UNIVERSIDADE TÉCNICA DE LISBOA

Faculdade de Medicina Veterinária

Structure and function relationships in novel cellulosomal enzymes and cohesin-dockerin complexes

Joana Luís Armada Brás

TESE DE DOUTORAMENTO EM CIÊNCIAS VETERINÁRIAS

ESPECIALIDADE DE CIÊNCIAS BIOLÓGICAS E BIOMÉDICAS

CONSTITUIÇÃO DO JURI

PRESIDENTE DO JURI:

Doutor Rui Manuel de Vasconcelos e Horta Caldeira

VOGAIS:

Doutor José António Mestre Prates

Doutor Edward A. Bayer

Doutor Carlos Mendes Godinho de Andrade Fontes

Doutor Victor Manuel Diogo de Oliveira Alves

Doutora Ana Luísa Moreira de Carvalho

ORIENTADOR

Doutor Carlos Mendes Godinho de Andrade Fontes

CO-ORIENTADOR

Doutor José António Mestre Prates

2012

LISBOA

À minha Família

Agradecimentos

Uma tese de doutoramento não é apenas o culminar de muitos anos de trabalho. É uma longa viagem com momentos altos e baixos, uma maratona cuja meta é o conhecimento científico. O interesse crescente pelo trabalho científico e pela obtenção de conhecimento são dois dos factores importantes para o sucesso desta etapa. Contudo, só existem se estivermos rodeados das pessoas certas. Na verdade, não podia ter escolhido melhor grupo de investigação para fazer o meu doutoramento. O meu sucesso como investigadora é também resultado do envolvimento de várias pessoas e entidades às quais quero agradecer:

À Fundação para a Ciência e Tecnologia pelo financiamento da minha bolsa de doutoramento;

À Faculdade de Medicina Veterinária e ao CIISA por me terem aceite como estudante de doutoramento e por terem disponibilizado meios físicos e materiais para a realização de todo o trabalho.

Ao Professor Doutor Carlos Fontes, por ser um excelente orientador, por ser um verdadeiro exemplo do que é ser um investigador de alto nível e por isso o líder e o elemento chave do grupo de investigação. Agradeço o seu entusiasmo constante, a sua boa disposição, os seus ensinamentos, as suas críticas sempre construtivas e a sua amizade;

Ao Professor Doutor José Prates por ter sido um co-orientador exemplar, pela sua simpatia e pelos seus ensinamentos na área da bioquímica;

Ao Professor Doutor Luís Ferreira, pela sua simpatia, entusiasmo, sentido crítico e disponibilidade para conversar sobre temas inerentes às áreas de estudo do grupo;

À Helena Santos pela ajuda incondicional sempre que necessário, pela sua paciência e capacidade de organização. Sem ela o laboratório não seria tão funcional e simultaneamente um local tão acolhedor. Acima de tudo, pela sua amizade e companheirismo que são para mim tão importantes;

À Teresa Ribeiro por ser muito mais do que uma colega de laboratório, pela boa disposição, pelas conversas sérias ou divertidas, pela troca de ideias, por todo o apoio incondicional e sobretudo pela amizade;

À Vânia Fernandes e à Ana Sofia Luís pela boa disposição, amizade e por serem um exemplo de dedicação;

À Márcia Correia e à Benedita Pinheiro pela óptima disposição no laboratório, espírito de equipa, amizade e por terem sido mentoras impecáveis sobretudo durante o meu período de adaptação ao laboratório;

À Catarina Guerreiro, à Patrícia Ponte, à Virginia Pires, ao Fernando Dias e à Maria José Centeno por terem sido "seniores" excepcionais, por me terem transmitido boas práticas laboratoriais, por se terem mostrado sempre disponíveis e pela boa disposição;

Ao Pedro Bule, à Kate Cameron, à Immacolata Venditto, à Mónica Costa e ao Luís Serrano por serem uma nova geração de investigadores cheios de boa disposição e potencial; a word in english to Imma and Kate: thanks for your good nature and sense of humour;

Ao Dr. Shabir Najmudin, por toda a ajuda durante os últimos anos, por se encontrar sempre disponível para ensinar, esclarecer dúvidas, discutir resultados e por ser uma óptima companhia para trabalhar até tarde. Em particular, por ser uma pessoa inspiradora com um surpreendente espírito aventureiro;

Ao Professor Doutor Victor Alves, pela excelente colaboração com o nosso grupo e dedicação exemplar à área da cristalografia de proteínas;

À D. Paula pela sua disponibilidade, boa disposição e amizade. É um elemento fundamental para o bom funcionamento do Departamento de Produção Animal e Segurança Alimentar;

To Dr. Harry Gilbert and Dr. Ed Bayer for collaborating with our group and for being such inspiring scientists.

To Dr. David Bolam from the Institute for Cell and Molecular Biosciences, Newcastle University for supervising my work during training in the use of Isothermal Titration Calorimetry; thank you for your hospitality, support and guidance;

Ao grupo liderado pela Professora Doutora Maria João Romão da Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa, particularmente à Doutora Ana Luísa Carvalho pela colaboração e apoio na resolução de estruturas cristalográficas;

À Susana Martins, à Cristina Monteiro e à Inês Viegas pela amizade, boa disposição, sentido de humor e pelos almoços repletos de gargalhadas e de conversas animadas;

A todos os meus amigos que têm acompanhado o meu percurso, pela amizade, pelos bons momentos, por saber que estarão sempre presentes. Em particular, agradeço à Inês Silva e à Ana Rita Duarte por terem dado comigo os primeiros passos no mundo da investigação. Obrigada por todos os bons momentos passados em Göttingen: o tempo passa, mas as recordações e a vossa amizade ficam e conseguem sempre arrancar-me um sorriso;

À minha família mais próxima, ao meu tio Rui e aos meus primos por toda amizade, carinho e momentos em família que nunca são demais;

À Catarina pela amizade e por teres sido uma pessoa fundamental na minha educação; por me teres ensinado tanto de forma tão simples e fácil de compreender;

À tia Maria José, por toda a amizade, pela companhia, pelas conversas animadas, e pelos almoços e lanches na Ajuda que tornaram os meus dias de trabalho ainda mais felizes;

Aos meus pais, César Luís e Maria Margarida o meu agradecimento por tudo: pela educação, pelos valores, pela amizade, pelas oportunidades, pelos conselhos e pelo apoio incondicional; Sem voces não teria chegado aqui; convosco chegarei ainda mais longe;

Às minhas quatro irmãs: Ana, Teresa, Helena e Inês pela nossa forte irmandade, pela amizade, pelas brincadeiras, pelos anos em que crescemos juntas, pela cumplicidade e por me transmitirem sempre a energia necessária para ser feliz;

Ao Pedro, pelo apoio e ajuda incondicional, pelas passagens por Monsanto ao fim-de-semana para eu ir ao laboratório; Pelo amor e amizade, entusiasmo contagiante e sonhos que se tornam realidade a cada dia que passa; Por partilharmos não apenas um projecto de vida, mas também o gosto pela ciência, que tanto nos preenche o coração.

This work was funded by Fundação para a Ciência e a Tecnologia, grant SFRH/BD/38667/2007 from Ministério da Ciência, Tecnologia e Ensino Superior

RESUMO

Estrutura e função de novas enzimas celulossomais e de novos complexos coesina-doquerina.

A parede celular vegetal constitui uma das principais fontes de carbono do planeta, sendo por isso um extraordinário recurso energético para muitos microrganismos. As limitações energéticas características dos ecossistemas anaeróbios conduziram à evolução de complexos multi-enzimáticos de elevada eficiência, denominados celulossomas, os quais coordenam a degradação dos hidratos de carbono da parede celular vegetal. O *Clostridium thermocellum* produz um celulossoma relativamente bem caracterizado já que a bactéria apresenta uma das maiores taxas de crescimento em celulose. A organização do celulossoma é efectuada por uma proteína não-catalítica multi-modular. Esta proteína de integração possui uma série de módulos repetidos, denominadas coesinas do tipo I. Na extremidade C-terminal das enzimas celulosomais existem módulos doquerina do tipo I, os quais se ligam fortemente às coesinas do tipo I. As proteínas de integração celulossomal podem conter ainda uma doquerina do tipo II que reconhece especificamente coesinas do tipo II localizadas no envelope celular permitindo, assim, a fixação dos celulossomas à superfície da bactéria. No presente trabalho foram desenvolvidas várias metodologias inovadoras com o intuito de determinar a estrutura e a função de novas enzimas celulossomais e de novos complexos coesina-doquerina. Os protocolos de biologia molecular e bioquímicos aqui descritos (capítulo 2) permitem ultrapassar as dificuldades inerentes à cristalização e, por conseguinte, à resolução da estrutura de complexos coesina-doquerina. Com base nestas metodologias, foram elucidadas as estruturas tridimensionais de dois novos complexos coesina-doquerina do tipo I (*Ct*CohOlpC-Doc124A e *Ct*CohOlpA-Doc918). A análise destas estruturas revelou que as suas doquerinas são atípicas, uma vez que não possuem a simetria estrutural característica dos módulos doquerina do tipo I que lhes confere um modo de ligação duplo. Com efeito, estas novas doquerinas apresentam um modo de ligação simples e parecem ligar-se preferencialmente a coesinas localizadas na superfície da bactéria (capítulo 3). A Doc124A é a doquerina da *Ct*Cel124A, uma endo-celulase que possui um enrolamento super-helicoidal e que atua em sinergia com a principal exo-celulase celulossomal, a Cel48S, na hidrólise da celulose. A estrutura tridimensional da *Ct*Cel124A em complexo com duas moléculas de celotriose sugere que a enzima pode ter como substrato alvo a interface entre as formas cristalina e amorfa da celulose (capítulo 5). Para além da elucidação destas estruturas, foi também resolvida a estrutura de um novo complexo coesina-doquerina do tipo II (*Ct*CohScaC2-XDocCipB). Os resultados sugerem que as doquerinas do tipo II apresentam duas interfaces de ligação a coesinas que expressam diferentes especificidades (capítulo 4). Por último, é apresentada a estrutura tridimensional de uma proteína celulossomal penta-modular (*Ct*Xyl5A) e é elucidada a sua função (capítulo 6). Esta proteína é um dos componentes celulossomais de maiores dimensões e compreende uma GH5, dois módulos de ligação a hidratos de carbono (CBMs) das famílias 6 e 13, um módulo de fibronectina do tipo III, um CBM da família 62 e ainda uma doquerina do tipo I. A GH5 apresenta um enrolamento canónico em barril $(\alpha/\beta)_8$ e possui especificidade para os arabinoxilanos, tendo sido por isso definida como uma arabinoxilanase. O CBM6 apresenta um enrolamento em β-sanduiche e possui afinidade para os produtos de reação gerados pela GH5 e para xilo-oligossacáridos não-decorados. A estrutura penta-modular desta proteína revelou uma grande flexibilidade no domínio CBM62.

Palavras-chave: Celulossoma, *Clostridium thermocellum,* coesina, doquerina, glicósido hidrolase, módulos de ligação a hidratos de carbono.

ABSTRACT

Structure and function relationships in novel cellulosomal enzymes and cohesin-dockerin complexes.

Plant cell walls are the most abundant source of organic carbon on earth, providing an extraordinary supply of energy for various microorganisms. The energetic constrains posed by anaerobic ecosystems lead to the evolution of highly efficient multi-enzymatic complexes, termed cellulosomes, which orchestrate the deconstruction of structural carbohydrates. *Clostridium thermocellum* cellulosome has been extensively studied as the bacterium exhibits one of the highest growth rates on cellulose. Cellulosomes are assembled by a large non-catalytic multi-modular scaffoldin which contains repeated type I cohesins. Type I dockerin modules, usually located at the C-terminus of enzymes, bind tenaciously to type I cohesins. Scaffoldins may contain a type II dockerin which specifically recognizes type II cohesins located at the cell envelope, allowing the cell surface attachment of cellulosomes. Here a combination of methodologies was applied to study the structure and function relationships of novel cellulosomal enzymes and cohesin-dockerin complexes. Innovative molecular biology and biochemical protocols that can be applied to crystallize and solve the structure of cohesin-dockerin complexes are described in chapter 2. In addition, the crystal structures of two novel type I cohesin-dockerin complexes (CtCohOlpC-Doc124A and *Ct*CohOlpA-Doc918) are described here. They revealed that the two dockerins are unusual since they lack the structural symmetry that supports the dual binding mode typical of type I modules. Thus, these dockerins present a single binding mode and seem to bind preferentially to cohesins located at the bacterium cell surface and not to cellulosomes (chapter 3). Doc124A is the dockerin of *Ct*Cel124A, an endo-acting cellulase with a superhelical fold that acts in synergy with the major cellulosomal exo-cellulase, Cel48S, during cellulose hydrolysis. The crystal structure of *Ct*Cel124A in complex with two cellotriose molecules suggests that the enzyme may target the interface between crystalline and amorphous cellulose (chapter 5). In addition, the structure of a novel type II cohesin-dockerin complex (*Ct*CohScaC2-XDocCipB) was solved. The functional importance of specific dockerin residues was determined. Type II dockerins are suggested to present two different cohesin-binding faces that express different specificities (chapter 4). Finally, the crystal structure of a penta-modular cellulosomal protein (*Ct*Xyl5A), previously of unknown function, was assessed (chapter 6). This protein is one of the largest cellulosomal components and comprises a GH5, two CBMs from families 6 and 13, a fibronectin type III-like module, a CBM from family 62 and a type I dockerin. *Ct*GH5 has a canonical $(\alpha/\beta)_8$-barrel fold and displays specificity for arabinoxylans and as such, is defined as an arabinoxylanase. *Ct*CBM6 adopts a β-sandwich fold and displays affinity for the reaction products generated by *Ct*GH5 and for undecorated xylooligosaccharides. In addition, the penta-modular structure revealed a great flexibility for the *Ct*CBM62 domain.

Key-words: Cellulosome, *Clostridium thermocellum*, cohesin, dockerin, glycoside hydrolase, carbohydrate-binding module.

This thesis was based on the following manuscripts:

Brás, J. L. A.; Carvalho, A. L.; Nadjmudim, S.; Alves, V. D.; Prates, J. A. M.; Ferreira, L. M. A.; Gilbert, H. J. and Fontes C. M. G. A. (2012) *Escherichia coli* Expression, Purification, Crystallization and Structure Determination of Bacterial Cohesin-Dockerin Complexes. *Methods in Enzymology*, Volume 510, Cellulases, Chapter 21, 395-415.

Brás J. L. A.; Alves V. D.; Carvalho A. L.; Najmudin S.; Prates J. A. M.; Ferreira L. M. A.; Bolam D. N.; Gilbert H. J. and Fontes C. M. G. A. (2012) Novel *Clostridium thermocellum* Type I Cohesin-Dockerin Complexes Reveal a Single Binding Mode. *The Journal of Biological Chemistry.* Published on November 1, 2012 as Manuscript M112.407700.

Brás, J. L. A.; Viegas, A.; Pinheiro, B. A.; Ribeiro, T.; Cuskin, F.; Najmudin, S.; Alves,.V. D.; Prates, J. M. A.; Romão, M. J.; Gilbert, H. J.; Carvalho, A. L. and Fontes, C. M. G. A. (2012) Do *Clostridium thermocellum* type II dockerins present two different cohesin-binding faces?. Work in progress.

Brás J. L. A.; Cartmell A.; Carvalho A. L.; Verzé G.; Bayer E. A.; Vazana Y.; Correia M. A.; Prates J. A.; Ratnaparkhe S.; Boraston A. B.; Romão M. J.; Fontes C. M.; Gilbert H. J. (2011) Structural insights into a unique cellulase fold and mechanism of cellulose hydrolysis. *Proceedings of the National Academy of Sciences of the United States of America*, 108(13), 5237-5242.

Correia M. A.; Mazumder K.; Brás J. L. A.; Firbank S. J.; Zhu Y.; Lewis R. J.; York W. S.; Fontes C. M; Gilbert H. J. (2011) Structure and function of an arabinoxylan-specific xylanase. *The Journal of Biological Chemistry* 286(25), 22510-22520.

Brás, J. L. A.; Correia, M. A. S.; Romão, M. J.; Prates, J.; Fontes, C. M. G. A.; Najmudin, S. (2011) Purification, crystallization and preliminary X-ray characterization of the penta-modular arabinoxylanase *Ct*Xyl5A from *Clostridium thermocellum. Acta Crystallogr Sect F Struct Biol Cryst Commun.* 67(Pt 7):833-6.

INDEX

List of Tables XXI

List of Figures XXIII

List of Abbreviations and Symbols XXV

1. Bibliographic review and objectives 1

1.1. Introduction 1

1.2. Plant Cell Wall 2

1.2.1. Plant Cell Wall Components 2

1.2.1.1. Cellulose 3

1.2.1.2. Hemicellulose 4

1.2.1.3. Pectin 5

1.2.2. Plant Cell Wall Models 5

1.2.3. Plant Cell Wall Hydrolysis 7

1.2.4. Carbohydrate-Active Enzymes database (CAZy) 8

1.3. The Cellulosome: Architecture and Function 9

1.3.1. The scaffoldin 11

1.3.2. Cellulosome Cell surface attachment in *C. thermocellum* 12

1.3.3. The Diversity of Cellulosomes 14

1.3.4. Transcriptional Regulation in Cellulolytic Bacteria 18

1.3.5. The Type I Cohesin-Dockerin Interaction 20

1.3.6. The Type II Cohesin-Dockerin Interaction 25

1.3.7. Cohesin-dockerin Specificity 27

1.3.8. Cellulosome Structural Organization 28

1.3.9. Cellulosomal Enzymes 30

1.3.9.1. Glycoside hydrolases 30

1.3.9.2. Polysaccharide lyases 32

1.3.9.3. Carbohydrate esterases 32

1.3.10. Linker regions and Non-catalytic Modules 33

1.3.11. Carbohydrate Binding Modules (CBMs) 33

1.3.11.1. CBMs functions 34

1.3.11.2. CBMs Classification and Nomenclature 35
1.3.11.3. CBMs and Multivalency 37
1.3.11.4. Ligand Binding Specificity 38
1.3.12. Cellulolytic Machinery: affinity and bioenergy applications 39
1.3.12.1. Potential applications of CBMs and cohesin-dockerin complexes 39
1.3.12.2. Bioenergy prodution from lignocellulosic materials 39
1.3.12.3. Other applications 41
1.4. Objectives 43
1.5. Thesis Outline 43
2. Methods for the production, expression and structure determination of bacterial cohesin-dockerin complexes. 45
2.1. *Escherichia coli* Expression, Purification, Crystallization and Structure determination of Bacterial Cohesin-Dockerin Complexes. 45
2.1.1. Introduction 46
2.1.2. Cloning of cohesin and dockerin genes in prokaryotic expression vectors 46
2.1.2.1. Cloning genes encoding dockerin and cohesin modules through PCR 47
2.1.2.2. Producing synthetic dockerin or cohesin genes 47
2.1.2.3. Producing genes constructs for protein co-expression 48
2.1.3. Expression & purification of cohesin-dockerin complexes in *E. coli* 49
2.1.3.1. Protein Expression 49
2.1.3.2. Protein Purification 50
2.1.4. The dual binding mode and the crystallization of cohesin-dockerin complexes 51
2.1.5. X-ray crystallography of cohesin-dockerin complexes 52
2.1.5.1. Crystallization 54
2.1.5.1.1. Screening of crystallization conditions 55
2.1.5.1.2. Crystals Optimization 55
2.1.5.2. X-ray data collection and reduction 56
2.1.5.3. Model building and structure refinement 58
2.1.6. Summary 61
3. Novel type I cohesin dockerin complexes 63

3.1. Novel *Clostridium thermocellum* Type I Cohesin-Dockerin Complexes Reveal a Single Binding Mode 63
3.1.1. Introduction 64
3.1.2. Material and Methods 66
3.1.2.1. Cloning and expression 66
3.1.2.2. Protein Purification 67
3.1.2.2.1. Cohesin-Dockerin Complexes 67
3.1.2.2.2. Unbound Cohesins and Dockerins 67
3.1.2.3. Isothermal Titration Calorimetry 68
3.1.2.4. Crystallization and Data Collection 68
3.1.2.5. Structure Determination and Refinement 69
3.1.3. Results and Discussion 71
3.1.3.1. Expression and Crystallization of novel Coh-Doc complexes 71
3.1.3.2. Structure of type I Coh-Doc complexes 72
3.1.3.2.1. Structure of OlpA and OlpC type I Cohesins 73
3.1.3.2.2. Structure of Type I Dockerins 74
3.1.3.2.3. Novel Type I Coh-Doc Complex Interfaces 75
3.1.3.3. Probing the Importance of Contact Residues in Dockerins 80
3.1.3.3.1. CohOlpA-Doc918 Complex 80
3.1.3.3.2. CohOlpC-Doc124A Complex 82
3.1.4. Conclusions 83
4. The Structure of a novel Type II Cohesin-Dockerin Complex 85
4.1. Do *Clostridium thermocellum* type II dockerins present two different cohesin-binding faces? 85
4.1.1. Introduction 86
4.1.2. Material and Methods 89
4.1.2.1. Cloning, expression and purification 89
4.1.2.2. Site-Directed Mutagenesis 91
4.1.2.3. Analysis of complex formation in solution 91
4.1.2.4. Isothermal titration calorimetry of cohesin–dockerin binding 91
4.1.2.5. Complex Crystallization 91
4.1.2.6. X-Ray Data Collection and Processing 92

4.1.2.7. Structure Determination, Refinement and Model Building 92
4.1.3. Results and Discussion 93
4.1.3.1. Novel type II cohesin and dockerin domains in *C. thermocellum* proteins 93
4.1.3.2. Are the novel *C. thermocellum* type II cohesins and dockerins functional? 97
4.1.3.3. The structure of ScaC2 cohesin in complex with CipB XDockerin (ScaC2-CipBXDoc). 99
4.1.3.4. The ScaC2-CipBXDoc complex interface 103
4.1.3.5. Comparison of *C. thermocellum* type II cohesin-dockerin complexes 103
4.1.3.6. Functional importance of residues at the surface of the type II dockerin for cohesin recognition 105
4.1.4. Conclusions 109
5. Structural insights into a unique cellulase fold and mechanism of cellulose hydrolysis 111
5.1. *Ct*Cel124 a novel cellulase from *Clostridium thermocellum* cellulosome 111
5.1.1. Introduction 112
5.1.2. Material and Methods 113
5.1.2.1. Cloning, expression and purification of *Ct*Cel124$_{CD}$ 113
5.1.2.2. Enzyme assays 114
5.1.2.3. Binding of the *Ct*Cel124$_{CD}$ mutant E96A to ligands 114
5.1.2.4. Crystallization of *Ct*Cel124$_{CD}$ and data collection 115
5.1.2.5. Phasing, Model Building and Refinement 115
5.1.3. Results and Discussion 116
5.1.3.1. Catalytic properties of *Ct*Cel124 116
5.1.3.2. Affinity of *Ct*Cel124 for cellulose and cellooligosaccharides 118
5.1.3.3. Synergy between GH48S and *Ct*Cel124 119
5.1.3.4. Crystal structure of *Ct*Cel124 120
5.1.3.5. Structural similarity of *Ct*Cel124 to other glycoside hydrolases 122
5.1.3.6. Active site and catalytic mechanism of *Ct*Cel124 122
5.1.3.7. The substrate binding cleft of *Ct*Cel124 124
6. Structure of a penta-modular cellulosomal arabinoxylanase 127
6.1. The structure and function of an arabinoxylan-specific xylanase 127

6.1.1. Introduction 128
6.1.2. Experimental Procedures 129
6.1.2.1. Cloning, expression and purification of components of *Ct*Xyl5A 129
6.1.2.2. Mutagenesis 129
6.1.2.3. Enzyme assays 130
6.1.2.4. Oligosaccharide analysis 130
6.1.2.4.1. Preparation of the partially methylated alditol acetates 130
6.1.2.4.2. GC-EIMS analysis 130
6.1.2.4.3. Preparation of per-O-methylated oligoglycosyl alditols 130
6.1.2.4.4. MALDI-TOF mass spectrometry (MALDI-TOF-MS) 131
6.1.2.4.5. ESI-MS 131
6.1.2.4.6. NMR spectroscopy 131
6.1.2.5. Isothermal Titration Calorimetry 131
6.1.2.6. Crystallography 131
6.1.3. Results 132
6.1.3.1. Expression and purification of CtXyl5A 132
6.1.3.2. *Ct*Xyl5A is an arabinoxylanase 132
6.1.3.3. Characterization of the reaction products generated by *Ct*Xyl5A from arabinoxylan 133
6.1.3.4. Binding of *Ct*Xyl5A to arabinoxylan 139
6.1.3.5. *Ct*CBM6 specificity 139
6.1.3.6. Crystal structure of *Ct*GH5-CBM6 141
6.1.3.6.1. *Ct*GH5 141
6.1.3.6.2. *Ct*CBM6 143
6.1.3.6.3. The linker connecting *Ct*GH5 with *Ct*CBM6 144
6.1.4. Discussion 146
6.2. Purification, crystallization and preliminary X-ray characterization of the penta-modular arabinoxylanase *Ct*Xyl5A from *Clostridium thermocellum* 147
6.2.1. Introduction 148
6.2.2. Materials and Methods 149
6.2.2.1. Protein Expression and Purification 149
6.2.2.2. Crystallization 150

6.2.2.3. Data collection and processing 151
6.2.3. Brief description of the penta-modular cellulosomal arabinoxylanase 154
7. General Discussion and Future Perspectives 157
8. Bibliographic References 165
Annexes ..... A

LIST OF TABLES

Table 1.1| Suspected recognition residues of different dockerin domains derived from cellulosomal components of different species. 28
Table 3.1| Primers used to obtain the genes encoding the cohesin and the dockerin derivatives used in the present study. Engineered restriction sites are shown in bold. 66
Table 3.2| Protein sequences of dockerins DocCel124A and Doc918 and respective synthesized mutants. 67
Table 3.3| Data collection and refinement statistics 70
Table 3.4| Network of polar interactions in Novel Type-I Coh-Doc complex interfaces. 79
Table 3.5| Thermodynamics of type I dockerin-cohesin interactions. 81
Table 4.1| Primers used to clone the genes encoding the cohesin and dockerin derivatives produced in the present study. 90
Table 4.2| X-ray data and structure quality statistics for the *C. thermocellum* SdbC2-CipBXDoc complex. 93
Table 4.3| Identification of preferred cohesin and dockerin partners. 98
Table 4.4| Thermodynamics of type II dockerin-cohesin interactions. 106
Table 5.1| Catalytic activity of *wild type* and mutants of *Ct*Cel124$_{CD}$. 118
Table 6.1| Catalytic activity of *Ct*Xyl5A and its variants. 133
Table 6.2| Binding of *Ct*Xyl5A derivatives to polysaccharides and oligosaccharides. 140
Table 6.3| Data collection statistics. 152

LIST OF FIGURES

Figure 1.1| Simplified model of the primary cell wall.....2
Figure 1.2| Molecular representation of the structure of cellulose.....4
Figure 1.3| Alternative models of cell wall structure 6
Figure 1.4| Ultrastructure of the *C. thermocellum* cell surface.....9
Figure 1.5| Schematic representation of polycellulosomes bound to cellulose cell surface. ..10
Figure 1.6| Organization of *C. thermocellum* cellulases and hemicellulases in cellulosomes. 13
Figure 1.7| Schematic representation of the *B. cellulosolvens* cellulosome system.....14
Figure 1.8| Schematic representation of the *A. cellulolyticus* cellulosomal components.....15
Figure 1.9| The complexity of *R. flavefaciens* strain FD-1 cellulosome. 17
Figure 1.10| Structure of the type I Coh-Doc complex.....22
Figure 1.11| The dual binding mode of the Xyn10B dockerin. 24
Figure 1.12| Structure of the type II Cohesin-Xdockerin complex (CohSdbA-CipAXDoc).....26
Figure 1.13| Structure of the *C. thermoc*ellum CipA scaffoldin Cohesin I9-X-Dockerin II trimodular fragment in complex with the SdbA Cohesin II module. 29
Figure 1.14| Overall view of the structure of the cellobiohydrolase CelS from family 48.....31
Figure 1.15| Structures of the three different CBM types based on topology of carbohydrate binding site. 37
Figure 2.1| DNA construct containing cohesin and dockerin genes cloned in tandem under the control of separate T7 promoter and terminator regions. 49
Figure 2.2| Purification of cohesin-dockerin complexes. 51
Figure 2.3| Crystals of cohesin-dockerin complexes and X-ray analysis.....58
Figure 2.4| The dual binding mode of cohesin-dockerin complexes.....61
Figure 3.1| *C. thermocellum* cellulosome. 72
Figure 3.2| Structure of novel type-I Cohesin-Dockerin complexes, CohOlpC-Doc124A and CohOlpA-Doc918. 73
Figure 3.3| Structure superposition and important contact residues. 75
Figure 3.4| Electrostatic surface potential for the Coh-Doc molecules. 78
Figure 3.5| Alignment of Doc918 primary sequence and its mutant derivatives (A) and examples of ITC experiments (B). 81
Figure 3.6| Alignment of Doc124A primary sequence and its mutant derivatives (A) and examples of ITC experiments (B). 83
Figure 4.1| Organization of *C. thermocellum* cellulosome. 88
Figure 4.2| Alignment of Type II dockerins fused to their neighboring X domain. 94
Figure 4.3| Alignment of *C. thermocellum* 18 type II cohesin modules.....96

Figure 4.4| Detection type II cohesin-dockerin complex formation by non-denaturing gel electrophoresis..... 97
Figure 4.5| Preference of CipA XDoc domain for cohesin partners. 99
Figure 4.6| Structure of the ScaC2-CipBXDoc type II cohesin-dockerin complex. 100
Figure 4.7| Complex interfaces between the dockerin and the X module (A and B) and between the dockerin and the cohesin (C, D, E and F). 102
Figure 4.8| Examples of the isothermal titration calorimetry (ITC) experiments between the *wild-type* CipB Xdoc, its mutant derivatives Phe124A, Leu147A and Phe148A and the *wild-type* cohesins ScaE6 A) and ScaC2 B). 107
Figure 5.1| Catalytic activity of *Ct*Cel124$_{CD}$..... 117
Figure 5.2| Crystal structure of *Ct*Cel124..... 121
Figure 5.3| Overlay of the structural fold and active site of *Ct*Cel124 and GH23 enzymes. 124
Figure 6.1| Schematic of xylan..... 129
Figure 6.2| Analysis of the reaction products generated by *Ct*Xyl5A from arabinoxylan. 134
Figure 6.3| NMR analysis of the oligosaccharides generated by *Ct*Xyl5A..... 136
Figure 6.4| ESI-MS of the tetrasaccharides in Fraction 1. 137
Figure 6.5| The structure of the tetrasaccharides generated by *Ct*Xyl5A. 138
Figure 6.6| Representative ITC data of *Ct*GH5-CBM6 to oligosaccharides..... 141
Figure 6.7| Crystal structure of *Ct*GH5-CBM6. 144
Figure 6.8| Superimposition of *Ct*GH5 and the cellulase *Ba*Cel5A..... 145
Figure 6.9| Superimposition of *Ct*CBM6 and *Cm*CBM6..... 145
Figure 6.10| Domain architecture of *Ct*Xyl5A..... 149
Figure 6.11| Coomasie Brilliant Blue-stained 14%PAGE gel evaluation of protein purity.... 150
Figure 6.12| Crystals of *Ct*Xyl5A obtained by hanging-drop vapour diffusion in the presence 40 % (v/v) 2-methyl-2,4-pentandiol and 10-20 % isopropanol. 151
Figure 6.13| Representative diffraction pattern of a *Ct*Xyl5A crystal (the outer circle corresponds to 2.64 Å resolution). 153
Figure 6.14| Crystal structure of the penta-modular cellulosomal arabinoxylanase. 154
Figure 6.15| Small Angle X-ray Scattering (SAXS)..... 155
Figure 6.16| The SAXS model of the *Ct*Xyl5A. 155

LIST OF ABBREVIATIONS AND SYMBOLS

%	Percentage
2D	Two dimensional
4-O-GlcA	4-O-methyl glucuronic acid
Å	Angstrom
A. cellulolyticus	*Acetivibrio cellulolyticus*
A_{600}	Absorbance at 600 nanometers
Ace	Acetate
AGE	Affinity gel electrophoresis
Ala	Alanine (A)
Araf	Arabinofuranose
Arg	Arginine (R)
Asn	Asparagine (N)
Asp	Aspartic acid (D)
***Ba*Cel5A**	*Bacillus agaradhaerens* cellulase 5A
B. agaradhaerens	*Bacillus agaradhaerens*
B. cellulosolvens	*Bacteroides cellulosolvens*
bp	Base-pair
BSA	Bovine serum albumin
C. acetobutylicum	*Clostridium acetobutylicum*
C. cellulolyticum	*Clostridium cellulolyticum*
C. cellulovorans	*Clostridium cellulovorans*
C. japonicus	*Cellvibrio japonicus*
C. josui	*Clostridium josui*
C. mixtus	*Cellvibrio mixtus*
C. thermocellum	*Clostridium thermocellum*
C. thermosaccharolyticum	*Clostridium thermosaccharolyticum*
$CaCl_2$	Calcium Chloride
cal	Calorie
CAZymes	Carbohydrate-active enzymes
CBD	Cellulose-binding domain
CBM	Carbohydrate-binding module
CBM13	Carbohydrate binding module from family 13
CBM3a	Cellulose binding module
CBM3a-coh	CBM3a fused to the type I cohesin3 from CipA
CBM6	Carbohydrate binding module from family 6
CBM62	Carbohydrate binding module from family 62
cbpA	*C. cellulovorans* scaffoldin gene
CCP4	Collaborative Computational Project Number 4
CE	Carbohydrate esterase
$Cel124_{CD}$	Celullase from family 124
Cel48S	Major *exo*-cellulase from *C. thermocellum* cellulosome
CelS	*C. thermocellum* major cellulosomal cellobiohydrolase
CF	Cationized ferritin
CipA	*C. thermocellum* Cellulosome integrating protein
cipA	*C. thermocellum* cellulosomal gene
CipA2	Second cohesin from CipA scaffoldin
CipBXDoc	Type II dockerin from CipB scaffoldin with its X module

***Cj*Abf51A**	GH51 arabinofuranosidase from *C. japonicus*
***Cj*Xyn10A**	GH10 xylanase from *C. japonicus*
CMC	Carboxymethyl cellulose
***Cm*Lic5A**	*C. mixtus* lichenase 5A
Coh	Cohesin
Coh-Doc	Cohesin-dockerin complex
CohOlpA-Doc918A	*C. thermocellum* Cohesin OlpA in complex with Dockerin 918
CohOlpC-Doc124A	*C. thermocellum* Cohesin OlpC in complex with Dockerin 124A
Cys	Cysteine (C)
Da	Dalton
ΔG	Gibbs energy
ΔH	Change in Entalphy of a system
DNSA	3,5-dinitrosalicylic acid
Doc	Dockerin
Doc124A	Dockerin 124A
Doc918	Dockerin 918
DocXyn10B	Xylanase 10B dockerin
Dp	Polymerization degree
ΔS	Entropy change of a system
DTT	Dithiothreitol
E. coli	*Escherichia coli*
EDTA	Ethylenediaminetetraacetic acid
ESFR	European Synchrotron Radiation Facility
ESI-MS	Electrospray ionization - mass spectrometry
Ext	*C. thermocellum* anchoring scaffoldin
Fae	Ferulate
Fn_3	Fibronectin type-III like module
FPLC	Fast protein liquid chromatography
g	gram
GC-EIMS	Gas Chromatography with Electron Impact Mass Spectrometry
GFP	Green Fluorescent protein
GH	Glycoside hydrolase
GH5	Glycoside hydrolase from family 5
Gln	Glutamine (Q)
Glu	Glutamic acid (E)
Gly	Glycine (G)
h	hour
H_2O	Water molecule
HB	Hydrogen bond
HEPES	Hydroxyethyl piperazineethanesulfonic acid
His	Histidine (H)
His_6-tag	Six Histidines tag
HPAEC	High-Performance Anion-Exchange Chromatography
Hz	Hertz
Ile	Isoleucine (I)
IMAC	Immobilized Metal Affinity Chromatography
IPTG	Isopropyl β-D-1-thiogalactopyranoside
ITC	Isothermal Titration Calorimetry
K	kelvin
Ka	Association constant

Kcat	Catalytic constant
Km	Michaelis constant
L	litre
LB	Luria Bertani
LC-MS/MS	Liquid chromatography-tandem mass spectrometry
Leu	Leucine (L)
Lys	Lysine (K)
M	molar
m/z	Mass-to-charge ratio
MALDI-TOF	Matrix-assisted laser desorption/ionization - time-of-flight
MeCN	Acetonitrile
MES	2-(*N*-morpholino)ethanesulfonic acid
Met	Methionine (M)
min	minute
mol	Unit of measurement to express amount of a chemical substance
MPD	2 methyl-2,4-pentanediol
MR	Molecular replacement
Mr	Molecular weight
mRNA	Messenger RNA
N	Normality
NaCl	Sodium Chloride
NMR	Nuclear magnetic resonance
ºC	Celcius degree
OlpA	*C. thermocellum* cell surface protein
olpA	*C. thermocellum* scaffoldin gene
OlpB	*C. thermocellum* anchoring scaffoldin
olpB	*C. thermocellum* scaffoldin gene
OlpC	*C. thermocellum* cell surface protein
orf2	*C. thermocellum* scaffoldin gene
ORF2p	*C. thermocellum* anchoring scaffoldin
PASC	Phosphoric acid-swollen *cellulose*
pcel124	DNA encoding *Ct*Cel124
PCR	Polymerase chain reaction
PD-10	Gel filtration collumns GE Healthcare
Pdb	Protein data bank
PEG	Polyethylene glycol
pH	negative decimal logarithm of the hydrogen ion activity in a solution
Phe	Phenylalanine (F)
pKa	Acid dissociation constant
PL	Polyssacharide Lyase
PMAA	Partially methylated alditol acetate
Pro	Proline (P)
R	Universal gas constant
R. flavefaciens	*Ruminococcus flavefaciens*
RC	Regenerated cellulose
Rpm	Rotation per minute
S. cerevisae	*Saccharomyces cerevisae*
SAD	Single wavelenght Anomalous Dispersion
SAXS	Small-angle X-ray scattering
SB	Salt-Bridge

ScaA	Anchoring scaffoldin
ScaA-CipAXDoc	Cohesin from ScaA in complex with XDockerin CipA
ScaB	Anchoring scaffoldin
ScaB3	Third cohesin from the anchoring scaffoldin ScaB
ScaC	Anchoring scaffoldin
ScaC1	First cohesin from the anchoring scaffoldin ScaC
ScaC2	Second cohesin from the anchoring scaffoldin ScaC
ScaC2-CipBXDoc	Cohesin 2 from ScaC in complex with XDockerin CipB
ScaD	Anchoring scaffoldin
ScaD-Unk	Anchoring scaffoldin with an unknown module
ScaE	Anchoring scaffoldin
ScaE6	Sixth cohesin from the anchoring scaffoldin ScaE
SdbA	*C. thermocellum* anchoring scaffoldin
SDS-PAGE	Sodium Dodecyl Sulfate-Polyacrylamide Gel Electrophoresis
Ser	Serine (S)
Ser-Thr	Serine-Threonine pair
SLH	S-layer homology module
SSM	Secondary Structure Matching
T	Absolute temperature
T. reesei	*Trichoderma reesei*
TEM	Transmission Electron Microscopy
TFA	Trifluoroacetic acid
Thr	Threonine (T)
Tris	2-Amino-2-hydroxymethyl-propane-1,3-diol
Tyr	Tyrosine (Y)
U	Enzymatic unities
V	Volt
Val	Valine (V)
w/v	Weight per volume
w/w	Weight per weight
Wat	H_2O bridge
Xyl5A	Arabinoxylanase 5A from *C. thermocellum* cellulosome
Xylp	Xylopyranose

1. BIBLIOGRAPHIC REVIEW AND OBJECTIVES

1.1. Introduction

Today, society faces the challenge of finding alternative and renewable energy sources to the conventional fossil fuels. 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 organic carbon on earth. Photosynthetically fixed carbon is recycled by numerous microbial enzymes that hydrolyse cell wall polysaccharides and play an important role in nature 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 individualy in the hydrolysis of structural polysaccharides. The released products are then used as a carbon and energy source by the cells. In contrast, the energetic constraints posed by anaerobic ecosystems lead to the evolution of a remarkably highly efficient supramolecular complex, termed Cellulosome, which is attached to the microorganism and efficiently degrades a variety of plant cell wall polysaccharides. Anaerobic organisms display a lower production protein capacity and thus the improved efficiency that results from enzyme assembly leads to a higher relative capacity to degrade the lignocellulose biomass. The cellulosome of the Gram-positive thermophilic bacterium *Clostridium thermocellum* is the paradigm for such enzymatic nanomachines. This large extracellular enzymatic complex comprises various cellulases, hemicellulases and pectinases anchored to a large non-enzymatic multimodular primary scaffoldin protein attached to the cell wall of the microrganism. Additionally, this scaffoldin contains a noncatalytic Carbohydrate-Binding Module (CBM) that anchors the entire complex onto crystalline cellulose.

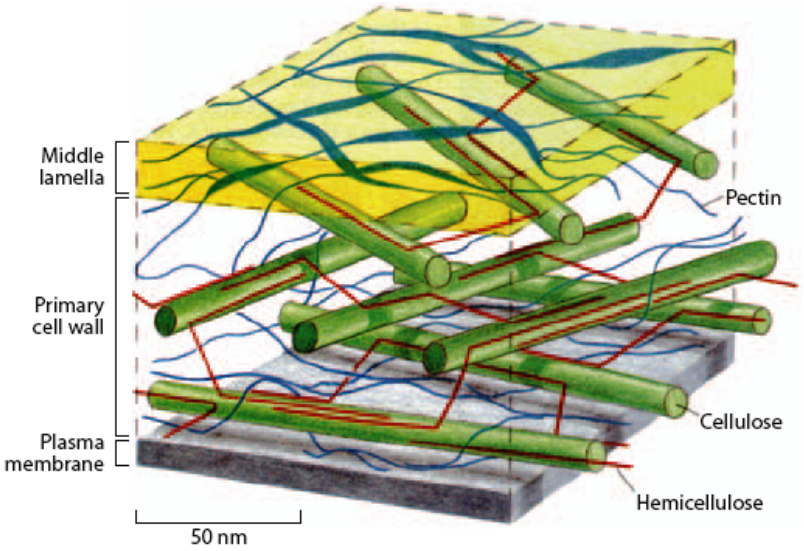
This chapter begins with a general review on plant cell wall composition, with particular focus on cellulose and different polysaccharide constituents, followed by a description of the different mechanisms required for plant cell wall degradation. Subsequently, cellulosome complexity and functionality will be analysed according to the current knowledge on the structure and function of the different cellulosomal components, particularly, Cohesin and Dockerin modules, Glycoside-hydrolases (GHs) and CBMs. 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 Components

Plant cells are encapsulated within a complex and fibrous wall whose properties are crucial to both the form and function of plants. The plant cell wall acts as an exoskeleton to give plant cell its shape and to allow high turgor pressures (Cosgrove, 1997). Plants have two cell wall types with different functions and composition that are termed the primary and the secondary cell walls. Primary cell walls (Figure 1.1) surround cells capable of growth, providing mechanical strength but allowing, at the same time, the cell to expand. They are composed of cellulose microfibrils embedded in a highly hydrated gel-like matrix of non-cellulosic polysaccharides and glycoproteins. The cellulose-hemicellulose network co-exists with a network consisting of pectic polysaccharides (Popper *et al.*, 2011). Secondary cell walls are thickened structures that surround specialized cells, such as vessel elements or fiber cells. (Keegstra, 2010). The secondary walls are strengthened by the incorporation of lignin, a phenolic polymer, which cements and anchors the cellulose microfibrils among other matrix polysaccharides. This stiffens the walls, preventing biochemical degradation and physical damage (Popper, 2008).

Figure 1.1| Simplified model of the primary cell wall.



(Scheller & Ulvskov, 2010)

The most abundant component found in all plant cell walls is cellulose. It consists of a collection of β-1,4-linked glucan chains that interact with each other via hydrogen bonds to form a crystalline microfibril (Somerville, 2006). The latter contains the crystalline allomorphs, cellulose Iα and Iβ (Brett & Waldron, 1996). In addition to cellulose, plant cell walls contain several matrix polysaccharides that are grouped into two general categories: the pectic

polysaccharides which include homogalacturonan, and rhamnogalacturan I and II (Harholt, Suttangkakul, & Scheller, 2010) and the hemicellulosic polysaccharides that include xyloglucans, glucomannans, xylans, and mixed-linkage glucans (Scheller & Ulvskov, 2010). Plant cell walls also contain many proteins and glycoproteins, including various enzymes and structural proteins (Keegstra, 2010).

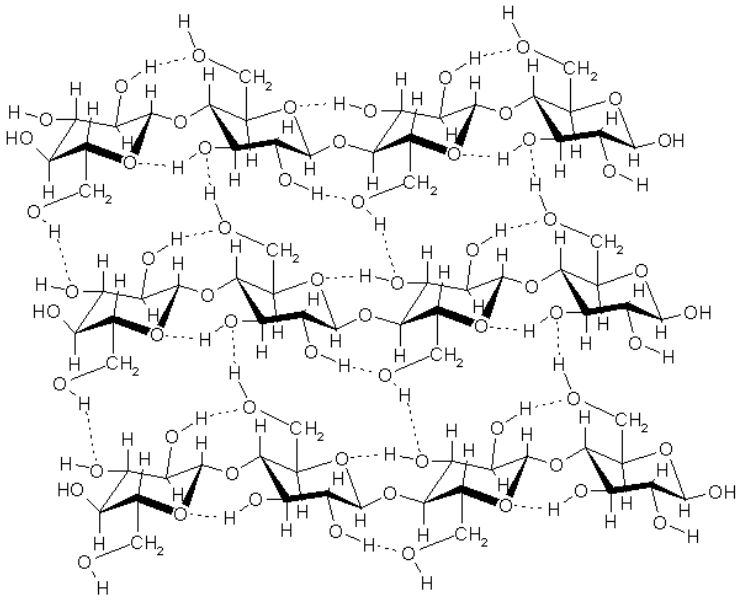
1.2.1.1. Cellulose

Cellulose is the most widespread biopolymer on the planet. Its primary structure is of an unbranched β–1,4 linked D-glucan. Many parallel glucans form a crystalline microfibril that is mechanically strong and highly resistant to enzymatic attack (Cosgrove, 2005). Cellulose has a chemical repeating unit of D-glucopyranose, linked by β-1,4 glycosidic bonds linkages in a "zigzag" arrangement between oxygen bridges. The structural repeat is the disaccharide cellobiose (Figure 1.2). Cellulose biogenesis results from the coordinated action of enzymatic polymerization, followed by the extrusion and crystallization of the nascent cellulose microfibrils. The combination of these events leads to the production of whisker-like crystalline microfibrils, wherein the cellulose chains are packed in a parallel fashion. The microfibrils are then assembled into superstructures, such as cell walls, fibers and pellicles, held together predominantly by strong interactions: hydrogen bonds and hydrophobic stacking between the sugar rings (Bayer, Chanzy, Lamed, & Shoham, 1998).

Natural crystalline cellulose is named cellulose I, or native cellulose, and comprises the two forms Iα and Iβ, in which these chains lie parallel (Jamal, Nurizzo, Boraston, & Davies, 2004). The Iα form consists of a single-chain triclinic structure, whereas the Iβ form is monoclinic and is characterized by two parallel chains. The density and stability of the Iα form was shown to be lower and its enzymatic or chemical reactivity is therefore higher (Bayer *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 may be obtained from cellulose I by either of two processes: regeneration, which is the solubilization of cellulose I in a solvent, followed by reprecipitation by dilution in water to give cellulose II, or mercerization, which is the process of swelling native fibres in concentrated sodium hydroxide, to yield cellulose II on removal of the swelling agent (OSullivan, 1997). In addition, many model sources of natural crystalline cellulose such as microcrystalline cellulose (for example Avicel™, extracted from plant fiber), bacterial microcrystalline cellulose (purified from *Acetobacter xylinum* pellicles), 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 (OSullivan, 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 *et al.*, 1998). With respect to cellulose biosynthesis, it is also known that it takes place at the nonreducing end of the growing chain (Bayer *et al.*, 1998).

Figure 1.2| Molecular representation of the structure of cellulose



The β-configuration allows cellulose to form very long, straight chains. Fibrils are formed by parallel chains that interact with one another through hydrogen bonds.
Adapted from: http://chemphys.gcsu.edu/~metzker/Common/Structures/Carbohydrates/cellulose.png

1.2.1.2. Hemicellulose

Hemicelluloses are highly branched polysaccharides that are hydrogen-bonded to the surface of cellulose microfibrils. 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). Hemicelluloses are structurally homologous to cellulose but contain β-linked backbones decorated with a variety of sugars and acetyl groups, which explains why these polymers are not crystalline (Gilbert, 2010). The detailed structure of the hemicelluloses and their abundance vary widely between different species and cell types (Scheller & Ulvskov, 2010). Xyloglucan, xylan, arabinoxylan, mannan and glucomannan are examples of these hemicellulosic polysaccharides. They have a backbone composed of β-1,4-D-pyranosyl linked residues with O4 in equatorial orientation (Cosgrove, 2005).

Xyloglucan is the most abundant hemicellulosic polysaccharide in the plant cell wall of non-gramineas and has a semirigid backbone of β-1,4-glucan that is decorated with xylose branches on 3 out of 4 glucose residues. The xylose can also be appended with galactose

and fucose residues (Cosgrove, 2005). Less branched xyloglucans are known to be less soluble (Scheller & Ulvskov, 2010).

