA. A T7 terminator sequence (to terminate transcription of the dockerin gene) and a T7 promoter sequence (to control transcription of the cohesin gene) were incorporated between the sequences of the two genes. This construct contained *Nhe*I and *Nco*I recognition sites at the 5' end and *Xho*I and *Sal*I at the 3' end specifically tailored to allow subcloning into pET-28a (Merck Millipore, Germany), such that the sequence encoding a six-residue His tag could be introduced either at the N-terminus of the dockerin (through digestion with *Nhe*I and *Sal*I, incorporating the additional sequence MGSSHHHHHHSSGLVPRGSHMAS N-terminal of the Doc) or at the C-terminus of the cohesin (by cutting with *Nco*I and *Xho*I, which incorporates the additional sequence LEHHHHHH C-terminal of the Coh). Thus, as a result of this strategy two pET28a plasmid derivatives were produced: one leading to the expression of dockerin with an engineered hexa-histidine tag and a second derivative where the engineered tag is attached to the cohesin. The two separate plasmids were used to express *Rf*CohScaB5-DocScaA

complexes in *E. coli*. Recombinant DocScaA and CohScaB5 primary sequences are presented in Table 5.1.

Table 5.1 Recombinant protein sequences of *Rf*CohScaB5, *Rf*DocScaA and mutant variants of the latter produced for the interaction studies.

Protein	Sequence
*Rf*DocScaA WT	MSDKIIHLTDDSFDTDVLKADGAILVDFWAEWCGPCKMIAPILDEIADEYQGKLTVAKLNIDQNPGTAPKYGIRGIPTLLLFKNGEVAATKV GALSKGQLKEFLDANLAGSGSGHMHHHHHHSSMTSLYKKAGFGNTLKPVWGDVNCDGDVNVADVVLLNKWLNNNADYAMTDQGKV NADCFNPQDANGGAVDASKVDLTKADSDAIIKSVVHLITLPAKG
*Rf*DocScaA N661A	AGFGNTLKPVWGDVNCDGDV**A**VADVVLLNKWLNNNADYAMTDQGKVNADCFNPQDANGGAVDASKVDLTKADSDAIIKSVVHLIT…
*Rf*DocScaA V662A	AGFGNTLKPVWGDVNCDGDVN**A**ADVVLLNKWLNNNADYAMTDQGKVNADCFNPQDANGGAVDASKVDLTKADSDAIIKSVVHLIT…
*Rf*DocScaA V666A	AGFGNTLKPVWGDVNCDGDVNVADV**A**LLNKWLNNNADYAMTDQGKVNADCFNPQDANGGAVDASKVDLTKADSDAIIKSVVHLIT…
*Rf*DocScaA N669A	AGFGNTLKPVWGDVNCDGDVNVADVVLL**A**KWLNNNADYAMTDQGKVNADCFNPQDANGGAVDASKVDLTKADSDAIIKSVVHLIT…
*Rf*DocScaA K670A	AGFGNTLKPVWGDVNCDGDVNVADVVLLN**A**WLNNNADYAMTDQGKVNADCFNPQDANGGAVDASKVDLTKADSDAIIKSVVHLIT…
*Rf*DocScaA V721A	AGFGNTLKPVWGDVNCDGDVNVADVVLLNKWLNNNADYAMTDQGKVNADCFNPQDANGGAVDASKVDLTKADSDAIIKSV**A**HLIT…
*Rf*DocScaA H722A	AGFGNTLKPVWGDVNCDGDVNVADVVLLNKWLNNNADYAMTDQGKVNADCFNPQDANGGAVDASKVDLTKADSDAIIKSVV**A**LIT…
*Rf*DocScaA N661A + N669A	AGFGNTLKPVWGDVNCDGDV**A**VADVVLL**A**KWLNNNADYAMTDQGKVNADCFNPQDANGGAVDASKVDLTKADSDAIIKSVVHLIT…
*Rf*DocScaA V662A + V666A	AGFGNTLKPVWGDVNCDGDVN**A**ADV**A**LLNKWLNNNADYAMTDQGKVNADCFNPQDANGGAVDASKVDLTKADSDAIIKSVVHLIT…
*Rf*DocScaA V662A + V721A	AGFGNTLKPVWGDVNCDGDVN**A**ADVVLLNKWLNNNADYAMTDQGKVNADCFNPQDANGGAVDASKVDLTKADSDAIIKSV**A**HLIT…
*Rf*DocScaA V666A + V721A	AGFGNTLKPVWGDVNCDGDVNVADV**A**LLNKWLNNNADYAMTDQGKVNADCFNPQDANGGAVDASKVDLTKADSDAIIKSV**A**HLIT…
*Rf*DocScaA V662A + V666A + V721A	AGFGNTLKPVWGDVNCDGDVN**A**ADV**A**LLNKWLNNNADYAMTDQGKVNADCFNPQDANGGAVDASKVDLTKADSDAIIKSV**A**HLIT…
*Rf*CohScaB5 WT	MGSSHHHHHHSSGLVPRGSHMASKNVTPATGSAEWVIPTVNAKPGEKVTMDVVVKNSAIEVAGAQFNIKQTAPIAYGSAASGDAYAAIV PNETEQYYAFGEGIGKGIKAADGAKIITLTFNVPADCAKGTYPVKWSNAFITDTNGNKITDKITLTDGAIVVGDTPVV

The mutated residues are highlighted in black. The underline sequences correspond to the Dockerin's TrxA-His6x and the Cohesin's His6x tags.

To produce the recombinant cohesins and dockerins individually, two distinct cloning methods were used. Digesting the previously described cohesin-tagged version of the pET28 derivatives with BglII allowed removal of the dockerin sequence. Plasmid integrity was reconstituted by re-ligation. This strategy gave a pET28a derivative encoding the recombinant cohesin CohScaB5 fused to a C-terminal hexa-histidine tag. The DocScaA-encoding gene was cloned into the pHTP2 vector (NZYtech, Lisbon, Portugal) using NZYEasy Cloning & Expression System (NZYtech, Lisbon, Portugal), following the manufacturer's protocol. Dockerin genes were isolated by PCR using *R. flavefaciens* FD-1 genomic DNA as a template and the primers shown in Table S5.1 (Annexes). The recombinant dockerin encoded by the pHTP2 derivatives contained an N-terminal thioredoxin A and an internal hexa-histidine tag for increased protein stability and solubility. Sequences of all plasmids produced were verified by Sanger sequencing. To identify the Doc residues that modulate Coh recognition, several TrxADocScaA protein derivatives were produced using site directed mutagenesis. PCR amplification of the Doc containing plasmid using the primers presented in Table S5.1 (Annexes), allowed the

production of seven DocScaA protein derivatives, namely N661A, V662A, V666A, N669A, K670A, V721A, H722A. Each of the newly generated gene sequence was fully sequenced to confirm that only the desired mutation accumulated in the nucleic acid.

5.2.2. Expression and purification of recombinant proteins

Preliminary expression screens revealed that when the polyhistidine tag was located at the Doc N-terminal end in *Rf*CohScaB5-DocScaA complexes, the expression levels of both Coh and Doc were higher. Tagging the cohesin resulted in the accumulation of large levels of unbound cohesin in the purification product suggesting that cohesin was expressed at higher levels than dockerins or that untagged dockerin was less stable. Therefore, the pET28a derivative encoding the protein complex with the tagged dockerin was subsequently selected to produce the *Rf*CohScaB5-DocScaA protein complex in large quantities. Recombinant BL21 (DE3) *E. coli* were grown at 37ºC to an OD_{600} of 0.5. Recombinant protein expression was induced by the addition of 1 mM isopropyl *β*-D-1-thiogalactopyranoside followed by incubation at 19ºC for 16 hours. Cells were harvested by 15 min centrifugation at 5000 x *g* and resuspendend in 20 mL of immobilized-metal affinity chromatography (IMAC) binding buffer (50 mM HEPES, pH 7.5, 10 mM imidazole, 1 M NaCl, 5 mM $CaCl_2$). Cells were then disrupted by sonication and the cell-free supernatant recovered by 30 min centrifugation at 15,000 x *g*. After loading the soluble fraction into a HisTrap^{TM} nickel-charged Sepharose column (GE Healthcare, UK), initial purification was carried out by IMAC in a FPLC system (GE Healthcare, UK) using conventional protocols with a 35 mM imidazole wash and a 35-300 mM imidazole elution gradient. Fractions containing the purified cohesin–dockerin complex were buffer exchanged into 50 mM HEPES, pH 7.5, containing 200 mM NaCl, 5 mM $CaCl_2$ using a PD-10 Sephadex G-25M gel-filtration column (Amersham Pharmacia Biosciences, UK). A further purification step by gel-filtration chromatography was performed by loading the Coh-Doc complexes onto a HiLoad 16/60 Superdex 75 (GE Healthcare, UK) at a flow rate of 1 ml min^{-1}. Fractions containing the purified complex were then concentrated with Amicon Ultra-15 centrifugal devices with a 10-kDa cutoff membrane (Millipore, USA) and washed three times with molecular biology grade water (Sigma) containing 0.5 mM $CaCl_2$. The protein concentration was estimated in a NanoDrop 2000c spectrophotometer (Thermo Scientific, USA) using a molar extinction coefficient (ε) of 31 065 M^{-1} cm^{-1}. Final protein concentration was adjusted to 45 $mg.mL^{-1}$. The protein complex was stored in molecular biology grade water containing 0.5 mM $CaCl_2$. The purity and molecular mass of the recombinant complexes were confirmed by 14 % (w/v) SDS–PAGE.

The TrxADocScaA mutant derivatives and CohScaB5 used in native PAGE and ITC experiments were expressed as described before and purified with IMAC using nickel-charged Sepharose His GraviTrap gravity-flow columns (GE Healthcare, UK). For the ITC experiments, the recombinant cohesin and dockerins were buffer exchanged to 50 mM HEPES pH 7.5, 0.5 mM $CaCl_2$ and 0.5 mM TCEP using PD-10 Sephadex G-25M gel filtration columns (GE Healthcare, UK).

5.2.3. Nondenaturing gel electrophoresis (NGE)

For the NGE experiments, the proteins were kept in the IMAC elution buffer (50 mM HEPES, pH 7.5, 300 mM imidazole, 1 M NaCl, 5 mM $CaCl_2$). Each of the TrxADocScaA variants, at a concentration of 30 μM, was incubated in the presence and absence of 30 μM CohScaB5 for 30 min at room temperature and separated on a 10 % native polyacrilamide gel. Electrophoresis was carried out at room temperature. The gels were stained with Comassie Blue. Complex formation was detected by the presence of an additional band displaying a lower electrophoretic mobility than the individual modules.

5.2.4. Isothermal titration calorimetry

All ITC experiments were carried out at 308.14 K. The purified TrxADocScaA variants and CohScaB5 were diluted to the required concentrations and filtered using a 0.45 μm syringe filter (PALL). During titrations, the Doc constructs were stirred at 307 revolutions/min in the reaction cell and titrated with 28 successive 10 μL injections of CohScaC at 220 s intervals. Integrated heat effects, after correction for heats of dilution, were analyzed by nonlinear regression using a single-site model (Microcal ORIGIN version 7.0, Microcal Software, USA). The fitted data yielded the association constant (K_A) and the enthalpy of binding (ΔH). Other thermodynamic parameters were calculated using the standard thermodynamic equation: $\Delta RTlnK_A=\Delta G=\Delta H-T\Delta S$.

5.2.5. X-ray crystallography, structural determination and refinement

Optimal crystallization conditions were obtained by using the sitting-drop vapor-diffusion method with an Oryx8 robotic nanodrop dispensing system (Douglas Instruments, UK; (Bule, Correia, *et al.*, 2014). The commercial kits Crystal Screen, Crystal Screen 2, PEG/Ion and PEG/Ion 2 (Hampton Research, California, USA), JCSG+ HT96 (Molecular Dimensions,UK) and an in-house screen (80 factorial) were used for the screening. 0.6 μl drops of 12.5, 25 and 45 mg ml^{-1} *Rf*CohScaB5-DocScaA were mixed with 0.6 μl reservoir solution at room

temperature per well containing 50 µl of the crystallization solution. The resulting plates were then stored at 292 K. Crystal formation was observed in 2 conditions (0.1 M HEPES pH 7.5, 1.2 M sodium citrate; 2.1 M DL-malic acid pH 7.0) after a period of approximately 180 days from setting up the plates (maximum dimensions ~50 x 50 x 20 µm). These crystals were cryoprotected with mother solution containing 20–30 % glycerol and flash-cooled in liquid nitrogen. Preliminary X-ray diffraction experiments revealed that these crystals were of very poor quality mainly due to high mosaicity. Optimization plates based on the 2 original hits were set up. Two additive plates (one for each original condition) were also set up using the HT Additive Screen (Hampton Research, California, USA). The additive screen drops consisted of 0.8µl protein + 0.8µl optimization condition + 0.2µl stock additive solution. This approach generated several good quality crystals. X-ray diffraction data were collected on beamline PROXIMA-1 at the Soleil Synchrotron, Saint-Aubin, France using a PILATUS 6M detector (Dectris Ltd) from crystals cooled to 100 K with a Cryostream (Oxford Cryosystems Ltd). A systematic grid search was carried out on all of these crystals to select the best diffracting part of each crystal. EDNA (Winter & McAuley, 2011) and iMosflm (Battye *et al.*, 2011) were used for strategy calculation during data collection. All data sets were processed using the Fast_dp and xia2 (Winter, 2010) packages, which use the programs XDS (Kabsch, 2010), POINTLESS and SCALA (Evans, 2006) from the CCP4 suite (Winn *et al.*, 2011). Data-collection statistics are given in Table 5.2.

The best diffracting crystal was formed in one of the additive screen conditions (0.1 M HEPS pH 7.5, 1.2 M Sodium Citrate, 4% v/v acetonitrile). It diffracted to a resolution of 1.4 Å and belonged to the monoclinic spacegroup $P2_1$. Phaser MR was used to carry out molecular replacement (McCoy *et al.*, 2007). The best solution was found using a cohesin from *R. flavefaciens* strain 17 ScaB (unreleased) and an ensemble of 3 *R. flavefaciens* FD-1 dockerins (Doc1a from 5M2O, Doc1b from 5M2S and Doc3 from 5LXV) made with Dali (Holm & Rosenstrom, 2010). The cohesins had a sequence identity of 33.0 % and the dockerins between 22% (Doc3) and 34% (Doc1b). Two copies of the heterodimer *Rf*CohScaB5-DocScaA complex were present in the asymmetric unit. The partially obtained model was completed with Buccaneer (ref) and with manual modeling in COOT. It was then refined using REFMAC5 (Murshudov *et al.*, 2011) and PDB REDO (Joosten *et al.*, 2014) interspersed with model adjustment in COOT. The final round of refinement was performed using the TLS/restrained refinement procedure using each module as a single group giving the final model (Protein Data Bank code 5N5P, Table 5.2). The root mean square deviation of bond lengths, bond angles, torsion angles and other indicators were continuously monitored using validation tools in COOT and MOLPROBITY. A summary of the refinement statistics is provided in Table 5.2.

Table 5.2 X-ray crystallography data collection and refinement statistics for *Rf*CohScaB5-DocScaA.

Data collection	
Beamline	*PROXIMA-1, Soleil*
Space Group	*$P12_11$*
Wavelength (Å)	*0.82*
Unit-cell parameters	
a, b c (Å)	*30.09, 142.90, 46.59*
α, β , γ (°)	*90, 90.75, 90*
Vm♯ ($Å^3$ Da^{-1})	*1.89*
Solvent Content (%)	*35.10*
Resolution limits (Å)	*46.58 – 1.98 (2.072 – 1.98)*
No. of observations	*182195 (13279)*
No. of unique observations	*26476 (2602)*
Multiplicity	*6.9 (6.9)*
Completeness (%)	*99.9 (99.7)*
$<I/\sigma(I)>$	*8.85 (3.96)*
$CC_{1/2}$†	*0.995 (0.974)*
Wilson B-factor	*22.71*
Rmerge ‡	*0.098 (0.294)*
Structure refinement	
R-work § , R-free ¥	*0.1819, 0.2142*
No. of Non-H atoms	*3519*
Macromolecules	*3287*
Ligands	*7*
Water	*225*
Protein residues	*449*
RMS(bonds)	*0.010*
RMS(angles)	*1.4*
Ramachandran favored (%)	*96*
Ramachandran outliers (%)	*0*
Clash score	*0.61*
Average B-factor	*33.80*
macromolecules	*33.80*
ligands	*39.30*
solvent	*33.50*
PDB accession code	*5N5P*

Values in parenthesis are for the highest resolution shell. # Matthews coefficient (Matthews, 1968). † $CC_{1/2}$ = the correlation between intensities from random half-dataset (Diederichs & Karplus, 2013) ‡ $R_{merge} = \Sigma_{hkl} \Sigma_i |I_i(hkl) - <I(hkl)>| / \Sigma_{hkl} \Sigma_i I_i(hkl)$, where $I_i(hkl)$ is the *i*th intensity measurement of reflection *hkl*, including symmetry-related reflections and $<I(hkl)>$ is its average. §$R_{work} = \Sigma_{hkl}||F_{obs}|-|F_{calc}||/\Sigma_{hkl}|F_{obs}|$. ¥$R_{free}$ as R_{work}, but summed over a 5% test set of reflections.

5.3. Results and Discussion

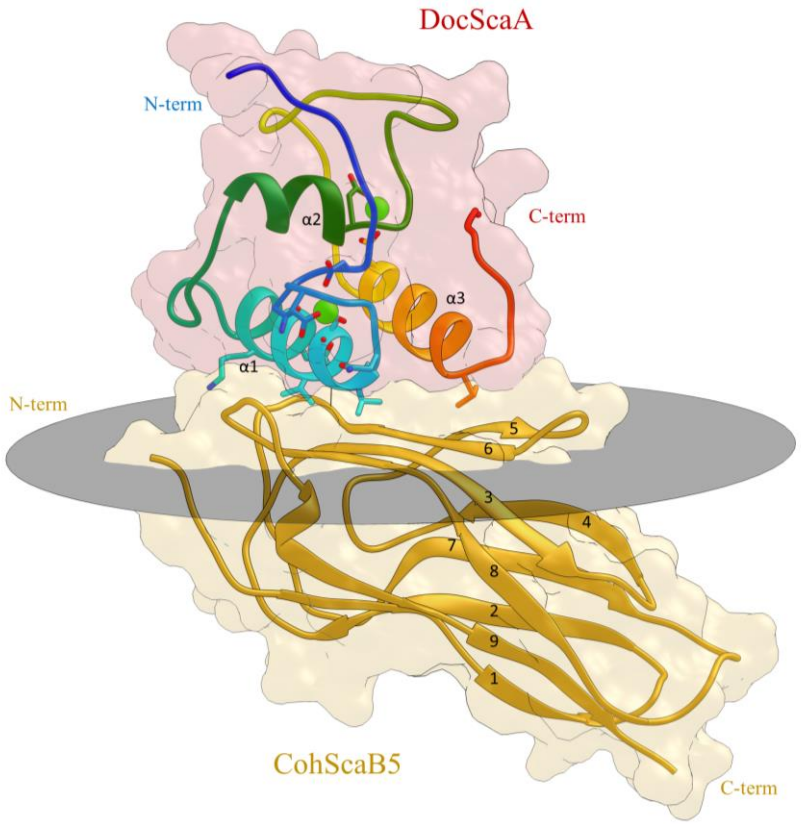
Previous studies have shown that the dockerin of *R. flavefaciens* primary scaffoldin ScaA (*Rf*DocScaA) interacts exclusively with Cohs 5 to 9 of scaffoldin ScaB (Israeli-Ruimy *et al.*, 2017; Rincon *et al.*, 2003). Likewise, ScaB Cohs 5 – 9 specifically recognize *Rf*DocScaA. Thus, CohScaB-DocScaA interaction is highly specific and central for *R. flavefaciens* cellulosome organization. Out of the 5 possible *Rf*DocScaA-CohScaB complexes, the one involving the fifth ScaB cohesin (*Rf*CohScaB5) with *Rf*DocScaA displayed the highest levels of expression (Israeli-Ruimy *et al.*, 2017). To gain insight into the molecular mechanisms of cellulosome assembly, the X-ray crystal structures of *R. flavefaciens* DocScaA in complex with the fifth cohesin from ScaB (CohScaB5), *Rf*CohScaB5-DocScaA, was determined. Established co-

expression strategies for the production and purification of Coh-Doc complexes generated sufficient quantity of highly pure protein complexes to obtain good quality crystals.

5.3.1. Structure of a novel *R. flavefaciens* Coh-Doc complex

*Rf*CohScaB5-DocScaA crystal structure was solved by molecular replacement, as described in the experimental procedures section (Figure 5.2). The best crystals belonged to space group $P2_1$ with unit cell dimensions of a = 30.1 Å, b = 142.9 Å and c = 46.6 Å. *Rf*CohScaB5-DocScaA complex displayed an elongated comma shape with overall dimensions of 60 x 50 x 25 Å and includes residues 740 – 877 from *Rf*CohScaB5 and 548 – 730 from *Rf*DocScaA. The structure included 2 molecules of the heterodimer in the asymmetric unit, with each Doc coordinating 2 calcium ions, as well as 1 acetonitrile and 225 water molecules. The dimer resulted from interactions between two *Rf*CohScaB5 modules (chains A and C). Thus, chain A CohScaB5 O of Thr-743, Oγ of Ser-745 and Oδ1 of Asn-769 interact via hydrogen bonds with chain B CohScaB5 Oγ1 of Thr-752, N of Asp-869 and N of Leu-867, respectively. Thirty non-bonded contacts also contribute to the dimerization (not shown). The biological relevance of these crystallographic interactions, if any, is presently unclear. Data collection and structure refinement statistics are shown in Table 5.2.

Figure 5.2 Structure of *Rf*CohScaB5-DocScaA complex.



Structure of *Rf*CohScaB5-DocScaA complex with the dockerin in dark red and the cohesin in gold. The molecular surface of each module is represented in transparent colors. Under the transparent molecular surface and above the grey oval disk that marks the plane defined by the Coh 8-3-6-5 β-sheets, a ribbon representation shows the three Doc α-helices labeled α1, α2 and α3. Below the grey oval disk a ribbon representation of the cohesin shows each of the 9 β-strands, labeled from 1 to 9. Ca^{2+} ions are depicted as purple spheres.

5.3.2. Structure of ScaB Coh5

*Rf*CohScaB5 displays an overall typical elliptical structure with nine β-strands, which form two β-sheets aligned in an elongated β-barrel that displays a classical "jelly-roll fold". The two sheets comprise β-strands 9, 1, 2, 7, 4 on one face and β-strands 8, 3, 6, 5 on the opposite face. Strands 1 and 9 align parallel to each other, thus completing the jelly-roll, while the other β-strands are antiparallel (Figure 5.2). Interestingly, except for a very poorly defined 3_{10}-helix formed by residues Thr862 to Lys864, there are no structural motifs other than β-strands (Figure 5.2). This contrasts with several bacterial cohesins where β-flaps are commonly found interrupting β-strand 8 or 4, like in *Acetivibrio cellulolyticus* (PDB code 4UYP), *Pseudobacteroides cellulosolvens* (PDB code 1TYJ) or *R. flavefaciens* ScaC Coh (PDB code 5LXV) (Bule *et al.*, 2016; Cameron, Najmudin, *et al.*, 2015; Noach *et al.*, 2005). The distinct α-helix commonly found between β-strands 4 and 5 in other cohesins is also absent. This particularity is shared with the recently described structures of *Rf*CohScaB3 (PDB code 5AOZ) and *Rf*CohScaA2 (PDB code 5M2S) which are, according to a structural similarity search using the PDBeFold server (http://www.ebi.ac.uk/msd-srv/ssm/), the closest functionally relevant *Rf*CohScaB5 structural homologs (with a Z-score of 8.1, r.m.s.d of 1.78 Å and sequence identity of 27% over 127 aligned residues and Z-score of 7.9, r.m.s.d of 1.76 Å and sequence identity of 23% over 127 aligned residues, respectively; (Bule *et al.*, 2017). Other structural homologs include the type I *Acetivibrio cellulolyticus* CohScaC3 (PDB code 4UYP) with a Z-score of 8.0, r.m.s.d of 1.81 Å and 14% sequence identity over 125 aligned residues and the type I *Pseudobacteroides cellulosolvens* CohScaB7 (PDB code 4UMS), with a Z-score of 9.0, r.m.s.d of 1.87 Å and sequence identity of 20% over 129 aligned residues.

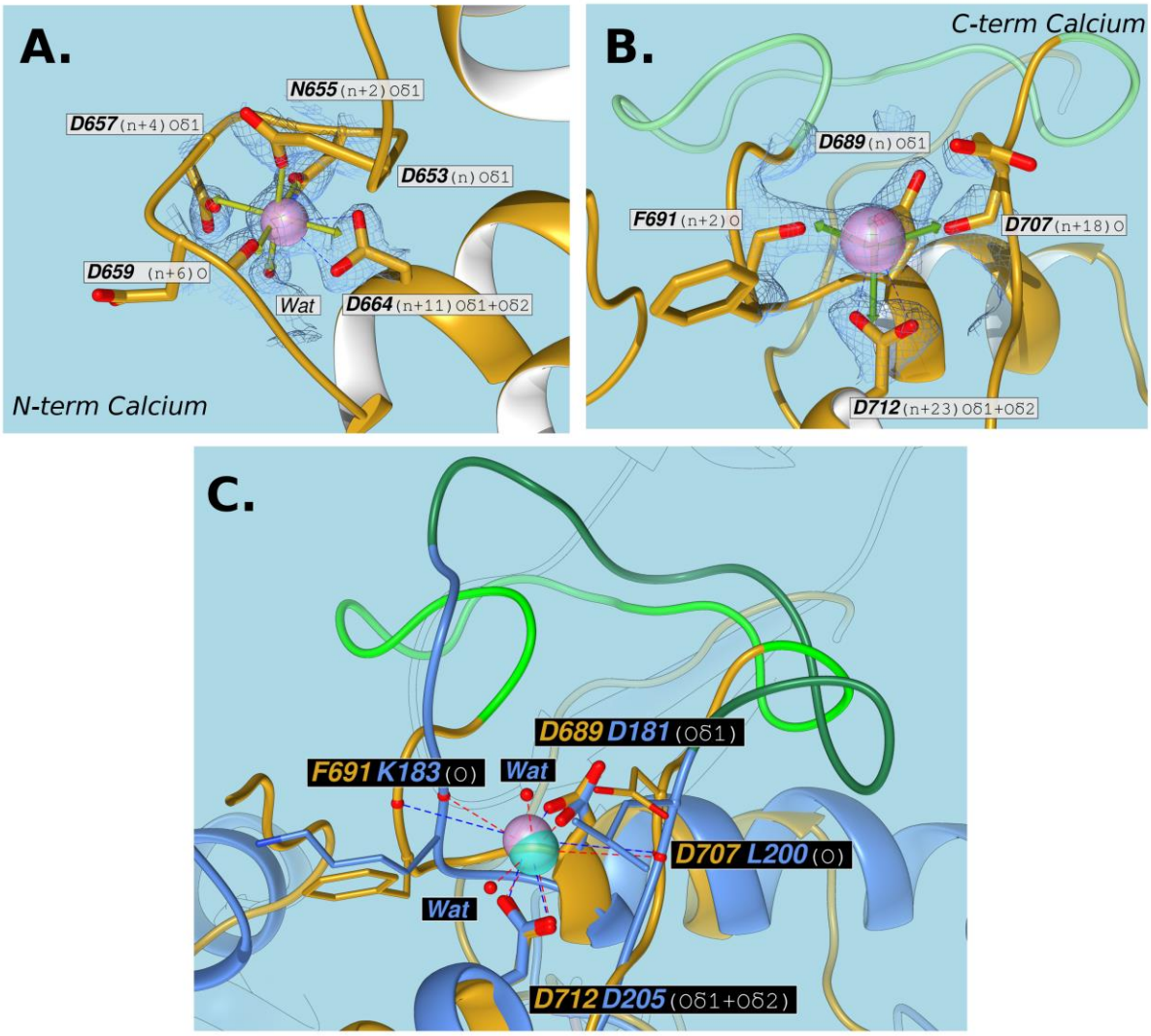
5.3.3. Structure of ScaA Doc

*Rf*DocScaA possesses a total of three α-helices. Two α-helices are arranged in an antiparallel orientation forming a planar surface on one face of the Doc, which interacts with CohScaB5 (Figure 5.2). One helix extends from residues Val-662 to Asp-677 (helix 1) and the other from Lys-710 to Leu-723 (helix 3). These two helices comprise portions of the two classic dockerin

repeating segments, each containing a bound calcium ion in loops located at opposite ends of the module. However, much like in *R. flavefaciens* ScaB XDoc (Salama-Alber *et al.*, 2013) the second repeating segment consists of an atypical variation of the EF-hand motif due to a large insertion in the calcium binding loop (Figure 5.2). This is the most defining characteristic of this module and will be further discussed below. Connecting these two structural elements is yet another α-helix (helix 2) extending from Asp-682 to Asp-689. The overall tertiary structure, with the exception of the loop insertion, bears some similarities to enzyme associated dockerins from *C. thermocellum* (PDB code 3P0D: Z-score of 6.9, r.m.s.d of 1.24 Å and 25% sequence identity over 65 aligned residues; PDB code 2CCL: Z-score of 6.5, r.m.s.d of 1.44 Å and 26% sequence identity over 61 aligned residues) and *R. flavefaciens* (PDB code 5M2O: Z-score of 7.6, r.m.s.d of 1.31 Å and 27% sequence identity over 68 aligned residues). The Ca^{2+} coordination in the N-terminal loop follows the typical *n, n+2, n+4, n+6, n+11* plus a water molecule (at the *n+8* position) pattern. Thus, the Ca^{2+} ion located at the N-terminus is coordinated by the side chains of Asp-653, Asn-655, Asp-657 and Asp-664 (both the Oδ1 and Oδ2), the latter belonging to α-helix 1 (Figure 5.3A,) The octahedral geometry of the coordination is completed by the main chain carbonyl of Asp-659 and one water molecule (*n+8, via* Asn-661) (Figure 5.3A). Contrastingly, the pattern of Ca^{2+} coordination in the C-terminal repeat is displaced due to the 12 residue long loop insertion between Pro-693 and Ser-704 (Figure 5.3 B). A phenylalanine residue replaces the usual Asn/Asp at position *n+2* and provides a backbone carbonyl oxygen ligand. The Asn/Asp at position *n+4* and water at position *n+8* are absent (Figure 5.3 B). Therefore, the coordination follows an atypical *n, n+2, n+18 (*at the *n+6* position*), n+23 (*at the *n+11* position*),* pattern with no water molecules involved. This means that, instead of a typical octahedral geometry, the C-terminal Ca^{2+} coordination adopts a tetrahedral configuration involving the side chains of residues Asp-689 and Asp-712 (both the Oδ1 and Oδ2) and completed by the main chain carbonyl groups of Phe-691 and Asp-707 (Figure 5.3C). A similar atypical calcium binding loop disruption has been observed in *R. flavefaciens Rf*XDocCttA structure in complex with *Rf*CohScaE, where a 13-residue long insertion in the C-terminal loop also alters the calcium coordination pattern in the X-module associated dockerin of the CttA protein, although the octahedral geometry is maintained thanks to the contribution of 2 water molecules (Figure 5.3C) (Salama-Alber *et al.*, 2013). CttA is believed to constitute the Carbohydrate-Binding Module that allows *R. flavefaciens* to be anchored to the plant cell wall (Salama-Alber *et al.*, 2013). In *Rf*XDocCttA, it was found that the loop insert, together with two other inserts, serves as structural buttresses stabilizing the X-Module-Doc relationship. However, there is no X-module associated with *Rf*DocScaA and therefore the 12 residue flap function remains unknown. Interestingly, the *Rf*DocScaA loop

insert, although having a similar location, has no primary structure homology with the *Rf*XDocCttA insert.

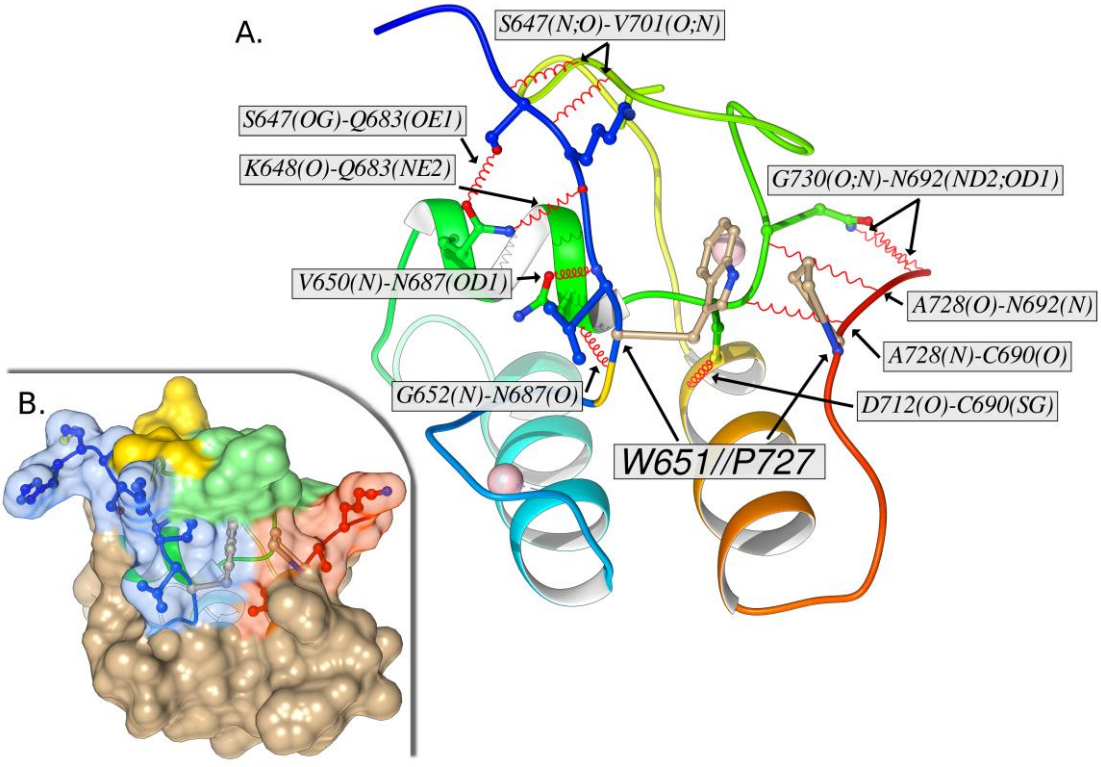
Figure 5.3 Calcium coordination geometry at the N-terminal and C-terminal F-hand motifs of *R. flavefaciens* DocScaA.



Panels A and B show a representation of *Rf*DocScaA N- and C-terminal Ca^{2+} coordination regions, respectively. In both panels the amino-acid residues involved in the metal coordination are depicted as sticks, surrounded by a mesh representation of the Refmac5 maximum-likelihood σ_A–weighted $2F_o-F_c$ electron density map contoured at 1σ (0.46 electrons/A^3). The labels show the *Rf*DocScaA residue and coordination position numbers and also the atoms involved. Both calcium ions are depicted as purple spheres and are overlaid with an idealized geometry representation (green arrows), which is octahedral for the N-terminal Ca^{2+} (Panel A) and tetrahedral for the C-terminal Ca^{2+} (Panel B). A single water molecule (Wat) completes the coordination sphere of the N-terminal Ca^{2+} ion (Panel A). The bidentate nature of the Asp-664 and Asp-712 coordination is highlighted with blue dashed lines (Panel A). The 12-residue insert at the C-terminal calcium coordination loop is colored in light green (Panel B). Panel C depicts the overlay of the C-terminal Ca^{2+} of *Rf*DocScaA (purple) with the C-terminal Ca^{2+} of group 4 dockerin *Rf*DocCttA (cyan), whose coordination is also disrupted by a 13-residue long insert (dark green), but maintains an octahedral geometry due to the contribution of 2 water molecules (Wat). The structure of *Rf*DocScaA is colored tan and the structure of *Rf*DocCttA is colored blue.

A recent study suggests the existence of an intramolecular clasp between the N-terminal and C-terminal ends of DocScaA, that contributes to increase the module's stability (Slutzki *et al.*, 2013). Based on an *in silico* model of DocScaA from *R. flavefaciens* strain 17, the authors predicted a stacking interaction between an N-terminal tryptophan and a C-terminal proline (Slutzki *et al.*, 2013). By mutating those two residues a reduction of the dockerin's thermal and chemical stability was observed (Slutzki *et al.*, 2013). The X-ray crystal structure of *Rf*DocScaA, observed here in complex with *Rf*CohScaB5, revealed the same stacking interaction between Trp-651 and Pro-727 (Figure 5.4) suggesting this may indeed be a crucial contact to maintain the dockerin structural integrity. Furthermore, this kind of aromatic interactions are commonly involved in protein structure stabilization and similar intramolecular clasps have been identified in other known dockerins (Adams *et al.*, 2006; Currie *et al.*, 2012; Waters, 2002). Additional intramolecular contacts established by both end of the protein, such as the hydrogen bonds between Cys-690 and Asp-712/Ala-728 and between Asn-687 and Val-650/Gly-652 should also provide additional structural stabilization to *Rf*DocScaA and contribute to its compact and globular conformation.

Figure 5.4 The most important intramolecular contacts for the stabilization of the dockerin module.

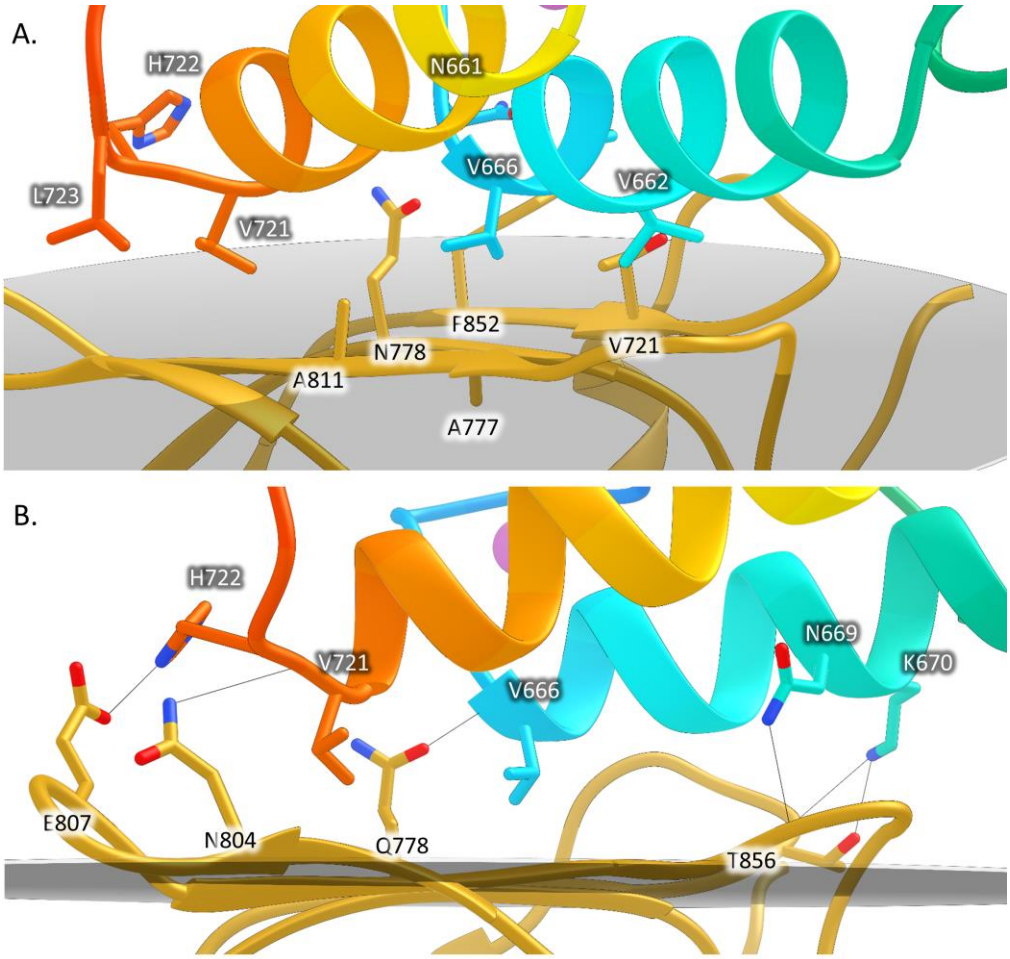


Panel A shows the structure of *Rf*DocScaA represented in color ramped style from the blue N-terminus to the red C-terminus. The sidechains of residues Trp-651 and Pro-727, which make an important stacking interaction are represented as ball & stick and colored tan. The most important hydrogen bond contacts involved in structure stabilization are represented as red springs and the residues making those contacts have their sidechains highlighted in ball&stick representation. Gly-652 is highlighted in yellow due to the lack of a sidechain. In panel B the molecular surface *Rf*DocScaA is represented in tan and shows the dockerin globular conformation supported by the extensive network of intramolecular contacts established by two ends of the protein, both between themselves and with other regions. The N- and C-terminal regions are highlighted in blue and red, respectively.

5.3.4. *Rf*CohScaB5-DocScaA complex interface

*Rf*DocScaA helices 1 and 3 make numerous contacts with the *Rf*CohScaB5 planar surface established by β-sheets 8-3-6-5 (Figure 5.5A,B). Although the Coh-interacting platform is predominantly flat, the loop connecting β-strands 8 and 9 is elevated from the plane defined by strands 8-3-6-5, thus positioning itself in close proximity to the N-terminus of *Rf*DocScaA helix-1. A slight elevation is also observed in the loop connecting β-strands 6 and 7, promoting the interaction with the middle to the C-terminal portion of helix-1. This means that the entire length of *Rf*DocScaA helix-1 interacts with the Coh surface. In contrast, helix-3 binds the Coh platform predominantly through the C-terminus. Thus, *Rf*DocScaA display a similar mechanism of Coh recognition to Group1 Docs that also bind to ScaA or ScaB Cohs, predominantly through one helix (Bule *et al.*, 2017).

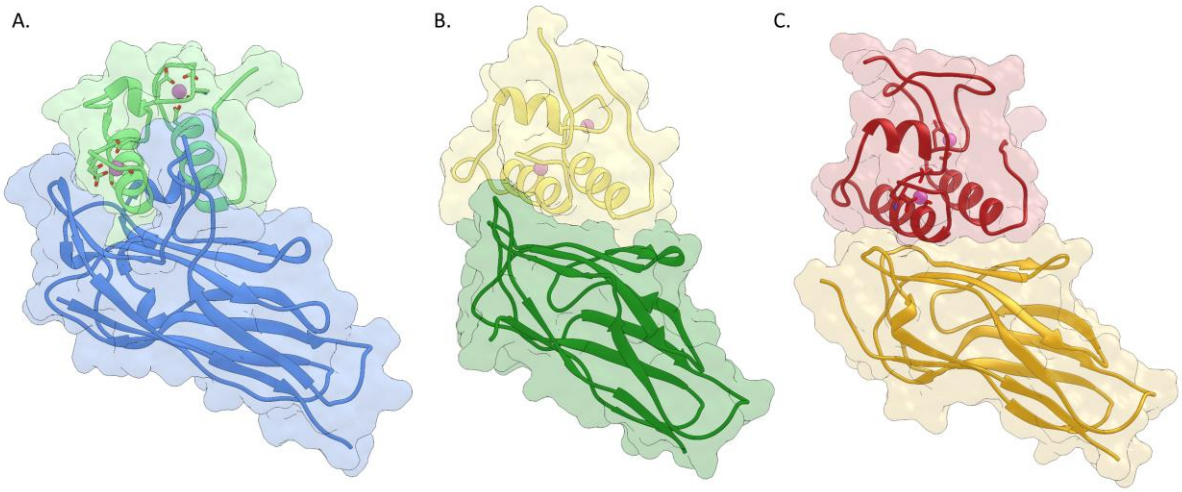
Figure 5.5 Cohesin-dockerin interface of *Rf*CohScaB5-DocScaA



Structure of *Rf*CohScaB5-DocScaA complex with a detailed view of the Coh-Doc interface showing the main polar interactions (Panel A) and main hydrophobic contacts (Panel B). In both panels the most important residues involved in Coh-Doc recognition are depicted in stick configuration, with a dark background label for the Doc residues and a light background label for the Coh residues, using the DocScaA and CohScaB5 numbering. Solid black lines mark hydrogen-bonds interactions. Ca^{2+} ions are depicted as purple spheres. In all panels, the transparent grey disk marks the plane defined by the 8-3-6-5 β-sheet, where the β-strands form a distinctive dockerin-interacting plateau.

However, *R. flavefaciens* group 3 and group 6 Docs interact with their Coh partners through the entire length of their two helixes as previously observed in the *R. flavefaciens Rf*CohScaC-Doc3 complex. Thus, in *Rf*CohScaC-Doc3 the two Doc3 α-helices (helix 1 and helix 3) make similar contributions to CohScaC recognition (Bule *et al.*, 2016) (Figure 5.6).

Figure 5.6 Structure of the three *R. flavefaciens* Coh-Doc complex specificities responsible for cellulosomal assembly.



Panel A depicts the structure of *Rf*CohScaC-Doc3 with the dockerin in light green and the cohesin in blue. This complex is responsible for recruiting group 3 and 6 dockerin associated enzymes *via* the ScaC adaptor scaffoldin to *R. flavefaciens* cellulosome. Panel B displays the structure of *Rf*CohScaB3-Doc1a with the dockerin in light yellow and the cohesin in dark green. This interaction is responsible for the integration of group 1 dockerin associated proteins directly to primary scaffoldins ScaA and ScaB. Group 1 Docs are the major Doc group in *R. flavefaciens*. Panel C shows the structure of *Rf*CohScaB5-DocScaA with the dockerin in dark red and the cohesin in gold. This interaction is responsible for attaching up to 5 ScaA primary scaffoldins onto a single ScaB primary/adaptor scaffoldin.

A large network of polar (Table 5.3) and hydrophobic interactions (Table 5.4) were identified stabilizing the *Rf*CohScaB5-DocScaA complex interface (Figure 5.5A,B). The interactions between the α-helix 1 of *Rf*DocScaA and the Coh are dominated by residues Val-662, Ala-663, Val-666, Leu-667, Asn-669 and Lys-670 while the main contacting residues of α-helix 3 are Ile-717, Val-720, Val-721, His-722 and Leu-723. The side chains of Val-662 and Val-666 at positions 11 and 15 dominate the hydrophobic recognition by contacting with *Rf*CohScaB5 hydrophobic platform formed by Ala-775/777 and Phe812/852 (Figure 5.5A). The highly hydrophobic character of α-helix 1 interaction is reinforced by the contacts established by Ala-663 and the aliphatic regions of Lys-670, Asn-673, Asn-661 and Asn-669 of *Rf*DocScaA. The hydrogen bond network established by α-helix 1 is dominated by the interaction of Asn-669 with Glu-814 of *Rf*CohScaB5 and Lys-670 with Thr-856 (both Oγ1 and Oδ1) and Asn-857 of *Rf*CohScaB5 (Figure 5.5B). An extra hydrogen bond is established between *Rf*DocScaA Val-666 main chain N and *Rf*CohScaB5 Gln-778. In α-helix-3 the contacts are dominated by the important hydrophobic interactions involving Val-721, whose side chain is positioned in the hydrophobic pocket created by Ala-811, Tyr-809, Tyr-810 and the aliphatic region of Ans-804 of *Rf*CohScaB5. Lys-710, Ile-717, Val-720, His-722 and Leu-723 reinforce the hydrophobic

contacts of α-helix-3. The close proximity of the C-terminal portion of α-helix-3 also allows the establishment of an important hydrogen bond between Val-721 of *Rf*DocScaA and Asn-804 of *Rf*CohScaB5. In addition, a salt bridge is established between the Nδ1 atom of *Rf*DocScaA His-722 and the Oε1 atom of *Rf*CohScaB5 Glu-807.

Table 5.3 . Main polar contacts between *Rf*CohScaB5 and *Rf*DocScaA.

DocScaA					CohScaB5		
	Atom	Residue	Residue #		Atom	Residue	Residue #
Hydrogen Bonds							
H1	N	VAL	662	< >	OE1	GLN	778
H1	ND2	ASN	669	< >	O	GLU	814
H1	NZ	LYS	670	< >	O	THR	856
H1	NZ	LYS	670	< >	OG1	THR	856
H1	NZ	LYS	670	< >	OD1	ASN	857
H3	O	VAL	721	< >	ND2	ASN	804
Salt Bridges							
H3	ND1	HIS	722	< >	OE1	GLU	807

Table was made using the PDBePISA server. Dockerin residues are marked as belonging either to helix 1 (H1) or to helix 3 (H3) interfaces.

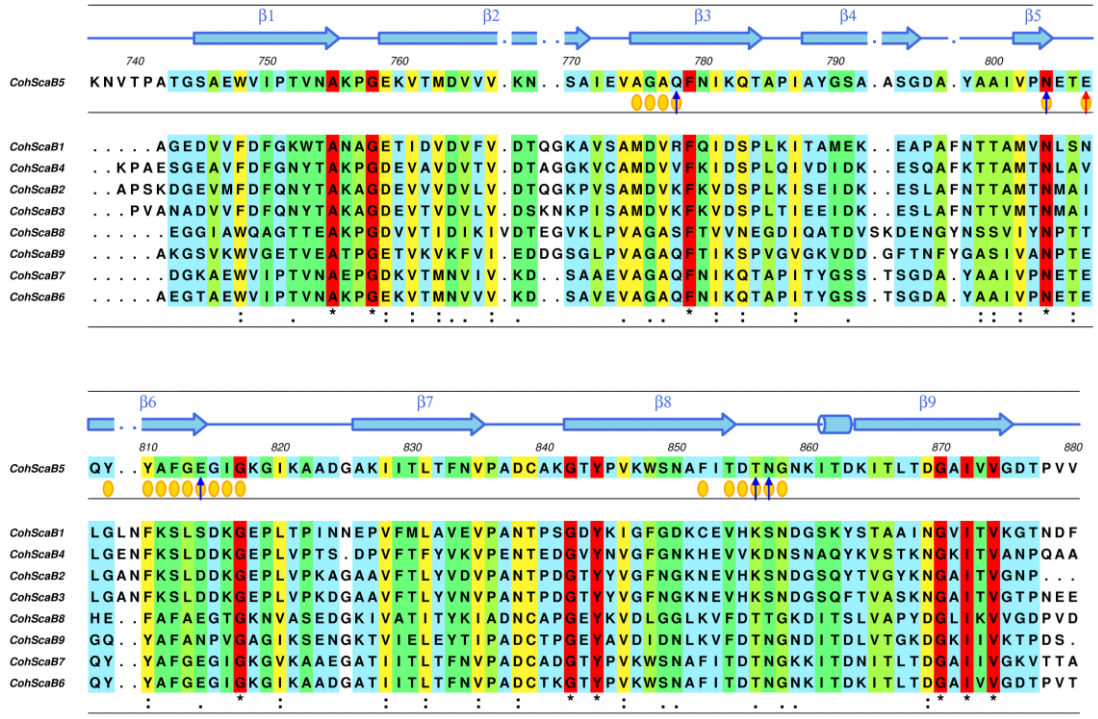
The structure of *Rf*CohScaB5-DocScaA protein complex allowed for the first time visualizing the residues of ScaB cohesin 5 that recognize ScaA. Previous work (Israeli-Ruimy *et al.*, 2017) revealed that ScaB cohesins 5 to 9 display a similar binding specificity as these Cohs bind exclusively to *Rf*ScaA Doc. Alignment of the primary sequences of Cohs 5 to 9 (Figure 5.7) revealed why a conservation in binding specificity is observed in these five Cohs. Thus, CohScaB5 residues Gln-778, Asn-804, Glu-807 and Thr-856, whose sidechains establish the main hydrogen bonds with DocScaA, are conserved in ScaB Cohs 6, 7 and 9. Interestingly, CohScaB8 Gln-778 and Glu-807 are replaced by hydroxy amino acids (Figure 5.7). Whether this fact has implications in the affinity for DocScaA remains to be explored. Furthermore, CohScaB5 Ala-811 is also conserved in CohScaB6 to 9. Ala-811 lies in the hydrophobic pocket that accommodates the sidechains of DocScaA Val-662 and Val-721. In CohScaB1-4, this Ala is replaced by a Lys that will not allow these very important hydrophobic contacts and very likely result in steric clash with CohScaA. Thus, Ala-Lys replacement is an important determinant of Coh-Doc specificity within *R. flavefaciens* cellulosome.

Table 5.4 Main hydrophobic contacts between *Rf*CohScaB5 and *Rf*DocScaA.

DocScaA					CohScaB5
		Residue	Residue #		Residues
		ASN	661	< >	GLN778 (5), PHE852 (7),
H1		VAL	662	< >	GLY776, ALA777 (4), GLN778 (5), PHE812, PHE852, THR854 (3)
H1		ALA	663	< >	THR854 (3), GLY858 (2)
H1		VAL	666	< >	ALA775 (4), GLY817 (3), THR854, ASP855, THR856
H1		LEU	667	< >	ASN857 (2)
H1		ASN	669	< >	GLU814 (4), GLY815 (7), ILE816 (6), GLY817
H1		LYS	670	< >	ILE816 (3), THR856 (5), ASN857 (4)
H1		ASN	673	< >	ILE816 (6)
	H3	LYS	710	< >	GLY815
	H3	ILE	717	< >	PHE812, GLY813
	H3	VAL	720	< >	GLN778 (3), TYR809 (3),
	H3	VAL	721	< >	ASN804 (3), TYR809 (6), TYR810, ALA811 (2)
	H3	HIS	722	< >	ASN804 (2), GLU807 (6), TYR809 (3)
	H3	LEU	723	< >	PRO803 (2), ASN804 (3)

Table was made using the PDBePISA server. Some of the dockerin residues are marked as belonging either to helix 1 (H1) or to helix 3 (H3) interfaces.

Figure 5.7 Multiple sequence alignment of *R. flavefaciens* ScaB cohesins 5 to 9.

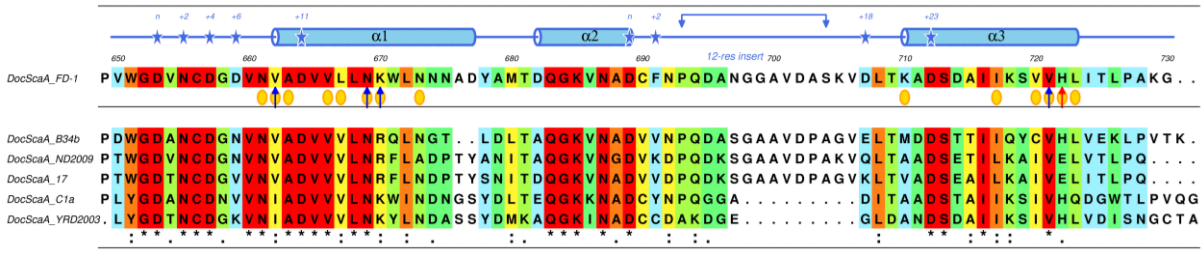


The primary sequence background is colored according to the ALSCRIPT Calcons convention, implemented in ALINE (Bond & Schüttelkopf, 2009): red, identical residues; orange to blue, lowering color-ramped scale of conservation. Above and below the alignment lies a cartoon representation of the secondary structure of *Rf*CohScaB5 (blue color) (Coh-Doc complex PDB codes: 5N5P). Residues involved in molecular interactions with the Doc partner are represented as follows: blue arrow for hydrogen bonds, red arrow for salt bridges and yellow circles for hydrophobic contacts.

Within *R. flavefaciens* cellulosome *Rf*DocScaA displays an exclusive binding specificity as it is the only Doc that is able to recognize ScaB cohesins 5 to 9. The alignment of *Rf*DocScaA with the Doc sequences of ScaA scaffoldins recently discovered in diverse *R. flavefaciens*

strains has enabled the degree of conservation of residues involved in the specific ScaB Coh recognition to be identified (Figure 5.8). Thus, within the 5 ScaA Doc homologues analyzed, residues Asn-661, Val-666, N669 and Val-721 are completely conserved and Val-662 is replaced by an Ile in 2 strains. This conservation reinforces the the importance of these residues for the DocScaA's ability to recognize CohScaB5.

Figure 5.8 Multiple sequence alignment of *Rf*DocScaA with its closest primary structure homologues.

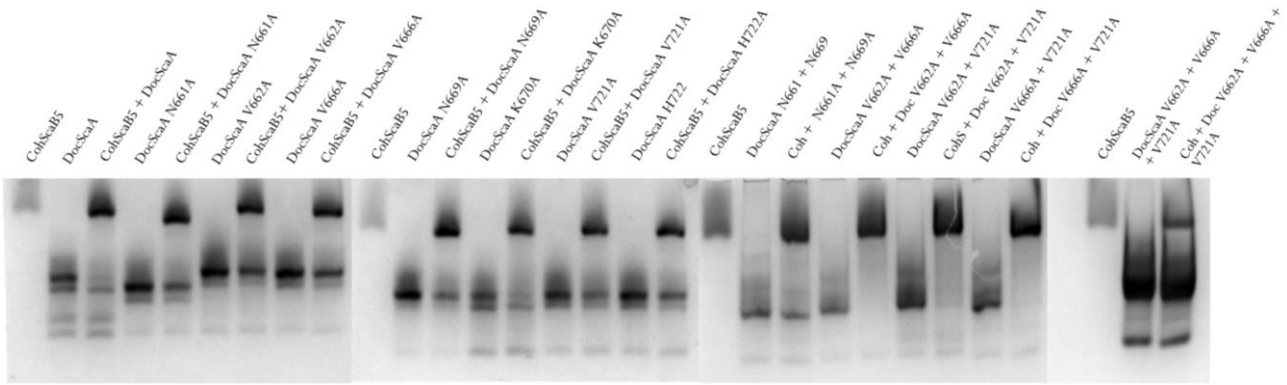


The primary sequence background is colored according to the ALSCRIPT Calcons convention, implemented in ALINE (Bond & Schüttelkopf, 2009): red, identical residues; orange to blue, lowering color-ramped scale of conservation. Above the alignment lies a cartoon representation of the secondary structure of *Rf*DocScaA from strain FD-1 (blue color) (Coh-Doc complex PDB code: 5N5P). Also, the residues involved in molecular interactions with the Coh partner are represented as follows: blue arrow for hydrogen bonds, red arrow for salt bridges and yellow circles for hydrophobic contacts.

5.3.5. *Rf*ScaA presents a single Coh-binding interface

The importance of *Rf*DocScaA residues for Coh recognition was initially probed through non-denaturing gel electrophoresis (NGE) (Figure 5.9). The data revealed that single mutant derivatives of *Rf*DocScaA retain the capacity to interact with their protein partner, suggesting that the amino acid substitutions explored in this study had a marginal impact in affinity.

Figure 5.9 Binding affinity of CohScaB5 to DocScaA and its mutant derivatives as determined by NGE.



In Panel A the lanes marked CohScaB5 were loaded with the Coh. Adjacent lanes were loaded with the dockerin mutant derivatives and with both Coh and Doc modules after 60-min incubation at equimolar

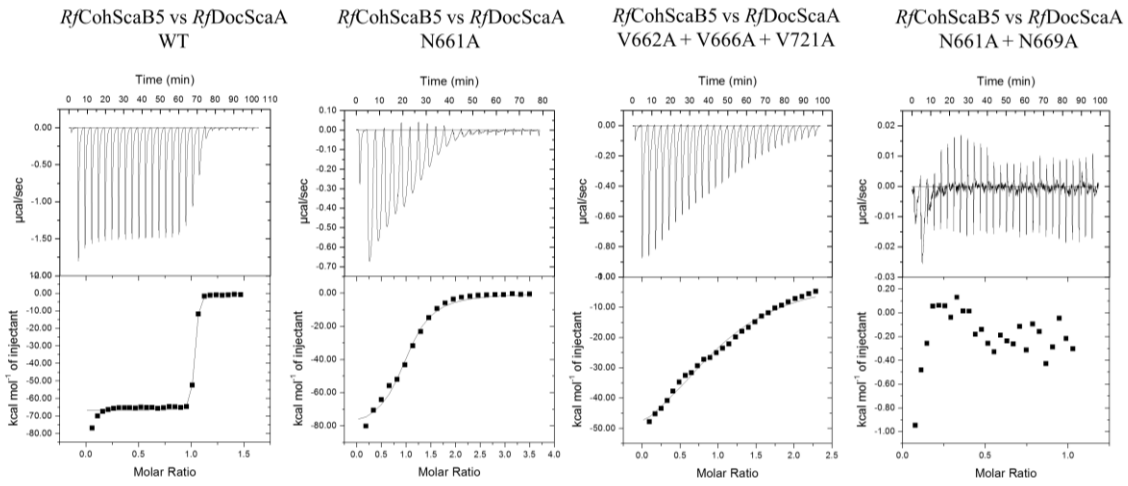
concentrations. The appearance of a band with a different migration pattern in lanes containing the complex represents a positive result (e.g. CohScaB5 + DocScaA3), while a negative result (e.g. Coh + N661A + N669A) is given by the presence of only the individual dockerin and cohesin bands. Thus, to gain more insight into the driving forces of Coh-Doc recognition, the binding thermodynamics of *Rf*DocScA to *Rf*CohScaB5 were assessed by isothermal titration calorimetry (ITC) at 308 K, consistent with the approximate temperature of the rumen. The data, presented in Table 5.5 and exemplified in Figure 5.10, revealed a macromolecular association with a 1:1 stoichiometry and a K_a of ~ 10^8 M^{-1}, an affinity similar to other Coh-Doc interactions of *R. flavefaciens*. The affinity might possibly be even higher since the error associated to the calculated K_a is high, probably due to the real value being too close to the upper sensitivity range of the technique. The interaction was driven by changes in enthalpy with the reduction in entropy having a negative impact on affinity.

Table 5.5 Thermodynamics of interaction between wild type CohScaB5 and wild type and mutant variants of ScaDocA.

Dockerin	K_a M^{-1}	ΔG^o kcal mol^{-1}	ΔH kcal mol^{-1}	$-T\Delta S^o$ kcal mol^{-1}	*N*
DocScaA WT	4.02E8 ± 1.69E8	-12.14	-66.66 ± 0.585	54.51	1.01
DocScaA N661A	2.64E6 ± 4.49E5	-9.15	-82.45 ± 2.800	73.30	0.98
DocScaA V662A	4.78E8 ± 5.03E7	-12.32	-68.38 ± 0.110	56.05	1.02
DocScaA V666A	4.07E8 ± 1.86E8	-12.20	-56.56 ± 0.391	44.35	1.02
DocScaA N669A	3.16E8 ± 6.59E7	-12.12	-73.42 ± 0.322	61.29	1.01
DocScaA K670A	3.67E8 ± 4.33E7	-12.25	-75.09 ± 0.184	62.83	1.02
DocScaA V721A	5.10E8 ± 1.54E8	-12.35	-48.70 ± 0.199	36.34	0.95
DocScaA H722A	2.73E8 ± 7.38E7	-12.07	-72.44 ± 0.443	60.36	1.00
DocScaA N661 + N669	*Nb*	*Nb*	*Nb*	*Nb*	*Nb*
DocScaA V662 + V666	1.98E7 ± 1.44E6	-10.19	-67.48 ± 0.281	57.29	0.98
DocScaA V662 + V721	6.24E7 ± 9.36E6	-11.05	-64.33 ± 0.455	53.28	1.08
DocScaA V666 + V721	2.25E7 ± 1.58E6	-10.33	-56.53 ± 0.238	46.2	1.01
DocScaA V662 + V666 + V721	2.91E5 ± 3.79E4	-7.83	-64.81 ± 3.59	56.98	1.04

All Thermodynamic parameters were determined at 308 K.

Figure 5.10 Binding affinity of CohScaB5 to DocScaA mutant derivatives and wild type partners as determined by ITC.



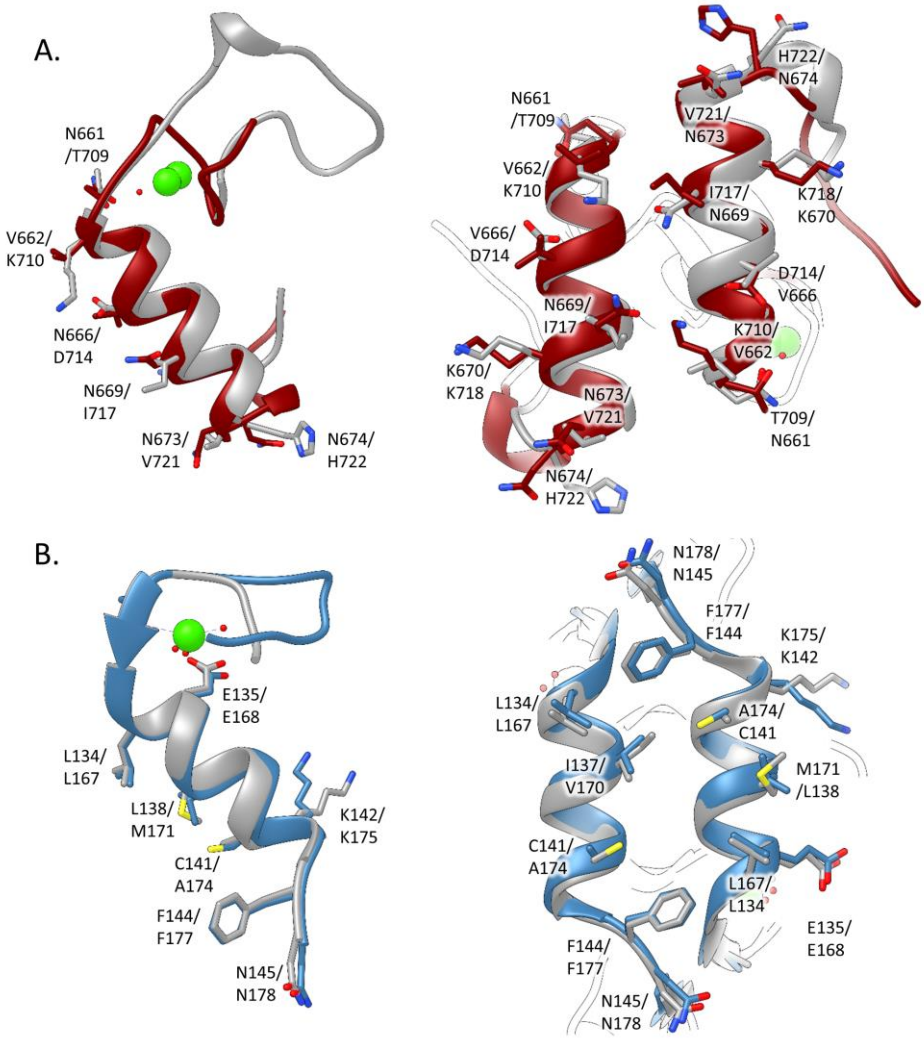
Binding isotherms for: Panel A, *Rf*CohScaB5 vs *Rf*DocScaA; Panel B, *Rf*CohScaB5 vs *Rf*DocScaA N661A; Panel C, *Rf*CohScaB5 vs *Rf*DocScaA triple mutant; and Panel D, *Rf*CohScaB5 vs *Rf*DocScaA N661A + N669A double mutant. The upper part of each panel shows the raw heats of binding, whereas the lower parts comprise the integrated heats after correction for heat of dilution. The curve represents the best fit to a single-site binding model. The corresponding thermodynamic parameters are shown in Table 5.5.

The affinity of *Rf*DocScaA mutant derivatives described above for *Rf*CohScaB5 was also explored by ITC. Thus, alanine substitution of *Rf*DocScaA residue Asn-661 resulted in a ~100-fold reduction in the affinity for *Rf*CohScaB3 (Table 5.5, Figure 5.10). Even though the alanine substitutions of residues Val-662, Asn-669, Lys-670 and His-722 did not result in a decreased K_a, the associated standard errors were lowered, which may indicate a reduction in affinity, enough to place its value within the ITC's sensitivity range. The low impact that the alanine substitution of Val-662, Val-666 and Val-721 had in affecting the affinity may reflect the inherent hydrophobic nature of the alanine side-chain and its ability to significantly compensate the disrupted interaction. Overall, single mutations on the dockerin contacting residues seem to have little to no effect on the affinity for its Coh partner. However, combining any 2 of the tested valine mutations (Val-662, Val-666, Val-721) into *Rf*DocScaA double mutants resulted in a ~10-fold reduction in the affinity for *Rf*CohScaB5. Mutating all 3 *Rf*DocScaA valines, led the K_a to drop approximately 1000 times lower than for the estimated wild type interaction (Figure 5.10). The *Rf*DocScaA double mutant derivative where Asn-661 and Asn-669 where replaced by alanines completely lost its capacity for *Rf*ScaBCoh5 recognition (Figure 5.10). These data suggest that both polar and hydrophobic interactions play an important role in stabilizing the *Rf*CohScaB5-DocScaA interaction, with particularly relevant contributions provided by Val-662, Val-666 and Val-721.

A close inspection of the *Rf*CohScaB5-DocScaA complex structure, suggests that *Rf*DocScaA residue Asn-661 does not plays a critical role in *Rf*CohScaB5 recognition when compared with other residues such as Val-662 and Val-666, although it does establish important hydrophobic contacts with *Rf*CohScaB5 Phe-852. However, Asn-661 is critically involved in the coordination of the N-terminal Ca^{2+}, which may explain the decreased affinity observed after alanine substitution. Thus, the alanine side-chain is unable to contribute for calcium coordination leading to an improper dockerin fold and resulting conformational changes that may hinder the interaction between the two modules. The thermogram resulting from the interaction between *Rf*ScaBCoh5 and the N661A *Rf*DocScaA mutant is displayed in Figure 5.10. Interestingly, the peaks appear to be broader, which can indicate that the binding reaction is slower than the one between *Rf*ScaBCoh5 and the wild type *Rf*DocScaA and therefore the heat signal is given over a longer period of time. Thus, the decreased affinity revealed by *Rf*DocScaA single and multiple mutant derivatives where Asn-661 was replaced by an alanine may reflect an improper Doc fold rather than the importance of the residue to Coh recognition. The observation that the Asn-661/Asn-669 mutant did not bind to its target Coh suggests that *Rf*DocScaA presents a single-binding mode; although Asn-661 substitution affected calcium coordination, it is plausible that under this conditions the symmetry related helix-3 could replace helix-1 supporting the recognition of *Rf*ScaBCoh5 through a symmetry related interface. When Docs present a dual-binding mode, mutation of a single or two residues positioned in the same helix usually has no effect on affinity, as a symmetry related functional binding site can assume Coh recognition involving a 180º rotation of the Doc when binding its protein partner. In addition, it is usually impossible to crystalize dual-binding mode complexes as these present conformational heterogeneity that precludes crystal formation. Thus, this initial observation strongly suggested that *Rf*DocScaA presents a single-binding mode. To analyze the nature of structural symmetry observed within *Rf*DocScaA, the structure of *Rf*DocScaA was overlaid with itself after rotation of 180º in the Coh plane (Figure 5.11). The overlay suggests that residues Asn-661 and Asn-669 are replaced by Thr-709 and Ile-717, respectively, when the Doc is rotated by 180º suggesting a disruption of the capacity of *Rf*DocScaA to recognize the Coh at these positions (Figure 5.11). In addition, the symmetry related residues for valines 662, 666 and 721 are all of polar nature and therefore do not allow establishment of the extensive hydrophobic platform created by this critical valine triad. Thus, overall these observations suggest that *Rf*CohScaB5-DocScaA interaction of the single-binding mode type due to the asymmetric nature of *Rf*DocScaA. This contrasts with a large majority of Coh-Doc interactions where a dual binding mode is observed, including those involving the binding of primary to adaptor scaffoldins as is the case for *Rf*CohScaB5-DocScaA. Thus, the symmetrical nature of

Acetivibrio cellulolyticus DocScaA, which was previously shown to display a dual binding mode by binding to cohesin *Ac*CohScaB3 in two distinct orientations (data not published) (4U3S, 4WI0), is easily demonstrated when its structure is overlayed with itself after a 180º rotation (Figure 5.11).

Figure 5.11 Non-symmetric and symmetric nature of Docs as exemplified by the structures of single binding mode *Rf*DocScaA and dual binding mode *Ac*DocScaA.



Panel A, *R. flavefaciens* Group5 Doc (DocScaA). Panel B, *A. cellulolyticus* DocScaA (*Ac*DocScaA). The left image of each panel shows an overlay of the N-terminal and C-terminal dockerin repeats. In both cases it is apparent that the 2 repeats are similar at the main-chain atoms but only the *Ac*DocScaA (Panel B) shows conservation in the side chains, allowing the dual-binding mode. The right image of each panel shows a comparison of the two putative binding surfaces by overlaying the dockerins with a version of themselves rotated by 180º (in grey), showing a lack of conservation in the key contacting residues in the *R. flavefaciens* dockerins (Panel A). Contrary to the *Ac*DocScaA (*B*), lack of internal symmetry in *Rf*DocScaA and the involvement of both α1 and α3 helices in cohesin recognition suggest that they display a single cohesin-binding platform.

5.4. Conclusions

This paper represents our latest contribution in understanding the structural nature of the Coh-Doc interactions used to assemble the highly complex cellulosomes operating in the rumen of mammals and exemplified by those secreted by *R. falvefaciens*. Recruitment of enzymes for *R. flavefaciens* cellulosome involves groups 1, 3 and 6 enzyme-borne Docs. Groups 3 and 6 Docs present essentially the same specificity, although a reversed binding mode, and recruit primarily hemicellulases to the multi-enzyme complex by binding the Coh of ScaC adaptor scaffoldin. ScaC contains a group 1 Doc that, like the remaining 95 group 1 Docs, specifically binds Cohs of primary scaffoldin ScaA as well as Coh 1 to 4 of adaptor scaffoldin ScaB. Thus, group 1 Docs represent the major group of Docs: those that recruit a large number of enzymes to ruminal cellulosomes. Work present in the previous Chapters reveals that group 1, 3 and 6 Docs essentially display a single binding mode mechanism. This contrasts with previous observations on the cellulosomes of *C. thermocellum* (Carvalho *et al.*, 2007) and *C. cellulolyticum* (Pinheiro *et al.*, 2008), which revealed that Docs used to recruit microbial enzymes to bacterial multi-enzyme complexes display a dual-binding mode. The structure of dual-binding mode Docs presents a 2-fold internal symmetry that allow binding to the Coh partner in two 180º-related alternate positions. The fact that Docs, in general, possess two different Coh-interacting platforms displaying identical specificities suggests that the dual-binding mode could contribute to enhance the conformational flexibility of the quaternary architecture of the highly populated multi-enzyme complex. In this Chapter, we have elucidated the structure of ScaA group 5 Doc bond to Coh 5 of ScaB. The data revealed that, like group 1, 3 and 6 Docs, ScaA doc 5 lacks the internal symmetry previously observed in all cellulosomal Docs. Thus, taken together, the data presented here and in the previous two Chapters reveals that the dual binding mode is not universal to all cellulosomal systems and, surprisingly, the most complex cellulosome described to data is assembled using single-binding mode Docs. This is rather a puzzling observation as the dual-binding mode was believed to improve flexibility of highly complex and populated cellulosomal systems. While it is possible, as suggested in the previous Chapter, that the dual-binding mode mechanism has evolved to enable the Docs in cellulosomes with a limited scaffoldin repertoire to explore a larger space by having alternate conformations, it is also possible that the dual binding mode represents an adaptation to the physic-chemical properties of different ecological niches. The fact that CAZymes have spread through bacteria and fungi essentially through horizontal gene transfer, suggest that the same mechanism operated to exchange the other components of cellulosomal systems (Shterzer & Mizrahi, 2015) Thus, all Docs evolved from a common ancestral sequence and were likely acquired through horizontal gene transfer. In addition, either ruminal Docs lost their ability to present a dual-binding mode

or Docs involved in the assembly of cellulosomes present in the soil have acquired a dual binding mode mechanism. Either way, the biochemical factors that constitute the driving selective force for the evolution of dual versus single-binding mode Docs remain to be elucidated.