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-D-xylan linked 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 a 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

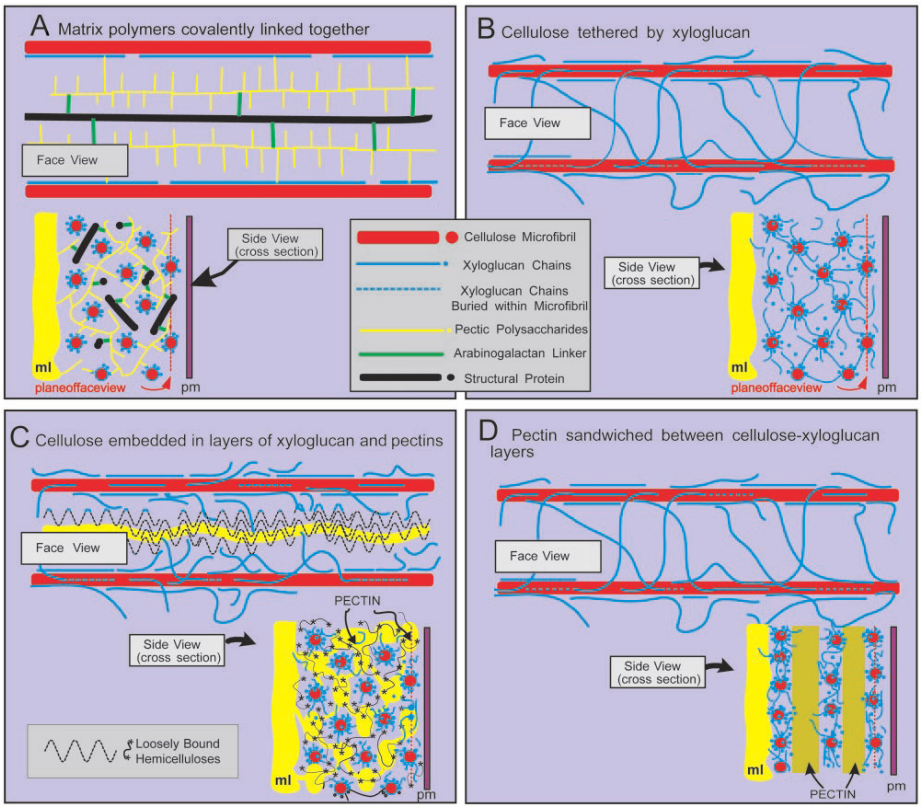
Pectin, the most soluble of the cell wall polysaccharides, constitute a heterogeneous group of polysaccharides, characteristically containing acidic sugars, such as glucuronic acid and galacturonic acid (Cosgrove, 1997). Ramnogalacturan I consists of alternative residues of α-1,4 D-galacturonic acid and α-1,2 L-rhamnose, decorated primarily with arabinan and galactan side chains. It has been suggested that ramnogalacturan I functions as a scaffold to which other pectins, such as ramnogalacturan II and homogalacturonan are covalently attached as side chains (Somerville *et al.*, 2004). Homogalacturonan, the simplest of these polymers, comprises a linear chain of α-1,4 D-galacturonic acid residues, whereas xylogalacturonan are often methyl esterified, a modification that blocks the acidic group and reduces their ability to form gels. Rhamnogalacturonan II is a complex pectin domain that contains 11 different sugar residues and forms dimers through borat esters. The neutral arabinans and arabinogalactans are also linked to the acidic pectins and it has been proposed that they promote wall flexibility and that they bind to the surface of cellulose (Cosgrove, 2005).

1.2.2. Plant Cell Wall Models

Over the years, several models have been proposed to explain the interactions established between plant cell wall components. In 1973, Keegstra *et al.*, proposed that matrix polymers, consisting of xyloglucan, pectic polysaccharides, and structural proteins, are covalently

linked to form a giant macromolecular network. Besides that, cellulose was proposed to be bonded to the matrix via H-bonding to xyloglucans (Keegstra, Talmadge, Bauer, & Albersheim, 1973). Later, an alternative model proposed by Hayashi (1989) and Fry (1989) defended that cellulose microfibrils may be tethered together directly via long xyloglucan chains. Pectic polysaccharides and structural proteins are imagined to form co-extensive, but independent, networks that physically entangle the cellulose-xyloglucan network, but that are not covalently bonded to it. Although this is the most popular model, two different models have been lately proposed. The first is the multicoat model, explained by Talbott and Ray, in which each microfibril is coated by a series of progressively less-tighly bound polysaccharides layers. In addition, the linkage between microfibrils is made indirectly through non-covalent bindings between the distinctive polysaccharide layers (Talbott & Ray, 1992). The second is the stratified wall model, conceived by Ha *et al.* (1997), in which pectic layers serve as spacers between cellulose and hemicellulose lamellae (Ha, Apperley, & Jarvis, 1997).

Figure 1.3| Alternative models of cell wall structure



A) Keegstra *et al.* model, in which the matrix polymers are all covalently linked to one another and anchor cellulose by hydrogen bonding to xyloglucans. **B)** The "tethered network" model of Hayashi and Fry in which single xyloglucan chains span the gap between microfibrils ant tether them together. **C)** The "multicoat" model of Talbott and Ray in which cellulose is coated with successively looser layers of matrix polysaccharides. **D)** The "stratified" wall model of Ha *et al.* in which pectic layers serve as spacers between cellulose and hemicellulose lamellae. pm: Plasma membrane; ml: middle lamella. Adapted from Cosgrove (2001).

In summary, most of these models have focused on understanding the organization of the components in primary cell walls, which would allow regulated reorganization of wall components during cell growth and differentiation (Figure 1.3). Hemicellulosic polysaccharides are known to bind tightly to cellulose microfibrilis via hydrogen bonds (through OH groups on the sugars) and most wall models have incorporated this interaction as one important feature of cell wall architecture (Keegstra, 2010).

1.2.3. Plant Cell Wall Hydrolysis

Plant cell wall polysaccharides, primarily cellulose and hemicelluloses are a major reservoir of carbon and energy. However, since they have a high chemical and physical complexity, only a restricted number of microorganisms have acquired the ability to deconstruct these structural carbohydrates (Fontes & Gilbert, 2010). In this way, the microbial degradation of plant cell wall is a complex process in which an extensive battery of hydrolytic enzymes attacks a heterogeneous insoluble and highly recalcitrant substrate. The requirement for a consortium of enzymes reflects the physical association of the polysaccharides within the plant cell wall, which demands that the catalytic entities act in synergy to degrade this composite structure (Gilbert, 2007).

Enzyme secretion by plant cell wall degrading microorganisms can be divided along two lines. Thus, in aerobes, enzymes are either secreted into the extracellular milieu or are located on the outer membrane. Although these enzymes do not physically associate, they do display extensive biochemical synergy. Besides that, it was also found that many of these biocatalysts possess a multi-modular structure composed of a catalytic module linked to one or more CBMs, which improve enzyme efficacy by targeting the catalytic module to surfaces of insoluble substrates (Gilbert, 2007; Gilbert, Stålbrand, & Brumer, 2008). Alternatively, the plant cell wall degrading enzymes in most anaerobic bacteria and fungi associate into a large multienzyme complex (with a molecular weight higher than 3MDa), termed Cellulosome. The catalytic modules are grafted onto a non-catalytic scaffold protein that also contains CBMs, thus creating an intimate link between the cell and the substrate surface (Fontes & Gilbert, 2010). It is believed that the anaerobic environment impose a greater selective pressure for the evolution of these highly efficient nanomachines (Bayer, Belaich, Shoham, & Lamed, 2004).

Although only a single type of reaction, hydrolysis of β-1,4-glycosidic bonds, is required to convert cellulose to soluble products, degradation 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. A more detailed characterization of each enzyme is given below.

1.2.4. Carbohydrate-Active Enzymes database (CAZy)

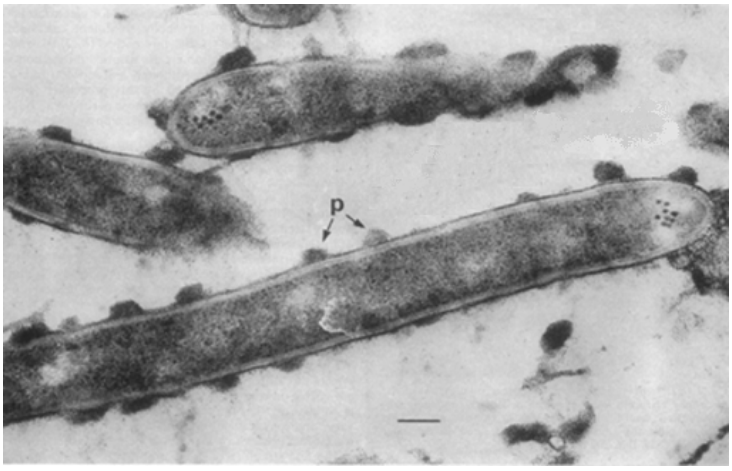
A large variety of enzymes that act on complex carbohydrates and also several CBMs have been identified and characterized in the last years. In order to organize all this knowledge, these proteins have been grouped into sequence-based families on the continuously updated Carbohydrate-Active enZymes database (CAZy) (Cantarel *et al.*, 2009). Nearly 20 years ago, the first foundation for a family classification of CAZymes was made to organize cellulases into several distinct families based on amino-acid sequence similarity. Later, the family classification system based on protein sequence and structure similarities, was extended to all known glycoside hydrolases (GHs), and subsequently extended to all CAZymes involved in the synthesis, degradation and modification of glycoconjugates (Cantarel *et al.*, 2009). In 1998, the classification of CAZymes became available on the web. Since the classification is based on amino-acid sequence similarities, it is possible to correlate the sequence with both enzyme mechanisms and protein fold. Significant sequence similarity (usually over 30%) is a strong indication of folding similarities. Therefore, members of one family most likely share the same folding characteristics. Thus, if the three-dimensional structure of one member is known, it is possible to do homology modelling and deduce structural insights for other family members. Consequently, these families are used to conservatively classify proteins of uncharacterized function (which only known feature is sequence similarity). In this way, this classification method avoids overprediction of enzyme activities (Cantarel *et al.*, 2009). Additionally, Henrissat *et al* (1998) also proposed an enzyme terminology for glycoside hydrolases that is based on the target substrate. The enzymes can be named according to their target substrate (following the three-letter standard used in bacterial genetics for genes) followed by the family number and by a uppercase letter that corresponds to the order in which the catalytic domain was first reported. For example, a family 5 GH will be named Cel5 or Man5, depending on its substrate that could be cellulose or mannose, respectively, and by Cel5A or Cel5B if there were two catalytic domains with the same specificity, but reported at different times. The microorganism abbreviation is also included before the enzyme name, in order to differentiate similar enzymes of different origins (Henrissat, 1998). At present, CAZy covers approximately 300 protein families in the following classes of enzymes activities: Glycoside hydrolases (GHs), Polysaccharide lyases (PLs), Carbohydrate esterases (CEs), CBMs and also Glycosyltransferases (May 2012).

1.3. The Cellulosome: Architecture and Function

In the early 1980s, multienzyme complexes, known as Cellulosomes, were identified in many anaerobic cellulolytic microorganisms. These complexes were shown to be dedicated to the efficient degradation of cellulose and hemicelluloses (Bayer, Morag, & Lamed, 1994). Cellulosomes were identified in anaerobic bacteria from the genera *Clostridium, Acetivibrio, Bacteroides* and *Ruminococcus*, that colonize various environmental niches, including the soil, wood chip piles, sewage and the rumen (Doi, Kosugi, Murashima, Tamaru, & Han, 2003). Bayer and Lamed were the first to observe cellulosomes as large protuberances on the surface of *Clostridium thermocellum,* an anaerobic, thermophilic, cellulolytic, gram positive bacterium (Figure 1.4). Later, on the basis of combined biochemical, immunochemical, ultrastructural and genetic techniques it was possible to identify and describe this multi-enzymatic complex (Bayer *et al.*, 1998).

The cellulosome was found to be initially located in protuberances present on the cell surface and to be subsequently released into the culture medium. Because all known sequences of cellulosome polypeptides start with a typical signal peptide, they are most likely secreted individually through a general secretion pathway. Thus, attachment of catalytic components probably occurs on the surface of the cells (Béguin & Lemaire, 1996).

Figure 1.4| Ultrastructure of the *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. Adapted from Bayer & Lamed (1986).

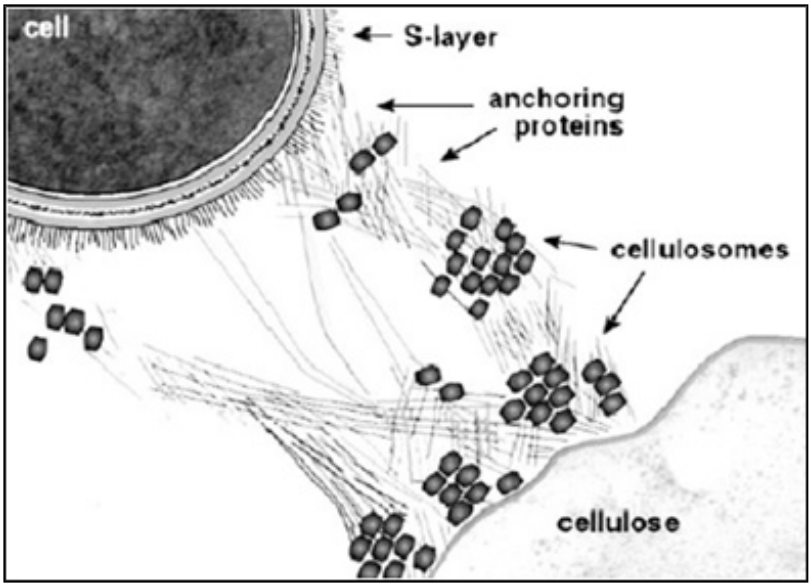
In the following years, several cellulase genes from this bacterium were cloned and sequenced and their modular structure was determined. Initially, the cellulosome was believed to exclusively degrade cellulose, but soon it was recognized that the complex contains not only cellulases but also a large array of hemicellulases and even pectinases (Bayer *et al.*, 1998; Fontes & Gilbert, 2010).

The principal component of the *C. thermocellum* cellulosome is a scaffoldin subunit, named CipA. This is a large enzyme-integrating protein that contains nine highly conserved

modules, known as type I cohesins, which incorporate the different enzymes. The cellulosomal catalytic components contain noncatalytic modules, called dockerins, which bind to the cohesin modules, through a very tight protein:protein interaction. The C-terminal region of CipA also contains a type II dockerin that, through its interaction with type II cohesin modules located in proteins anchored to the bacterial peptidoglycan layer, tethers the cellulosome on the cell surface of the prokaryote (Bayer *et al.*, 2004; Xu *et al.*, 2003). Although structurally related, there is no cross-specificity between type I and type II cohesin-dockerin partners ensuring a reliable mechanism for cell-surface attachment and cellulosome assembly (Leibovitz & Béguin, 1996). CipA also contains a family 3 CBM (CBM3) that interacts tightly with crystalline cellulose and thus, plays a key role in bringing the cellulosome into close proximity with its target substrate, the plant cell wall (Bayer *et al.*, 2004; Gilbert, 2007).

The physical association of the enzymes within cellulosomes, or in polycellulosomes, is believed to potentiate the biochemical synergy between these enzymes, which probably means that cellulosomes are more efficient at deconstructing plant structural polysaccharides (Figure 1.5), when compared to the “free” enzyme systems produced by aerobic bacteria and fungi. The synergistic effects are due to the targeting effect of the scaffoldin-born CBM, the proximity effect of the enzymes and also due to the elimination of substrate inhibition from the quick uptake of released products (Goyal, Tsai, Madan, DaSilva, & Chen, 2011). Indeed, *C. thermocellum* exhibits one of the highest rates of cellulose utilization known and the cellulosome of the bacterium is reported to display a specific activity against cellulose that is 50-fold higher than the corresponding aerobic *Tricoderma* system (Demain, Newcomb, & Wu, 2005).

Figure 1.5| 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 *et al.* (1998).

The genome sequences of *C. thermocellum, Clostridium acetobutylicum, Ruminococcus flavefaciens* and *Clostridium cellulolyticum* are already known. These sequences will provide a complete view of the molecular components of the cellulosome of each organism (Fontes & Gilbert, 2010).

1.3.1. The scaffoldin

The defining components that distinguish the cellulosome from free enzymes systems are the cohesin containing-scaffoldin on the one hand, and the dockerin-containing enzymes on the other hand. The scaffoldin, the large modular integrating protein that binds the various cellulosomal enzymes, contains several cohesin modules, named type I cohesins, for the incorporation of the various cellulosomal biocatalysts which have a complementary module, the type I dockerin domain. However, cohesin modules are not restricted to the enzyme integrating scaffoldins, usually termed primary scaffoldins, since other cohesins may serve supplementary functions as anchoring or adaptor proteins (Bayer *et al.*, 2004).

The scaffoldin of *Clostridium cellulovorans* was the first to be sequenced. By that time, the cellulose-binding function of the cellulosome was recognized, but the meaning of the repeating elements was still unknown (Shoseyov, Takagi, Goldstein, & Doi, 1992).

Later, when the scaffoldin from *C. thermocellum* was sequenced, the relationship between the cohesins and the dockerins was revealed. *C. thermocellum* primary scaffoldin, named CipA, is an 1853 amino acid non-catalytic polypeptide which contains nine highly conserved type I cohesins. Thus, type I cohesins recognize the type I dockerins from the catalytic subunits (Béguin & Aubert, 1994; Felix & Ljungdahl, 1993; Gerngross, Romaniec, Kobayashi, Huskisson, & Demain, 1993). Mesophilic bacteria as *Clostridium josui,* and *C. cellulolyticum* were also shown to contain scaffoldins that appear to resemble the *C. cellulovorans* scaffoldin, rather than the *C. thermocellum*, since the later contains a C-terminal dockerin that is lacking in the others (Kakiuchi *et al.*, 1998; Pagès *et al.*, 1999). Since *C. thermocellum* CipA C-terminal dockerin displays different cohesin specificity when compared with the type I dockerins located in the cellulosomal enzymes it was named as type II dockerin. It was shown that this type II dockerin interacts with type II cohesins of proteins which also contain C-terminal S-layer homology (SLH) modules that mediate attachment to the cell surface. Thus, proteins that contain an SLH module were termed anchoring scaffoldins and mediate the cell surface attachment of primary scaffoldins.

Therefore, bacterial cellulosomes may be categorized into two major types: those that contain a single primary scaffoldin and those that present multiple types of scaffoldins including primary and anchoring scaffoldins (Bayer *et al.*, 2004).

Cellulosomes that assemble via a single primary scaffoldin are characteristic of most mesophilic *Clostridia - C. cellulolyticum, C. cellulovorans, C. josui, and C. acetobutylicum,*

among others. These are considered the simplest cellulosomes: only contain type I cohesins and lack type II dockerins. Thus, they are not attached to the bacterial cell surface. This may explain why these enzyme complexes are found in the culture media, particularly in the late exponential and stationary growth phases (Fontes & Gilbert, 2010). In addition, many of the cellulosomal genes are arranged in a cluster on the genome, consisting of the scaffoldin gene followed sequentially downstream by the various dockerin-containing enzyme genes (Bagnara-Tardif *et al.*, 1992; Kakiuchi *et al.*, 1998; Nölling *et al.*, 2001; Tamaru, Karita, Ibrahim, Chan, & Doi, 2000).

Bacteria expressing cell-surface cellulosomes, such as *C. thermocellum, Acetivibrio cellulolyticus* and *Bacteroides cellulosolvens* contain a single primary scaffoldin and usually multiple anchoring scaffoldins. A more elaborate arrangement is observed wherein 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 CipA-containing operon (Lemaire, Ohayon, Gounon, Fujino, & Béguin, 1995). *A. cellulolyticus* and *B. cellulosolvens* also contain dockerin-containing scaffoldins and SLH-bearing anchoring proteins (Xu *et al.*, 2003). Significantly, type I and type II cohesin-dockerin partners do not interact, ensuring a clear distinction between the mechanism for cellulosome assembly and cell-surface attachment (Leibovitz & Béguin, 1996). In addition, the cellulosome system characterized by multiple scaffoldins usually bears a single cellulose-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).

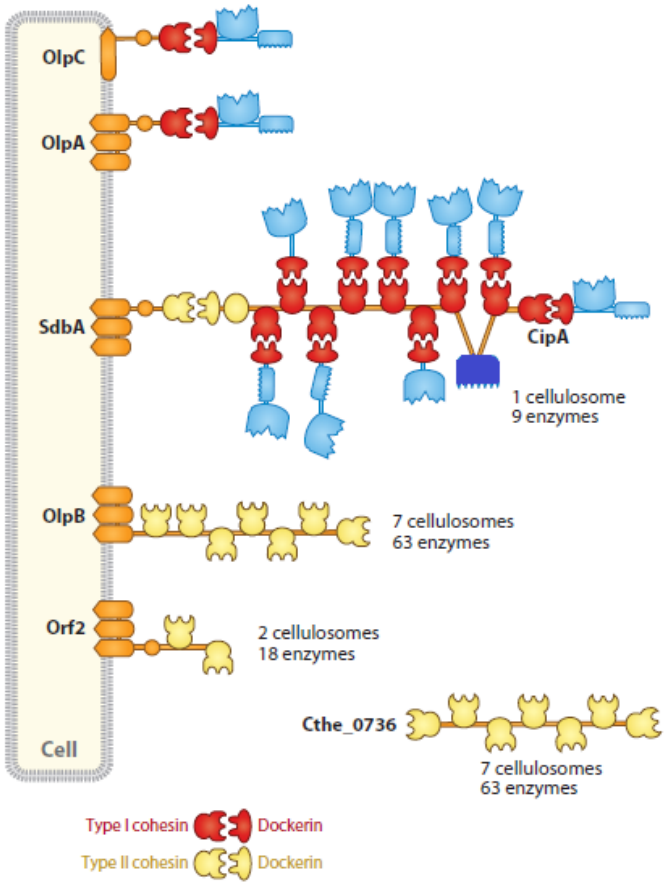
1.3.2. Cellulosome Cell surface attachment in *C. thermocellum*

As explained above, *C. thermocellum, A. cellulolyticus* and *B. cellulosolvens* contain multiple anchoring scaffoldins. The majority of these contain the SLH modules, which are threefold reiterated segments, that are usually present in the majority of S-layer proteins. SLH modules may bind the peptidoglycan layer or secondary cell wall polysaccharides (Xu, Bayer, *et al.*, 2004). In *C. thermocellum,* cellulosomes were shown to be located at the cell surface in the early stages of growth. The CipA type II dockerin recognizes type II cohesins located in cell-surface proteins (Leibovitz & Béguin, 1996; Leibovitz, Ohayon, Gounon, & Béguin, 1997). There are three described anchoring scaffoldins containing type II cohesins: SdbA, OlpB, and ORF2p. 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. SdbA, ORF2p, and OlpB present one, two or seven type II cohesins, respectively (Leibovitz &

Béguin, 1996). In contrast, a fourth anchoring scaffoldin that contains seven type II cohesins, Cthe_0736, is believed to be exclusively extracellular (Figure 1.6). The presence of tandem-repeated type II cohesins in anchoring scaffoldins allows for the formation of polycellulosomes that may contain up to 63 catalytic subunits (Fontes & Gilbert, 2010).

Additionally, type I cohesins modules from *C. thermocellum* were identified in two cell surface proteins, OlpA and OlpC, suggesting that cellulosomal enzymes can also adhere directly, and individually, onto the bacterial cell envelope (Figure 1.6) (Fontes & Gilbert, 2010; Salamitou, Lemaire, *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 CipA cohesins than to OlpC, the dominant type I cohesin-containing cell surface protein (Raman *et al.*, 2009). In this way, it was suggested that cellulosomal enzymes may transiently interact with the bacterium's cell surface through the binding to OlpC, before they are assembled into the multi-enzyme complexes (Pinheiro *et al.*, 2009).

Figure 1.6| Organization of *C. thermocellum* cellulases and hemicellulases in cellulosomes.

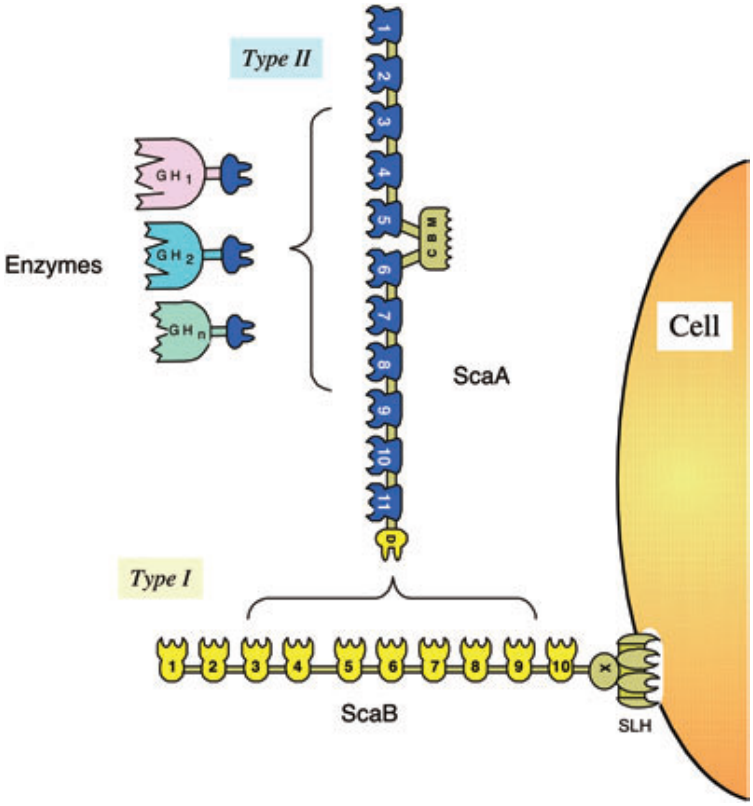


The *C. thermocellum* scaffoldin (CipA) contains nine type I cohesins and thus organizes a multiprotein complex with nine enzymes (blue), through the binding to type I dockerins. CipA also contains a cellulose-specific family 3 CBM (dark blue). The C-terminal type II dockerin domain of CipA binds specifically type II cohesin domains found in cell-surface proteins SdbA, OlpB, and Orf2 (orange) or in the extracellular Cthe_ 0736. Cellulosomal enzymes may adhere directly to the bacterium cell surface by binding the single type I cohesin domains found in OlpA and OlpC. The linkers joining the modules in the scaffoldin and catalytic subunits are shown as orange and blue lines, respectively. Adapted from Fontes & Gilbert (2010).

1.3.3. The Diversity of Cellulosomes

Studies on anaerobic organisms that do not belong to the genus *Clostridium* have revealed fundamental differences in the organization of cellulosome complexes and in the nature of the cohesin-dockerin pairs. An example of this is provided by *B. cellulosolvens* which has a two-component scaffoldin arrangement similar to that of *C. thermocellum,* except that the types of cohesins carried by the primary and anchoring scaffoldins are reversed (McLean *et al.*, 2000; Xu, Bayer, *et al.*, 2004). The cellulosome of *B. cellulosolvens* comprises a primary scaffoldin, named ScaA that contains 11 type II cohesins with a C-terminal type I dockerin, and an anchoring scaffoldin, named ScaB that bears 10 type I cohesins (Figure 1.7). Altogether, this system is able to comprise a total of 110 dockerin-containing enzymes into the cellulosome complex (Xu, Bayer, *et al.*, 2004).

Figure 1.7| Schematic representation of the *B. cellulosolvens* cellulosome system.

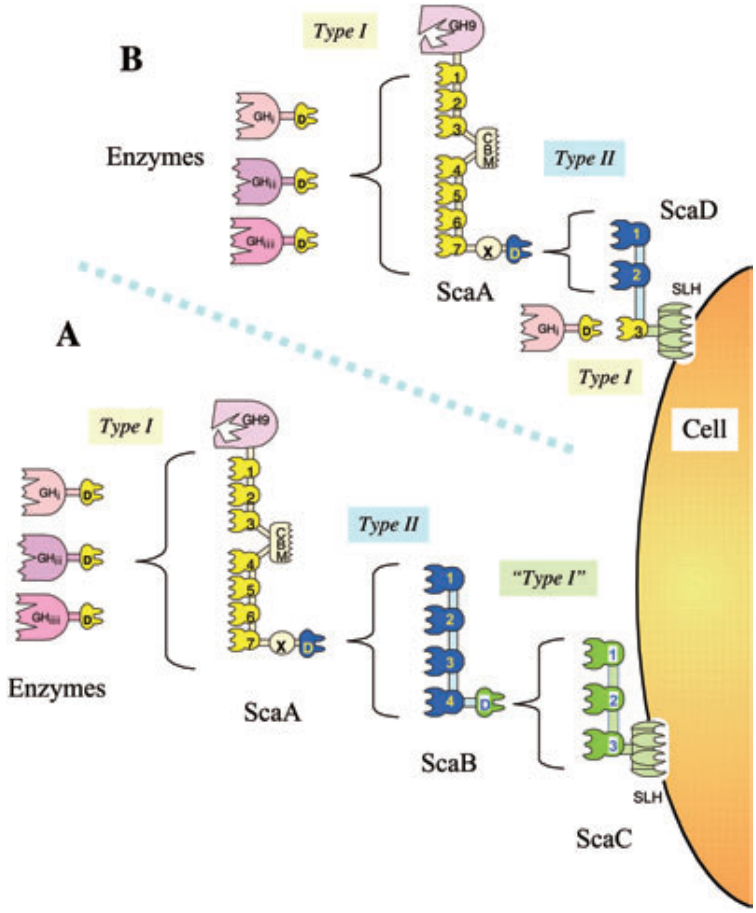


This cellulosome includes two known scaffoldins, ScaA (primary scaffoldin) and ScaB (anchoring scaffoldin) with 11 and 10 cohesins, respectively. The types of cohesins carried by the primary and anchoring scaffoldins are reversed. Adapted from Bayer, Lamed, White, & Flint, 2008.

In *A. cellulolyticus*, experimental evidence indicates that its cellulosome system comprises four different scaffoldins (Figure 1.8). In its capacity to integrate dockerin-containing enzymes, ScaA can be considered a primary scaffoldin. In contrast, ScaB essentially plays the role of an adaptor protein, which mediates the interaction between ScaA (and its attached enzymes) and ScaC. *A. cellulolyticus* ScaB is the first example of an adaptor protein. ScaC, on the other hand, clearly plays the role of an anchoring scaffoldin by virtue of

its C-terminal SLH module (Xu *et al.*, 2003). Later, Xu *et al.* (2004) succeeded in completing the sequence of a final gene of the *A. cellulolyticus* cluster, *scaD*. ScaD contains, in the same polypeptide chain, two different types of cohesin, type I and type II, which exhibit 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 and thus, it can hypothetically act as an anchoring protein. Since each primary scaffoldin represents eight enzymes, the ScaD-anchored cellulosome system of *A. cellulolyticus* would carry up to 17 enzymes, in addition to the 96 enzymes that can be apparently assembled through the ScaC-anchored system. Thus, one can predict that at least 113 enzymes can be incorporated into an *A. cellulolyticus* cellulosome (Xu *et al.*, 2004).

Figure 1.8| Schematic representation of the *A. cellulolyticus* cellulosomal components.



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

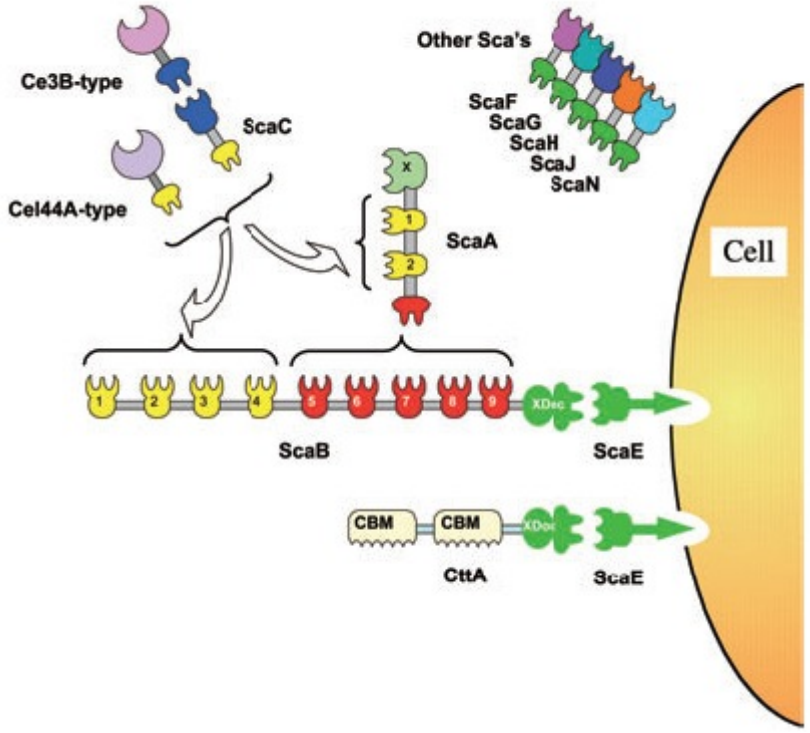
More recently, it was realized that *R. flavefaciens*, a prominent rumen bacteria, presents the most intricate and potentially versatile cellulosomal complex described to date (Figure 1.9). Initial characterization of *R. flavefaciens* cohesins suggested clear sequence and structural differences to the previously described type I and type II cohesin-dockerin complexes.Thus cohesin-dockerin pairs of *R. flavefaciens* were termed type III modules (Ding *et al.*, 2001). *R. flavefaciens* produces a cellulosome complex that is known to involve the cohesin-containing structural components ScaA, ScaB, ScaC and ScaE, together with interacting enzymes and unidentified proteins that carry dockerin and cohesin domains. In strain FD-1, the anchoring scaffoldin ScaE contains a single cohesin domain. Unlike *Clostridia*, in which cell-surface attachment is mediated through the noncovalent binding of SLH modules to the S-layer of the host cell, in *R. flavefaciens* ScaE is covalently attached to the cell surface through a sortase-mediated transpeptidation.

Additionally, ScaE interacts with the ScaB C-terminal dockerin via a novel cohesin-dockerin interaction (Rincon *et al.*, 2005). Furthermore, ScaE may bind a protein termed CttA, which carries two putative CBMs that mediate the primary anchorage to insoluble substrates (Rincon *et al.*, 2007). ScaB contains nine cohesins which present two different specificities: four recognize 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 thus, amplifies the number of enzymes in the *R. flavefaciens* cellulosome. ScaC, a small dockerin-bearing protein that also possesses a single divergent cohesin domain, was shown to bind to both ScaA and ScaB via its dockerin and also binds to a different, and yet unknown, group of dockerins via its divergent cohesin. ScaC has therefore been proposed to serve as an adaptor protein that enhances the repertoire of subunits present in the cellulosome (Rincón *et al.*, 2004). In addition, *R. flavefaciens* FD-1 strain encodes for more than 200 dockerin-containing proteins, including homologues of the scaffoldin Sca proteins (Bayer *et al.* 2008). However, previous studies revealed that cellulosome structural organization varies between strains of this bacterium, which may reflect the complexity of the rumen ecosystem and the diversity of the lignocellulosic substrate (Jindou *et al.*, 2006, 2008).

The central factor contributing to the enhanced amplification of both *A. cellulolyticus* and *R. flavefaciens* systems are the adaptor proteins. In contrast, in *C. thermocellum* cellulosomal system, hypotetical polycellulosomes may contain up to 63 catalytic units, since nine enzymes can be incorporated into the CipA scaffoldin, and up to seven CipA molecules can be assembled with the seven type II cohesins of the anchoring scaffoldin OlpB. The role of amplified enzyme incorporation into a cellulosome presumably reflects the already described proximity effect of the cellulosome which is one of the key factors for efficient digestion of recalcitrant forms of cellulose. Concentration of complementary cellulolytic enzymes on the

surface of the substrate and in the vicinity of the bacterial cell surface should thus enhance the synergistic action of the enzymes (Ding *et al.*, 2001; Xu *et al.*, 2003).

Figure 1.9| 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 (yellow) 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.

Together with bacteria, fungi have also been found to degrade cellulose and other plant cell wall fibers. Anaerobic fungi represent a special group of microorganisms inhabiting the gastro-intestinal tract of ruminants and most non-ruminant herbivores. These fungi, along with anaerobic bacteria produce a range of cellulolytic and hemicellulolytic enzymes, which were also found to be organized in a cellulosome (Ljungdahl, 2008). The size of fungal cellulosomes is estimated to range between 3 to 80 MDa (Ali *et al.*, 1995). Most of the anaerobic fungal lignocellulolytic enzymes are associated with the cellulosome. Many of the identified cellulases were also shown to have a CBM which is connected to the catalytic domain by a flexible linker. In addition, all enzymes have noncatalytic subunits, known as fungal dockerin domains, which allow binding to putative scaffolding proteins that remain to be identified. Interestingly, several fungal dockerin domains have been identified so far and their amino acid sequences were shown to be significantly unrelated to their bacterial

counterparts (Ljungdahl, 2008; Nagy *et al.*, 2007). Furthermore, unlike bacterial proteins, in anaerobic fungi, such as *Piromyces equi*, the dockerins of cellulosomal enzymes are often present in tandem copies. Generally, each enzyme contains two copies of a dockerin module which contains about 40 highly conserved residues joined together by short linker sequences (Ljungdahl, 2008). Previous studies have shown that cellulosome assembly in fungi may not involve cohesin modules. Thus, it has been proposed that the dockerin domains bind the carbohydrate decorations on glycoside hydrolase cellulosomal β-glucosidase 3 (GH3), which was identified as a potential scaffoldin (Nagy *et al.*, 2007). More work is required in order to unveil several unresolved issues concerning the molecular basis for the assembly of fungal cellulosomes.

Interestingly, the major product of cellulose digestion by fungal cellulosomes is glucose, whereas in the case of bacterial cellulosomes, cellobiose is the major sugar released (Dijkerman, Op den Camp, Van der Drift, & Vogels, 1997). Despite many advantages, such as synergistic activity between the cellulosome components and an efficient hydrolytic activity on both cellulose and hemicellulose, fungal cellulosomes are much less well characterized compared to bacterial cellulosomes.

1.3.4. Transcriptional Regulation in Cellulolytic Bacteria

In most cellulolytic organisms, cellulase synthesis is repressed in the presence of easily metabolized soluble carbon sources and induced in the presence of cellulose. Induction of cellulases appears to be effected by soluble products generated from cellulose by cellulolytic enzymes synthesized constitutively at a low level. These products are presumably converted into true inducers by transglycosylation reactions (Béguin & Aubert, 1994).

In the simplest cellulosome systems, such as *C. cellulovorans* and *C. cellulolyticum,* the scaffoldin gene is followed downstream by a series of genes that code for dockerin-bearing enzymes. Thus, several of these genes are transcribed as polycistronic mRNA units. In *C. cellulovorans*, the promoter regions of the cellulosomal genes are highly conserved (Han, Yukawa, Inui, & Doi, 2003). Furthermore, it was shown that most of the cellulosomal genes, including the scaffoldin gene *cbpA* are expressed efficiently when cells are grown on high-molecular-weight natural polymers, such as cellulose, xylan and pectin. However, the expression is reduced in cells grown on cellobiose or fructose and reduced even further on carbon sources such as mannose and lactose. It is thus likely that the expression of the cellulosomal genes in this bacterium is coordinated and that some catabolite repression is involved (Han, Cho, Yukawa, Inui, & Doi, 2004).

Concerning complex cellulosome systems, the scaffoldin genes are organized into multiple scaffoldin gene clusters on the bacterial chromosome. Thus, in *C. thermocellum cipA, olpB, orf2* and *olpA* genes are located in tandem on the chromosome, whereas *sdbA* is located

elsewhere on the genome. The proteins OlpB, Orf2p and SdbA serve as anchoring proteins that connect the CipA scaffoldin to the cell surface. In contrast, OlpA selectively binds individual cellulases to the cell surface. The expression of these cellulosomal genes has been studied at the transcriptional level. In this way, it was shown that mRNA levels of scaffoldin genes *olpB, orf2* and *cipA* vary with growth rate. Thus, under carbon limitation mRNA levels measured were high. A similar dependence on growth rate was observed for the major cellulosomal cellobiohydrolase CelS and also for different endoglucanases from *C. thermocellum.* So, the overall cellulase expression is higher when the cells are grown on cellulose versus cellobiose. In contrast, the *sdb*A expression observed was low and not influenced by growth rate (Bayer, Setter, & Lamed, 1985; Bayer *et al.*, 2004). Later, Martin *et al.* (2007) reported an increased expression of the anchor protein OlpB, exoglucanases CelS and CelK, and GH9 endoglucanase CelJ during growth on cellulose, as compared to cellobiose-grown cellulosomes. For the same growth conditions, a lowered expression of endoglucanases from glycoside hydrolase families GH8 (CelA) and GH5 (CelB, CelE, CelG) and hemicellulases (XynA, XynC, XynZ and XghA) was also verified during growth on cellulose (Gold & Martin, 2007). However, most expression or cellulosomal composition studies only investigated growth on two model substrates, crystalline cellulose and cellobiose. Therefore, Raman *et al.* (2009) investigated qualitative and quantitative changes in cellulosome composition of *C. thermocellum* during growth on a wide variety of growth substrates ranging from crystalline cellulose, amorphous cellulose and cellobiose to combinations of cellulose with pectin and xylan. In addition, cellulosomal expression profile was investigated during growth on dilute-acid pretreated switchgrass, a natural biomass substrate for cellulosic ethanol production. Quantitative proteomics was obtained by Liquid Chromatography-Tandem Mass Spectrometry (LC-MS/MS). In this study, 16 new cellulosomal components were identified, and while many of these new subunits were low abundant proteins, two proteins Cthe_0435 (see chapters 3 and 5 of this thesis) and Cthe_0452 (latter named OlpC, a potential anchor protein containing one type I cohesin) appeared to be quite abundant. An increased expression of Cthe_0452 (OlpC) during growth on cellobiose was also observed (Raman *et al.*, 2009). This may be related to a similar pattern in expression observed for Cthe_0435, as the dockerin module of the latter is known to interact specifically with the cohesin on Cthe_0452 (OlpC) (Pinheiro *et al.*, 2009). OlpA, which contains one type I cohesin module, has been suggested to play an intermediary role in the assembly of the cellulosome complex by binding the catalytic units prior to their transfer and assembly on the scaffoldin CipA. Compared to the Cthe_0452 (OlpC) protein, OlpA was less abundant under all growth conditions. Concerning type II cohesins (see chapter 4 of this thesis), apart from Cthe_0735 (with one cohesin domain) which was detected only during growth on cellobiose, all other proteins were detected under all growth conditions. Generally, the abundance of cohesin II containing anchor proteins was inversely

proportional to the number of cohesin modules borne by them. Thus, SdbA (with one cohesin) was shown to be the most abundant and OlpB or Cthe_0736, both containing 7 type II cohesins, were the least abundant, in all growth conditions (Raman *et al.*, 2009). However, this result is in clear contrast with the previous statement that OlpB was the most prominent anchor protein during growth on cellulose (Gold & Martin, 2007). Cthe_0736, a new type II cohesin containing protein (with 7 cohesin II domains) that lacks the surface layer homology domain, was shown to have increased expression during growth on all substrates, except on pretreated biomass (Raman *et al.*, 2009).

As for exoglucanases, CelS from GH48 family is the major component in *C. thermocellum* cellulosome. Quantitative proteomics showed lower expression of all four already known exoglucanases GH48 (CelS), GH9 (CelK, CbhA) and GH5 (CelO) during growth on amorphous cellulose than on cellulose. Additionally, GH9 proteins CelK and CbhA were both expressed at higher levels during growth on pretreated switchgrass, as compared to cellulose. Since these two enzymes attack the cellulose chain from the reducing end of the chain, this increased expression may suggest an enhanced need for exo synergy between both enzymes. Concerning endoglucanases, CelA (GH8) was found to be one of the most abundant proteins, during growth on various substrates. GH9 endoglucanase showed a decreased expression in the absence of crystalline cellulose, whereas GH5 endoglucanase showed decreased expression in the absence of amorphous or crystalline cellulose. This might suggest an important role for the GH9 in the decrystallization of crystalline cellulose. In this way, GH9 was proposed to attack the crystalline surface of cellulose fibrils aiding the creation of amorphous cellulose regions which can be hydrolysed by GH5 endoglucanase (Raman *et al.*, 2009). Xylanases were suggested to play a vital role in exposing the cellulosome preferred substrate, cellulose, in plant cell walls through the degradation of hemicelluloses (Moraïs *et al.*, 2011). Xylanase transcription was shown to be growth rate independent and to increase on cellobiose. Since *C. thermocellum* is unable to utilize the pentose sugars produced by the action of xylanases and other hemicellulases, the apparent role of hemicellulases is to expose cellulose to the action of cellulases. When the organism is not getting energy from cellulose, as when it is grown on cellobiose, in general, it appears to prepare itself to mine cellulose from plant wall materials, hemicellulose and lignin (Gold & Martin, 2007).

1.3.5. The Type I Cohesin-Dockerin Interaction

The cohesin-dockerin interaction is crucial for biomass convertion by anaerobic organisms, because the enzyme complexes synthesized by these organisms are among the most potent hydrolytic enzyme systems known so far. Previous studies showed that the cohesin-dockerin pair represents one of the strongest protein:protein interactions known in nature (Bayer *et al.*,

2004; Carvalho *et al.*, 2003). The first structures of cellulosomal components to be revealed were of the type I cohesins from the scaffoldins of *C. thermocellum* and *C. cellulolyticum*. These 147-residues type I cohesins were shown to form a nine-stranded β-sandwich in an elongated shape, with a β-barrel and jelly-roll topology. The two sheets of the sandwich are composed of strands 8, 3, 6, 5 and 9, 1, 2, 7, 4, respectively, where β-strand 9 (C-terminus) and β–strand 1 (N terminus) run parallel and the remaining strands are all anti-parallel. The nine β-strands are assembled around an extensive aromatic core (Carvalho *et al.*, 2003; Shimon *et al.*, 1997).

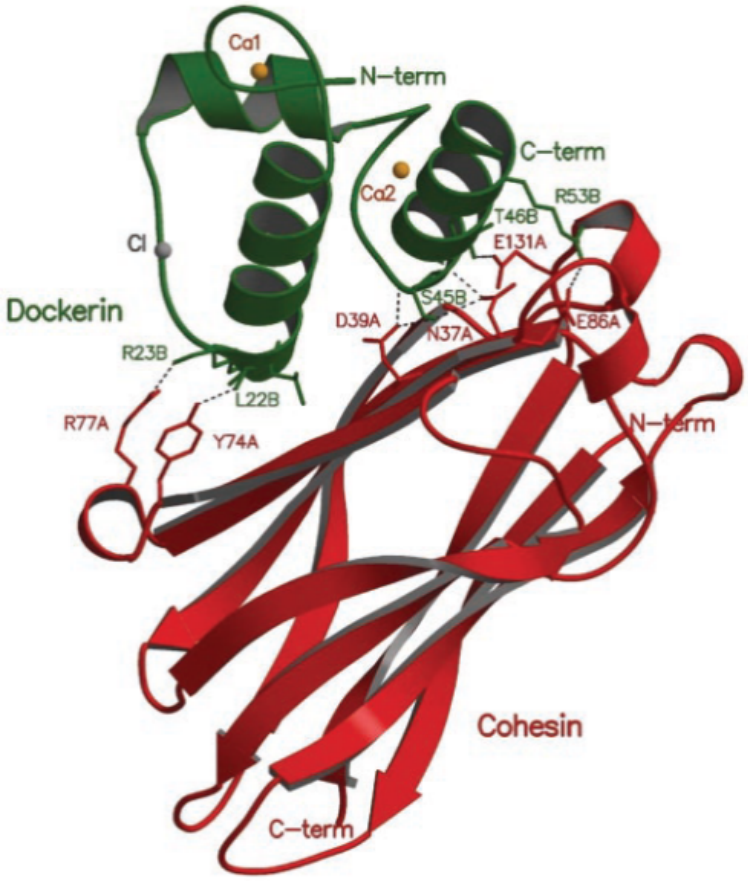
The dockerin sequences consist of about 70 amino acid residues that comprise a 22-residue tandemly repeated sequence, separated by a distinctive short segment of 9-15 amino acids (Bayer *et al.*, 1994). An NMR solution structure of the free *C. thermocellum* type I dockerin module from cellobiohydrolase Cel48S was reported by Lytle *et al.* (2001) and revealed that the first 12 residues of the duplicated sequences bear remarkable resemblance to the calcium-binding loop of the EF-hand motif, in which all the calcium-binding residues like aspartic acids and asparagines, are highly conserved (Lytle, Volkman, Westler, Heckman, & Wu, 2001). In this context, calcium dependence of functional dockerins was demonstrated experimentally (Choi & Ljungdahl, 1996) and both duplicated segments were shown to be involved in cohesin binding (Fierobe *et al.*, 1999). Besides that, the presence of the duplicated segments suggested that the structure of these modules display a twofold symmetry. When the crystal structure of the first type I cohesin-dockerin complex was revealed, this prediction was shown to be right.

The crystal structure of a representative type I dockerin from *C. thermocellum* Xyn10B in complex with the second cohesin module of CipA scaffoldin, obtained by Carvalho *et al.* (2003), provided the first insights into the mechanism by which cellulosome are assembled (Figure 1.10). Thus, it was revealed that the dockerin module contains three α-helices, with helices 1 and 3 comprising the first and the second duplicated segment, respectively. Each duplicated segment displays remarkable structural conservation and also contributes an F-hand calcium-binding motif and so, 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. In the absence of the cohesin, dockerins display a great structural flexibility (Lytle *et al.*, 2001) and are largely disordered in the absence of calcium (Lytle, Volkman, Westler, & Wu, 2000).

In the crystal structure of the type I complex, the cohesin interacts with its dockerin partner primarily along one face of its flattened β-barrel. Although the dockerin presents a remarkable internal symmetry, the detailed crystal structure of the complex also 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 binding. While hydrophobic forces dominate cohesin-dockerin recognition, 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 1.10| Structure of the type I Coh-Doc complex



The protein-protein complex is formed between a cohesin molecule (red) and a Ca^{2+}-bound dockerin (green). The residues involved in domain contacts are shown as stick models. The two Ca^{2+} binding sites of the dockerin domain are represented as orange spheres (Carvalho *et al.*, 2003).

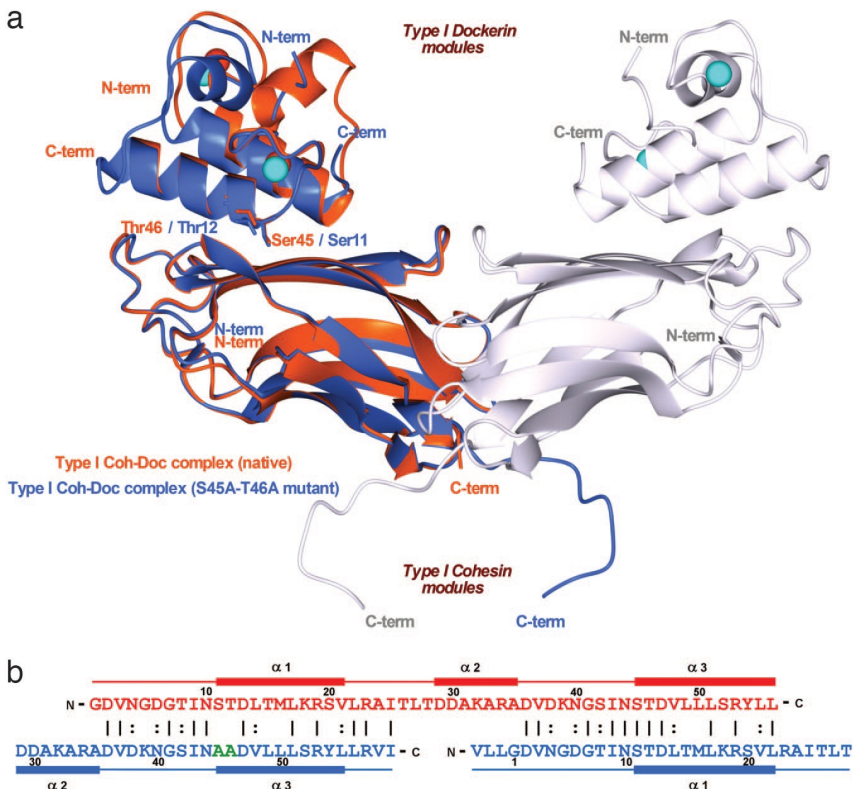
However, Carvalho *et al.* (2007) revealed later that the type I dockerin can bind to its cohesin partner through two distinct surfaces (Figure 1.11). The *C. thermocellum* type I cohesin-dockerin complex was proved to have dual binding mode, because when mutating one of the dockerin's helices, the binding with this helix was disrupted and the reverse binding, which seems to have equal importance, could occur. This second binding mode was indeed observed when the Ser-Thr pair of helix 3 was substituted by two alanines. According to the prediction, the mutated dockerin was rotated by 180º with helix 1 in the position of helix 3, and the Ser-Thr pair in the first duplicated segment dominating the hydrogen bond network. In essence, the equivalent residues in helix 1 of the mutant and helix 3 in the wild-type dockerin interact with the cohesin module and so, an almost perfect overlapping was observed (Carvalho *et al.*, 2007). Additional truncation and mutation experiments were performed in order to confirm whether the cohesin-dockerin interaction is symmetrical. 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 reduced the affinity significantly. These data are then mutually compatible with the crystal structure of type I cohesin-dockerin complex explained above (Karpol, Barak, Lamed, Shoham, & Bayer, 2008).