Chapter 6

Type I Coh-Doc complexes of *Acetivibrio cellulolyticus*

Designing a dockerin with dual binding specificity based on the structure of the type I cohesin-dockerin complex of *Acetivibrio cellulolyticus* cellulosome

Pedro Bule[a], Kate Cameron[a], Luís M.A. Ferreira[a], Steven P. Smith[d], Harry J. Gilbert[e], Edward A. Bayer[b], Shabir Najmudin[a], Carlos M.G.A. Fontes[a] and Victor D. Alves[a,1]

[a] *CIISA – Faculdade de Medicina Veterinária, ULisboa, Pólo Universitário do Alto da Ajuda, Avenida da Universidade Técnica, 1300-477 Lisboa, Portugal*; [b] *Department of Biomolecular Sciences, The Weizmann Institute of Science, Rehovot 76100 Israel;* [c] *UCIBIO-REQUIMTE, Departamento de Química, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, 2829-516 Caparica, Portugal.* [d] *Department of Biomedical and Molecular Sciences, Queen's University, Kingston, ON K7L 3N6, Canada*; [e] *Institute for Cell and Molecular Biosciences, Newcastle University, The Medical School, Newcastle upon Tyne NE2 4HH, United Kingdom.* ***[1]Corresponding author***

Adapted from a manuscript in preparation

Abstract

The cellulosome is a remarkable intricate nanomachine developed by anaerobic bacteria to deconstruct complex carbohydrates. Integration of cellulosomal components usually occurs through the binding of type-I dockerin modules, located at the C-terminus of the cellulosomal enzymes, to cohesin modules located in a primary scaffoldin subunit. Cellulosomes are usually recruited to the surface of bacteria via type-II cohesin-dockerin interactions established between

primary and cell-surface anchoring scaffoldin subunits. It is now well established that type-I dockerins usually display a dual binding mode that is believed to increase conformational flexibility during cellulosome assembly. Unusually, *Acetivibrio cellulolyticus* produces a highly complex cellulosome comprising an adaptor scaffoldin, ScaB, which mediates the interaction between the primary scaffoldin, ScaA, through type-II cohesin-dockerin interactions and the anchoring scaffoldin, ScaC, via type-I cohesin-dockerin interactions. Here, we report the crystal structure of the type-I dockerin of a cellulosomal enzyme in complex with a type-I ScaA cohesin in two distinct orientations. The enzyme-borne dockerin displays internal structural symmetry, which supports the presence of two essentially identical binding surfaces. A mutagenesis study allowed identifying the residues that modulate type I cohesin-dockerin specificity in *A. cellulolyticus*. This knowledge was used to engineer a dockerin presenting two different cohesin binding interfaces; one that binds ScaA cohesins and the second one that binds ScaC cell surface cohesins. Thus, the generation of a dockerin with two different binding affinities illustrates how structure function studies can be used to generate novel specificities in bacterial dockerins.

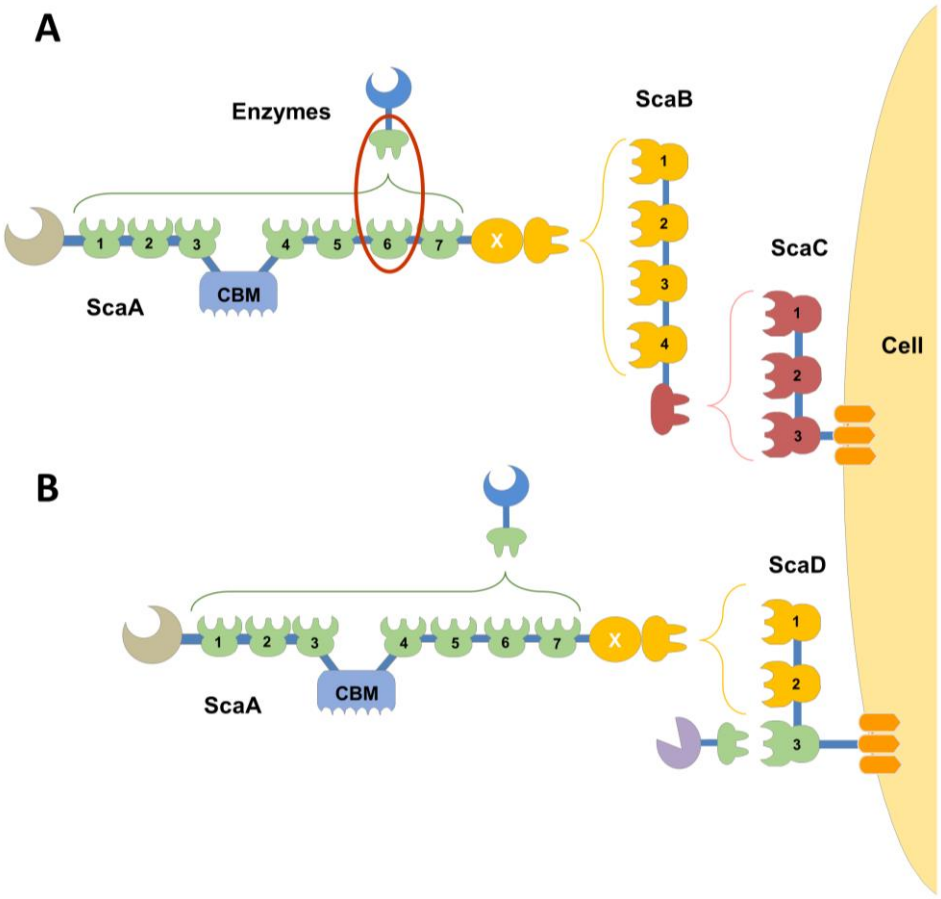
6.1. Introduction

Plant cell wall polysaccharides, primarily cellulose and hemicelluloses, are a major reservoir of carbon and energy (Fontes & Gilbert, 2010). As the demand for renewable sources for energy and novel molecules for the chemical industry increases, so does the environmental and industrial significance of these abundant structural molecules. The deconstruction of the plant cell wall requires, however, an extensive array of hydrolytic enzymes to attack this heterogeneous, predominantly insoluble and highly recalcitrant substrate (Gilbert, 2007). In nature, the microbial degradation of plant cell wall and its conversion to sugars and other byproducts is a key step of the carbon cycle. Specialized anaerobic bacteria have adopted an elaborate strategy to degrade structural plant carbohydrates, through the organization of enzymes into multiprotein complexes, termed cellulosomes (Bayer *et al.*, 2004). Typically, the molecular integration of microbial biocatalysts into these extremely elaborate nanomachines results from the binding of enzyme-borne type I dockerin (Doc) modules to reiterated type I cohesin (Coh) domains located in a large non-catalytic protein, termed scaffoldin, thus promoting enzyme synergism and protein stability. In addition, recruitment of cellulosomes to the bacterial cell surface via divergent type II Coh-Doc interactions allows the immediate uptake of released sugars, which are used by microbes as an energy source (Fontes & Gilbert, 2010; Leibovitz & Béguin, 1996). The protein:protein interaction established between the Coh and Doc modules exhibit one of the strongest affinities found in nature, close to that of a

covalent bond, and play a crucial role for both cellulosome assembly and cell-surface attachment (Bayer *et al.*, 2004; Carvalho *et al.*, 2003; Stahl *et al.*, 2012). In addition, the organization and structural architecture of cellulosomes are defined by the specificity of the different Coh and Doc modules (Ding *et al.*, 2000).

The mesophilic anaerobic bacterium *Acetivibrio cellulolyticus* produces a highly efficient cellulosome capable of hydrolyzing a range of cellulosic materials. These include crystalline cellulose, which is degraded with a higher efficacy than that of the *Aspergillus niger* and *Trichoderma viride* systems (Khan, 1980; Lamed, Naimark, Morgenstern, & Bayer, 1987). Initial sequencing of an *A. cellulolyticus* gene cluster identified four tandem scaffoldin genes (scaA, scaB, scaC, and scaD) (Xu *et al.*, 2003; Xu, Barak, *et al.*, 2004). The primary scaffoldin ScaA (where the enzymes of the cellulosome are recruited) shares the main traits found in the primary ScaA scaffoldin of the canonical cellulosome of *Clostridium thermocellum*. Thus, *A. cellulolyticus* ScaA contains an internal family 3 Carbohydrate-Binding Module (CBM3), flanked by seven type I Coh domains, a single X-module and a divergent C-terminal type II Doc (Ding *et al.*, 1999). Downstream of ScaA are genes encoding for an adaptor and an anchoring scaffoldin, ScaB and ScaC, respectively. ScaB was found to contain four type II Coh modules, which interact with the C-terminal type II Docs of the ScaA, and a divergent C-terminal type I Doc which in turn interacts with the type I Coh modules found on the ScaC scaffoldin. ScaB essentially plays the role of an adaptor protein, which mediates the interaction between ScaA (and its incorporated enzymes) and ScaC. This ScaB scaffoldin was the first example of an adaptor protein discovered in nature (Xu *et al.*, 2003). In turn, ScaC acts as an anchoring scaffoldin by virtue of its C-terminal SLH module (Figure 6.1) (Xu *et al.*, 2003). The recent sequencing of the *A. cellulolyticus* CD2 genome identified numerous additional cellulosomal components, gene regulatory elements and cell anchoring modules (identified by the presence of signature Docs or Cohs sequences), suggestive of a much more elaborate and sophisticated cellulosome system than originally observed (Dassa *et al.*, 2012). The genome of *A. cellulolyticus* encodes 143 Doc-containing proteins, which is considerably more than that observed in most clostridial bacteria, but fewer than the 220 cellulosomal proteins encoded by the *Ruminococcus flavefaciens* FD-1 genome (Berg Miller *et al.*, 2009).

Figure 6.1 Architecture of *A. cellulolyticus* cellulosome.



The scheme is color coded to highlight the different Coh-Doc specificities. In A) Doc-containing enzymes are incorporated into the ScaA scaffoldin through interaction with the seven ScaA Cohs (light green). ScaB plays the role of an adaptor protein that mediates between the ScaA Doc (yellow) and the Cohs of the anchoring scaffoldin (red) ScaC. The entire complex is attached to the cell surface *via* the SLH module of ScaC (orange). ScaA contains also a CBM (blue) and a GH9 (light brown) catalytic module. In B) An additional mechanism of cellulosome attachment; ScaA is bound to the type II Cohs of ScaD (yellow), that can also accept a single enzyme via its third type I Coh (light green). The SLH module of ScaD serves to anchor the alternative complex to the cell surface.

Although structurally related, there is no cross-specificity between type I and type II Coh-Doc partners, which allows for the efficient assembly and cell-surface attachment of bacterial cellulosomes (Miras *et al.*, 2002; Schaeffer *et al.*, 2002). Structural studies on type-I complexes from several organisms, including *Clostridium thermocellum* (Brás *et al.*, 2012; Carvalho *et al.*, 2003, 2007) and *Clostridium cellulolyticum* (Pinheiro *et al.*, 2008), revealed that the primary sequence duplication displayed by type-I Docs supports a dual-binding mode, based on the interaction of two 180º-symmetry-related binding interfaces. It was recently shown that the sequence and structural symmetry within the ScaB *A. cellulolyticus* type I Doc allows it to bind ScaC Cohs in two different orientations (Cameron, Najmudin, *et al.*, 2015). This symmetry is also evident in the enzyme-borne Docs of *A. cellulolyticus* that interact with ScaA, therefore suggesting a putative dual-binding mode capability for these interactions. Although very closely

related, the enzyme-borne and ScaB type I Docs do not display cross-specificity. Thus, Coh-contacting residues at positions 11 and 12 (numbering established considering the first Gly of each calcium binding loop as residue 1), which are traditionally recognized as specificity determinants (Bayer *et al.*, 2004), are different in the two type I Docs. Differences at these key residues may explain why there is a lack of cross-specificity between the type I-Doc interactions that modulate the binding of ScaB onto ScaC or the cellulosomal enzymes onto ScaA (Hamberg *et al.*, 2014; Xu *et al.*, 2003).

Although the striking symmetry of the duplicated Doc segments supports a potential dual-binding mode, it does not necessarily mean that both Doc orientations are capable of binding to the Coh. In this context, mutagenesis studies, combined with structural and affinity-based binding data, usually provide confirmation for the dual binding mode between a given Coh-Doc pair. Here, we report the structure of the protein complex established between the sixth Coh from ScaA (*Ac*CohScaA6) and a family 5 glycoside hydrolase associated type I Doc (*Ac*DocCel5) from *A. cellulolyticus*. A comprehensive biochemical analysis guided by the crystal structure confirmed that the enzyme-borne Docs of *A. cellulolyticus* that bind to ScaA also display a dual binding mode. By combining these data with previous information on the ScaB Doc interaction with ScaC Coh, a specificity hybrid Doc with the ability to recognize both ScaA and ScaC Cohs was successfully designed.

6.2. Experimental Procedures

6.2.1. Gene synthesis and DNA cloning

Docs are inherently unstable when produced in *Escherichia coli*. To promote Doc stability, *A. cellulolyticus* DocCel5 of protein WP_010249057 (residues 502 - 573) was co-expressed *in vivo* with the sixth Coh of ScaA, *Ac*CohScaA6 (AAF06064; residues 1472 – 1611). The immediate binding of *Ac*DocCel5 to *Ac*CohScaA6 is believed to confer the necessary Doc stabilization. The genes encoding the two proteins were designed with a codon usage optimized to maximize expression in *E. coli*, synthesized *in vitro* (NZYTech Ltd, Lisbon, Portugal) and cloned into pET28a (Merck Millipore, Germany) under the control of separate T7 promoters. The *Ac*DocCel5-encoding gene was positioned at the 5’ end and the *Ac*CohScaA6-encoding gene at the 3’ end of the artificial DNA. A T7 terminator sequence (to terminate transcription of the Doc gene) and a T7 promoter sequence (to control transcription of the Coh gene) were incorporated between the sequences of the two genes. This construct contained specifically tailored NheI and NcoI recognition sites at the 5’ end and XhoI and SalI at the 3’ end to allow subcloning the nucleic acid into pET-28a (Merck Millipore, Germany) such that the sequence

encoding a six-residue His tag could be introduced either at the N-terminus of the Doc (through digestion with NheI and SalI, incorporating the additional sequence MGSSHHHHHHSSGLVPRGSHMAS at the N-terminus of the *Ac*DocCel5) or at the C-terminus of the *Ac*CohScaA6 (by cutting with NcoI and XhoI, which incorporates the additional sequence LEHHHHHH at the C-terminus of the Coh). To block the dual binding mode and promote the structural homogeneity required for protein crystallization, two different genes were synthetized, each with a distinct Doc mutant: mutant M1 with the S15I and I16N amino acid changes and mutant M2 with the S51I and L52N replacements. These substitutions represent residue changes to amino acids present in Type I Docs of *A. cellulolyticus* that do not bind to ScaA, but rather to the cell-surface anchoring scaffoldin ScaC. In addition, these residues are located, respectively, at the N-terminal and C-terminal Coh recognition sites. Thus, as a result of this strategy four pET28a plasmid derivatives were produced: pET28DtC M1 and M2 with the engineered tag in the Doc and pET28DCt M1 and M2 where the engineered tag is attached to the Coh. The four plasmids were used to express *Ac*CohScaA6-DocCel5 M1 and M2 complexes in *E. coli*. Recombinant *Ac*DocCel5 and *Ac*CohScaA6 primary sequences are presented in Table 6.1.

Table 6.1 Recombinant protein sequences of *Ac*DocCel5, *Ac*DocScaB, *Ac*CohScaA6, *Ac*CohScaC3 and mutants of both dockerins, produced for the interaction studies.

Protein	Sequence
*Ac*DocCel5 WT	DVKPGDVDGNGSINSIDFALMRNYLLGNLKDFPAEDDIKAGDLNGDKSINSLDFAIMRMYLLGMITKFSV
*Ac*DocCel5 M1 (S15I, I16N)	DVKPGDVDGNGSININDFALMRNYLLGNLKDFPAEDDIKAGDLNGDKSINSLDFAIMRMYLLGMITKFSV
*Ac*DocCel5 M2 (S51I, L52N)	DVKPGDVDGNGSINSIDFALMRNYLLGNLKDFPAEDDIKAGDLNGDKSININDFAIMRMYLLGMITKFSV
*Ac*DocCel5 M1 + M2	DVKPGDVDGNGSININDFALMRNYLLGNLKDFPAEDDIKAGDLNGDKSININDFAIMRMYLLGMITKFSV
*Ac*DocCel5 (N14R, S15I, I16N, F18A, A19V, N23D)	DVKPGDVDGNGSIRINDAVLMRDYLLGNLKDFPAEDDIKAGDLNGDKSINSLDFAIMRMYLLGMITKFSV
*Ac*DocCel5 (N14R, S15I, I16N, F18A, A19V, M21I, N23D)	DVKPGDVDGNGSIRINDAVLIRDYLLGNLKDFPAEDDIKAGDLNGDKSINSLDFAIMRMYLLGMITKFSV
*Ac*DocScaB WT	KFIYGDVDGNGSVRINDAVLIRDYVLGKINEFPYEYGMLAADVDGNGSIKINDAVLVRDYVLGKIFLFPVEEKEE
*Ac*DocScaB M7	KFIYGDVDGNGSVNSIDFVYIRQYVLGKINEFPYEYGMLAADVDGNGSIKINDAVLVRDYVLGKIFLFPVEEKEE
*Ac*DocScaB M8	KFIYGDVDGNGSVNSIDFVYIRQYVLGKINEFPYEYGMLAADVDGNGSINSIDFVYVRQYVLGKIFLFPVEEKEE
*Ac*CohScaA6 WT	QTGFNLSIDTVEGNPGSSVVVPVKLSGISKNGISTADFTVTYDATKLEYISGDAGSIVTNPGVNFGINKESDGKLK VLFLDYTMSTGYISTDGVFANLNFNIKSSAAIGSKAEVSISGTPTFGDSTLTPVVAKVTNGAVN
*Ac*CohScaC3 WT	LQVDIGSTSGKAGSVVSVPITFTNVPKSGIYALSFRTNFDPQKVTVASIDAGSLIENASDFTTYYNNENGFASMTF EAPVDRARIIDSDGVFATINFKVSDSAKVGELYNITTNSAYTSFYYSGTDEIKNVVYNDGKIEVIA

The mutated residues are highlighted in black.

To produce recombinant *Ac*CohScaA6 and *Ac*DocCel5 individually, an ELISA-based system designed to probe Coh-Doc affinities that requires fusion with xylanase or carbohydrate-binding modules (CBMs) was selected, as it allows production of highly stable and functional Coh and Doc derivatives (Barak *et al.*, 2005). Thus, sequences encoding each of the 2 modules were amplified from *A. cellulolyticus* genomic DNA by PCR, using NZYProof polymerase (NZYTech Ltd, Portugal) and the primers shown in Table S6.1 (Annexes). The M1 and M2

Doc mutants were amplified from the previously described synthetized DNA constructs. Following gel purification, the *Ac*DocCel5 encoding amplicon was inserted into a Xylanase-Doc cassette in pET9d plasmid after digestion with KpnI and BamHI and ligation with T4-ligase. The resulting expressed products consist of His-tagged *Ac*DocCel5 fused to the xylanase T-6 from *Geobacillus stearothermophilus* at the N-terminus of the polyhistidine tag (Xyn *Ac*DocCel5). The *Ac*CohScaA6 encoding gene was cloned into CBM-Coh cassettes in pET28a after digestion with BamHI and XhoI restriction enzymes. This resulted in His-tagged *Ac*CohScaA6 recombinant derivative fused to a CBM3a from the *Clostridium thermocellum* scaffoldin ScaA (CBM *Ac*CohScaA6) (Handelsman *et al.*, 2004). Xyn *Ac*DocScaB and CBM *Ac*CohScaC3 were produced for a previous study, following the same approach (Cameron, Najmudin, *et al.*, 2015).

For the specificity switch experiments, several Xyn*Ac*DocCel5 protein derivatives were produced using site directed mutagenesis (Table S6.1 Annexes) . Each of the newly generated gene sequence was fully sequenced to verify that only the desired mutation accumulated in the nucleic acid chain. The *Ac*DocScaB mutants (M7 and M8) were produced for a previous study using previously published primers (Cameron, Najmudin, *et al.*, 2015).

6.2.2. Expression and purification of recombinant proteins

Preliminary expression screens revealed that when the polyhistidine tag was located at the Doc N-terminal end of the *Ac*CohScaA6-DocCel5 complex, the expression levels of both Coh and Doc were higher. Tagging the Coh resulted in the accumulation of large levels of unbound Coh in the purification product suggesting that Coh was expressed at higher levels than Docs. Consequently, the plasmid pET28DtC was used to transform *E. coli* BL21 (DE3) cells in order to produce *Ac*CohScaA6-DocCel5 complex in large quantities. Transformed *E. coli* were grown at 37ºC to an OD_{600} of 0.5. Recombinant protein expression was induced by the addition of 1 mM isopropyl *β*-D-1-thiogalactopyranoside followed by incubation at 19ºC for 16 hours. Cells were harvested by 15 min centrifugation at 5000 x g and resuspended in 20 mL of IMAC binding buffer (50 mM HEPES, pH 7.5, 10 mM imidazole, 1 M NaCl, 5 mM $CaCl_2$). Cells were then disrupted by sonication and the cell free supernatant recovered by 30 min centrifugation at 15,000 x g. After loading the soluble fraction into a HisTrap^{TM} nickel charged sepharose column (GE Healthcare, UK), initial purification was carried out by IMAC in a FPLC system (GE Healthcare, UK) using conventional protocols with a 35 mM imidazole wash and a 35-300 mM imidazole gradient. The buffer of all recovered fractions containing the purified Coh–Doc complex was exchanged into 50 mM HEPES, pH 7.5, containing 200 mM NaCl, 5 mM $CaCl_2$ using a PD-10 Sephadex G-25M gel-filtration column (Amersham Pharmacia Biosciences,

UK). A further purification step by gel-filtration chromatography was performed by loading the samples onto a HiLoad 16/60 Superdex 75 (GE Healthcare, UK) at a flow rate of 1 ml min^{-1}. Fractions containing the purified complex were then concentrated with Amicon Ultra-15 centrifugal devices with a 10 kDa cutoff membrane (Millipore, USA) and washed three times with molecular biology grade water (Sigma) containing 0.5 mM $CaCl_2$. The protein concentration was estimated in a NanoDrop 2000c spectrophotometer (Thermo Scientific, USA) using a molar extinction coefficient (ε) of 8,940M^{-1} cm^{-1}. The final protein concentration was adjusted to 12 mg/mL for Xyn*Ac*DocCel5 M2 and 15 mg/mL for Xyn*Ac*DocCel5 M1, in molecular biology grade water containing 0.5 mM $CaCl_2$. The purity and molecular mass of the recombinant complex was confirmed by 14% (w/v) SDS–PAGE.

CBMCohs, XynDocs and respective protein derivatives used in ITC and native PAGE experiments were expressed as described above and purified with IMAC by nickel charged sepharose His GraviTrap gravity-flow columns (GE Healthcare, UK). After IMAC, the recombinant Coh and Docs were buffer exchanged to 50 mM HEPES pH 7.5, 0.5 mM $CaCl_2$ and 0.5 mM TCEP using PD-10 Sephadex G-25M gel filtration columns (GE Healthcare, UK).

6.2.3. Nondenaturing gel electrophoresis (NGE)

For the NGE experiments each of the Xyn*Ac*DocCel5 and Xyn*Ac*DocScaB variants, at a concentration of 30 μM, was incubated in the presence and absence of 30 μM CBM *Ac*CohScaA6 or CBM*Ac*CohScaC3 for 30 min at room temperature and separated on a 10% native polyacrilamide gel. Electrophoresis was carried out at room temperature. The gels were stained with Comassie Blue. Complex formation was detected by the presence of an additional band displaying a lower electrophoretic mobility than the individual modules.

6.2.4. Isothermal titration calorimetry

All ITC experiments were carried out at 308 K. The purified Xyn*Ac*DocCel5, Xyn*Ac*DocScaB, CBM *Ac*CohScaA6 or CBM*Ac*CohScaC3 variants were diluted to the required concentrations and filtered using a 0.45 μm syringe filter (PALL). During titrations, the Doc constructs were stirred at 307 revolutions/min in the reaction cell and titrated with 28 successive 10 μL injections of Coh at 220 s intervals. Integrated heat effects, after correction for heats of dilution, were analyzed by nonlinear regression using a single-site model (Microcal ORIGIN version 7.0, Microcal Software, USA). The fitted data yielded the association constant (K_A) and the enthalpy of binding (ΔH). Other thermodynamic parameters were calculated using the standard thermodynamic equation: $\Delta RTlnK_A=\Delta G=\Delta H-T\Delta S$.

6.2.5. X-ray crystallography, structural determination and refinement

The crystallization conditions were set up using the sitting-drop vapor-diffusion method with an Oryx8 robotic nanodrop dispensing system (Douglas Instruments, UK; (Bule, Correia, *et al.*, 2014). The commercial kits Crystal Screen, Crystal Screen 2, PEG/Ion and PEG/Ion 2 (Hampton Research, California, USA), JCSG+ HT96 (Molecular Dimensions,UK) and an in-house screen (80 factorial) were used for the screening. Precisely 0.7 µl drops of 15 and 12 mg/mL mg ml^{-1} of *Ac*CohScaA6-DocCel5 M1 and M2, respectively, were mixed with 0.7 µl reservoir solution at room temperature per well containing 50 µl of the crystallization solution. The resulting plates were then stored at 292 K. Crystal formation was observed in 4 different conditions for *Ac*CohScaA6-DocCel5 M1 and in 1 condition for *Ac*CohScaA6-DocCel5 M2, within approximately 15 days (maximum dimension ~120 x 100 x 30 µm). All the crystals were obtained from the initial screens. These crystals were cryoprotected with mother solution containing 20–30 % glycerol or with 100 % Paratone-N (Hampton Research, USA) and flash-cooled in liquid nitrogen. Data were collected on beamline ID29 at the European Synchrotron Radiation Facility, Grenoble, France, using a PILATUS 6M detector (Dectris Ltd) from crystals cooled to 100 K using a Cryostream (Oxford Cryosystems Ltd). iMOSFLM (Battye *et al.*, 2011) was used for strategy calculation during data collection. All data sets were processed using iMOSFLM (Battye *et al.*, 2011) and AIMLESS (Evans, 2006) from the CCP4 suite (Collaborative Computational Project, Number 4, 1994; (Winn *et al.*, 2011). Data collection statistics are given in Table 6.2.

The best diffracting *Ac*CohScaA6-DocCel5 M1 crystals were the ones formed in the condition composed of 0.2 M sodium thiocyanate, 20% (w/v) PEG 3350, pH 6.9 and diffracted to a resolution of 1.57 Å. The crystals from the other three conditions did not diffract at all. The crystal belongs to the orthorhombic space group $P2_12_12_1$. The best diffracting *Ac*CohScaA6-DocCel5 M2 crystals were those formed in the condition composed by 0.2M $CaCl_2$, 0.1M HEPES, pH 7.5 and 28% PEG400. The crystal belongs to the monoclinic space group $P2_1$. BALBES was used to carry out molecular replacement (Long *et al.*, 2008). The best solution for *Ac*CohScaA6-DocCel5 M1 was found using the type I Coh-Doc complex from *C. cellulolyticum* (PDB entries 2vn5/6 with sequence identity of 36.9% with the Coh and 32.8% with the Doc; (Pinheiro *et al.*, 2008)) producing at the end of the BALBES run a R_{factor} and R_{free} of 35.7% and 40.6%, respectively, and a Q-factor of 0.719 after REFMAC5 refinement (Murshudov *et al.*, 2011). An ARP/wARP (Langer, Cohen, Lamzin, & Perrakis, 2008) run after BALBES gave a model of 400 residues in 6 chains, with an estimated correctness of 99.9%. Two copies of the heterodimer *Ac*CohScaA6-DocCel5 M1 complex are present in the

asymmetric unit. This model was adjusted and refined using REFMAC5 and PDB REDO (Joosten *et al.*, 2014) interspersed with model adjustment in COOT to give the final structure (Protein Data Bank code 5NRK, Table 6.2) (Emsley & Cowtan, 2004). The final round of refinement was performed using the TLS/restrained refinement procedure using each module as a single group. The root mean square deviation of bond lengths, bond angles, torsion angles and other indicators were continuously monitored using validation tools in COOT and MOLPROBITY (Chen *et al.*, 2010). A summary of the refinement statistics is shown in Table 6.2. The best solution for *Ac*CohScaA6-DocCel5 M2 was found using the *Ac*CohScaA6-DocCel5 M1 refined model. The refinement process was as described above for *Ac*CohScaA6-DocCel5 M1 (Protein Data Bank code 5NRM, Table 6.2).

Table 6.2 X-ray crystallography data collection and refinement statistics for *Ac*CohScaA6-Gh5Doc.

Dataset	***Ac*CohScaA6-DocCel5 M1**	***Ac*CohScaA6-DocCel5 M2**
Data Colection		
Beamline	*ESRF ID29*	*ESRF ID29*
Space Group	*$P2_12_12_1$*	*$P2_1$*
Wavelength (Å)	*0.9763*	*0.9763*
Unit-cell parameters		
a, b, c (Å)	*46.539, 79.809, 112.159*	*30.49, 59.95, 51.26*
α, β, γ (º)	*90, 90, 90*	*90, 106.88, 90*
$Vm^{\sharp}$ ($Å^3$ Da^{-1})	*2.11*	*2.67*
Solvent Content (%)	*42*	*53.98*
Resolution limits (Å)	*65.03 - 1.45 (1.502 - 1.45)*	*49.05 - 1.4 (1.45 - 1.4)*
No. of observations	*1417284 (78998)*	*485764 (47950)*
No. of unique observations	*74084 (7157)*	*33519 (3358)*
Multiplicity	*19.0 (11.0)*	*14.2 (14.3)*
Completeness (%)	*99.00 (95.16)*	*96.33 (94.58)*
<I/σ(I)>	*18.9 (3.5)*	*7.0 (2.2)*
$CC1/2^{\dagger}$	*0.998 (0.912)*	*0.995 (0.503)*
Wilson B-factor	*10.75*	*12.78*
Rmerge ‡	*0.057 (0.203)*	*0.083 (0.599)*
Strtucture Refinement		
R-work § , R-free $^{¥}$	*0.144, 0.170*	*0.178, 0.205*
No. of Non-H atoms	*3831*	*1819*
Macromolecules	*3224*	*1762*
Ligands	*27*	*3*
Water	*580*	*231*
Protein residues	*421*	*209*
RMS(bonds)	*0.013*	*0.020*
RMS(angles)	*1.44*	*2.08*
Ramachandran favored (%)	*97*	*98*
Ramachandran outliers (%)	*0*	*0*
Clash score	*15.70*	*11.97*
Average B-factor	*15.70*	*16.80*
macromolecules	*13.60*	*15.50*
ligands	*23.50*	*15.20*
solvent	*26.90*	*25.60*
PDB accession code	*5NRK*	*5NRM*

Values in parenthesis are for the highest resolution shell. # Matthews coefficient (Matthews, 1968). † $CC_{1/2}$ = the correlation between intensities from random half-dataset (Diederichs & Karplus, 2013) ‡ $R_{merge} = \Sigma_{hkl} \Sigma_i |I_i(hkl) - \langle I(hkl) \rangle| / \Sigma_{hkl} \Sigma_i I_i(hkl)$, where $I_i(hkl)$ is the *i*th intensity measurement of reflection *hkl*, including symmetry-related reflections and $\langle I(hkl) \rangle$ is its average. §$R_{work} = \Sigma_{hkl} ||F_{obs}| - |F_{calc}|| / \Sigma_{hkl} |F_{obs}|$. ¥$R_{free}$ as R_{work}, but summed over a 5% test set of reflections.

6.3. Results and Discussion

Previous studies have shown that the type I Doc of *Acetivibrio cellulolyticus* ScaB binds specifically to the type I Cohs of ScaC, but not to those of ScaA. Similarly, enzyme borne type I Docs specifically bind to the nine type I Cohs of ScaA (and to one in ScaD), but not to those of ScaC (Hamberg *et al.*, 2014). In adherence to the canonical cellulosomal organizational framework, there are thus two distinct specificities within type I Coh-Doc complexes of the *A. cellulolyticus* cellulosomal system, one responsible for recruiting enzymes to ScaA and a second one responsible for the anchoring of cellulosomes to the cell wall surface (Figure 6.1). A recent study explored the structural and biochemical nature of one of these specificities by studying the interaction between the Doc of adaptor scaffoldin ScaB (*Ac*DocScaB) and the third Coh of anchoring scaffoldin (*Ac*CohScaC3) (Cameron, Najmudin, *et al.*, 2015). In order to gain further insights into the molecular mechanisms of *A. cellulolyticus* cellulosomal assembly, the structure of the Coh-Doc complex that recruits cellulosomal enzymes to ScaA was investigated by solving the X-ray crystal structure of the sixth ScaA Coh (*Ac*CohScaA6) in complex with the Doc of a family 5 glycoside hydrolase (*Ac*DocCel5). Established co-expression strategies for the production and purification of Coh-Doc complexes (Carvalho *et al.*, 2003) allowed generation of sufficient amount of highly pure protein complexes that gave good quality crystals.

6.3.1. Expression and Crystallization of *A. cellulolyticus* Coh-Doc Complexes

Analysis of the *Ac*DocCel5 sequence revealed a high degree of internal symmetry, which suggested that this Doc contained two identical Coh binding interfaces. Since a dual binding mode implies that two different complex conformations will be present in solution, this will likely compromise crystallization. It is well established that in type I Docs residues at positions 11 and 12 of each one of the two duplicated segments present a key role in Coh recognition and act as specificity determinants (Bayer *et al.*, 2004). Thus, to force a single binding mode and therefore promote homogeneity in the final product, two Doc mutants were created. *Ac*DocCel5 mutations used for the crystallization experiments were designed to replace the putative recognition residues in relative positions 11 and 12 (Ser-15/Ile-16 and Ser-51/Leu52) with those of the ScaB Doc (Ile-Asn), rather than the commonly applied alanine substitution. These amino

acid changes were performed based on previous data that revealed a lack of cross-reaction between these two type I Coh-Doc complexes (Hamberg *et al.*, 2014). The sequence of the resulting Docs is displayed in Table 6.1. Preliminary experiments evaluated the levels of expression of the *Ac*CohScaA6-DocCel5 M1 and *Ac*CohScaA6-DocCel5 M2 complexes, wherein the histidine tag was located either on the N-terminus of the Doc or the C-terminus of the Coh. The data (not shown) revealed higher protein yield with Doc tagged complexes. Thus, the recombinant plasmids encoding the Doc tagged complexes were selected to produce highly pure Coh-Doc complexes for crystallization. Both *Ac*CohScaA6-DocCel5 variants (M1 and M2) resulted in high quality crystals.

6.3.2. Structure of a novel *A. cellulolyticus* Coh-Doc complex

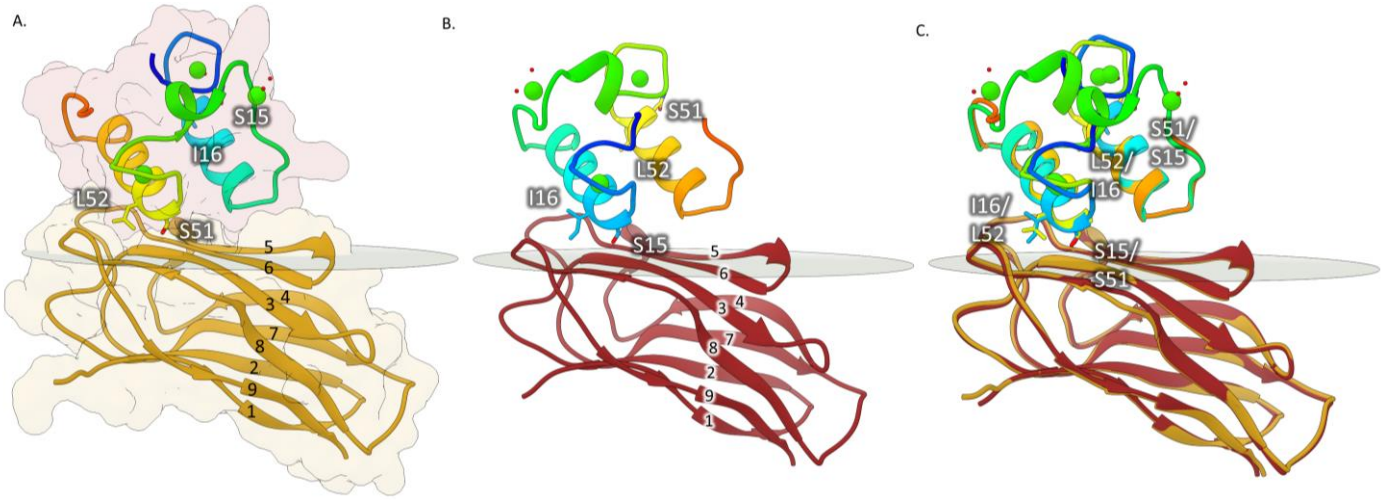
*Ac*CohScaA6-DocCel5 M1 and *Ac*CohScaA6-DocCel5 M2 structures were solved by molecular replacement, as described in the experimental procedures section (Figure 6.2). The best *Ac*CohScaA6-DocCel5 M1 crystal belonged to space group $P2_12_12_1$ with unit cell dimensions of a = 46.5 Å, b = 79.8 Å and c = 112.2 Å. The structure included 2 molecules of the Coh-Doc heterodimer in the asymmetric unit, with each Doc coordinating 3 calcium ions, as well as 3 thiocyanate, 2 glycerol and 580 water molecules. The dimer results mainly from several interactions between both Docs (2 H-bonds and 47 non-bonded contacts), but also from interactions between the Doc of one heterodimer with the Coh of the second heterodimer and vice-versa. A total of 10 H-bonds and 78 non-bonded contacts contribute to the dimerization. The biological relevance of these crystallographic interactions is currently unknown. In contrast, the *Ac*CohScaA6-DocCel5 M2 crystal structure consisted of one heterodimer together with 3 Doc coordinated calciums and 231 water molecules. The *Ac*CohScaA6-DocCel5 M2 crystal belonged to space group $P2_1$ with unit cell dimensions of a = 30.5 Å, b = 59.9 Å and c = 51.3 Å. In both structures, the complex displayed an elongated comma shape with overall dimensions of approximately 61 x 30 x 32 Å. Data collection and refinement statistics are given in Table 6.2.

6.3.3. Structure of *Ac*CohScaA6 in complex with *Ac*DocCel5

*Ac*CohScaA6 type I Coh in complex with its cognate Doc presents an elliptical structure comprising two β-sheets aligned in an elongated β-sandwich in a classic jellyroll fold. The two sheets are composed of β-strands 9,1,2,7 and 4 on one face and β-strands 8, 3, 6, 5 on the other face. Strands 1 and 9 align parallel to each other, thus completing the jelly-roll, while the other β-strands are antiparallel (Figure 6.2). β-strand 8 is interrupted by a small β-hairpin which spans

residues Gly-118 to Pro-120 and there is a small α-helix just before β-strand 5. The two closest functionally relevant structural homologues to *Ac*CohScaA6, based on a structural similarity search using the PDBeFold server (http://www.ebi.ac.uk/msd-srv/ssm/), are the type I Cohs from *C. thermocellum Ct*CohOlpA (PDB entry 3UL4, with a Z-score of 12.9, r.m.s.d. of 1.1 Å and sequence identity of 27% over 138 aligned residues) and *Ct*CohScaA (PDB entry 1AOH, with a Z-score of 12.1, r.m.s.d. of 1.3 Å and sequence identity of 30% over 138 aligned residues). Other structural homologues include the type I Cohs from *C. cellulolyticum* (PDB entry 2VN5) and *Pseudobacteroides cellulosolvens* (PDB entry 4UMS).

Figure 6.2 Structures of the *A. cellulolyticus* cohesin-dockerin complexes.



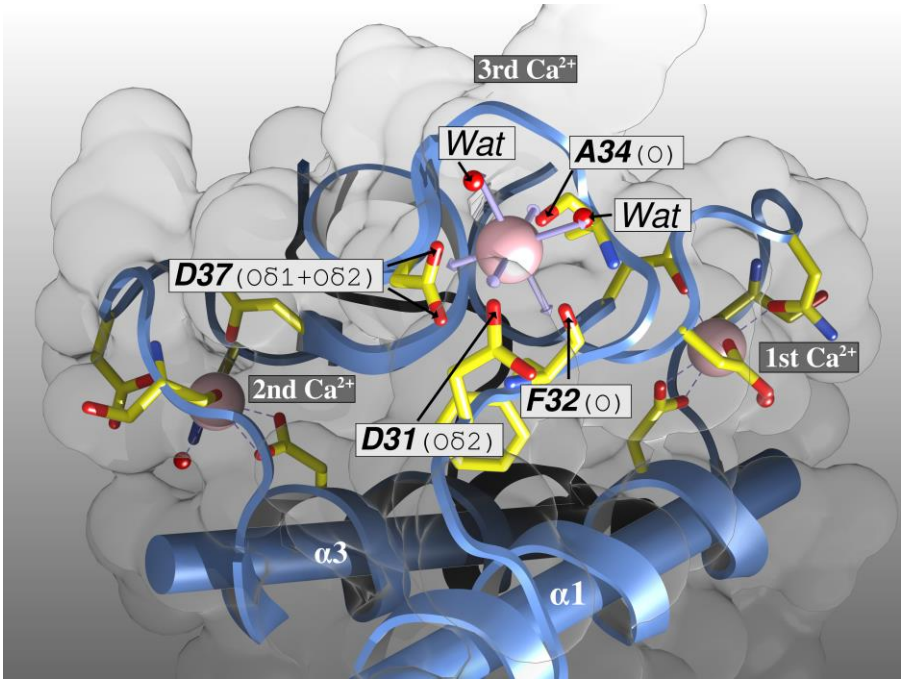
A. Structure of *Ac*CohScaA6-DocCel5 M1 with the Doc color-ramped from N-terminus (blue) to C terminus (red) and the Coh in gold. Ser-51 and Leu-52 that dominate Coh recognition and engineered residues Ile-15 and Asn-16, to force a single binding mode, are labeled and shown as stick configuration. Ca^{2+} ions are depicted as purple spheres. B. Structure of *Ac*CohScaA6-DocCel5 M2 with the Doc color-ramped from N terminus (blue) to C terminus (red) and the Coh in burgundy. Ser-15 and Ile-16 that dominate Coh recognition and engineered residues Ile-51 and Asn-52, to force a single binding mode, are again labeled and shown as stick representations. C. Overlay of the two binding modes showing the high degree of overall similarity reflecting the internal 2-fold symmetry of the Doc module. The transparent gray disk marks the plane defined by the 8-3-6-5 β-sheet, where the β-strands form a distinctive Doc-interacting plateau.. A. also depicts a representation of the molecular surface contour of the Coh and Doc, respectively. Ca^{2+} ions are depicted as green spheres.

6.3.4. Structure of *Ac*DocCel5 in complex with *Ac*CohScaA6

In both complexes, the *Ac*DocCel5 Doc displays an identical structure that comprises two α-helices arranged in an antiparallel orientation ranging from residue Ile-15 to Leu-25 (helix-1) and from Ser-51 to Leu-61 (helix-3), respectively. These two helices comprise portions of the two classic Doc repeating segments, each containing a bound calcium ion in loops located at opposite ends of the module. The loop connecting these secondary structures contains a six-residue α-helix extending from Asp-37 to Gly-41 (helix-2). The overall structure of *A.*

cellulolyticus DocCel5 is very similar *C. thermocellum* type I Doc Doc435 (PDB entry 4DH2, with a Z-score of 8.5, r.m.s.d. of 0.5 Å and sequence identity of 46% over 68 aligned residues), but it also shares high homology with the type I Doc from *A. cellulolyticus* adaptor scaffoldin ScaB (PDB entry 4UYP, with a Z-score of 6.5, r.m.s.d. of 1.2 Å and sequence identity of 44% over 68 aligned residues), with whom it does not share any known ligand specificity. The Ca^{2+} ion located at the Doc N terminus is coordinated by the side chains of residues Asp-6, Asp-8, Asn-10, and Asp-17 (both the Oδ1 and Oδ2), the latter belonging to the N-terminal α-helix (helix-1) of this module. The octahedral geometry of the coordination of this Ca^{2+} ion is fulfilled by the main chain carbonyl of Ser-12 and by a water molecule. The second Ca^{2+} site stabilizes the loop connecting the internal and C-terminal α-helix (helices 2 and 3) of the Doc module. This Ca^{2+} ion is coordinated by the side chains of residues Asp-42, Asn-44, Asp-46, and Asp-53 (both the Oδ1 and Oδ2), as well as by the carbonyl of Ser-48, with the octahedral geometry also completed by a water molecule. Thus, both Ca^{2+} sites show the n, n+2, n+4, n+6 (main-chain O atom), n + 11 and a water molecule completing the coordination pattern. Interestingly, there is a third calcium atom bound to *Ac*DocCel5 which is coordinated by a loop between helix-1 and helix-2. Although this calcium does not appear to be relevant to the Doc function, as it is located in a distant position relative to the Coh, it is consistently present bound to both Docs of the *Ac*CohScaA6-DocCel5 M1 complex dimer structure and also to the Doc of the *Ac*CohScaA6-DocCel5 M2 structure (Figure 6.3). Thus, this unusual third Ca^{2+}, not previously observed in Coh-Doc complexes, seems to display a stabilizing role and presents the typical octahedral geometry coordination through the side chains of Asp-31 and Asp-37, the main chain carbonyl O atoms of Phe-32 and Ala-34 and by two water molecules (Figure 6.3).

Figure 6.3 Octahedral coordination of the third calcium from *A. cellulolyticus* AcDocCel5



The calcium ions are labelled and depicted as pink spheres and the amino-acid residues involved in the metal coordination are shown as sticks. The third calcium is overlaid with an idealized geometry representation (purple arrows) and the residues are labelled, including the contribution of 2 water molecules (Wat). The secondary structure of *Ac*DocCel5 is colored in blue, with axis-aligned cylinders along helices -1 and -3 (labelled) while the molecular surface is drawn in transparent grey. Figure prepared with the UCSF Chimera program (Pettersen *et al.*, 2004).

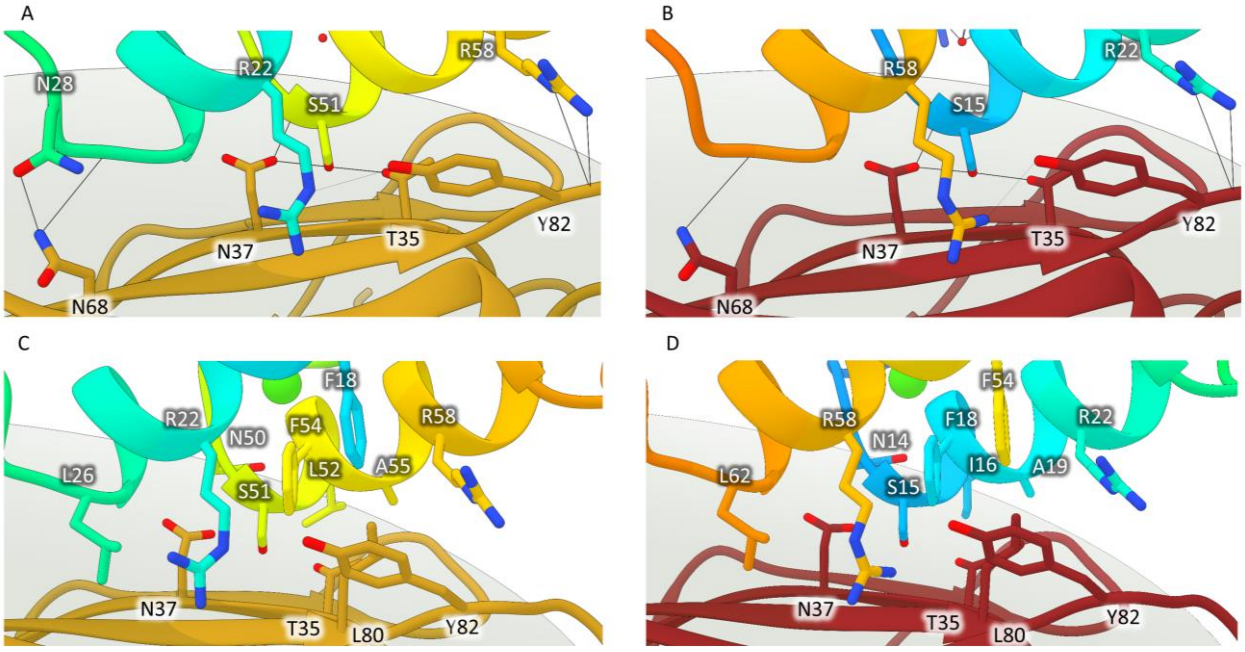
6.3.5. *A. cellulolyticus* type I CohScaA6-DocCel5 M1 and CohScaA6-DocCel5 M2 Interfaces

In the two *Ac*CohScaA6-DocCel5 complexes solved here, *Ac*DocCel5 interacts with the 8-3-5-6 *β*-sheet of the *Ac*CohScaA6 *β*-sandwich via helices 1 and 3. The Doc contacting surface of *Ac*CohScaA6 presents a predominantly flat rectangular shape, whose angles are slightly elevated towards the Doc and correspond to the loops between *β*-strands 4 and 5, 5 and 6, and 8 and 9 and the *β*-hairpin that interrupts *β*-strand 8. In the *Ac*CohScaA6-DocCel5 M1 structure, helix-3 dominates the Doc's interaction with the Coh. Contacts are established by the entire length of helix-3, while only the C-terminal portion of helix-1 interacts with the Coh. In contrast, in the *Ac*CohScaA6-DocCel5 M2 the exact opposite happens: Coh contacts are established by the entire length of helix-1 and the C-terminal portion of helix-3, in a helix-1 dominated interaction. The structures of *Ac*CohScaA6-DocCel5 M1 and M2 were found to be very similar to each other, with a backbone r.m.s.d. of 0.5 Å (Figure 6.2). Furthermore, helix-1 and helix-3 of *Ac*DocCel5 M1 overlapped almost perfectly with helix-3 and helix-1 of *Ac*DocCel5 M2, respectively, as a result of a 180º rotation in relation to the Coh, imposed by the symmetrically-related opposite mutations (Figure 6.2). In contrast, helix-2 that bridges the

duplicated segments has two distinct spacial positions when both structures are overlaid. This suggests that the Doc internal structural symmetry supports the Coh recognition through two highly similar binding interfaces. This dual binding mode, resulting from a near-perfect 2-fold internal structural symmetry, has been extensively described in a variety of type I Coh-Doc complexes (Cameron, Najmudin, *et al.*, 2015; Carvalho *et al.*, 2003, 2007), including the *A. cellulolyticus* CohScaC3-DocScaB (Cameron, Najmudin, *et al.*, 2015; Carvalho *et al.*, 2003, 2007).

The intermolecular interfaces include several hydrogen bonds (Table 6.3, Figure 6.4A, B) and also a large network of hydrophobic interactions that play a key role in *Ac*CohScaA6-DocCel5 M1 and M2 complex assembly (Table 6.4, Figure 6.4C, D). The DocCel5 residues at the complex interface located in helices 1 and 3 remain practically unchanged upon the 180° rotation of the Doc module over the CohScaA6 surface, reflecting the internal symmetry of the ScaB Doc (Figure 6.4 & 6.5).

Figure 6.4 Cohesin-dockerin interface of *Ac*CohScaA6-DocCel5 M1 and *Ac*CohScaA6-DocCel5 M2.



Structure of *Ac*CohScaA6-DocCel5 M1 (gold cohesin) and *Ac*CohScaA6-DocCel5 M2 (burgundy cohesin) complexes with a detailed view of the Coh-Doc interface showing the main polar interactions (Panel A, C) and main hydrophobic contacts (Panel B, D). In all panels the most important residues involved in Coh-Doc recognition are depicted in stick configuration, with a dark background label for the Doc residues and a light background label for the Coh residues, using the *Ac*DocCel5 and *Ac*CohScaA6 numbering. Solid black lines mark hydrogen-bonds interactions. The Docs are shown color-ramped from N-terminus (blue) to C terminus (red). Ca^{2+} ions are depicted as green spheres. In all panels, the transparent grey disk marks the plane defined by the 8-3-6-5 β-sheet, where the β-strands form a distinctive Doc-interacting plateau.

Table 6.3 Main polar contacts between *Ac*CohScaB6 and both *Ac*DocCel5 mutants.

DocCel5 M1					CohScaB6				DocCel5 M2			
	Atom	Residue	Residue #		Atom	Residue	Residue #		Atom	Residue	Residue #	
Hydrogen Bonds												
H1	NE	Arg	22	< >	OH	Tyr	82	< >	NH1	Arg	58	H3
H1	O	Leu	26	< >	ND2	Asn	68	< >	O	Leu	62	H3
	OD1	Asn	28	< >	ND2	Asn	68					
H3	OG	Ser	51	< >	OG1	Thr	35	< >	OG	Ser	15	H1
H3	N	Ser	51	< >	OD1	Asp	37	< >	N	Ser	15	H1
H3	OG	Ser	51	< >	OD1	Asp	37	< >	OG	Ser	15	H1
H3	NE	Arg	58	< >	O	Tyr	82	< >	NE	Arg	22	H1
H3	NH2	Arg	58	< >	O	Tyr	82	< >	NH2	Arg	22	H1

Table was made using the PDBePISA server. Dockerin residues are marked as belonging either to helix 1 (H1) or to helix 3 (H3) interfaces.

Table 6.4 Main hydrophobic contacts between *Ac*CohScaB6 and *Ac*DocCel5.

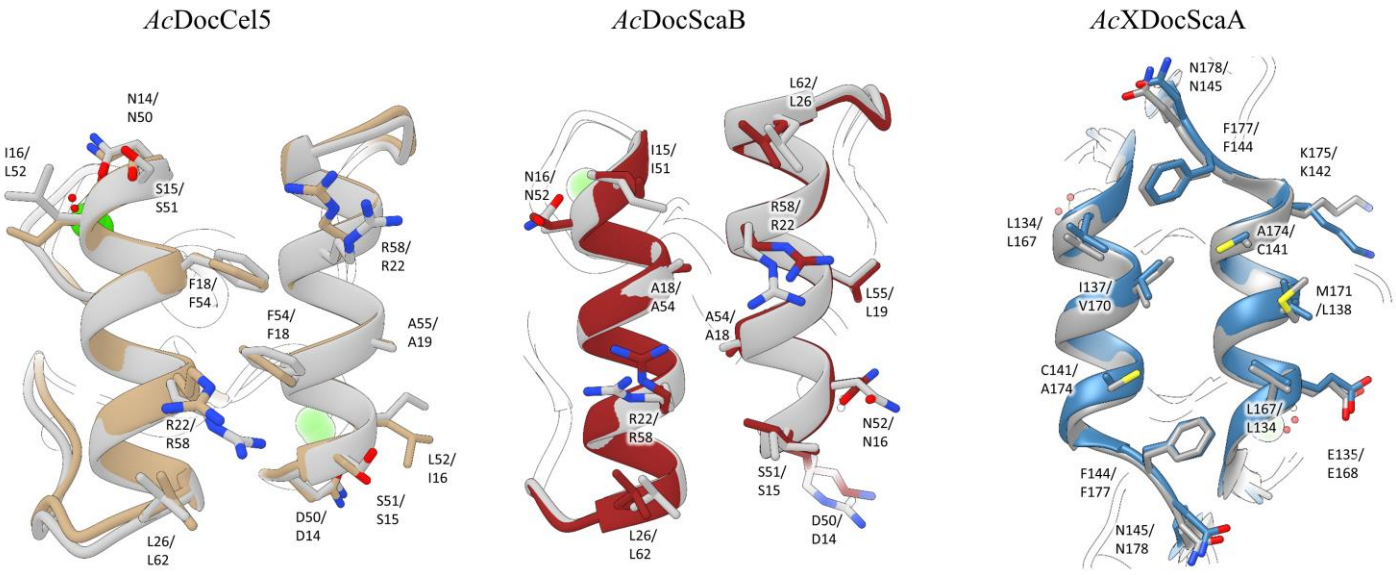
DocCel5 M1				CohScaB6
	Residue	Residue #		Residues
H1	Phe	18	< >	Tyr82 (5)
H1	Arg	22	< >	Phe65 (2), Gly66 (2), Leu78, Tyr82 (2),
H1	Leu	25	< >	Lys76 (2)
H1	Leu	26	< >	Gly66, Ile87, Asn68 (3), Lys76 (6)
	Gly	27	< >	Glu70 (2)
	Asn	28	< >	Asn68 (4)
	Asn	50	< >	Asp37 (3), Thr121 (3)
H3	Ser	51	< >	Thr35 (9), Ala36 (4), Asp37(9)
H3	Leu	52	< >	Thr35 (2), Gly123
H3	Phe	54	< >	Leu78 (3), Tyr82
H3	Ala	55	< >	The35, Leu80 (2), Leu127
H3	Arg	58	< >	Tyr82 (10)
H3	Met	59	< >	Met84 (2), Leu127
H3	Leu	62	< >	Tyr82
H3	Met	64	< >	Thr83 (2), Met84
DocCel5 M2				**CohScaB6**
	Residue	Residue #		Residues
H1	Asn	14	< >	Asp37 (4), Thr121 (5)
H1	Ser	15	< >	Thr35 (5), Ala36 (7), Asp37 (8), Thr121
H1	Ile	16	< >	Thr5, Thr121, Pro129
H1	Phe	18	< >	Leu78, Leu80 (3), Tyr82
H1	Ala	19	< >	Thr35, Leu80 (2), Met84, Leu127
H1	Arg	22	< >	Leu80, Tyr82 (10), Met84
H1	Asn	23	< >	Met84 (2), Leu127
H1	Leu	26	< >	Tyr82, Thr83
H3	Phe	54	< >	Leu80 (2), Tyr82 (6)
H3	Arg	58	< >	Val63, Phe65 (2), Gly66 (2), Leu78, Tyr82 (6)
H3	Leu	61	< >	Lys76 (2), Leu78
H3	Leu	62	< >	Gly66 (2), Asn68 (3), Lys76 (6), Leu78
H3	Gly	63	< >	Glu70 (2), Lys76
H3	Met	64	< >	Asn68

Table was made using the PDBePISA server. Some of the dockerin residues are marked as belonging either to helix 1 (H1) or to helix 3 (H3) interfaces.

Therefore, the interactions between the dominant Doc helix and the Coh are mainly established by residues Asn-50/14, Ser-51/15, Leu-52/Ile-16, Phe-54/18, Ala-55/19 and Arg-58/22 (on the

*Ac*CohScaA6-DocCel5 M1/M2 structures, respectively) while the main contacting residues in the non-dominant helix are Phe-18/54, Arg22/58 and Leu-26/62 (again from M1/M2 structures, respectively). The side-chains of residues Phe-18/54, Leu-26/62, Leu-52/16, Phe-54/18 and Ala-55/19 dominate the hydrophobic recognition of the Coh. Phe-18/54, Phe-54/18 and Ala-55/19 are at the core of the interaction where, together with the aliphatic regions of Arg58/22 and Ser-51/15, establish numerous non-bonded interactions with β-sheets 3 and 6 of *Ac*CohScaA6 including with the side-chains of residues Thr-35, Leu-80 and Tyr-82. Leu-26/62 and Leu-52/16 are located in opposing extremities of the interaction interface. Leu-26/62 together with the aliphatic region of Arg-22/58 contact mainly with β-sheet 5 of *Ac*CohScaA6, while Leu-52/16 together with the aliphatic region of Asn-50/14 establishes several non-bonded contacts with β-sheet 8 of *Ac*CohScaA6. The hydrogen bond network established by the dominant helix of *Ac*DocCel5 is supported by the interactions between the Oγ atom of Ser-51/15 and *Ac*CohScaA6 residues Thr-35 and Asn-37, between the N atom of Ser-51/15 and *Ac*CohScaA6 Asn-37 and between Arg-58/22 (both Nε and NH2 atoms) and Tyr-82 of *Ac*CohScaA6 (Table 6.3). In the non-dominant interacting helix, the hydrogen bonds are established between *Ac*DocCel5 Arg-22/58 and Leu-26/62 and *Ac*CohScaA6 Tyr-82 and Asn-68, respectively. In the *Ac*CohScaA6-DocCel5 M1 structure, one extra H-bond was observed between Doc Asn-28 and Coh Asn-68 (Table 6.3).

Figure 6.5 Symmetric nature of *A. cellulolyticus* dockerins exemplified by structures of different specificities.



From left to right: *Ac*DocCel5 (brown), *Ac*DocScaB (green) and *Ac*XDocScaA (blue) structures overlayed with a 180º rotated version of themselves, showing conservation of key Coh interacting residues. This suggests that *A. cellulolyticus* cellulosome is assembled exclusively *via* dual binding mode Coh.-Doc interactions, therefore having a highly dynamic architecture.

6.3.6. Thermodynamics of the dual binding mode

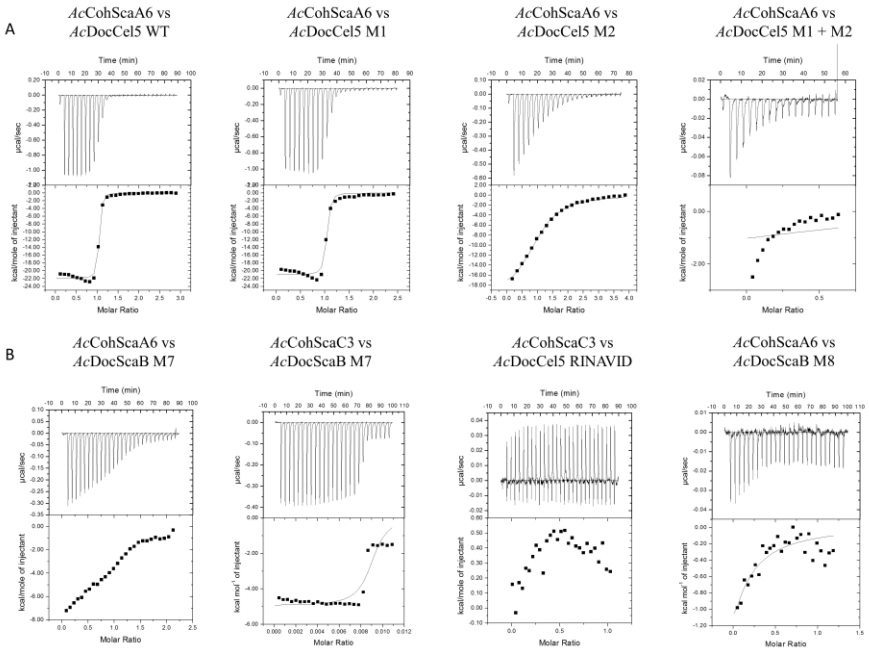
Previous studies revealed that type I complexes of other cellulosome systems that display a dual-binding mode, such as of *C. thermocellum* or *C. cellulolyticum*, have no preference for a particular binding orientation (Carvalho *et al.*, 2007; Pinheiro *et al.*, 2008). Thus, affinity between Cohs and Docs is similar whether the Doc module binds its protein partner *via* the N-terminal or the C-terminal helix. To establish if a similar mechanism operates during *Ac*CohScaA6-DocCel5 recognition, the binding thermodynamics between *Ac*CohScaA6 and the wild type, M1, M2 and M1+M2 variants of *Ac*DocCel5 were determined using isothermal titration calorimetry (ITC). The data, presented in Table 6.5 and exemplified in Figure 6.6, revealed a macromolecular association with a 1:1 stoichiometry and a K_a of ~ 10^8 M^{-1}, an affinity similar to other type I Coh-Doc interactions (Pinheiro *et al.*, 2008). This stoichiometry suggests that one Coh can only bind one Doc, suggesting that the heterodimer observed in the structure of *Ac*CohScaA6-DocCel5 M1, which suggested that a second Doc could eventually bind to previously formed complex, does not represents a biological functional possibility. As expected, the *Ac*DocCel5 M1+M2 mutant, in which both N-terminal and C-terminal residues at positions 11 and 12 were substituted, did not bind *Ac*CohScaA6. Interestingly, both M1 and M2 mutations resulted in a decreased affinity for *Ac*CohScaA6. While in *Ac*CohScaA6-DocCel5 M1 interaction affinity decreased by only 0.6 times when compared with that of the wild type complex, the K_a value registered for the Coh interaction with *Ac*DocCel5 M2 mutant is over 160 times lower than the one obtained with the wild type Doc. Even though the binding interface of both M1 and M2 mutants is virtually identical, the subtle differences in the interface observed in the two protein complexes may result in relatively weaker contribution of non-bonded contacts when helix-1 dominates the interaction (87 non-bonded contacts in total *versus* 99 for the *Ac*DocCel5 M1 mutant). Alternatively, the fact that *Ac*DocCel5 is fused to an unrelated protein module to provide additional stability may lead to a steric effect of the protein partner and thus justify the observed decrease in affinity. It is rather unlikely that there is a preferential binding orientation for the *Ac*CohScaA6-DocCel5 interaction, favoring the conformation in which the Docs N-terminal α-helix dominates Coh recognition. In addition, these observations suggest that there is no truncation on the Doc sequence that could justify a difference in affinity as the N-terminal binding mode seems to be favored versus the C-terminal interaction.

Table 6.5 Thermodynamics of interaction between wild type *Ac*CohScaA6 and *Ac*CohScaC3, and various variants of *Ac*DocGh5 and *Ac*DocScaB.

***Ac*CohScaA6**					
Dockerin	$K_a\ M^{-1}$	ΔG^o kcal mol^{-1}	ΔH kcal mol^{-1}	$-T\Delta S^o$ kcal mol^{-1}	*N*
***Ac*DocGh5 WT**	*1.12E8 ± 3.17E7*	*-11.35*	*-21.91 ± 0.20*	*10.56*	*1.00*
***Ac*DocGh5M1 (S15I I16N)**	*6.72E7 ± 2.75E7*	*-11.04*	*-21.02 ± 0.52*	*9.97*	*1.00*
***Ac*DocGh5M2 (S51I L52N)**	*6.89E5 ± 4.89E4*	*-8.31*	*-20.63 ± 0.52*	*12.32*	*1.00*
***Ac*DocGh5M1+M2**	*Nb*	*Nb*	*Nb*	*Nb*	*Nb*
***Ac*DocGh5 RINAVID**	*2.04E6 ± 4.77E5*	*-9.12*	*-21.75± 2.59*	*12.62*	*0.97*
***Ac*DocScaB WT**	*Nb*	*Nb*	*Nb*	*Nb*	*Nb*
***Ac*DocScaB M7**	*4.18E5 ± 4.73E4*	*-7.92*	*-8.17 ± 0.25*	*0.25*	*0.99*
***Ac*DocScaB M8**	*Nb*	*Nb*	*Nb*	*Nb*	*Nb*
AcCohScaC3					
Dockerin	$K_a\ M^{-1}$	ΔG^o kcal mol^{-1}	ΔH kcal mol^{-1}	$-T\Delta S^o$ kcal mol^{-1}	*N*
***Ac*DocGh5WT**	*Nb*	*Nb*	*Nb*	*Nb*	*Nb*
***Ac*DocGh5M1 (S15I I16N)**	*Nb*	*Nb*	*Nb*	*Nb*	*Nb*
***Ac*DocGh5M2 (S51I L52N)**	*Nb*	*Nb*	*Nb*	*Nb*	*Nb*
***Ac*DocGh5M1+M2**	*Nb*	*Nb*	*Nb*	*Nb*	*Nb*
***Ac*DocGh5 RINAVID**	*Nb*	*Nb*	*Nb*	*Nb*	*Nb*
***Ac*DocScaB WT**	*Nd*	*Nd*	*Nd*	*Nd*	*Nd*
***Ac*DocScaB M7**	*3.63E6 ± 1.68E6*	*-9.24*	*-4.937 ± 0.14*	*-4.31*	*0.89*
***Ac*DocScaB M8**	*Nb*	*Nb*	*Nb*	*Nb*	*Nb*

All Thermodynamic parameters were determined at 308 K. *Nb* – No binding. *Nd* – Affinitty too high to accurately determine thermodynamic parameters.

Figure 6.6 Binding affinity of *Ac*CohScaA6 and *Ac*CohScaC3 to *Ac*DocCel5 and *Ac*DocScaB wild type and mutant derivatives as determined by ITC.



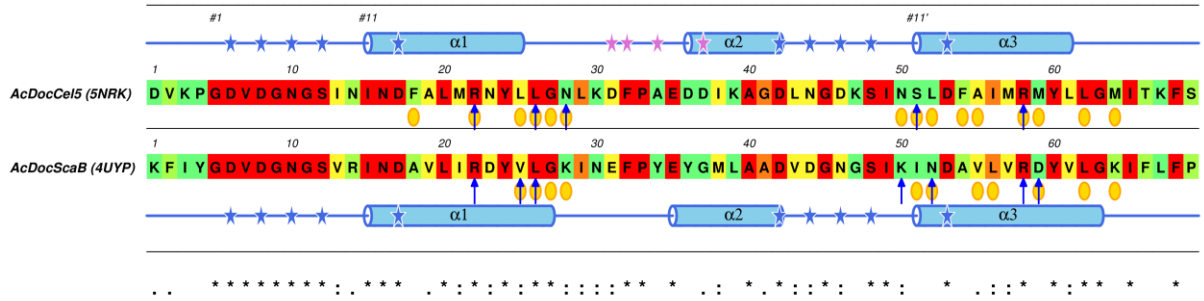
Example binding isotherms for: Panel A, *Ac*CohScaA6 vs *Ac*DocCel5 specificity determinant mutants for the dual binding mode evaluation; Panel B, the specificity exchange experiments. The upper part of each panel shows the raw heats of binding, whereas the lower parts comprise the integrated heats after correction for heat of dilution. The curve represents the best fit to a single-site binding model. The corresponding thermodynamic parameters are shown in Table 6.5.

6.3.7. Developing a specificity hybrid *A. cellulolyticus* type I Doc

Overall, the structure and mode of interaction of the *Ac*CohScaA6-DocCel5 complex are very similar to those of the previously characterized *Ac*CohScaC3-DocScaB complex that displays a different specificity (Cameron, Najmudin, *et al.*, 2015). Both *Ac*DocCel5 and *Ac*DocScaB possess the ability to bind their Coh partners in two different orientations, resulting in Coh-Doc complex configurations that are highly superposable (r.m.s.d. of 1.1 Å and 1.0 Å for the helix-1 dominated interaction and the helix-3 dominated interactions, respectively) (Figure 6.5). Crucial interacting residues are generally located at the same relative positions and, especially at the N-terminal Doc repeats, there is a high degree of conservation between *Ac*DocCel5 and *Ac*DocScaB. Thus, even some key interacting residues such as Arg-22, Leu-26 and Arg-58 are conserved between the two Docs. In spite of those resemblances, *Ac*DocScaB displays a distinct Coh specificity when compared with *Ac*DocCel5, whose binding properties should represent those of the remaining *A. cellulolyticus* type I Docs that recruit enzymes to the cellulosome.

Considering the similarities between *A. cellulolyticus* type I complexes of distinct specificities, an attempt to alter the ability of type I Docs to recognize a specific Coh was carried. The aim was to create an *Ac*DocScaB mutant capable of recognizing ScaA Cohs, an *Ac*DocCel5 mutant capable of binding to ScaC Cohs and an hybrid type I Doc that accumulated both specificities *via* its 2 distinct Coh binding interfaces. A structural alignment between *Ac*DocCel5 and *Ac*DocScaB (Figure 6.7) revealed the main divergent residues between both Docs first repeats and was used to design an *Ac*DocCel5 mutant where residues Asn-14, Ser-15, Ile-16, Phe-18, Ala-19, Met-21 and Asn-23 were replaced by *Ac*DocScaB residues Arg-14, Ile-15, Asn-16, Ala-18, Val-19, Ile-21 and Asp-23, generating *Ac*DocCel5 RINAVID mutant.

Figure 6.7 Multiple sequence alignment of *Ac*DocCel5 and *Ac*DocScaB in a C-terminus (helix-3) dominated Coh-Doc interface

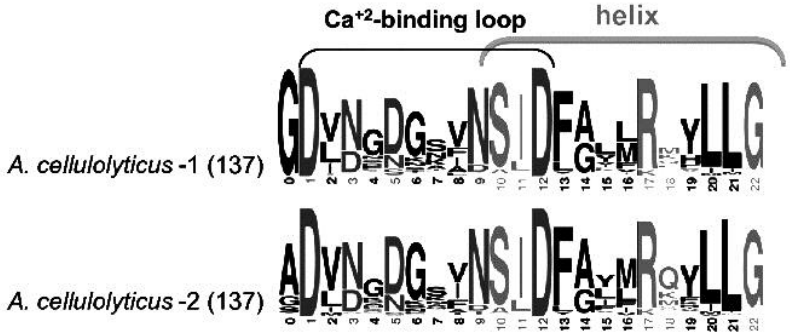


Based on the AcCohScaB6-DocCel5 M1 complex (PDB code: 5NRK, on top) and AcCohScaC3-DocScaB (PDB code: 4UYP, on the bottom), there is a cartoon representation of the secondary structure. The residues involved in the molecular interactions with the respective Coh partner are highlighted as follows: blue arrow for polar contacts, and yellow circles for hydrophobic contacts. While the residues involved in the coordination with the first and second calcium ions are marked with a blue star, and

those implicated on binding the third calcium (AcDocCel5) are shown with a purple star. Bellow the alignment, the ClustalO consensus symbols represent the position conservation status. The primary sequence background is colored according to the ALSCRIPT Calcons convention, implemented in ALINE (Bond & Schüttelkopf, 2009): red, identical residues; orange to blue, lowering color-ramped scale of conservation.

For the *Ac*DocScaB mutants, instead of directly replacing the key residues with those of *Ac*DocCel5, a consensus sequence based on 137 *A. cellulolyticus* Doc sequences (that also bind to *Ac*CohScaA6) was used (Figure 6.8) (Dassa *et al.*, 2012).

Figure 6.8 Sequence conservation pattern of type I dockerin modules.

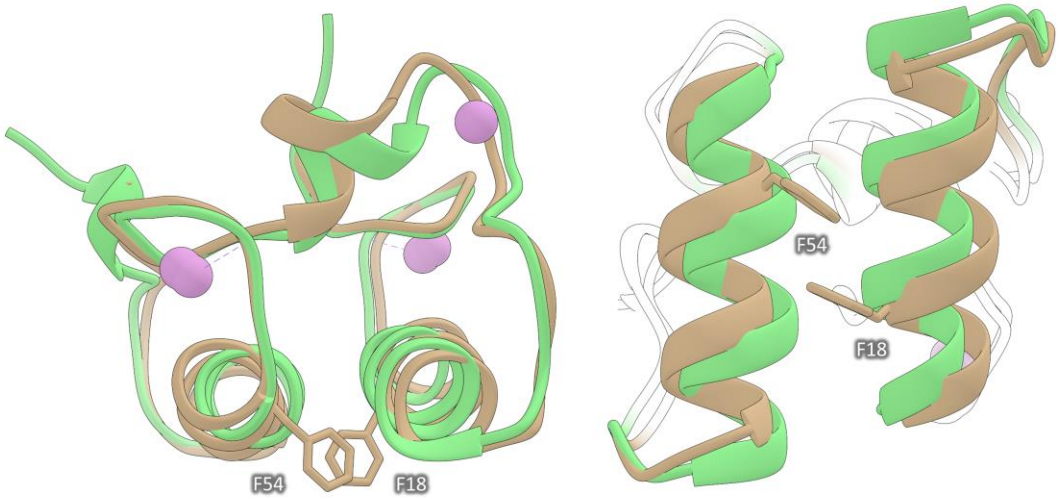


The two internal Doc repeats of *A. cellulolyticus* (based on 137 sequences) are represented by sequence logos. The bigger the letter the more conserved that residue is. (Adapted from Dassa *et al* (Dassa *et al.*, 2012)).

Residues Arg-14, Ile-15, Asn-16, Ala-18, Leu-20, and Asp-23 were thus replaced on *Ac*DocScaB by Asn-14, Ser-15, Ile-16, Phe-18, Tyr-20, and Gln-23 of the type I Doc consensus. This generated *Ac*DocScaB M7, whose sequence is displayed in Table 6.1. By duplicating these substitutions in *Ac*DocScaB second repeat, therefore adding mutations K50N, I51S, N52L, A54F, L56Y and D59Q to M7, mutant M8 was generated (Table 6.1). It was predicted that *Ac*DocCel5 RINAVID and *Ac*DocScaB M7 would be able to recognize both *Ac*CohScaA6 and *Ac*CohScaC3 and that *Ac*DocScaB M8 would completely switch specificity and only be able to bind *Ac*CohScaA6. The ability for these Doc derivatives to bind the two different Coh counterparts was initially probed by non-denaturing gel electrophoresis (NGE) (data not shown). Data suggested that *Ac*DocCel5 RINAVID could still recognize *Ac*CohScaA6 while it did not acquire the ability to bind *Ac*CohScaC3, but that *Ac*DocScaB M7 could indeed recognize both Cohs. Interestingly, based on the NGE analysis, *Ac*DocScaB M8 did not seem to be able to bind any of the Cohs (data not shown). In order to confirm these results and further explore the thermodynamics of these interactions, ITC was carried out at 308K. The data, presented in Table 6.5 and exemplified in Figure 6.6, confirm the results suggested by the NGE

analysis. *Ac*DocScaB RINAVID mutant can still bind to *Ac*CohScaA6 with an affinity constant of 2.04E6 while it failed to bind *Ac*CohScaC3. On the other hand, *Ac*DocScaB M7 binds to *Ac*CohScaA6 with an affinity constant of 4.18E5 and to *Ac*CohScaC3 with an affinity constant of 3.63E6, meaning that the attempt to create a specificity hybrid type I Doc was successful. In contrast, *Ac*DocScaB M8 did not show affinity for *Ac*CohScaA6. After close inspection of both *Ac*DocScaB and *Ac*DocCel5 Doc structures it becomes apparent that the gap between the N-terminal and C-terminal helices backbone chains is narrower in *Ac*DocScaB, with approximately 4.7 Å at its narrowest point *versus* the 6.5 Å in *Ac*DocCel5 (Figure 6.9). The fact that helices 1 and 3 are interacting closely in *Ac*DocScaB might not allow enough space to accommodate the mutations on both Doc repeats without steric clashes, especially between Phe-18 and Phe-54 (Figure 6.9). Therefore, *Ac*DocScaB M8 might have had its structural integrity compromised by such clashes, which explains why this Doc derivative was unable to bind to *Ac*CohScaA6.