It could be argued that the two binding modes enable type I dockerins to interact with two cohesin modules simultaneously, providing a possible explanation for the formation of polycellulosomes described in *C. thermocellum.* However, the stoichiometry of the binding of a variety of type I cohesin-dockerin complexes is consistently 1:1, suggesting that the two binding sites are not able to recognize their ligands simultaneously. It seems that the dual binding mode may be responsible for the introduction of quaternary flexibility into the multi-enzyme complex and for the enhancement of the substrate targeting and synergistic interactions between some enzymes, particularly exo- and endo-acting cellulases (Carvalho *et al.*, 2003; Gilbert, 2007). In this way, the dual binding mode may reduce the steric constrains that are likely to be imposed by assembling a large number of different catalytic and noncatalytic domains into a single cellulosome. Quaternary flexibility could be provided by the proline-threonine rich linker sequences that join the dockerins to their catalytic modules. Indeed, probing cellulosome components by small angle X-ray scattering supports the proposal that the intermodule 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 optimize 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). Thus, 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 structure of the complex and reveals that disruption of the cohesin-dockerin complex requires the mutation of the two cohesin binding interfaces simultaneously.

Figure 1.11| The dual binding mode of the 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, A45 and A46, are shown in green. Because of internal 2-fold symmetry of each dockerin module, the two structures overlap almost perfectly in their $\alpha 1/\alpha 3$ regions. (Carvalho *et al.* 2007).

The genome of *C. thermocellum* ATCC 27405 encodes 72 polypeptides containing type I dockerin sequences. Inspection of dockerin sequences at the two ligand binding sites revealed a strong conservation of the amino acids that mediate cohesin recognition, particularly those at positions 11 and 12 that are occupied by hydroxyl amino acids. However, there are at least four dockerins, which are located in proteins Cthe_0435 Cthe_0918, Cthe_0258 and Cel9D-Cel44A (Cthe_0624), in which the usually conserved Ser-Thr pair is replaced in one of the duplicated segments by non-hydroxy residues. Significantly, two of these dockerins, Doc-258 and Doc-435, appear to bind preferentially to the type I cohesin Cthe_0452 (OlpC), rather than to CipA cohesins. Thus, it was suggested that a particular set of enzymes might preferentially bind directly to the bacterium cell surface rather to the CipA scaffoldin. By contrast, the dockerin domains of the Cthe_0624 and Cthe_0918 are primarily cellulosomal, since they bind preferentially to the cohesins of CipA. The

biological significance of *C. thermocellum* targeting a set of enzymes to the cell envelop instead of the CipA scaffoldin remains unknown. It is thought that perhaps the increased activity reflects synergistic interactions between catalytic components of the cellulosome and enzymes directly appended to the surface of the bacterium (Pinheiro *et al.*, 2009).

1.3.6. The Type II Cohesin-Dockerin Interaction

Attachment of cellulosomes to the bacterial cell surface is a key mechanism for the optimal uptake of glucosidic nutrients and consequently for the viability of the microbe. In *C. thermocellum*, the type II dockerin tethers 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 SdbA, OlpB, Orf2p and Cthe_0736 (Fontes & Gilbert, 2010; Leibovitz & Béguin, 1996).

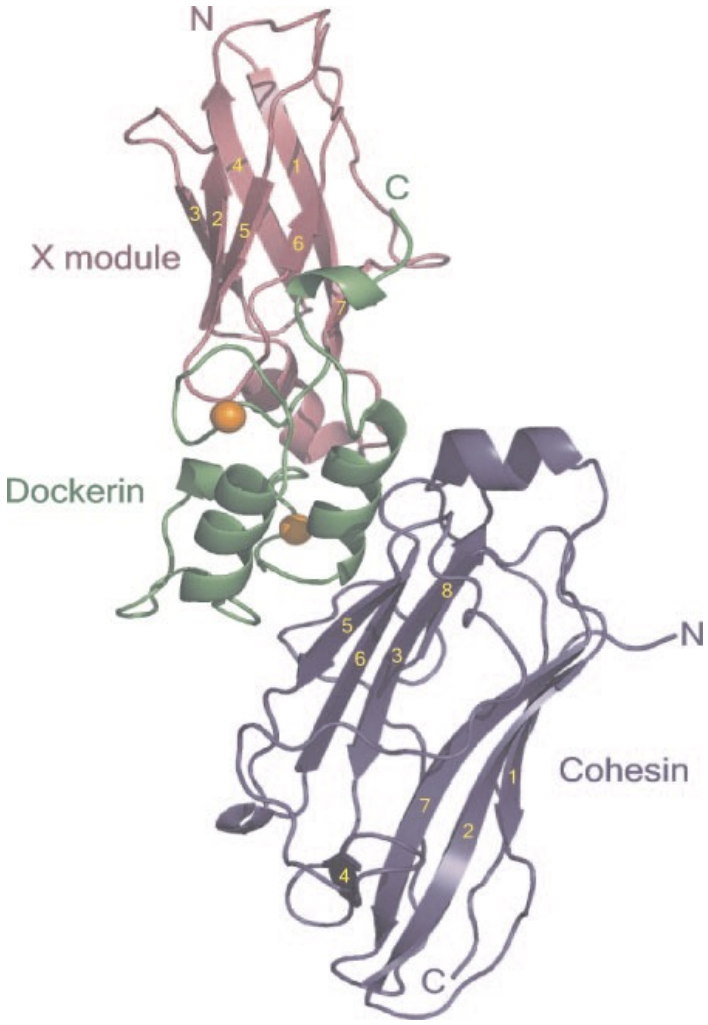
The first type II cohesin crystal structure to be obtained was the type II cohesin of *B. cellulosolvens* of scaffoldin ScaA shortly followed by the structure of the type II cohesin from *C. thermocellum* anchoring protein SdbA. Both structures had the same jelly-rol topology observed in type I cohesins with the exception of the presence of a α-helix, between β-strand 6 and 7 and of two "β-flaps". The sequences of these three secondary elements, as well as the rest of the structural elements, are more conserved among all type II cohesins than among type I cohesins (Carvalho *et al.*, 2005; Noach *et al.*, 2005).

The crystal structure of the *C. thermocellum* heterodimeric SdbA type II cohesin in complex with the type II CipA dockerin was obtained by Smith and colleagues (Figure 1.12). As proved before, the type II cohesin displays a typical jellyroll fold. Data showed that the cohesin does not undergo significant conformational changes upon ligand binding (Adams, Pal, Jia, & Smith, 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 dockerin also interacts with a neighbouring module of unknown function, named X-module, which adopts an immunoglobulin-like fold. Unlike type I dockerins, in which ligand 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 and thus binding is predominantly hydrophobic. There is an extensive hydrogen-bonding network that involves residues from the X module, both dockerin helices and the 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 CipA scaffoldin a rigid and elongated conformation. Besides interacting with the type II dockerin, the CipA 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-Xdockerin interaction in solution. Titration of the Xdockerin 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 that 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 stabilization 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. Thereby, 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 1.12| Structure of the type II Cohesin-Xdockerin complex (CohSdbA-CipAXDoc).



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).

1.3.7. Cohesin-dockerin Specificity

Although structurally related, there is no cross-specificity between type I and type II cohesin-dockerin partners, which allows for the correct assembly and cell-surface attachment of bacterial cellulosomes (Miras *et al.*, 2002; Schaeffer *et al.*, 2002). Concerning type I cohesin-dockerin interactions, it is known that the sequence duplication displayed by type I dockerins from a variety of organisms, beyond *C. thermocellum*, indicates that the dual binding mode may be replicated in other microbial cellulosomes (see Table 1.1). Analysis of the aligned dockerin sequences suggested a correlation with defined Ser-Thr pair (positions 11 and 12) found in both duplicated segments, which appeared to 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 invariant within each species but divergent between them (Pagès *et al.*, 1997). The same interspecies effect also exists between the cohesin-dockerin interaction of *C. thermocellum* and *C. josui*, owing to the intimate relationship and near identity of the component sequences from the latter strain, with those of *C. cellulolyticum* (Jindou *et al.*, 2004). Thus, ligand specificities in type I cohesin-dockerin interactions were shown to vary between species (Mechaly *et al.*, 2001). Mutagenesis studies 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 recognized 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. In this way, the very tight interaction between cohesins and dockerins is known to be species specific, although there is a considerable similarity in sequence and structure among cohesins and dockerins of different species (Bayer *et al.*, 2004). Taken together, these data suggest that cellulosomal enzyme sharing is not an evolutionary driver in organisms depending on plant structural carbohydrates as an energy source. However, one can also argue that microorganisms inhabiting the same ecological niche might, in particular circumstances, benefit from the sharing of cellulosomal enzymes. An example that supports this last argument 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 with what was observed for the type I interaction, in type II cohesin-dockerin complexes there is relatively extensive cross-species plasticity. For example, the type II cohesin of the *C. thermocellum* anchoring scaffoldin SdbA binds not only to the *C. thermocellum* CipA type II dockerin but also to both *B. cellulosolvens* and *A. cellulolyticus* type II dockerins. 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 of the promiscuous type II cohesin-dockerin interaction remains unknown.

Table 1.1| 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

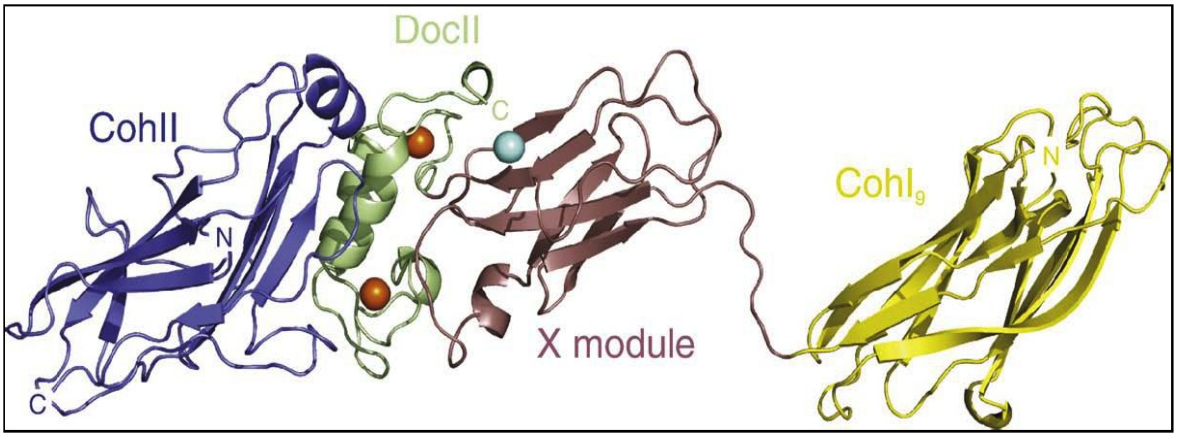
Scaffoldin-borne dockerins are highlighted in grey. [a] Consensus residues represent the dominant amino acids that appear in the designated position from the indicated group of cellulosomal enzymes. Adapted from (Bayer *et al.*, 2004).

1.3.8. Cellulosome Structural Organization

The unique arrangement of the enzymatic subunits in the cellulosome complex, made possible by the scaffoldin subunit, promotes enhanced substrate degradation relative to the enzymes free in solution. Whereas the high-resolution structures of several cellulosomal components have been elucidated, the structural organization of the complete cellulosome remains poorly understood. Nevertheless, substantial efforts have been undertaken to elucidate the tertiary and quaternary structural features of the cellulosome. In this way, electron microscopy initial studies 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 the type I cohesin modules of the scaffoldin allow for optimal positioning of the enzymatic subunits onto the substrate. In

contrast, 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). In addition, 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 CipA scaffoldin has an elongated topology (Adams *et al.*, 2006). Recently, Smith and colleagues determined the X-ray crystal structure of the largest multimodular portion of the *C. thermocellum* cellulosome complex to date, which contains the C-terminal trimodular fragment of the CipA scaffoldin (the ninth cohesin I, connected by a linker to the Xdockerin II) bound to the SdbA type II cohesin (Figure 1.13). The structure reveals an elongated topology with a flexible 13-residue linker connecting the ninth type I cohesin module and the X module, indicating that this region is dynamic. This flexibility could allow the ninth type I cohesin to explore conformational space, including coming into closer proximity with the type II Xdockerin-cohesin region. Furthermore, a dimer interface was observed between CipA and a second, symmetry-related CipA molecule within the crystal structure. This binding is mediated by van der Waals forces and hydrogen-bonding contacts between a type I cohesin and a X module of a symmetry mate and results in two intertwined scaffoldins. Data showed that scaffoldin homodimerization appears to limit the degree of freedom between the different protein modules. Nevertheless, the intrinsic flexibility of the linker between the X module and the type I cohesin in each CipA monomer should accommodate conformational changes of different protein modules (Adams *et al.*, 2010). Taken together, these studies suggest the existence of several possible conformations of the linker sequences when bound to the neighbouring cohesin modules, thus indicating that structural changes in the linker regions may contribute to modulate the overall conformation of the cellulosome.

Figure 1.13| Structure of the *C. thermo*cellum CipA scaffoldin Cohesin I9-X-Dockerin II trimodular fragment in complex with the SdbA Cohesin II module.



The backbone ribbon representation of the complex depicts SdbA cohesin II in blue, dockerin II in green, X module in rose and the ninth cohesin I in yellow. The calcium ions and chloride ion appear as orange and cyan spheres, respectively. The modules are identified and the N and C termini are labelled accordingly (Adams *et al.*, 2010).

More recently, Llorca and colleagues (2011) revealed by cryo-electron microscopy that a large fragment of the cellulosome presents a very compact conformation in solution. Thus, a three-dimensional structure of a *C. thermocellum* mini-cellulosome that comprises three consecutive cohesin CipA modules (third, fourth and fifth) bound to three Cel8A cellulases, through their native dockerin modules, was solved (García-Alvarez *et al.*, 2011). Unlike to what was observed by Hammel, *et al.* (2005), the structure revealed that the linker regions between cohesin modules showed a restricted flexibility. This compact conformation is thought to be a result of the stabilization of specific contacts between cohesin modules by the linkers. Additionally, the cellulosome revealed an antiparallel disposition of the catalytic cores of Cel8A, which alternately project the enzymes into opposite directions from the third, fourth and fifth cohesin modules. Whereas a parallel conformation would restrict the access of the enzymes to their substrate to one side of the scaffoldin protein, the antiparallel arrangement could facilitate the accessibility of the catalytic domains for the substrate in a larger range of orientations (García-Alvarez *et al.*, 2011).

Future work is required in order to obtain novel insights into the higher-order scaffoldin interactions which are behind cellulosome modular architecture and polycellulosome formation. Understanding the intricacy of cellulosomes evolved by anaerobic microbes might sustain the development of an effective process for the conversion of lignocellulosic biomass to bio-ethanol.

1.3.9. Cellulosomal Enzymes

1.3.9.1. Glycoside hydrolases

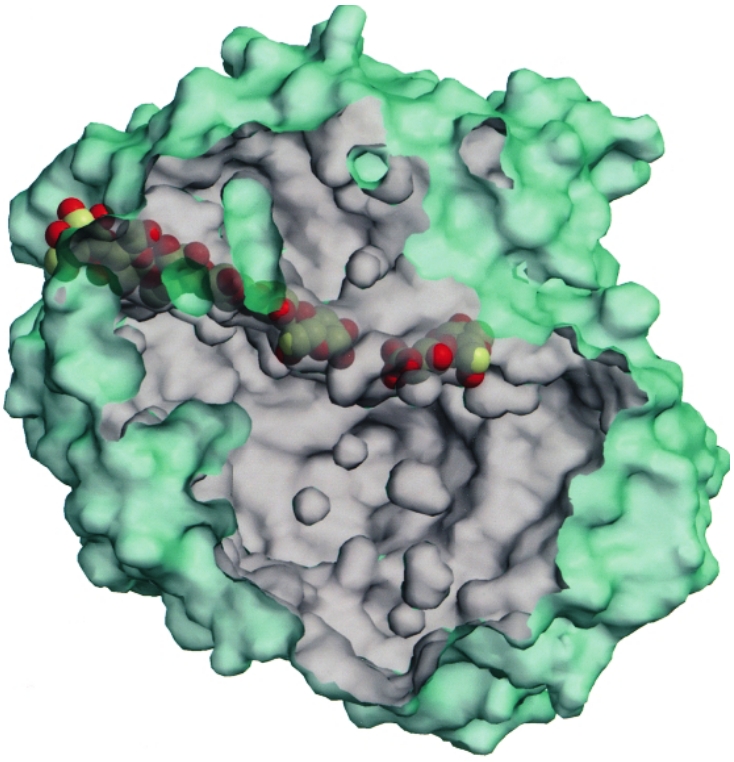
Glycoside hydrolases (GHs) are key enzymes in the hydrolysis of plant structural carbohydrates. These proteins attack β-glycosidic or α-glycosidic bonds and can be classified as retaining or inverting enzymes. Retaining enzymes catalyse either transglycosylation or hydrolysis reactions with retention of configuration at the anomeric center. In contrast, inverting enzymes do not catalyse transglycosylation reactions, only hydrolysis with the inversion of configuration at the anomeric center (McCarter & Withers, 1994).

Glycoside hydrolases acting on a particular substrate also differ in the products they release from a specific carbohydrate. Exo-acting enzymes remove units of one or more sugars from the ends of polysaccharide chains. Endo-acting enzymes hydrolyse random bonds within the chains, thereby producing more ends for the exoenzymes to act on (Warren, 1996). For long, linear substrates, such as cellulose, the ends are a limiting factor, so in order to have an efficient hydrolysis of the substrate, three major classes of enzymes are generally required to work in synergy: exo-β−1,4-glucanases, which release cellobiose; endo-β−1,4-glucanases, which degrade regions of amorphous cellulose and finally, β−glucosidases, which degrade

short oligosaccharides such as cellobiose and cellotriose to glucose (Wood & Ingram, 1992). As such, endo-β−1,4-glucanases create new ends from which exo-β−1,4-glucanases can release cellobiose from either the reducing or nonreducing end of the cellulose chains (Gilbert, 2010). Moreover, exocellulases have a processive mode of action which gives them the ability to move from amorphous regions into the crystalline structures of cellulose. Nevertheless, this model was shown to be inconsistent with other cellulase degradative systems in which exocellulases were shown to also display endo activity (Varrot, Schülein, & Davies, 1999).

The crystal structure of the family 48 exocelulase (Cel48A) from *C. thermocellum*, which is the most abundant cellulosomal enzyme in this bacterium, revealed an $(\alpha/\alpha)_6$ barrel fold with a tunnel-shaped substrate-binding region (Figure 1.14). The enzyme attacks the reducing end of cellulose chains. The polymer threads through the tunnel and the glycosidic bond between the second and the third glucose residue is cleaved, releasing cellobiose. In contrast, cellulase 8A, an endocellulase, contains a groove-shaped substrate-binding region with an open cleft, explaining why this enzyme can cleave the internal regions of the polysaccharide (Fontes & Gilbert, 2010; Guimarães, Souchon, Lytle, Wu, & Alzari, 2002).

Figure 1.14| Overall view of the structure of the cellobiohydrolase CelS from family 48.



Cut-through of the molecular surface of CelS the major enzymatic component of the *C. thermocellum* cellulosome. The structure shows the substrate binding tunnel with a bound cellohexaose molecule and the open cleft with a bound cellobiose molecule (Guimarães *et al.*, 2002).

A classification of glycoside hydrolases in families, based on amino acid sequence similarities, has been proposed by Henrissat *et al.* (1998). Because there is a direct relationship between sequences and folding similarities, this classification reflects the structural features of these enzymes better than their sole substrate specificity. In addition, it helps to reveal the evolutionary relationships between CAZYmes while providing a convenient tool to derive mechanistic information, thus illustrating the difficulty of deriving relationships between family membership and substrate specificity. The CAZy database provides a continuously updated list of the glycoside hydrolase families. Because the fold of proteins is better conserved than their sequences, some of the families can be grouped in *clans* when new sequences are found to be related to more than one family, when the sensitivity of sequence comparison methods is increased or when structural determinations demonstrate the resemblance between members of different families (Henrissat, 1998). Particularly, various folds have been observed in different cellulase families: $(\beta/\alpha)_8$-barrel in families GH5, GH51 and GH44; distorted (α/β)-barrel in family GH6; β-jelly roll fold in families GH7 and GH12; $(\alpha/\alpha)_6$-barrel represented by families GH8, GH9 and GH48; β_6-barrel in family GH45 and finally, sevenfold β-propeller in family GH74.
According to CAZy, the GHs catalytic modules are currently classified into 130 different families based on amino acid sequence similarities (April 2012).

1.3.9.2. Polysaccharide lyases

Polysaccharide lyases (PLs) are a group of enzymes that 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 PLs have been invented more than once during evolution from totally different scaffolds (Lombard *et al.*, 2010). They are presently found in 22 families in CAZy (April 2012).

1.3.9.3. Carbohydrate esterases

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).

1.3.10. Linker regions and Non-catalytic Modules

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 connection between cohesin modules within the scaffoldin unit and also to connect the dockerins with the catalytic subunits. The exact role of the cellulosomal linkers has yet to be described, although, as described above, it is assumed that they contribute to the architecture and action of the cellulosome. Thus, linkers supply the protein subunits with flexibility and provide spacers between the enzymatic modules that could hypothetically enhance interactions with the substrates (Noach *et al.*, 2009). Generally linker regions are rich in amino acids serine and threonine (Coutinho & Reilly, 1994) and may be glycosylated in the original organisms, which confer protection against proteolysis (Tomme, Warren, & Gilkes, 1995).

Other modules, such as fibronectin-like or imunoglobuline-like sequences with unknown function have been identified. Previous studies of fibronectin type III-like modules, which were identified in *C. thermocellum* cellulosome, have indicated that they might function as ligand-binding modules, as a compact form of peptide linkers or spacers between other domains, as cellulose-disrupting modules or even as proteins that help large enzyme complexes remain soluble (Alahuhta *et al.*, 2010).

Nevertheless, most of the characterized non-catalytic binding modules are CBMs. As explained below, their main function is to direct the appended catalytic regions to their target substrates.

1.3.11. Carbohydrate Binding Modules (CBMs)

CBMs are non-catalytic domains most commonly representing parts of modular glycoside hydrolysing or modifying enzymes that recognise different carbohydrates.

Initially, these non-catalytic polysaccharide-recognizing modules were defined as Cellulose-binding domains (CBDs) since the first examples of these protein domains bound to crystalline cellulose as their primary ligand (Gilkes, Warren, Miller, & Kilburn, 1988). Later, in order to reflect the diverse ligand specificity of these modules, a more inclusive term, CBM, was assigned.

Hence, CBMs were defined as a contiguous amino acid sequence, within a carbohydrate-active enzyme, with autonomous folding and skilled recognition for a specific carbohydrate motif. A few exceptions are CBMs in cellulosomal scaffoldin proteins, which are responsible for the primary targeting of the entire cellulosome to the crystalline cellulose and also CBMs

that are found isolated, as a single protein. CBMs which are part of scaffoldin subunits are usually from family 3. An example of this is the CBM3 of *C. thermocellum* scaffoldin CipA. This type of CBM binds strongly to the crystalline cellulose surface and accounts for the primary targeting of the cellulosome to its substrate (Bayer *et al.*, 2004). In addition, an example of a free CBM is the small olive pollen protein, Ole e 10, which has the ability to bind several polysaccharides (Barral *et al.*, 2005). The biological role of this CBM remains, however, elusive.

The requirement of CBMs within large modular enzymes sets this class of carbohydrate-binding protein apart from other non-catalytic sugar binding proteins such as lectins and sugar transport proteins. Lectins, which occur ubiquitously in nature, are homologous to CBMs in the sense that they also bind carbohydrates. They may bind to a soluble carbohydrate or to a carbohydrate moiety that is a part of a glycoprotein or glycolipid. However, these proteins fulfil an important role in the immune system that has been extensively studied over the last years (McGreal, Martinez-Pomares, & Gordon, 2004).

1.3.11.1. CBMs functions

Invariably, the main role of CBMs is to recognize and bind specifically to carbohydrates. The biological consequence of this event results in different functions, such as enhanced hydrolysis of insoluble substrates as a consequence of a closer proximity between the catalytic domain with the substrate, polysaccharide structure disruption and cell surface protein anchoring (Boraston, Bolam, Gilbert, & Davies, 2004; Guillén, Sánchez, & Rodríguez-Sanoja, 2010). The hydrolysis of insoluble polysaccharides, such as cellulose, requires initially that glycoside hydrolases which are responsible for substrate hydrolysis approach and anchor to the polysaccharide chains. The CBM function in these GHs is to increase the effective enzyme concentration on the polysaccharide surface and consequently to enhance enzymatic activity (Bolam *et al.*, 1998). Previous studies proved that maintaining enzymes in proximity with the insoluble substrates leads to a more rapid degradation of polysaccharides. In this way, it is clearly established that 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. For instance, CBM22 was shown to change the specificity of a glycoside hydrolase family 10 xylanase such that it displayed primarily β-1,4-β-1,3-glucanase activity (Araki *et al.*, 2004). Thus, CBMs not only recognize and attach the enzymes to their substrate but can also regulate their catalytic properties.

Previous studies defend that CBMs may also have a disruption function. In some cases, binding of CBMs to a crystalline substrate leads to polysaccharide disorganization and improvement of substrate availability. Gao *et al.* (2001) suggested that the binding of the cellulose-binding domains (CBD) to cotton fibers leads to structural changes and release of short fibers (Gao, Chen, Wang, Zhang, & Liu, 2001). Later, the same group also found that the attachment of a CBM to cotton fibers promotes severe weakening of the cellulose-interchain hydrogen bonds (Wang, Zhang, & Gao, 2008). Therefore, an enzyme located next to CBMs, which possess disruption functions may have facilitated hydrolysis of recalcitrant substrates, giving it an advantage over other enzymes. However, this disruptive function does not seems to be common in CBMs (Guillén *et al.*, 2010).

For efficient polysaccharide hydrolysis, there is a need of a dynamic interaction between CBMs and their substrates, where the catalytic domain is first positioned in proximity to substrate through the CBM. Then, the catalytic domain is able to hydrolyze the polysaccharide chains inserted in the active site. Besides that, CBMs can also be relocated to new regions on the ligand allowing a continuous hydrolysis of the substrate (Guillén *et al.*, 2010).

Generally, CBMs can be found in proteins that recognize polysaccharides such as cellulose, chitin, β-glucan, starch, glycogen, pullulan, xylan and many other different sugars such as mannan, fucose, lactose, galactose, polygalacturonic acid, among others (Tomme *et al.*, 1998).

1.3.11.2. CBMs Classification and Nomenclature

Similar to the catalytic modules of glycoside hydrolases, CBMs are divided into families on CAZy database. The family classification of CBMs is expected to aid in the identification of novel CBMs. In some cases, the family classification will allow predicting the binding specificity while aiding in identifying functional residues and revealing evolutionary relationships. As it was described for GHs, domain classification in a specific CBM family is also predictive of a specific polypeptide fold. In the last update of the CAZy database 25243 CBMs were grouped into 64 families (April 2012). A CBM is named by its family, e.g. the family 6 CBM from *C. thermocellum* XynZ would be called CBM6, but one may also include the organism and even the enzyme from which it is derived. Thus, for this example, the CBM would be defined as *Ct*CBM6 or *CtXynZ*CBM6. If glycoside hydrolases contain tandem CBMs belonging to the same family, a number corresponding to the position of the CBM in the enzyme relative to the N-terminus is included (Boraston *et al.*, 2004).

Because the fold of proteins is better conserved than their sequences, some of the CBM families can be grouped into superfamilies or clans by using the criteria of conservation of the protein fold, catalytic machinery and also mechanism of glycoside bond cleavage. CBM families were then classified into seven structural family folds (β-sandwich, β-trefoil, Cysteine knot, Unique, OB fold, Hevein fold and also Hevein-like fold). Thus, common folds are

observed in proteins with different specificities (Guillén *et al.*, 2010). The dominant fold among CBMs is the β-sandwich followed by the β-trefoil. The β-sandwich fold comprises two β-sheets, each consisting of three to six antiparallel β-strands. An example of a β-sandwich conformation is the family 11 CBM from *C. thermocellum* (Carvalho *et al.,* 2004). The β-trefoil fold contains 12 β-sheet, forming six hairpin turns. A β-barrel structure is formed by six of the strands, attendant with three hairpin turns. The other three hairpin turns form a triangular cap on one end of the β-barrel called the “hairpin triplet”. The subunit of this fold, named a trefoil domain, is a contiguous amino acid sequence with a fourth β-strand, two-hairpin structures having a trefoil shape. Each trefoil domain contributes one hairpin to the β-barrel and one hairpin to the hairpin triplet. In this way, each of the three trefoil subdomains comprises a carbohydrate-binding site. *C. thermocellum* family 42 CBM is an example of a β-trefoil fold in the CBM families (Ribeiro *et al.*, 2010).

In terms of function, CBMs have been divided into three general categories: type-A CBMs, which bind strongly to insoluble polysaccharide surfaces; type-B CBMs, which bind to soluble glycan chains and type-C CBMs that bind to small saccharides (Figure 1.15).

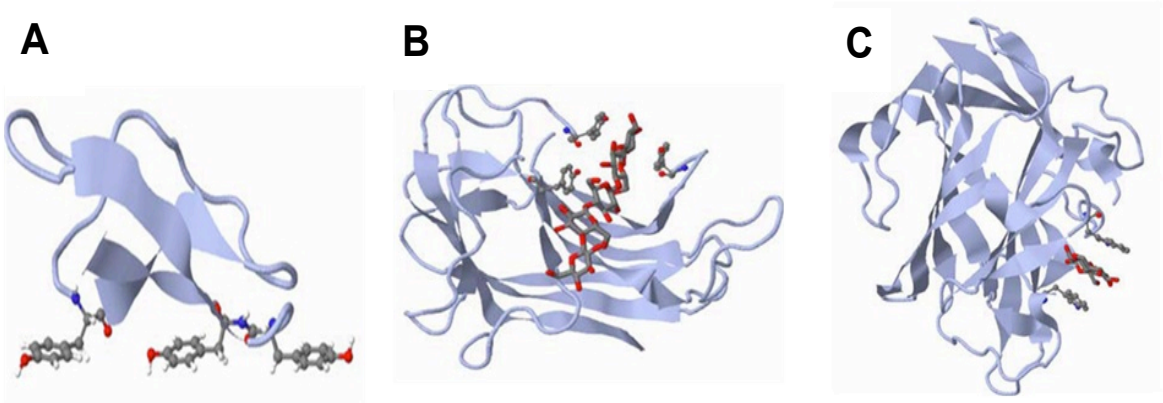
Type A CBMs include members of CBM families 1, 2a, 3, 5 and 10 that bind to insoluble, highly crystalline polysaccharides, such as cellulose, chitin or mannan. Thus, type A CBMs show little or no affinity for soluble carbohydrates providing a distinctive property when compared with the other CBM types. These CBMs have a flat or platform-like hydrophobic surface composed of aromatic residues that recognizes the carbohydrate ligand. Thus, the planar conformation of the type A binding site reflects the architecture of the crystalline polysaccharides that has a flat surface. Hydrogen bonds have little effect in ligand recognition which is dominated by stacking interactions. Additionally, the interaction of type A CBMs is associated with positive entropy, demonstrating that the thermodynamic forces that drive the binding of CBMs to crystalline ligands are relatively unique among carbohydrate binding proteins (Boraston *et al.*, 2004; Guillén *et al.*, 2010).

Type B CBMs bind amorphous cellulose or soluble complex carbohydrates such as xylan or xyloglucan, for example. These CBMs allocate the carbohydrate chain in a distinctive cleft in which aromatic residues interact with the single polysaccharide chain. Aromatic side chains form twisted or sandwich platforms. The orientation of these amino acids was shown to be a key determinant of ligand specificity. Biochemical studies revealed that the binding capacity of these CBMs is determined by the degree of polymerization of the carbohydrate ligand. Thus, the affinity was shown to be higher for hexasaccharides and much lower for oligosaccharides with a degree of polymerization of three or less. Therefore, type B CBMs are usually described as “chain binders”. Furthermore, type B CBMs comprise several sub-sites that are able to accommodate the individual sugar units of the polymeric ligand. Among others, CBMs from families 2b, 4, 6, 15, 17, 20, 22, 27, 28, 29, 34 and 36 are included in this type B group and in general, these proteins have evolved binding site topographies that are

able to interact with individual glycan chains rather than crystalline surfaces. In contrast with what was observed in type A CBMs, direct hydrogen bonds play a key role in defining the affinity and ligand specificity in type B CBMs (Boraston *et al.*, 2004; Guillén *et al.*, 2010; Hashimoto, 2006).

Finally, type C CBMs, also known as lectin-like CBMs, only bind mono-, di-, or trisaccharides due to steric restriction in their binding site. Thus, type C CBMs lacks the extended binding-site grooves of type B CBMs. Although some type B and type C CBMs have a very similar fold, it is apparent that the hydrogen-bonding network between protein and ligand is more extensive in type C CBMs than type B CBMs. Type C CBMs include examples from families 9, 13, 14, 18 and 32, among others. Particularly, CBMs from families 13 and 32 appear to be more prevalent in bacterial toxins or enzymes that attack eukaryotic cell surfaces or matrix glycans (Boraston *et al.*, 2004).

Figure 1.15| Structures of the three different CBM types based on topology of carbohydrate binding site.



A) Type A CBM - CBM1 from *Trichoderma reesei* cellobiohydrolase I (PDB code 1CBH); **B)** Type B CBM - CBM4 from *Cellulomonas fimi* endo-1,4-glucanase C (PDB code 1GU3) and **C)** Type C CBM - CBM9 from *Thermotoga maritime* xylanase 10A (PDB code 1I82). Adapted from Guillén *et al.*(2010).

1.3.11.3. CBMs and Multivalency

Carbohydrate binding proteins can also be classified into two general groups based on their affinity for carbohydrates and their modes of carbohydrate recognition. In this way, group I comprises proteins which bind carbohydrates tightly ($Ka>10^6$ M^{-1}) in binding sites that completely enclose the carbohydrate ligand. In contrast, group II comprises proteins that bind carbohydrates more weakly ($Ka<10^6$ M^{-1}), in open binding sites that leave significant portions of the carbohydrate ligand exposed to solvent when bound (Quiocho, 1986). All CBM-carbohydrate interactions are included in group II. Proteins from this group appear well suited to bind cell-surface glycans, oligosaccharides or polysaccharides that cannot be completely enclosed in the binding site. Additionally, these weak relations are often compensated by multiple clustered carbohydrate-binding sites that can result from a single protein having multiple binding sites or from the association of two or more univalent carbohydrate-binding

proteins into multivalent quaternary structures (in random or in tandem). Interestingly, the appearance of multiple CBMs seems to occur more often in thermo- or hyperthermophilic enzymes. This may allow overcoming the loss of binding affinity that accompanies most molecular interactions at elevated temperatures (Boraston *et al.*, 2004). On the one hand, the carbohydrate-binding proteins from group II may have evolved to have weak binding because this may be advantageous for the function of these proteins. On the other hand, the weak binding may be a result of restrictions on the number of direct interactions between the protein and the sugar (Boraston *et al.*, 2004).

1.3.11.4. Ligand Binding Specificity

Aromatic amino acids, especially tryptophan and tyrosine, form stacking interactions with the sugar rings resulting in strong van der Waals interactions that stabilize the structure. In addition, the side chain of other polar amino acid residues may form hydrogen bonds with the sugar ligand, helping to stabilize the interaction. Different studies suggest that the orientation of the aromatic side-chains is responsible for the different ligand specificities of the CBM families, since it defines the topology of their ligand binding platforms. Subtle changes in the topology of the binding sites dictate ligand specificity and explain why CBMs with apparent similar structure recognize different ligands. Thus, CBMs appear to have carbohydrate-recognition sites which mirror the solution conformations of their target ligands and consequently minimize the energetic penalty paid upon binding (Boraston *et al.*, 2004; Guillén *et al.*, 2010). Besides aromatic residues orientation and positioning in the binding sites of CBMs, other interactions, such as, direct hydrogen bonds and calcium-mediated co-ordination also play a significant role in CBM ligand recognition. Concerning hydrogen bonds, it is known that their relative importance varies depending on the CBM type. As explained above, in type A CBMs hydrogen bonds play only a minor role in ligand recognition (McLean *et al.*, 2000). However, in type B and type C CBMs, mutagenesis studies in which hydrogen-bond residues were replaced with alanine, lead to a significant decrease in affinity or even to a complete knock-out of the binding (Pell *et al.*, 2003; Xie *et al.*, 2001). Although many CBMs are metalloproteins, the role of metal ions, such as calcium, in CBM-ligand interactions has only recently been described. Thus, there are several studies that have revealed that ligand recognition can be calcium-dependent (Bolam *et al.*, 2004; Jamal-Talabani *et al.*, 2004).

In conclusion, CBMs play a pivotal role in degradative enzymes that mediate the recycling of photosynthetically fixed carbon in the biosphere. Thus, understanding CBM mechanisms of ligand recognition will provide novel insights for the development of new carbohydrate-binding technologies and to manipulate carbohydrate-ligand interactions (Boraston *et al.*, 2004; Guillén *et al.*, 2010).

1.3.12. Cellulolytic Machinery: affinity and bioenergy applications

1.3.12.1. Potential applications of CBMs and cohesin-dockerin complexes

Cellulosomal architecture provides a biological model to design enzymatic complexes that synergistically combine multiple catalytic subunits in order to achieve higher specific activities than would be obtained using free enzymes. In this way, multimeric enzymatic complexes may have industrial applications of relevance for an emerging carbon economy. In addition, the use of designer cellulosomes in a broad spectrum of unconventional applications in research, medicine and industry have also been suggested in the last years.

Thus, chimaeric proteins incorporating affinity domains from the cellulosome have potential diverse applications including protein purification, microarray technology and cell substrate. The biological and physicochemical properties of membrane bound proteins can be studied by tagging the domain of interest with a cellulose CBM for immobilization on cellulosic substrates (Nordon, Craig, & Foong, 2009). For example, Nahálka *et al* (2006) developed a thermally reversible, enzyme-binding system suitable for regenerating batch enzymatic processes. The CBM from *C. cellulovorans* was fused with thermophilic enzymes from *Pyrococcus furiosus*. Enzyme was active and free in the reaction mixture at 80-90ºC and deactivated and immobilized by affinity adsorption on cellulose at 30-40ºC (Nahálka & Gemeiner, 2006).

The potential of employing the cohesin-dockerin interaction for use in affinity chromatography has attracted attention of many groups. Karpol *et al.* (2009) have 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 include the CBM and the adjacent cohesin, such that the cohesin bound to a mutated dockerin and its host protein, and the CBM bound the cellulose column. However, even though the dockerin was mutated, it still retained high levels of affinity for its complementary cohesin, yet enabled complete dissociation of the dockerin from the CBM-cohesin affinity column (Karpol *et al.*, 2009). Later, Bayer *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 to be very efficient and robust, the affinity tag was 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, Karpol, Barak, Lamed, & Bayer, 2010).

1.3.12.2. Bioenergy prodution from lignocellulosic materials

As mentioned before, lignocellulose is the most abundant renewable natural resource and substrate available for conversion to fuels. On a worldwide basis, terrestrial plants produce

1.3×10^{10} metric tons of wood per year, which is equivalent to 7×10^{9} metric tons of coal or about two-thirds of the world's energy requirement. Furthermore, available cellulosic feedstocks from agriculture and other sources are about 180 million tons per year. Moreover, tremendous amounts of cellulose are available as municipal and industrial wastes. Thus, there is great interest in the use of cellulosic biomass as a renewable source of energy via breakdown of carbohydrates that can be converted to sugars and then fuel, namely bioethanol. Consequently, microorganisms that metabolize cellulose have gained prominence in recent years (Demain *et al.*, 2005).

Lignocellulose is difficult to hydrolase not only because it is associated with hemicellulose and lignin, but also because much of it has a tightly packed crystalline structure. Neverthless, the bioconversion of lignocellulosic residues to valuable materials by fungi, such as *Trichoderma reesei* and *Saccharomyces cerevisae,* has been developed recently. This process requires four steps which include pretreatment, de-polymerization (saccharification) of cellulose and hemicellulose to soluble monomer sugars (hexoses and pentoses) by hydrolysis (especially using *T. reesei* enzymes), conversion of the sugars to ethanol in a fermentation process mediated by yeast (particularly *S. cerevisaei*) and finally separation and purification of the products. Hydrolysis and fermentation steps can be performed separately or simultaneously. There are advantages and disadvantages for both options, but it seems that, when both processes are performed simultaneously, there is no production of enzyme-inhibiting end-products (cellobiose and glucose) during hydrolysis, which avoids the costly addition of β-glucosidase. However, further work is required in order to not only optimize each step but also to minimize all bioconversion steps into one step in a single reactor using one or more microorganisms and reducing the production costs (Dashtban, Schraft, & Qin, 2009).

It was probably because all eyes were turned to the *Trichoderma-Saccharomyces* concept that the industrial potential of cellulolytic, thermophilic anaerobic bacteria, such as *C. thermocellum*, was not properly taken into account. *C. thermocellum* breaks down cellulose, with the formation of cellobiose and cellodextrins as main products. Cellobiose can be further utilized by bacteria and the final products are ethanol, acetic acid, lactic acid, hydrogen and carbon dioxide (Lamed & Zeikus, 1980). Besides having an efficient cellulase system, *C. thermocellum* anaerobiosis is an advantage because one of the most expensive steps in industrial fermentations is that of providing adequate oxygen transfer for cellulase production. Additionally, growth at a high temperature facilitates the recovery of ethanol (Lynd, Cushman, Nichols, & Wyman, 1991). The cost of the cellulase production in the *Trichoderma-Saccharomyces* process is very high, whereas in a direct clostridial coculture process the enzyme cost would be rather limited because biocatalysts would be produced by the fermenting organism in the course of ethanol production. A coculture consisting of *C. thermocellum* and *Clostridium thermosaccharolyticum* was shown to have great potential. On

the one hand, *C. thermocellum* serves as a cellulase and hemicellulase producer. On the other hand, the hemicellulose-derived pentoses can be utilized by *C. thermosaccharolyticum* but not by *C. thermocellum.* In addition*, C. thermosaccharolyticum* uses cellobiose faster and is a better ethanol producer. The thermophilic ethanol fermentation is a single process step which consists of four biologically mediated events: cellulase and hemicellulase production, cellulose and hemicellulose hydrolysis and hexose and pentose fermentation. One of the main problems of this dual-culture process is the production of side products like acetate and lactate. These products decrease the yield of ethanol and can slow cell growth. To overcome this problem, metabolic engineering research (involving the elimination of metabolic branches in both members of the coculture) is ongoing (Demain *et al.*, 2005).

1.3.12.3. 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 utilization, 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 functioning to produce more efficient biocatalysts for feed supplementation.

In the last years, several groups developed different chimaeric cellulosomes to target a multitude of potential applications (Nordon *et al.*, 2009). For instance, Mingardon *et al.* (2007) incorporated a family 6 fungal cellulase from *Neocallimastix patriciarum* into bacterial minicelllosomes derived from *C. cellulolyticum* and the enzyme complex revealed a up to 26-fold increase in activity over the free enzymes (Mingardon *et al.*, 2007). In order to enhance enzymatic degradation of xylan from wheat straw and to study the synergistic action among different xylanases, Bayer *et al.* (2011) incorporated the entire xylanolytic system of the bacterium *Thermobifida fusca* into defined artificial cellulosome complexes. Data demonstrated that xylanolytic designer cellulosomes displayed enhanced synergistic activities on a natural recalcitrant wheat substrate and could thus serve in the development of advanced systems for improved degradation of lignocellulosic material (Moraïs *et al.*, 2011). Chen *et al.* (2011) engineered a yeast consortium displaying a functional minicellulosome. The basic design of the consortium consisted of four different engineered yeast strains capable of either displaying a trifunctional scaffoldin, carrying three divergent cohesin domains from *C. thermocellum*, *C. cellulolyticum*, and *R. flavefaciens*, or secreting one of the three corresponding dockerin-tagged enzymes: an endoglucanase (from *C. thermocellum* with its native dockerin), an exocellulase (from *T. reesei* fused with a dockerin from *C. cellulolyticum*) and a β-glucosidase (from *Thermoascus aurantiacus* fused with a

dockerin from *R. flavefaciens)*. The secreted cellulases were docked onto the displayed trifunctional scaffoldin in a highly organized manner based on the specific interaction of the three cohesin-dockerin pairs employed and resulting in the assembly of a functional minicellulosome on the yeast surface. The resulting consortium was demonstrated to utilize Phosphoric Acid Swollen Cellulose (PASC) for growth and also for ethanol production (Goyal *et al.*, 2011; Tsai, Goyal, & Chen, 2010).

The conversion of plant biomass to biofuels or fine chemicals is a complex problem that requires a multidisciplinary approach to be properly addressed. Thus, one can anticipate that in order to take advantage of the cellulosome hydrolytic potential, more than one technology and process should be combined to achieve environmental and economic goals. It should be noted that the cellulosome technology should not be restricted to the optimization of plant carbohydrate hydrolysis but rather extended to all biological systems that might benefit from enzymatic proximity.

1.4. Objectives

The work presented here aims to elucidate several unresolved questions concerning the structure, function and importance of novel Cohesin-Dockerin complexes, Glycoside-hydrolases and Carbohydrate Binding Modules from *C. thermocellum* cellulosome. Specifically, the main goals of this project can be summarized in the following points:

- To develop novel methodologies for the production, purification and structure determination of cohesin-dockerin complexes.

- To determine the structural determinants of specificity of novel type I cohesin-dockerin complexes from *C. thermocellum* cellulosome.

- To evaluate the molecular modulators of specificity in type II cohesin-dockerin pairs and so, to extend our knowledge on the mechanisms of cellulosome cell surface attachment in *C. thermocellum*.

- Modules of unknown function appended to dockerins may comprehend important cellulosomal enzymes. Thus, we aim to explore the biochemical properties and the crystal structure of novel cellulosomal enzymes from *C. thermocellum*.

- To solve the crystal structure of a penta-modular cellulosomal protein, providing important clues about the functional importance of modularity in cellulosomes.

1.5. Thesis Outline

Taking in mind the objectives stated above and with the purpose of presenting and discussing the results obtained with clarity, this thesis was divided into six chapters. The first chapter comprises a state of the art review. Several concepts concerning plant cell wall composition and degradation, the complexity and functionality of different cellulosome components, particularly cohesin-dockerin complexes, catalytic and non-catalytic components are revised. In addition, a general description of the current biotechnological applications of cellulosomes is performed. Chapters 2, 3, 5 and 6 are organized in papers based on scientific manuscripts, already published or submitted to international peer reviewed journals. Chapter 4 is also based on a scientific manuscript which is currently in preparation.

Finally, the last chapter discusses and integrates the results presented in each of the previous chapters. Future perspectives for the scientific knowledge attained with this work are also approached.

2. METHODS FOR THE PRODUCTION, EXPRESSION AND STRUCTURE DETERMINATION OF BACTERIAL COHESIN-DOCKERIN COMPLEXES.

2.1. *Escherichia coli* Expression, Purification, Crystallization and Structure determination of Bacterial Cohesin-Dockerin Complexes∞.

Joana L. A. Brás,* Ana. Luísa Carvalho,† Aldino Viegas,† Shabir Nadjmudim,* Victor D. Alves,* José A.M. Prates*, Luís M. A. Ferreira,* Maria J. Romão,† Harry J. Gilbert,‡ and Carlos M. G. A. Fontes*

* CIISA-Faculdade de Medicina Veterinária, Pólo Universitário do Alto da Ajuda, Avenida da Universidade Técnica, 1300-477 Lisboa, Portugal

† REQUIMTE/CQFB, Departamento de Química, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, 2829-516 Caparica, Portugal

‡ Institute for Cell and Molecular Biosciences, Newcastle University, The Medical School, Newcastle upon Tyne NE2 4HH, United Kingdom.

Adapted from: Brás *et al.* (2012) Methods in Enzymology, Volume 510; Chapter 21; 395-415

Abstract

Cellulosomes are highly efficient nanomachines that play a fundamental role during the anaerobic deconstruction of plant cell wall complex carbohydrates. The assembly of these complex nanomachines results from the very tight binding of repetitive cohesin modules, located in a non-catalytic molecular scaffold, and dockerin domains located at the C-terminus of the enzyme components of the cellulosome. The number of enzymes found in a cellulosome varies but may reach more than 100 catalytic subunits if cellulosomes are further organized in polycellulosomes, through a second type of cohesin-dockerin interaction. Structural studies have revealed how the cohesin-dockerin interaction mediates cellulosome assembly and cell-surface attachment, while retaining the flexibility required to potentiate catalytic synergy within the complex. Methods that might be applied for the production, purification and structure determination of cohesin-dockerin complexes are described here.

∞ The student contributed in the following methodologies: cloning, expression, purification and crystallization.