Figure 6.9 Overlay of *Ac*DocCel5 and *Ac*DocScaB.



Structures of *Ac*DocCel5 (brown) and *Ac*DocScaB (light green) are overlaid showing the size difference in the gap between the Coh contacting helices. *Ac*DocCel5 residues Phe-18 and Phe-54 are highlighted.

6.4. Conclusions

It is now well established that type-I Coh-Doc interactions are essential to recruit cellulosomal enzymes onto primary scaffoldins, which in turn is attached to the cell surface via a type-II Coh-Doc pair. In *A. cellulolyticus*, a second type I Coh-Doc specificity is responsible for the

attachment of an unusual adaptor scaffoldin ScaB to the bacterial cell surface. Previous work revealed that the type-I Coh-Doc complexes that recruit ScaB to the cell envelop presents a dual binding mode resulting from the presence of two identical Coh-binding faces as it is characteristic of the majority of cellulosomal type-I Coh–Doc complexes. Here, we reveal that the type I Coh-Doc complexes that recruit enzymes to the cellulosome of *A. cellulolyticus* also present a dual-binding mode suggesting that flexibility in Coh recognition seems to be a general feature of type-I Doc modules, including those that recruit cellulosomes into the cell surface. The structure of *Ac*DocCel5 revealed an internal symmetry that supports the presence of two virtually identical Coh-binding faces. Due to the high degree of homology shared by the two different type I Coh-Doc specificities discovered in *A. cellulolyticus*, an engineered Doc with the capacity to bind two different Coh, using the two different binding surfaces was produced. Thus, this work exemplifies how structural studies can inform protein engineering to design novel Doc with different specificities. Since the Coh-Doc platform represents a very useful target to engineer novel nano-machines for a variety of applied processes that might benefit from enzyme proximity, these results reveal how protein engineering can be used to produce novel specificities for cellulosome construction.

Chapter 7

General discussion and future perspectives

Since the early 1980's, when the cellulosome was first described by Raffi Lamed and Ed Bayer (Lamed *et al.*, 1983), a tremendous amount of knowledge has been gathered regarding these highly specialized nanomachines. Their definition has changed from simply cellulose degrading structures to "multi-enzyme complexes produced by anaerobic bacteria for the efficient deconstruction of plant cell wall polysaccharides" (Smith & Bayer, 2013). Recent biochemical characterisation of cellulosomal components and accumulated genomic and metagenomic information has confirmed the sophistication of cellulosomes, which supports a diversity of catalytic activities such as cellulase, hemicellulase, pectate lyase and carbohydrate esterase (Bayer & Lamed, 2006; Bayer *et al.*, 1994; Lamed *et al.*, 1983; Smith & Bayer, 2013). In addition, data derived from molecular biology, bioinformatics, biochemistry and structural biology has allowed a deeper understanding of the molecular basis for cellulosome assembly and function (Bayer, Chanzy, *et al.*, 1998). However, while some questions are being answered others are still emerging. Up until recently, the architecture and organization of the cellulosome was thought to rely on two types of cohesin-dockerin interaction, namely type I and type II. While that is true for the majority of the better described cellulosomal systems, such as those of *C. thermocellum* and *C. cellulolyticum*, the development of metagenomics has allowed the identification of additional cellulosome producing bacteria with distinct organizational strategies. Certain species, such has *R. flavefaciens* and *R. champanellensis*, where found to be capable of producing highly complex cellulosomes whose assembly relies on the interaction of numerous cohesin and dockerin modules that diverge from the classical type I and II modules. These Coh-Doc complexes, collectively classified as type III, are still poorly described and present an excellent opportunity to expand our current knowledge concerning cellulosome structure and function. Understanding the intricacy of cellulosomes evolved by anaerobic microbes may assist the development of a variety of novel biotechnological applications such

as the conversion of lignocellulosic biomass to bio-ethanol or the development of new affinity based molecular biology tools. The main goal of this work was to identify and characterize, both structurally and functionally, novel type III cohesin-dockerin complexes by exploring the *R. flavefaciens* cellulosome.

Development of novel innovative molecular biology and biochemical approaches is crucial to achieve significant scientific advances in the cellulosome field. High-throughput methodologies and automation, of otherwise very time-consuming and laborious protocols, allow generating tremendous amounts of data, required to thoroughly analyse highly complex biological systems. The present work was designed to take advantage of those resources by building a large library of cohesins and dockerins from *R. flavefaciens* and implementing three different strategies to probe novel protein-protein specificities in ruminal cellulosomes. Therefore, the work described in Chapter 2 represents a medium/high-throughput approach aimed at screening novel cohesin-dockerin complex specificities by testing a sample of 1463 possible interactions (19 cohesins *versus* 77 dockerins) out of the putative 4683 possible combinations in *R. flavefaciens* proteome (21 cohesins *versus* 223 dockerins) (Rincon *et al.*, 2010). In addition to revealing novel cohesin-dockerin specificities, this approach also allowed identifying highly stable and expressible cohesin-dockerin complexes, best suited for both structural and biochemical analysis. The development of an innovative *in vivo* screening strategy involving the co-expression of both cohesin and dockerin encoding genes in the same *E. coli* cells was crucial for this. Dockerins are highly unstable and very susceptible to proteolysis when expressed individually in heterologous hosts. Co-expression of dockerins with their protein partners is believed to confer stabilization (Cameron, Najmudin, *et al.*, 2015). When this approach is used to obtain protein complexes for crystallization studies, his-tags are usually fused to cohesins, which are normally expressed at higher levels than dockerins. However, recent work by K. Cameron (data not published) suggested that integration of his-tags in the dockerin sequence might result in higher levels of expression. Thus, in this thesis, a cloning strategy that allowed testing fusion of the His-tag to either the cohesin or the dockerin modules was devised to identify the preferable approach to generate high levels of each Coh-Doc complex for protein crystallography. The application of these methodologies resulted, generally, in high yields of stable cohesin-dockerin complexes, an initial pre-requisite for obtaining good quality crystals. The development of molecular biology strategies that allow the expression of different cellulosome components in the same plasmid and under the control of different promoters could also be useful for other applications that span the crystallization of cohesin-dockerin complexes. This strategy could be used to create self-assembling mini-cellulosomes *in vivo* by cloning different cellulolytic enzymes containing their endogenous

dockerins and a mini-scaffoldin containing a series of cohesin modules, in the same vector. These mini-cellulosomes can have several applications, such as the production of biofuels (Cha *et al.*, 2007) or the improvement of the nutritive value of cereal-based diets for poultry (Cha *et al.*, 2007; Ribeiro *et al.*, 2008) which could potentially benefit from the presentation of the enzymes in close proximity, leading to higher enzyme synergy and stability. Finally, the automation of several key steps on the crystallization process have allowed testing several hundreds of crystallization conditions, increasing the chances of obtaining good quality crystals. Ultimately, the strategies adopted in this work have allowed the successful structural and biochemical characterization of all Coh-Doc specificities identified in the *R. flavefaciens* cellulosome that had not previously been characterized (Chapters 3, 4 and 5).

The draft genome of *R. flavefaciens* strain FD-1 revealed the presence of 223 dockerin-encoding ORFs (Dassa *et al.*, 2014; Rincon *et al.*, 2010). This is triple the number of cellulosomal components observed for clostridial species, rendering *R. flavefaciens* cellulosome the most intricate described so far and the perfect subject to expand our knowledge on the bacterial mechanisms of cohesin-dockerin interaction. The 223 *R. flavefaciens* FD-1 dockerin sequences exhibit great sequence diversity that ranges between 20 % to 98 % identity. Thus, based on primary sequence homology, *R. flavefaciens* FD-1 dockerins were grouped into six distinct major families and eleven sub-groups. Each group exhibits unique and recognizable features. Some dockerins resemble known type I dockerins described in *Clostridia* (groups 3 and 6) while others are exclusive to *R. flavefaciens* FD-1 (groups 1 and 2). Thus, one of this this thesis main goals was to ascertain the functional significance of dockerin classification into different groups. Work presented in Chapter 2 revealed several novel specificities within the different *R. flavefaciens* FD-1 cohesin-dockerin groups, providing a snapshot of the intricate *R. flavefaciens* cellulosome system molecular organization. The data correlates well with the group classification, which was therefore found to be functionally relevant. Contrastingly, the second order of classification of the six dockerin groups into 11 subgroups seems to be functionally redundant. Group 1 dockerins, which are mainly associated with various catalytic modules, were found to bind cohesins ScaA1-2 and ScaB1-4. Group 3 and 6, comprising mostly hemicellulase associated dockerins, seems to have a specific affinity for the cohesin of adaptor scaffoldin ScaC. ScaA's dockerin, the single member of group 5, showed affinity for ScaB cohesins 5-9, as previously suggested (Rincon *et al.*, 2003; Slutzki *et al.*, 2013). Finally, group 2 and 4 dockerins were found to bind ScaH and anchoring scaffoldin ScaE, completing the full picture of *R. flavefaciens* Coh-Doc specificities. The *in vivo* co-expression system used in the Coh-Doc complex mining was essential to identify the best candidates for subsequent structural and biochemical work described in Chapters 3 to 5. Therefore, the complexes formed between

CohScaC and Doc3, CohScaB3 and Doc1a, CohScaA2 and Doc1b and between CohScaB5 and DocScaA displayed exceptional stability and were easily expressed while presenting very high protein yields. They represent three of the four identified Coh-Doc specificities in the *R. flavefaciens* cellulosome. The structural attributes of the fourth *R. flavefaciens* Coh-Doc specificity were described in a previous study involving the complex formed between CttA's XDoc (a group 4 dockerin) and the cohesin of ScaE (Alber *et al.*, 2009). The *Rf*CohScaE-XDocCttA complex structure revealed an atypical Coh-Doc fold, as predicted by the dockerin's divergent sequence that contains three cryptic inserts. A combined functional role for the three enigmatic dockerin inserts was established, whereby the extraneous segments serve as structural buttresses that support the extended conformation of the dockerin associated X-module, through the establishment of an extensive network of intermodular interactions (Alber *et al.*, 2009). The dockerin also possesses a second atypical calcium-binding loop that is disrupted by one of the inserts, altering the pattern of Ca^{2+} ion coordination. The structure of the X-module does not share similarities with other known X-modules from cellulolytic bacteria and, unlike the X module in the type-II Coh:XDoc interaction of *C. thermocellum* (Adams *et al.*, 2006), it does not participate in cohesin recognition or in dockerin stabiliazation. In contrast, group 4 dockerin associated X-module seems to act as a spacer, separating the CBM modules of CttA and the bacterial cell wall (Alber *et al.*, 2009).

Out of all the *R. flavefaciens* Coh-Doc complex structures described in the present work, the structure of ScaC's cohesin in complex with a group 3 dockerin, described in Chapter 3, is the one that shares the highest level of similarity with complexes from other bacteria. In fact, a phylogenetic analysis places ScaC's cohesin on a separate branch from all other *R. flavefaciens* type III cohesins. CohScaC's branch is actually closer to the type I cohesins branch that includes cohesins from other organisms such as *C. thermocellum* and *C. cellulolyticum* (Carvalho *et al.*, 2003; Pinheiro *et al.*, 2008). In spite of that, *Rf*CohScaC-Doc3 presents some distinctive structural features when compared with other type I Coh-Doc complexes, namely two significant α-helices and a large β-hairpin insertion that significantly increases the dockerin contacting surface. The structure of the 2 complexes involving group 1 dockerins bound to CohScaB3 and CohScaA2, described in Chapter 4, display striking structural similarities with each other, presenting a r.m.s.d. of 0.45 Å over 136 main chain carbon atoms. Even more striking is the complete conservation of the main Doc and Coh interacting residues between both complexes, whose positions completely overlap when they are superposed. This is coherent with the data shown in Chapter 2, suggesting that group 1 dockerins have similar specificities, displaying affinity for both ScaA cohesins and ScaB cohesin 1-4. Lastly, the structure of *Rf*CohScaB5-DocScaA complex, described in Chapter 5, exhibits several unique

features. The dockerin of ScaA displays a very compact and globular conformation promoted by numerous intramolecular interactions not previously observed in dockerins. In addition, similarly to the group 4 XDoc of the CttA protein, a 12-residue long insert is present in the second calcium binding loop, disrupting the pattern of calcium coordination. Thus, the geometry of the Ca^{2+} coordination in ScaA Doc adopts an unusual tetrahedral geometry instead of the classic octahedral geometry.

Previous structure-function studies in Coh-Doc complexes of the cellulosomes of *C. thermocellum* (Carvalho *et al.*, 2003, 2007) and *C. cellulolyticum* (Pinheiro *et al.*, 2008) revealed that Docs used to recruit the microbial enzymes to these highly intricate multi-enzyme complexes display a dual-binding mode. In addition, recent reports revealed that the recruitment of cellulosomes to the *P. cellulosolvens* (Cameron, Weinstein, *et al.*, 2015) and *A. cellulolyticus* cell surfaces is also mediated by Docs that display a dual-binding mode (Cameron, Najmudin, *et al.*, 2015; Brás *et al.*, 2016). The structure of dual-binding mode Docs presents a 2-fold internal symmetry that allows binding to the Coh partner in two 180º-related alternate positions. The fact that Docs, in general, possess two different Coh-interacting platforms displaying identical specificities suggests that the dual-binding mode could contribute to enhance the conformational flexibility of the quaternary architecture in highly populated multi-enzyme complexes. This was supported by the observation that non-cellulosomal Docs that recruit single enzymes directly to the cell surface of *C. thermocellum* present a single-binding mode (Brás *et al.*, 2012). In contrast, the data presented in this work suggest that the recruitment of enzymes into *R. flavefaciens* cellulosome, whether directly into the primary scaffoldins ScaA and ScaB or indirectely *via* adaptor scaffoldins such as ScaC, is performed exclusively through single binding mode Coh-Doc interactions. Thus, the enzyme-associated group 1, 3 and 6 dockerins do not seem to possess the internal symmetry required to support a dual binding mode and do interact with their cognate cohesins through a single protein-binding interface. Thus, single-binding mode Docs seems to completely dominate both cellulosome assembly and cellulosome cell surface attachment in *R. flavefaciens*. Much like the enzyme associated dockerins, the dockerins of ScaA and ScaB also seem to lack the required internal symmetry to support a dual binding mode. As such, a complete *R. flavefaciens* cellulosome can be assembled and attached to the cell wall solely *via* single binding mode Coh-Doc interactions.

In general, data presented in this work questions the hypothesis that the dual-binding mode mechanism provides the conformational flexibility required to accommodate a large number of enzymes acting in close proximity. This property was believed to be of intrinsic importance to degrade plant cell walls in which the topology of different composite structures varies between plants and during the degradative process. It is possible that widespread presence of adaptor

scaffoldins and the lower complexity presented by ScaA primary scaffoldin, which only contains two cohesins, may reduce the steric constraints imposed by enzyme assembly and this may reduce the need for Docs displaying a dual-binding mode. Interestingly, as described in Chapter 2, some dockerins belonging to group 4, which do not contain the characteristic upstream X-module, can interact with the ScaE anchoring scaffoldin (located at the cell surface) in two different orientations. One of such dockerins belongs to the adaptor scaffoldin ScaH. Considering that ScaB dockerin can interact with the single cohesin of ScaH, it is likely that the dual binding mode can be incorporated into the attachment of *R. flavefaciens* cellulosome to the cell wall via ScaH, adding flexibility to its structure. In addition, it seems that ScaH Doc presents affinity for its own cohesin. Incorporation of a variable number of ScaH adaptor scaffoldins at the cell surface could lead to the positioning of several cellulosomes at different distances from the cell wall. This could probably be advantageous to the bacteria in periods requiring high cellulolytic activity and consequent increased expression of cellulosomal components, by distancing the multiple cell wall attached cellulosomes from each other, thus reducing non-productive steric clashes.

Overall, this report reveals that type III Coh-Doc interactions involved in the assembly of *R. flavefaciens* cellulosome, are very diverse and present unique features not previously described in the now well described type I and II complexes. Four distinct specificities were identified and thoroughly characterized in *R. flavefaciens* cellulosome. Novel Coh-Doc structures were described containing several unique structural features that result in very stable and specific interactions. The key residues for each interaction were recognized revealing major differences between each new type III complex described and the previously characterized type I and II complexes. Monovalent adaptor scaffoldins with unique functions such as integration of hemicelluloses (ScaC) and incorporation of a dual binding mode Doc (ScaH) that, to date, are unique to the *R. flavefaciens* cellulosome, represent one of the most curious findings of this work. Another interesting discovery relates to the capacity to strongly bind to cohesins of ScaH and ScaE revealed by a group 2 dockerin with a truncated sequence, as this Doc only possesses one of the two duplicated segments universally observed in Docs. This is probably the strongest protein-protein interaction, involving such a small peptide, ever to be described. Alltogether, a substantial amount of new information was gathered from studying these novel Cohs and Docs, proving that type III Coh-Doc complexes present a great opportunity to expand our understanding of these unique, very strong and specific interactions.

In the near future, further developments are expected in the ruminal cellulosome field. This will involve research into the structural mechanisms behind both the interaction of the small group 2 dockerins with their cognate cohesins and the group 4 dockerins that seem to present a dual

binding mode. The development of a mini-cellulosome with multiple valences using only *R. flavefaciens* Cohs and Docs is also envisaged, with focus on potential biotechnological applications working at mesophilic temperatures. The advantage of using the *R. flavefaciens* cellulosomal system to develop designer cellulosomes rest on the fact that up to 4 different specificities can be incorporated in a single cellulosome, using cohesins and dockerins that have similar optimal conditions of expression and stability. This allows integration of up to 4 different catalytic modules in a single nanomachine.

The work described in Chapter 6 results from applying the methodologies described in the previous Chapters to answer some questions regarding type I Coh-Doc interactions of *A. cellulolyticus*. It was recently shown that the sequence and structural symmetry within the ScaB *A. cellulolyticus* type I dockerin allows it to bind ScaC cohesins in two different orientations (Cameron, Najmudin, *et al.*, 2015). This symmetry is also evident in the enzyme-borne dockerins of *A. cellulolyticus* that interact with ScaA, therefore suggesting a putative dual-binding mode capability for these interactions. Chapter 6 describes the structures of two complexes displaying the Doc of a glycoside hydrolase bound to the sixth cohesin of ScaA, in two different orientations. Thus, again the data confirms the widespread significance of the dual binding mode in Docs of non-ruminal cellulosomes. Although very closely related, in *A. cellulolyticum* the enzyme-borne and ScaB type I Docs do not display cross-specificity. Thus, Coh-contacting residues at positions 11 and 12, which are traditionally recognized has specificity determinants (Bayer *et al.*, 2004), are different in the two type I Docs. Differences at these key residues may explain why there is a lack of cross-specificity between the type I-Doc interactions that modulate the binding of ScaB into ScaC or the cellulosomal enzymes into ScaA (Hamberg *et al.*, 2014; Xu *et al.*, 2003). Given the similarities between these structures and the published structures of DocScaB in complex with a CohScaC3, the knowledge gathered on these two complexes was used to design a specificity hybrid dockerin with the ability to recognize both ScaA and ScaC cohesins, using two different cohesin-binding interfaces. By changing key residues of the N-terminal repeat of DocScaB by those of the enzyme bearing type I dockerins, the specificity hybrid was successfully created, as DocScaB retains its capacity to bind CohScaC3 via its C-terminal Coh interface while it binds CohScaA6 via its N-terminal Coh interface. The specificity hybrid is an example of the kind of manipulation that is possible to achieve by continuously gathering information about Coh-Doc interactions.

A brief overview of the variety of cellulosomes identified to date demonstrates the sophistication and diversity of structural mechanisms that was evolved to organize these highly intricate multi-protein complexes. It is evident that these systems are exceptionally varied in size, structural organisation and nature of their different modular components. Continued input

from structural biology initiatives will enable deriving functional implication and will aid in the description of new and more accurate cellulosome component structures and functions (Cameron, 2015). As mentioned before, the potential contained in the Coh-Doc interaction is remarkable. Such a strong and highly specific protein:protein interaction may be explored in several ways: from the development of mini-cellulosomes that can be used in biofuel production, waste management and animal nutrition to the incorporation of cohesin and dockerins in affinity based systems with applications in research, medical diagnosis or even pharmaceutics. However, to fully harness the Coh-Doc interaction potential, a deep understanding of the mechanisms behind such a unique system is essential, which requires a continuous effort to expand our knowledge on this subject. Fortunately, the advent of automation and high-throughput methodologies suggests a very promising future for cellulosome research

Bibliographic References

Adams, J. J., Currie, M. A., Ali, S., Bayer, E. A., Jia, Z., & Smith, S. P. (2010). Insights into higher-order organization of the cellulosome revealed by a dissect-and-build approach: crystal structure of interacting Clostridium thermocellum multimodular components. *Journal of Molecular Biology*, *396*(4), 833–839. https://doi.org/10.1016/j.jmb.2010.01.015

Adams, J. J., Gregg, K., Bayer, E. A., Boraston, A. B., & Smith, S. P. (2008). Structural basis of Clostridium perfringens toxin complex formation. *Proceedings of the National Academy of Sciences*, *105*(34), 12194–12199. https://doi.org/10.1073/pnas.0803154105

Adams, J. J., Pal, G., Jia, Z., & Smith, S. P. (2006). Mechanism of bacterial cell-surface attachment revealed by the structure of cellulosomal type II cohesin-dockerin complex. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(2), 305–310.

Adams, J. J., Pal, G., Yam, K., Spencer, H. L., Jia, Z., & Smith, S. P. (2005). Purification and crystallization of a trimodular complex comprising the type II cohesin-dockerin interaction from the cellulosome of Clostridium thermocellum. *Acta Crystallographica. Section F, Structural Biology and Crystallization Communications*, *61*(Pt 1), 46–48. https://doi.org/10.1107/S1744309104025837

Adams, J. J., Webb, B. A., Spencer, H. L., & Smith, S. P. (2005). Structural Characterization of Type II Dockerin Module from the Cellulosome of *Clostridium thermocellum* : Calcium-Induced Effects on Conformation and Target Recognition †. *Biochemistry*, *44*(6), 2173–2182. https://doi.org/10.1021/bi048039u

Adams, P. D., Afonine, P. V., Bunkóczi, G., Chen, V. B., Davis, I. W., Echols, N., … Zwart, P. H. (2010). PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallographica. Section D, Biological Crystallography*, *66*(Pt 2), 213–221. https://doi.org/10.1107/S0907444909052925

Alber, O., Noach, I., Rincon, M. T., Flint, H. J., Shimon, L. J. W., Lamed, R., … Bayer, E. A. (2009). Cohesin diversity revealed by the crystal structure of the anchoring cohesin from *Ruminococcus flavefaciens*. *Proteins*, *77*(3), 699–709. https://doi.org/10.1002/prot.22483

Arfi, Y., Shamshoum, M., Rogachev, I., Peleg, Y., & Bayer, E. A. (2014). Integration of bacterial lytic polysaccharide monooxygenases into designer cellulosomes promotes enhanced cellulose degradation. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(25), 9109–9114. https://doi.org/10.1073/pnas.1404148111

Artzi, L., Bayer, E. A., & Moraïs, S. (2016). Cellulosomes: bacterial nanomachines for dismantling plant polysaccharides. *Nature Reviews Microbiology*, *15*(2), 83–95. https://doi.org/10.1038/nrmicro.2016.164

Artzi, L., Dassa, B., Borovok, I., Shamshoum, M., Lamed, R., & Bayer, E. A. (2014). Cellulosomics of the cellulolytic thermophile Clostridium clariflavum. *Biotechnology for Biofuels*, *7*(1), 100. https://doi.org/10.1186/1754-6834-7-100

Artzi, L., Morag, E., Barak, Y., Lamed, R., & Bayer, E. A. (2015). Clostridium clariflavum: Key Cellulosome Players Are Revealed by Proteomic Analysis. *mBio*, *6*(3), e00411-00415. https://doi.org/10.1128/mBio.00411-15

Artzi, L., Morag, E., Shamshoum, M., & Bayer, E. A. (2016). Cellulosomal expansin: functionality and incorporation into the complex. *Biotechnology for Biofuels*, *9*, 61. https://doi.org/10.1186/s13068-016-0474-5

Aurilia, V., Martin, J. C., Scott, K. P., Mercer, D. K., Johnston, M. E. ., & Flint, H. J. (2000). Organisation and Variable Incidence of Genes Concerned with the Utilization of Xylans

in the Rumen Cellulolytic Bacterium *Ruminococcus flavefaciens*. *Anaerobe*, *6*(6), 333–340. https://doi.org/10.1006/anae.2000.0358

Barak, Y., Handelsman, T., Nakar, D., Mechaly, A., Lamed, R., Shoham, Y., & Bayer, E. A. (2005). Matching fusion protein systems for affinity analysis of two interacting families of proteins: the cohesin-dockerin interaction. *Journal of Molecular Recognition: JMR*, *18*(6), 491–501. https://doi.org/10.1002/jmr.749

Barrière, Y., Guillet, C., Goffner, D., & Pichon, M. (2003). Genetic variation and breeding strategies for improved cell wall digestibility in annual forage crops. A review. *Animal Research*, *52*(3), 193–228. https://doi.org/10.1051/animres:2003018

Battye, T. G. G., Kontogiannis, L., Johnson, O., Powell, H. R., & Leslie, A. G. W. (2011). *iMOSFLM* : a new graphical interface for diffraction-image processing with *MOSFLM*. *Acta Crystallographica Section D Biological Crystallography*, *67*(4), 271–281. https://doi.org/10.1107/S0907444910048675

Bayer, E. A., and Lamed, R. (2006). The cellulosome saga: Early history. In *Cellulosome* (pp. 11–46). New York: Nova Science Publishers, Inc.

Bayer, E. A., Belaich, J.-P., Shoham, Y., & Lamed, R. (2004). The cellulosomes: multienzyme machines for degradation of plant cell wall polysaccharides. *Annual Review of Microbiology*, *58*, 521–554. https://doi.org/10.1146/annurev.micro.57.030502.091022

Bayer, E. A., Chanzy, H., Lamed, R., & Shoham, Y. (1998). Cellulose, cellulases and cellulosomes. *Current Opinion in Structural Biology*, *8*(5), 548–557.

Bayer, E. A., Kenig, R., & Lamed, R. (1983). Adherence of Clostridium thermocellum to Cellulose. *Journal of Bacteriology*, *156*(2), 818–827.

Bayer, E. A., & Lamed, R. (1986). Ultrastructure of the cell surface cellulosome of Clostridium thermocellum and its interaction with cellulose. *Journal of Bacteriology*, *167*(3), 828–836.

Bayer, E. A., Lamed, R., & Himmel, M. E. (2007). The potential of cellulases and cellulosomes for cellulosic waste management. *Current Opinion in Biotechnology*, *18*(3), 237–245. https://doi.org/10.1016/j.copbio.2007.04.004

Bayer, E. A., Lamed, R., White, B. A., & Flint, H. J. (2008). From cellulosomes to cellulosomics. *The Chemical Record*, *8*(6), 364–377. https://doi.org/10.1002/tcr.20160

Bayer, E. A., Morag, E., & Lamed, R. (1994). The cellulosome--a treasure-trove for biotechnology. *Trends in Biotechnology*, *12*(9), 379–386. https://doi.org/10.1016/0167-7799(94)90039-6

Bayer, E. A., Shimon, L. J., Shoham, Y., & Lamed, R. (1998). Cellulosomes-structure and ultrastructure. *Journal of Structural Biology*, *124*(2–3), 221–234. https://doi.org/10.1006/jsbi.1998.4065

Bedford, M. R. (2000). Exogenous enzymes in monogastric nutrition — their current value and future benefits. *Animal Feed Science and Technology*, *86*(1–2), 1–13. https://doi.org/10.1016/S0377-8401(00)00155-3

Béguin, P., & Aubert, J. P. (1994). The biological degradation of cellulose. *FEMS Microbiology Reviews*, *13*(1), 25–58.

Bégum, P., & Lemaire, M. (1996). The Cellulosome: An Exocellular, Multiprotein Complex Specialized in Cellulose Degradation. *Critical Reviews in Biochemistry and Molecular Biology*, *31*(3), 201–236. https://doi.org/10.3109/10409239609106584

Ben David, Y., Dassa, B., Borovok, I., Lamed, R., Koropatkin, N. M., Martens, E. C., … Moraïs, S. (2015). Ruminococcal cellulosome systems from rumen to human. *Environmental Microbiology*, *17*(9), 3407–3426. https://doi.org/10.1111/1462-2920.12868

Berg Miller, M. E., Antonopoulos, D. A., Rincon, M. T., Band, M., Bari, A., Akraiko, T., … White, B. A. (2009). Diversity and strain specificity of plant cell wall degrading enzymes revealed by the draft genome of *Ruminococcus flavefaciens* FD-1. *PloS One*, *4*(8), e6650. https://doi.org/10.1371/journal.pone.0006650

Bhat, S., Goodenough, P. W., Bhat, M. K., & Owen, E. (1994). Isolation of four major subunits from Clostridium thermocellum cellulosome and their synergism in the hydrolysis of crystalline cellulose. *International Journal of Biological Macromolecules*, *16*(6), 335–342.

Blanchette, C., Lacayo, C. I., Fischer, N. O., Hwang, M., & Thelen, M. P. (2012). Enhanced cellulose degradation using cellulase-nanosphere complexes. *PloS One*, *7*(8), e42116. https://doi.org/10.1371/journal.pone.0042116

Bolam, D. N., Ciruela, A., McQueen-Mason, S., Simpson, P., Williamson, M. P., Rixon, J. E., … Gilbert, H. J. (1998). Pseudomonas cellulose-binding domains mediate their effects by increasing enzyme substrate proximity. *The Biochemical Journal*, *331 (Pt 3)*, 775–781.

Bomble, Y. J., Beckham, G. T., Matthews, J. F., Nimlos, M. R., Himmel, M. E., & Crowley, M. F. (2011). Modeling the self-assembly of the cellulosome enzyme complex. *The Journal of Biological Chemistry*, *286*(7), 5614–5623. https://doi.org/10.1074/jbc.M110.186031

Bond, C. S., & Schüttelkopf, A. W. (2009). *ALINE* : a WYSIWYG protein-sequence alignment editor for publication-quality alignments. *Acta Crystallographica Section D Biological Crystallography*, *65*(5), 510–512. https://doi.org/10.1107/S0907444909007835

Boraston, A. B., Bolam, D. N., Gilbert, H. J., & Davies, G. J. (2004). Carbohydrate-binding modules: fine-tuning polysaccharide recognition. *The Biochemical Journal*, *382*(Pt 3), 769–781. https://doi.org/10.1042/BJ20040892

Boraston, A. B., Kwan, E., Chiu, P., Warren, R. A. J., & Kilburn, D. G. (2003). Recognition and hydrolysis of noncrystalline cellulose. *The Journal of Biological Chemistry*, *278*(8), 6120–6127. https://doi.org/10.1074/jbc.M209554200

Borne, R., Bayer, E. A., Pagès, S., Perret, S., & Fierobe, H.-P. (2013). Unraveling enzyme discrimination during cellulosome assembly independent of cohesin-dockerin affinity. *The FEBS Journal*, *280*(22), 5764–5779. https://doi.org/10.1111/febs.12497

Brás, J. L. A., Alves, V. D., Carvalho, A. L., Najmudin, S., Prates, J. A. M., Ferreira, L. M. A., … Fontes, C. M. G. A. (2012). Novel Clostridium thermocellum type I cohesin-dockerin complexes reveal a single binding mode. *The Journal of Biological Chemistry*, *287*(53), 44394–44405. https://doi.org/10.1074/jbc.M112.407700

Brás, J. L. A., Pinheiro, B. A., Cameron, K., Cuskin, F., Viegas, A., Najmudin, S., … Fontes, C. M. G. A. (2016). Diverse specificity of cellulosome attachment to the bacterial cell surface. *Scientific Reports*, *6*, 38292. https://doi.org/10.1038/srep38292

Brett, C. T., & Waldron, K. W. (1996). *Physiology and Biochemistry of Plant Cell Walls* (2nd ed.). Springer.

Brown, R. M. (2004). Cellulose structure and biosynthesis: What is in store for the 21st century? *Journal of Polymer Science Part A: Polymer Chemistry*, *42*(3), 487–495. https://doi.org/10.1002/pola.10877

Brulc, J. M., Yeoman, C. J., Wilson, M. K., Berg Miller, M. E., Jeraldo, P., Jindou, S., … White, B. A. (2011). Cellulosomics, a Gene-Centric Approach to Investigating the Intraspecific Diversity and Adaptation of *Ruminococcus flavefaciens* within the Rumen. *PLoS ONE*, *6*(10), e25329. https://doi.org/10.1371/journal.pone.0025329

Bryant, M. P., Small, N., Bouma, C., & Robinson, I. M. (1958). Characteristics of ruminal anaerobic celluloytic cocci and Cillobacterium cellulosolvens n. sp. *Journal of Bacteriology*, *76*(5), 529–537.

Buffetto, F., Ropartz, D., Zhang, X. J., Gilbert, H. J., Guillon, F., & Ralet, M.-C. (2014). Recovery and fine structure variability of RGII sub-domains in wine (Vitis vinifera Merlot). *Annals of Botany*. https://doi.org/10.1093/aob/mcu097

Bule, P., Alves, V. D., Israeli-Ruimy, V., Carvalho, A. L., Ferreira, L. M. A., Smith, S. P., … Fontes, C. M. G. A. (2017). Assembly of *Ruminococcus flavefaciens* cellulosome

revealed by structures of two cohesin-dockerin complexes. *Scientific Reports*, *7*(1). https://doi.org/10.1038/s41598-017-00919-w

Bule, P., Alves, V. D., Leitão, A., Ferreira, L. M. A., Bayer, E. A., Smith, S. P., … Fontes, C. M. G. A. (2016). Single-binding mode integration of hemicellulose degrading enzymes via adaptor scaffoldins in *Ruminococcus flavefaciens* cellulosome. *Journal of Biological Chemistry*, jbc.M116.761643. https://doi.org/10.1074/jbc.M116.761643

Bule, P., Correia, A., Cameron, K., Alves, V. D., Cardoso, V., Fontes, C. M. G. A., & Najmudin, S. (2014). Overexpression, purification, crystallization and preliminary X-ray characterization of the fourth scaffoldin A cohesin from *Acetivibrio cellulolyticus* in complex with a dockerin from a family 5 glycoside hydrolase. *Acta Crystallographica. Section F, Structural Biology Communications*, *70*(Pt 8), 1065–1067. https://doi.org/10.1107/S2053230X14013181

Bule, P., Ruimy-Israeli, V., Cardoso, V., Bayer, E. A., Fontes, C. M. G. A., & Najmudin, S. (2014). Overexpression, crystallization and preliminary X-ray characterization of *Ruminococcus flavefaciens* scaffoldin C cohesin in complex with a dockerin from an uncharacterized CBM-containing protein. *Acta Crystallographica. Section F, Structural Biology Communications*, *70*(Pt 8), 1061–1064. https://doi.org/10.1107/S2053230X14012667

Burton, R. A., & Fincher, G. B. (2014). Evolution and development of cell walls in cereal grains. *Frontiers in Plant Science*, *5*, 456. https://doi.org/10.3389/fpls.2014.00456

Caffall, K. H., & Mohnen, D. (2009). The structure, function, and biosynthesis of plant cell wall pectic polysaccharides. *Carbohydrate Research*, *344*(14), 1879–1900. https://doi.org/10.1016/j.carres.2009.05.021

Cameron, K. (2015). *Structure and function relationships in novel cohesin-dockerin complexes* (Doutoramento). Lisboa, Lisboa.

Cameron, K., Najmudin, S., Alves, V. D., Bayer, E. A., Smith, S. P., Bule, P., … Fontes, C. M. G. A. (2015). Cell-surface Attachment of Bacterial Multienzyme Complexes Involves Highly Dynamic Protein-Protein Anchors. *Journal of Biological Chemistry*, *290*(21), 13578–13590. https://doi.org/10.1074/jbc.M114.633339

Cameron, K., Weinstein, J. Y., Zhivin, O., Bule, P., Fleishman, S. J., Alves, V. D., … Najmudin, S. (2015). Combined crystal structure of a type-I cohesin, mutation and affinity-binding studies reveal structural determinants of cohesin-dockerin specificity. *Journal of Biological Chemistry*, jbc.M115.653303. https://doi.org/10.1074/jbc.M115.653303

Cantarel, B. L., Coutinho, P. M., Rancurel, C., Bernard, T., Lombard, V., & Henrissat, B. (2009). The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. *Nucleic Acids Research*, *37*(Database issue), D233-238. https://doi.org/10.1093/nar/gkn663

Cao, H., & Yin, Y. (2014). Rapid Evolution of Cellulosome Modules by Comparative Analyses of Five Clostridiales Genomes. *BioEnergy Research*. https://doi.org/10.1007/s12155-014-9474-0

Carpita, N., Ralph, J., & McCann, M. (2015). Cell Walls. In B. Buchanan, W. Gruissem, & R. Jones (Eds.), *Biochemistry & Molecular Biology of Plants* (2nd ed., pp. 45–110). Oxford, UK: Wiley Blackwell.

Carvalho, A. L., Dias, F. M. V., Nagy, T., Prates, J. A. M., Proctor, M. R., Smith, N., … Gilbert, H. J. (2007). Evidence for a dual binding mode of dockerin modules to cohesins. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(9), 3089–3094. https://doi.org/10.1073/pnas.0611173104

Carvalho, A. L., Dias, F. M. V., Prates, J. A. M., Nagy, T., Gilbert, H. J., Davies, G. J., … Fontes, C. M. G. A. (2003). Cellulosome assembly revealed by the crystal structure of the cohesin-dockerin complex. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(24), 13809–13814. https://doi.org/10.1073/pnas.1936124100

Carvalho, A. L., Pires, V. M. R., Gloster, T. M., Turkenburg, J. P., Prates, J. A. M., Ferreira, L. M. A., … Gilbert, H. J. (2005). Insights into the structural determinants of cohesin-dockerin specificity revealed by the crystal structure of the type II cohesin from Clostridium thermocellum SdbA. *Journal of Molecular Biology*, *349*(5), 909–915. https://doi.org/10.1016/j.jmb.2005.04.037

Cha, J., Matsuoka, S., Chan, H., Yukawa, H., Inui, M., & Doi, R. H. (2007). Effect of multiple copies of cohesins on cellulase and hemicellulase activities of Clostridium cellulovorans mini-cellulosomes. *Journal of Microbiology and Biotechnology*, *17*(11), 1782–1788.

Chakrabarty, A. M., Demain, A. L., & Tiedje, J. M. (1997). *Gastrointestinal Microbiology, Volume 1: Gastrointestinal Ecosystems and Fermentations*. Boston: Springer US. Retrieved from http://public.eblib.com/choice/publicfullrecord.aspx?p=3069528

Chauvaux, S., Matuschek, M., & Beguin, P. (1999). Distinct affinity of binding sites for S-layer homologous domains in Clostridium thermocellum and Bacillus anthracis cell envelopes. *Journal of Bacteriology*, *181*(8), 2455–2458.

Chen, C., Cui, Z., Song, X., Liu, Y.-J., Cui, Q., & Feng, Y. (2016). Integration of bacterial expansin-like proteins into cellulosome promotes the cellulose degradation. *Applied Microbiology and Biotechnology*, *100*(5), 2203–2212. https://doi.org/10.1007/s00253-015-7071-6

Chen, C., Cui, Z., Xiao, Y., Cui, Q., Smith, S. P., Lamed, R., … Feng, Y. (2014). Revisiting the NMR solution structure of the Cel48S type-I dockerin module from Clostridium thermocellum reveals a cohesin-primed conformation. *Journal of Structural Biology*, *188*(2), 188–193. https://doi.org/10.1016/j.jsb.2014.09.006

Chen, V. B., Arendall, W. B., Headd, J. J., Keedy, D. A., Immormino, R. M., Kapral, G. J., … Richardson, D. C. (2010). *MolProbity* : all-atom structure validation for macromolecular crystallography. *Acta Crystallographica Section D Biological Crystallography*, *66*(1), 12–21. https://doi.org/10.1107/S0907444909042073

Choi, S. K., & Ljungdahl, L. G. (1996). Structural role of calcium for the organization of the cellulosome of Clostridium thermocellum. *Biochemistry*, *35*(15), 4906–4910. https://doi.org/10.1021/bi9524631

Cooper, G. M., & Hausman, R. E. (2009). *The Cell: A Molecular Approach, Fifth Edition* (5th Edition). Sinauer Associates Inc.

Cosgrove, D. J. (1997). Assembly and enlargement of the primary cell wall in plants. *Annual Review of Cell and Developmental Biology*, *13*, 171–201. https://doi.org/10.1146/annurev.cellbio.13.1.171

Cosgrove, D. J. (2005). Growth of the plant cell wall. *Nature Reviews. Molecular Cell Biology*, *6*(11), 850–861. https://doi.org/10.1038/nrm1746

Costa, M., Fernandes, V. O., Ribeiro, T., Serrano, L., Cardoso, V., Santos, H., … Fontes, C. M. G. A. (2014). Construction of GH16 *β*-glucanase mini-cellulosomes to improve the nutritive value of barley-based diets for broilers. *Journal of Agricultural and Food Chemistry*, *62*(30), 7496–7506. https://doi.org/10.1021/jf502157y

Coutinho, P. M., Deleury, E., Davies, G. J., & Henrissat, B. (2003). An Evolving Hierarchical Family Classification for Glycosyltransferases. *Journal of Molecular Biology*, *328*(2), 307–317. https://doi.org/10.1016/S0022-2836(03)00307-3

Cowtan, K. (2006). The Buccaneer software for automated model building. 1. Tracing protein chains. *Acta Crystallographica. Section D, Biological Crystallography*, *62*(Pt 9), 1002–1011. https://doi.org/10.1107/S0907444906022116

Currie, M. A., Adams, J. J., Faucher, F., Bayer, E. A., Jia, Z., & Smith, S. P. (2012). Scaffoldin Conformation and Dynamics Revealed by a Ternary Complex from the Clostridium thermocellum Cellulosome. *Journal of Biological Chemistry*, *287*(32), 26953–26961. https://doi.org/10.1074/jbc.M112.343897

Dassa, B., Borovok, I., Lamed, R., Henrissat, B., Coutinho, P., Hemme, C. L., … Bayer, E. A. (2012). Genome-wide analysis of *Acetivibrio cellulolyticus* provides a blueprint of an

elaborate cellulosome system. *BMC Genomics*, *13*(1), 210. https://doi.org/10.1186/1471-2164-13-210

Dassa, B., Borovok, I., Ruimy-Israeli, V., Lamed, R., Flint, H. J., Duncan, S. H., … Bayer, E. A. (2014). Rumen cellulosomics: divergent fiber-degrading strategies revealed by comparative genome-wide analysis of six ruminococcal strains. *PloS One*, *9*(7), e99221. https://doi.org/10.1371/journal.pone.0099221

Demain, A. L., Newcomb, M., & Wu, J. H. D. (2005). Cellulase, Clostridia, and Ethanol. *Microbiol. Mol. Biol. Rev.*, *69*(1), 124–154. https://doi.org/10.1128/mmbr.69.1.124-154.2005

Demishtein, A., Karpol, A., Barak, Y., Lamed, R., & Bayer, E. A. (2010). Characterization of a dockerin-based affinity tag: application for purification of a broad variety of target proteins. *Journal of Molecular Recognition: JMR*, *23*(6), 525–535. https://doi.org/10.1002/jmr.1029

Devillard, E., Bera-Maillet, C., Flint, H. J., Scott, K. P., Newbold, C. J., Wallace, R. J., … Forano, E. (2003). Characterization of XYN10B, a modular xylanase from the ruminal protozoan Polyplastron multivesiculatum, with a family 22 carbohydrate-binding module that binds to cellulose. *Biochemical Journal*, *373*(2), 495–503. https://doi.org/10.1042/bj20021784

Diederichs, K., & Karplus, P. A. (2013). Better models by discarding data? *Acta Crystallographica Section D Biological Crystallography*, *69*(7), 1215–1222. https://doi.org/10.1107/S0907444913001121

Ding, S. Y., Bayer, E. A., Steiner, D., Shoham, Y., & Lamed, R. (1999). A novel cellulosomal scaffoldin from *Acetivibrio cellulolyticus* that contains a family 9 glycosyl hydrolase. *Journal of Bacteriology*, *181*(21), 6720–6729.

Ding, S. Y., Rincon, M. T., Lamed, R., Martin, J. C., McCrae, S. I., Aurilia, V., … Flint, H. J. (2001). Cellulosomal scaffoldin-like proteins from *Ruminococcus flavefaciens*. *Journal of Bacteriology*, *183*(6), 1945–1953. https://doi.org/10.1128/JB.183.6.1945-1953.2001

Ding, S.-Y., Bayer, E. A., Steiner, D., Shoham, Y., & Lamed, R. (2000). A Scaffoldin of the Bacteroides cellulosolvens Cellulosome That Contains 11 Type II Cohesins. *Journal of Bacteriology*, *182*(17), 4915–4925. https://doi.org/10.1128/JB.182.17.4915-4925.2000

Doi, R. H., & Kosugi, A. (2004). Cellulosomes: plant-cell-wall-degrading enzyme complexes. *Nature Reviews. Microbiology*, *2*(7), 541–551. https://doi.org/10.1038/nrmicro925

Doi, R. H., Kosugi, A., Murashima, K., Tamaru, Y., & Han, S. O. (2003). Cellulosomes from mesophilic bacteria. *Journal of Bacteriology*, *185*(20), 5907–5914.

Emsley, P., & Cowtan, K. (2004). Coot: model-building tools for molecular graphics. *Acta Crystallographica. Section D, Biological Crystallography*, *60*(Pt 12 Pt 1), 2126–2132. https://doi.org/10.1107/S0907444904019158

Evans, P. (2006). Scaling and assessment of data quality. *Acta Crystallographica Section D Biological Crystallography*, *62*(1), 72–82. https://doi.org/10.1107/S0907444905036693

Felix, C. R., & Ljungdahl, L. G. (1993). The cellulosome: the exocellular organelle of Clostridium. *Annual Review of Microbiology*, *47*, 791–819. https://doi.org/10.1146/annurev.mi.47.100193.004043

Fierobe, H. P., Pagès, S., Bélaïch, A., Champ, S., Lexa, D., & Bélaïch, J. P. (1999). Cellulosome from Clostridium cellulolyticum: molecular study of the Dockerin/Cohesin interaction. *Biochemistry*, *38*(39), 12822–12832.

Fierobe, H.-P., Mingardon, F., Mechaly, A., Bélaïch, A., Rincon, M. T., Pagès, S., … Bayer, E. A. (2005). Action of designer cellulosomes on homogeneous versus complex substrates: controlled incorporation of three distinct enzymes into a defined trifunctional scaffoldin. *The Journal of Biological Chemistry*, *280*(16), 16325–16334. https://doi.org/10.1074/jbc.M414449200

Flint, H. J., & Bayer, E. A. (2008). Plant Cell Wall Breakdown by Anaerobic Microorganisms from the Mammalian Digestive Tract. *Annals of the New York Academy of Sciences*, *1125*(1), 280–288. https://doi.org/10.1196/annals.1419.022

Flint, H. J., Bayer, E. A., Rincon, M. T., Lamed, R., & White, B. A. (2008). Polysaccharide utilization by gut bacteria: potential for new insights from genomic analysis. *Nature Reviews Microbiology*, *6*(2), 121–131. https://doi.org/10.1038/nrmicro1817

Fontes, C. M. G. A., & Gilbert, H. J. (2010). Cellulosomes: Highly Efficient Nanomachines Designed to Deconstruct Plant Cell Wall Complex Carbohydrates. *Annual Review of Biochemistry*, *79*(1), 655–681. https://doi.org/10.1146/annurev-biochem-091208-085603

García-Alvarez, B., Melero, R., Dias, F. M. V., Prates, J. A. M., Fontes, C. M. G. A., Smith, S. P., … Llorca, O. (2011). Molecular architecture and structural transitions of a Clostridium thermocellum mini-cellulosome. *Journal of Molecular Biology*, *407*(4), 571–580. https://doi.org/10.1016/j.jmb.2011.01.060

Gefen, G., Anbar, M., Morag, E., Lamed, R., & Bayer, E. A. (2012). Enhanced cellulose degradation by targeted integration of a cohesin-fused *β*-glucosidase into the Clostridium thermocellum cellulosome. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(26), 10298–10303. https://doi.org/10.1073/pnas.1202747109

Gerngross, U. T., Romaniec, M. P., Kobayashi, T., Huskisson, N. S., & Demain, A. L. (1993). Sequencing of a Clostridium thermocellum gene (cipA) encoding the cellulosomal SL-protein reveals an unusual degree of internal homology. *Molecular Microbiology*, *8*(2), 325–334.

Gerwig, G. J., Kamerling, J. P., Vliegenthart, J. F., Morag, E., Lamed, R., & Bayer, E. A. (1993). The nature of the carbohydrate-peptide linkage region in glycoproteins from the cellulosomes of Clostridium thermocellum and Bacteroides cellulosolvens. *The Journal of Biological Chemistry*, *268*(36), 26956–26960.

Gilbert, H. J. (Ed.). (1999). *Recent advances in carbohydrate bioengineering*. Cambridge, UK: Royal Society of Chemistry.

Gilbert, H. J. (2007). Cellulosomes: microbial nanomachines that display plasticity in quaternary structure: Cohesin dockerin recognition. *Molecular Microbiology*, *63*(6), 1568–1576. https://doi.org/10.1111/j.1365-2958.2007.05640.x

Gilbert, H. J. (2010). The Biochemistry and Structural Biology of Plant Cell Wall Deconstruction. *Plant Physiology*, *153*(2), 444–455. https://doi.org/10.1104/pp.110.156646

Gilbert, H. J., Stålbrand, H., & Brumer, H. (2008). How the walls come crumbling down: recent structural biochemistry of plant polysaccharide degradation. *Current Opinion in Plant Biology*, *11*(3), 338–348. https://doi.org/10.1016/j.pbi.2008.03.004

Gilkes, N. R., Warren, R. A., Miller, R. C. J., & Kilburn, D. G. (1988). Precise excision of the cellulose binding domains from two Cellulomonas fimi cellulases by a homologous protease and the effect on catalysis. *The Journal of Biological Chemistry*, *263*(21), 10401–10407.

Goyal, G., Tsai, S.-L., Madan, B., DaSilva, N. A., & Chen, W. (2011). Simultaneous cell growth and ethanol production from cellulose by an engineered yeast consortium displaying a functional mini-cellulosome. *Microbial Cell Factories*, *10*(1), 89. https://doi.org/10.1186/1475-2859-10-89

Guillén, D., Sánchez, S., & Rodríguez-Sanoja, R. (2009). Carbohydrate-binding domains: multiplicity of biological roles. *Applied Microbiology and Biotechnology*, *85*(5), 1241–1249. https://doi.org/10.1007/s00253-009-2331-y

Guillén, D., Sánchez, S., & Rodríguez-Sanoja, R. (2010). Carbohydrate-binding domains: multiplicity of biological roles. *Applied Microbiology and Biotechnology*, *85*(5), 1241–1249. https://doi.org/10.1007/s00253-009-2331-y

Gunnoo, M., Cazade, P.-A., Galera-Prat, A., Nash, M. A., Czjzek, M., Cieplak, M., … Thompson, D. (2016). Nanoscale Engineering of Designer Cellulosomes. *Advanced Materials (Deerfield Beach, Fla.)*, *28*(27), 5619–5647. https://doi.org/10.1002/adma.201503948

Haimovitz, R., Barak, Y., Morag, E., Voronov-Goldman, M., Shoham, Y., Lamed, R., & Bayer, E. A. (2008). Cohesin-dockerin microarray: Diverse specificities between two complementary families of interacting protein modules. *Proteomics*, *8*(5), 968–979. https://doi.org/10.1002/pmic.200700486

Hamberg, Y., Ruimy-Israeli, V., Dassa, B., Barak, Y., Lamed, R., Cameron, K., … Fried, D. B. (2014). Elaborate cellulosome architecture of *Acetivibrio cellulolyticus* revealed by selective screening of cohesin-dockerin interactions. *PeerJ*, *2*, e636. https://doi.org/10.7717/peerj.636

Hammel, M., Fierobe, H.-P., Czjzek, M., Finet, S., & Receveur-Bréchot, V. (2004). Structural insights into the mechanism of formation of cellulosomes probed by small angle X-ray scattering. *The Journal of Biological Chemistry*, *279*(53), 55985–55994. https://doi.org/10.1074/jbc.M408979200

Hammel, M., Fierobe, H.-P., Czjzek, M., Kurkal, V., Smith, J. C., Bayer, E. A., … Receveur-Bréchot, V. (2005). Structural basis of cellulosome efficiency explored by small angle X-ray scattering. *The Journal of Biological Chemistry*, *280*(46), 38562–38568. https://doi.org/10.1074/jbc.M503168200

Handelsman, T., Barak, Y., Nakar, D., Mechaly, A., Lamed, R., Shoham, Y., & Bayer, E. A. (2004). Cohesin-dockerin interaction in cellulosome assembly: a single Asp-to-Asn mutation disrupts high-affinity cohesin-dockerin binding. *FEBS Letters*, *572*(1–3), 195–200. https://doi.org/10.1016/j.febslet.2004.07.040

Hashimoto, C., Kim, D. R., Weiss, L. A., Miller, J. W., & Morisato, D. (2003). Spatial regulation of developmental signaling by a serpin. *Developmental Cell*, *5*(6), 945–950.

Hasunuma, T., Okazaki, F., Okai, N., Hara, K. Y., Ishii, J., & Kondo, A. (2013). A review of enzymes and microbes for lignocellulosic biorefinery and the possibility of their application to consolidated bioprocessing technology. *Bioresource Technology*, *135*, 513–522. https://doi.org/10.1016/j.biortech.2012.10.047

Henrissat, B. (1998). Glycosidase families. *Biochemical Society Transactions*, *26*(2), 153–156.

Henrissat, B., & Davies, G. J. (2000). Glycoside Hydrolases and Glycosyltransferases. Families, Modules, and Implications for Genomics. *Plant Physiology*, *124*(4), 1515–1519. https://doi.org/10.1104/pp.124.4.1515

Hespell, R. B., Akin, D. E., & Dehority, B. A. (1997). Bacteria, fungi and protozoa of the rumen. In R. I. Mackie, B. A. White, & R. Isaacson (Eds.), *Gastrointestinal Microbiology* (Vol. 2, pp. 59–186). New York: Chapman and Hall.

Himmel, M. E., & Bayer, E. A. (2009). Lignocellulose conversion to biofuels: current challenges, global perspectives. *Current Opinion in Biotechnology*, *20*(3), 316–317. https://doi.org/10.1016/j.copbio.2009.05.005

Himmel, M. E., Ding, S.-Y., Johnson, D. K., Adney, W. S., Nimlos, M. R., Brady, J. W., & Foust, T. D. (2007). Biomass Recalcitrance: Engineering Plants and Enzymes for Biofuels Production. *Science*, *315*(5813), 804–807. https://doi.org/10.1126/science.1137016

Hobson, P. N., & Stewart, C. S. (1997). *The Rumen Microbial Ecosystem*. Dordrecht: Springer Netherlands. Retrieved from http://public.eblib.com/choice/publicfullrecord.aspx?p=3102914

Holm, L., & Rosenstrom, P. (2010). Dali server: conservation mapping in 3D. *Nucleic Acids Research*, *38*(Web Server), W545–W549. https://doi.org/10.1093/nar/gkq366

Hong, W., Zhang, J., Feng, Y., Mohr, G., Lambowitz, A. M., Cui, G.-Z., … Cui, Q. (2014). The contribution of cellulosomal scaffoldins to cellulose hydrolysis by Clostridium

thermocellum analyzed by using thermotargetrons. *Biotechnology for Biofuels*, *7*(1), 80. https://doi.org/10.1186/1754-6834-7-80

Horino, H., Fujita, T., & Tonouchi, A. (2014). Description of Anaerobacterium chartisolvens gen. nov., sp. nov., an obligately anaerobic bacterium from Clostridium rRNA cluster III isolated from soil of a Japanese rice field, and reclassification of Bacteroides cellulosolvens Murray *et al.* 1984 as Pseudobacteroides cellulosolvens gen. nov., comb. nov. *International Journal of Systematic and Evolutionary Microbiology*, *64*(Pt 4), 1296–1303. https://doi.org/10.1099/ijs.0.059378-0

Hungate, R. E. (1966). *The rumen and its microbes*. New York: Academic Press. Retrieved from http://site.ebrary.com/id/10954424

Hyeon, J. E., Kang, D. H., & Han, S. O. (2014). Signal amplification by a self-assembled biosensor system designed on the principle of dockerin-cohesin interactions in a cellulosome complex. *The Analyst*, *139*(19), 4790–4793. https://doi.org/10.1039/c4an00856a

Irving, J. A., Steenbakkers, P. J. M., Lesk, A. M., Op den Camp, H. J. M., Pike, R. N., & Whisstock, J. C. (2002). Serpins in prokaryotes. *Molecular Biology and Evolution*, *19*(11), 1881–1890.

Israeli-Ruimy, V., Bule, P., Jindou, S., Dassa, B., Moraïs, S., Borovok, I., … Bayer, E. A. (2017). Complexity of the *Ruminococcus flavefaciens* FD-1 cellulosome reflects an expansion of family-related protein-protein interactions. *Scientific Reports*, *7*, 42355. https://doi.org/10.1038/srep42355

Jamal, S., Nurizzo, D., Boraston, A. B., & Davies, G. J. (2004). X-ray crystal structure of a non-crystalline cellulose-specific carbohydrate-binding module: CBM28. *Journal of Molecular Biology*, *339*(2), 253–258. https://doi.org/10.1016/j.jmb.2004.03.069

Jindou, S., Borovok, I., Rincon, M. T., Flint, H. J., Antonopoulos, D. A., Berg, M. E., … Lamed, R. (2006). Conservation and divergence in cellulosome architecture between two strains of *Ruminococcus flavefaciens*. *Journal of Bacteriology*, *188*(22), 7971–7976. https://doi.org/10.1128/JB.00973-06

Jindou, S., Brulc, J. M., Levy-Assaraf, M., Rincon, M. T., Flint, H. J., Berg, M. E., … Borovok, I. (2008). Cellulosome gene cluster analysis for gauging the diversity of the ruminal cellulolytic bacterium *Ruminococcus flavefaciens*. *FEMS Microbiology Letters*, *285*(2), 188–194. https://doi.org/10.1111/j.1574-6968.2008.01234.x

Jindou, S., Soda, A., Karita, S., Kajino, T., Béguin, P., Wu, J. H. D., … Ohmiya, K. (2004). Cohesin-dockerin interactions within and between Clostridium josui and Clostridium thermocellum: binding selectivity between cognate dockerin and cohesin domains and species specificity. *The Journal of Biological Chemistry*, *279*(11), 9867–9874. https://doi.org/10.1074/jbc.M308673200

Joosten, R. P., Long, F., Murshudov, G. N., & Perrakis, A. (2014). The PDB_REDO server for macromolecular structure model optimization. *IUCrJ*, *1*(Pt 4), 213–220. https://doi.org/10.1107/S2052252514009324

Kabsch, W. (2010). *XDS*. *Acta Crystallographica Section D Biological Crystallography*, *66*(2), 125–132. https://doi.org/10.1107/S0907444909047337

Kakiuchi, M., Isui, A., Suzuki, K., Fujino, T., Fujino, E., Kimura, T., … Ohmiya, K. (1998). Cloning and DNA sequencing of the genes encoding Clostridium josui scaffolding protein CipA and cellulase CelD and identification of their gene products as major components of the cellulosome. *Journal of Bacteriology*, *180*(16), 4303–4308.

Kang, S., Barak, Y., Lamed, R., Bayer, E. A., & Morrison, M. (2006). The functional repertoire of prokaryote cellulosomes includes the serpin superfamily of serine proteinase inhibitors. *Molecular Microbiology*, *60*(6), 1344–1354. https://doi.org/10.1111/j.1365-2958.2006.05182.x

Karmakar, M., & Ray, R. R. (2011). Current Trends in Research and Application of Microbial Cellulases. *Research Journal of Microbiology*, *6*(1), 41–53. https://doi.org/10.3923/jm.2011.41.53

Karpol, A., Barak, Y., Lamed, R., Shoham, Y., & Bayer, E. A. (2008). Functional asymmetry in cohesin binding belies inherent symmetry of the dockerin module: insight into cellulosome assembly revealed by systematic mutagenesis. *The Biochemical Journal*, *410*(2), 331–338. https://doi.org/10.1042/BJ20071193

Karpol, A., Jobby, M. K., Slutzki, M., Noach, I., Chitayat, S., Smith, S. P., & Bayer, E. A. (2013). Structural and functional characterization of a novel type-III dockerin from *Ruminococcus flavefaciens*. *FEBS Letters*, *587*(1), 30–36. https://doi.org/10.1016/j.febslet.2012.11.012

Karpol, A., Kantorovich, L., Demishtein, A., Barak, Y., Morag, E., Lamed, R., & Bayer, E. A. (2009). Engineering a reversible, high-affinity system for efficient protein purification based on the cohesin-dockerin interaction. *Journal of Molecular Recognition: JMR*, *22*(2), 91–98. https://doi.org/10.1002/jmr.926

Kataeva, I. A., Uversky, V. N., Brewer, J. M., Schubot, F., Rose, J. P., Wang, B.-C., & Ljungdahl, L. G. (2004). Interactions between immunoglobulin-like and catalytic modules in Clostridium thermocellum cellulosomal cellobiohydrolase CbhA. *Protein Engineering Design and Selection*, *17*(11), 759–769. https://doi.org/10.1093/protein/gzh094

Keegstra, K. (2010). Plant cell walls. *Plant Physiology*, *154*(2), 483–486. https://doi.org/10.1104/pp.110.161240

Khan, A. W. (1980). Cellulolytic Enzyme System of *Acetivibrio cellulolyticus*, a Newly Isolated Anaerobe. *Microbiology*, *121*(2), 499–502. https://doi.org/10.1099/00221287-121-2-499

Kim, D.-M., Nakazawa, H., Umetsu, M., Matsuyama, T., Ishida, N., Ikeuchi, A., … Kumagai, I. (2012). A nanocluster design for the construction of artificial cellulosomes. *Catalysis Science & Technology*, *2*(3), 499. https://doi.org/10.1039/c2cy00371f

Kirby, J., Aurilia, V., McCrae, S. I., Martin, J. C., & Flint, H. J. (1998). Plant cell wall degrading enzyme complexes from the cellulolytic rumen bacterium *Ruminococcus flavefaciens*. *Biochemical Society Transactions*, *26*(2), S169–S169. https://doi.org/10.1042/bst026s169

Kleine, J., & Liebl, W. (2006). Comparative characterization of deletion derivatives of the modular xylanase XynA of Thermotoga maritima. *Extremophiles: Life Under Extreme Conditions*, *10*(5), 373–381. https://doi.org/10.1007/s00792-006-0509-0

Koike, S., & Kobayashi, Y. (2001). Development and use of competitive PCR assays for the rumen cellulolytic bacteria: Fibrobacter succinogenes, Ruminococcus albus and *Ruminococcus flavefaciens*. *FEMS Microbiology Letters*, *204*(2), 361–366. https://doi.org/10.1111/j.1574-6968.2001.tb10911.x

Krause, D. O., Bunch, R. J., Smith, W. J. M., & McSweeney, C. S. (1999). Diversity of Ruminococcus strains: a survey of genetic polymorphisms and plant digestibility. *Journal of Applied Microbiology*, *86*(3), 487–495. https://doi.org/10.1046/j.1365-2672.1999.00688.x

Krause, D. O., Denman, S. E., Mackie, R. I., Morrison, M., Rae, A. L., Attwood, G. T., & McSweeney, C. S. (2003). Opportunities to improve fiber degradation in the rumen: microbiology, ecology, and genomics. *FEMS Microbiology Reviews*, *27*(5), 663–693. https://doi.org/10.1016/S0168-6445(03)00072-X

Kretsinger, R. H., & Nockolds, C. E. (1973). Carp Muscle Calcium-binding Protein: II. STRUCTURE DETERMINATION AND GENERAL DESCRIPTION. *Journal of Biological Chemistry*, *248*(9), 3313–3326.

Krissinel, E., & Henrick, K. (2004). Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallographica. Section D,*

Biological Crystallography, *60*(Pt 12 Pt 1), 2256–2268. https://doi.org/10.1107/S0907444904026460

Kuhad, R. C., Gupta, R., & Singh, A. (2011). Microbial Cellulases and Their Industrial Applications. *Enzyme Research*, *2011*, 1–10. https://doi.org/10.4061/2011/280696

Lamed, R., Morag (Morgenstern), E., Mor-Yosef, O., & Bayer, E. A. (1991). Cellulosome-like entities inBacteroides cellulosolvens. *Current Microbiology*, *22*(1), 27–33. https://doi.org/10.1007/BF02106209

Lamed, R., Naimark, J., Morgenstern, E., & Bayer, E. A. (1987). Specialized cell surface structures in cellulolytic bacteria. *Journal of Bacteriology*, *169*(8), 3792–3800.

Lamed, R., Setter, E., & Bayer, E. A. (1983). Characterization of a cellulose-binding, cellulase-containing complex in Clostridium thermocellum. *Journal of Bacteriology*, *156*(2), 828–836.

Langer, G., Cohen, S. X., Lamzin, V. S., & Perrakis, A. (2008). Automated macromolecular model building for X-ray crystallography using ARP/wARP version 7. *Nature Protocols*, *3*(7), 1171–1179. https://doi.org/10.1038/nprot.2008.91

Leibovitz, E., & Béguin, P. (1996). A new type of cohesin domain that specifically binds the dockerin domain of the Clostridium thermocellum cellulosome-integrating protein CipA. *Journal of Bacteriology*, *178*(11), 3077–3084.

Leibovitz, E., Ohayon, H., Gounon, P., & Béguin, P. (1997). Characterization and subcellular localization of the Clostridium thermocellum scaffoldin dockerin binding protein SdbA. *Journal of Bacteriology*, *179*(8), 2519–2523.

Lemaire, M., Miras, I., Gounon, P., & Béguin, P. (1998). Identification of a region responsible for binding to the cell wall within the S-layer protein of Clostridium thermocellum. *Microbiology (Reading, England)*, *144 (Pt 1)*, 211–217.

Lemaire, M., Ohayon, H., Gounon, P., Fujino, T., & Béguin, P. (1995). OlpB, a new outer layer protein of Clostridium thermocellum, and binding of its S-layer-like domains to components of the cell envelope. *Journal of Bacteriology*, *177*(9), 2451–2459.

Levy-Assaraf, M., Voronov-Goldman, M., Rozman Grinberg, I., Weiserman, G., Shimon, L. J. W., Jindou, S., … Frolow, F. (2013). Crystal structure of an uncommon cellulosome-related protein module from *Ruminococcus flavefaciens* that resembles papain-like cysteine peptidases. *PloS One*, *8*(2), e56138. https://doi.org/10.1371/journal.pone.0056138

Lombard, V., Bernard, T., Rancurel, C., Brumer, H., Coutinho, P. M., & Henrissat, B. (2010). A hierarchical classification of polysaccharide lyases for glycogenomics. *The Biochemical Journal*, *432*(3), 437–444. https://doi.org/10.1042/BJ20101185

Lombard, V., Golaconda Ramulu, H., Drula, E., Coutinho, P. M., & Henrissat, B. (2014). The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Research*, *42*(D1), D490–D495. https://doi.org/10.1093/nar/gkt1178

Long, F., Vagin, A. A., Young, P., & Murshudov, G. N. (2008). *BALBES* : a molecular-replacement pipeline. *Acta Crystallographica Section D Biological Crystallography*, *64*(1), 125–132. https://doi.org/10.1107/S0907444907050172

Lytle, B. L., Volkman, B. F., Westler, W. M., Heckman, M. P., & Wu, J. H. (2001). Solution structure of a type I dockerin domain, a novel prokaryotic, extracellular calcium-binding domain. *Journal of Molecular Biology*, *307*(3), 745–753. https://doi.org/10.1006/jmbi.2001.4522

Lytle, B., Myers, C., Kruus, K., & Wu, J. H. (1996). Interactions of the CelS binding ligand with various receptor domains of the Clostridium thermocellum cellulosomal scaffolding protein, CipA. *Journal of Bacteriology*, *178*(4), 1200–1203.

Mader, S. S., Windelspecht, M., & Cognato, A. (2013). *Biology*. New York, NY: McGraw-Hill.

Mahalingeswara Bhat, K., & Wood, T. M. (1992). The cellulase of the anaerobic bacterium Clostridium thermocellum: Isolation, dissociation, and reassociation of the cellulosome. *Carbohydrate Research*, *227*, 293–300. https://doi.org/10.1016/0008-6215(92)85079-F

Matano, Y., Hasunuma, T., & Kondo, A. (2012). Display of cellulases on the cell surface of Saccharomyces cerevisiae for high yield ethanol production from high-solid lignocellulosic biomass. *Bioresource Technology*, *108*, 128–133. https://doi.org/10.1016/j.biortech.2011.12.144

Matthews, B. W. (1968). Solvent content of protein crystals. *Journal of Molecular Biology*, *33*(2), 491–497. https://doi.org/10.1016/0022-2836(68)90205-2

McCoy, A. J., Grosse-Kunstleve, R. W., Adams, P. D., Winn, M. D., Storoni, L. C., & Read, R. J. (2007). Phaser crystallographic software. *Journal of Applied Crystallography*, *40*(Pt 4), 658–674. https://doi.org/10.1107/S0021889807021206

McCoy, J., & La Ville, E. (2001). Expression and purification of thioredoxin fusion proteins. *Current Protocols in Protein Science*, *Chapter 6*, Unit 6.7. https://doi.org/10.1002/0471140864.ps0607s10

Mechaly, A., Fierobe, H. P., Belaich, A., Belaich, J. P., Lamed, R., Shoham, Y., & Bayer, E. A. (2001). Cohesin-dockerin interaction in cellulosome assembly: a single hydroxyl group of a dockerin domain distinguishes between nonrecognition and high affinity recognition. *The Journal of Biological Chemistry*, *276*(13), 9883–9888. https://doi.org/10.1074/jbc.M009237200

Meng, E. C., Pettersen, E. F., Couch, G. S., Huang, C. C., & Ferrin, T. E. (2006). Tools for integrated sequence-structure analysis with UCSF Chimera. *BMC Bioinformatics*, *7*, 339. https://doi.org/10.1186/1471-2105-7-339

Meng, X., & Ragauskas, A. J. (2014). Recent advances in understanding the role of cellulose accessibility in enzymatic hydrolysis of lignocellulosic substrates. *Current Opinion in Biotechnology*, *27*, 150–158. https://doi.org/10.1016/j.copbio.2014.01.014

Miras, I., Schaeffer, F., Béguin, P., & Alzari, P. M. (2002). Mapping by site-directed mutagenesis of the region responsible for cohesin-dockerin interaction on the surface of the seventh cohesin domain of Clostridium thermocellum CipA. *Biochemistry*, *41*(7), 2115–2119.

Mitsuzawa, S., Kagawa, H., Li, Y., Chan, S. L., Paavola, C. D., & Trent, J. D. (2009). The rosettazyme: a synthetic cellulosome. *Journal of Biotechnology*, *143*(2), 139–144. https://doi.org/10.1016/j.jbiotec.2009.06.019

Mohnen, D. (2008). Pectin structure and biosynthesis. *Current Opinion in Plant Biology*, *11*(3), 266–277. https://doi.org/10.1016/j.pbi.2008.03.006

Morag, E., Bayer, E. A., & Lamed, R. (1991). Anomalous dissociative behavior of the major glycosylated component of the cellulosome of Clostridium thermocellum. *Applied Biochemistry and Biotechnology*, *30*(2), 129–136.

Morag, E., Halevy, I., Bayer, E. A., & Lamed, R. (1991). Isolation and properties of a major cellobiohydrolase from the cellulosome of Clostridium thermocellum. *Journal of Bacteriology*, *173*(13), 4155–4162.

Morag, E., Lapidot, A., Govorko, D., Lamed, R., Wilchek, M., Bayer, E. A., & Shoham, Y. (1995). Expression, purification, and characterization of the cellulose-binding domain of the scaffoldin subunit from the cellulosome of Clostridium thermocellum. *Applied and Environmental Microbiology*, *61*(5), 1980–1986.

Moraïs, S., Ben David, Y., Bensoussan, L., Duncan, S. H., Koropatkin, N. M., Martens, E. C., … Bayer, E. A. (2016). Enzymatic profiling of cellulosomal enzymes from the human gut bacterium, Ruminococcus champanellensis, reveals a fine-tuned system for cohesin-dockerin recognition. *Environmental Microbiology*, *18*(2), 542–556. https://doi.org/10.1111/1462-2920.13047

Moraïs, S., Heyman, A., Barak, Y., Caspi, J., Wilson, D. B., Lamed, R., … Bayer, E. A. (2010). Enhanced cellulose degradation by nano-complexed enzymes: Synergism between a scaffold-linked exoglucanase and a free endoglucanase. *Journal of Biotechnology*, *147*(3–4), 205–211. https://doi.org/10.1016/j.jbiotec.2010.04.012

Moraïs, S., Morag, E., Barak, Y., Goldman, D., Hadar, Y., Lamed, R., … Bayer, E. A. (2012). Deconstruction of lignocellulose into soluble sugars by native and designer cellulosomes. *mBio*, *3*(6). https://doi.org/10.1128/mBio.00508-12

Mori, Y. (1992). Purification and characterization of an endoglucanase from the cellulosomes (multicomponent cellulase complexes) of Clostridium thermocellum. *Bioscience, Biotechnology, and Biochemistry*, *56*(8), 1198–1203.

Mori, Y., Ozasa, S., Kitaoka, M., Noda, S., Tanaka, T., Ichinose, H., & Kamiya, N. (2013). Aligning an endoglucanase Cel5A from Thermobifida fusca on a DNA scaffold: potent design of an artificial cellulosome. *Chemical Communications (Cambridge, England)*, *49*(62), 6971–6973. https://doi.org/10.1039/c3cc42614a

Mosbah, A., Belaïch, A., Bornet, O., Belaïch, J.-P., Henrissat, B., & Darbon, H. (2000). Solution structure of the module X2_1 of unknown function of the cellulosomal scaffolding protein CipC of Clostridium cellulolyticum. *Journal of Molecular Biology*, *304*(2), 201–217. https://doi.org/10.1006/jmbi.2000.4192

Murshudov, G. N., Skubák, P., Lebedev, A. A., Pannu, N. S., Steiner, R. A., Nicholls, R. A., … Vagin, A. A. (2011). *REFMAC* 5 for the refinement of macromolecular crystal structures. *Acta Crystallographica Section D Biological Crystallography*, *67*(4), 355–367. https://doi.org/10.1107/S0907444911001314

Nash, M. A., Smith, S. P., Fontes, C. M., & Bayer, E. A. (2016). Single versus dual-binding conformations in cellulosomal cohesin–dockerin complexes. *Current Opinion in Structural Biology*, *40*, 89–96. https://doi.org/10.1016/j.sbi.2016.08.002

Navarre, W. W., & Schneewind, O. (1994). Proteolytic cleavage and cell wall anchoring at the LPXTG motif of surface proteins in Gram-positive bacteria. *Molecular Microbiology*, *14*(1), 115–121. https://doi.org/10.1111/j.1365-2958.1994.tb01271.x

Ndeh, D., Rogowski, A., Cartmell, A., Luis, A. S., Baslé, A., Gray, J., … Gilbert, H. J. (2017). Complex pectin metabolism by gut bacteria reveals novel catalytic functions. *Nature*. https://doi.org/10.1038/nature21725

Noach, I., Frolow, F., Alber, O., Lamed, R., Shimon, L. J. W., & Bayer, E. A. (2009). Intermodular linker flexibility revealed from crystal structures of adjacent cellulosomal cohesins of *Acetivibrio cellulolyticus*. *Journal of Molecular Biology*, *391*(1), 86–97. https://doi.org/10.1016/j.jmb.2009.06.006

Noach, I., Frolow, F., Jakoby, H., Rosenheck, S., Shimon, L. W., Lamed, R., & Bayer, E. A. (2005). Crystal structure of a type-II cohesin module from the Bacteroides cellulosolvens cellulosome reveals novel and distinctive secondary structural elements. *Journal of Molecular Biology*, *348*(1), 1–12. https://doi.org/10.1016/j.jmb.2005.02.024

Noach, I., Lamed, R., Xu, Q., Rosenheck, S., Shimon, L. J. W., Bayer, E. A., & Frolow, F. (2003). Preliminary X-ray characterization and phasing of a type II cohesin domain from the cellulosome of *Acetivibrio cellulolyticus*. *Acta Crystallographica. Section D, Biological Crystallography*, *59*(Pt 9), 1670–1673.