2.1.1. Introduction

In anaerobic ecosystems, recycling of photosynthetically fixed plant cell wall carbon is mediated by an extensive repertoire of microbial modular enzymes that are organized in multi-enzyme complexes termed cellulosomes (Fontes & Gilbert, 2010). Integration of cellulases and hemicellulases in these highly efficient nanomachines represents a powerful mechanism for targeting multi-enzymes to a localized region of the substrate, while also promoting enzyme synergy. The quaternary assembly of cellulosomes, exemplified by the *Clostridium thermocellum* complex, is dictated by highly ordered protein:protein interactions between cohesins and dockerins (Bayer *et al.*, 2004). Cohesins are found as repetitive domains in a large non-catalytic molecular scaffold (defined as the scaffoldin), while dockerins are part of the cellulosomal catalytic subunits. The multifunctional scaffoldin contains, in addition to the various cohesin domains, a divergent dockerin that specifically interacts with cohesins found in polypeptides located at the bacterial surface (Leibovitz & Béguin, 1996). Thus, dockerins that are components of the enzymes only recognize scaffoldin cohesins. These interactions were termed as type I. In contrast, scaffoldin dockerins exclusively bind cell surface cohesins and these complexes were termed as type II. The structures of several cohesin-dockerin complexes have started to reveal the molecular determinants responsible for the high affinity and tight specificity displayed by these protein:protein interactions (Adams *et al.*, 2006; Carvalho *et al.*, 2003; Carvalho *et al.* 2007; Pinheiro *et al.*, 2008). The different methods that can be applied for this purpose are described below.

2.1.2. Cloning of cohesin and dockerin genes in prokaryotic expression vectors

Dockerins are usually present as a single copy at the C-terminus of cellulosomal cellulases and hemicellulases. These non-catalytic domains consist of ~70 amino acids that contain two duplicated segments, each of about 22 residues (Fontes & Gilbert, 2010). The first 12 residues of each duplicated segment resemble the calcium-binding loop of F-hand motifs in which the calcium-binding residues, asparagine or aspartate, are highly conserved (Pagès *et al.*, 1997). Calcium was shown to play a critical stabilizing role in dockerin structures (Choi & Ljungdahl, 1996). In addition, calcium is required for dockerin function (Fontes & Gilbert, 2010). Thus, in the presence of EDTA, dockerins are unable to interact with cohesins. Dockerins are highly unstable when produced as discrete entities in *Escherichia coli,* being very susceptible to proteolysis and degradation. However, high levels of dockerin expression in *E. coli* can be obtained when these unstable domains are co-expressed *in vivo* with their cognate cohesin partners (Carvalho *et al.*, 2003; Carvalho *et al.*, 2007). It is believed that the

binding of dockerins to cohesins after the small recombinant domain (7-9 kDa) is properly folded fulfils a critical role in dockerin stabilization in *E. coli*. Co-expression of dockerins with cohesins might be obtained either through the cloning of both encoding genes in the same plasmid, or through the cloning of the two genes in different, but compatible, plasmids.

2.1.2.1. Cloning genes encoding dockerin and cohesin modules through PCR

The genes encoding dockerin and cohesin domains are amplified by PCR from bacterial genomic DNA using previously designed primers. Primers should contain engineered restriction sites for direct cloning into the appropriate vectors.

1. Set up a 50 µl PCR reaction using 50-200 ng of bacterial genomic DNA, 0.4 µM of each respective primer, 0.4 mM dNTPs, 2.5 U of a proofreading DNA polymerase in 1× of the respective buffer as recommended by the thermostable polymerase manufacturer.
2. Run the PCR reaction in a conventional thermocycler using 30 amplification cycles with an annealing temperature that is 5 ºC lower than the primers melting point.
3. When the amplification reaction is finished, subject 5-10 µl of each reaction to agarose electrophoresis, using a 1-1.5 % (w/v) gel, to confirm that the amplification reaction was successful.
4. Purify the amplified genes from nucleotides and unincorporated primer dimers in silica columns following the manufacturer's instructions. Recover the purified PCR products in 50 µl of elution buffer (5-10 mM Tris-HCl, pH 8.5 or water).
5. Clone the genes into a blunt ended prokaryotic vector following the manufacturer instructions. Sequence the cloned genes to verify that no mutations have accumulated during the amplification.

2.1.2.2. Producing synthetic dockerin or cohesin genes

Under some circumstances the lack of bacterial genomic DNA, or inappropriate codon usage for obtaining high levels of gene expression in *E. coli*, might entail the production of synthetic genes in addition to the strategy described in section 2.1.2.1.

1. Select the primary sequences of the required cohesin and dockerin genes and design the genes encoding the respective proteins with a codon usage that is compatible with high level of expression in *E. coli* (gene design might be performed using Genedesigner by DNA2.0; https://www.dna20.com/genedesigner2; (Villalobos, Ness, Gustafsson, Minshull, & Govindarajan, 2006).This dedicated software excludes undesired internal restriction sites, repetitive regions or putative regulatory sequences.
2. Divide the designed gene into overlapping oligonucleotides (20 bp overlap and 40 bp in length). One can use a dedicated software to design primers with overlapping

regions with similar melting temperatures (for example Gene2oligo http://berry.engin.umich.edu/gene2oligo; (Rouillard *et al.*, 2004). Design upstream and downstream primers incorporating the engineered restriction sites that will be used for the subsequent cloning reactions.

3. Assemble a 50 µl PCR reaction using 25 pmoles of the upstream and downstream primers and 0.25 pmoles of the internal primers.
4. Perform the PCR reaction as described in Section 2.1 (point 2) using a proofreading thermostable DNA polymerase. Perform a standard PCR cycle using a 55 ºC annealing temperature and an extension period of at least 1 min/kb.
5. Check the result of the PCR reaction through agarose gel electrophoresis. Clone the PCR product of the estimated size into a blunt-ended vector as described above.
6. Sequence the synthetic gene to confirm that no mutations have accumulated during the amplification.

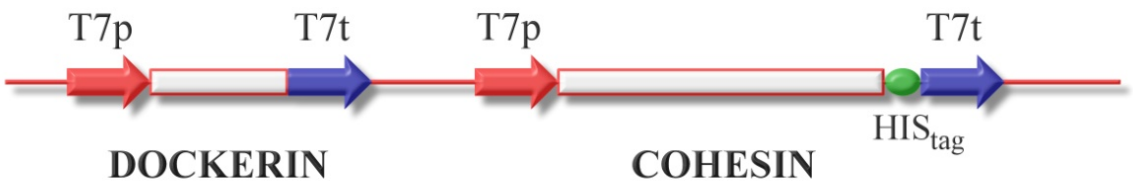
2.1.2.3. Producing genes constructs for protein co-expression

Following our established strategy, cohesins and dockerins are cloned under the control of separated promoters in the same plasmid (Figure 2.1), allowing the simultaneous expression of the two proteins in *E. coli*. Recombinant cohesins usually contain N- or C-terminal His_6-tags and co-expressed dockerins contain no additional vector-derived primary sequence. Following this approach, and assuming the *in vivo* binding of cohesins to dockerins, Immobilized Metal Affinity Chromatography (IMAC) can be used to purify the cohesin-dockerin complexes and unbound cohesins (Figure 2.2). After the first chromatographic step unbound dockerins are eliminated. A second purification step is, however, required to obtain purified cohesin-dockerin complexes (Figure 2.2). Although we usually use this approach, we anticipate that the reverse strategy of including a His_6-tag in the dockerin might also be successful.

1. Subclone the cohesin gene into pET21a (Novagen) or any other suitable vector so that the gene is under the control of a T7 promoter and a T7 terminator. The generated plasmid is termed pET21a_coh. The cloned gene should have no 3´-end stop codon and should be in frame with the vector His_6-tag sequence. This strategy will ensure that the recombinant cohesin will contain a C-terminal His tag.
2. Subclone the dockerin gene from the standard cloning vector into the prokaryotic expression vector pET3a (Novagen). Other vectors with similar properties can also be used. However, if using pET3a the gene can only be subcloned in *Nde*I and *Bam*HI restriction sites. The dockerin gene should contain a stop codon at the 5´-end, just before the engineered restriction site.

3. To allow excision of the dockerin gene with the T7 terminator sequences, use site-directed mutagenesis to engineer a *Bgl*II restriction sequence at the 5´-end of the T7 terminator sequence. Use a commercially available kit (NZYTech, for example) and follow the manufacturers' instructions.
4. Sub-clone the dockerin gene under the control of T7 promoter and a T7 terminator by digesting the generated pET3a-doc-mut plasmid with *Bgl*II. Purify the digested gene following agarose gel electrophoresis. Ligate into pET21a_coh previously digested with *Bgl*II and dephosphorylated.
5. Perform a restriction analysis of the generated plasmids to evaluate the orientation of the Coh and Doc genes. (Figure 2.1) However, gene orientation usually does not affect expression levels.

Figure 2.1| DNA construct containing cohesin and dockerin genes cloned in tandem under the control of separate T7 promoter and terminator regions.



2.1.3. Expression & purification of cohesin-dockerin complexes in *E. coli*

2.1.3.1. Protein Expression

Expression of cohesin and dockerin genes under the control of T7 promoters requires the use of *E. coli* DE3 strains.

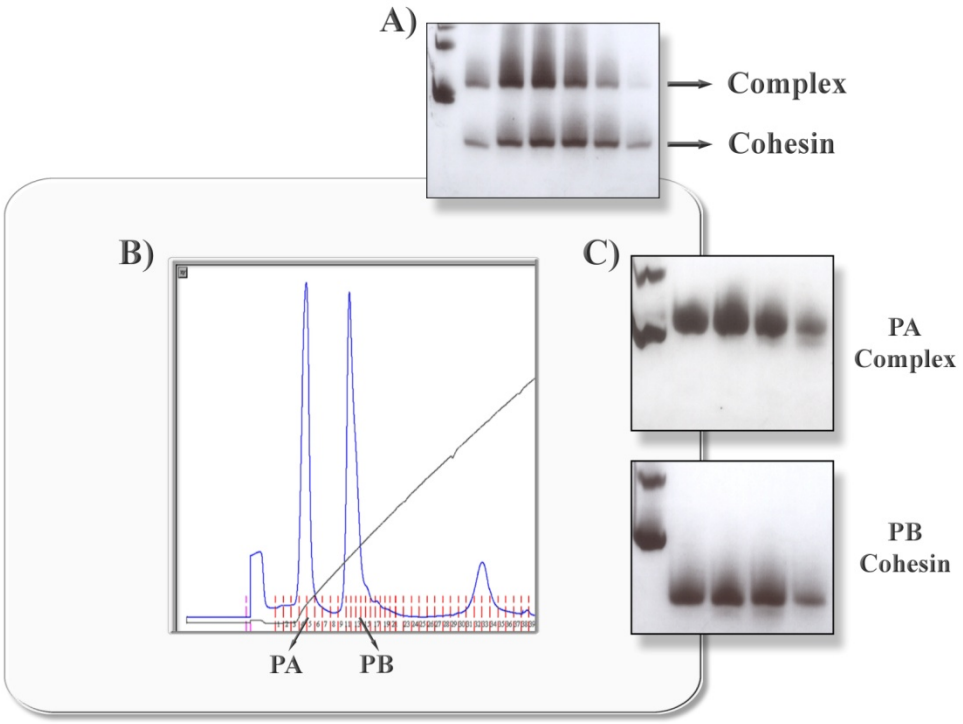
1. Plasmids containing the dockerin and cohesin genes organized in tandem and under the control of separate T7 promoter and terminator sequences are used to transform BL21 (DE3) cells.
2. Recombinant *E. coli* cells are grown in LB media supplemented with the appropriate antibiotic at 37 ºC untill OD_{550} 0.5. Gene expression is induced by the addition of 1 mM isopropyl-β-D-thiogalactopyranoside (IPTG). Induced cells are further incubated for 16 hours at 19 ºC.
3. Centrifuge the cell suspension at 5000×g for 15 min at 4 ºC.
4. Resuspend collected cells in 50 mM Na-Hepes buffer, pH 7.5, containing 1 M NaCl, 10 mM imidazole and 5 mM $CaCl_2$. Disrupt bacterial membranes and cell wall through ultrasonication.
5. Collect cell-free extracts through centrifugation at 15000×g for 30 min at 4 ºC.

2.1.3.2. Protein Purification

For crystallography, cohesin-dockerin complexes are usually purified through three purification steps using an FPLC chromatography system (Figure 2.2). All procedures, unless otherwise indicated, are carried out at 4 ºC.

1. Unbound cohesin and cohesin-dockerin complexes are initially purified through IMAC essentially as described by (Pinheiro *et al.*, 2008) in HisTrap^{TM} HP 5 ml columns (GE Healthcare). The column is equilibrated with 50 mM NaHepes buffer, pH 7.5, containing 1 M NaCl, 10 mM imidazole and 5 mM $CaCl_2$ and after loading the *E. coli* extracts the column is extensively washed with the same buffer. Proteins are eluted from the column in a gradient of the equilibration buffer and 50 mM NaHepes buffer, pH 7.5, containing 1 M NaCl, 300 mM imidazole and 5 mM $CaCl_2$.
2. Fractions containing the protein-protein complexes are selected following 10% native gel electrophoresis and 16% SDS-PAGE (Figure 2.2). The complex is usually co-purified with unbound cohesin (Figure 2.2). A control, consisting of purified cohesin, should be incorporated in the native gel to allow the identification of the cohesin-dockerin complex band.
3. IMAC purified proteins are buffer exchanged in PD-10 Sephadex G25M gel filtration columns (GE Healthcare) into 20 mM Tris-HCl buffer, pH 8.0, containing 2 mM $CaCl_2$.
4. Proteins are subjected to a further purification step by anion exchange chromatography using a column loaded with Source 30Q media (GE Healthcare).
5. Separation of the cohesin-dockerin complexes from the individual cohesin is achieved through the application of a 0-1 M NaCl elution gradient.
6. Protein fractions are analysed through affinity gel electrophoresis. Fractions containing cohesin-dockerin complexes purified from isolated cohesin are selected and, if required, further purified by gel filtration chromatography (Figure 2.2).
7. Before gel filtration chromatography, protein fractions are buffer exchanged into 20 mM Na-Hepes buffer, pH 7.5, containing 200 mM NaCl and 2 mM $CaCl_2$ as described above.
8. The protein is concentrated with an Amicon 10 kDa cut off molecular-weight centrifugal membrane to approximately 25 mg/ml.
9. The protein-protein complex is loaded into an HiLoad 16/60 Superdex 75 column (GE Healthcare) previously equilibrated with 20 mM Na-Hepes buffer, pH 7.5, containing 200 mM NaCl and 2 mM $CaCl_2$.
10. Purity of eluted protein fractions is evaluated through SDS-PAGE and native gel electrophoresis, as mentioned before. Selected fractions are pooled and concentrated as described above. Pure complexes are buffer exchanged by washing with 2 mM $CaCl_2$ and concentrated to 6-12 mg/ml.

Figure 2.2| Purification of cohesin-dockerin complexes.



A) Purified fractions collected after the first purification step through IMAC are analysed through native acrylamide gel electrophoresis. Affinity chromatography generated both unbound cohesin and the cohesin-dockerin complex suggesting that the dockerin is expressed at a lower level. **B)** Ion exchange chromatography is used to separate the protein-protein complex from unbound cohesin. **C)** Purified fractions collected after ion exchange are analysed through native acrylamide gel electrophoresis. Fractions from peak A, analysed in the first gel, contain protein-protein complex. Fractions from peak B contain individual cohesin that was overexpressed.

2.1.4. The dual binding mode and the crystallization of cohesin-dockerin complexes

Dockerins are usually highly symmetrical molecules and generally contain two cohesin binding interfaces (Fontes & Gilbert, 2010). Dockerin primary sequence is a tandem duplication of a 22-residue segment that displays remarkable structural conservation. Thus, in *C. thermocellum*, the structure of the first duplicated segment, containing the N-terminal helix, can be precisely superimposed over the C-terminal helix (helix-3). Several mutagenesis studies informed by the structure of cohesin-dockerin complexes (Carvalho *et al.*, 2003; Carvalho *et al.*, 2007), revealed that the *C. thermocellum* type I dockerin contains two ligand binding sites that display similar affinity to its protein partner. Thus, in one complex helix-3 dominates cohesin recognition with Ser-45 and Thr-46 playing a central role in the polar interactions between the two protein partners (Carvalho *et al.*, 2003). In the second binding mode, the dockerin is rotated 180° relative to the cohesin and helix-1, rather than helix-3, plays a central role in complex formation (Carvalho *et al.*, 2007). Thus, the equivalent residues to Ser-45 and Thr-46 in the N-terminal helix, Ser-11 and Thr-12,

dominate the hydrogen-bonding interactions between the dockerin and its cohesin partner in this second binding mode. Similar observations were made in cohesin-dockerin interactions of *C. cellulolyticum* (Pinheiro *et al.*, 2008). Although dockerins present two cohesin-interacting surfaces, only one of these sites interacts with a cohesin at a defined moment.

The dual binding mode expressed by dockerins poses significant obstacles to cohesin-dockerin complex crystallization. Thus, initial attempts to crystallise the purified dockerin-cohesin complexes were unsuccessful. This observation likely reflects the dynamic binding of the two potential dockerin binding sites to the cohesin. To encourage a single binding mode between the protein partners, two variants of the dockerin should be constructed in which the function of site 1 or site 2 is disrupted by the introduction of mutations at the residue pairs that dominate cohesin recognition at each binding site. To achieve this, a commercial site-directed mutagenesis kit should be employed and the residues should be changed to alanine or large bulky amino acids such as glutamine.

2.1.5. X-ray crystallography of cohesin-dockerin complexes

Solving the three-dimensional structure of a macromolecule by X-ray crystallography involves several steps. Once a target protein is expressed and purified, it needs to be crystallized and single crystals must be obtained. A single crystal is built-up by translationally repeating units, called unit cells. Each unit cell is characterized by three cell axes *a*, *b*, *c* and by three angles *α*, *β* and *γ*. Crystals are subjected to X-rays and diffraction data are measured, processed and analyzed. Diffraction data are recorded as a range of diffraction patterns, containing spots (reflections), each attributed an *hkl* index. Each reflection *hkl* is produced by families of imaginary planes passing by all atoms in the crystal lattice. Thus, diffraction (the production of a spot in the diffraction pattern) only occurs when a constructive interference of the scattered radiation occurs, satisfying Bragg's law. The diffraction experiment produces a list of intensities (I_{hkl}) and associated errors for each reflection recorded. Each reflection can also be regarded as a scattered wave, with an amplitude, phase and frequency (the same as the incident radiation). Mathematically, a wave can be described as a Fourier series, formulated as a function of the electron density $\rho(x,y,z)$ of all atoms in the unit cell, Eq. (1).

(1)

$$F_{hkl} = \int_V \rho(x,y,z) e^{[2\pi i(hx+ky+lz)]} dV$$

where F_{hkl} (the structure factor) also possesses an associated amplitude, phase and frequency. Applying the Fourier Transform to Eq. (1), the Electron Density function is obtained,

(2)

$$\rho(x,y,z) = (1/V)\sum_{h}\sum_{k}\sum_{l}|F_{hkl}|e^{2\pi i\alpha_{hkl}}e^{[-2\pi i(hx+ky+lz)]}$$

where α_{hkl} is the phase angle of reflection *hkl*, $|F_{hkl}|$ is the structure factor amplitude (proportional to √I), *(x,y,z)* are the fractional atomic coordinates in the unit cell and *V* is the volume of the unit cell. The diffraction experiment provides the values of $|F_{hkl}|$, however, the phase information is lost, and this is known as the Phase Problem in crystallography. The Fourier Transform marks the frontier between what is known as the reciprocal space (the space of the diffraction pattern, *hkl* reflections and structure factors) and the real space (the space of the electron density and atomic coordinates). Solving the phase problem requires the initial phases to be estimated. There are several methods available that allow the indirect determination of the phase angles but, in this paper, only the Molecular Replacement (MR) method is used (Long, Vagin, Young, & Murshudov, 2008). This is the method of choice when a structure of a similar protein to the one of interest is available. Similarity is evaluated by the primary sequence similarity between the two proteins; the higher the sequence similarity, the higher the chance that the proteins will share a common fold. The known structure is referred to as the search model. It is used to locate the position of the protein of interest in the unit cell. Particular electron density maps (where the phases are attributed a value of 0, known as Patterson maps) are calculated for both proteins and superposed for comparison to find a match. This is usually done in two steps: rotation of the search model to find the molecule's orientation, and translation of the model to find the position in the unit cell. The correct positioning of the search model in the unit cell provides the phase estimates that solve the Phase Problem and enable calculation of the first electron density map. After this, model adjustments will have to be made, in order to bring the model structure as close as possible to the structure of the protein of interest. In an iterative way, new electron density maps are calculated after model adjustments and these should improve with the addition of correct features to the model. Validation methods are used to monitor every step of model building and adjustment, comparing calculated structure factors $|F_{calc}|$, from the model, with observed structure factors $|F_{obs}|$, measured in the diffraction experiment, according to the expression:

(3)

$$\frac{\sum_{hkl}\left||F_{obs,hkl}| - |F_{calc,hkl}|\right|}{\sum_{hkl}F_{obs,hkl}}$$

This is known as the *R* factor. For cross-validation, 5 to 10% of the unique reflections are arbitrarily excluded from the refinement process and used to calculate the free *R* factor (R_{free}). This is used to prevent over-fitting of the model, as *R* and R_{free} should not differ by more than 5%. Other validation criteria ensure that the final structure is in agreement with

known chemical and geometrical parameters. In the following sections, a step-by-step procedure to solve the crystal structure of cohesin-dockerin complexes is presented.

2.1.5.1. Crystallization

The goal of crystallization is to obtain a well-ordered crystal, capable of producing a diffraction pattern, when subjected to X-rays. In the crystallization process, the purified protein gradually precipitates in an aqueous solution and, under the appropriate conditions, protein molecules adopt a consistent orientation, aligning themselves in repeating blocks of "unit cells". Protein crystallization is difficult because of the fragile nature of protein crystals, dominated by large channels of solvent, and where non-covalent interactions are responsible for sustaining the crystal lattice (Matthews, 1968). However, the fragility of protein crystals is not the only problem to overcome in crystallization. Since so much variation exists, each individual protein requires very specific and unique conditions to produce a well-ordered crystal. Many environmental factors have to be considered, like protein purity and concentration, pH, temperature and precipitants.

Vapor Diffusion is usually the method of choice for protein crystallization and the one we describe here. The two variations of the method are known as the hanging drop and sitting drop techniques. Basically, a drop containing purified protein, buffer, and precipitant is allowed to equilibrate against a similar reservoir solution, containing the precipitant in higher concentrations. In a sealed environment, water vapor diffuses from the drop to the reservoir, bringing the concentration of the precipitant in the drop closer to its concentration in the reservoir. The drop slowly loses volume, concentrating the protein slowly enough to permit the consistent orientation of protein molecules to form a crystal. The search for initial crystallization conditions is usually done using commercial screens, which provide trial formulations, often selected from known crystallization conditions for proteins.

Screening for suitable crystallization conditions requires the setup of many different conditions. Therefore, it is a common practice nowadays to use automated systems, known as crystallization robots. Still based on the vapor diffusion method, these systems have the major advantage of screening a large number of crystallization conditions while using a much smaller amount of protein solution, when compared with the traditional setups. Besides, setting up drops with a robotic system is a quick, reproducible and less error-prone procedure. These equipments are computer-controlled, running software that not only gives instructions to operate the system, but can also be used to program different experiments, from screening to optimization. Drop volume can vary from 0.2 to 10 μl, using, for example, plates of 96 wells. An imaging system is useful for monitoring subsequent crystal growth.

2.1.5.1.1. Screening of crystallization conditions

The crystallization of a new cohesin-dockerin complex will require the search for initial conditions, which is most efficiently done using an automated nano-drop dispensing system, as described in the following protocol.

1. Prepare your protein complex for crystallization by buffer exchanging it into 10-20mM Tris-HCl, pH 8, and 2 mM $CaCl_2$. Low concentration of buffer guaranties that the crystallization pH will be determined by the candidate precipitant solution and not the protein solution. Different pHs can cause different packings. It is very important that the complex is close to 97% pure. $CaCl_2$ is required for stabilization of the dockerin module.
2. Pre-load the plate's reservoirs with the solutions of your screens of choice.
3. If you are using an automated system, follow the equipment's instructions to load the protein. The volume of dispensed protein can vary from 0.1 to 10 μl.
4. Program the software to perform your screening. Depending on the system, it will dispense nano-drops of precipitant solutions, mixing them with the protein solution.
5. Seal the plate, visualize it under the microscope, take note of any early precipitation and store the plate at the chosen temperature (4 or 20 ºC are commonly used, but different temperatures may be explored). Depending on the amount of purified complex, many screening trials can be set up.
6. If you are setting up your drops "by hand", you can use the procedure described below to prepare a screening of crystallization conditions, with one different condition per well.

2.1.5.1.2. Crystals Optimization

Once the preliminary conditions are found, crystals may need optimization (which can be performed automatically in some systems) or scaling up to produce crystals with bigger dimensions (Figure 2.3). Larger crystals may be more suitable for subsequent manipulation.

1. Scaling up usually means preparing the drops using bigger volumes of protein and precipitant solutions, but it may also require adjustment of the preliminary crystallization conditions. At least, drop proportion should be assayed, but pH conditions may also change. Prepare a plate for crystallization (Linbro plates are commonly used). Grease the top of the reservoir well with high-vacuum grease.
2. Fill the reservoir with 500-700 μl of precipitant solution (found in the automated screening experiments).
3. In a lamella, dispense 2 μl of protein solution (different volumes can be used). Since lamellas are big enough, one suggestion would be to assay two different protein concentrations per lamella.
4. Pipette 2 μl of reservoir solution and mix it in the protein drop.

5. Flip the lamella, putting the fresh drop facing the reservoir solution. This describes the hanging drop method. In the sitting drop technique, the drop is dispensed on top of a support inside the reservoir. In the process of optimizing crystallization conditions, both techniques can be used.
6. Seal off the system using the high-vacuum grease to make it airtight.
7. Visualize the prepared drops under the microscope, take note of any early precipitation and store the plate at the temperature of interest.

2.1.5.2. X-ray data collection and reduction

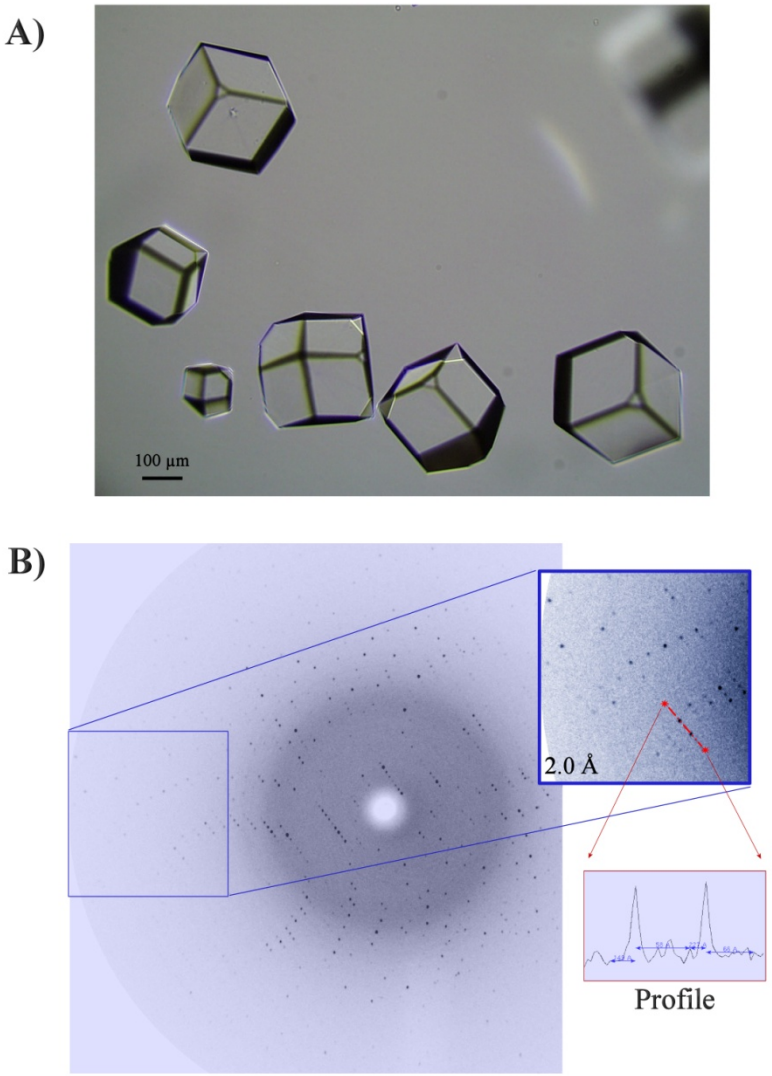
After a period of 5 to 6 days for crystal growth at room temperature (Figure 2.3), crystals can be cryo-protected, mounted and flash-cooled in a dry nitrogen stream or directly in liquid nitrogen. Exposure to intense X-rays can cause radiation damage to the crystals, which is minimized by lowering the temperature to approximately 100K. The cryo-protectant prevents the formation of disordered ice in the crystal's solvent channels, which can damage the crystal or interfere in diffraction by forming ice rings (visible in the diffraction pattern).

1. Stabilize the crystal by transferring it to a harvesting solution, similar to the crystallization solution, but where the precipitant is in higher concentration; this prevents dissolution of crystals.
2. Transfer the crystal to a second drop (~10μl) containing 20-40% (v/v) of glycerol added to the harvesting solution, used previously.
3. Crystals can now be flash-cooled in a dry nitrogen stream or directly in liquid nitrogen, bringing the crystals to very low temperatures (around 100K) and facilitating crystal handling and storage. These procedures require the use of specific cryo-tools (Pflugrath, 2004).
4. Crystal testing and data collection can be performed using graphite monochromated CuK_{α} radiation from a rotating anode generator and an imaging plate system as detector. Under this radiation, cohesin-dockerin crystals typically diffract beyond 2.5 Å resolution.
5. For higher resolution, X-ray diffraction data can be collected using a synchrotron radiation source in a beamline dedicated to macromolecular crystallography (e.g. ID14 at the European Synchrotron Radiation Facility).
6. The availability of 3D structures of the isolated cohesin (e.g. PDB ID code 1ANU), means that structure solution can be attempted by the well-established method of MR. Therefore, a single-wavelength data collection with high completeness is, in theory, sufficient to solve the 3D structure. Start by collecting a few frames, each with 1º oscillation angle (this is just a suggested starting value).
7. Using program MOSFLM (Leslie, 2006), autoindex the frames, determining the symmetry and unit cell parameters of the crystal, and the orientation of the crystal

relative to the beam. Depending on the complex, space groups may differ (e.g. cubic $P2_13$ in the case of native *C. thermocellum* type I cohesin-dockerin complex, and monoclinic $P2_1$ in the case of Ser45A, Thr46A mutant) and the amount of unique data will depend on the inner symmetry of the crystals.

8. Redundancy of data may also be crucial for solving a structure by MR methods. However, in high intensity beams, such as synchrotron sources, crystal decay can occur due to radiation damage, even for cryo-protected crystals, so the optimal strategy of data collection will have to take into account several of these factors. The Strategy option in the program MOSFLM is helpful to calculate the best data collection strategy, determining the best phi angles and ranges to obtain a complete data set. This is followed by refinement of crystal parameters (unit cell and mosaic spread), detector and beam parameters.
9. Proceed to obtain a data set as complete as possible, by collecting the necessary number of frames. Depending on the cell constants and the visible resolution limits, the oscillation angle and the detector-to-crystal distance will have to be adjusted, such that good spot separation is obtained and a significant number of full reflections are measured in each diffraction image.
10. After a complete data collection, use MOSFLM to integrate the several diffraction patterns (obtained over a wide range of rotations) into a list of indices (*hkl*), each with a measured intensity and an associated uncertainty.
11. .Data integration is followed by scaling with program SCALA (Philip Evans, 2006) from the CCP4 suite of programs (Winn *et al.*, 2011), to put all observations on a common scale. Check the overall quality of the data by analysing the agreement between equivalent reflections. Check the value of R_{merge} against batch number to detect outliers. At this stage, you can already select the arbitrary fraction of data (5-10% of the reflections) to be used to calculate R_{free} during structure refinement.
12. Use program TRUNCATE from the CCP4 suite of programs (Winn *et al.*, 2011), to convert intensities (I_{hkl}) to structure factor amplitudes (F_{hkl}) and to generate useful intensity statistics. Data quality indicators are factors R_{merge} and R_{pim}, mean $[(I)/\sigma(I)]$, completeness and multiplicity of data (Weiss, 2001). Possible twinning problems or the presence of translational NCS (non-crystallographic symmetry) can also be diagnosed, at this stage.

Figure 2.3| Crystals of cohesin-dockerin complexes and X-ray analysis.



A) Cubic-shaped native Coh-Doc Type I complex crystals, around 0.2 x 0.2 x 0.2 mm, obtained in PEG/Ion Screen HR2-126 from Hampton Research (condition 18).

B) Diffraction pattern obtained from a mutant Cohesin-dockerin Type I complex crystal, using graphite monochromated CuKα radiation from an Enraf-Nonius FR591 rotating anode generator and a MAR-Research 300mm imaging plate detector. Inset image depicts high resolution reflections and their profile along the dotted red line.

2.1.5.3. Model building and structure refinement

Molecular replacement provides the solution to the phase problem and it is now possible to calculate an electron density map, which should be inspected for secondary structure features. The quality of this electron density map depends on the quality of the initial phases, as well as the resolution and quality of the measured data. The correctly placed cohesin model will have to be adjusted to the features observed in the electron density, which usually includes adjusting side-chains or any parts that are not conserved between the search model and the cohesin of interest (some amino-acid residues may need to be altered, according to the primary sequence). Since MR was not performed using a search model for the dockerin,

this smaller module can be built from the observation of the electron density map. This can be done by the experimenter, using a graphics program (e.g. COOT (Emsley & Cowtan, 2004) and the direct observation of the electron density, or automatically, using dedicated software (e.g. ARP/wARP (Langer, Cohen, Lamzin, & Perrakis, 2008) or AutoBuild in PHENIX (Terwilliger *et al.*, 2008)). Model improvement and addition of correct features will bring the model closer to the real structure and new improved phases can be calculated. These are used, in an iterative way, to calculate a new electron density map, which may reveal new features that will have to be added to the model. The iterative process is validated by comparing calculated structure factors $|F_{calc}|$, from the model, with observed structure factors $|F_{obs}|$, measured in the diffraction experiment. After adjusting and building missing parts of the complex, water molecules and/or other ligands will have to be identified and added. Overall, model building and improvement is done in real space, using graphics programs, while refinement of atomic positions, temperature factors (a measure that indicates how much an atom oscillates around a specific position) and occupancies (the fraction of molecules where the atom occupies that specific position in the crystal) is performed in the reciprocal space (e.g. program phenix. refine). A suggested procedure for model building and refinement of cohesin-dockerin complexes is presented. These steps are a simple guidance; crystallographic software includes many other options that can be used in different ways, depending on the model building and refinement context.

1. After structure solution with PHASER (McCoy *et al.*, 2007), the coordinates of the positioned cohesin can be loaded in COOT for visualization. Use the unit cell and space group information to visualize also the cell content and symmetry. Confirm that there are no clashes and there is enough space to fit the dockerin module.
2. Load the file output by PHASER containing the information to calculate electron density maps. PHASER outputs a file containing column labels FWT/PHWT (amplitude and phase for $2m|F_{obs}|-D|F_{calc}|\exp(i\alpha_{calc})$ map) and DELFWT/PHDELWT (amplitude and phase for $m|F_{obs}|-D|F_{calc}|\exp(i\alpha_{calc})$ map) to calculate sigmaA-weighted electron density maps. The difference map $mF_{obs}-DF_{calc}$ will show positive density where features must be added to the model and negative density where atoms have to be removed from the model. The double difference map $2mF_{obs}-DF_{calc}$ is used to minimize bias from the model.
3. Use the modeling tools from COOT to draw a skeleton in the $2mF_{obs}-DF_{calc}$ map. Try to identify the part of the electron density corresponding to the dockerin module and the characteristic α-helices. Identify the N- and C-terminus and confirm the connectivity of the polypeptide chain, making sure that the correct identification of helices is not impaired by the inner symmetry of the dockerin modules (i.e. make sure you can unambiguously identify some amino-acid residues in each α-helix). If resolution permits and phases are sufficient, most of the polypeptide chain can be

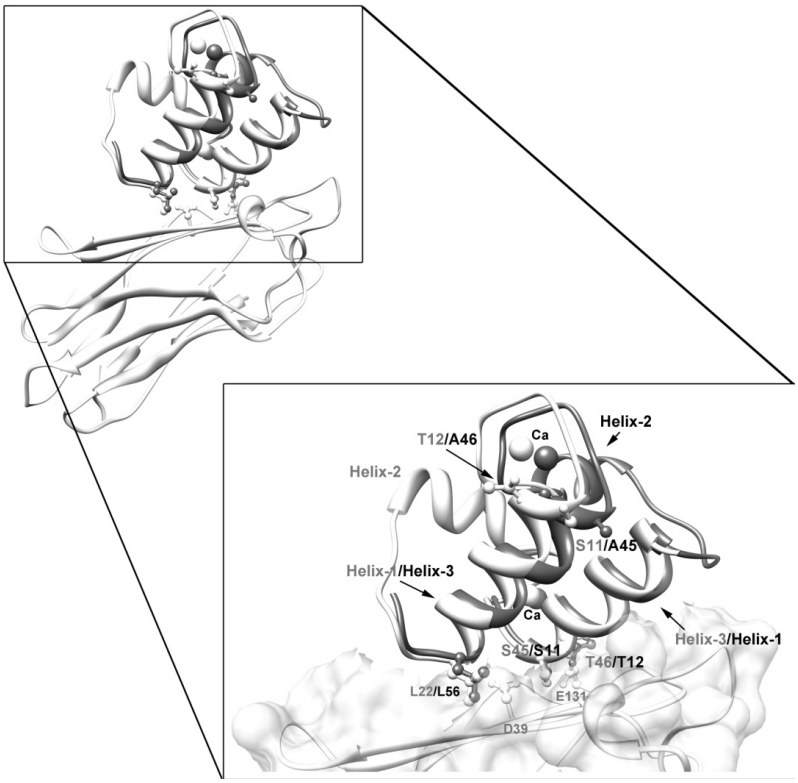
identified and no connectivity errors will be introduced. Good enough resolution may allow identification of the side chains of several amino-acid residues.

4. Use COOT to place Cα carbons at 3.8 Å intervals, followed by a main-chain trace of the polypeptide chain. The presence of some characteristic residues may help to identify local stretches of sequence and side-chains may be introduced.
5. Several solvent molecules may be indicated by positive peaks in the mF_{obs}-DF_{calc} map and should be added to the model. This can also be done automatically using dedicated software; nevertheless, solvent molecules should be visually inspected and kept if matching several criteria, like temperature factors, map sigma level and contacts.
6. After this, the model can be refined with program phenix-refine from the PHENIX software suite (Adams *et al.*, 2010), by uploading the coordinate file of the improved model and the file containing observed structure factors $|F_{obs}|$ (and associated uncertainties), measured in the diffraction experiment.
7. In Refinement settings, choose to refine individual sites and individuals ADPs, as refinement strategy. The model x,y,z coordinates and temperature factors (also known as atomic displacement parameters, ADP) are refined and used to calculate new phase angles, hopefully more accurate than the experimental phases.
8. Considering that cohesin-dockerin complexes are composed of separate domains, choose option TLS parameters to automatically find suitable TLS (Translation/Libration/Screw) groups, using program phenix.find_tls_groups (Winn, Isupov, & Murshudov, 2001). The program will analyze the crystal structure of the complex, searching for evidence of flexibility, e.g. local or inter-domain motions. Individual chains are partitioned into multiple segments that are modeled as rigid bodies undergoing TLS vibrational motion. Each group, having a different number of segments, is scored according to its ability to explain the observed atomic displacement parameters ("B values") obtained in the crystallographic refinement.
9. If two or more copies of the complex are present, NCS restraints should also be imposed. NCS groups can be found automatically by phenix.refine or be defined by the user.
10. Phenix.refine will output several files, including the refined model, various maps, structure factors and complete statistics. Load the model and maps in COOT and proceed with new adjustments to the model, according to the new features in the improved electron density (use the several validation tools provided by COOT).

2.1.6. Summary

Anaerobic microbes produce a remarkably efficient nanomachine to deconstruct plant cell wall polysaccharides, which was termed, when discovered more than 20 years ago, as the cellulosome. Cellulases and hemicellulases are assembled into multi-enzyme complexes through a high affinity interaction established between type I dockerin domains of the modular enzymes and type I cohesin modules of a non-catalytic scaffoldin (Figure 2.4). It is believed that integration of the microbial biocatalysts into cellulosomes potentiates catalysis through the maximization of enzyme synergism afforded by enzyme proximity and efficient substrate targeting. Substantial structural and functional evidence exists suggesting that cellulosomal dockerins display a dual cohesin binding interface. The dual binding mode expressed by cohesin-dockerin complexes may introduce enhanced flexibility in the quaternary organization of the multi-enzyme complex thus potentiating the hydrolysis of a predominantly insoluble substrate. Recently it has become apparent that the cohesin-dockerin interaction is quite widespread in nature and may fulfil a large range of, mostly currently unknown, functions which remain to be described.

Figure 2.4| The dual binding mode of cohesin-dockerin complexes.



Ribbon representation of the superposition of the dockerin modules of Type I Cohesin-dockerin native complex (light gray) with the S45A-T46A mutant complex (dark grey) in *C. thermocellum*. For simplification only one cohesin module is represented. The inset shows a more detailed view of the cohesin-dockerin contacts and of the almost perfect superposition of helices 1 and 3 of both complexes. In the mutant complex, helix-1 (containing Ser-11 and Thr-12) dominates binding whereas, in the native 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 ball-and-stick models.

3. NOVEL TYPE I COHESIN DOCKERIN COMPLEXES

3.1. Novel *Clostridium thermocellum* Type I Cohesin-Dockerin Complexes Reveal a Single Binding Mode∞

Joana L. A. Brás[a,*], Victor D. Alves[a,*], Ana Luísa Carvalho[†], Shabir Najmudin[*], José A. M. Prates[*], Luís M. A. Ferreira[*], David N. Bolam[‡], Harry J. Gilbert[‡] and Carlos M. G. A. Fontes[*]

*Centro Interdisciplinar de Investigação em Sanidade Animal, Faculdade de Medicina Veterinária, Universidade Técnica de Lisboa,1300-477 Lisboa, Portugal; [†]REQUIMTE–CQFB, Departamento de Química, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, 2829-516 Monte da Caparica, Portugal; [‡] Institute for Cell and Molecular Biosciences, Newcastle University, The Medical School, Newcastle upon Tyne NE2 4HH, United Kingdom; [a] Equal contribution.

Adapted from: Brás *et al.* (2012) The Journal of Biological Chemistry, doi:10.1074/jbc.M112.407700

Abstract

Protein:protein interactions play a pivotal role in a large number of biological processes exemplified by the assembly of the cellulosome. Integration of cellulosomal components occurs through the binding of type I cohesin modules located in a non-catalytic molecular scaffold to type I dockerin domains located at the C-terminus of cellulosomal enzymes. The majority of type I dockerins display internal symmetry reflected by the presence of two essentially identical cohesin binding surfaces. Here we report the crystal structures of two novel *Clostridium thermocellum* type I cohesin-dockerin complexes (CohOlpC-Doc124A and CohOlpA-Doc918). The data revealed that the two dockerins, Doc918 and Doc124A, are unusual because they lack the structural symmetry required to support a dual binding mode. Thus, in both cases cohesin recognition is dominated by residues located at positions 11, 12 and 19 of one of the dockerin binding surfaces. The alternative binding mode is not possible (Doc918) or highly limited (Doc124A) as residues that assume the critical interacting positions, when dockerins are reoriented by 180º, make steric clashes with the cohesin. In common with a third dockerin (Doc258) that also presents a single binding mode, Doc124A directs the appended cellulase, Cel124A, to the surface of *C. thermocellum* and not to cellulosomes because it binds preferentially single type I cohesins located at the cell envelop. Although there are few exceptions, such as Doc918 described here, these data suggest that there is a considerable selective pressure for the evolution of a dual binding mode in type I dockerins that direct enzymes into cellulosomes.

∞ The student contributed in the following methodologies: cloning and expression, protein purification, isothermal titration calorimetry and crystallization.

3.1.1. Introduction

Biological nanomachines combining a range of complimentary enzyme activities are critical to cellular function. Cellulosomes are one of nature´s most elaborate and highly efficient multi-enzyme complexes that actively deconstruct cellulose and hemicellulose, two of the most abundant polymers on Earth (Fontes & Gilbert, 2010; Bayer *et al.*, 2004; Bayer *et al.* 2008). Thus, these elaborate nanomachines play a major role in carbon re-cycling and provide an opportunity to explore the largely untapped energy provided by plant biomass by the bioenergy and bioprocessing sectors. It is now well established that the complex physical and chemical structure of plant cell walls restrict their access to hydrolytic enzymes. Aerobic microorganisms that utilize plant biomass as a significant nutrient express extensive repertoires of degradative enzymes, primarily, glycoside hydrolases but also lyases and esterases, which attack the structural polysaccharides of the plant cell wall. In contrast, anaerobes, due to environmental selective pressures, have a lower protein producing capacity and organize enzymes into cellulosomes, which enhance enzyme synergy and substrate targeting (see Bayer *et al.*, 2004; Fontes & Gilbert, 2010), for review).

The cellulosome of the thermophilic bacterium *Clostridium thermocellum* has been extensively explored (Bayer & Lamed, 1986; Béguin & Alzari, 1998). It consists of a large non-catalytic multi-modular protein, termed CipA that contains nine tandemly repeated type I cohesins that recognize type I dockerins located in the cellulosomal enzymes (Salamitou *et al.* 1994; Tokatlidis *et al.* 1991). Type I cohesins of CipA display a very high level of sequence identity. It was thus suggested that there is little discrimination by the dockerins and their protein receptors presented by the cellulosome scaffold (Salamitou *et al.* 1992). Primary scaffoldins such as CipA, may also contain a C-terminal divergent type II dockerin that specifically recognizes type II cohesins located at the bacterium envelop, thereby providing a mechanism for the cell surface attachment of cellulosomes (Leibovitz & Béguin, 1996). Thus, different cohesin-dockerin (Coh-Doc) specificities (in *C. thermocellum* of type I and type II) are responsible for the correct assembly of the multi-enzyme complex (type I) and its direct attachment to the organism (type II), respectively.

Structural studies on type I Coh-Doc complexes of *C. thermocellum* (Carvalho *et al.*, 2003; Carvalho *et al.*, 2007) and *C. cellulolyticum* (Pinheiro *et al.* 2008), a mesophilic bacterium that produces a cellulosome analogous to *C. thermocellum*, provided insights into the molecular determinants of protein:protein recognition that mediate the assembly of these protein complexes. Dockerins fold into two α-helices and EF hand-like calcium-binding motifs, each corresponding to one of the two duplicated segments (Carvalho *et al.*, 2003; Pinheiro *et al.* 2008). The structure of the N-terminal α-helix and EF motif can be precisely superimposed over the structure of the C-terminal α-helix and EF motif, leading to an internal two fold symmetry in the dockerin molecule (Carvalho *et al.*, 2007). The implications of this

internal symmetry were realized when it was observed that type I dockerins present two cohesin binding surfaces, as they can bind their cognate protein module either through the analogous N- or C-terminal α-helices (Carvalho *et al.*, 2007). In *C. thermocellum* type I dockerins, residues that dominate the hydrogen bond network with cohesins are located at positions 11 and 12 of the calcium binding loop and are usually a Ser-Thr pair (Carvalho *et al.*, 2003). When the dockerin is 180º reverse oriented, the equivalent residues (Ser45 and Thr46) in the C-terminal dockerin helix participate in cohesin recognition (Carvalho *et al.*, 2007). Significantly, in *C. cellulolyticum* the Ser-Thr dyad symmetry in *C. thermocellum* dockerins is replaced by hydrophobic residues, which accounts for the lack of affinity between protein partners from different species. The dockerin dual binding mode may reduce the steric constraints that are likely to be imposed by assembling a large number of different catalytic modules into a single cellulosome. In addition, the switching of the binding mode between two conformations may also introduce quaternary flexibility into multi-enzyme complexes thus enhancing substrate targeting and the synergistic interactions between some enzymes, particularly exo- and endo-acting cellulases.

Currently, it is unclear whether the dual-binding mode displayed by *C. thermocellum* and *C. cellulolyticum* dockerins is universal to all cellulosomal enzymes. The genome sequence of *C. thermocellum* ATCC 27405 encodes 72 polypeptides containing type I dockerin sequences. Alignment of the 72 dockerin sequences at the two ligand binding sites revealed a strong conservation of the amino acids that mediate cohesin recognition (particularly Ser11, Thr12 and a Lys-Arg motif at positions 18 and 19). Recently we described the identification of four dockerins, of proteins Cthe_0435 (Cel124A), Cthe_0918, Cthe_0258 and Cthe_0624 (Cel9D-Cel44A), which deviate from the canonical *C. thermocellum* motifs, at least in one of the cohesin binding interfaces (Pinheiro *et al.* 2009). Here we describe the structure of two complexes in which two different type I cohesins are bound to these unusual dockerin domains. The data indicate that a cohort of *C. thermocellum* type I dockerins display a single binding mode. The possible biological significance for the single binding mode displayed by these dockerins is discussed

3.1.2. Material and Methods

3.1.2.1. Cloning and expression

DNA encoding type I dockerins of Cthe_0435 (Cel124A, residues 31-112) and Cthe_0918 (residues 1146-1209) and type I cohesins of Cthe_0452 (OlpC, residues 108-258) and Cthe_3080 (OlpA, residues 30-177) were amplified by PCR from *C. thermocellum* genomic DNA using the thermostable DNA polymerase NZYDNAChange (NZYTech Ltd) (see Table 3.1). Genes encoding the type I dockerin domains of Cthe_0435 and Cthe_0918, here termed Doc124A and Doc918, respectively, were ligated into *Nde*I_*Bam*HI-digested pET3a (Novagen). Genes encoding cohesin domains termed CohOlpC and CohOlpA, which derive from proteins Cthe_0452 and Cthe_3080, respectively, were ligated into *Nhe*I_*Xho*I-restricted pET21a (Novagen). Recombinant cohesins contained a C-terminal His-6 tag. To express the dockerin and the cohesin genes in the same plasmid, the recombinant pET3a derivative was digested with *Bgl*II and *Bam*HI, to excise the dockerin gene under the control of the T7 promoter, which was subcloned into the B*gl*II site of recombinant pET21a so that both genes were organized in tandem. In this way it was possible to express both Doc124A and Coh452 in the same plasmid, and also Doc918 with Coh3080. Doc124A and Doc918 were also subcloned into pET32a vector (Merck, Germany) restricted with *Eco*RI and *Xho*I. Recombinant dockerins were expressed in fusion with thioredoxin, to improve dockerin solubility and stability. OlpA and OlpC cohesins were also cloned into *Bgl*II and *Eco*RI digested pRSETa (Invitrogen). Mutant derivatives of both dockerins were synthesized (NZYTech Ltd) with codon usage optimized for expression in *Escherichia coli* (Table 3.2). The synthesized genes contained engineered *Eco*RI and *Xho*I recognition sequences at the 5´ and 3´ ends, respectively, which were used for subsequent subcloning into pET-32a (Merck, Germany), as described above.

Table 3.1| Primers used to obtain the genes encoding the cohesin and the dockerin derivatives used in the present study. Engineered restriction sites are shown in bold.

Clone	Sequence (5' → 3')	Direction
DocCel124A	CTC**CATATG**TGGAATAAGGCAGTTATT	Forward
	CAC**GGATCC**TTACGAATTGTAAGAGTTC	Reverse
Doc918	CTC**CATAT**GGTTGTGCTTAATGGTGAC	Forward
	CAC**GGATCC**CTATATAGTTATAAGTCC	Reverse
CohOlpC	CTC**GCTAGC**GTTGTGGCAATTCATG	Forward
	CAC**CTCGAG**TTTTTCAATTTCCAC	Reverse
CohOlpA	CTC**GCTAGC**CAAACAAACACCATTGAA	Forward
	CAC**CTCGAG**TGCCTCCGGAGCGGATGC	Reverse

Table 3.2| Protein sequences of dockerins DocCel124A and Doc918 and respective synthesized mutants.

	Protein Sequence N- → C-terminal
Doc124A	WNKAVIGDVNADGVVNISDYVLMKRYILRIIADFPADDDMWVGDVNGDNVINDIDCNYLKRYLLHMIREFPKNSYNSA
Doc124A m1	WNKAVIGDVNADGVVNISDYVLM**AA**YILRIIADFPADDDMWVGDVNGDNVINDIDCNYLKRYLLHMIREFPKNSYNSA
Doc124A m2	WNKAVIGDVNADGVVNISDYVLM**AA**YILRIIADFPADDDMWVGDVNGDNVINDIDCNYL**AA**YLLHMIREFPKNSYNSA
Doc124A m3	WNKAVIGDVNADGVVN**D**SDY**NY**MKRYILRIIADFPADDDMWVGDVNGDNVINDIDCNYLKRYLLHMIREFPKNSYNSA
Doc124A m4	WNKAVIGDVNADGVVN**D**SDY**NY**MKRYILRIIADFPADDDMWVGDVNGDNVIN**I**IDC**VL**LKRYLLHMIREFPKNSYNSA
Doc124A m5	WNKAVIGDVNADGVVNISDYVLMKRYILRIIADFPADDDMWVGDVNGDNVINDIDCNYLKRY**QQ**HMIREFPKNSYNSA
Doc124A m6	WNKAVIGDVNADGVVNISDYVLMKRY**QQ**RIIADFPADDDMWVGDVNGDNVINDIDCNYLKRY**QQ**HMIREFPKNSYNSA
Doc918	VVLNGDLNRNGIVNDEDYILLKNYLLRGNKLVIDLNVADVNKDGKVNSTDCLFLKKYILGLITI
Doc918 m1	VVLNGDLNRNGIVNDEDYILLKNYLLRGNKLVIDLNVADVNKDGKVN**DE**DCLFLK**N**YILGLITI
Doc918 m2	VVLNGDLNRNGIVN**ST**DYILLK**K**YLLRGNKLVIDLNVADVNKDGKVN**DE**DCLFLK**N**YILGLITI
Doc918 m3	VVLNGDLNRNGIVNDEDYILLKNYLLRGNKLVIDLNVADVNKDGKVN**QQ**DCLFLKKYILGLITI

Mutations are shown in bold.

3.1.2.2. Protein Purification

3.1.2.2.1. Cohesin-Dockerin Complexes

The Coh-Doc complexes CohOlpC-Doc124A and CohOlpA-Doc918 were expressed in *E. coli* Tuner cells, grown at 37 °C to OD_{600} 0.5. Recombinant protein expression was induced by adding isopropyl-β-D-thiogalactopyranoside (IPTG) to a final concentration of 0.2 mM and incubation for 16 h at 19 ºC. The recombinant proteins were purified by immobilized metal ion affinity chromatography (IMAC) using Sepharose columns charged with nickel (HisTrap™). Fractions containing the purified Coh-Doc complexes were buffer exchanged, using PD-10 Sephadex G-25M gel filtration columns (Amersham Pharmacia Biosciences), into 20 mM Tris-HCl buffer, pH 8.0, containing 2 mM $CaCl_2$. A further purification step by anionic exchange chromatography was performed by using a column loaded with Source 30Q matrix and a gradient elution of 0–1 M NaCl (GE Healthcare). Fractions containing the purified complex were then concentrated with Amicon 10-kDa molecular-mass centrifugal membranes and washed three times with 2 mM $CaCl_2$. The final protein concentration was adjusted to 21 g/l in 2 mM $CaCl_2$ for CohOlpC-Doc124A complex and was 16 g/l for CohOlpA-Doc918 complex.

3.1.2.2.2. Unbound Cohesins and Dockerins

Dockerins Doc124A, Doc918 and the respective mutant derivatives cloned in pET32a were expressed in *E. coli* Origami cells. CohOlpC and CohOlpA cloned in pRSETa vector were expressed in *E. coli* Tuner cells. Growth was performed at 37°C to mid-exponential phase

(OD_{600}=0.5) in Luria broth. Recombinant protein expression was induced with 1 mM (Origami) or 0.2 mM (Tuner) IPTG and incubation for 16 h at 19 ºC. The recombinant proteins were purified by IMAC as described above and buffer exchanged into 50 mM Na-Hepes buffer, pH 7.5, containing 2 mM $CaCl_2$ and then subjected to gel filtration using a HiLoad 16/60 Superdex 75 column (GE Healthcare) at a flow rate of 1 ml/min.

3.1.2.3. Isothermal Titration Calorimetry

ITC experiments were carried out essentially as described previously (Carvalho *et al.* 2007; Pinheiro *et al.* 2008), except that the titrations were at 55 ºC, and proteins were in 50 mM Na-HEPES buffer, pH 7.5, containing 2 mM $CaCl_2$. During titration the dockerin (40 μM) was stirred at 300 rev/min in the reaction cell, which was injected with 28 successive 10 μl aliquots of ligand comprising cohesin (180 μM) at 200 s intervals. Integrated heat effects, after correction for heats of dilution, were analysed by non-linear regression using a single site-binding model (Microcal ORIGIN, Version 5.0; Microcal Software). The fitted data yielded the association constant (K_A) and the enthalpy of binding (ΔH). Other thermodynamic parameters were calculated by using the standard thermodynamic equation: $\Delta RT\ln K_A = \Delta G = \Delta H - T\Delta S$.

3.1.2.4. Crystallization and Data Collection

Protein crystals were obtained using the hanging-drop, vapor-diffusion method. CohOlpC-Doc124A complex crystals grew in 2 M ammonium sulphate, pH 4.6 (condition 32 of Crystal Screen HR2-110 from Hampton Research) in drops with 7 g/liter of protein and were harvested after 5-7 weeks at 19 ºC. CohOlpA-Doc918 complex crystals grew in 0.2 M lithium sulphate, 10% w/v PEG 8000+, 10% w/v PEG 1000, pH 7.5 (condition 14 of Clear Strategy Screen I MD1-14 from Molecular Dimensions) in drops with 16 g/l of protein and were harvested after 3-5 weeks at 19 ºC. Crystals were cryo-cooled with paratone in liquid nitrogen prior to data collection at beamline ID14-2 at the European Synchrotron Radiation Facility (ESRF, Grenoble, France) at 100 K using an ADSC QUANTUM 4R CCD detector and at a wavelength of 0.9330 Å.

The two datasets were integrated using MOSFLM (Leslie & Powel, 2007) and scaled with SCALA (Evans, 2006) from the CCP4 suite (Winn *et al.*, 2011). CohOlpC-Doc124A crystals belong to $P3_221$ space group, with cell constants a = *b* = 90.76 Å, *c* = 135.07 Å and diffracted beyond 1.75 Å resolution. Matthews (Matthews, 1968) coefficient calculations suggested the presence of two molecules in the asymmetric unit (2.84 $Å^3$/Da and 56.8 % of solvent content). CohOlpA-Doc918 crystals belong to $I4_122$ space group, with cell constants *a* = *b* = 130.05 Å, *c* = 70.19 Å and diffracted beyond 1.95 Å resolution. Matthews coefficient calculations suggested the presence of one molecule in the asymmetric unit (3.0 $Å^3$/Da and 59 % of solvent content).

3.1.2.5. Structure Determination and Refinement

Structure determination of CohOlpC-Doc124A was based on two datasets that were processed in MOSFLM (Leslie & Powel, 2007) merged and combined with SORTMTZ from the CCP4 suite (Winn *et al.*, 2011) and scaled in SCALA (Evans, 2006). Phasing was performed by molecular replacement (MR) with the program BALBES (Long *et al.,* 2008) using a search model based on the following PDB ID code structures: 2ccl, 1aoh, 2vn6, 1nv8 and 1ixh. Density modification, together with non-crystallographic symmetry (NCS) averaging, was done with DM program from the CCP4 suite (Winn *et al.*, 2011). ARP/wARP Langer *et al.,* 2008) was used to automatically build the protein model. The asymmetric unit composition and crystal packing was adjusted, based on the results from the PISA server at the European Bioinformatics Institute (Krissinel & Henrick, 2007) and the PDBSET program from the CCP4 suite (Winn *et al.*, 2011). Model completion, editing and initial validation were carried out in COOT (Emsley, Lohkamp, Scott, & Cowtan, 2010). Initial restrained refinement of the molecular model was done using REFMAC 5.5 (Murshudov *et al.,*1997) and water molecules were added using COOT. The final cycles of refinement were done with the program PHENIX.REFINE from the PHENIX suite (Adams *et al.,* 2010). The two molecules in the asymmetric unit are arranged as a dimer of heterodimers, the later comprised of chains A/B and C/D. All atoms in the protein could be properly assigned and refined, apart from a few initial (first 6 and 7 residues from chains A and C, respectively, and the first 6 residues from chains B and D) and final residues (last 6 and 7 residues from chains A and C, respectively, and the last 6 and 8 residues from chains B and D, respectively) in the polypeptide chains. The final model also includes 461 water molecules and 4 calcium ions. R-work and R-free converged to 18.1 % and 21.5 %, respectively. Model assessment and validation were carried out by PHENIX.POLYGON (Urzhumtseva, Afonine, Adams, & Urzhumtsev, 2009) and MOLPROBITY (Chen *et al.,* 2010) from within PHENIX suite and PROCHECK (Laskowski, MacArthur, Moss, & Thornton, 1993). According to these programs the final model contains 99.5 % of the residues in mostly favored and allowed regions of the Ramachandran plot and 0.5 % of the residues in generously allowed regions of the plot.

Structure determination of CohOlpA-Doc918 was similarly done by MR with BALBES using structures with the PDB ID codes 2ccl and 2vn6 as models. Density modification with NCS was done with the PARROT (Zhang, Cowtan, & Main, 1997) program from the CCP4 suite. ARP/wARP was used to automatically build the protein model. Model completion, editing and initial validation were also carried out in COOT. The dockerin start model had to be manually rebuilt due to a mis-tracing error by ARP/wARP, originating from the presence of dockerin's two duplicated segments that share a striking sequence similarity and strong structural conservation. Refinement procedures were done as described for CohOlpC-Doc124A. R-work and R-free converged to 17.5 % and 20.6 %, respectively. Two chains were found in

the asymmetric unit, arranged as a dimer (A/B). Protein residues could be properly assigned and refined, apart from the first 5 and last 10 residues from chain A and the first 2 residues from chain B. The final model includes 224 water molecules and 2 calcium ions. Model assessment and validation using the above-mentioned tools produced a final model with 100 % of the residues in the mostly favored and allowed regions of the Ramachandran plot. Data collection and refinement details data for the two complete structures are summarized in Table 3.3.

Table 3.3| Data collection and refinement statistics

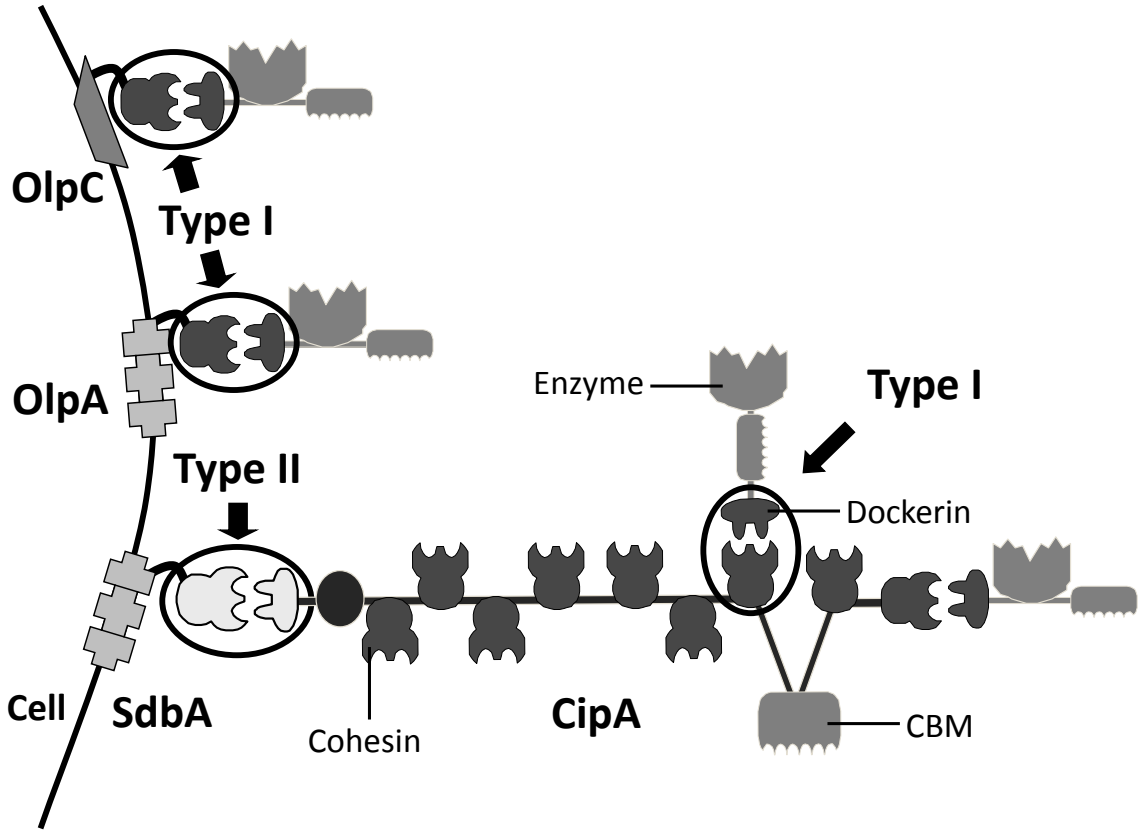
Crystal		***CohOlpC-Doc124A***	***CohOlpA-Doc918***
Space Group		$P3_221$	$I4_122$
Unit cell parameters			
a = b, c (Å)		90.76,135.07	130.05,70.19
α, β, γm(°)		90, 90, 120	90, 90, 90
Mathews parameter ($Å^3$/Da)		2.81	3.0
Data collection statistics			
X-ray source		ESRF, ID14-2	ESRF, ID14-2
Wavelength (Å)		0.933	0.933
No. of unique reflections		66027	22341
Resolution limits (Å)		78.60 – 1.75 (1.80 – 1.75)	91.96 – 1.95 (2.05 – 1.95)
Completeness (%)		99.9 (99.6)	100 (100)
Redundancy		14.9 (8.8)	14.4 (13.8)
Average I/σ(I)		18.7 (0.9)	23.3 (1.8)
Rsym (%)		0.090 (0.887)	0.080 (0.420)
Refinement statistics			
Resolution limits (Å)		39.30 - 1.80	32.79 - 1.95
R-work		0.181	0.175
R-free		0.215	0.206
No. protein residues in the asymmetric unit		456	205
No. water molecules in the asymmetric unit		461	178
No. atoms in the asymmetric unit		7844	1824
rmsd bond length (Å)		0.014	0.014
rmsd bond angles (°)		1.296	1.314
Average temperature factor ($Å^2$)	protein main chain	26.4	28.2
	protein side chain	34.5	33.8
	water molecules	41.9	37.5
	ligands	84.2	50.6
	calcium	24.5	20.0
Ramachandran plot	residues in most favored regions (%)	91.1	89.4
	residues in additionally allowed regions (%)	8.4	10.6
	residues in generously allowed regions (%)	0.5	0
PDB codes		4dh2	3ul4

3.1.3. Results and Discussion

3.1.3.1. Expression and Crystallization of novel Coh-Doc complexes

In a previous study (Pinheiro *et al.*, 2009) *C. thermocellum* type I dockerins were shown to bind to the nine type I cohesins of CipA and the single cohesin domains of the cell surface proteins OlpA or OlpC (Figure 3.1). The majority of *C. thermocellum* dockerins (68 out of 72), exemplified by the well characterized dockerin of Xyn10B (Carvalho *et al.,* 2003; *Carvalho et al.,* 2007), display a distinctive internal symmetry which is compatible with a dual binding mode. These dockerins display preferential recognition for OlpA and CipA cohesins. In contrast, two *C. thermocellum* dockerins, from the protein of unknown function, Cthe_0258, and the recently described cellulase, Cel124A (Brás *et al.,* 2011), bind preferentially to the cell envelop cohesin of OlpC (Pinheiro *et al.,* 2009). A third dockerin, of bi-functional cellulase Cel9D-Cel44A, displays two cohesin-binding interfaces with different specificities; the dockerin can interact with *C. cellulolyticum* cohesins through the N-terminal interface and with *C. thermocellum* counterparts through the C-terminal binding site. A fourth dockerin, from the protein of unknown function Cthe_0918, binds equally well to CipA and OlpA cohesins and displays a lower affinity to the cohesin of OlpC. The primary sequences of these four dockerins lack the distinctive symmetry at the binding interfaces, which may explain, at least for dockerins of Cthe_0258, Cel124A and Cel9D-Cel44A, the observed differences in ligand specificity (Pinheiro *et al.,* 2009). We are now in the position to explore the full range of structural determinants of ligand specificity in type I Coh-Doc complexes of *C. thermocellum*. Dockerin dual binding mode does not favor the crystallization of protein complexes and the usual approach used to study the Coh-Doc interaction involves the inactivation of one of the cohesin-binding interfaces through site-directed mutagenesis (Carvalho *et al.,* 2007; Pinheiro *et al.,* 2008). Since the four unusual dockerins described above do not seem to present more than one binding interface, wild-type proteins were used for these studies. Here, we have used established strategies for the production of Coh-Doc complexes that involves the co-expression of both proteins in *E. coli* cells (Brás *et al.,* 2012).

Figure 3.1| *C. thermocellum* cellulosome.

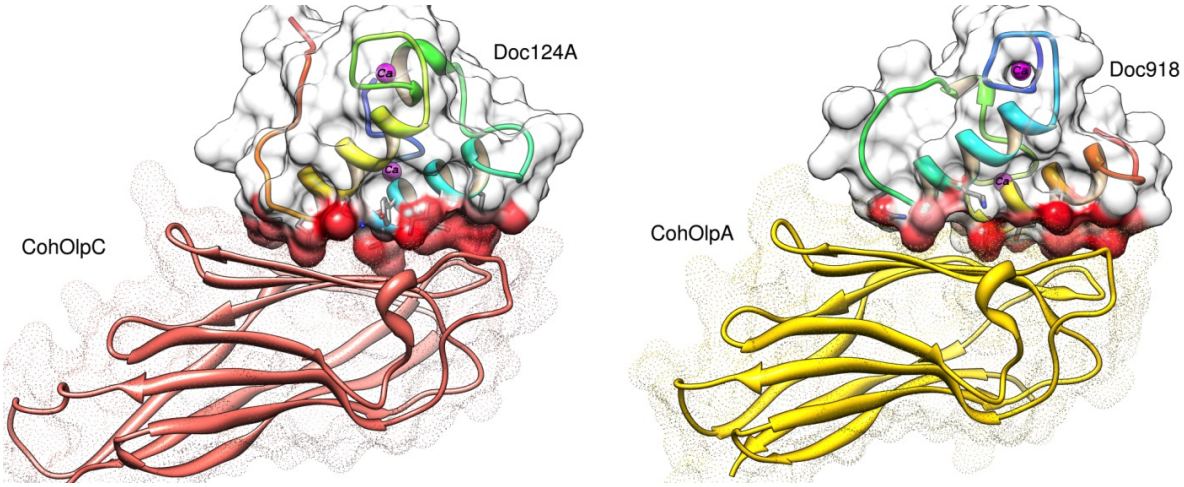


C. thermocellum scaffoldin (CipA) contains nine type I cohesins and thus organizes a multi-enzyme complex that incorporates nine enzymes. The C-terminal type II dockerin of CipA binds specifically to type II cohesin domains found in cell-surface proteins. Individual enzymes may also adhere directly to the bacterium cell envelop by binding to the single type I cohesins found in OlpA and OlpC.

3.1.3.2. Structure of type I Coh-Doc complexes

The structures of OlpA type I cohesin bound to the dockerin of protein Cthe_0918 (CohOlpA-Doc918) and of the OlpC type I cohesin in complex with the dockerin of Cel124A (CohOlpC-Doc124A) were solved to 1.95 Å and 1.75 Å resolution, respectively (Figure 3.2). In *C. thermocellum*, OlpA and OlpC cohesins are the only two type I cohesins that do not belong to CipA and show significant deviations in the putative residues that participate in dockerin recognition, when compared with the 9 highly homologous cohesins of CipA (Pinheiro *et al.,* 2009) (for details see Figure S3.2 in annex).

Figure 3.2| Structure of novel type-I Cohesin-Dockerin complexes, CohOlpC-Doc124A and CohOlpA-Doc918.



The dockerin structure is rainbow colored (N-term:blue, C-term:red). Calcium ions are magenta spheres. The cohesin surface is depicted as dots while the dockerin surface is white solid with the main contact surface area highlighted in red. CohOlpA-Doc918 shows a C-terminal (helix-3) dominated Coh-Doc interface while CohOlpC-Doc124A reveals a N-terminal (helix-1) dominated Coh-Doc interface.

3.1.3.2.1. Structure of OlpA and OlpC type I Cohesins

The 146-residue OlpA cohesin in complex with its cognate dockerin displays an elongated nine-stranded flattened β-sandwich structure, defined by two β-sheets (A and B) in a classical jelly-roll topology (see Figure S3.3 in annex) (Carvalho *et al.*, 2003). β-sheet A is comprised of the following β-strands (and respective residues): 4 (51-58), 7 (99-108), 2 (22-30), 1 (9-17) and 9 (138-146); while β-sheet B includes β-strands: 5 (69-74), 6 (79-85), 3 (37-45) and 8 (115-129). β-strand 8 is the longest and partly integrates both β-sheets. β-sheet B forms a distinctive planar plateau amid a molecule with a flattened and overall curved cylindrical shape (Figure S3.3_C). The OlpC cohesin domain with 159 amino-acid residues exhibits the same type I cohesin architecture, where β-sheet A is comprised of the same β-strands (and respective residues): 4 (60-67), 7 (108-117), 2 (31-39), 1 (15-27) and 9 (149-161); while β-sheet B includes β-strands: 5 (78-83), 6 (88-94), 3 (48-54) and 8 (129-142), the latter also contributing to both β-sheets. Both cohesins contain a highly hydrophobic core (Figure S3.3_B).

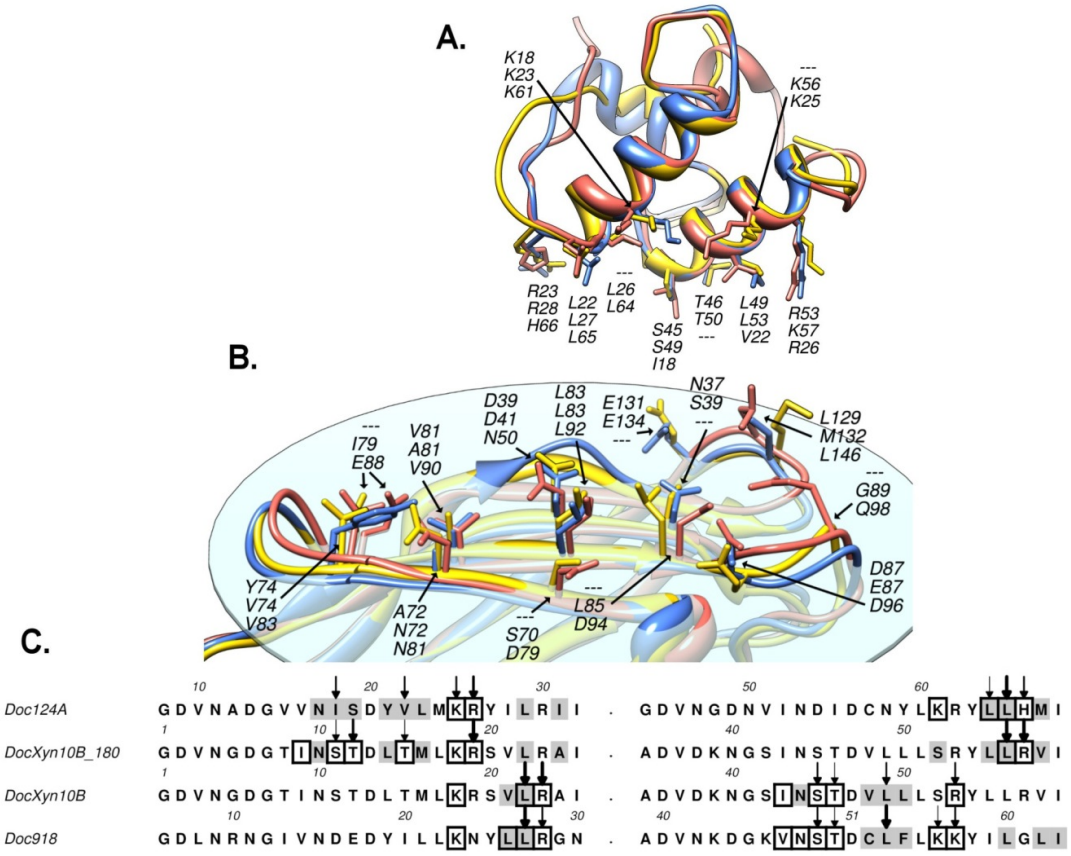
Both structures reveal a striking similarity (Figure S3.3_A). In comparison with the second cohesin of CipA (CohCipA2), a structure superposition with CohOlpA has an r.m.s.d. of 1.04 Å (between 136 C_α pairs for a 35.3 % of sequence identity) and 1.18 Å with CohOlpC (133 C_α pairs, 38.3 % sequence identity). The two novel cohesins superpose with each other with a r.m.s.d. of 1.14 Å (139 C_α pairs, 34.5 % sequence identity). However, by comparing the protein volume encircled by their molecular surface, CohOlpA and CohOlpC have a 4 % and 13 % increase, respectively, over CohCipA2, the latter having a volume of 165,100 $Å^3$. Noteworthy structural divergences occur between β-strands 4 and 5 (which include a small α-

helix) where both CohOlpA and CohOlpC have a shorter loop than CohCipA2, and on the loop between β-strands 7 and 8 that, compared to CohCipA2, is slightly longer in CohOlpA and considerably larger in CohOlpC, increasing the main longitudinal axis length of this 2 proteins by roughly 2 Å and 10 Å, respectively (Figure S3.3_A). The β-sheet B interface area, evaluated on the basis of its solvent-accessible area when in complex with their cognate dockerin (PDBePISA) (Krissinel & Henrick, 2007), was: 686 $Å^3$ for CohCipA2, 803 $Å^3$ for CohOlpA and 729 $Å^3$ for CohOlpC.

3.1.3.2.2. Structure of Type I Dockerins

The structure of dockerins of Cthe_0918 and Cel124A, here termed Doc918 and Doc124A, respectively, are organized in two α-helices, arranged in an antiparallel orientation (N-terminal or helix-1 and C-terminal or helix-3) connected through an extended loop displaying a small helix (helix-2) (Figure 3.3). In Doc918, helix-1 is composed of residues 15 to 27, helix-2 extends between residues 35-39 and helix-3 from residues 48-60. In Doc124A the respective residues are: helix-1 (17-29), helix-2 (39-45) and helix-3 (53-65). Helix-2 constitutes a linker between the other two helices both of which provide the two putative cohesin binding interfaces. In fact the linker region, limited by the distal end of helix-1 and the C-terminus of helix-2, contains a large amount of the structural variability found among the core Cα trace of these dockerins. The linker in Doc918 is less structured than in DocXyn10B, presenting a single turn on its α-helix, similarly to a type I *C. cellulolyticum* dockerin (Pinheiro *et al.,* 2008), albeit the latter has a much smaller linker. The internal sequence duplication and near-perfect 2-fold symmetry was quantified by an internal superposition between helices-1 and -3 within each structure. Doc918 shows a r.m.s.d of 0.57 Å for 23 C_α pairs and in Doc124A both segments overlap almost as well, with an r.m.s.d of 0.66 Å for 26 C_α pairs. Lack of conservation in the key contacting residues when the two putative binding surfaces are compared would preclude a dual binding mode, which is discussed below. Both dockerins contain two Ca^{2+} ions coordinated by several residues in canonical EF-hand loop motifs. The coordination of the two calcium ions is similar to the metal ions observed in the type I dockerins of *C. thermocellum* and *C. cellulolyticum* in complex with their cognate protein partners.

Figure 3.3| Structure superposition and important contact residues.



A) Dockerin superposition between Xyn10B (blue), Doc918 (yellow) and a 180°-rotated Doc124A (salmon). Residues with a significant contribution to the Coh-Doc contact surface area are shown in stick representation and numbered according to the PDB records 1ohz, 3ul4 and 4dh2. **B)** Cohesin superposition between CohCipA2 (blue), CohOlpA (yellow) and a CohOlpC (salmon). Important residues to the Coh-Doc contact interface are shown above the β-sheet B plane, in stick representation and numbered according to the respective PDBs. **C)** Dockerin sequence alignment and interacting residues. Residues with a significant contribution to the Coh-Doc contact surface area are marked with a top variable-width small arrow. Residues involved in hydrophobic interactions are shown with a grey background while a box highlights residues with polar interactions. DocXyn10B_180 denotes a 180° binding interface rotation.

3.1.3.2.3. Novel Type I Coh-Doc Complex Interfaces

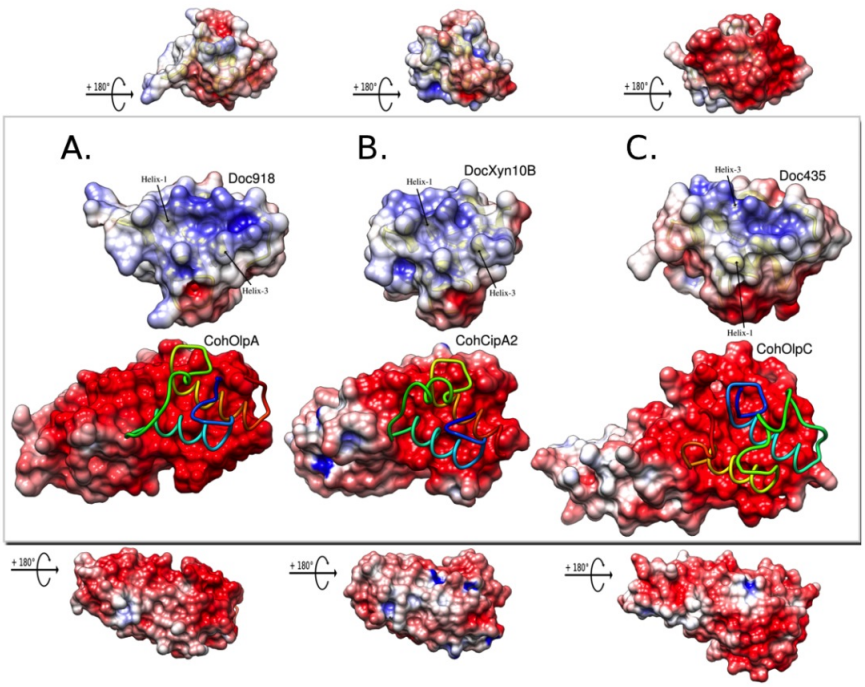
In contrast with what was previously observed for other type I complexes, the dockerins described here in complex with their protein partners, seem to present a single binding mode. Thus, in the CohOlpA-Doc918 complex, binding is dominated by the Doc918 C-terminal helix (Figure 3.2). In contrast, in CohOlpC-Doc124A complex, binding is orchestrated by the dockerin N-terminal (Figure 3.2). In these two novel Coh-Doc structures, the complex interface has a significant hydrophobic nature. Using the solvation free energy gain at complexation, calculated by PDBePISA ($\Delta^{i}G$ in kcal/mole (Krissinel & Henrick 2007)), the CohOlpA-Doc918 interaction is more hydrophobic (-10.6 kcal/mol) than that of CohOlpC-Doc124A (-7.7 kcal/mol) while both complexes exceed the CohCipA-DocXyn10B value of -6.4 kcal/mol. However, the negative values upon binding are less significant than those of

highly hydrophobic *C. cellulolyticum* type I complex (PDB ID 2vn6) with -14.9 kcal/mol. These differences reflect the numerous hydrophobic residues, involved in the Coh-Doc complex interface. Thus, the numbers of cohesin and dockerin hydrophobic residues implicated in the interface of the CohOlpA-Doc918 are greater than in the CohCipA2-DocXyn10B complex. Although the hydrophobic contact network of CohOlpC-Doc124A is also extensive, the hydrophobic residues that contribute to the heterodimer interface are contributed primarily by Doc124A. The major hydrophobic contact residues located at the surface of cohesins CipA2, OlpA, and OlpC include a completely conserved leucine (Leu83, Leu83, and Leu92, respectively), which is assisted by upstream hydrophobic residues Val81, Ala81, and Val90 and downstream by Ala85, Leu85, and a divergent Asp94 in OlpC, respectively. Other important contributors correspond to Leu129, Met132, and Leu146, respectively. With respect to the dockerins Xyn10B-α3/Xyn10B-α1, 918, and 124A, the major hydrophobic contact residues are Leu22/Leu56, Leu27, and Leu65, respectively, at position 22 of the less interacting binding interface. In addition, in position 15 of the dominating interface, residues Leu49/Thr15, Leu53, and Val22 make a significant contribution to cohesin recognition. The above mentioned conserved leucine located at the surface of the three cohesins is part of an important hydrophobic pocket formed in CohCipA2 by Ala72, Tyr74, Val81, and Leu83, which is occupied by Leu22 or Leu56 from DocXyn10B in the two possible binding modes, respectively. Using the same relative structural positioning order, for CohOlpA, we find Asn72, Ala81, and Leu83, which accommodate Leu27 from Doc918. As for CohOlpC, residues Asn81, Val90, and Leu92 form a hydrophobic pocket that is occupied by the equivalent Doc124A residue, Leu65, found in the opposite C-terminal interface. The heterodimer interfaces are assisted by a network of direct and bridged hydrogen bonds and salt bridge interactions (described in detail in Table 3.4). Compared with DocXyn10B (α3) in a similar C-terminal binding conformation (Carvalho *et al.,* 2003), Doc918 reveals a more imbalanced distribution of polar bonds, favoring helix-3 residues. Although the Ser/Thr dyads of both complexes share an equivalent contribution, the main difference occurs at the Lys56/Lys57 pair of Doc918 that contribute with one salt bridge and two direct H-bonds, whereas in DocXyn10B (α3), the equivalent Ser52 makes no polar bonds, and Arg53 establishes a single salt bridge. Again, in comparison with the N-terminal bound DocXyn10B (α1), Doc124A reveals some striking differences with respect to the relevant Ser-Thr pair, which is replaced by a divergent Ile18-Ser19 motif. In Doc124A the N-terminal binding face interacts with CohOlpC, through significant hydrophobic contacts. The only direct polar interactions mediated by helix-1 occur via positions 18 and 19 (Lys25-Arg26), through six direct bonds (two salt bridges from Lys25 and four H-bonds from Arg26) and a couple of water bridged H-bonds involving Ile18. In contrast, the N-terminally based interface of DocXyn10B reveals a hydrogen bond network around, and dominated by the conserved Ser-Thr pair and also some involvement of residues 18 and 19. Also in contrast to DocXyn10B

(α1), Doc124A presents in the opposite interface (α3) a stronger polar contribution participated in by four residues, Lys61 (one salt bridge), Leu64 (one H-bond), Leu65 (one H-bond), and His66 (one salt bridge), whereas in DocXyn10B, Leu56 and mainly Arg57 make polar contacts with the CipA cohesin (Figure. 3.3, *A* and *C*; detailed contacts in supplemental Figures S3.1 and S3.3). Extending the comparison of Doc124A to *C. cellulolyticum* type I complex (2vn5/2vn6) in an analogous binding conformation, the major difference consists of a much subdued polar interaction network found in the latter, especially at the α3 interface, where only positions 22 and 23 reveal direct contacts (Pinheiro *et al.,* 2009). The cohesin-interacting residues can be grouped into three regions corresponding to β-strands β3, β-strands β5/ β6, and the loop between β-strands β8 and β9 (Figure 3.3*B*; detailed contacts in supplemental Figure S3.2). Around the β3 region, the important interactions are quite similar among CohCipA2/CohOlpA, because equivalent residues Asn37/Ser39 and Asp39/Asp41, respectively, establish relevant polar contacts with the dockerin Ser/Thr pair. Conversely, the equivalent CohOlpC residue Ser48 does not display any polar contacts, and Asn50, equivalent to CohCipA2 Asp39, establishes a single H-bond with the dockerin. In the β5/ β6 cohesin region, notable differences between CohCipA2, CohOlpA, and CohOlpC occur, respectively, at Arg77, Asp77, and Asp86 residues; Arg77 makes an H-bond with its target dockerin, whereas the equivalent acidic residues of the other two cohesins are not implicated on the interface. In the β8-loop-β9 region, the corresponding residues Asn127, Asn130, and Phe144 in CohCipA2, OlpA, and OlpC, respectively, reveal some differences in their capacity to recognize the dockerin protein partner. In a helix-3-dominated binding, CohCipA2-Asn127 does not exhibit any contacts with its dockerin, whereas CohOlpA-Asn130 makes two bridged H-bonds. However, in a helix-1-dominated binding, CohCipA2 uses its Asn127 to make two H-bonds with Doc-Arg19, whereas in CohOlpC, the backbone of Phe144 establishes two H-bonds with Doc-Arg26. In addition, in the Glu131/Glu134/Pro148 position of the cohesins, both acidic residues from CohCipA2 and CohOlpA form an H-bond with the critical threonine found at position 12 of the dockerin, whereas CohOlpC-Pro148 does not contribute to dockerin recognition. Further analysis of the differences between the canonical type I cohesin and this work on cell-bound cellulosomal cohesins was based on the predicted negative hydrogen bond accepting regions in an electrostatic surface potential evaluation using the Poisson-Boltzmann electrostatics calculation on the PDB2PQR server (Dolinsky *et al*., 2004) and visualization of the results in UCSF Chimera (Pettersen *et al.,* 2004) (Figure 3.4). As reported previously (Carvalho *et al.,* 2005), cohesins are strikingly negatively charged in the binding interface plateau, whereas dockerins present a suitable complementary positive-to-neutral surface. Compared with CohOlpC and CohCipA2, CohOlpA shows an elongated polar region that extends beyond the binding interface. As described for the typeII cohesin of SdbA (Carvalho *et al.,* 2005), the opposite cohesin surfaces in CohOlpC and CohOlpA are more positively/neutrally charged, which was

suggested to be important to promote a tighter interaction of cell surface cohesins to the negatively charged peptidoglycan layer. Analysis of the type I Coh-Doc interfaces provides significant insights into the previously described tight binding of Doc124A to CohOlpC, in comparison with the lower affinity displayed by this dockerin toward the cohesins of CipA and OlpA (Pinheiro *et al.,* 2009) The hydrophobic nature of Ile18 at the critical position 11 of the Doc124A interface, establishes a strong network of apolar contacts with CohOlpC, namely with Asn50, Asn140, and Cys142. These CohOlpC pivotal residues are replaced by an aspartate (position 50) and by small residues, namely glycine or alanine, at the other two positions in CohCipA and CohOlpA cohesins. The aspartate residue, equivalent to CohOlpC-Asn50, found in both OlpA and CipA cohesins is also highly relevant for the recognition of typical type I dockerins because, together with Asn37, it establishes conserved hydrogen bonds with the canonical serine residues usually found at position 11. In CohOlpC, the latter residues are replaced with Ser48 and Asn50, respectively, whose side chains were found more than 4Å apart and thus presumably unavailable for H-bond formation. In addition, dockerin position 12 of the binding interface makes some relevant polar contacts with the mentioned residue of CohCipA2-Asn37and also Glu131. Again, in CohOlpC, the equivalent Ser48 side chain orientation is unsuitable for those contacts, and the Pro148, which substitutes Glu131, is manifestly non-reactive.

Figure 3.4| Electrostatic surface potential for the Coh-Doc molecules.



In each panel the central image shows the top cohesin binding plateau with a licorice model of the bound dockerin in N-to-C rainbow color ramped style. Above, there is a view of the molecular dockerin's binding surface. The top and bottom images are smaller scaled and orthogonal representations, respectively, of the dockerin and cohesin. **A)** CohOlpA and Doc918; **B)** CohCipA2 and DocXyn10B; **C)** CohOlpC and Doc124A. The electrostatic potential is contoured in UCSF Chimera from -6 (red) to +6 (blue) (arbitrary Chimera units).

Table 3.4| Network of polar interactions in Novel Type-I Coh-Doc complex interfaces.

	Dockerin Doc124A	**HB: H Bond SB: Salt bridge Wat: H_2O bridge**	**Cohesin OlpC**
Helix-1	Lys B25 NZ	SB	Asp A94 OD2
	Lys B25 NZ	SB	Asp A96 OD1
	Arg B26 NE	HB	Gln A98 OE1
	Arg B26 NH2	HB	Gln A98 OE1
	Arg B26 NH1	HB	Phe A144 O
	Arg B26 NH2	HB	Phe A144 O
Helix-3	Lys B61 NZ	SB	Asp A79 OD2
	Leu B64 O	HB	Asn A50 ND2
	Leu B65 O	HB	Asn A81 ND2
	His B66 ND1	SB	Glu A88 OE1
	Dockerin Doc918	**HB: H Bond SB: Salt bridge Wat: H_2O bridge**	**Cohesin OlpA**
Helix-1	Lys B23 NZ	Wat A193	Ser A70 OG
	Leu B26 O	Wat A179	Val A82 O
	Leu B27 O	HB	Asn A72 ND2
	Arg B28 NH1	Wat A188	Asp A41 OD2
Helix-3	Val B47 O	Wat A199	Asp A41 OD2
	Asn B48 ND2	Wat A189	Glu A134 OE2
	Ser B49 N	HB	Asp A41 OD1
	Ser B49 OG	HB	Asp A41 OD1
	Ser B49 OG	HB	Ala A126 O
	Ser B49 OG	Wat A167	Ser A39 OG
	Thr B50 OG1	HB	Glu A134 OE2
	Thr B50 OG1	Wat A167	Ser A39 OG
	Lys B56 NZ	HB	Glu A87 O
	Lys B56 NZ	SB	Glu A87 OE2
	Lys B57 NZ	HB	Gly A89 O
	Lys B57 NZ	Wat A196	Asn A130 OD1
	Lys B57 NZ	Wat A207	Asn A130 OD1

3.1.3.3. Probing the Importance of Contact Residues in Dockerins

To identify the dockerin residues that are involved in cohesin recognition, a mutagenesis study based on previously described type I complex structures was implemented. Previous data suggest that the implications of single changes in dockerin activity may be relatively modest, so the strategy used here involved the change of particular groups of residues that are believed to play a cooperative role in cohesin recognition (Carvalho *et al.,* 2003; Carvalho *et al.,* 2007; Pinheiro *et al.,* 2008; Pinheiro *et al.,* 2009).

3.1.3.3.1. CohOlpA-Doc918 Complex

Site-directed mutagenesis and ITC data of the CohOlpA-Doc918 complex (Figure 3.5 and Table 3.5), show that the replication of the relevant residue environment from dockerin helix-1 into the C-terminal helix-3, which dominates ligand recognition (mutant Doc918_m1: Ser49Asp, Thr50Glu and Lys57Asn) precluded any binding, which reinforces a vital role for the Ser-Thr motif, similarly to what is acknowledged for the canonical type I Coh-Doc interaction (Carvalho *et al.,* 2003). The drastic decrease in affinity obtained with mutant Doc918_m3 (Ser49Gln, Thr50Gln) also support a major role for the dockerin ST motif. Its importance was also predicted from the network of interacting residue contacts, both in terms of their contribution to the total contact surface area and also from the formation of important polar contacts. The Lys57Asn mutation also emphasizes the relative importance of a basic residue such as lysine or an arginine residue (akin to the Arg53 from DocXyn10B) in this position for efficient binding.

The Doc918_m2 mutant design provided additional insights into the pivotal residues mediating cohesin recognition. Essentially, using the inactive dockerin, Doc918_m1, an attempt was made to force the alternate helix-1 binding mode. Thus, the non-functional N-terminal helix of Doc918_m1 was engineered to restore an N-terminal cohesin-binding interface by introducing the three pivotal residues identified in Helix-3 in the corresponding positions of helix-1. ITC data showed that, although with a ten-fold reduction in affinity, this strategy indeed allowed binding through helix-1. In CohOlpC-Doc124A and CohCipA2-DocXyn10B (i.e. the Ser45Ala/Thr46Ala mutant of DocXyn10B) (Carvalho *et al.* 2007) helix-1 dominated binding interfaces, there is a bulky, positively charged residue in helix-3 (His66 and Arg57, respectively) which provide polar and hydrophobic interactions to the interface, but which is replaced by a Gly61 in Doc918 when binding was engineered at the N-terminal face (Figure 3.3). This divergent substitution could thus contribute to the reduced affinity displayed by the Doc918_m2 mutant. Overall the data presented here confirm that Doc918 presents a single protein-binding interface that is dominated by the C-terminal helix and where Ser49, Thr50 and Lys57 dominate cohesin recognition.

Table 3.5| Thermodynamics of type I dockerin-cohesin interactions.

Cohesin	Dockerin	K_a M^{-1}	ΔG^o kcal mol^{-1}	ΔH^o kcal mol^{-1}	$T\Delta S^o$ kcal mol^{-1}
OlpC	Doc124A	3.07E7 ±3.57E6	-11.24±0.18	-40.70±0.18	-29.46
	Doc124A _m1	6.91E4 ±1.10E4	-7.16±0.47	-9.08±0.47	-1.92
	Doc124A _m2	*Nd*	*Nd*	*Nd*	*Nd*
	Doc124A _m3	7.40E6 ±3.43E5	-10.44±0.09	-28.29±0.09	-17.85
	Doc124A _m4	6.12E7 ±7.20E6	-11.84±0.04	-28.90±0.04	-17.06
	Doc124A _m5	1.68E6 ±1.50E5	-9.34±0.34	-32.38±0.34	-23.04
	Doc124A _m6	1.50E5 ±1.39E4	-7.77±2.80	-39.86±2.80	-32.09
OlpA	Doc918	8.77E7 ±1.31E7	-11.80±0.22	-47.89±0.22	-36.09
	Doc918_m1	*Nd*	*Nd*	*Nd*	*Nd*
	Doc918_m2	6.07E6 ±5.42E5	-10.18±0.22	-30.49±0.22	-20.31
	Doc918_m3	2.47E6 ±3.52E5	-9.62±0.39	-27.24±0.39	-17.62

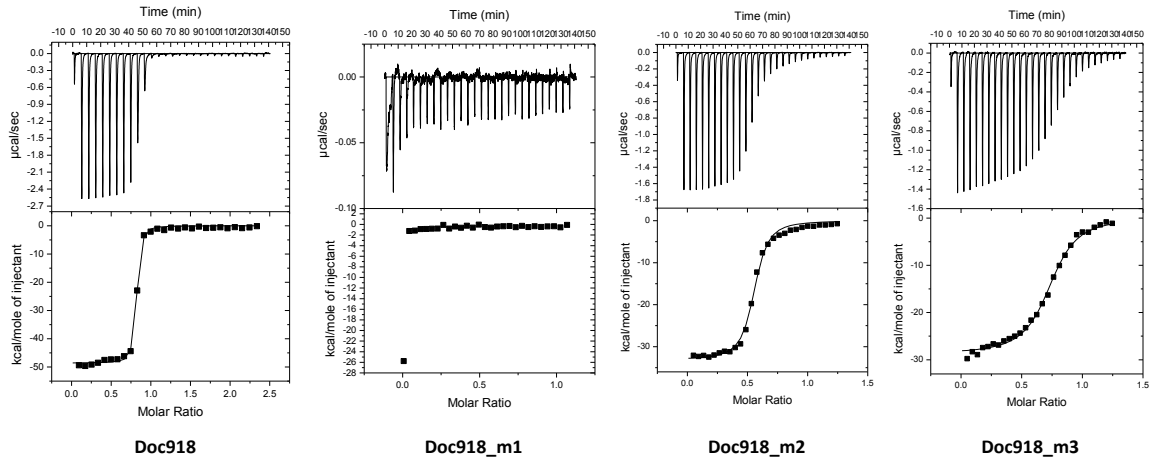
Thermodynamic parameters were determined at 328.15 K. *Nd* means that the values were too low to be determined.

Figure 3.5| Alignment of Doc918 primary sequence and its mutant derivatives (A) and examples of ITC experiments (B).

A)

```
Doc918      GDLNRNGIVNDEDYILLKNYLLRGN···ADVNKDGKVNSTDCLFLKKYILGLITI
Doc918_m1   GDLNRNGIVNDEDYILLKNYLLRGN···ADVNKDGKVNDEDCLFLKNYILGLITI
Doc918_m2   GDLNRNGIVNSTDYILLKKYLLRGN···ADVNKDGKVNDEDCLFLKNYILGLITI
Doc918_m3   GDLNRNGIVNDEDYILLKNYLLRGN···ADVNKDGKVNQQDCLFLKKYILGLITI
```

B)



A) Alignment of Doc918 with its mutant derivatives. Mutations are highlighted in grey. **B)** Examples of the ITC experiments with CohOlpA and Doc918 and its mutant derivatives expressing different affinities. The upper parts of each panel show the raw heats of binding, whereas the lower parts are the integrated heats after correction for heat dilution. The curve represents the best fit to a single-site binding model.