Noach, I., Levy-Assaraf, M., Lamed, R., Shimon, L. J. W., Frolow, F., & Bayer, E. A. (2010). Modular arrangement of a cellulosomal scaffoldin subunit revealed from the crystal structure of a cohesin dyad. *Journal of Molecular Biology*, *399*(2), 294–305. https://doi.org/10.1016/j.jmb.2010.04.013

O Cuív, P., Gupta, R., Goswami, H. P., & Morrison, M. (2013). Extending the cellulosome paradigm: the modular Clostridium thermocellum cellulosomal serpin PinA is a broad-spectrum inhibitor of subtilisin-like proteases. *Applied and Environmental Microbiology*, *79*(19), 6173–6175. https://doi.org/10.1128/AEM.01912-13

Ohara, H., Karita, S., Kimura, T., Sakka, K., & Ohmiya, K. (2000). Characterization of the Cellulolytic Complex (Cellulosome) from *Ruminococcus albus*. *Bioscience, Biotechnology, and Biochemistry*, *64*(2), 254–260. https://doi.org/10.1271/bbb.64.254

Ohara, H., Noguchi, J., Karita, S., Kimura, T., Sakka, K., & Ohmiya, K. (2000). Sequence of *egV* and Properties of EgV, a *Ruminococcus albus* Endoglucanase Containing a

Dockerin Domain. *Bioscience, Biotechnology, and Biochemistry*, *64*(1), 80–88. https://doi.org/10.1271/bbb.64.80

Osiro, K. O., de Camargo, B. R., Satomi, R., Hamann, P. R. V., Silva, J. P., de Sousa, M. V., … Noronha, E. F. (2017). Characterization of Clostridium thermocellum (B8) secretome and purified cellulosomes for lignocellulosic biomass degradation. *Enzyme and Microbial Technology*, *97*, 43–54. https://doi.org/10.1016/j.enzmictec.2016.11.002

O'Sullivan, A. C. (1997). Cellulose: the structure slowly unravels. *Cellulose*, *4*(3), 173–207. https://doi.org/10.1023/A:1018431705579

Pabst, M., Fischl, R. M., Brecker, L., Morelle, W., Fauland, A., Köfeler, H., … Léonard, R. (2013). Rhamnogalacturonan II structure shows variation in the side chains monosaccharide composition and methylation status within and across different plant species. *The Plant Journal*, n/a-n/a. https://doi.org/10.1111/tpj.12271

Pagès, S., Bélaïch, A., Bélaïch, J. P., Morag, E., Lamed, R., Shoham, Y., & Bayer, E. A. (1997). Species-specificity of the cohesin-dockerin interaction between Clostridium thermocellum and Clostridium cellulolyticum: prediction of specificity determinants of the dockerin domain. *Proteins*, *29*(4), 517–527.

Pagès, S., Bélaïch, A., Fierobe, H. P., Tardif, C., Gaudin, C., & Bélaïch, J. P. (1999). Sequence analysis of scaffolding protein CipC and ORFXp, a new cohesin-containing protein in Clostridium cellulolyticum: comparison of various cohesin domains and subcellular localization of ORFXp. *Journal of Bacteriology*, *181*(6), 1801–1810.

Paterson, G. K., & Mitchell, T. J. (2004). The biology of Gram-positive sortase enzymes. *Trends in Microbiology*, *12*(2), 89–95. https://doi.org/10.1016/j.tim.2003.12.007

Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., & Ferrin, T. E. (2004). UCSF Chimera--a visualization system for exploratory research and analysis. *Journal of Computational Chemistry*, *25*(13), 1605–1612. https://doi.org/10.1002/jcc.20084

Pinheiro, B. A., Brás, J. L. A., Najmudin, S., Carvalho, A. L., Ferreira, L. M. A., Prates, J. A. M., & Fontes, C. M. G. A. (2012). Flexibility and specificity of the cohesin–dockerin interaction: implications for cellulosome assembly and functionality. *Biocatalysis and Biotransformation*, 1–7. https://doi.org/10.3109/10242422.2012.681854

Pinheiro, B. A., Gilbert, H. J., Sakka, K., Sakka, K., Fernandes, V. O., Prates, J. A. M., … Fontes, C. M. G. A. (2009). Functional insights into the role of novel type I cohesin and dockerin domains from Clostridium thermocellum. *The Biochemical Journal*, *424*(3), 375–384. https://doi.org/10.1042/BJ20091152

Pinheiro, B. A., Proctor, M. R., Martinez-Fleites, C., Prates, J. A. M., Money, V. A., Davies, G. J., … Gilbert, H. J. (2008). The Clostridium cellulolyticum dockerin displays a dual binding mode for its cohesin partner. *The Journal of Biological Chemistry*, *283*(26), 18422–18430. https://doi.org/10.1074/jbc.M801533200

Poole, D. M., Morag, E., Lamed, R., Bayer, E. A., Hazlewood, G. P., & Gilbert, H. J. (1992). Identification of the cellulose-binding domain of the cellulosome subunit S1 from Clostridium thermocellum YS. *FEMS Microbiology Letters*, *78*(2–3), 181–186.

Popper, Z. A. (2008). Evolution and diversity of green plant cell walls. *Current Opinion in Plant Biology*, *11*(3), 286–292. https://doi.org/10.1016/j.pbi.2008.02.012

Popper, Z. A., Michel, G., Hervé, C., Domozych, D. S., Willats, W. G. T., Tuohy, M. G., … Stengel, D. B. (2011). Evolution and diversity of plant cell walls: from algae to flowering plants. *Annual Review of Plant Biology*, *62*, 567–590. https://doi.org/10.1146/annurev-arplant-042110-103809

Poudel, S., Giannone, R. J., Rodriguez, M., Raman, B., Martin, M. Z., Engle, N. L., … Hettich, R. L. (2017). Integrated omics analyses reveal the details of metabolic adaptation of Clostridium thermocellum to lignocellulose-derived growth inhibitors released during the deconstruction of switchgrass. *Biotechnology for Biofuels*, *10*(1). https://doi.org/10.1186/s13068-016-0697-5

Ragauskas, A. J., Williams, C. K., Davison, B. H., Britovsek, G., Cairney, J., Eckert, C. A., … Tschaplinski, T. (2006). The path forward for biofuels and biomaterials. *Science (New York, N.Y.)*, *311*(5760), 484–489. https://doi.org/10.1126/science.1114736

Ramakrishnan, V., Finch, J. T., Graziano, V., Lee, P. L., & Sweet, R. M. (1993). Crystal structure of globular domain of histone H5 and its implications for nucleosome binding. *Nature*, *362*(6417), 219–223. https://doi.org/10.1038/362219a0

Raman, B., Pan, C., Hurst, G. B., Rodriguez, M., Jr, McKeown, C. K., Lankford, P. K., … Mielenz, J. R. (2009). Impact of pretreated Switchgrass and biomass carbohydrates on Clostridium thermocellum ATCC 27405 cellulosome composition: a quantitative proteomic analysis. *PloS One*, *4*(4), e5271. https://doi.org/10.1371/journal.pone.0005271

Ravachol, J., Borne, R., Meynial-Salles, I., Soucaille, P., Pagès, S., Tardif, C., & Fierobe, H.-P. (2015). Combining free and aggregated cellulolytic systems in the cellulosome-producing bacterium Ruminiclostridium cellulolyticum. *Biotechnology for Biofuels*, *8*, 114. https://doi.org/10.1186/s13068-015-0301-4

Ravachol, J., Borne, R., Tardif, C., de Philip, P., & Fierobe, H.-P. (2014). Characterization of all family-9 glycoside hydrolases synthesized by the cellulosome-producing bacterium Clostridium cellulolyticum. *The Journal of Biological Chemistry*, *289*(11), 7335–7348. https://doi.org/10.1074/jbc.M113.545046

Ribeiro, T., Ponte, P. I. P., Guerreiro, C. I. P. D., Santos, H. M., Falcão, L., Freire, J. P. B., … Lordelo, M. M. (2008). A family 11 carbohydrate-binding module (CBM) improves the efficacy of a recombinant cellulase used to supplement barley-based diets for broilers at lower dosage rates. *British Poultry Science*, *49*(5), 600–608. https://doi.org/10.1080/00071660802345749

Ridley, B. L., O'Neill, M. A., & Mohnen, D. (2001). Pectins: structure, biosynthesis, and oligogalacturonide-related signaling. *Phytochemistry*, *57*(6), 929–967. https://doi.org/10.1016/S0031-9422(01)00113-3

Rincon, M. T., Cepeljnik, T., Martin, J. C., Barak, Y., Lamed, R., Bayer, E. A., & Flint, H. J. (2007). A novel cell surface-anchored cellulose-binding protein encoded by the sca gene cluster of *Ruminococcus flavefaciens*. *Journal of Bacteriology*, *189*(13), 4774–4783. https://doi.org/10.1128/JB.00143-07

Rincon, M. T., Cepeljnik, T., Martin, J. C., Lamed, R., Barak, Y., Bayer, E. A., & Flint, H. J. (2005). Unconventional Mode of Attachment of the *Ruminococcus flavefaciens* Cellulosome to the Cell Surface. *J. Bacteriol.*, *187*(22), 7569–7578. https://doi.org/10.1128/jb.187.22.7569-7578.2005

Rincon, M. T., Dassa, B., Flint, H. J., Travis, A. J., Jindou, S., Borovok, I., … White, B. A. (2010). Abundance and diversity of dockerin-containing proteins in the fiber-degrading rumen bacterium, *Ruminococcus flavefaciens* FD-1. *PloS One*, *5*(8), e12476. https://doi.org/10.1371/journal.pone.0012476

Rincon, M. T., Ding, S.-Y., McCrae, S. I., Martin, J. C., Aurilia, V., Lamed, R., … Flint, H. J. (2003). Novel organization and divergent dockerin specificities in the cellulosome system of *Ruminococcus flavefaciens*. *Journal of Bacteriology*, *185*(3), 703–713.

Rincon, M. T., Martin, J. C., Aurilia, V., McCrae, S. I., Rucklidge, G. J., Reid, M. D., … Flint, H. J. (2004). ScaC, an Adaptor Protein Carrying a Novel Cohesin That Expands the Dockerin-Binding Repertoire of the *Ruminococcus flavefaciens* 17 Cellulosome. *Journal of Bacteriology*, *186*(9), 2576–2585. https://doi.org/10.1128/JB.186.9.2576-2585.2004

Rosenberg, E. (2013). *The prokaryotes -prokaryotic physiology and biochemistry: prokaryotic physiology and biochemistry* (4th ed). New York: Springer.

Ruimy, V. (2013). *Organization of the Ruminococcus flavefaciens FD-1 cellulosome and features dictating recognition of its cohesin-dockerin interaction* (MsC Thesis). The Weizmann Institute of Science, Rehovot, Israel.

Sakka, K., Sugihara, Y., Jindou, S., Sakka, M., Inagaki, M., Sakka, K., & Kimura, T. (2011). Analysis of cohesin-dockerin interactions using mutant dockerin proteins. *FEMS Microbiology Letters*, *314*(1), 75–80. https://doi.org/10.1111/j.1574-6968.2010.02146.x

Salama-Alber, O., Gat, Y., Lamed, R., Shimon, L. J. W., Bayer, E. A., & Frolow, F. (2012). Crystallization and preliminary X-ray characterization of a type III cohesin-dockerin complex from the cellulosome system of *Ruminococcus flavefaciens*. *Acta Crystallographica. Section F, Structural Biology and Crystallization Communications*, *68*(Pt 9), 1116–1119. https://doi.org/10.1107/S1744309112033088

Salama-Alber, O., Jobby, M. K., Chitayat, S., Smith, S. P., White, B. A., Shimon, L. J. W., … Bayer, E. A. (2013). Atypical Cohesin-Dockerin Complex Responsible for Cell Surface Attachment of Cellulosomal Components: BINDING FIDELITY, PROMISCUITY, AND STRUCTURAL BUTTRESSES. *Journal of Biological Chemistry*, *288*(23), 16827–16838. https://doi.org/10.1074/jbc.M113.466672

Salamitou, S., Lemaire, M., Fujino, T., Ohayon, H., Gounon, P., Béguin, P., & Aubert, J. P. (1994). Subcellular localization of Clostridium thermocellum ORF3p, a protein carrying a receptor for the docking sequence borne by the catalytic components of the cellulosome. *Journal of Bacteriology*, *176*(10), 2828–2834.

Salamitou, S., Tokatlidis, K., Béguin, P., & Aubert, J. P. (1992). Involvement of separate domains of the cellulosomal protein S1 of Clostridium thermocellum in binding to cellulose and in anchoring of catalytic subunits to the cellulosome. *FEBS Letters*, *304*(1), 89–92.

Schaeffer, F., Matuschek, M., Guglielmi, G., Miras, I., Alzari, P. M., & Béguin, P. (2002). Duplicated dockerin subdomains of Clostridium thermocellum endoglucanase CelD bind to a cohesin domain of the scaffolding protein CipA with distinct thermodynamic parameters and a negative cooperativity. *Biochemistry*, *41*(7), 2106–2114.

Scheller, H. V., & Ulvskov, P. (2010). Hemicelluloses. *Annual Review of Plant Biology*, *61*, 263–289. https://doi.org/10.1146/annurev-arplant-042809-112315

Schoeler, C., Bernardi, R. C., Malinowska, K. H., Durner, E., Ott, W., Bayer, E. A., … Gaub, H. E. (2015). Mapping Mechanical Force Propagation through Biomolecular Complexes. *Nano Letters*, *15*(11), 7370–7376. https://doi.org/10.1021/acs.nanolett.5b02727

Schoeler, C., Malinowska, K. H., Bernardi, R. C., Milles, L. F., Jobst, M. A., Durner, E., … Nash, M. A. (2014). Ultrastable cellulosome-adhesion complex tightens under load. *Nature Communications*, *5*, 5635. https://doi.org/10.1038/ncomms6635

Schubert, C. (2006). Can biofuels finally take center stage? *Nature Biotechnology*, *24*(7), 777–784. https://doi.org/10.1038/nbt0706-777

Schubot, F. D., Kataeva, I. A., Chang, J., Shah, A. K., Ljungdahl, L. G., Rose, J. P., & Wang, B.-C. (2004). Structural Basis for the Exocellulase Activity of the Cellobiohydrolase CbhA from *Clostridium thermocellum* †. *Biochemistry*, *43*(5), 1163–1170. https://doi.org/10.1021/bi030202i

Shimon, L. J., Frolow, F., Yaron, S., Bayer, E. A., Lamed, R., Morag, E., & Shoham, Y. (1997). Crystallization and preliminary X-ray analysis of a cohesin domain of the cellulosome from Clostridium thermocellum. *Acta Crystallographica. Section D, Biological Crystallography*, *53*(Pt 1), 114–115. https://doi.org/10.1107/S090744499601164X

Shimon, L. J., Pagès, S., Belaich, A., Belaich, J. P., Bayer, E. A., Lamed, R., … Frolow, F. (2000). Structure of a family IIIa scaffoldin CBD from the cellulosome of Clostridium cellulolyticum at 2.2 A resolution. *Acta Crystallographica. Section D, Biological Crystallography*, *56*(Pt 12), 1560–1568.

Shoham, Y., Lamed, R., & Bayer, E. A. (1999). The cellulosome concept as an efficient microbial strategy for the degradation of insoluble polysaccharides. *Trends in Microbiology*, *7*(7), 275–281.

Shoseyov, O., Takagi, M., Goldstein, M. A., & Doi, R. H. (1992). Primary sequence analysis of Clostridium cellulovorans cellulose binding protein A. *Proceedings of the National Academy of Sciences of the United States of America*, *89*(8), 3483–3487.

Showalter, A. M. (1993). Structure and Function of Plant Cell Wall Proteins. *THE PLANT CELL ONLINE*, *5*(1), 9–23. https://doi.org/10.1105/tpc.5.1.9

Shterzer, N., & Mizrahi, I. (2015). The animal gut as a melting pot for horizontal gene transfer [1]. *Canadian Journal of Microbiology*, *61*(9), 603–605. https://doi.org/10.1139/cjm-2015-0049

Simpson, P. J., Xie, H., Bolam, D. N., Gilbert, H. J., & Williamson, M. P. (2000). The Structural Basis for the Ligand Specificity of Family 2 Carbohydrate-binding Modules. *Journal of Biological Chemistry*, *275*(52), 41137–41142. https://doi.org/10.1074/jbc.M006948200

Slutzki, M., Barak, Y., Reshef, D., Schueler-Furman, O., Lamed, R., & Bayer, E. A. (2012). Measurements of relative binding of cohesin and dockerin mutants using an advanced ELISA technique for high-affinity interactions. *Methods in Enzymology*, *510*, 417–428. https://doi.org/10.1016/B978-0-12-415931-0.00022-7

Slutzki, M., Jobby, M. K., Chitayat, S., Karpol, A., Dassa, B., Barak, Y., … Bayer, E. A. (2013). Intramolecular clasp of the cellulosomal *Ruminococcus flavefaciens* ScaA dockerin module confers structural stability. *FEBS Open Bio*, *3*, 398–405. https://doi.org/10.1016/j.fob.2013.09.006

Smith, S. P., & Bayer, E. A. (2013). Insights into cellulosome assembly and dynamics: from dissection to reconstruction of the supramolecular enzyme complex. *Current Opinion in Structural Biology*, *23*(5), 686–694. https://doi.org/10.1016/j.sbi.2013.09.002

Somerville, C. (2006). Cellulose Synthesis in Higher Plants. *Annual Review of Cell and Developmental Biology*, *22*(1), 53–78. https://doi.org/10.1146/annurev.cellbio.22.022206.160206

Somerville, C., Bauer, S., Brininstool, G., Facette, M., Hamann, T., Milne, J., … Youngs, H. (2004). Toward a systems approach to understanding plant cell walls. *Science (New York, N.Y.)*, *306*(5705), 2206–2211. https://doi.org/10.1126/science.1102765

Spinelli, S., Fiérobe, H.-P., Belaïch, A., Belaïch, J.-P., Henrissat, B., & Cambillau, C. (2000). Crystal structure of a cohesin module from Clostridium cellulolyticum: implications for dockerin recognition. *Journal of Molecular Biology*, *304*(2), 189–200. https://doi.org/10.1006/jmbi.2000.4191

Stahl, S. W., Nash, M. A., Fried, D. B., Slutzki, M., Barak, Y., Bayer, E. A., & Gaub, H. E. (2012). Single-molecule dissection of the high-affinity cohesin-dockerin complex. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(50), 20431–20436. https://doi.org/10.1073/pnas.1211929109

Steenbakkers, P. J. M., Irving, J. A., Harhangi, H. R., Swinkels, W. J. C., Akhmanova, A., Dijkerman, R., … Op den Camp, H. J. M. (2008). A serpin in the cellulosome of the anaerobic fungus Piromyces sp. strain E2. *Mycological Research*, *112*(Pt 8), 999–1006. https://doi.org/10.1016/j.mycres.2008.01.021

Stern, J., Moraïs, S., Lamed, R., & Bayer, E. A. (2016). Adaptor Scaffoldins: An Original Strategy for Extended Designer Cellulosomes, Inspired from Nature. *mBio*, *7*(2), e00083. https://doi.org/10.1128/mBio.00083-16

Terwilliger, T. C., Adams, P. D., Read, R. J., McCoy, A. J., Moriarty, N. W., Grosse-Kunstleve, R. W., … Hung, L.-W. (2009). Decision-making in structure solution using Bayesian estimates of map quality: the *PHENIX AutoSol* wizard. *Acta Crystallographica Section D Biological Crystallography*, *65*(6), 582–601. https://doi.org/10.1107/S0907444909012098

Terwilliger, T. C., Grosse-Kunstleve, R. W., Afonine, P. V., Moriarty, N. W., Zwart, P. H., Hung, L. W., … Adams, P. D. (2008). Iterative model building, structure refinement and density modification with the PHENIX AutoBuild wizard. *Acta Crystallographica.*

Section D, Biological Crystallography, *64*(Pt 1), 61–69. https://doi.org/10.1107/S090744490705024X

Tokatlidis, K., Dhurjati, P., & Béguin, P. (1993). Properties conferred on Clostridium thermocellum endoglucanase CelC by grafting the duplicated segment of endoglucanase CelD. *Protein Engineering*, *6*(8), 947–952.

Tokatlidis, K., Salamitou, S., Béguin, P., Dhurjati, P., & Aubert, J. P. (1991). Interaction of the duplicated segment carried by Clostridium thermocellum cellulases with cellulosome components. *FEBS Letters*, *291*(2), 185–188.

Tomme, P., Warren, R. A. J., Gilkes, N. R., & Poole, R. K. (1995). Cellulose Hydrolysis by Bacteria and Fungi. In *Advances in Microbial Physiology* (Vol. Volume 37, pp. 1–81). Academic Press. Retrieved from http://www.sciencedirect.com/science/article/B7CTD-4SD21X0-5/2/1926adaee22aa17e2bb16766ddbb3aba

Tormo, J., Lamed, R., Chirino, A. J., Morag, E., Bayer, E. A., Shoham, Y., & Steitz, T. A. (1996). Crystal structure of a bacterial family-III cellulose-binding domain: a general mechanism for attachment to cellulose. *The EMBO Journal*, *15*(21), 5739–5751.

Valbuena, A., Oroz, J., Hervás, R., Vera, A. M., Rodríguez, D., Menéndez, M., … Carrión-Vázquez, M. (2009). On the remarkable mechanostability of scaffoldins and the mechanical clamp motif. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(33), 13791–13796. https://doi.org/10.1073/pnas.0813093106

Vazana, Y., Barak, Y., Unger, T., Peleg, Y., Shamshoum, M., Ben-Yehezkel, T., … Bayer, E. A. (2013). A synthetic biology approach for evaluating the functional contribution of designer cellulosome components to deconstruction of cellulosic substrates. *Biotechnology for Biofuels*, *6*(1), 182. https://doi.org/10.1186/1754-6834-6-182

Venditto, I., Bule, P., Thompson, A., Sanchez-Weatherby, J., Sandy, J., Ferreira, L. M. A., … Najmudin, S. (2015). Expression, purification, crystallization and preliminary X-ray analysis of CttA, a putative cellulose-binding protein from *Ruminococcus flavefaciens*. *Acta Crystallographica Section F Structural Biology Communications*, *71*(6), 784–789. https://doi.org/10.1107/S2053230X15008249

Waeonukul, R., Pason, P., Kyu, K. L., Sakka, K., Kosugi, A., Mori, Y., & Ratanakhanokchai, K. (2009). Cloning, sequencing, and expression of the gene encoding a multidomain endo-beta-1,4-xylanase from Paenibacillus curdlanolyticus B-6, and characterization of the recombinant enzyme. *Journal of Microbiology and Biotechnology*, *19*(3), 277–285.

Warren, R. A. J. (1996). MICROBIAL HYDROLYSIS OF POLYSACCHARIDES. *Annual Review of Microbiology*, *50*(1), 183–212. https://doi.org/10.1146/annurev.micro.50.1.183

Waters, M. L. (2002). Aromatic interactions in model systems. *Current Opinion in Chemical Biology*, *6*(6), 736–741. https://doi.org/10.1016/S1367-5931(02)00359-9

Weimer, P. J. (1996). Why Don't Ruminal Bacteria Digest Cellulose Faster? *Journal of Dairy Science*, *79*(8), 1496–1502. https://doi.org/10.3168/jds.S0022-0302(96)76509-8

Weinstein, J. Y., Slutzki, M., Karpol, A., Barak, Y., Gul, O., Lamed, R., … Fried, D. B. (2015). Insights into a type III cohesin-dockerin recognition interface from the cellulose-degrading bacterium *Ruminococcus flavefaciens*: COHESIN-DOCKERIN RECOGNITION IN *Ruminococcus flavefaciens*. *Journal of Molecular Recognition*, *28*(3), 148–154. https://doi.org/10.1002/jmr.2380

White, B. A., Lamed, R., Bayer, E. A., & Flint, H. J. (2014). Biomass utilization by gut microbiomes. *Annual Review of Microbiology*, *68*, 279–296. https://doi.org/10.1146/annurev-micro-092412-155618

Wilchek, M., Bayer, E. A., & Livnah, O. (2006). Essentials of biorecognition: The (strept)avidin–biotin system as a model for protein–protein and protein–ligand interaction. *Immunology Letters*, *103*(1), 27–32. https://doi.org/10.1016/j.imlet.2005.10.022

Winn, M. D., Ballard, C. C., Cowtan, K. D., Dodson, E. J., Emsley, P., Evans, P. R., … Wilson, K. S. (2011). Overview of the *CCP* 4 suite and current developments. *Acta Crystallographica Section D Biological Crystallography*, *67*(4), 235–242. https://doi.org/10.1107/S0907444910045749

Winter, G. (2010). *xia2* : an expert system for macromolecular crystallography data reduction. *Journal of Applied Crystallography*, *43*(1), 186–190. https://doi.org/10.1107/S0021889809045701

Winter, G., & McAuley, K. E. (2011). Automated data collection for macromolecular crystallography. *Methods*, *55*(1), 81–93. https://doi.org/10.1016/j.ymeth.2011.06.010

Xu, Q., Barak, Y., Kenig, R., Shoham, Y., Bayer, E. A., & Lamed, R. (2004). A Novel *Acetivibrio cellulolyticus* Anchoring Scaffoldin That Bears Divergent Cohesins. *Journal of Bacteriology*, *186*(17), 5782–5789. https://doi.org/10.1128/JB.186.17.5782-5789.2004

Xu, Q., Bayer, E. A., Goldman, M., Kenig, R., Shoham, Y., & Lamed, R. (2004). Architecture of the Bacteroides cellulosolvens cellulosome: description of a cell surface-anchoring scaffoldin and a family 48 cellulase. *Journal of Bacteriology*, *186*(4), 968–977.

Xu, Q., Gao, W., Ding, S.-Y., Kenig, R., Shoham, Y., Bayer, E. A., & Lamed, R. (2003). The Cellulosome System of *Acetivibrio cellulolyticus* Includes a Novel Type of Adaptor Protein and a Cell Surface Anchoring Protein. *Journal of Bacteriology*, *185*(15), 4548–4557. https://doi.org/10.1128/JB.185.15.4548-4557.2003

Xu, Q., Resch, M. G., Podkaminer, K., Yang, S., Baker, J. O., Donohoe, B. S., … Bomble, Y. J. (2016). Dramatic performance of Clostridium thermocellum explained by its wide range of cellulase modalities. *Science Advances*, *2*(2), e1501254. https://doi.org/10.1126/sciadv.1501254

Zhao, G., Li, H., Wamalwa, B., Sakka, M., Kimura, T., & Sakka, K. (2006). Different binding specificities of S-layer homology modules from Clostridium thermocellum AncA, Slp1, and Slp2. *Bioscience, Biotechnology, and Biochemistry*, *70*(7), 1636–1641. https://doi.org/10.1271/bbb.50699

Annexes

Table S2. 1 Dockerin modules of *R. flavefaciens* strain FD-1 selected for the microarray study.

Accession No.	Group	Architecture of parent protein	Primers used
ZP_06142678	1a	SIGN-GH9-CBM3-Doc	5' gctac**ggtacc**t GAG CGT GTT ACT CTG TGG 3' cgccag***ggatcc***TTA TCA GTT ATA GCT CTC GGG
ZP_06142769	1a	SIGN-GH11-CBM22-GH10-Doc-CBM22-CE4	5' gctac**ggtacc**t GTA ACA CTC TGG GGC GAT GCT 3' cgccag***ggatcc*** TTA TGC GAT ATA TGT CTT ATT TGA TGC
ZP_06142857	1a	SIGN-GH11-CBM22-Doc-GH11-CE3	5' gctac**ggtacc**t ACA CTC TGG GGC GAT GCC 3' cgccag***ggatcc*** TTA CTG ATA ATT TGA TCT TGA GGC
ZP_06142983	1a	SIGN-UNK-CE12-CBM13-Doc-CBM35-CE12	5' gctac**ggtacc**t GAG GCT GTT CAG AAG TTC 3' cgccag***ggatcc***TTA TTC GGG CTC ATA GTA AAC
ZP_06144535	1a	SIGN-Coh- Doc (ScaO)	5' gctac**ggtacc**t TCT GTA ACT TCA ACA GTC AAA G 3' cgccag***ggatcc*** TTAACT CTC CAC AAA CTC CCA GT
ZP_06145360	1a	SIGN-GH48-Doc	5' gctac**ggtacc**t GTT CTC TGG GGC GAT GCT 3' cgccag***ggatcc***TTA TGA CTC AGG GAG CTT AGT
ZP_06145505	1a	SIGN-Coh-Doc (ScaM)	5' gctac**ggtacc**t TTA GAG ATA GTT CTT GAT GAA CC 3' cgccagggatccTTA ATC AAG CTT CAG CAG TTT TTT C
ZP_06142866	1b	SIGN-GH9-UNK(CBM?)-UNK(CBM?)-Doc	5' gctac**ggtacc**t GCT ACT ATC GTT GGT GAC 3' cgccag***ggatcc***TTA TTA CTT AGT TGT TGG GAG AG
ZP_06142991	1b	SIGN-Coh-Doc (ScaE-like)	5' gctac**ggtacc**t GTC GGC GAC TAC AAT GCA 3' cgccag***ggatcc*** TTA ATC TTC GGG GAG CGA AGG
ZP_06145705	1b	SIGN-GH43-UNK-CBM13-CBM13-Doc	5' gctac**ggtacc**t GGA CTT GCA GGC GAT ACC 3' cgccag***ggatcc***TTA TCA GCT TGT CAG CTT GTC
CAK18894	1b	SIGN-Coh-Doc (ScaC)	5' gctac**ggtacc**t CCC GAT CAG GCT ACT CTG 3' cgccag***ggatcc*** TTA TCA AAG TTC TGT GAT GAG AG
ZP_06142105	1c	SIGN-UNK-LamGL(CBM?)-Doc	5' gctac**ggtacc**t GCC GGT ATT CTC TGG GGC 3' cgccag***ggatcc*** TTA TTA TTT GCT ATA GGA TTC GGG
ZP_06145497	1d	SIGN-Coh-Coh-Doc (ScaJ)	5' gctac**ggtacc**t ACT GCT GCT GAG CCT GTA 3' cgccag***ggatcc*** TTA ATG TCA TTA TTC AAG CTT CAG
ZP_06141916	3	SIGN-GH43-X19-CBM22-Doc-CE1	5' gctac**ggtacc**t TCC GGT GAC GTT CAG TAT ATC 3' cgccag***ggatcc*** TTA GGC AGG CTG ACT TTC TCC
ZP_06144896	3	SIGN-GH11-UNK-Doc	5' gctac**ggtacc**t TAT GAG ATC ATG GGT GAC 3' cgccag***ggatcc*** TTA CTT TTG GGA AGC CTT GTC
ZP_06142181	4a	SIGN-Peptidase-UNK-Doc	5' gctac**ggtacc**t CTC ACA CTG CTT CTG AAA CGT 3' cgccag***ggatcc*** TTA CTA ATT TAT TAC AGA TGA TTT AGC
ZP_06142361	4a	SIGN-Coh-Doc (ScaH)	5' gctac**ggtacc**t AAA CCG CAG TAC CGC CTC 3' cgccag***ggatcc*** TTA TCA ACC TCT GAG AGG CTG
CAK18896	4a	SIGN-Coh-Coh-Coh-Coh-Coh-Coh-Coh-Coh-UNK-Coh-UNK-Doc (ScaB)	5' aatt**ggtacc**aACTACAGCAACAAATTCCGGTG 3' taat***ggatcc***TTAACCGAATCTGTTTGGAAC
CAK18897	4a	SIGN-CBM-CBM-Doc (CttA)	5' aatt**ggtacc**aAACACTGTTACATCAGCTG 3' ttaa***ggatcc***TTATTCTTCTTCAGCATCGCC
ZP_06144588	4a	SIGN-Coh-Doc (ScaF)	5' gctac**ggtacc**t GAT GAA ACT ACT GAG TAT AAG 3' cgccag***ggatcc*** TTA TGG AGA ATT ATG AGC CTG
ZP_06145744	4a	SIGN-LRR-Coh-Doc (ScaI)	5' gctac**ggtacc**t GCG GTT ATT ATC GGC GAT 3' cgccag***ggatcc*** TTA TCT GCT TGC GTT TAT AAA TTC
CAK18895	5	SIGN-UNK-Coh-Coh-Doc (ScaA)	5' gctac**ggtacc**t CCA AGC GGC AAC ACA CTC 3' cgccag***ggatcc***TTA TTA GCC CTT AGC AGG GAG
ZP_06143476	6a	SIGN-UNK(LbetaH-LamGL)-Doc	5' gctac**ggtacc**t GAA GCA GAC AGT TTC ATT ATG 3' cgccag***ggatcc***TTA TTA TTG TTT CAG AAG TTC ACG
ZP_06142906	6b	SIGN-Doc-SERPIN	5' gctac**ggtacc**t GCT CTC GAA CCG CCA AGG 3' cgccag***ggatcc***TTA AGG ATG AGC GCT TTC AAT GCC
ZP_06143078	6b	SIGN-GH5-CBM32-CBM32-Doc	5' gctac**ggtacc**t GGA CAG AAA TCA GCT GAG 3' cgccag***ggatcc***TTA TTA TTT GTT GAG TAT TTT TCT GAG

Table S2. 2 Dockerin modules of *R. flavefaciens* strain FD-1 selected for the *in vivo* study.