3.1.3.3.2. CohOlpC-Doc124A Complex

As described above, the Doc124A Lys25-Arg26 pair dominates the polar binding network with OlpC cohesin, whereas Lys61 makes an important salt bridge with Asp79 present at the surface of the cohesin. Thus, Doc124A mutants m1 and m2 were used to explore the importance of helix-1 Lys25-Arg26 and helix-3 Lys61-Arg62 pairs, by mutating them separately (m1) or simultaneously (m2) to Ala (Figure 3.6 and Table 3.5). As expected, based on these multiple polar contacts, the lesion in helix-1 (m1) caused an ~400-fold decrease in affinity. In addition, the additive effect of mutating the two basic pairs at helix-1 and helix-3 simultaneously (m2) led to complete loss in cohesin recognition, confirming the importance of Lys61 in heterodimer formation. Thus, the basic pair at helix-1 plays a key role in cohesin recognition, and the massive reduction in affinity suggests a single binding mode for Doc124A. However, because the helix-3 Lys61-Arg62 pair is in a position symmetry-related to that of Lys25-Arg26 in helix-1, it is also possible that, following a 180° rotation of the dockerin, these latter residues could participate in a lower affinity cohesin recognition mediated by helix-3. Under these circumstances, the lower affinity of m1 would result from substitution of the critical Ile by an Asp at position 11 and by the loss of a putative Lys25-mediated salt bridge at the other helix. Data presented above suggest that Doc124A could eventually present two cohesin binding interfaces expressing different affinities. To explore this possibility, Doc124A Ile18, Val22, and Leu23, which are part of the hydrophobic platform of the helix-1 binding interface, were mutated to replicate their symmetry related counterparts in helix-3 (m3) (Figure 3.6 and Table 3.5). The data revealed that these mutations lead to a reduction in the capacity of Doc124A to bind its cohesin partner. TheDoc124A_m4 mutant introduces into the m3 background, in which helix-1 binding is reduced, the mutations Asp54Ile, Asn58Val, and Tyr59Leu, with the intention of promoting a reversal in binding through the C-terminal helix (Figure 3.6 and Table 3.5). ITC results show an 8-fold increase in affinity over the m3 mutant, similar to the wild type dockerin, suggesting that although a dual binding mode is not feasible in the native form of Doc124A, in the m4 mutant, binding is probably dominated by the C-terminal interface. Thus, overall, the data suggest that Doc124A presents a single binding mode driven by helix-1. The importance of the hydrophobic network established between Doc124A and OlpC was further explored in the mutant m5, which investigated the role of a second residue pair, Leu64-Leu65, in the interactions established with the cohesin (Figure 3.6 and Table 3.5). As described above, the Doc124A dockerin presents a symmetry-related pair at helix-1, Ile28-Leu29, which could be involved in a similar interaction if binding was mediated by the helix-3 lower affinity interface. The importance of this pair was explored in m6. The knock-out of the Leu64-Leu65helix-3 pair (m5) induced a 10-fold decrease in affinity, confirming the relevance of these residues in binding the cohesin when helix-1 is the dominant binding face. Indeed, it is reason able to assume that the loss of Leu64 and Leu65 in m5 reorientates the major binding face to helix-

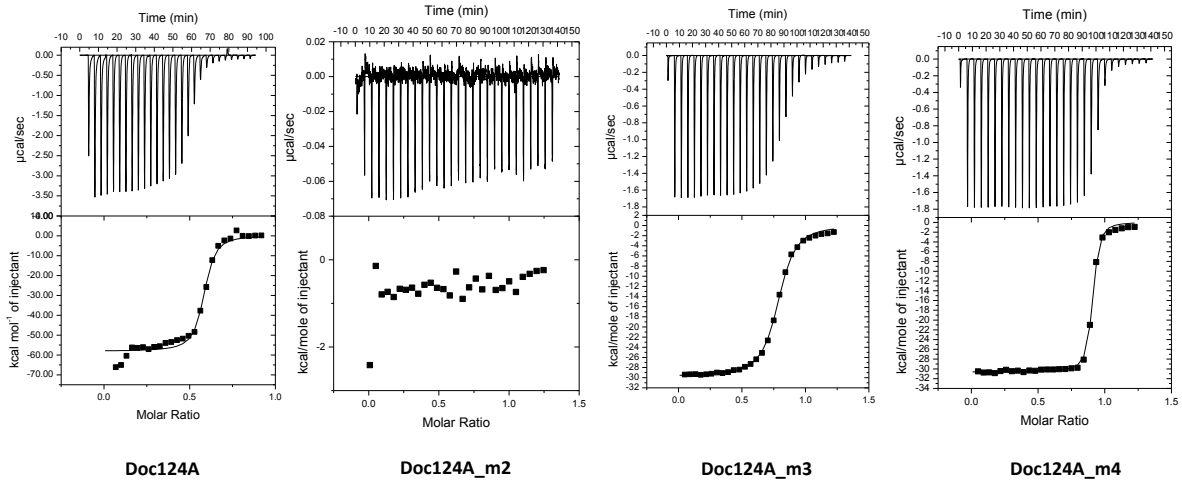
3. Consistent with this view is the further reduction in affinity by the concurrent mutation of the proximal helix-1 residue pair (m6).

Figure 3.6| Alignment of Doc124A primary sequence and its mutant derivatives (A) and examples of ITC experiments (B).

A)

```
Doc124A      GDVNADGVVNISDYVLMKRYILRII   GDVNGDNVINDIDCNYLKRYLLHMI
Doc124A_m1   GDVNADGVVNISDYVLMAAYILRII   GDVNGDNVINDIDCNYLKRYLLHMI
Doc124A_m2   GDVNADGVVNISDYVLMAAYILRII   GDVNGDNVINDIDCNYLAAYLLHMI
Doc124A_m3   GDVNADGVVNDSDYNYMKRYILRII   GDVNGDNVINDIDCNYLKRYLLHMI
Doc124A_m4   GDVNADGVVNDSDYNYMKRYILRII   GDVNGDNVINIIDCVLLKRYLLHMI
Doc124A_m5   GDVNADGVVNISDYVLMKRYILRII   GDVNGDNVINDIDCNYLKRYQQHMI
Doc124A_m6   GDVNADGVVNISDYVLMKRYQQRII   GDVNGDNVINDIDCNYLKRYQQHMI
```

B)



A) Alignment of Doc124A with its mutant derivatives. Mutations are highlighted in grey. **B)** Examples of the ITC experiments with CohOlpC and Doc124A and its mutant derivatives expressing different affinities. The upper parts of each panel show the raw heats of binding, whereas the lower parts are the integrated heats after correction for heat dilution. The curve represents the best fit to a single-site binding model**.**

3.1.4. Conclusions

The structure of two type I Coh-Doc complexes presented here show that unlike the large majority of *C. thermocellum* dockerins, the dockerins of cellulase Cel124A and of the unknown protein Cthe_0918 present a single cohesin-binding surface. Although Doc124A C-terminal binding face also displays the capacity to bind the cognate cohesin, the significantly lower affinity presented by the alternate binding favors a single-binding mode. The structures of the two dockerins were solved in complex with the two unique cell surface type I cohesins of *C. thermocellum*, OlpA and OlpC, which direct plant cell wall hydrolytic enzymes directly to the cell surface. A recent study revealed that cellulosomes act in synergy with enzymes located at the bacterium cell envelop, which include the abundant Cel124A endo-cellulase that targets cellulose crystalline-amorphous junctions (Lu *et al.,* 2006). The fact that high quality crystals for both complexes were obtained using wild type dockerins was an initial

good indication that these dockerins present essentially a single interacting surface. The structures of the two complexes revealed that the critical positions 11 and 12 of the dockerin non-interacting interface are occupied predominantly with acidic residues (Glu and Asp). Acidic residues are not suitable for interacting with the highly negatively charged cohesin platform. Site-directed mutagenesis data demonstrate the importance of the Ser/Ile-Thr motif residue at positions 11 and 12, and the Lys/Lys-Arg pair at positions 18 and 19 in cohesin recognition. Inspection of the primary sequences of dockerins of Cthe_0258, which recognizes OlpC with higher affinity, and cellulase Cel9D-Cel44A, which binds both *C. thermocellum* and *C. cellulolyticum* cohesins, also revealed unsuitable substitutions at one of the dockerin binding faces which should render one of the dockerin faces inactive for the recognition of *C. thermocellum* cohesins. It is presently unclear why a subset of four dockerins, the two described here and those from Cthe_0258 and Cel9D-Cel44A, have not evolved the dual binding mode characteristic of the other 68 *C. thermocellum* cellulosomal enzymes and extensively described for Xyn10B dockerin. While Cthe_0258 and Cel124A dockerins were shown to direct the appended enzymes to the cell surface, since they bind predominantly OlpC cohesin, the dockerin of Cel9D-Cel44A is believed to present two cohesin binding interfaces with different cohesin specificities (the N-terminal face binds *C. cellulolyticum*-like cohesins and the C-terminal interface their *C. thermocellum* counterparts). Thus, together these data suggest that a dual binding mode is of primary importance for enzymes binding CipA, the multi-modular cohesin scaffolding responsible for cellulosome assembly in *C. thermocellum*. An exception to this general rule is the dockerin of Cthe_0918 which recognize CipA cohesins with higher affinity. The elucidation of the functional role of the protein domain appended to Cthe_0918 dockerin would help to clarify this issue. Nevertheless, the presence of two cohesin binding interfaces in dockerins integrated in multi-enzyme complexes may result in an increment of cellulosome capacity to adjust its catalytic machinery to a highly insoluble and recalcitrant substrate.

4. THE STRUCTURE OF A NOVEL TYPE II COHESIN-DOCKERIN COMPLEX

4.1. Do *Clostridium thermocellum* type II dockerins present two different cohesin-binding faces? ∞

Joana L. A. Brás*, Aldino Viegas†, Benedita Pinheiro†, Teresa Ribeiro*, Fiona Cuskin‡, Shabir Najmudin*, Victor D. Alves*, José A. M. Prates*, Maria João Romão†, Harry J. Gilbert‡, Ana L. Carvalho† and Carlos M. G. A. Fontes*

*Centro Interdisciplinar de Investigação em Sanidade Animal, Faculdade de Medicina Veterinária, Universidade Técnica de Lisboa,1300-477 Lisboa, Portugal; †REQUIMTE–CQFB, Departamento de Química, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, 2829-516 Monte da Caparica, Portugal; ‡ Institute for Cell and Molecular Biosciences, Newcastle University, The Medical School, Newcastle upon Tyne NE2 4HH, UK.

Adapted from a manuscript in preparation.

Abstract

Anaerobic cellulolytic bacteria organize a diverse consortium of enzymes in highly efficient multi-enzyme complexes termed cellulosomes. Cellulosomes are assembled by a large non-catalytic multi-modular protein, termed scaffoldin, which contains repeated type I cohesin domains in tandem that tenaciously bind type I dockerin modules usually located at the C-terminus of enzymes. Scaffoldins may contain a type II dockerin that specifically recognizes type II cohesins located at the cell envelope resulting in cellulosome cell-surface attachment. The structure of type I cohesin-dockerin complexes revealed that dockerins contain a remarkable internal symmetry that sustains two different cohesin-binding faces presenting a similar specificity. Dockerin surface symmetry is responsible for a dual binding mode that is believed to confer significant flexibility to cellulosomes. Here we describe the structure of a novel type II cohesin-dockerin complex from *Clostridium thermocellum*. Similarly to what was previously described for a homologous type II structure, the dockerin presents an asymmetry that would preclude a dual binding mode such similar to that described for type I complexes. However, when the two dockerin structures were overlaid it was observed that there is considerable variation in the residues that provide the most important cohesin contacts. Homology is only restored when either dockerin is overlaid with the 2-fold symmetry related homologue. These observations suggest that, in contrast to what was observed for type I cohesin-dockerin complexes, type II dockerins present two different cohesin-binding faces that express different specificities. This property may confer to type II dockerins the capacity to bind the range of highly diverse type II cohesins identified in *C. thermocellum*.

∞ The student contributed in the following methodologies: cloning, expression and purification, analysis of the complex formation in solution and isothermal titration calorimetry.

4.1.1. Introduction

Cellulosomes are remarkably efficient nanomachines produced by anaerobic microbes to deconstruct plant structural carbohydrates (Bayer *et al.*, 2004; Fontes & Gilbert, 2010). Cellulose and hemicellulose are the most abundant polymers on earth and thus cellulosomes play a key role in carbon turnover. In addition, cellulosomes are central elements for the production of readily uptake sugars in the gastrointestinal tract of a variety of mammals. It is now well established that highly ordered protein:protein interactions between dockerins and cohesins are responsible for both cellulosome assembly and cellulosome cell-surface attachment. Typically, cellulosomal enzymes contain one conserved dockerin domain which tenaciously binds to one of the various cohesin domains found within a protein macromolecular scaffold (Miras *et al.*, 2002; Schaeffer *et al.*, 2002). The assembly of the catalytic components into a complex enhances the synergistic interactions between enzymes with complementary activities and potentiates enzyme-substrate targeting as scaffoldins usually contain cellulose-binding domains. Both these effects contribute to the more efficient plant cell wall degradation.

The cellulosome of the anaerobic bacterium *Clostridium thermocellum* has been extensively studied (Bayer, *et al.*, 2008; Béguin & Alzari, 1998) (Figure 4.1). The protein that mediates the assembly of *C. thermocellum* cellulosome is the scaffoldin subunit termed CipA, which contains nine highly conserved type I cohesins and a family 3 carbohydrate-binding module (CBM) that attaches the cellulosome onto crystalline cellulose (Salamitou, Raynaud, *et al.*, 1994; Tokatlidis, Salamitou, Béguin, Dhurjati, & Aubert, 1991). Type I dockerins, located in cellulosomal enzymes, primarily glycoside hydrolases (GHs) but also carbohydrate esterases and polysaccharide lyases, bind extremely tightly to CipA cohesins thus anchoring the enzymes on to the macromolecular scaffold. Dockerins are usually located at the C-terminus of the enzymes and contain two duplicated segments each of about 22 amino acid residues (Salamitou, Tokatlidis, Béguin, & Aubert, 1992). Significantly, CipA also contains a C-terminal type II dockerin domain, which does not recognize CipA internal type I cohesins. Instead, CipA dockerin binds specifically to type II cohesins located in proteins at the cell-surface, which are usually termed anchoring scaffoldins (Leibovitz & Béguin, 1996). In *C. thermocellum* there are three anchoring scaffoldins located at the bacterium cell-surface which contain one (SdbA, Cthe_1307 (Leibovitz *et al.*, 1997), two (Orf2, Cthe_3079 (Leibovitz & Béguin, 1996) or seven (OlpB, Cthe_3078) (Lemaire *et al.*, 1995) type II cohesins. In order to harmonize the nomenclature of *C. thermocellum* proteins containing type II cohesins we propose to rename scaffoldins SdbA, OlpB and Orf2 as ScaA, ScaB and ScaC, respectively (from Scaffoldin). The presence of type II cohesins in tandem in anchoring scaffoldins contributes for the formation of polycellulosomes which, in *C. thermocellum*, may contain up to 63 different enzymes (when seven CipA molecules that contain nine cohesin domains are bound to ScaB) (Fontes & Gilbert, 2010). Lack in cross-

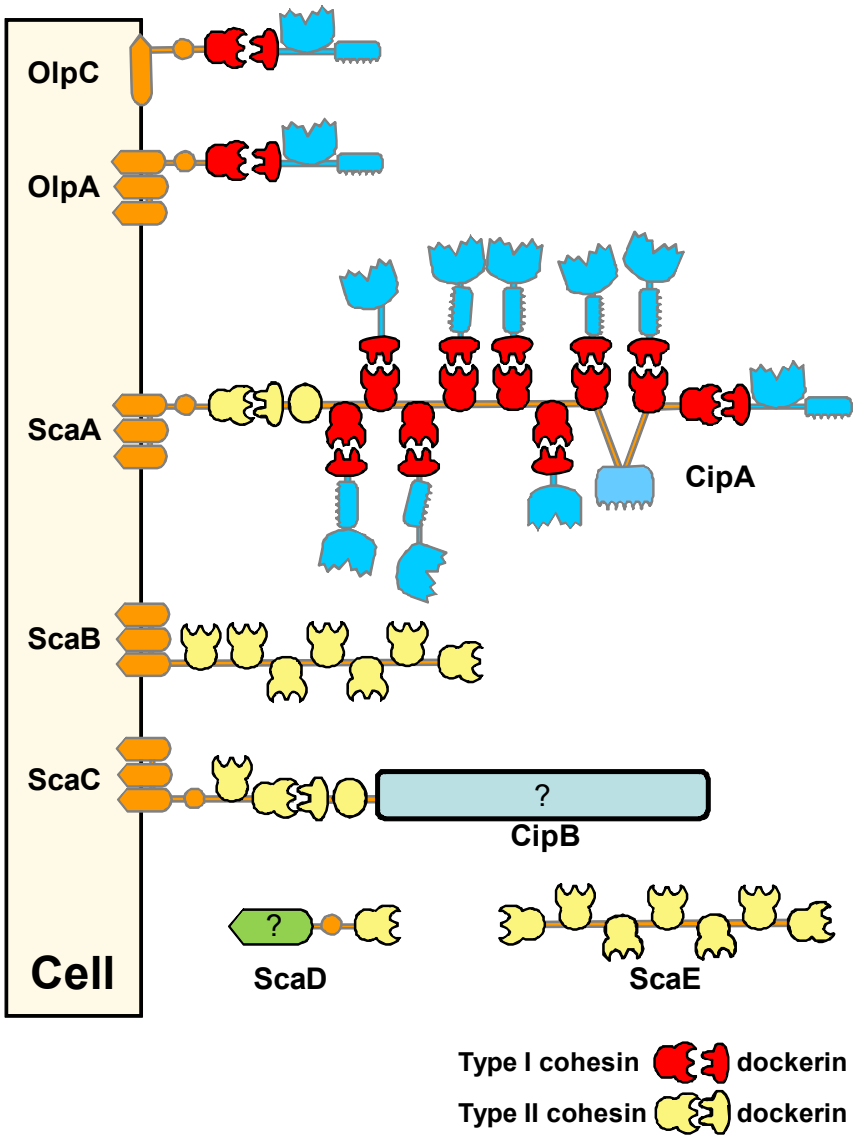
specificity between type I and type II cohesin–dockerin partners, ensures a clear distinction between the mechanism of cellulosome assembly and cell-surface attachment/polycellulosome assembly.

The crystal structures of both type I and type II *C. thermocellum* Coh-Doc complexes have been determined providing considerable evidence for the mechanisms modulating the differences in Coh-Doc specificities (Adams *et al.*, 2006; Carvalho *et al.*, 2003; Carvalho *et al.*, 2007). Dockerins fold in three α-helices that correspond to the first and second duplicated segments, respectively. In contrast, cohesins display a flattened β-barrel fold, which is defined by two β-sheets, one of which represents the dockerin-binding surface. In type I Coh-Doc complexes an extensive hydrogen-bonding network is observed at the Coh-Doc interface which comprehends predominantly one of the dockerin α-helices and the 8-3-6-5 face of the cohesin (Carvalho *et al.*, 2003; Carvalho *et al.*, 2007; Pinheiro *et al.*, 2008). Type I dockerins display a remarkable internal 2-fold symmetry that was shown to possess significant functional importance. Thus, dockerins can bind cohesins either by the N or C-terminal α-helix, revealing that this small module contains two different cohesin-binding interfaces (Carvalho *et al.*, 2007; Pinheiro *et al.*, 2008). The dockerin dual-binding mode is believed to confer considerable flexibility to cellulosome assembly. In contrast to the type I Coh-Doc interaction, in type II complexes both dockerin α-helices contact with the cohesin module through their entire length (Adams *et al.*, 2006). In addition, in type II complexes the interaction surfaces are significantly less charged and binding is predominantly hydrophobic (Adams *et al.*, 2006). Given the involvement of both dockerin helices in cohesin recognition and the lack in symmetry in the hydrophobic and hydrogen bond contacts identified in the type II dockerin of CipA, it was proposed that, unlike the type I interactions a dual binding mode does not operate in type II complexes. Moreover, a module of unknown function, termed X domain found at the N-terminus of type II dockerins, which is absent in type I dockerins, was found to improve dockerin stability and cohesin enhanced recognition (Adams *et al.*, 2006).

To extend our knowledge on the mechanisms of cellulosome assembly and cell surface attachment, *C. thermocellum* (A.T.C.C. 27405) proteome was screened for the presence of unknown proteins containing type II cohesins or dockerins. Here we have identified two novel *C. thermocellum* proteins containing one (Cthe_0735) or seven (Cthe_0736) previously unknown type II cohesins. Following the nomenclature proposed here, the proteins were termed ScaD and ScaE, respectively. In addition, *C. thermocellum* proteome contains a large 2177-residue extracellular protein (Cthe_1806; termed CipB) that possesses a C-terminal type II dockerin fused to an X module. The 1.98 Å resolution structure of the CipB dockerin together with its neighboring X module in complex with the second type II cohesin from ScaC is reported here. The structure reveals the dominant effects of dockerin residues Phe124, Leu147 and Phe148 in cohesin recognition. CipA Asn122, which performs three important

hydrogen bond contacts with the cohesin of ScaA, is replaced in CipB by a glycine residue. This amino acid change would preclude binding of CipA dockerin to ScaC2 cohesin as Asn122 would clash with cohesin residue Phe136. Nevertheless, CipB dockerin may interact with ScaA cohesin if rotated by 180º at the interface plan. These observations together with the asymmetric nature of *C. thermocellum* type II dockerins suggest that these modules present two different cohesin-binding interfaces.

Figure 4.1| Organization of *C. thermocellum* cellulosome.



C. thermocellum scaffoldin (CipA) contains nine type I cohesin domains and thus organizes a multi-protein complex with 9 enzymes. The C-terminal type II dockerin domain of CipA binds, specifically, type II cohesin domains found in cell surface proteins (ScaA, ScaB and ScaC) or to the extracellular ScaD and ScaE scaffoldins described in this work. Since the anchoring scaffoldins ScaB, ScaC and ScaE contain more than one type II cohesin domains, they effectively contribute for the assembly of polycellulosomes that may contain up to 63 catalytic sub-units. Nevertheless, cellulosomal enzymes may adhere directly to the bacterium cell surface by binding the single type I cohesin domains found in OlpA and OlpC.

4.1.2. Material and Methods

4.1.2.1. Cloning, expression and purification

Genes encoding dockerin and cohesin domains were amplified from *C. thermocellum* genomic DNA using the thermostable DNA polymerase NZYDNAChange (NZYTech; see Table 4.1 for primer sequences). Amplified DNA was directly cloned into pNZY28 (NZYTech) and sequenced to ensure that no mutations were accumulated during the amplification. Genes encoding type II dockerins of CipA (residues 1691-1853) and CipB (residues 2015-2177) fused to their respective X modules were cloned into pET32a (Novagen) in *Eco*RI and *Xho*I sites (see Table 4.1). The resulting recombinant proteins contain an N-terminal thioredoxin domain. In contrast, the cohesin domain of ScaA (ScaA; residues 27-201), the third cohesin of ScaB (ScaB3; residues 404-565), the first cohesin of ScaC (ScaC1; residues 34-204), the second cohesin of ScaC (ScaC2; residues 205-364), the unique cohesin of ScaD (ScaD; residues 104-268) and the sixth cohesin of ScaE (ScaE6, residues 974-1136) were subcloned into pET21a (Novagen) restricted with *Nhe*I and *Xho*I (see Table 4.1). The type I dockerin from Xyn10B and the second cohesin of CipA (type I) were expressed as described previously (Pinheiro *et al.*, 2008). The gene encoding the unknown N-terminal domain of ScaD, termed ScaD-Unk, was amplified as described above using the primer pair specified in Table 4.1. The gene was subsequently cloned into pET32a (Novagen) using engineered *EcoR*I and *Xho*I sites. All recombinant proteins contained an internal or a C-terminal His_6-tag. For crystallization studies the type II cohesin-dockerin complex was generated in vivo. Thus, the gene encoding the X domain fused to the type II dockerin of CipB was amplified through PCR containing engineered *Nde*I-*Bam*HI sites (see Table 4.1 for primer sequences) and cloned into pET3a (Novagen), generating pET3aCipBXDoc. The dockerin gene under the control of the vector T7 promoter was extracted from pET3aCipBXDoc by digesting the recombinant plasmid with *Bgl*II and *Bam*HI restriction enzymes. The resulting gene was cloned into *Bgl*II site of pET21a derivative containing the gene encoding ScaC2 such that the two genes were fused in tandem. The resulting construct, termed pET21a_ScaC2-CipBXDoc, encodes a recombinant cohesin that contains a C-terminal His_6-tag while the encoded XDoc module has no appended tags.

Escherichia coli Origami cells, transformed with pET32a derivatives, and BL21 cells, transformed with pET21a derivatives, respectively, were grown at 37 ºC to mid-exponential phase (OD_{600}=0.6). Recombinant protein expression was induced by adding 1 mM IPTG (isopropyl β-D-thiogalactoside) and incubated for further 16 h at 19 ºC. Soluble recombinant proteins were purified by immobilized metal ion affinity chromatography as described previously (Pinheiro *et al.*, 2008). Fractions containing the purified proteins were buffer exchanged using PD-10 Sephadex G-25M gel-filtration columns (GE Healthcare) into 50 mM Hepes, pH 7.5, containing 100 mM NaCl and 5 mM $CaCl_2$. For isothermal calorimetry, a

further purification step by size exclusion chromatography using a Superdex 75 column was performed and the proteins were buffer exchanged into the same buffer without NaCl. SDS-PAGE indicated that all the recombinant proteins were highly pure (>95%). For crystallography, the complex was purified by metal ion affinity chromatography as described above, buffer exchanged into 20 mM Tris-HCl, pH 8.0, containing 2 mM $CaCl_2$, and then further purified by anionic exchange chromatography using a Source 30Q column and a gradient elution of 0-1 M NaCl, to separate the complex from unbound cohesin. Fractions containing the protein complex were buffer exchanged and then concentrated in 2 mM $CaCl_2$ to a final concentration of 20 g/l.

Table 4.1| Primers used to clone the genes encoding the cohesin and dockerin derivatives produced in the present study.

Clone	Sequence (5' → 3')	Direction
CipA XDoc	CTC**GAATTC**AATAAACCTGTAATAGAA	Forward
	CAC**CTCGAG**TTACTGTGCGTCGTAATC	Reverse
CipB XDoc	CTC**GAATTC**AACAACGATAGTACTG	Forward
	CAC**CTCGAG**TTAATAGCGGGAAGGT	Reverse
ScaA	CTC**GCTAGC**AGGGCAGATAAAGCCTCG	Forward
	CAC**GTCGAC**CTCATAAGGCTCGTCACC	Reverse
ScaB6	CTC**GCTAGC**GATTCCTATGTGATAATG	Forward
	CAC**CTCGAG**TACCACTATTTCCCCAGG	Reverse
ScaC1	CTCGCTAGCGAGACTTCGAGTATACCT	Forward
	CAC**CTCGAG**AATCGTTATTGCAAGTTC	Reverse
ScaC2	CTCGCTAGCGCACACATTGCTTTGGAAC	Forward
	CAC**CTCGAG**AATCACAGTAATTTTGTCG	Reverse
ScaD	CTCGCTAGCGCCGAAGCAAATATTCAA	Forward
	CACCTCGAGATTAACTTTTACCCCTTT	Reverse
ScaE6	CTC**GCTAGC**GATGCCGTATCATCGGGC	Forward
	CAC**CTCGAG**GTTGATTGGTTCGGGCTG	Reverse
ScaC2-CipBXDoc	CTC**CATATG**AACAACGATAGTACTG	Forward
	CAC**GGATTC**TTAATAGCGGGAAGGT	Reverse
CipB XDoc M114A	GCAAGACAATGCTATTAAT**GC**GGTGGATGTGATGGAAATATCC	Forward
	GGATATTTCCATCACATCCACC**GC**ATTAATAGCATTGTCTTGC	Reverse
CipB XDoc M118A	GCTATTAATATGGTGGATGTG**GC**GGAAATATCCAAAG	Forward
	CTTTGGATATTTCC**GC**CACATCCACCATATTAATAGC	Reverse
CipB XDoc S121A	GTGGATGTGATGGAAATA**G**CCAAAGTTTTTGGCAC	Forward
	GTGCCAAAAACTTT**GGC**TATTTCCATCACATCCAC	Reverse
CipB XDoc F124A	GGAAATATCCAAAGTT**GC**TGGCACAAGAGCCGGAGATG	Forward
	CATCTCCGGCTCTTGTGCCA**GC**AACTTTGGATATTTCC	Reverse
CipB XDoc L147A	GGACGGAGCAATCAAT**GC**ATTTGATATAGCTATAGTTATCAGGC	Forward
	GCCTGATAACTATAGCTATATCAAAT**GC**ATTGATTGCTCCGTCC	Reverse
CipB XDoc F148A	GGACGGAGCAATCAATTTA**GC**TGATATAGCTATAGTTATCAGGC	Forward
	GCCTGATAACTATAGCTATATCA**GC**TAAATTGATTGCTCCGTCC	Reverse
CipB XDoc I154A	GATATAGCTATAGTT**GA**CAGGCATTTTAACGCATTACC	Forward
	GGTAATGCGTTAAAATGCCTG**TC**AACTATAGCTATATC	Reverse
ScaD-Unk	CTC**GAATTC**GCTGAGCCCACAATATCCGGAG	Forward
	CAC**CTCGAG**TTAATCATTATTTGCCGTATTTG	Reverse

Engineered restriction sites and mutation points are shown in bold.

4.1.2.2. Site-Directed Mutagenesis

Site-directed mutagenesis was carried out using the PCR-based NZYTech site-directed mutagenesis kit (NZYTech Ltd) according to the manufacturer's instructions, using DNA of pET32a derivative containing the gene encoding CipB XDoc as the template. The sequence of the primers used to generate these mutants is displayed in Table 4.1. The mutated DNA sequences were sequenced to ensure that only the appropriate mutations had been incorporated into the nucleic acids.

4.1.2.3. Analysis of complex formation in solution

Complex formation was initially evaluated through non-denaturing PAGE (Pinheiro *et al.*, 2009). The two putatively interacting proteins were combined in 50 mM Hepes buffer (pH 7.5), containing 100 mM NaCl and 5 mM $CaCl_2$ for 1 h at 25 ºC and complex formation analysed was evaluated in 10% non-denaturing polyacrilamide gels. For lower-affinity interactions, complexes were detected by increasing concentrations of a dockerin against a fixed concentration of cohesin. New bands appearing in the native gel were used as an indication of complex formation. The differences in the affinities of the various interacting proteins were assessed by combining, in the same solution, equimolar quantities of two dockerins and one cohesin (or two cohesins and one dockerin) and analyzing complex formation by non-denaturing gel electrophoresis. The levels of the protein containing two potential competing protein partners were reduced when compared with the potential partner.

4.1.2.4. Isothermal titration calorimetry of cohesin–dockerin binding

Isothermal titration calorimetry (ITC) was carried out essentially as described previously (Pinheiro *et al.*, 2009), except that proteins were dialyzed into 50 mM sodium Hepes, pH 7.5, containing 2 mM $CaCl_2$. Measurements were made at 328.15 K titrations between CipB XDoc (25 μM) and cohesin ScaC2 (250 μM). For titrations between CipB XDoc (20 μM) and cohesin ScaE6 (140 μM) measurements were made at 318.15 K since the protein precipitates at 328.5 K. During titration, the dockerin was stirred at 300 rev/min in the reaction cell, which was injected with 28 successive 10 μl aliquots of ligand comprising cohesin at 300 s intervals. Integrated heat effects, after correction for heats of dilution, were analysed by non-linear regression using a single site-binding model (Microcal ORIGIN, Version 5.0; MicrocalSoftware). The fitted data yield the association constant (K_a) and the enthalpy of binding (ΔH). Other thermodynamic parameters were calculated by using the standard thermodynamic equation: $RTlnK_a = -\Delta G = \Delta H - T\Delta S$.

4.1.2.5. Complex Crystallization

The Type II complex ScaC2-CipBXDoc was crystallized at 293 K by the hanging drop vapour diffusion method. The crystals were grown in 4% Tacsimate, pH 5.0, and 12% PEG 3350

over a period of 3-5 days and were cryoprotected with 30% glycerol. The space group was determined to be C121 with unit cell dimensions *a* = 116.67 Å, *b* = 78.63 Å, *c* = 35.80 Å, with *β* = 95.87°.

4.1.2.6. X-Ray Data Collection and Processing

Data from a single ScaC2-CipBXDoc crystal was collected at a wavelength of 0.9735 Å in the European Synchrotron Radiation Facility (ESRF), ID14-EH4 (Grenoble, France) to 1.98 Å resolution at 100 K. Data were processed and scaled with the software MOSFLM and SCALA from the CCP4 suite (Winn *et al.*, 2011). The Matthews coefficient of the ScaC2-CipBXDoc crystal is 2.2 $Å^3$ Da^{-1} for one ScaC2-CipBXDoc heterodimer in the asymmetric unit, with a solvent content of 43.13% (Table 4.2).

4.1.2.7. Structure Determination, Refinement and Model Building

The crystal structure of the ScaC2-CipBXDoc complex from *C. thermocellum* was solved by molecular replacement using the program PHASER (McCoy *et al.*, 2007) from the CCP4 suite (Winn *et al.*, 2011) and the ScaA type II Coh-XDoc complex (PDB ID code: 2b59) as model. The data was refined at 1.98 Å resolution and the final statistics are summarized in Table 4.2. Initial building of the complex into the electron density was performed using ARPwARP (Langer *et al.*, 2008) and the remaining residues were manually built using the COOT program (Emsley *et al.*, 2010). The refinement was performed using REFMAC (Skubák, Murshudov, & Pannu, 2004). Water molecules were added and final refinement included translation, liberation and screw-rotation (Adams *et al.*, 2010) of the two independent groups (molecules A and B). The final model has R_{cryst} = 18.7% and R_{free} = 24.7% and includes 322 water molecules and two calcium ions. The residues Met1 and Ala1 of the Coh module (chain A), Met1, Asn2, Asn3, Asp4, Ser5 and Thr6 of the X module (chain B) and Leu160, Pro161, Ser162, Arg163 and Tyr164 from the Doc module (chain B) are disordered and, hence, not observed. The side chains of residues Arg73, Lys158 and terminal His_6-tag from the Coh module and Glu63 and Lys85 from the X module are also disordered and, therefore, not observed. The structure was deposited in the Protein Data Bank under the accession code: 2vt9. All polypeptide chains are well defined in the electron density map (with the exception of the residues mentioned above) with average *B* factors of 16.6, 16.7 and 16.6 $Å^2$ for the cohesin, X and dockerin modules, respectively.

Table 4.2| X-ray data and structure quality statistics for the *C. thermocellum* SdbC2-CipBXDoc complex.

Data quality	Type II ScaC2-CipBXDoc
X-ray source	European Synchrotron Radiation Facility, ID14-EH4
Unit cell dimensions, Å	*a* =116.67, *b* =78.63, *c* =35.80, β = 95.87°
Space group	*C*121
Resolution of data, Å (outer shell)	39.31– 1.98 (2.09 - 1.98)
R_{pim} (outer shell), %	5.8 (10.4)
R_{merge} (outer shell), %*	9.4 (17.3)
Mean *I*/σ (*I*) (outer shell)	13.1 (5.5)
Completeness (outer shell), %	97.8 (97.2)
Multiplicity (outer shell)	3.6 (3.7)
Refinement	
Resolution used in refinement, Å	1.98
R_{cryst}/R_{free} (%)†	18.7 /24.7
rms deviation bonds, Å	0.01
rms deviation angles, º	1.2

*R_{merge} = Σ |*I*-<*I*>|/Σ <*I*>, where *I* is the observed intensity, and <*I*> is the statistically weighted average intensity of multiple observations. †R_{work} = Σ ||F_{calc}|– |F_{obs}||/Σ |F_{obs}|× 100, where F_{calc} and F_{obs} are the calculated and observed structure factor amplitudes, respectively (R_{free} is calculated for a randomly chosen 5% of the reflections).

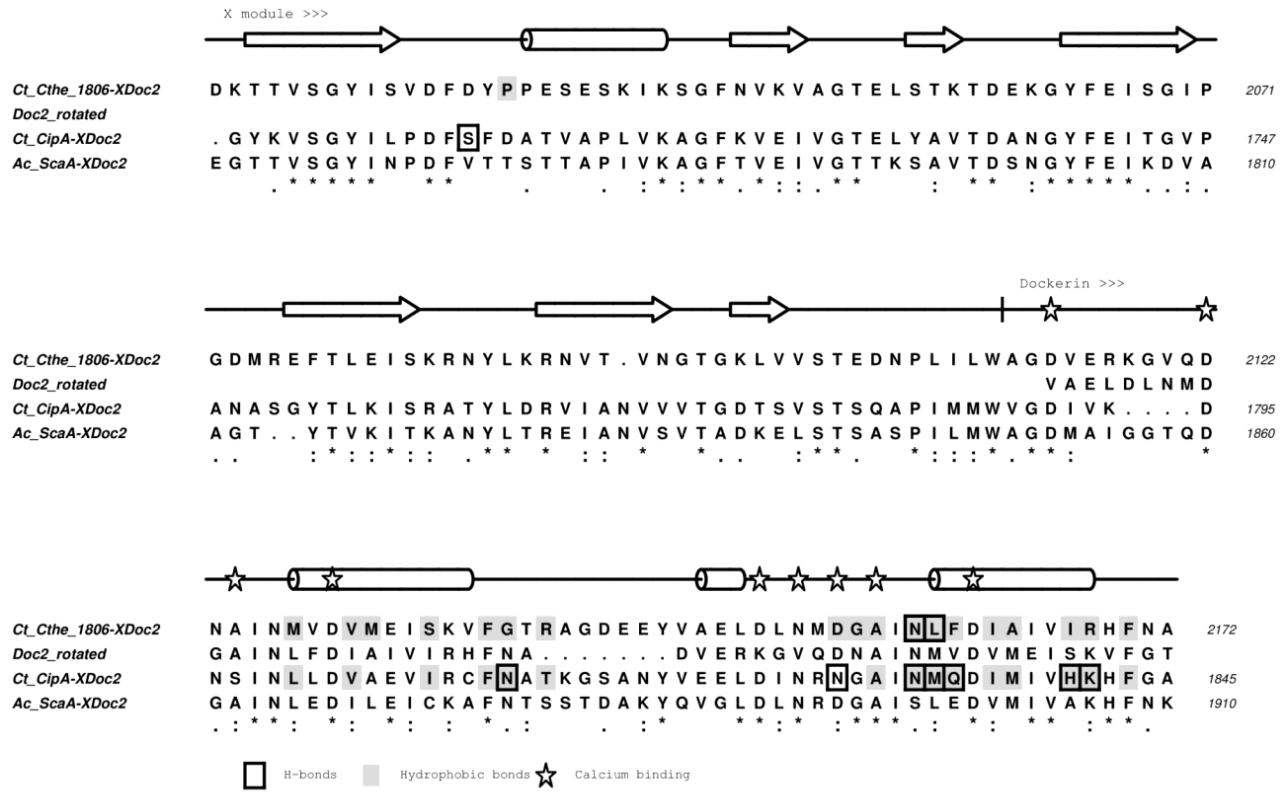
4.1.3. Results and Discussion

4.1.3.1. Novel type II cohesin and dockerin domains in *C. thermocellum* proteins

To identify the complete repertoire of cellulosomal proteins containing type II modules, *C. thermocellum* ATCC 27405 proteome was searched through BLAST (http://blast.ncbi.nlm.nih.gov) using the primary sequences of CipA type II dockerin or the type II cohesin of ScaA. The data revealed that, in addition to CipA, *C. thermocellum* contains a second protein containing a C-terminal type II dockerin (Cthe_1806), which was termed CipB. A recent transcriptomic study revealed that similarly to CipA, CipB expression decreases as the growth rate on cellobiose increases while scaffoldin gene expression was not influenced by the growth rate on cellulose (Riederer *et al.*, 2011). In addition, a comprehensive proteomic project developed to identify cellulosomal changes in response to different carbon sources revealed that CipB is upregulated when *C. thermocellum* is grown on cellulose particularly if this polysaccharide is mixed with xylan and/or pectin preparations

(Raman *et al.*, 2009). However, the levels of CipB in cellulosomes are relatively modest and ranged from 5%, when cellulose was the sole substrate, to 11-13%, when cellulose was mixed with xylan and/or pectin, of the levels of CipA (Raman *et al.*, 2009). CipB is a 2177 residue polypeptide containing a signal peptide followed by four domains of unknown function including an internal 41-residue motif that is repeated 19 times and contains three highly conserved cysteine amino acid residues. The four domains were expressed independently in *Escherichia coli* and screened for polysaccharide activity or CBM function against a significant range of carbohydrates, although the recombinant polypeptides were unable to act both as enzymes or CBMs against the substrates/ligands tested (data not shown). CipB contains at its C-terminus an X module followed by a putative type II dockerin in a location similar to CipA. Alignment of the X-Doc regions of CipA and CipB (Figure 4.2), which are 48% identical, revealed significant amino acid substitutions at positions that were previously shown to contain key amino acids recognizing the type II cohesin. Residues at positions 11 and 12 of the calcium binding loops were previously suggested to play a role in modulating cohesin recognition (Adams *et al.*, 2006). In CipA, Met144 and Gln145 occupy such important positions at the second duplicated segment and dominate the protein:protein interaction with type II cohesins. These residues are replaced by Leu147 and Phe148, respectively, in CipB thus introducing significant differences in particular at position 12 of the calcium binding region (Figure 4.2).

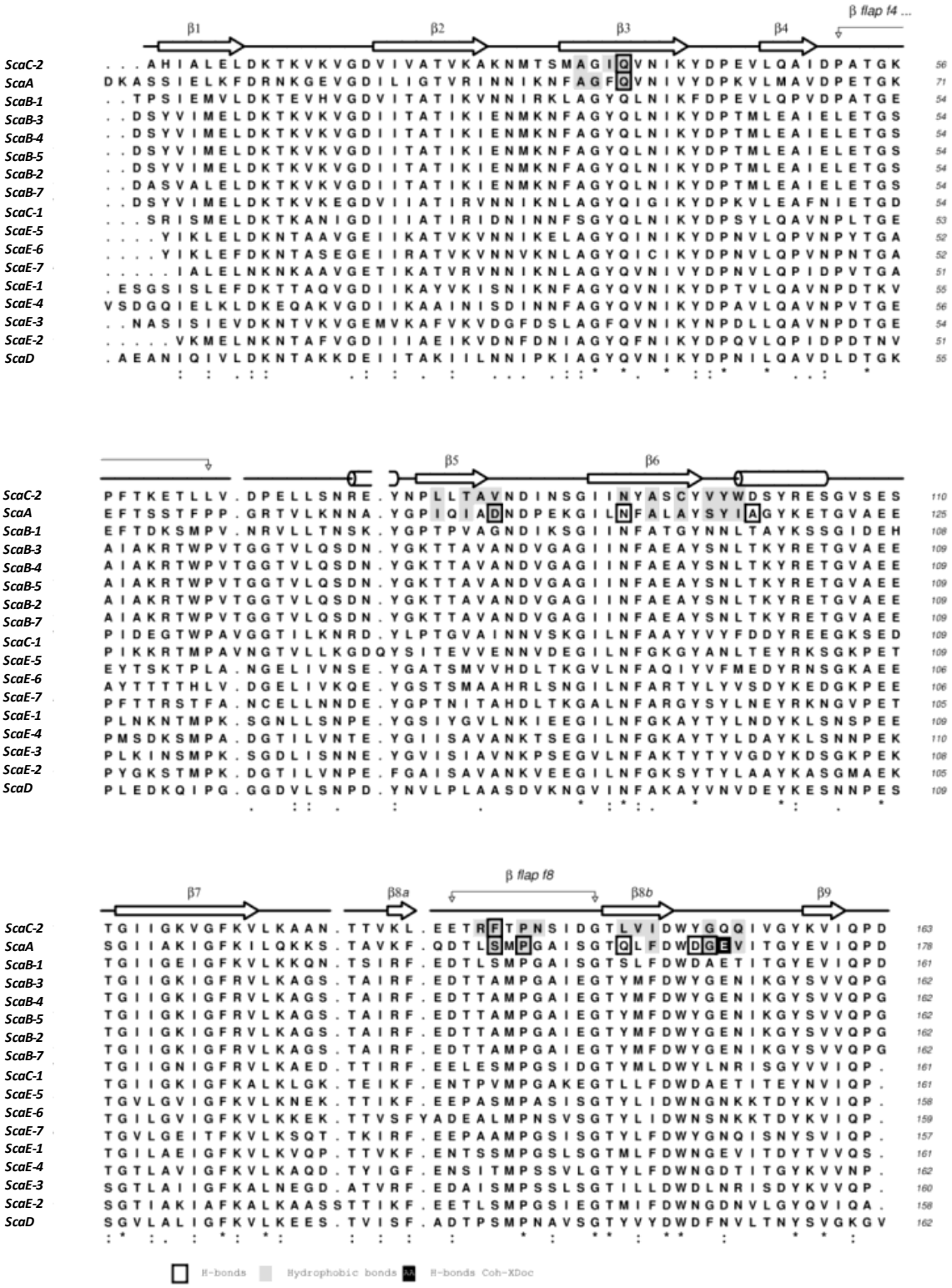
Figure 4.2| Alignment of Type II dockerins fused to their neighboring X domain.



In addition, Asn122 that dominates the hydrogen bond network with the cohesin in the first dockerin helix of CipA is replaced in CipB by a glycine residue. The implications of these amino acid substitutions in cohesin recognition remain to be determined (see below). However, if CipA dockerin is aligned with CipB dockerin when the two duplicated segments have been reversed then much more conservation in the residues that made the main contacts with the type II cohesin is apparent (Figure 4.2). This suggests that, in contrast to what was observed for CipA dockerin, in CipB the dockerin might interact with ScaA cohesin primarily through the N-terminal helix.

The BLAST search of *C. thermocellum* proteome with type II ScaA cohesin sequence revealed that, in addition to the previously described scaffoldins (ScaA, ScaC and ScaB) containing type II domains, two additional polypeptides (here termed ScaD and ScaE) contain putative type II cohesin sequences. Alignment of the 18 type II cohesin domains of *C. thermocellum* (Figure 4.3) revealed that although pivotal residues for dockerin recognition by ScaA cohesin (Adams *et al.*, 2006), such as Gln52, Asn106 and Pro153, are conserved in all sequences there is considerable deviation in the other protein-interacting amino acids. Thus, it is suggested that if functional ScaD and ScaE type II cohesins might present different ligand recognition platforms when compared with ScaA and that considerable differences in the binding faces of type II cohesins might exist (see below). The genes encoding the two novel proteins ScaD and ScaE are organized in a putative operon and thus should be co-ordinately expressed. The two proteins contain one (ScaD) or seven (ScaE) cohesin domains. ScaE contains a typical signal peptide but does not present an apparent cell surface tag and thus this protein is believed to be exclusively extracellular. In contrast, ScaD contains an N-terminal domain of unknown function followed by the C-terminal putative type II cohesin. To determine the role of the N-terminal domain of ScaD, here termed ScaD-Unk, in carbohydrate metabolism or cellulosomal function, the module was purified to electrophoretic homogeneity and its biochemical properties evaluated. Affinity-gel electrophoresis analysis revealed that ScaD-Unk displays no significant affinity for xyloglucan, the β1,4-β1,3-mixed glucans barley β-glucan and lichenan, the β1,4-glucan HEC (hydroxyethylcellulose), konjac glucomannan, oat spelt xylan, the β1,3-glucans laminarin and curdlan, carob galactomannan, potato galactan, pullulan or pustulan (results not shown). In addition, the domain was unable to depolymerase the above referred polysaccharides, suggesting that ScaD is not a plant-cell-wall-degrading enzyme (results not shown). The capacity of ScaD-Unk to interact with *C. thermocellum* cell-wall fractions was investigated. Pull down experiments revealed that ScaD-Unk is unable to bind *C. thermocellum* secondary cell wall polymers or peptidoglycan (data not shown). Thus, the role of the N-terminal domain of ScaDC in cellulosome assembly or cell surface attachment remains elusive.

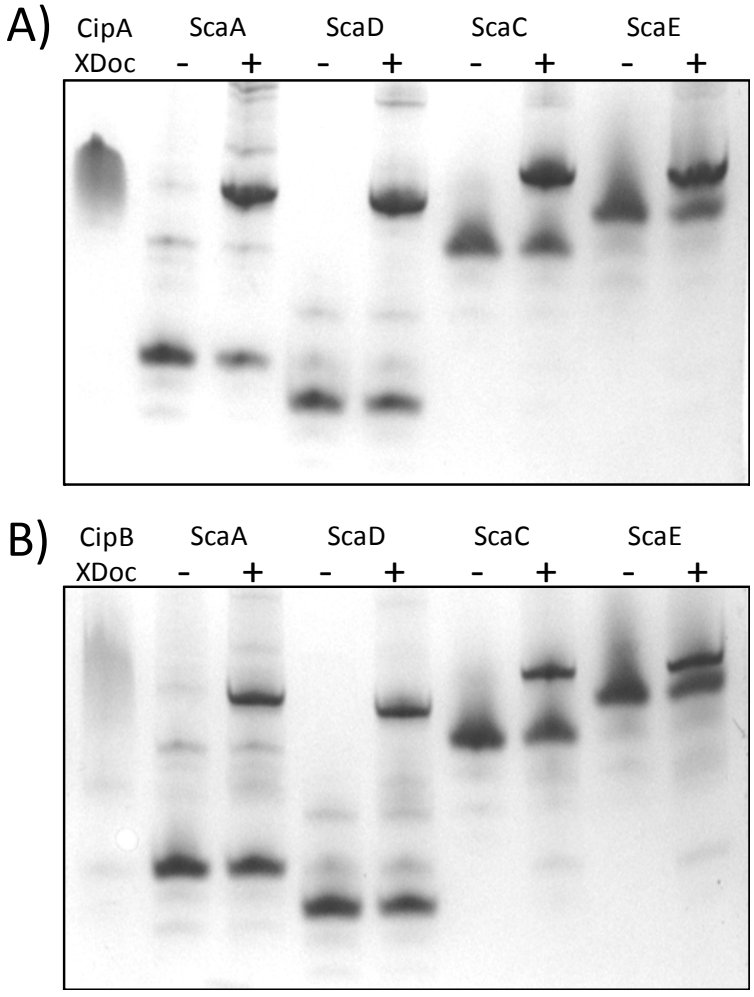
Figure 4.3| Alignment of *C. thermocellum* 18 type II cohesin modules.



4.1.3.2. Are the novel *C. thermocellum* type II cohesins and dockerins functional?

The capacity of *C. thermocellum* type II dockerins, while fused to their neighbouring X modules, to recognize a range of type II cohesins selected from the potentially five anchoring scaffoldins (see above) was initially explored through non-denaturing gel electrophoresis. The data, exemplified in Figure 4.4, reveals that both CipA and CipB type II dockerins bind to the type II cohesins of scaffoldins ScaA, ScaC, ScaD and ScaE. This observation confirms that both the dockerin of CipB and the cohesins of ScaD and ScaE are functional type II domains. In addition, the dockerin of CipB and the cohesins of ScaD and ScaE were unable to interact with type I counterparts (data not shown) confirming that no cross-reactivity exists between type I and II modules.

Figure 4.4| Detection type II cohesin-dockerin complex formation by non-denaturing gel electrophoresis.



A) XDoc module of CipA was probed against the cohesins of ScaA, ScaC, ScaD and ScaE. **B)** XDoc module of CipB was probed against the cohesins of ScaA, ScaC, ScaD and ScaE. Cohesins were used at a double molar concentration in relation to dockerins.

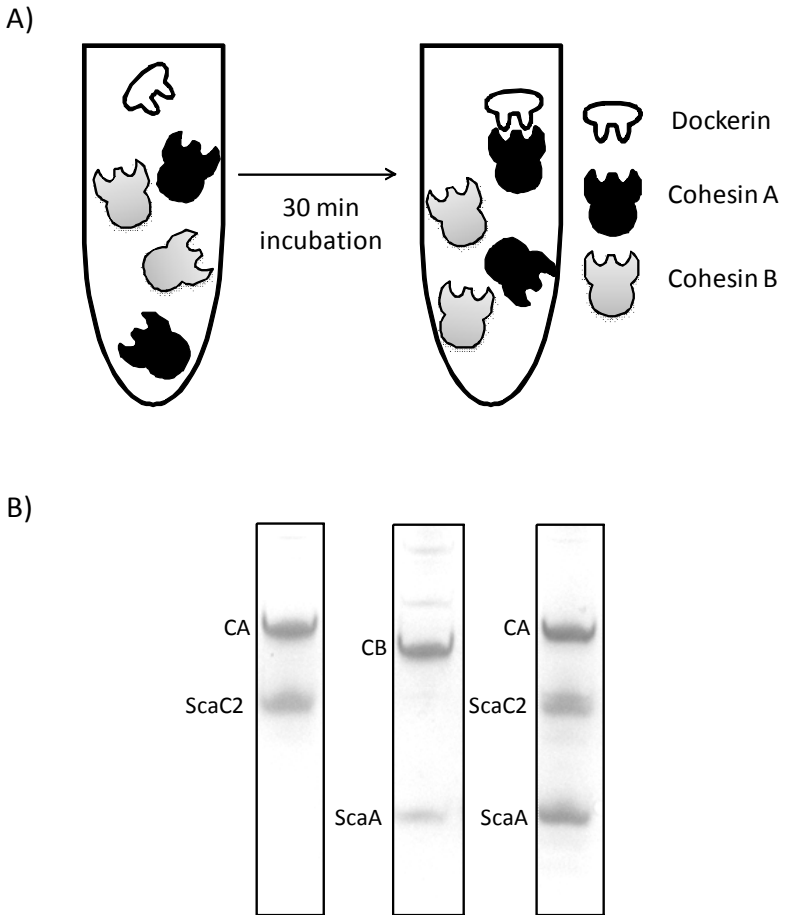
The preference of type II dockerins or type II cohesins for specific protein partners was analysed by mixing one dockerin with two cohesins, when dockerin preference was analysed, or one cohesin with two dockerins, when cohesin preference was evaluated, and assessing complex formation by native gel electrophoresis. The data are fully presented in Table 4.3 and are exemplified in Figure 4.5. When dockerin preference was evaluated, the levels of dockerin were in deficit when compared with those of their individual protein counterparts that were joined in equimolar concentrations (the same applied when cohesin preference was evaluated). Results revealed that the tested type II cohesins have no preference for either CipA or CipB dockerins. Thus, when the two dockerins are mixed exclusively with one cohesin present at a lower concentration, the formation of the two complexes is always observed. This is quite a meaningful result and suggests that both proteins, CipA or CipB, bind equally well to all anchoring scaffoldins. This way, anchoring scaffoldins containing more than one type II cohesin contribute for the production of polycellulosomes that may contain both CipA and CipB. This observation suggests that polycellulosome composition may depend primarily from the levels of expression of CipA and CipB, which are usually 10 times higher for CipA (Raman *et al.*, 2009). In contrast, the type II dockerins displayed a clear preference for specific cohesin partners. Thus, the type II cohesin of ScaC2 was the preferred protein partner for both CipA and CipB dockerins. In contrast the cohesin of ScaE6 was never selected as protein partner when any of the other cohesins was present. Both dockerins displayed a similar affinity profile and have no distinct difference in relation to the proteins partners. Integration of all data concerning dockerin preference allowed ranking the affinities of the dockerins for the various cohesins (ScaB3 was excluded from this analysis as it did not migrate in native gels). Thus, the two type II dockerins preferred the following cohesins, from the most to the less preferred: ScaC2>ScaA>ScaD>ScaE6.

Table 4.3| Identification of preferred cohesin and dockerin partners.

Dockerin	Dockerin binding preference
CipA XDoc	ScaC2 > ScaA> ScaD > ScaE6
CipB XDoc	ScaC2 > ScaA> ScaD> ScaE6
Cohesin	**Cohesin binding preference**
ScaA	CipA XDoc = CipB XDoc
ScaB	CipA XDoc = CipB XDoc
ScaC1	CipA XDoc = CipB XDoc
ScaC2	CipA XDoc = CipB XDoc
ScaD	CipA XDoc = CipB XDoc
ScaE6	CipA XDoc = CipB XDoc

One dockerin and two cohesins *or vice versa* were mixed at the same time and the resulting cohesin-dockerin complex formed was analyzed by native gel electrophoresis. In this way, it was possible to analyze the preference for protein partners of both cohesin and dockerins.

Figure 4.5| Preference of CipA XDoc domain for cohesin partners.



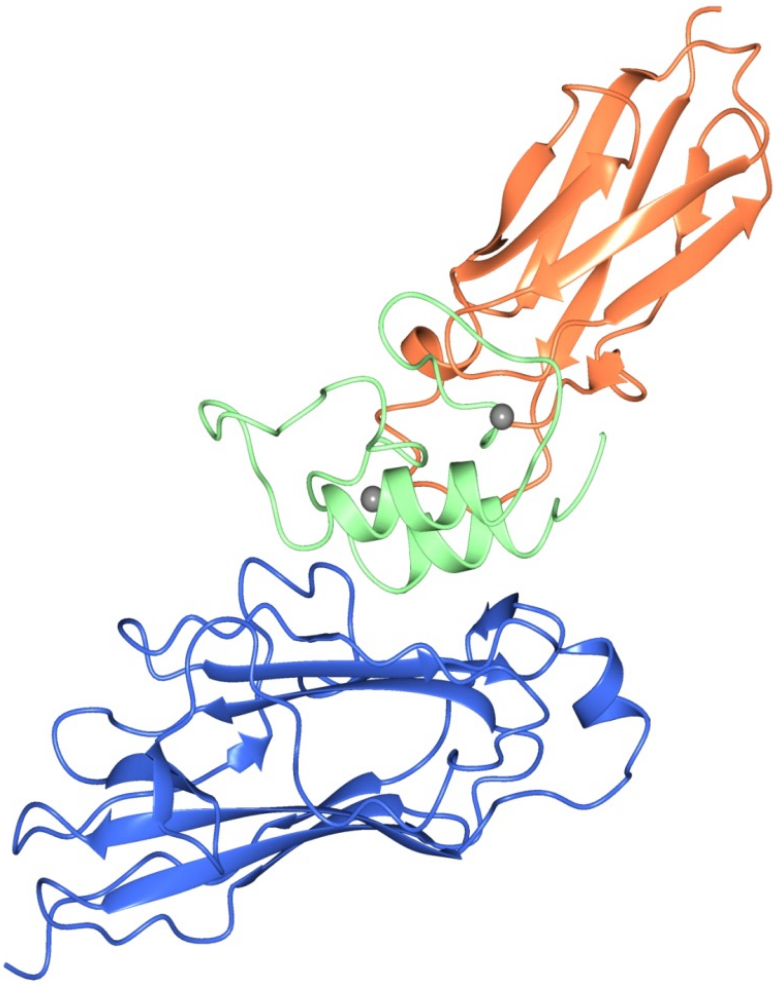
A) In the first panel the method used to detect preferential partners for dockerin is illustrated. The dockerin is mixed with a double molar concentration of two potential cohesin partners and after a 30 min incubation the complex formed is visualized through native gel electrophoresis. **B)** Example of one of such experiments where CipA XDoc domain is mixed with ScaC2, forming complex A (CA) or ScaA, forming complex B (CB). When the XDoc domain is mixed with the two cohesins exclusively complex A (CA) is formed, revealing that CipA XDoc domain displays a preference for binding ScaC2.

4.1.3.3. The structure of ScaC2 cohesin in complex with CipB XDockerin (ScaC2-CipBXDoc).

The crystal structure of the ScaC2 type II cohesin in complex with the X-Doc domains of CipB was solved to a 1.98 Å resolution by molecular replacement (Figure 4.6), using the crystal structure of the previously described type II cohesin as search model (PDB ID code 2bm3). The high degree of similarity of this structure when compared to the ScaA-XDocCipA type II complex is reflected by the low r.m.s.d values between them - 1.12 Å for 166 Cα atoms of the whole complex, 0.86 Å for 156 Cα atoms of the Coh alone, 0.87 Å for 127 Cα atoms of the XDoc module, 0.77 Å for 83 Cα atoms of the X module alone and 0.78 Å for 44 Cα atoms of the Doc alone.

The type II cohesin of the ScaC2-CipBXDoc complex (Figure 4.6) forms a flattened, elongated 9-stranded β-barrel with a jelly-roll topology. The nine β-strands define two β-sheets – the first β-sheet is defined by strands 8-3-6-5 (front face) and the second is defined by strands 9-1-2-7 (back face). Its core is highly hydrophobic. The α-helical crowning observed between strands 6 and 7 and the two β-flap regions that disrupt the normal progression of strands 4 and 8, respectively, are a common feature in this type of structure (Haimovitz *et al.*, 2008; Noach *et al.*, 2005). These β-flap regions are thought to be involved in the type II interaction and specificity (see below). As described above, the structure displays striking similarity to the complexed type II cohesin of ScaA. In addition, comparing the structure of ScaC2 with that of unbound ScaA type II Coh (PDB ID code 2bm3 (Carvalho *et al.*, 2005)) shows that, similar to what happens to other related structures, the cohesin does not undergo significant conformational changes upon binding as revealed by the r.m.s.d of 0.77 $Å^2$ (for 155 Cα atoms) between both structures.

Figure 4.6| Structure of the ScaC2-CipBXDoc type II cohesin-dockerin complex.



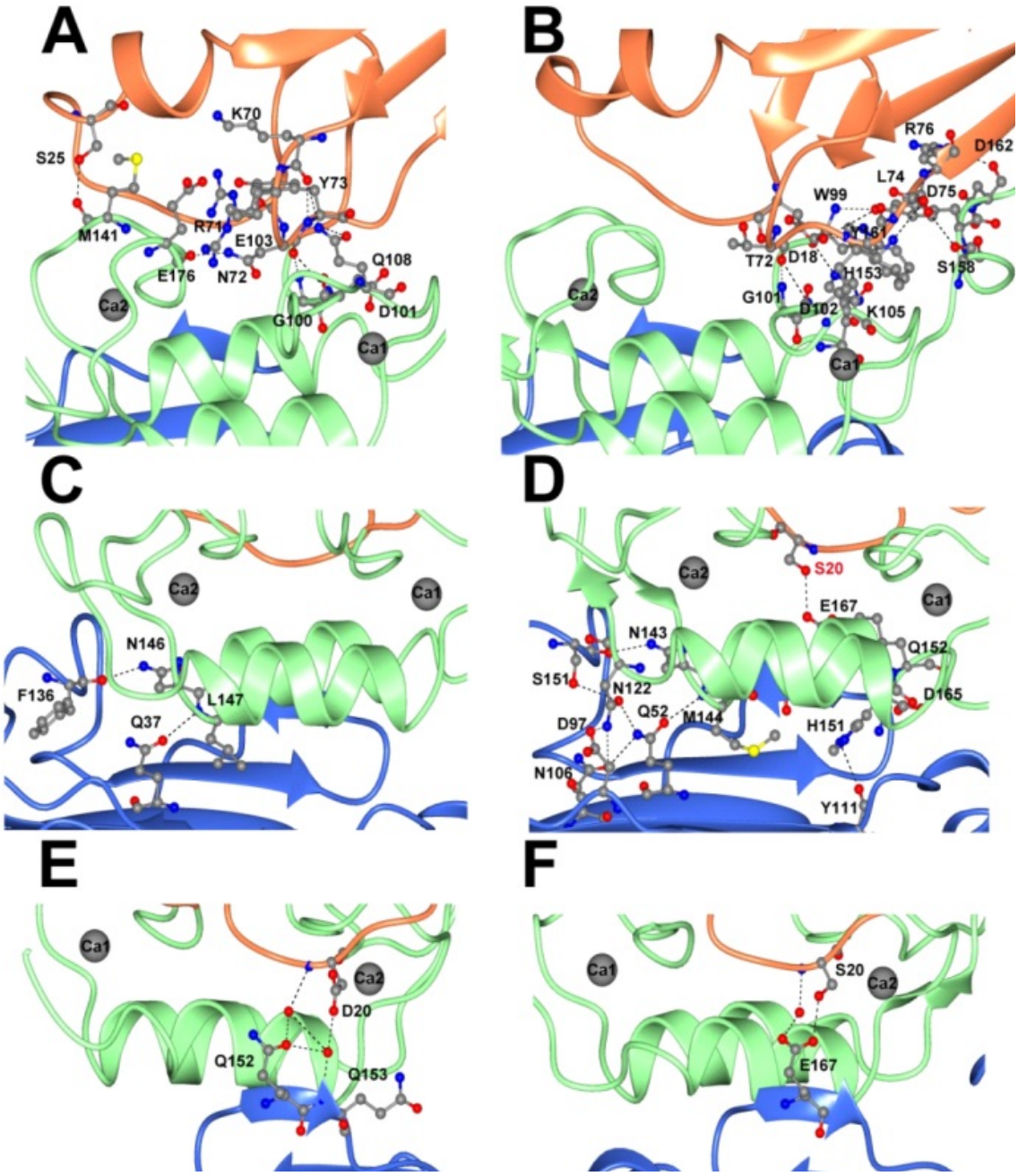
X-ray structure of the novel type II cohesin-dockerin complex. The CipB type II dockerin together with its neighboring X domain are depicted in green and orange, respectively. The second cohesin from the ScaC cell surface protein is depicted in blue.

The XDoc module was modelled as one single polypeptide chain (chain B) of 164 amino acids (the first 98 belonging to the X module and the remaining to the dockerin). The module X subunit is composed by seven β-strands arranged into two β-sheets (1-4-7 and 2-3-5-6) and a small α-helix connecting strands 1 and 2. The overall fold of this subunit and the β-sheet topology are similar to Ig-like module of avian carboxypeptidase D domain II (PDB ID code: 1qmu) with a backbone r.m.s.d. of 0.96 $Å^2$ to this module. The type II dockerin domain (residues 99-164) forms two loop-helix motifs, named EF-hand motifs (Adams, Webb, Spencer, & Smith, 2005) separated by a 23-residue linker that also forms a small helix. Helix 1 is defined by residues Met114 to Val123, helix 2 (the one in the linker) is defined by residues Ala135 to Asp138 and helix 3 is formed by residues Leu147 to His156. Helixes 1 and 3 are arranged in an antiparallel orientation that places the two calcium ions in opposite sides of the Doc module, similar to that observed for the previously described type II Doc (Adams *et al.*, 2006). Nonetheless, the linker in type II Doc is less structured than in the type I Doc, comprising only one turn in contrast with the three turns in the type I structures. The EF-hand motif loops bind to two calcium ions coordinated in a typical octahedral geometry. The first calcium ion, Ca1, is located near the C-terminus of the X module and is coordinated by residues Asp101 (Oγ1), Asp109 (Oγ1), Ala111 (backbone carbonyl), Asp116 (Oγ1 and Oγ2) and two water molecules (274 and 283). The second calcium, Ca2, is coordinated by residues Asp138 (Oγ1), Asn140 (Oγ1), Asp142 (Oγ1), Ala144 (backbone carbonyl), Asp149 (Oγ1 and Oγ2) and a water molecule (316). These calcium ions are fundamental for the folding stabilization of the dockerin and for cohesin recognition. Furthermore, in the absence of the cohesin subunit, it was shown that binding of calcium to the XDoc module induces homodimerization (Adams *et al.*, 2005).

The X module and the dockerin form an intimate hydrophobic interface (Figure 4.7B) involving residues Asp18, Phe19, Asp20, Tyr21, Pro22, Glu24, Ser25, Lys28, Ile29, Lys70, Arg71, Asn72, Ty73, Leu74, Lys75, Leu97 and Trp98 from the X module and residues Ala99, Gly100, Asp 101, Val102, Glu103, Gln108, Asn110, Ile112, Val134, Glu136, Leu137, Leu139, Asn140, Met141, Asp142, Ile152, Arg155, His156, Asn158 and Ala159 from the dockerin. These interactions include 9 hydrogen bonds and 5 salt bridges (Supplementary Table S4.1 – see annex). X modules were proposed to play a key role in the stabilization of type II dockerins (Adams *et al.*, 2006). The higher stability of the two domain linkage results from the extensive hydrophobic interface and hydrogen bond network established between the X module and the dockerin. These contacts potentiate cohesin recognition and significantly contribute to improve the affinity for type II cohesins (Ka>10^9 M^{-1}). When comparing the XDoc interface of the ScaC2-CipBXDoc with the one from ScaA-CipAXDoc, it is clear that there is a more extensive network of contacts in the complex described here. While in the complex including CipA dockerin the X module only interacts with the first

calcium-binding loop, in the complex described here the X domain stabilizes both calcium binding loops (Figure 4.7-A,B). In addition, the higher number of contacts in the ScaA-CipAXDoc complex should lead to a more stable and rigid structure that, as previously suggested, would reduce the entropic cost arising from a tightening of the isolated type II Doc structure upon type II cohesin binding (Adams *et al.*, 2006).

Figure 4.7| Complex interfaces between the dockerin and the X module (A and B) and between the dockerin and the cohesin (C, D, E and F).



A), C) and **E)** correspond to the ScaC2-CipBXDoc complex. **B), D)** and **F)** correspond to the ScaA-CipAXDoc complex.

4.1.3.4. The ScaC2-CipBXDoc complex interface

Similarly to what is observed in the previously described ScaA-CipAXDoc structure, both helices 1 and 3 of CipB dockerin interact with the planar surface formed by ScaC2 cohesin 8-3-6-5 face. However, the number of contacts with dockerin helix 3 predominates and is dominated by Asn146, Leu147 and Phe148. Overall the dockerin in the ScaC2-CipBXDoc complex lays slightly further away from the cohesin when compared with CipA in the type II complex described by Adams *et al.* (2007). The interaction surface is predominantly hydrophobic and is defined by residues Gly35, Ile36, Gln37, Asn76, Leu78, Thr80, Val82, Asp84, Asn91, Tyr92, Ala93, Ser94, Cys95, Tyr96, Val97, Tyr98, Trp99, Arg135, Phe136, Pro138, Asn139, Leu145, Val146, Ile147, Try150, Gly151 and Gln153 from the 8-3-6-5 face and loop region leading to the crowning helix between strands 6 and 7 of the cohesin module and residues Met114, Val117, Met118, Ser121, Phe124, Gly125, Thr126, Arg127, Asp142, Gly143, Ala144, Asn146, Leu147, Phe148, Ile150, Ala151, Ile154, Arg155 and Phe157 from the dockerin module. Interestingly, in contrast to the CohScaA-XDocCipA, type II complex (PDB ID code: 2b59) where an extensive hydrogen bond network exists, in the *C. thermocellum* type II protein heterodimer described here there is no significant hydrogen bonding present at the complex interface. In fact, only two hydrogen bonds can be identified in this complex between Nδ2 of Asn146 dockerin and O of Phe136 cohesin and N of Leu147 dockerin and Oε1 of Gln37 cohesin (Figure 4.7-C,D). Furthermore, Asp20 from the X module forms three water mediated hydrogen bonds with residue Gln152 of the cohesin (Figure 4.7-E,F). At this position of the ScaA-XDocCipA complex there is a hydrogen bond established between residues Ser20 (X module) and Glu167 (Doc). Taken together the data reveals that the type II complex interface described here is more hydrophobic than the contact region established within the ScaA-CipAXDoc complex (Adams *et al.*, 2006). The lower number of hydrogen bond contacts in the complex described here confirms that the two type II complex interfaces known are relatively different.

4.1.3.5. Comparison of *C. thermocellum* type II cohesin-dockerin complexes

The structure of CipA dockerin in complex with ScaA cohesin (PDB ID code: 2b59) was superimposed with the structure of the type II complex described here. The primary sequences of CipA and CipB dockerins are quite different at the putative interacting residues and somehow the alignment presented in Figure 4.2 suggests that Helix 1 could constitute the main interacting helix in CipB. Since there is a remarkable variation in cohesin sequences and considering that ScaC2 cohesin presented the highest levels of affinity for both cohesins we decided to solve the structure of the ScaC2-CipBXDoc complex. When the two complexes are overlayed there are clear regions where topological variations at both cohesins and dockerins are apparent. The superposition of the two dockerins revealed that CipB Gly125 is substituted in CipA dockerin by a Asn residue. CipA Asn122 makes three

very important hydrogen bond contacts with the ScaA cohesin, which are obviously missing in the ScaC2-CipBXDoc complex. This difference also reflects inherent topological differences in ScaA cohesin β-flap interrupting strand 8, where ScaC2 Phe136, which is positioned towards the centre of the complex, is replaced by a Ser residue that allows the β-flat to be positioned more distantly from the cohesin. The position of Phe136 at the ScaC2 will clash with the side chain of CipA Asn122, if the two complexes are overlayed, strongly suggesting that CipA dockerin will be unable to bind the cohesin of ScaC2, at least through this orientation. However, when CipA dockerin is overlayed with its molecule rotated by 180º residues Asn122 and Gly155 will superimpose very well, and Gly155 from CipA will overlay very well with Gly125 of the non rotated CipB dockerin. In addition, conservation of other cohesin-contact residues is higher when the structure of CipB dockerin is superimposed with CipA molecule rotated by 180º. Overall these observations suggest that CipA dockerin should bind ScaC2 cohesin in a reverse orientation when compared with ScaA. In addition, when the sequence of CipB dockerin is overlayed with its 180º rotated molecule Asn158 superimposes with Gly125, suggesting that the rotated dockerin will be adequate to bind the ScaA cohesin. Overall these observations suggest that *C. thermocellum* dockerins contain two different cohesin binding interfaces that will discriminate from the different type II cohesins. These observations are further supported by data presented by Haimovitz *et al.* (Haimovitz *et al.*, 2008) while exploring the differences in specificities in type II dockerins from different species. While comparing the capacity of the type II dockerin from *A. cellulolyticus* to recognize type II cohesins from *C. thermocellum*, it was observed that this dockerin only bond to the first cohesin of ScaB and ScaA. In contrast the dockerin was unable to recognize the second cohesin of ScaC (the one in the complex presented here) and the fourth cohesin of ScaB. Interestingly, in contrast to what is observed for *C. thermocellum*, the type II dockerin of *A. cellulolyticus* is highly symmetric with a strong conservation of the putative residues that participate in cohesin recognition at the two interfaces. The fact that an Asn residue exclusively occupies the position 22 of the two calcium binding loops supports the capacity of these modules to bind ScaA while being unable to interact with the ScaC2 cohesin. At the equivalent position occupied by Phe136 in the ScaC2 cohesin in ScaB4 is occupied by an Ala while in ScaB1 it is a Ser as observed for ScaA. Thus, overall the data suggest that position 136 of type II cohesins is important to modulate the capacity to recognize type II dockerins, with binding mediated by a dockerin Gly if the cohesin residue is hydrophobic or by an Asn if the position 136 is occupied by a Ser or a Thr. The above presented arguments suggest that indeed type II dockerins present two different cohesin binding interfaces and will bind different cohesins in a different manner. Interestingly, this hypothesis is also supported by the affinities studies presented above. The data suggest that the two *C. thermocellum* dockerins recognize the various *C. thermocellum* cohesins with similar affinities; the qualitative competition experiments suggest that the

cohesins could not discriminate the two different dockerins. Thus, this is only possible if the two cohesin binding interfaces are highly homologous a possibility that is much more favoured when the two dockerins are overlaid with their rotated homologues.

Another major difference when the two type II complexes are compared is observed while analysing the orientation the side chain of the highly conserved Tyr located at cohesin helix. This helix seems to play a critical role for positioning the bound type II dockerin in the complex. Thus, in ScaC2 the side chains of the adjacent residues Asp100 and Ser101 force the side chain of Tyr98 to move into the direction of the dockerin binding face. This change in the side chain conformation of Tyr implies that Phe of the type II dockerin will move away from the cohesin. In contrast, in ScaA, the adjacent residues to the pivotal Tyr are Gly and Ala and thus the side chain of Tyr is in a flat position allowing the Phe to be in closer proximity of the cohesin. Overall, this change contributes to the closer proximity of the dockerin observed in the ScaA-CipAXDoc complex.

4.1.3.6. Functional importance of residues at the surface of the type II dockerin for cohesin recognition

The binding CipB XDockerin to their various type II cohesin partners was assessed by ITC. Initial experiments were performed with the type II cohesin of ScaC2. As reported by Adams *et al.* (2006) the very high affinity that characterizes the interaction of the two dockerins for these two cohesins ($K_a \geq 10^9$ M^{-1}) did not allow an accurate determination of K_a. Data presented above revealed that ScaE6 is the cohesin to which the two *C. thermocellum* type II dockerins bind with the lowest affinities. Thus, ITC experiments were performed using the ScaE6 cohesin. The data, presented in Table 4.4 and exemplified in Figure 4.8, revealed that the affinity of the interaction between the CipB XDoc module and the sixth cohesin of the extracellular scaffoldin is much smaller when compared with the ScaC2cohesin and present a K_a at 318.15 K of 1.27 x 10^7 M^{-1} with a ΔH of -8.27 kcal mol^{-1} and a $T\Delta S$ of -2,04 kcal mol^{-1}. It is interesting to note that the apparently hydrophobic nature of the type II interaction is associated with a gain in enthalpy. While it is currently unclear why the thermodynamic forces driving ligand binding are not reflected in the nature of the amino acids that mediate cohesin-dockerin recognition, the thermodynamic parameters are likely to be influenced by changes in solvation, which cannot easily be explained by static crystal structures.

It is difficult to rationalize why the ScaE6 cohesin presents a lower capacity to bind the two type II dockerins from *C. thermocellum*. However, the presence of an Arg at the middle of β-strand 6 of ScaE, which is an important platform for dockerin recognition in the two type II structures revealed, may contribute to destabilize the cohesin binding interface (see alignment of Figure 4.3). In ScaD, which also displays a lower capacity to bind XDoc

sequences, this Arg is replaced by a Lys. The functional importance of these amino acid residues substitutions to modulate dockerin recognition remains, however, to be determined. These basic substitutions are also conserved in all cohesins from ScaE protein suggesting that all seven cohesins should display a lower capacity to bind type II dockerins. The biological significance of the lower affinity of all ScaE cohesins to type II dockerin containing proteins remains to be determined. It is possible that binding to the extracellular scaffoldin only occurs when all the cell surface proteins are already saturated with CipA or CipB.

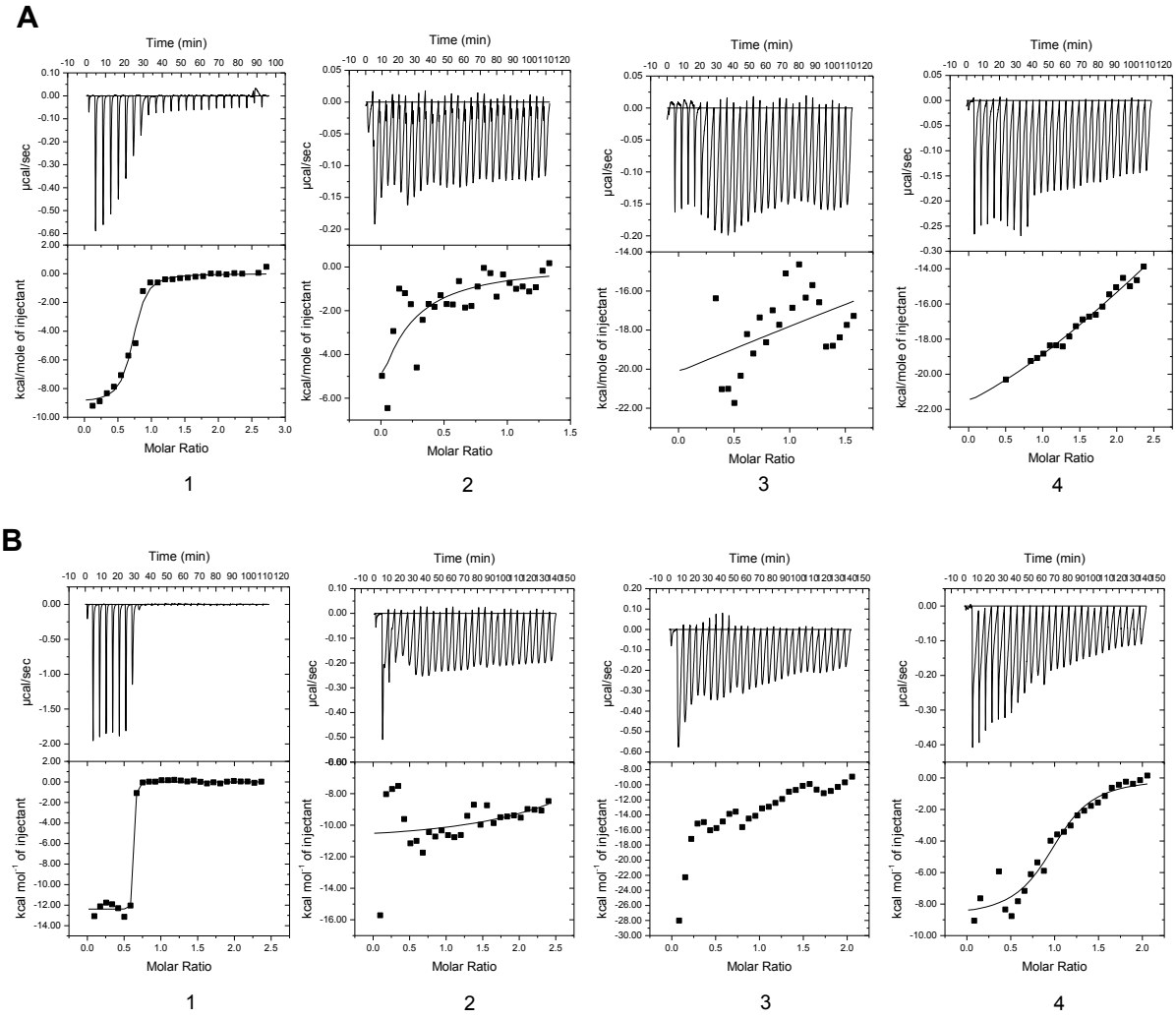
Table 4.4| Thermodynamics of type II dockerin-cohesin interactions.

Cohesin	Dockerin	K_a M^{-1}	ΔG^o kcal mol $^{-1}$	ΔH^o kcal mol $^{-1}$	$T\Delta S^o$ kcal mol $^{-1}$
ScaE6	CipB	1.21E7 ± 2.96E6	-6.23±0.19	-8.27±0.19	-2.04
	M114A	7.17E6±6.25E5	-9.99±1.06	-12.99±1.06	-3.00
	M118A	6.96E6±1.73E5	-9.95±0.62	-25.54±0.62	-15.59
	S121A	9.16E6±2.49E5	-3.72±0.16	-6.93±0.16	-3.21
	F124A	*Nd*	*Nd*	*Nd*	*Nd*
	L147A	*Nd*	*Nd*	*Nd*	*Nd*
	F148A	*Nd*	*Nd*	*Nd*	*Nd*
	I154A	1.08E7±1.92E6	-10.25±0.94	-17.26±0.94	-7.01
ScaC2	CipB	*Id*	-11.32±0.12	-12.39±0.12	1.07
	M114A	*Id*	-11.83±0.39	-10.43±0.39	1.40
	M118A	*Id*	-11.57±0.17	-8.66±0.17	2.91
	S120A	*Id*	-10.84±0.28	-10.28±0.28	0.56
	F124A	*Nd*	*Nd*	*Nd*	*Nd*
	L147A	*Nd*	*Nd*	*Nd*	*Nd*
	F148A	6.19E5±1.90E5	-8.42±0.48	-8.83±0.48	-0.42
	I154A	1.20E6±2.59E5	-9.13±0.74	-9.75±0.74	-0.63

Thermodynamic parameters were determined at 318.15 K (ScaE-6) or 328.15 K (ScaC-2). *Nd* means that the values were too low to be determined. *Id* means that the values were too high to be determined by ITC.

The importance of key residues identified at the dockerin cohesin binding region was probed by ITC. Site-directed mutagenesis data (Table 4.4) revealed that Ala substitution of CipB dockerin residues Phe124, Leu147 and Phe148 results in the complete abolition of ScaE6 binding. In contrast, dockerin amino acid substitutions Met114Ala, Met118Ala, Ser121Ala and Ile154Ala had little influence on the affinity of the dockerin for cohesin ScaE6. Affinities were also lower, than the detection limit, for Phe124A and Leu147A substitutions when dockerin mutant derivatives were probed against ScaC2 cohesin and K_a~10^5-10^6 were observed for Phe148Ala and Ile154Ala proteins (Figure 4.8).

Figure 4.8| Examples of the isothermal titration calorimetry (ITC) experiments between the *wild-type* CipB Xdoc, its mutant derivatives Phe124A, Leu147A and Phe148A and the *wild-type* cohesins ScaE6 A) and ScaC2 B).



The upper parts of each panel show the raw heats of binding, whereas the lower parts are the integrated heats after correction for heat of dilution. The curve represents the best fit to a single-site binding model. **1)** Cohesin plus CipB XDoc *wild-type*. **2)** Cohesin plus CipB XDoc Phe124A. **3)** Cohesin plus CipB XDoc Leu147A. **4)** Cohesin plus CipB XDoc Phe148A.

Overall these data suggest that Phe124, Leu147 and Phe148 dominate cohesin recognition by CipB type II dockerin. Inspection of the equivalent amino acid residues in the type II dockerins of CipA and *A. cellulolyticus* revealed that Phe124 is conserved in the three proteins. Surprisingly, there is considerable deviation in residues at position 11, occupied by a Leu in CipB and *A. cellulolyticus* but Met in CipA dockerin, and 12 of the second calcium binding loop which is occupied by Glu in the *A. cellulolyticus* protein. In addition, as described above, *A. cellulolyticus* dockerin is highly symmetric being the putative cohesin-interacting residues presented at the N- and C-terminal duplicated segments conserved. In contrast, *C. thermocellum* type II dockerins lack this internal symmetry. Inspection of CipA dockerin N-terminal duplicated segment reveals more conservation at the 11 and 12 position

of the calcium binding loop. Together, these observations suggest that the lack in conservation these key residues in the two duplicated segments of *A. cellulolyticus* dockerin explain why this dockerin is unable to recognize ScaC2. In addition, the observation that the first calcium binding loop of CipA dockerin is more homologous to the equivalent region of the second duplicated segment of CipB suggests that CipA, as proposed above, should bind ScaC2 cohesin in the 180º reverse orientation as compared with CipB.

It is now well established that type I dockerins display extensive internal symmetry which spans over the N and C-terminal helices (Carvalho *et al.*, 2007; Pinheiro *et al.*, 2008). However, in type I interactions, cohesin recognition is mainly asymmetric being primarily mediated by residues at helix 1 or 3 (Carvalho *et al.*, 2007; Pinheiro *et al.*, 2008). The dual binding mode of type I dockerins may introduce significant flexibility in the quaternary structure of the cellulosome as the binding affinity of the two cohesin interacting surfaces of the dockerins is similar. As described above, the *A. cellulolyticus* type II dockerin also displays nearly identical segment repeats expressing conservation in most of the putative cohesin binding residues, in contrast to the asymmetric nature of *C. thermocellum* type II dockerins. Thus, it was proposed that *A. cellulolyticus* type II dockerin would participate in a dual binding mode, while the clostridial dockerins should recognize cohesins through a single binding mechanism (Adams *et al.*, 2006; Haimovitz *et al.*, 2008). The work presented here allowed for the first time probing the importance of specific dockerin residues in recognizing type II cohesins. The data suggest that single mutation of each of the three amino acid residues that dominate cohesin recognition have a dramatic effect in the affinities between the two modules. Hence, it is proposed that indeed CipB dockerin only presents a single ScaC2 cohesin binding region.

Inspection of the ScaC2-CipBXDoc complex overlaid with the 2-fold symmetry related dockerin reveals why this dockerin does not present a dual binding mode while interacting with ScaC2 cohesin. Thus, potential steric clashes between the dockerin Met114 with cohesin Ala93 and Cys95 and also the dockerin Met118 with the cohesin Ile147 are observed when the dockerin is rotated by 180º at the interface plane. In addition, when rotated, dockerin Ile154 would steric clash with cohesin residue Leu78. More importantly, binding of the dockerin following a 180º rotation would introduce the side chain of Val115 in the cohesin hydrophobic pocket occupied by Phe148 resulting in a less stable complex. Finally, as described above, Asn158 would be not allowed in the position occupied by Gly125 as it would steric clash with ScaC2 Phe136. Therefore, the results described here reveals that CipB dockerin presents a single ScaC2 binding interface.

4.1.4. Conclusions

Data presented here reveals that *C. thermocellum* CipB type II dockerin binds the second cohesin of ScaC through a single binding mode. The dockerin reveals a lack in conservation in the interacting residues presented at the two duplicated segments which should preclude a dual binding mode. A similar observation was deduced when the structure of the type II dockerin of CipA was solved in complex with the cohesin of ScaA. Thus, the structures of the two complexes suggest that the two dockerins would not be able to bind their protein counterparts when rotated by 180º at the interface plane. However, comparison of the two type II complexes revealed clear differences both at the cohesin and dockerin interacting residues, demonstrating that the nature of the two cohesin-dockerin binding interfaces is different. More importantly, there is a clear symmetry on the two dockerins when they are compared with its homologue domain rotated by 180º. These observations suggest that *C. thermocellum* type II dockerins present two different binding interfaces. Thus, in CipA dockerin the N-terminal face would be more appropriated to bind ScaC2 cohesin while the second duplicated was more appropriated to bind ScaA cohesin face, as described in the type II complex described by Adams *et al.* (2007). In contrast, In CipB dockerin, the second interface recognizes ScaC2 cohesin, while the N-terminal segment would be more adapted to interact with the ScaA cohesin. Thus, these data suggest that *C. thermocellum* type II dockerins present two different cohesins-binding faces each binding to a different type II cohesin scaffold. The confirmation of this hypothesis requires solving the structures of the ScaC2-CipAXDoc and ScaA-CipBXDoc type II complexes. This work is presently on-going.

5. STRUCTURAL INSIGHTS INTO A UNIQUE CELLULASE FOLD AND MECHANISM OF CELLULOSE HYDROLYSIS

5.1. *Ct*Cel124 a novel cellulase from *Clostridium thermocellum* cellulosome∞

Joana L. A. Brás[*,1], Alan Cartmell[‡,¶¶,1], Ana Luísa Carvalho[†,2], Genny Verzé[†,‡], Edward A. Bayer[¶], Yael Vazana[¶], Márcia A. S. Correia[†], José A. M. Prates[*], Supriya Ratnaparkhe[¶¶], Alisdair B. Boraston[§], Maria João Romão[†], Carlos M. G. A. Fontes[*] and Harry J. Gilbert[‡,¶¶]

*CIISA-Faculdade de Medicina Veterinária, Pólo Universitário do Alto da Ajuda, Avenida da Universidade Técnica, 1300-477 Lisboa, Portugal; [†]REQUIMTE/CQFB, Departamento de Química, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, 2829-516 Caparica, Portugal; [‡]Institute for Cell and Molecular Biosciences, Newcastle University, The Medical School, Newcastle upon Tyne NE2 4HH, United Kingdom; [¶¶]Complex Carbohydrate Research Centre, University of Georgia, Athens, Georgia 30602-4712, USA.;[‡]Biocrystallography Laboratory, Department of Biotechnology, University of Verona, Verona, Italy; [¶]Department of Biological Chemistry, The Weizmann Institute of Science, Rehovot 76100 Israel; [§]Biochemistry and Microbiology, University of Victoria, PO Box 3055 STN CSC, Victoria, BC, Canada. [1]Equal contribution.

Adapted from:Brás *et al.,* (2011) Proc Natl Acad Sci U S A, 108(13):5237-42.

Abstract

Clostridium thermocellum is a well characterized cellulose-degrading microorganism. The genome sequence of *C. thermocellum* contains a number of proteins that contain type I dockerin domains, which implies that they are components of the cellulose degrading apparatus, but display no significant sequence similarity to known plant cell wall degrading enzymes. Here we report the biochemical properties and crystal structure of one of these proteins designated *Ct*Cel124. The protein was shown to be an *endo*-acting cellulase that displays a single displacement mechanism and acts in synergy with Cel48S, the major cellulosomal *exo*-cellulase. The crystal structure of *Ct*Cel124 in complex with two cellotriose molecules, determined to 1.5 Å, displays a superhelical fold in which a constellation of α-helices encircle a central helix that houses the catalytic apparatus. The catalytic acid, Glu96, is located at the C-terminus of the central helix, while there is no candidate catalytic base. The substrate-binding cleft can be divided into two discrete topographical domains in which the bound cellotriose molecules display twisted and linear conformations, respectively, suggesting that the enzyme may target the interface between crystalline and disordered regions of cellulose.

∞ The student contributed in the following methodologies: cloning, expression and purification, enzyme assays and crystallization.

5.1.1. Introduction

The plant cell wall is an important biological and industrial resource. Deconstruction of this composite structure provides nutrients that are utilized by microorganisms from a variety of ecosystems. Indeed, mammalian herbivores derive a significant proportion of their energy from the hydrolysis of plant cell wall polysaccharides by their symbiotic microbiota. The deconstruction of the plant cell wall is also of growing environmental and industrial significance as the demand for renewable sources for bioenergy and substrates for the chemical industry increases (Himmel & Bayer, 2009). The major plant cell wall polysaccharide is cellulose, a β–1,4-glucose polymer (Nishiyama, Johnson, French, Forsyth, & Langan, 2008), which is hydrolyzed by a range of glycoside hydrolases (cellulases). These enzymes display *endo* (endo-β1,4-glucanase), *exo* (cellobiohydrolases that release cellobiose from cellulose) or *endo*-processive (cleaves internally and then acts in a processive manner on the generated product) modes of action. The classical paradigm for cellulose hydrolysis is the endoglucanase-cellobiohydrolase synergy model, which is based, primarily, on aerobic fungal cellulase systems. (Gilbert, 2010) This model, however, does not accurately reflect clostridial systems, where endo-processive GH9 enzymes are central to the degradative process (Tolonen, Chilaka, & Church, 2009), or *Cytophaga hutchinsonii*, which appears to lack classical cellobiohydrolases. Cellulases are currently grouped into 11 of the 123 glycoside hydrolase sequence-based families (GHs) within the CAZy database (Cantarel *et al.*, 2009). As there is limited conservation in the catalytic apparatus and the overall fold between the cellulase-containing families, these enzymes are generally thought to have evolved by convergent evolution (Gilbert, 2010).

Clostridium thermocellum is a well characterized cellulose-degrading microorganism (Bayer *et al.*, 2004; Fontes & Gilbert, 2010). The bacterium synthesizes a large multienzyme complex, known as the “cellulosome”, which catalyzes the degradation of the plant cell wall (Bayer *et al.*, 2004; Fontes & Gilbert, 2010). Enzymes are recruited into the *C. thermocellum* cellulosome through the interaction of their type I dockerin modules with the multiple cohesin domains present on the scaffoldin (defined as CipA) (reviewed in Bayer *et al.* (2004); Fontes & Gilbert (2010)). The genome of *C. thermocellum* encodes 72 proteins containing type I dockerins. These proteins, therefore, are likely to be components of the cellulosome and thus contribute to cellulose or, in a wider context, plant cell wall deconstruction. Synergy experiments have identified two cellulosomal enzymes, an exo-acting GH48 cellobiohydrolase that acts from the reducing end of cellulose chains and the cellotetraose producing endo-processive GH9 endoglucanase, Cel9R (Zverlov, Schantz, & Schwarz, 2005), as central components of the *C. thermocellum* cellulase system (Fierobe *et al.*, 2002). It has been suggested that, by generating cellotetraose as the major product from cellulose,

likely through the action of Cel9R, *C. thermocellum* minimizes the utilization of ATP during import of glucose units (Zhang & Lynd, 2005). The cellulase activity obtained by combining cellulosomal enzymes *in vitro*, however, is considerably lower than that displayed by the cellulosome presented on the surface of *C. thermocellum*. It is possible, therefore, that proteins, currently of unknown function, either in the cellulosome or displayed on the surface of *C. thermocellum*, make a significant contribution to the cellulose degrading capacity of the bacterium. While most of the cellulosomal proteins can be assigned to glycoside hydrolase, esterase or polysaccharide lyase families, 14 of the predicted type I dockerin containing proteins display little sequence similarity to enzymes in the CAZy database. Thus, some of these non-CAZy *C. thermocellum* proteins may comprise novel enzymes that target the hydrolysis of components of the plant cell wall such as cellulose. One of these hypothetical proteins, Cthe_0435 (hereafter designated as *Ct*Cel124), is upregulated when *C. thermocellum* is cultured on crystalline cellulose (Raman *et al.*, 2009), suggesting that *Ct*Cel124 plays a role in the hydrolysis of the glucose polymer.

Here we show that *Ct*Cel124 is an endoglucanase that acts in synergy with the major exocellulase of the *C. thermocellum* cellulosome to degrade crystalline cellulose. The crystal structure of *Ct*Cel124 reveals a substrate binding cleft in which the bound cellooligosaccharides adopt two distinct conformations, indicating that the enzyme targets the interface between crystalline and amorphous regions of cellulose. The active site of the cellulase displays structural conservation to GH23 enzymes, a family that contains inverting lysozymes and lytic transglycosylases.

5.1.2. Material and Methods

5.1.2.1. Cloning, expression and purification of *Ct*Cel124$_{CD}$

DNA encoding the C-terminal module of *Ct*Cel124 (designated *Ct*Cel124$_{CD}$; residues 131-350) was amplified from *C. thermocellum* strain ATCC 2745 genomic DNA by PCR, using primers listed in Table S5.2 (annex), and the resultant DNA was cloned into pET28a to generate pCel124. *Ct*Cel124$_{CD}$ encoded by pCel124 contains a His$_6$-tag. To generate *Ct*Cel124$_{CD}$ fused to the non-catalytic carbohydrate (cellulose)-binding module CBM3a (*Ct*Cel124$_{CD}$-CBM3a) overlapping PCR was used deploying primers that amplified the DNA sequences encoding the two modules. Expression of the proteins was achieved by adding isopropyl β-D-thiogalactopyranoside (IPTG) (1 mM final concentration) to mid-exponential phase cultures of *E. coli* BL21(DE3) harbouring pCel124 with incubation for a further 16 h at 37 °C. The His$_6$-tagged recombinant protein was purified from cell-free extracts by immobilized metal ion affinity chromatography (IMAC) using standard methodology (Pell *et al.*, 2004). To produce seleno-L-methionine *Ct*Cel124$_{CD}$ the *E. coli* methionine auxotroph B834(DE3) containing pCel124 was cultured as described by Charnock (Charnock *et al.*,

2000). The recombinant protein was purified as described above except all buffers were supplemented with 5 mm β-mercaptoethanol. The production of recombinant Cel48S and the CBM3a-Coh construct (CBM3a fused to the type I cohesin3 from CipA, the primary scaffoldin of *C. thermocellum*) are described previously (Barak *et al.*, 2005; Fierobe *et al.*, 2002; Haimovitz *et al.*, 2008).

5.1.2.2. Enzyme assays

In enzyme assays using soluble polysaccharides the reactions were carried out in 50 mM MES buffer, pH 5.5, containing 1 mg/ml BSA and 0.5 % of the target polysaccharide. Reactions, which were incubated at 50 ^{o}C contained, typically, 15 nM of enzyme. At regular time points aliquots were removed and reducing sugar present was determined (Miller, 1959), using glucose to construct the standard curve. In enzyme assays using Avicel (Sigma Chem. Co.) and phosphoric acid swollen cellulose [PASC; prepared as described previously (Hall *et al.*, 1995)] as substrates, the reaction mixture consisted of 500 nM enzyme in 100 mM acetate buffer, pH 5.0, 24 mM $CaCl_2$, and insoluble cellulose at a concentration of 2 % (w/v). The reactions, which were carried out at 50 ^{o}C, were terminated by immersing the sample tubes in ice water. After centrifugation to remove insoluble substrate, the supernatant was assayed for reducing sugar. When assaying Cel48S in the presence of CBM3a-Coh, equimolar quantities of the protein partners were incubated for 2 h at 37 ^{o}C (without the substrate) preceding the assay. This ensured that Cel48S was appended to the CBM3a. The hydrolysis of cellooligosaccharides (30 μM), conducted in 50 mM MES buffer, was assessed by the rate of substrate depletion using a previously described Dionex-HPAEC method to detect cellooligosaccharides (Hall *et al.*, 1995). HPAEC was also used to identify the reaction products released from PASC. The α and β configuration of the anomeric carbon of cellotriose, generated by the hydrolysis of cellopentaose by *Ct*Cel124$_{CD}$, was determined as described previously (Braun, Meinke, Ziser, & Withers, 1993; Hall *et al.*, 1995).

5.1.2.3. Binding of the *Ct*Cel124$_{CD}$ mutant E96A to ligands

Binding was determined by isothermal titration calorimetry (ITC) and by depletion binding isotherms. ITC measurements were made at 40 ^{o}C following standard procedures (Charnock *et al.*, 2000) using a Microcal ITC$_{200}$ calorimeter in 50 mM MES buffer, pH 5.5. During a titration experiment, the protein sample 88 μM was injected with 20 x 2 μl aliquots of 2 mM cellohexaose. For experiments with regenerated cellulose (RC), prepared as described previously (Boraston *et al.*, 2001), the ligand was in the cell at 29.4 mg/ml, and the protein (835 μM) was the titrant. Integrated heat effects, after correction for heats of dilution, were analyzed by nonlinear regression using a single site-binding model (Microcal Origin, version 2.9) Thermodynamic parameters were calculated using the standard thermodynamic equation: $-RT \ln K_a = \Delta G = \Delta H - T\Delta S$. Depletion binding isotherms were performed in 50

mM MES buffer, pH 7.0, at protein concentrations ranging from 1 to 30 μM. Protein was added to 1 mg of RC in a final volume of 1 ml and incubated for 1 h with gentle mixing at 40 ^{o}C. The polysaccharide was centrifuged at 13000*g* for 1 min, and the A_{280} of the supernatant was measured to quantify the amount of free protein remaining after binding. Bound protein was calculated from the total minus the free protein. The data were analyzed by nonlinear regression using a standard one-site binding model (GraphPad Prism, v5), and the N_0 and K_a values were obtained from the regressed isotherm data.