Accession No.	Group	Architecture of parent protein	Primers used
ZP_06141990	1a	UNK-Doc	5' ggggacaagtttgtacaaaaaagcaggcttc TCA GAA TAT TCC GCA CCT GTC 3' ggggaccactttgtacaagaaagctgggtc TTA TAA GCC GAG CAG TTT CAT CTG
ZP_06142678	1a	SIGN-GH9-UNK-CBM3_1-LNK-Doc	5' ggggacaagtttgtacaaaaaagcaggcttc GTT ACT CTG TGG GGA GAC GCT AAC 3' ggggaccactttgtacaagaaagctgggtc TCA GTT ATA GCT CTC GGG AAG CTC
ZP_06143384	1a	SIGN-UNK-GH44-UNK-LNK-Doc	5' ggggacaagtttgtacaaaaaagcaggcttc CCC GCA AAC GTA ACA TAC GGC 3' ggggaccactttgtacaagaaagctgggtc TTA TGC TTC GGG AAG CTT GTC
ZP_06143935	1a	SIGN-UNK-X159-X159-UNK-Doc	5' ggggacaagtttgtacaaaaaagcaggcttc CCG AAA CCG GAT CTT ACC GGT GAC 3' ggggaccactttgtacaagaaagctgggtc TTA TTT CTT CTC GGG TAA TTC GG
ZP_06144449	1a	SIGN-X70-CE12-CBM13-LNK-Doc-LNK-CBM35-CE12	5' ggggacaagtttgtacaaaaaagcaggcttc GAG GCT GTT CAG AAG TTC CCG GG 3' ggggaccactttgtacaagaaagctgggtc TCA AGC GGG CTC TAC CGG CTG TTT AG
ZP_06145345	1a	SIGN-UNK-Doc	5' ggggacaagtttgtacaaaaaagcaggcttc AAA GTT TCA GAA GTA AAG GGT GAC 3' ggggaccactttgtacaagaaagctgggtc TTA TAC GAG CTT GAG GAG GAT C
ZP_06145412	1a	SIGN-UNK-X159-X159-UNK-X159-X159-X159-X159-X159-X159-UNK-UNK-X159-X159-X159-UNK-X159-X159-X159-UNK-X159-X159-X159-UNK-Doc	5' ggggacaagtttgtacaaaaaagcaggcttc TAC GGC GAC GCT AAC CTT GAC 3' ggggaccactttgtacaagaaagctgggtc TTA TTC TTT ATC GGG AAG TGT GG
ZP_06141671	1a	SIGN-CBM4-X229-GH9-LNK-Doc	5' ggggacaagtttgtacaaaaaagcaggcttc AAT GTT ACT CTC TGG GGC GAC 3' ggggaccactttgtacaagaaagctgggtc TCA CTC TGG AAG ATT TCC GAT AAG
ZP_06142866	1b	SIGN-UNK.GH9-UNK-LNK-Doc	5' ggggacaagtttgtacaaaaaagcaggcttc ATC GTT GGT GAC GCT AAC TGC 3' ggggaccactttgtacaagaaagctgggtc TTA CTT AGT TGT TGG GAG AGT TG
ZP_06142991	1b	SIGN-Coh-Doc (ScaG)	5' ggggacaagtttgtacaaaaaagcaggcttc GTC GGC GAC TAC AAT GCA GGC 3' ggggaccactttgtacaagaaagctgggtc TTA ATC TTC GGG GAG CGA AGG
ZP_06144353	1b	SIGN-UNK-Doc	5' ggggacaagtttgtacaaaaaagcaggcttc GAT CAG GCT ACT CTG AGA GGC 3' ggggaccactttgtacaagaaagctgggtc TCA AAG TTC TGT GAT GTC
ZP_06144572	1b	SIGN-Coh-UNK-Doc (ScaC)	5' ggggacaagtttgtacaaaaaagcaggcttc GTT TCA GAA AAT GTA AAT GGC 3' ggggaccactttgtacaagaaagctgggtc TCA CTG CTC AAT ATC ATC TTT TAT ACC
ZP_06145705	1b	SIGN-UNK-GH43-UNK-CBM13-CBM13-LNK-Doc	5' ggggacaagtttgtacaaaaaagcaggcttc GAA GAA CAG GGA CTT GCA GG 3' ggggaccactttgtacaagaaagctgggtc TCA GCT TGT CAG CTT GTC AAC
ZP_06143931	1c	SIGN-UNK-Doc	5' ggggacaagtttgtacaaaaaagcaggcttc ATT ATA AAC GGC ATT GAA GGC 3' ggggaccactttgtacaagaaagctgggtc TCA GTC AAG CTT CAG CAG
ZP_06142374	1d	SIGN-UNK-Doc	5' ggggacaagtttgtacaaaaaagcaggcttc GCA TTG AAA ACT AAT AGT ATC 3' ggggaccactttgtacaagaaagctgggtc TTA TTC AAG CTT CAG CAG
ZP_06144548	1d	SIGN-UNK-Doc-UNK	5' ggggacaagtttgtacaaaaaagcaggcttc ACT GAC AGT GTA TTA TAC GGT GAC 3' ggggaccactttgtacaagaaagctgggtc TCA TAT ATC AGC AGC ATC ATT CAG
ZP_06145497	1d	SIGN-UNK-Doc (ScaJ)	5' ggggacaagtttgtacaaaaaagcaggcttc GCT GCT GAG CCT GTA AAT GGC 3' ggggaccactttgtacaagaaagctgggtc GTC TTA TTC AAG CTT CAG CAG
ZP_06143271	2	SIGN-UNK-LNK-Doc-LNK-UNK	5' ggggacaagtttgtacaaaaaagcaggcttc GGC GAT ATC AAC GGC GAT GGT ATC 3' ggggaccactttgtacaagaaagctgggtc TCA TGT TGT GGT ATC TTC AGC
ZP_06141956	3	SIGN-Doc-UNK	5' ggggacaagtttgtacaaaaaagcaggcttc GAT ATC CTC ACA CTT TTC GGC 3' ggggaccactttgtacaagaaagctgggtc TCA AAG GGT TCC GCC GAC GGG
ZP_06142964	3	X231-UNK-Doc	5' Ggggacaagtttgtacaaaaaagcaggcttc GGC GAT ATA AAC CTT GAC GGC 3' ggggaccactttgtacaagaaagctgggtc TTA TCC TAT AAG CAT TTT GCG
ZP_06143424	3	SIGN-X141-CBM6-Doc1	5' ggggacaagtttgtacaaaaaagcaggcttc GTA TAC GGC GAC CTT GAC GGT GAC 3' ggggaccactttgtacaagaaagctgggtc GTC TTA TTC AAC CGG GAG AGT TTT GCG
ZP_06145446	3	SIGN-CBM22-GH10-CBM22-Doc	5' ggggacaagtttgtacaaaaaagcaggcttc CAG GAA ATG ATC CTG GGT GAC ATC 3' ggggaccactttgtacaagaaagctgggtc TTA ATT TGC AGG AAA TTC TCT TAT C
ZP_06144588	4a	UNK-Coh-UNK-Doc (ScaF)	5' ggggacaagtttgtacaaaaaagcaggcttc TTC ACT GAG TAT AAG CTT GGC 3' ggggaccactttgtacaagaaagctgggtc TTA TGG AGA ATT ATG AGC CTG
ZP_06142016	4a	SIGN-UNK-Doc	5' ggggacaagtttgtacaaaaaagcaggcttc AAC GAG ATG AAC GCC GCA GGA GAC 3' ggggaccactttgtacaagaaagctgggtc TCA CAC AGA GCT CTG AGC ATA ATG
ZP_06142361	4a	SIGN-Coh-LNK-Doc (ScaH)	5' ggggacaagtttgtacaaaaaagcaggcttc AAA CCG CAG TAC CGC CTC GGA G 3' ggggaccactttgtacaagaaagctgggtc TCA ACC TCT GAG AGG CTG ATG
ZP_06143379	4a	SIGN-Doc-UNK-GH3	5' ggggacaagtttgtacaaaaaagcaggcttc GAG GGA AAT ACC CTC GGC GAC 3' ggggaccactttgtacaagaaagctgggtc TCA GAA GGA ATC AGT CAG CCC
ZP_06143695	4a	UNK-LNK-Doc	5' ggggacaagtttgtacaaaaaagcaggcttc GTA AAC ATC AGT TAT ACA TTA GG 3' ggggaccactttgtacaagaaagctgggtc TTA AAC ATT TTT GAG TGA ATC
ZP_06144357	4a	SIGN-Doc-UNK	5' ggggacaagtttgtacaaaaaagcaggcttc GAA ACT GAT ATC ATG CAC GGT G 3' ggggaccactttgtacaagaaagctgggtc TCA TAT AAC AGT GTC ATT TAC
ZP_06145744	4a	UNK-Coh-Doc (ScaI)	5' ggggacaagtttgtacaaaaaagcaggcttc GCG GTT ATT ATC GGC GAT GTC 3' ggggaccactttgtacaagaaagctgggtc TTA TCT GCT TGC GTT TAT AAA TTC
ZP_06145754	4a	SIGN-Doc-UNK	5' ggggacaagtttgtacaaaaaagcaggcttc GCC GGC GGC CAG ACT CAC GGC 3' ggggaccactttgtacaagaaagctgggtc TCA AAG GGA TTC AGT GTA GCC
ZP_06142815	4b	SIGN-UNK-X142-UNK-X142-UNK-Doc	5' ggggacaagtttgtacaaaaaagcaggcttc GCC TGC GAG GAC AAA ATG GGG 3' ggggaccactttgtacaagaaagctgggtc TTA TTT TCC CTC TGA TGC TGA TGC
ZP_06144573	5	SIGN-X148-LNK-Coh-LNK-Coh-LNK-Doc (ScaA)	5' ggggacaagtttgtacaaaaaagcaggcttc CCT GCA GAA ACA ACA ACT ACA G 3' ggggaccactttgtacaagaaagctgggtc TTA GCC CTT AGC AGG GAG TGT GAT G

ZP_06142459	6a	SIGN-X128-LNK-Doc-UNK	5' ggggacaagtttgtacaaaaaagcaggcttc GAT GAA ACT TTC ATC ATG GGT GAC 3' ggggaccactttgtacaagaaagctgggtc TCA GTT ATC TGA CAA CAG CAA ACG
ZP_06143476	6a	SIGN-X134-UNK-Doc	5' ggggacaagtttgtacaaaaaagcaggcttc GCA GAC AGT TTC ATT ATG GGT GAC 3' ggggaccactttgtacaagaaagctgggtc TTA TTG TTT CAG AAG TTC ACG
ZP_06144432	6a	SIGN-UNK-Doc	5' ggggacaagtttgtacaaaaaagcaggcttc ATA GAT GAT ACA GCT GAC AG 3' ggggaccactttgtacaagaaagctgggtc TTA CTG TTT CAG ATA TTC ACG
ZP_06145118	6a	SIGN-UNK-GH18-Doc	5' ggggacaagtttgtacaaaaaagcaggcttc AAG ACT TTC ATT GCA GGC GAT G 3' ggggaccactttgtacaagaaagctgggtc TCA TAG CAT TTC CTT TAT AAG
ZP_06142225	6b	SIGN-UNK-PL1-UNK-Doc	5' ggggacaagtttgtacaaaaaagcaggcttc AAC CCG GAT GTT GAG CCT GTT CCG 3' ggggaccactttgtacaagaaagctgggtc TTA TTT GCT GAG AGT ATC GAT TAT G
ZP_06142906	6b	SIGN-Doc-UNK	5' ggggacaagtttgtacaaaaaagcaggcttc TCT GCT CTC GAA CCG CCA AGG 3' ggggaccactttgtacaagaaagctgggtc TCA ATT TGC GTC AGC AAT GCC
ZP_06143324	6b	SIGN-Doc-UNK	5' ggggacaagtttgtacaaaaaagcaggcttc GAC GCC CCT GCT ATG ACG GGC 3' ggggaccactttgtacaagaaagctgggtc TCA ACC ATG AGG GAG CCT GCC
ZP_06144185	6b	SIGN-UNK-Doc	5' ggggacaagtttgtacaaaaaagcaggcttc TGT GAC TGT CAG ATA GGT GAC 3' ggggaccactttgtacaagaaagctgggtc TCA TAA AGC AGT TTG GCC TTC
ZP_06142338	-	SIGN-UNK-CBM13-Doc-GH43-UNK-GH43-UNK	5' ggggacaagtttgtacaaaaaagcaggcttc GGA GGC GAG GGT CAG AAA TTC 3' ggggaccactttgtacaagaaagctgggtc TCA GCT GTA GTC CTC GCT GTC
ZP_06142740	-	UNK-Doc-UNK	5' ggggacaagtttgtacaaaaaagcaggcttc AAA TAC ACT CCC TCG AAT GTA G 3' ggggaccactttgtacaagaaagctgggtc TCA GTC GTA AGC TCC TGT TGA TGC
ZP_06142981	-	SIGN-UNK-X159-X159-X159-X159-X159-X159-X159-X159-UNK-X159-X159-X159-UNK-LNK-Doc-UNK-Doc-UNK	5' ggggacaagtttgtacaaaaaagcaggcttc ACT TCA AAG GAT ACA CTT TAC GGC 3' ggggaccactttgtacaagaaagctgggtc TCA TAC GGG GAT AGC AGC CTC GCC
ZP_06144059	-	SIGN-UNK-Doc	5' ggggacaagtttgtacaaaaaagcaggcttc GTC CCC AAA TCA TCA GGC G 3' ggggaccactttgtacaagaaagctgggtc TTA GCC GCC GGA GAG CAG
ZP_06145331	-	Sign-GH11-LNK-CBM22-LNK-GH10-LNK-Doc-LNK-GH11-LNK-CE4	5' ggggacaagtttgtacaaaaaagcaggcttc CAG GTT TCT ACA TGG GGC GAT G 3' ggggaccactttgtacaagaaagctgggtc TCA CCA CTG ATC ATA TGG CT G

Table S2. 3 Cohesin modules of *R. flavefaciens* strain FD-1 selected for the *in vivo* study.

Accession No.	Scaffoldin	Primers used
ZP_06144573	ScaA	5' cacaccatgggagctagc CAG CCT GTT GCT AAT GCA GAC 3' cacactcgagtta TGG ATC ATC AAC AGG GTT ACC
ZP_06144574	ScaB	5' cacaccatgggagctagc CCT GTA GCT AAC GCT GAT G 3' cacactcgagtta GCC CTC CTC ATT AGG AGT ACC
ZP_06144574	ScaB	5' cacaccatgggagctagc aag aat gta aca cct gct aca g 3' cacactcgagtta AAC TAC AGG TGT ATC ACC AAC
ZP_06144574	ScaB	5' cacaccatgggagctagc GCT AAG GGT TCA GTA AAA TGG 3' cacactcgagtta TGA ATC AGG AGT CTT AAC
ZP_06144572	ScaC	5' cacaccatgggagctagc GCA GGC GAA ACA GTG 3' cacactcgagtta TAC TTC TGC TGA AGG AAC
ZP_06144576	ScaE	5' cacaccatgggagctagc CTC ACA GAC AGA GGA ATG 3' cacactcgagtta CTC AGG CTC ACC AGC CTT GAT TG
ZP_06142991	ScaG	5' cacaccatgggagctagc GCT GAC GGC GGT TTC ACA GAC 3' cacactcgagtta GAT ATA GCC GTC CTT CAT GCC
ZP_06142361	ScaH	5' cacaccatgggagctagc GCC TGC CCA GAT CGT GGA AAC 3' cacactcgagtta TTC GGA AGG AGC GGT TAT CTC
ZP_06144588	ScaF	5' cacaccatgggagctagc aat tca aca gat ctc acc 3' cacactcgagtta TTT TTT CTC GCC GAG TAT CCT G
ZP_06145744	ScaI	5' cacaccatgggagctagc AAG CCT GTG CTG CGC ATC 3' cacactcgagtta CGA GAA AAT GTG CTT GTT CAT TG

Table S2. 4 Set of primers used to generate G10A/R11A and G48A/R49A mutations in the XynDoc constructs of peptidase-Doc (ZP_06142181) and ScaH-Doc (ZP_06142361) for testing the dual-binding mode in these type III dockerins.

		ZP_06142181	**ZP_06142361**
KpnI site	F[1]	GCTAC**GGTACC**TCTCACACTGCTTCTGAAACGT	GCTAC**GGTACC**TAAACCGCAGTACCGCCTC
BamHI site	R[4]	CGCCAG**GGATCC**TTACTAATTTATTACAGATGATTTAGC	CGCCAG**GGATCC**TTATCAACCTCTGAGAGGCTG
G10A R11A	F[2]	GAACGGAATAGTAGACG**CCGC**AGATGCTACACTGGTGC	CAACGGAATTATTGAC**GCAGCT**GATGCGACCGCAGTCC
	R[3]	GCACCAGTGTAGCATCT**GCGG**CGTCTACTATTCCGTTC	GGACTGCGGTCGCATC**AGCTGC**GTCAATAATTCCGTTG
G48A R49A	F[2]	CAGCAATGACATCATCGACG**CAGC**AGATGCTACAGAAATACTTAC	GGATAACATGATAGAC**GCAGCT**GACGCTACACATATCC
	R[3]	GTAAGTATTTCTGTAGCATCT**GCTG**CGTCGATGATGTCATTGCTG	GGATATGTGTAGCGTC**AGCTGC**GTCTATCATGTTATCC

Table S2. 5 List of non-interacting dockerin modules, tested by the various strategies in this work.

	Accession No.	Group	Architecture of parental enzyme	CM	E	Iv
1	ZP_06144474	1a	UNK-Doc	X		
2	ZP_06142991	1b	Coh-Doc (ScaE-like)	X		
3	ZP_06144783	1c	UNK-Doc-UNK	X		
4	ZP_06143931	1c	UNK-Doc			X
5	ZP_06143761	1d	UNK-Doc	X		
6	ZP_06141956	3	Doc-UNK			X
7	ZP_06143670	3	Coh-Doc (ScaL)	X		
8	ZP_06143567	4a	LRR-Doc-UNK	X		
9	ZP_06143379	4a	Doc-UNK-GH3			*X*
10	ZP_06142016	4a	UNK-Doc			X
11	ZP_06144357	4a	Doc-UNK			X
12	ZP_06145754	4a	Doc-UNK			X
13	ZP_06142815	4b	UNK-Doc		X	X
14	ZP_06142816	4b	LRR-Doc	X		
15	ZP_06143103	6a	GH43-Doc		X	
16	ZP_06142225	6b	UNK-PL-UNK-Doc			X
17	ZP_06143324	6b	Doc-UNK			*X*
18	ZP_06142338	Unclassified	UNK-CBM13-Doc-GH43-UNK-GH43-UNK			X
19	ZP_06142740	Unclassified	UKN-Doc-UNK			*X*
20	ZP_06142981	Unclassified	UKN-Doc-UNK-Doc-UNK			*X*
21	ZP_06144059	Unclassified	UKN-Doc			X
22	ZP_06145331	Unclassified	GH11-CBM22-GH10-Doc-GH11-CE4			*X*

The last 3 lanes correspond to the three methods used: Cellulose-coated microarray (CM), ELISA (E), and recombinant *in-vivo* co-expression (Iv). Dockerins tested using the designated method were marked (X). Italicized *X* indicates that the given dockerin was insoluble.

Table S3. 1 Set of primers used for DNA isolation of Coh ScaC and Doc 3, in the overlapping PCR to remove the *β*-flap insert of the CBMCoh construct of CohScaC and to generate the mutations in the XynDoc constructs of Doc 3 and ORF 1435.

ID	Vector	Primers used
CohScaC	pET28a	5′ CACACAGGATCCGCAGGCGAAACAGTGCAG 3‘ CACACACTCGAGTACTTCTGCTGAAGGAAC
Doc 3	pET9d	5′ CACACAGGTACCTGTATACGGCGACCTTGAC 3‘ CACACAGGATCCTTAATATTCAACCGGGAG
Doc 3 F902A	pET9d	5′ GGCGAGGTTGACGTAGCCGATCTCATCCTCATG 3‘ CATGAGGATGAGATCGGCTACGTCAACCTCGCC
Doc 3 R908A	pET9d	5′ GATCTCATCCTCATGGCAAAAGCTGTAGAAAAC 3‘ GTTTTCTACAGCTTTTGCCATGAGGATGAGATC
Doc 3 H943A	pET9d	5′ CACAGCGAGTATCTCGCCGGCATACGCAAAACTC 3‘ GAGTTTTGCGTATGCCGGCGAGATACTCGCTGTG
CohScaC No Flap 1	pET28a	5′ CACACATATGGCAGGCGAAACAGTGCAG 3‘ AGAATAACCTGCACTGATTGCTTCAAGCTTGATAGG
CohScaC No Flap 2	pET28a	5′ CCTATCAAGCTTGAAGCAATCAGTGCAGGTTATTCT 3‘ CACACTCGAGTACTTCTGCTGAAGGAAC
ORF1425 F46A	pETG20 A	5′ GCAGCGCCTGATGTT GCC GACATGATCGCTCTC 3′ GAGAGCGATCATGTC GGC AACATCAGGCGCTGC
ORF1425 R52A	pETG20 A	5′ GACATGATCGCTCTC GCC AAAATGCTTATAGGA 3′ TCCTATAAGCATTTT GGC GAGAGCGATCATGTC

Table S3. 2 Primers used to isolate group 3 and 6 dockerins.

ID	Vector	Primers used
ORF341	pETG20 A	5'**GGGGACAAGTTTGTACAAAAAAGCAGGCTTC**GGTGACGTTCAGTATATC 3'**GGGGACCACTTTGTACAAGAAAGCTGGGTCTCA**ACTTTCTCCGTCAAAATC
ORF381	pETG20 A	5' **GGGGACAAGTTTGTACAAAAAAGCAGGCTTC**GATATCCTCACACTTTTC 3' **GGGGACCACTTTGTACAAGAAAGCTGGGTCTCA**AAGGGTTCCGCCGACGGG
ORF408	pETG20 A	5' **GGGGACAAGTTTGTACAAAAAAGCAGGCTTC**GACTCGCTTATCAGAGGC 3' **GGGGACCACTTTGTACAAGAAAGCTGGGTCTCA**AGAAGCGTTTAAAAAGTC
ORF464	pETG20 A	5' **GGGGACAAGTTTGTACAAAAAAGCAGGCTTC**ATATTCTACCAGAAGGGC 3' **GGGGACCACTTTGTACAAGAAAGCTGGGTCTCA**TCTGCCGAGAAGGTAATTC
ORF691	pETG20 A	5' **GGGGACAAGTTTGTACAAAAAAGCAGGCTTC**CCCCAGCCTGTTGCAGGC 3' **GGGGACCACTTTGTACAAGAAAGCTGGGTCTCA**TTCAGCAGGGAAGCTGG
ORF764	pETG20 A	5' **GGGGACAAGTTTGTACAAAAAAGCAGGCTTC**AGACAGTTCATCAGAGGC 3' **GGGGACCACTTTGTACAAGAAAGCTGGGTCTCA**TGGAGGCTCAGGTGAATTATC
ORF1425	pETG20 A	5' **GGGGACAAGTTTGTACAAAAAAGCAGGCTTC**ACTATCGGAAGAAAAGGC 3' **GGGGACCACTTTGTACAAGAAAGCTGGGTCTCA**TCCTATAAGCATTTTGCG
ORF1739	pETG20 A	5' **GGGGACAAGTTTGTACAAAAAAGCAGGCTTC**GATGTCATCCTCGGTGAC 3' **GGGGACCACTTTGTACAAGAAAGCTGGGTCTCA**TTTTAAATGTCTGCCAAG
ORF1934	pETG20 A	5' **GGGGACAAGTTTGTACAAAAAAGCAGGCTTC**GTTTCGGAGGTAAAGGG 3' **GGGGACCACTTTGTACAAGAAAGCTGGGTCTCA**GGTACATTCTGCATCG
ORF2174	pETG20 A	5' **GGGGACAAGTTTGTACAAAAAAGCAGGCTTC**ATTTATTCTATAGGTG 3' **GGGGACCACTTTGTACAAGAAAGCTGGGTCTCA**TGGTTCAATAAGTGAAC
ORF2390	pETG20 A	5' **GGGGACAAGTTTGTACAAAAAAGCAGGCTTC**ACAGGCAAGATCAAGGGC 3' **GGGGACCACTTTGTACAAGAAAGCTGGGTCTCA**GTAGTTCACAGGATCAGC
ORF3451	pETG20 A	5' **GGGGACAAGTTTGTACAAAAAAGCAGGCTTC**AGCGATTTCATATACGG 3' **GGGGACCACTTTGTACAAGAAAGCTGGGTCTCA**TCTCTTTTCAGTCACTGC
ORF3454	pETG20 A	5' **GGGGACAAGTTTGTACAAAAAAGCAGGCTTC**CCATATGAGATCATGGG 3' **GGGGACCACTTTGTACAAGAAAGCTGGGTCTCA**CTTTTGGGAAGCCTTGTC
ORF3729	pETG20 A	5' **GGGGACAAGTTTGTACAAAAAAGCAGGCTTC**ACCGTTCTTCATAAAGGC 3' **GGGGACCACTTTGTACAAGAAAGCTGGGTCTCA**CCCCTCAACAATGACAGG
ORF3865	pETG20 A	5' **GGGGACAAGTTTGTACAAAAAAGCAGGCTTC**AAGACCCTCATCCGTGGC 3' **GGGGACCACTTTGTACAAGAAAGCTGGGTCTCA**TACAGGAACCGCAGGC
ORF4012	pETG20 A	5' **GGGGACAAGTTTGTACAAAAAAGCAGGCTTC**CAGGAAATGATCCTGGG 3' **GGGGACCACTTTGTACAAGAAAGCTGGGTCTCA**ATTTGCAGGAAATTCTC
ORF4092	pETG20 A	5' **GGGGACAAGTTTGTACAAAAAAGCAGGCTTC**ACGGACGTAATTCTAGGC 3' **GGGGACCACTTTGTACAAGAAAGCTGGGTCTCA**TGAGCCTGTAAACGAGTC
ORF4112	pETG20 A	5' **GGGGACAAGTTTGTACAAAAAAGCAGGCTTC**GTTTACGCTTTAGGTG 3' **GGGGACCACTTTGTACAAGAAAGCTGGGTCTCA**ATTGCTGTGTGTAGTTATG
ORF4165	pETG20 A	5' **GGGGACAAGTTTGTACAAAAAAGCAGGCTTC**GGCATCCTCGGTGATGTC 3' **GGGGACCACTTTGTACAAGAAAGCTGGGTCTC**AAACAGGTGTCGGCCAC
ORF903	pETG20 A	5' **GGGACAAGTTTGTACAAAAAAGCAGGCTTC**GATGAAACTTTCATCATGGGTGAC 3' **GGGGACCACTTTGTACAAGAAAGCTGGGTCTC**TCAGTTATCTGACAACAGCAAACG
ORF1369	pETG20 A	5' **GGGGACAAGTTTGTACAAAAAAGCAGGCTTC**TCTGCTCTCGAACCGCCAAGG 3' **GGGGACCACTTTGTACAAGAAAGCTGGGTCTC**TCAATTTGCGTCAGCAATGCC

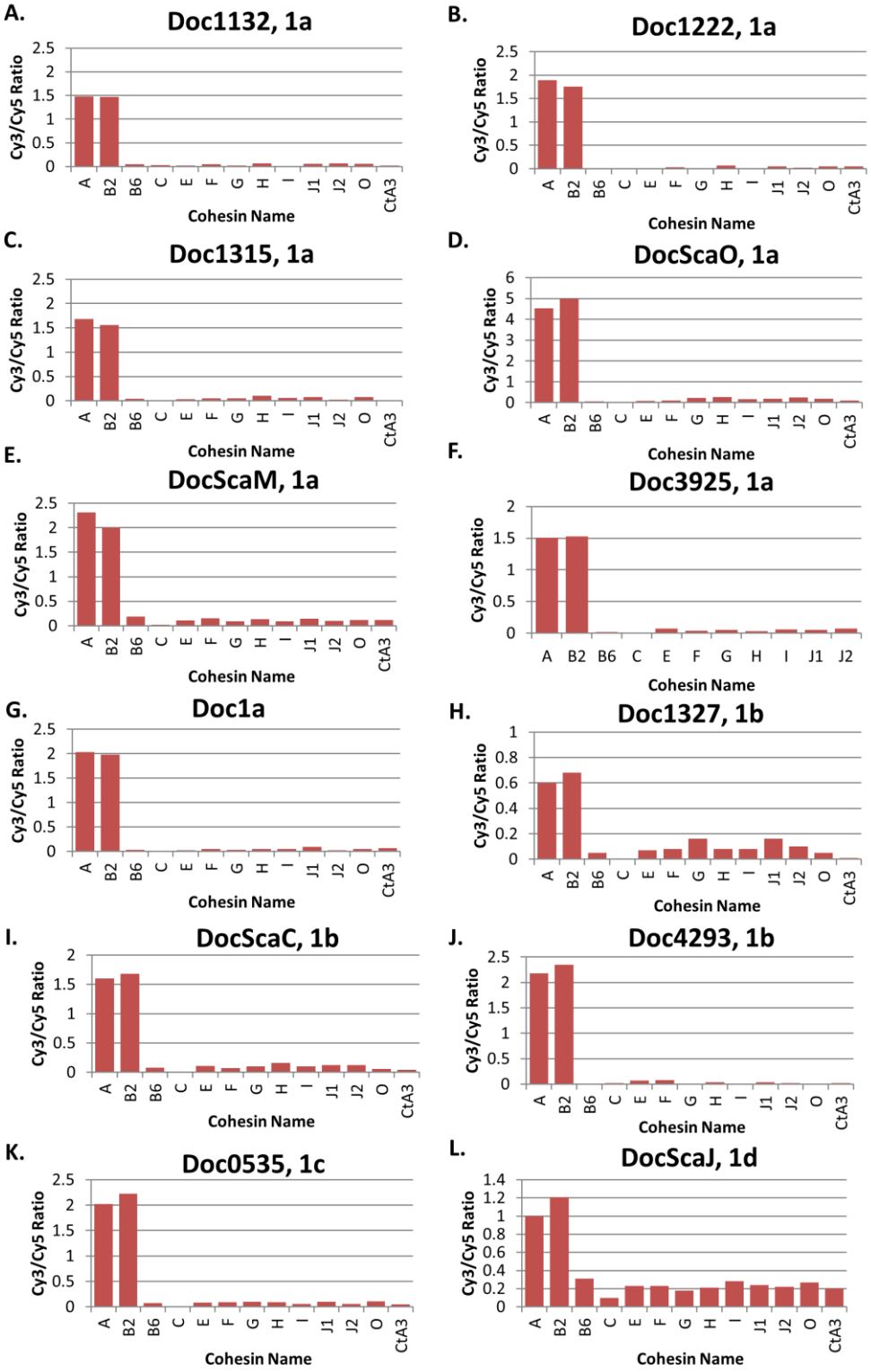
The underlined regions correspond to the recombination regions, designed to be used with the Gateway homologous recombination-mediated cloning protocol.

Table S4. 1 Primers used to isolate genes encoding *R. flavefaciens* dockerins *Rf*Doc1a and *Rf*Doc1b and to generate the Doc1a and CohScaB3 mutant derivatives.

Dockerin	Vector	Primers used
Doc1a	pHTP2	5' *TCAGCAAGGGCTGAGG*GTTCAGAAGTTCCCG 3' *TCAGCGGAAGCTGAGGTTA*TTCAACAGGCGGAAG
Doc1b	pHTP2	5' *TCAGCAAGGGCTGAGG*AATGTTACTCTCTGG 3' *TCAGCGGAAGCTGAGGTTA*CTCTGGAAGATTTCC
CohScaB3 A38Q	pET28a	5' GAACAAGCCAATCTCACAGATGGACGTTAAGTTC 3' GAACTTAACGTCCATCTGTGAGATTGGCTTGTTC
CohScaB3 N68A	pET28a	5' CAACAGTCATGACAGCCATGGCTATCCTTGG 3' CCAAGGATAGCCATGGCTGTCATGACTGTTG
CohScaB3 N75A	pET28a	5' GCTATCCTTGGTGCAGCCTTCAAGTCACTCGAC 3' GTCGAGTGACTTGAAGGCTGCACCAAGGATAGC
CohScaB3 K77A	pET28a	5' CTTGGTGCAAACTTCGCGTCACTCGACGATAAG 3' CTTATCGTCGAGTGACGCGAAGTTTGCACCAAG
CohScaB3 L79A	pET28a	5' GCAAACTTCAAGTCAGCCGACGATAAGGGCGAAC 3' GTTCGCCCTTATCGTCGGCTGACTTGAAGTTTGC
CohScaB3 E84A	pET28a	5' CTCGACGATAAGGGCGCACCGCTCGTTCCTAAG 3' CTTAGGAACGAGCGGTGCGCCCTTATCGTCGAG
CohScaB3 H121A	pET28a	5' GGAAAGAACGAAGTAGCCAAGAGCAACGACGG 3' CCGTCGTTGCTCTTGGCTACTTCGTTCTTTCC
CohScaB3 N124A	pET28a	5' GAAGTACACAAGAGCGCCGACGGTTCACAGTTC 3' GAACTGTGAACCGTCGGCGCTCTTGTGTACTTC
Doc1a I39A	pHTP2	5' GACGGAATAGTTGATGCTTCGGATGCAGTACTC 3' GAGTACTGCATCCGAAGCATCAACTATTCCGTC
Doc1a S40A	pHTP2	5' GGAATAGTTGATATTGCGGATGCAGTACTC 3' GAGTACTGCATCCGCAATATCAACTATTCC
Doc1a V43A	pHTP2	5' GATATTTCGGATGCAGCACTCATTATGCAGAC 3' GTCTGCATAATGAGTGCTGCATCCGAAATATC
Doc1a Q47A	pHTP2	5' GCAGTACTCATTATGGCGACTATGGCTAATCC 3' GGATTAGCCATAGTCGCCATAATGAGTACTGC
Doc1a K54A	pHTP2	5' GGCTAATCCAAGCGCATATCAGATGACCGAC 3' GTCGGTCATCTGATATGCGCTTGGATTAGCC
Doc1a Q83A	pHTP2	5' GATGCACAGTTCATAGCGAGCTATTGTCTGGGA 3' TCCCAGACAATAGCTCGCTATGAACTGTGCATC
Doc1a L87A	pHTP2	5' CATACAGAGCTATTGTGCGGGACTTGTTGAACTTC 3' GAAGTTCAACAAGTCCCGCACAATAGCTCTGTATG

Sequences used for plasmid recombination are in italic.

Figure 4. 1 Coh-binding range of *R. flavefaciens* group 1 dockerins.



Each bar graph represents the recognition profile of one dockerin from a different group 1 subgroup and 12 cohesins. The bar values correspond to the ratio between the measured Cy3 and Cy5 signals. Intensity values were calculated by Array Vision Evaluation 8.0 software and all data processing was made in ExcelIntensity values were calculated by Array Vision Evaluation software and all data processing was made in Excel.

Table S4. 2 Primers used to amplify the cohesins and group1 Docs used in the cellulose microarray assays.

Dockerin	Vector	Primers used
Doc1132_a	pET9d	5’ gctacggtacct GAG CGT GTT ACT CTG TGG 3’ cgccagggatcc TTA TCA GTT ATA GCT CTC GGG
Doc1222_a	pET9d	5’ gctacggtacct GTA ACA CTC TGG GGC GAT GCT 3’ cgccagggatcc TTA TGC GAT ATA TGT CTT ATT TGA TGC
Doc1315_a	pET9d	5’ gctacggtacct ACA CTC TGG GGC GAT GCC 3’ cgccagggatcc TTA CTG ATA ATT TGA TCT TGA GGC
Doc1_a	pET9d	5’ gctacggtacct GAG GCT GTT CAG AAG TTC 3’ cgccagggatcc TTA TTC GGG CTC ATA GTA AAC
DocScaO_a	pET9d	5’ gctac ggtacct TCT GTA ACT TCA ACA GTC AAA G 3’ cgccagggatcc TTAACT CTC CAC AAA CTC CCA GT
Doc3925_a	pET9d	5’ gctacggtacct GTT CTC TGG GGC GAT GCT 3’ cgccagggatcc TTA TGA CTC AGG GAG CTT AGT
DocScaM_a	pET9d	5’ gctac ggtacct TTA GAG ATA GTT CTT GAT GAA CC 3’ cgccagggatcc TTA ATC AAG CTT CAG CAG TTT TTT C
Doc1327_b	pET9d	5’ gctacggtacct GCT ACT ATC GTT GGT GAC 3’ cgccagggatcc TTA TTA CTT AGT TGT TGG GAG AG
Doc4293_b	pET9d	5’ gctacggtacct GGA CTT GCA GGC GAT ACC 3’ cgccagggatcc TTA TCA GCT TGT CAG CTT GTC
DocScaC_b	pET9d	5’ gctacggtacct CCC GAT CAG GCT ACT CTG 3’ cgccagggatcc TTA TCA AAG TTC TGT GAT GAG AG
Doc0535_c	pET9d	5’ gctacggtacct GCC GGT ATT CTC TGG GGC 3’ cgccagggatcc TTA TTA TTT GCT ATA GGA TTC GGG
DocScaJ_d	pET9d	5’ gctacggtacct ACT GCT GCT GAG CCT GTA 3’gccagggatcc TTA ATG TCA TTA TTC AAG CTT CAG
Cohesin	**Vector**	**Primers used**
CohScaA	pET28a	5’ gtccatggatcc CAG ACA AGT GGT ACT CCT TCC 3’ cagcttctcgag TTA AGC TGT TGT AGC AGA TGT TGT TGG ATC
CohScaB2	pET28a	5’ gtccatggatcc CAG ACA AGT GGT ACT CCT TCC 3’ cagcttctcgag TTA TGA GCC TGA ACC TGT TGT AGG
CohScaB6	pET28a	5’ gtccatggatcc ACT GAT ACA AAC GGT AAC AAG 3’ cagcttctcgag TTA TGT AAG AGT GAT CTT ATC AGT
CohScaC	pET28a	5’ gtccatggatcc GCT CCG GCA TTC GCT GCA 3’ cagcttctcgag TTA AGC CTT GGT GGT TGT TAC TTC
CohScaE	pET28a	5’ actaccatgg CGCTCACAGACAGAGGAATG 3’ actactcgag TGGCTCACCAGCCTTGATTGC
CohScaF	pET28a	5’ gtccatggatcc AAT TCA ACA GAT CTC ACC GAA GC 3’ cagcttctcgag TTA GCC AAG CTT ATA CTC AGT AG
CohScaG	pET28a	5’ gtccatggatcc AGC GGC GGA AGC AGT TCG 3’ cagcttctcgag TTATTC AAC TGT TAT AGT GCC GCC
CohScaH	pET28a	5’ gtccatggatcc GCC TGC CCA GAT CGT GGA 3’ cagcttctcgag TTA CGT TTC GGA AGG AGC GGT
CohScaI	pET28a	5’ gtccatggatcc GGC CCC GTA GTT CAG GGA AAG 3’ cagcttctcgag TTA ATC GGC AAC TAT CTC GAT GGC
CohScaJ1	pET28a	5’ gtccatggatcc GCT GAA ACA TCA ACA GCA 3’ cagcttctcgag TTA AGA AGT TTC GGT TGT AAC
CohScaJ2	pET28a	5’ gtccatggatcc TCT ACA AAA ACA AAC ACC CAA ACA 3’ cagcttctcgag TTA AGC AGC AGT AGT TGT TGT TAT TAC
CohScaO	pET28a	5’ gtccatggatcc GCG CCT GTT ACA ATA TCA G 3’ cagcttctcgag TTAAGT AGT ACT TAC CTG AGA A

Table S5. 1 Set of primers used to isolate the *Rf*DocScaA gene and to generate its mutant derivatives.

ID	Vector	Primers used
***Rf*DocScaA**	pETG20A	5′ GGGGACAAGTTTGTACAAAAAAGCAGGCTTC cctgcagaaacaacaactacag 3′ GGGGACCACTTTGTACAAGAAAGCTGGGTC ttagcccttagcagggagtgtgatg
***Rf*DocScaA N661A**	pETG20A	5′ ctgcgacggtgacgta**gcc**gtagctgacgttgttc 3′ gaacaacgtcagctacggctacgtcaccgtcgcag
***Rf*DocScaA V662A**	pETG20A	5′ gacggtgacgtaaac**gca**gctgacgttgttctc 3′ gagaacaacgtcagctgcgtttacgtcaccgtC
***Rf*DocScaA V666A**	pETG20A	5′ gtaaacgtagctgacgtt**gct**ctccttaacaagtgg 3′ ccacttgttaaggagagcaacgtcagctacgtttac
***Rf*DocScaA N669A**	pETG20A	5′ gacgttgttctcctt**gcc**aagtggctcaacaac 3′ gttgttgagccacttggcaaggagaacaacgtc
***Rf*DocScaA K670A**	pETG20A	5′ gttgttctccttaac**gcg**tggctcaacaacaatg 3′ cattgttgttgagccacgcgttaaggagaacaac
***Rf*DocScaA V721A**	pETG20A	5′ ctatcatcaagagcgta**gct**cacctcatcacactc 3′ gagtgtgatgaggtgagctacgctcttgatgatag
***Rf*DocScaA H722A**	pETG20A	5′ catcaagagcgtagtt**gcc**ctcatcacactccctg 3′ cagggagtgtgatgagggcaactacgctcttgatg
***Rf*DocScaA V662A + V666A**	pETG20A	5′ gtgacgtaaac**gca**gctgacgtt**gct**ctccttaacaagtg 3′ cacttgttaaggagagcaacgtcagctgcgtttacgtcac

The fraction in capital letters corresponds to the homologous recombination zone. Mutated codons are shown bold and underlined.

Table S6. 1 Set of primers used to isolate the *Ac*DocCel5 and *Ac*DocScaB genes and to generate their mutant derivatives.

ID	Vector	Primers used
***Ac*DocCel5 WT, M1 and M2**	pet9d	5′ cacacaGGTACCttcggatgtcaaaccgggc 3′ cacacaGGATCCttagacagagaatttggtaatc
***Ac*DocCel5 M1 + M2**	pet9d	5′ gacaaaagcatcaat**atcaac**gacttcgccattatg 3′ cataatggcgaagtc**gttgat**attgatgcttttgtc
***Ac*DocCel5** S15I, I16N, **N23D**	pet9d	5′ ttcgcactgatgcgt**gac**tatctgctgggcaac 3′ gttgcccagcagata**gtc**acgcatcagtgcggC
***Ac*DocCel5 N14R**, S15I, I16N, **F18A**, **A19V**, N23D	pet9d	5′ cggctcaatc**cgt**attaacgac**gccgta**ctgatgcg 3′ cgcatcag**tacggc**gtcgttaat**acg**gattgagccg
***Ac*DocCel5** N14R, S15I, I16N, F18A, A19V, **M21I**, N23D	pet9d	5′ cgacgccgtactg**att**cgtgactatctgctg 3′ cagcagatagtcacg**aat**cagtacggcgtcg
***Ac*CohScaA6**	pet28a	5′ cacacaGGATCCcaaacgggctttaatctg 3′ cacacaCTCGAGgacagcaccgttggtaac