5.1.2.4. Crystallization of *Ct*Cel124$_{CD}$ and data collection

Structure determination was by the Single Wavelength Anomalous Dispersion (SAD) method, using the anomalous diffraction of the selenium atoms, incorporated in the protein. *Ct*Cel124$_{CD}$ was crystallized with an equal volume (1 μl) of protein (60 mg/ml in solution with 10 mM cellohexaose) and reservoir solution (8 % (v/v) tacsimate pH 5.0 and 20 % (w/v) PEG 3350). Glycerol (30 % w/v) was added as the cryoprotectant. The SAD experiment was conducted on beamline ID14-EH4 at the European Synchrotron Radiation Facility (ESRF) at Grenoble, France, using an ADSC Quantum-4 CCD detector. Data were collected at 0.954 Å wavelength. A total of 120 images with 1° oscillation for 5 seconds were collected. The data were processed with DENZO and the HKL2000 package and scaled with SCALA (Winn *et al.*, 2011).

5.1.2.5. Phasing, Model Building and Refinement

Patterson maps were deconvoluted by calculation of anomalous Patterson maps. The positions of the seleniums were refined, and phases were calculated using SHARP/autoSHARP (La Fortelle & Bricogne, 1997). Subsequently, a cycle of phase improvement was applied using the program DM (Cowtan & Main, 1993) where phases were modified and extended to 1.5 Å resolution yielding a figure of merit of 0.91 and an interpretable electron-density map. A model comprising 243 amino acids was built from the initial map with program COOT (Emsley & Cowtan, 2004). Water molecules and alternative conformers were added using ARP/WARP («The CCP4 suite», 1994, p 4) , and refinement with REFMAC5 (Winn *et al.*, 2001) was performed as deemed appropriate from the behaviour of the cross-validation (R_{free}) subset of reflections (10 %). Solvent molecules were added in the final stages of refinement according to hydrogen bond criteria and only if their B factors refined to reasonable values and if they improved the R_{free}. The statistics for structure refinement are shown in Table S5.3 (see annex).

5.1.3. Results and Discussion

5.1.3.1. Catalytic properties of *Ct*Cel124

*Ct*Cel124 is highly upregulated when *C. thermocellum* is cultured on crystalline cellulose (Raman *et al.*, 2009), suggesting the protein may contribute to the metabolism of the polysaccharide. To test this hypothesis the biochemical properties of the 220 residue C-terminal module of the protein (designated *Ct*Cel124$_{CD}$) was assessed. The data, summarized in Table 5.1, show that the enzyme hydrolyzed barley β-glucan, a β-1,3-β-1,4 mixed linked glucan, phosphoric acid swollen cellulose (PASC) and carboxymethylcellulose. The specific activity of the enzyme against β-glucan was only four-fold higher than the value for PASC. The initial reaction products released from PASC ranged from cellotriose to cellohexaose (Figure 5.1). Such a profile is typical of *endo*-acting enzymes and thus *Ct*Cel124 appears to be an endo-β-1,4-glucanase. The difference in activity between the soluble and insoluble polysaccharides is relatively modest compared to, for example, GH5 endoglucanases and GH9 endo-processive endoglucanases, which generally display a much stronger preference for β-glucan (Hazlewood, Davidson, Laurie, Romaniec, & Gilbert, 1990).

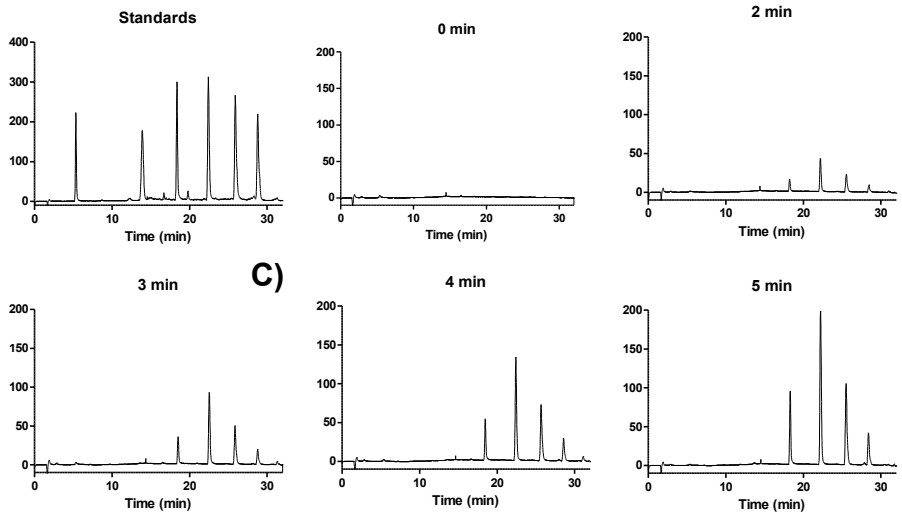
To explore the capacity of *Ct*Cel124 to disrupt the plant cell wall the catalytic module was incubated with sections of *Arabidopsis* stem, which were subsequently stained by calcofluor that binds predominantly to cellulose, and CBM9 fused to green fluorescent protein. CBM9 binds to the reducing end of cellulose and xylan chains and thus provides a direct read out of cellulose hydrolysis (Boraston *et al.*, 2001). The Calcofluor White staining data (Figure 5.1) show that the primary cell walls were considerably thinner and significantly disrupted, after cellulase treatment. Although CBM9 did not bind to untreated cell walls, after incubation with *Ct*Cel124$_{CD}$, the protein stained the secondary and primary cell walls (Figure 5.1). These data indicate that *Ct*Cel124 is able to attack cellulose embedded in cell walls. These promising data suggest that combining *Ct*Cel124, with other *exo*- and other *endo*-acting cellulases, may have a significant effect on cellulose degradation within intact cell walls.

*Ct*Cel124$_{CD}$ was approximately 20-fold more active against cellohexaose than cellopentaose (Table 5.1), but displayed no activity against cellotetraose. *Ct*Cel124$_{CD}$ hydrolyzed cellohexaose predominantly to cellotriose, whereas cellopentaose was converted exclusively to cellobiose and cellotriose (Figure S5.1 in annex). These data indicate that *Ct*Cel124 contains six dominant subsites extending from -3 to +3 (defined using standard nomenclature for subsite topology in glycoside hydrolases (Davies, Wilson, & Henrissat, 1997)). The small amount of cellotetraose and cellobiose, released from cellohexaose, is consistent with the weak -4 subsite (binding of cellohexaose from subsites -4 to +2 will generate cellotetraose

A)

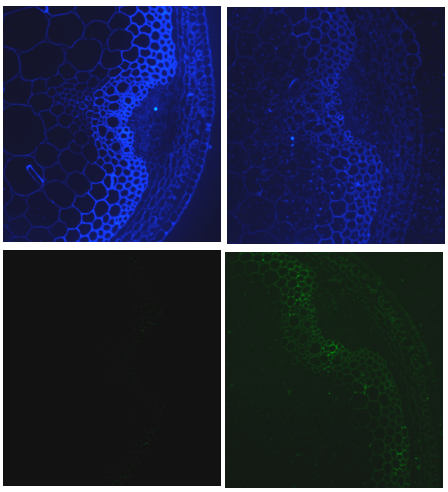
and cellobiose), identified through the crystallization of the enzyme in complex with substrate (see below).

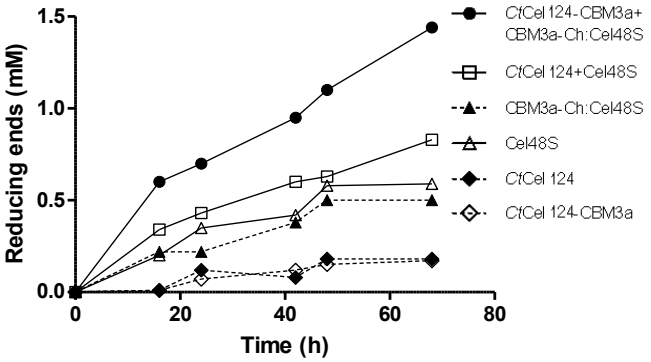
Figure 5.1| Catalytic activity of *Ct*Cel124$_{CD}$.



B)

C)





A) The HPAEC analysis of the reaction products generated by *Ct*Cel124$_{CD}$ from phosphoric acid swollen cellulose. The cellulose at 5 % (w/v) was incubated with 2 μM of *Ct*Cel124$_{CD}$ and at various time points the release of cellulooligosaccharides was analyzed by HPAEC.
B) The hydrolysis of sections of Arabidopsis stem tissue. Sections (i) and (iii) are untreated while (ii) and (iv) were incubated with 2 μM *Ct*Cel124$_{CD}$ for 16 h. Sections (i) and (ii) were stained with calcofluor, while sections (iii) and (iv) were probed with CBM9 fused to GFP.
C) The kinetics of Avicel hydrolysis by *Ct*Cel124$_{CD}$ and Cel48S. Dashed lines indicate a single enzyme activity and solid lines indicate activity of the two enzymes in combination. In Cel48S: CBM3a-Coh, Cel48S (which contains a type I dockerin) was pre-incubated with CBM3a-Coh (CBM3a fused to a type I cohesin) to generate Cel48S attached to CBM3a through the dockerin-cohesin interaction.

Table 5.1| Catalytic activity of *wild type* and mutants of *Ct*Cel124$_{CD}$.

Enzyme	Substrate	Specific activity[a]	k_{cat}/K_M (min^{-1} M^{-1})
Cel124$_{CD}$	β-glucan	$1.3 \times 10^4 \pm 4.0 \times 10^3$	-[b]
Cel124$_{CD}$	PASC	$3.2 \times 10^3 \pm 5.3 \times 10^2$	-
Cel124$_{CD}$	CMC	$9.6 \times 10^2 \pm 1.4 \times 10^2$	-
Cel124$_{CD}$	Avicel	$7.4 \times 10^{-1} \pm 2.8 \times 10^{-2}$	
Cel124$_{CD}$	Lichenan	$6.1 \times 10^3 \pm 2.1 \times 10^2$	
Cel124$_{CD}$	Chitin	NA[d]	
Cel124$_{CD}$	Chitosan	NA	
Cel124$_{CD}$	Cellohexaose[c]	-[b]	$1.0 \times 10^4 \pm 2.0 \times 10^2$
Cel124$_{CD}$	Cellopentaose	-	$4.5 \times 10^2 \pm 8.5 \times 10^1$
Cel124$_{CD}$	Cellotetraose	-	NA
Cel124$_{CD}$	Chitohexaose	-	NA
Cel124$_{CD}$	2,4 DNP-cellotriose	-	$1.1 \times 10^2 \pm 3.6 \times 10^0$
Cel124$_{CD}$ E96A	β-glucan	NA	
Cel124$_{CD}$ E96A	Cellohexaose	-	NA
Cel124$_{CD}$ E96A	2,4 DNP-cellotriose	-	$7.9 \times 10^1 \pm 1.2 \times 10^1$
Cel124$_{CD}$ N188A	Cellohexaose	-	$1.4 \times 10^2 \pm 3.1 \times 10^1$
S110A	Cellohexaose	-	$4.1 \times 10^3 \pm 1.0 \times 10^2$
S110E	Cellohexaose	-	$5.5 \times 10^3 \pm 6.6 \times 10^2$
S110D	Cellohexaose	-	$3.2 \times 10^3 \pm 3.7 \times 10^2$

[a]Specific activity is expressed as molecules of reducing sugar produced per molecule of enzyme per min. Assays were carried out at 50 ^{0}C using 0.5 % substrate for soluble polysaccharides and 2 % substrate for insoluble polysaccharides.
[b]Dash (-) indicates activity not assessed.
[c]The substrate concentration used was 30 μM, which is >>>K_M as the increase in rate was directly proportional to substrate concentration up to 100 μM, and thus provides a direct readout of k_{cat}/K_M.
[d]NA: no activity detected.

5.1.3.2. Affinity of *Ct*Cel124 for cellulose and cellooligosaccharides

The affinity of the inactive catalytic acid mutant (E96A) of *Ct*Cel124 for cellohexaose and regenerated cellulose (RC) was determined by ITC and depletion isotherms, (Figure S5.2 and Table S5.1 in annex). The affinity (association constant, K_A) for cellohexaose is 1.5±0.07 x 10^4 M^{-1} at 40 ^{o}C. Depletion binding isotherms showed that E96A had a K_A for RC of 3.9 (±0.5) x 10^5 M^{-1} at 40 ^{o}C. Thus the affinity of the enzyme for RC is ~10-fold higher than for cellohexaose, which spans the substrate binding cleft, suggesting that the enzyme is tailored to the conformation adopted by, at least, some regions of RC, and is not optimized to bind to the twisted structure adopted by cellooligosaccharides in solution (discussed within a structural context below).

5.1.3.3. Synergy between GH48S and *Ct*Cel124.

It is well established that *exo*- and *endo*-acting cellulases act in synergy to hydrolyze cellulose (see Gilbert (2010) for review). In addition, these enzymes normally contain cellulose-specific CBMs that potentiate catalysis by recruiting the cellulases to the surface of the insoluble substrate (Hall *et al.*, 1995). In the *C. thermocellum* cellulosome the most abundant *exo*-acting cellulase is Cel48S, while the crystalline cellulose-specific CBM (CBM3a) is supplied by the non-catalytic scaffoldin CipA (Bayer *et al.*, 2004; Fontes & Gilbert, 2010). To explore the possible synergy between *Ct*Cel124 and Cel48S, the capacity of the enzymes, individually and in combination, to release reducing sugar from Avicel was assessed. The data (Figure 5.1) showed that 1.3-fold more reducing sugar was released when the two enzymes were used in combination, compared to the additive value when the two enzymes were used in isolation. These data indicate that *Ct*Cel124 and Cel48S exhibit a degree of synergy when acting on highly crystalline cellulose. When both enzymes were appended to CBM3a, which binds to crystalline cellulose, more extensive synergy (1.9-fold) was observed between the two cellulases. Thus, it is possible that the CBM may target Cel48S and *Ct*Cel124 to similar regions of Avicel and, by so doing, potentiate the synergy between the two enzymes.

The observed synergy between Cel48S and *Ct*Cel124 is consistent with previous studies showing similar potentiation in cellulose hydrolysis when *endo*- or endo-processive cellulases were combined with the GH48 *exo*-acting cellulase (Fierobe *et al.*, 2002; Zhang, Sathitsuksanoh, & Zhang, 2010). The mechanisms by which *endo*- and *exo*-acting cellulases act in synergy have been extensively explored. The favored model, at least for fungal systems, proposes that the *endo*-acting enzymes target amorphous regions of cellulose creating new termini from which *exo*-acting cellobiohydrolases can extend substrate hydrolysis into the crystalline regions of the polysaccharide (Gilbert, 2010). Such a model, however, does not explain the low degree of synergy observed between some enzyme combinations, suggesting that new termini generated by endoglucanases are not always available to the cellobiohydrolases. Indeed the significance of the *C. thermocellum* GH48 enzymes, in the capacity of the cellulosome to solubilize crystalline cellulose, has been questioned by recent studies on mutants of the bacterium lacking these enzymes. The growth rate of the GH48 knockout mutants on cellulose, the cell yield of the variants, and the activity of the cellulosome against Avicel were reduced by 40, 60 and 35 %, respectively, compared to wild-type *C. thermocellum* (Olson *et al.*, 2010). These data suggest that Cel48S certainly contributes to cellulose degradation, but the classical *endo-exo* synergy model does not fully explain the capacity of the *C. thermocellum* cellulosome to completely solubilize crystalline cellulose.

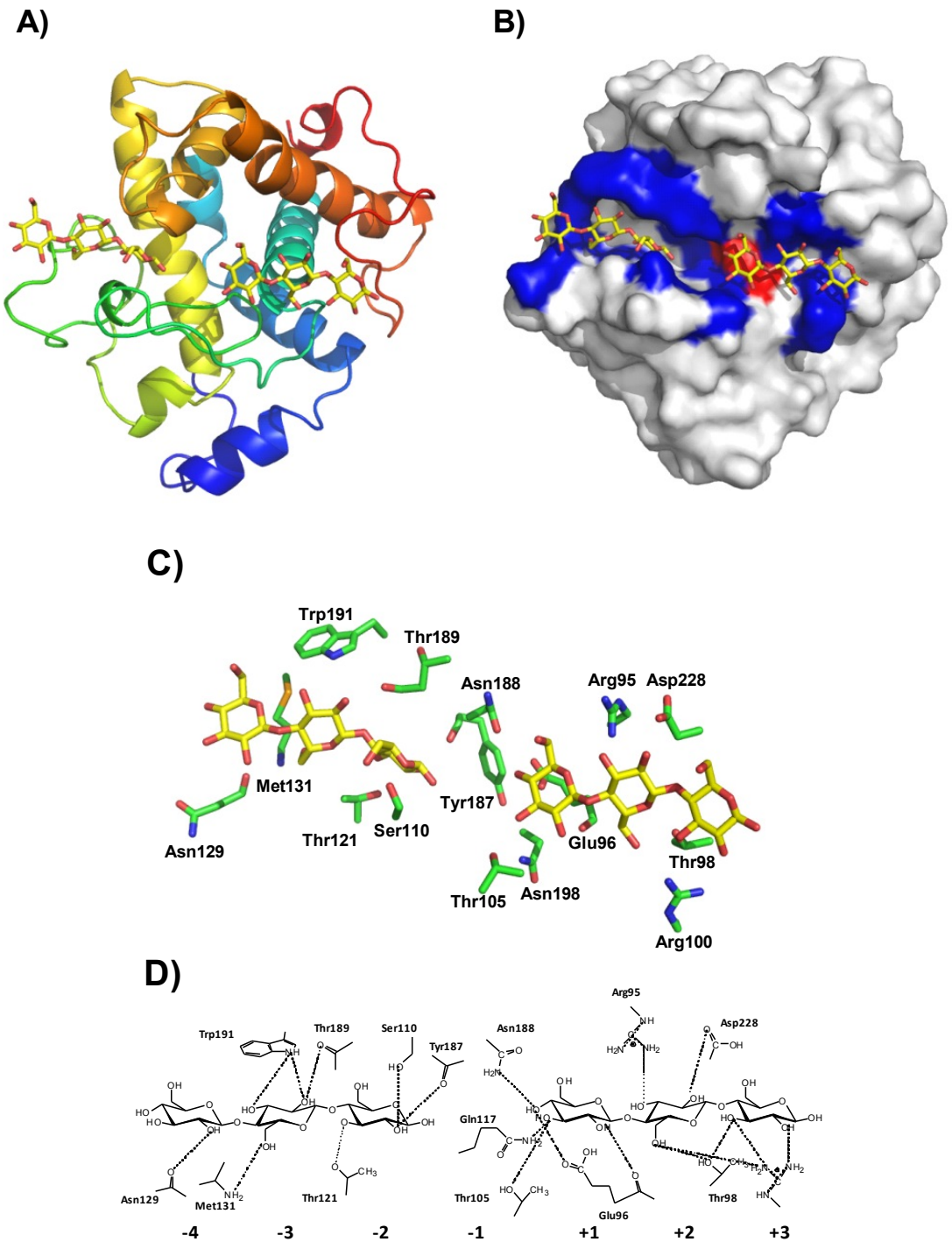
It should be emphasized that while *Ct*Cel124 contains a type I dockerin, the module displays a preference for the cohesin in the cell-envelop protein OlpC, rather than the cohesins in the

cellulosome scaffold protein, CipA (Pinheiro *et al.*, 2009). These data indicate that the enzyme is predominantly located at the cell surface of the bacterium. Lynd and colleagues showed that the cellulosome, when appended to the surface of *C. thermocellum*, is more efficient at cellulose degradation than when the complex is released into the culture media (Lu, Zhang, & Lynd, 2006). Thus, it is possible that this increased activity reflects synergistic interactions between catalytic components of the cellulosome and enzymes, such as *Ct*Cel124, which are predicted to be directly appended to the surface of the bacterium. For example, *Ct*Cel124 may create the chain ends at the amorphous-crystalline interface (see below) that are required by the cellulosomal *exo*-acting enzymes to hydrolyse cellulose.

5.1.3.4. Crystal structure of *Ct*Cel124

The X-ray crystal structure of *Ct*Cel124$_{CD}$ was determined in complex with two cellotriose (which arose through the hydrolysis of cellohexaose during crystallization). The crystal structure revealed a 210 amino acid α-helical protein containing eight α-helices and a small β-sheet comprising three anti-parallel β-strands. α-Helix-4 (α-H4) forms the hydrophobic core of the protein, while the other seven helices encircle the core helix (Figure 5.2). Thus *Ct*Cel124$_{CD}$ appears to display an α_8 superhelical fold. Such a structure has not previously been observed in cellulases families, which display the following folds: $(\beta/\alpha)_8$-barrel (GH5, GH51 and GH44), distorted α/β-barrel (GH6), β-jelly roll fold (GH7 and GH12), $(\alpha/\alpha)_6$-barrel (GH8, GH9 and GH48), β_6-barrel (GH45) and seven-fold β-propeller (GH74). It would appear, therefore, that the different cellulase families are in general the result of (functional) convergent evolution, a view reinforced by the superhelical fold displayed by the endoglucanase *Ct*Cel124.

Figure 5.2| Crystal structure of *Ct*Cel124.



A) The crystal structure of *Ct*Cel124$_{CD}$ color ramped from the N-terminus (blue) to the C-terminus (red) with cellotriose shown in stick format. **B)** The solvent accessible surface of *Ct*Cel124$_{CD}$ in complex with two cellotriose molecules occupying the subsites indicated. Amino acids that make direct interactions with the ligands are shaded blue, while the catalytic acid Glu96 is in red. **C)** The location of the residues that make direct polar contact with the two cellotriose molecules. The backbone carbonyl (Glu96, Asn126, Tyr187, Thr189) and amides (Met131) that make hydrogen bonds with the substrate are included. Amino acids (carbons in green) and ligand (carbons in yellow) are shown in stick format. **D)** A schematic of the direct polar interactions between the cellulase and the two ligand molecules. This figure and subsequent structural figures were constructed using PyMol (http://www.pymol.org/).

5.1.3.5. Structural similarity of *Ct*Cel124 to other glycoside hydrolases

A BLAST search reveals four proteins in the UNIPROT database that display significant sequence identitites (>44 %) with *e* values < e^{-36} (Figure S5.3). We therefore propose that *Ct*Cel124 comprises the founding member of a new CAZy family designated GH124. These homologous enzymes are derived from highly cellulolytic organisms that also assemble its plant cell wall degrading apparatus into cellulosomal structures. 3D structural comparisons with the PDB database reveal that black swan lysozyme G (Lyz23), a member of GH23, displays the closest structural similarity to *Ct*Cel124$_{CD}$. The proteins have a root mean square deviation of 2.1 Å for 115 aligned residues, which display 21 % sequence identity (7 % identity when the complete catalytic modules were compared). The secondary structural elements of *Ct*Cel124, apart from α-H1 and α-H3, are conserved in Lyz23 (Figure 5.3). The conformation and sequence of the critical loop connecting α-H4 and α-H5, which extends along one face of the substrate binding cleft of *Ct*Cel124 (see below) is not conserved in Lyz23. Furthermore, an additional helix, present in Lyz23 but not in *Ct*Cel124, would make steric clashes with the Glc at the +3 subsite. It is evident, therefore that significant differences in the topology and the residues in the substrate binding cleft of *Ct*Cel124 and Lyz23 explains why these two enzymes display very different substrate specificities.

5.1.3.6. Active site and catalytic mechanism of *Ct*Cel124

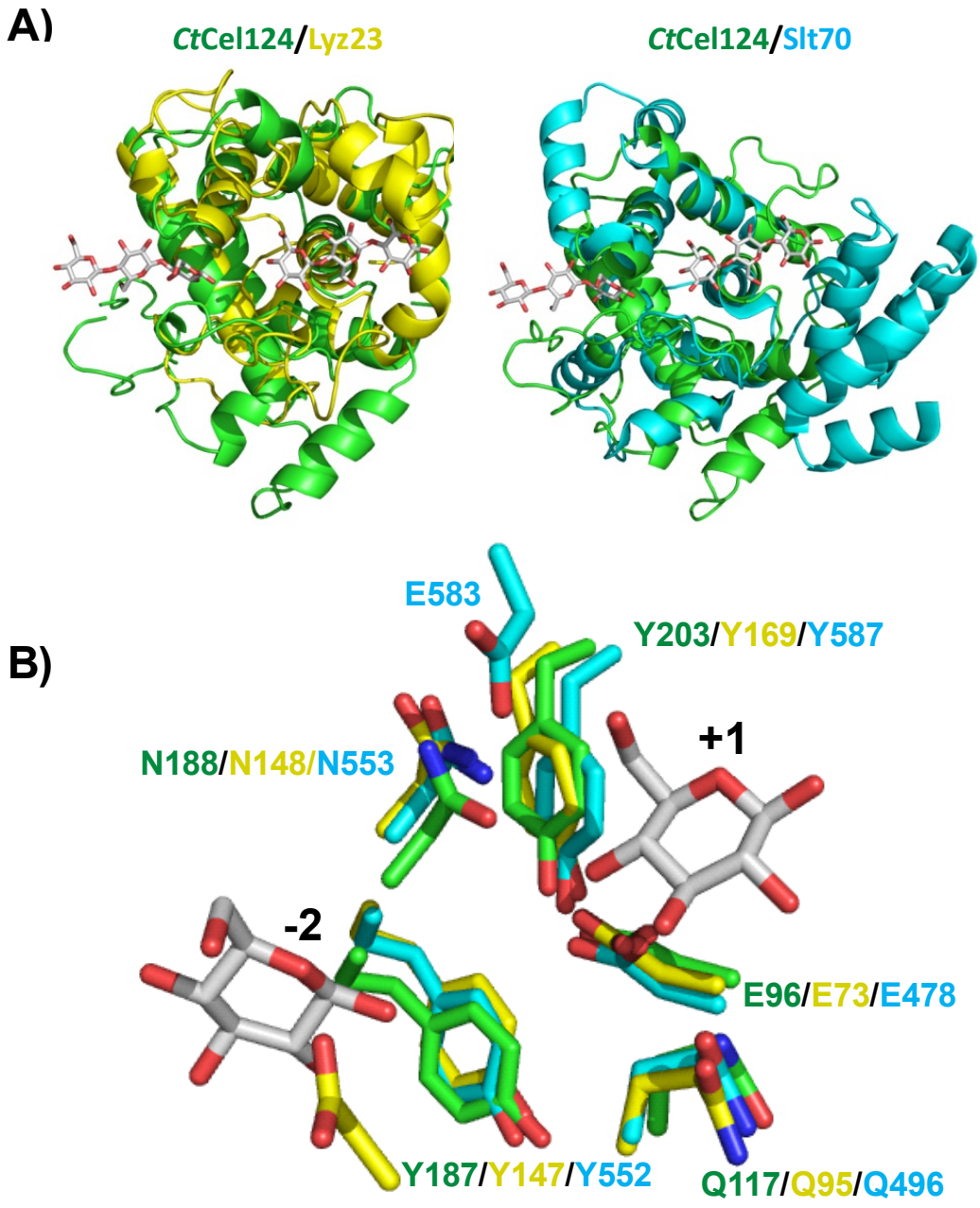
The active site of *Ct*Cel124 displays a high degree of structural similarity with Lyz23, which hydrolyses glycosidic bonds through a single displacement (inverting) mechanism, and the *E. coli* GH23 lytic transglycosylase Slt70, which cleaves glycosidic bonds through a water independent substrate assisted mechanism that ultimately leads to the formation of a 1,6-anhydro product (Figure 5.3) (see supplemental information Figure S5.4 in annex for details of the mechanism). Thus Glu96, Gln117, Tyr187, Asn188 and Tyr203, which compose the central core of the active site of *Ct*Cel124 are structurally conserved in the two GH23 enzymes. It is evident, however, that the active sites of both Lyz23 and Slt70 contain an additional, but distinct, acidic residue, Asp97 and Glu583, respectively, which are lacking in the cellulase. The structural similarities between the three enzymes present hypotheses regarding the catalytic mechanism of *Ct*Cel124. For example, it is possible that the cellulase cleaves glycosidic bonds through a lytic transglycosylase mechanism in which the 1,6-anhydro bond is hydrolyzed on enzyme. This mechanism, however, is disproven by the observation that chemically synthesized 1,6-anhydro-cellotriose is not hydrolyzed by the cellulase, and that chitin or chitosan are not substrates for the enzyme (Table 5.1 and Figure S5.5). Although the capacity of Slt70 to display a lytic activity is likely conferred, in part, by the catalytic acid base residue Glu478 (equivalent to Glu96 in *Ct*Cel124), there is no obvious acidic residue in the cellulase or, indeed, in Slt70 capable of stabilizing the negative charge of the C2 N-acetylgroup during substrate-assisted catalysis (Figure S5.4). Although the

structural similarity of *Ct*Cel124 and Lyz23 may indicate that the cellulase also acts through an inverting mechanism, the enzyme lacks an equivalent residue to Asp97 in Lyz23, the likely catalytic base (Koivula *et al.*, 2002), which is required to activate the catalytic water molecule that attacks C1 from the α face of the Glc at the -1 subsite. HPLC analysis of cellotriose, generated from cellopentaose by *Ct*Cel124, however, shows that the trisaccharide was primarily in its α configuration, (Figure S5.6) confirming that the cellulase cleaves glycosidic bonds by an acid-base single displacement mechanism leading to anomeric inversion. The lack of a candidate Asp or Glu catalytic base is evident in some inverting glycoside hydrolases. In GH6 enzymes the identity of the Brønsted base has remained particularly elusive, and a Grotthus-style mechanism, in which a remote amino acid activates an active site water *via* a string of solvent molecules, remains the likely mechanism (Koivula *et al.*, 2002). It is likely, therefore, that *Ct*Cel124 also hydrolyses glycosidic bonds through a Grotthus-style mechanism.

A central feature of the catalytic apparatus of *Ct*Cel124 is Glu96, which makes a hydrogen bond with *O4* of the Glc at +1 (Figures 5.2 and 5.3). The glutamate is underneath the β face of the +1 Glc and is thus in an ideal orientation to promote leaving group departure by donating a proton to the scissile glycosidic *O.* This is consistent with the observation that mutation of Glu96 (E96A) completely inactivates the enzyme against polysaccharides and oligosaccharides where the leaving group is poor. Against 2,4-nitrophenyl-cellotriose, in which the 2,4-dinitrophenolate leaving group does not require protonation (pKa ~3.5) (Damude, Withers, Kilburn, Miller, & Warren, 1995), the E96A mutation does not decrease activity (Table 5.1). This is again consistent with the view that Glu96 is the catalytic acid of *Ct*Cel124. The descending limb of the pH curve reports a single ionizing group, the catalytic acid, with a pKa of 6.8. The required modulation of the pKa of the Glu96 carboxylic acid (pKa of carboxylic acid groups in solution are ~4.0) is likely contributed in part through a polar interaction with the OH of Tyr203, although its apolar environment (Glu96 is in close proximity to Leu119, Tyr187, Tyr203 and the aliphatic chain of Gln117) will ensure that the carboxylic aid group of this residue is mostly protonated at pH 5.5. The equivalent residue in Lyz23 and Slt70, Glu73 and Glu478, respectively, are also likely to function as the catalytic acid during glycosidic bond cleavage.

To summarize the active site of *Ct*Cel124 comprises a basic structural platform that is capable of mediating glycosidic bond cleavage of gluco-configured substrates through a Grotthus-style mechanism. Adornment of this catalytic platform with Asp97 in Lys70 (Figure 5.3) enables the enzyme to exhibit a classical single displacement (inverting) acid-base mechanism. However, it is unclear how the C2 N-acetyl group in Slt70 is activated, which is believed to be an essential feature of its lytic transglycosylase activity.

Figure 5.3| Overlay of the structural fold and active site of *Ct*Cel124 and GH23 enzymes.



A) An overlay of *Ct*Cel124$_{CD}$ (green) and the GH23 black swan type G-lysozyme (yellow; PDB 1GBS); **B)** An overlay of the structural fold of *Ct*Cel124$_{CD}$ (green) and the GH23 *E. coli* lytic transglycosylase Slt70 (cyan; PDB 1QTE); **C)** An overlap of the three enzymes at the -1 active site. The silver-coloured sugars are derived from the two cellotriose molecules.

5.1.3.7. The substrate binding cleft of *Ct*Cel124

The superhelical fold provides a platform for the substrate binding cleft that extends across the top of the protein (Figure 5.2). The cleft houses the two molecules of cellotriose that bind to subsites -4 to -2 and +1 to +3, respectively. No ligand was evident in the -1 subsite. The extended loop, connecting α-H4 and α-H5, forms one wall of the binding cleft, while the other

wall consists of several different structural elements that include the C-terminal end of α-H7, the N-terminal region of α-H8 and the N-terminal region of the loop connecting these helices, (Figure 5.2). Unlike the majority of glycanases, and carbohydrate binding proteins in general, the substrate binding cleft of *Ct*Cel124 does not contain a significant hydrophobic platform; the cellulase makes numerous direct polar contacts with the two cellotriose molecules (Figure 5.2).

A striking feature of the substrate binding cleft is the different topologies displayed by its positive and negative subsites, respectively (Figure 5.2). Subsites -4 to -1 form a deep narrow cleft in which the bound trisaccharide is significantly twisted, Figure S5.7. In contrast, subsites +1 to +3 display a more open topology (Figure 5.2) and the conformation of the bound trisaccharide adopts an approximate two-fold screw axis, Figure S5.7. It is not possible to obtain the crystal structure of an enzyme in complex with insoluble polysaccharides such as cellulose. However, the conformation adopted by cellooligosaccharides such as cellotriose, on enzyme, provides insight into the likely topological features of cellulose bound to *Ct*Cel124$_{CD}$. Thus the helical structure of cellotriose bound to the negative subsites is distinct from the two fold screw axis displayed by glucan chains in crystalline cellulose. Indeed the twisted structure of cellotriose is adopted by cellooligosaccharides in solution (Sugiyama *et al.*, 2000). By contrast the linear conformation adopted by cellotriose bound to the distal positive subsites is similar to the structure of the glucan chains in crystalline cellulose (Nishiyama *et al.*, 2008). Thus, it is likely that the substrate binding cleft of *Ct*Cel124 is tailored to recognize specific substructures of cellulose, which are at the interface between crystalline and paracrystalline (or amorphous) regions of cellulose. Indeed, competition experiments between cellulose-specific CBMs indicate that these proteins recognize specific topological features of the polysaccharide. Thus, CBMs belonging to families 4, 17 and 28 recognize distinct amorphous or paracrystalline regions of cellulose (Boraston *et al.*, 2003; McLean *et al.*, 2002), while CBM2a, CBM3a and CBM1, which bind to crystalline cellulose, also display distinct specificities (Blake *et al.*, 2006). Many cellulose-degrading bacteria express a large number of *endo*-β-1,4-glucanases (Lykidis *et al.*, 2007; Weiner *et al.*, 2008), exemplified by *C. thermocellum* that has the potential to synthesize approximately 30 endoglucanases. The biological rationale for the expansion in this enzyme activity in *C. thermocellum* and other organisms is currently unclear. Cellulose, although chemically invariant, displays very different topologies ranging from highly crystalline structures to isolated highly twisted glucan chains (amorphous cellulose). It is possible that at least some of the endoglucanases expressed by a single organism are tailored to recognize specific cellulose substructures found in nature. *Ct*Cel124, by targeting the boundary between crystalline and amorphous regions of cellulose, may generate reaction products that comprise substrates for *exo*-acting cellulases that act on the non-reducing end of crystalline cellulose and the reducing end of isolated cellulose chains, respectively

.

6. STRUCTURE OF A PENTA-MODULAR CELLULOSOMAL ARABINOXYLANASE

6.1. The structure and function of an arabinoxylan-specific xylanase∞

Márcia A. S. Correia*, K. Mazumder¶¶, Joana L. A. Brás*, Susan J. Firbank‡, Yanping Zhu‡,¶¶, Richard J. Lewis‡, William S. York¶¶, Carlos M. G. A. Fontes* and Harry J. Gilbert‡,¶¶

* CIISA-Faculdade de Medicina Veterinária, Pólo Universitário do Alto da Ajuda, Avenida da Universidade Técnica, 1300-477 Lisboa, Portugal

‡ Institute for Cell and Molecular Biosciences, Newcastle University, The Medical School, Newcastle upon Tyne NE2 4HH, United Kingdom.

¶¶ Complex Carbohydrate Research Centre, University of Georgia, Athens, Georgia 30602-4712, USA.

Adapted from: Correia *et al.*, (2011) The Journal of Biological Chemistry Volume 286 (25), 22510–22520

Abstract

The enzymatic degradation of plant cell walls plays a central role in the carbon cycle and is of increasing environmental and industrial significance. The enzymes that catalyse this process include xylanases which degrade xylan, a β-1,4 xylose polymer that is decorated with various sugars. Although xylanases efficiently hydrolyse unsubstituted xylans, these enzymes are unable to access highly decorated forms of the polysaccharide, such as arabinoxylans that contain arabinofuranose decorations. Here we show that a *Clostridium thermocellum* enzyme, designated *Ct*Xyl5A, hydrolyses arabinoxylans but does not attack unsubstituted xylans. Analysis of the reaction products generated by *Ct*Xyl5A showed that all the oligosaccharides contain an O3 arabinose linked to the reducing end xylose. The crystal structure of the catalytic module (*Ct*GH5) of *Ct*Xyl5A, appended to a family 6 non-catalytic carbohydrate binding module (*Ct*CBM6), showed that *Ct*GH5 displays a canonical (α/β)8-barrel fold with the substrate binding cleft running along the surface of the protein. The catalytic apparatus is housed in the centre of the cleft. Adjacent to the -1 subsite is a pocket that could accommodate an L-arabinofuranose linked α-1,3 to the active site xylose, which is likely to function as a key specificity determinant. *Ct*CBM6, which adopts a β sandwich fold, recognizes the termini of xylo- and gluco-configured oligosaccharides, consistent with the pocket topology displayed by the ligand binding site. In contrast to typical modular glycoside hydrolases, there is an extensive hydrophobic interface between *Ct*GH5 and *Ct*CBM6, and thus the two modules cannot function as independent entities.

∞ The student contributed in the following methodologies: expression and purification of Xyl5A components and enzyme assays.

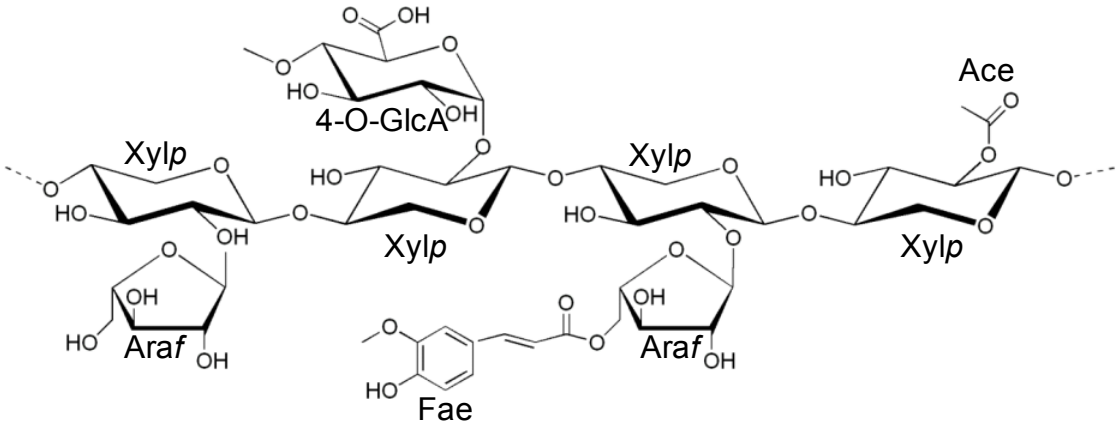
6.1.1. Introduction

The plant cell wall, which is an important biological and industrial resource, consists, primarily, of interlocking polysaccharides (see Brett & Waldron (1996) for review). The biological conversion of the polysaccharides within the plant cell wall to their constituent monosaccharides is central to its biological and industrial exploitation (Himmel & Bayer, 2009; Ragauskas *et al.*, 2006). An example of this chemical complexity is provided by xylan, which is the major hemicellulosic component of the wall. This polysaccharide comprises a backbone of β-1,4 xylose residues in their pyranose configuration (Xyl*p*), which are decorated at O2 with 4-O-methyl-D-glucuronic acid, at O2 and/or O3 with arabinofuranose (Ara*f*) residues, while the polysaccharide can also be extensively acetylated. In addition, the Ara*f* side chain decorations can also be esterified to ferulic acid that, in some species, provides a chemical link between hemicelluose and lignin, (Figure 6.1) (Brett & Waldron, 1996). The precise structure of xylans varies between plant species, tissues and during cellular differentiation (Heinze, 2005).

Reflecting the chemical complexity of plant structural polysaccharides, microbial plant cell wall degrading microorganisms express a large number of enzymes, often in excess of 100 biocatalysts, that target specific linkages within these carbohydrate polymers (Miller *et al.*, 2009; DeBoy *et al.*, 2008; Weiner *et al.*, 2008; Xu *et al.*, 2003). The majority of plant cell wall degrading enzymes are glycoside hydrolases, although polysaccharide lyases and carbohydrate esterases also contribute to the catabolic process. These enzymes are grouped into families based on sequence, structural and catalytic conservation, within the CAZy database (Cantarel *et al.*, 2009). As discussed in the accompanying paper, many of these enzymes are appended to non-catalytic carbohydrate binding modules (CBMs) that are also grouped into families on the CAZy database. The xylan backbone is hydrolyzed by xylanases, the majority of which are located in glycoside hydrolase families (GHs) 10 and 11, although they are also present in GH8 and GH30 (Gilbert, 2010; Gilbert *et al.*, 2008). The extensive decoration of the xylan backbone generally restricts the capacity of these enzymes to attack the polysaccharide prior to removal of the side chains (Pell *et al.*, 2004).

Here we report the biochemical properties and crystal structure of a GH5 enzyme that is appended to a family 6 CBM (*Ct*CBM6). The enzyme (defined as *Ct*Xyl5A) is an arabinoxylan-specific xylanase that utilizes Ara*f* decorations, appended to O3 of the Xyl*p* bound at the active site, as an essential specificity determinant. The capacity of *Ct*Xyl5A to also accommodate arabinose side chains in all the other subsites (in addition to the active site) within the substrate binding cleft enables the enzyme to hydrolyze highly decorated arabinoxylans. The functional significance of the specificity of the arabinoxylanase, in the context of the plant cell wall degrading apparatus of the host bacterium, is discussed.

Figure 6.1| Schematic of xylan.



Ace = acetate
Ara*f* = arabinofuranose
Xyl*p* = xylopyranosse
Fae = ferulate
4-O-GlcA = 4-O-methyl glucuronic acid

6.1.2. Experimental Procedures

6.1.2.1. Cloning, expression and purification of components of *Ct*Xyl5A

DNA encoding *Ct*GH5, *Ct*GH5-CBM6 and *Ct*CBM6 were amplified using primers, containing *NheI* and *XhoI* restriction sites, which are listed in supplemental information, Table S6.1 (annex). The amplified DNAs were cloned into *NheI/XhoI* restricted pET21a such that the encoded recombinant proteins contained a C-terminal His_6 tag. To express the *C. thermocellum* proteins, *Escherichia coli* strain BL21(DE3), harbouring appropriate recombinant plasmids, was cultured to mid-exponential phase in Luria broth at 37 ^{o}C, followed by the addition of isopropyl β-D-galactopyranoside at 1 mM to induce recombinant gene expression, and incubated for a further 5 h at 37 ^{o}C. The recombinant proteins were purified to >90 % electrophoretic purity by immobilized metal ion affinity chromatography (IMAC) using Talon^{TM} (Clontech), cobalt-based matrix, and elution with 100 mM imidazole, as described previously (Charnock *et al.*, 2002). When preparing the selenomethionine derivative of *Ct*GH5-CBM6 for crystallography, the proteins were expressed in *E. coli* B834 (DE3), a methionine auxotroph, cultured in media comprising 1 litre SelenoMet Medium Base^{TM}, 50 ml SelenoMet Nutrient Mix^{TM} (Molecular Dimensions) and 4 ml of a 10 mg/ml solution of L-selenomethionine. Recombinant gene expression and protein purification was as described above except that all purification buffers were supplemented with 10 mM β-mercaptoethanol.

6.1.2.2. Mutagenesis

Site-directed mutagenesis was carried out using the PCR-based QuikChange method (Stratagene) deploying the primers listed in annex, Table S6.1.

6.1.2.3. Enzyme assays

*Ct*Xyl5A and its derivatives were assayed for enzyme activity using the method of Miller (Charnock *et al.*, 1997; Miller, 1959) to detect the release of reducing sugar. The standard assay was carried out in 50 mM sodium phosphate buffer, pH 7.0, and the potential polysaccharide substrate was at 1 mg/ml. The reactions were initiated by the addition of enzyme up to 10 μM and incubated at 60 ^{o}C (unless otherwise stated) for up to 16 h. The identification of potential reaction products were also assessed by HPAEC using methodology described previously (Charnock *et al.*, 1997). The capacity of *Ct*GH5 and *Ct*GH5-CBM6 to hydrolyse xylooligosaccharides was assessed by HPAEC using 100 μM of oligosaccharide and 5 μM of protein.

6.1.2.4. Oligosaccharide analysis

Rye arabinoxylan (5 g) was digested to completion (no further increase in reducing sugar and change in the HAEPC product profile) with 3 μM of *Ct*Xyl5A at 60 ^{o}C for 48 h. The oligosaccharide products were partially purified by size exclusion chromatography using a Bio-Gel P2 column as described previously (Proctor *et al.*, 2005). The structures of the oligosaccharides were analyzed by NMR, electrospray ionization mass spectrometry (ESI-MS) and HPAEC in combination with selective enzyme treatment. Partially methylated alditol acetate derivatives of the glycosyl residues of the oligosaccharides were prepared and analyzed by gas chromatography electron impact mass spectrometry GC-EIMS.

6.1.2.4.1. Preparation of the partially methylated alditol acetates

The mixture of oligosaccharides (~500 μg) was per-O-methylated using the method of Ciucanu and Kerek (Ciucanu & Kerek, 1984). The per-O-methylated oligosaccharides were hydrolyzed with 2N TFA, reduced and acetylated to generate partially methylated alditol acetate (PMAA) derivatives (Carpita and Shea, 1989).

6.1.2.4.2. GC-EIMS analysis

PMAA derivatives were analyzed with a Hewlett Packard 5890 gas chromatograph - mass spectrometer. The PMAAs were separated with a SP 2330 column (30 m X 0.25 mm, 0.25 μm film thickness, Supelco) using the following temperature gradient: 80 ^{o}C for 2 min, 80-170^{o}C at 30 ^{o}C/min, 170-240^{o}C at 4 ^{o}C/min, 240 ^{o}C held for 20 min. Samples were ionized by electrons impact at 70eV.

6.1.2.4.3. Preparation of per-O-methylated oligoglycosyl alditols

The sample (~500 μg) was reduced with sodium borohydride to generate oligoglycosyl aditols, which were per-O-methylated as previously described (Mazumder & York, 2010).

6.1.2.4.4. MALDI-TOF mass spectrometry (MALDI-TOF-MS)

Positive ion MALDI-TOF mass spectra were recorded using an Applied Biosystems Voyager-DE biospectrometry workstation. Samples (1μl of a mg/ml solution) were mixed with an equal volume of matrix solution (0.1 M 2,5-dihydroxybenzoic acid and 0.03 M 1-hydroxyisoquinoline in aqueous 50 % MeCN) and dried on MALDI target plate. Typically, spectra from 200 laser shots were summed to generate a mass spectrum.

6.1.2.4.5. ESI-MS

The multiple stage ESI mass spectra were recorded in a Thermo Scientific LTQ XL ion trap mass spectrometer. Per-O-methylated oligoglycosyl alditols in methanol were diluted with 50 % acetonitrile-water containing 0.1 % TFA. Samples were infused through a fused silica capillary (150 μm i.d. X 363 μm o.d. X ~ 60 cm, Thermo Finnigan, USA) into the source at flow rate of 3 μl/min using the syringe pump provided with the instrument. The electrospray source was operated at a voltage of 5.0 KV and the capillary heater was set to 275 $^{\circ}$C. All the experiments were performed in the positive-ion mode.

6.1.2.4.6. NMR spectroscopy

Oligosaccharides (~2 mg) were dissolved in D_2O (0.5 ml, 99.9 %; Cambridge Isotope Laboratories). ^{1}H NMR spectra were recorded with Varian Inova NMR spectrometer operating at 500 MHz at 298 K. All two dimensional spectra were recorded using standard Varian pulse programs.

6.1.2.5. Isothermal Titration Calorimetry

The binding of *Ct*CBM6 to ligands was quantified by isothermal titration calorimetry (ITC), as described previously (Charnock *et al.*, 2000). Titrations were carried out in 50 mM Na/Hepes buffer, pH 7.5, containing 5 mM $CaCl_2$ at 25 ^{o}C. The reaction cell contained protein at 145 μM, while the syringe contained the monosaccharide or oligosaccharide at 5-15 mM, while polysaccharide, when used as the titrant, was at 3-5 mg/ml. The titrations were analyzed using Microcal Origin version 7.0 software to derive, n, K_a and ΔH values, while ΔS was calculated using the standard thermodynamic equation $-RT\ln K_a = \Delta G = \Delta H - T\Delta S$.

6.1.2.6. Crystallography

Proteins were crystallized using the hanging drop vapour technique at 20 °C with an equal volume (1 μl) of protein and reservoir solution. Native (10 mg/ml) and selenomethionine (3 mg/ml) *Ct*GH5-CBM6 crystallised in 16-24 % PEG 3000, 150 mM Na/citrate, pH 5.5. A *Ct*GH5-CBM6 construct containing 2 additional methionines, W391M/W397M, was produced to facilitate structure solution by selenomethionine SAD. Crystals were cryoprotected by the inclusion of 25 % glycerol in the crystallization solution and flash frozen in liquid nitrogen.

Diffraction data were collected at ID14.4 ESRF, Grenoble, France at the selenium *K* absorption edge to enable structure solution by SAD. The diffraction data were processed in MOSFLM (Leslie, 2006) and SCALA (Evans, 1993) and the heavy atom substructure was also solved using SHELXCDE (Sheldrick, 1990) as part of CCP4i, and an initial model was built in Arp/wArp (Lamzin & Wilson, 1993), which was completed manually in COOT (Emsley & Cowtan, 2004). The complete initial model was used to determine the structure of the wild type protein by molecular replacement and refined at higher resolution from data collected at the Diamond Light Source, UK. The crystal of the reported structure had been soaked in 20 mM "Fraction 1" in an attempt to obtain a structure of the enzyme in complex with carbohydrate although no sugar molecules, other than glycerol, were observed in the electron density.

All structures were refined to convergence using REFMAC5 (Murshudov *et al.*, 1997) with manual corrections being applied in COOT (Emsley & Cowtan, 2004). The data collection, phasing and refinement statistics are displayed in Table S6.2 and the PDB codes for the protein structures are as follows: 2y8k.

6.1.3. Results

6.1.3.1. Expression and purification of CtXyl5A

To investigate the function of the *Ct*GH5 and *Ct*CBM6 components of *Ct*Xyl5A, the modules were expressed as either individual entities or covalently linked. Whilst *Ct*CBM6 and *Ct*GH5-CBM6 were expressed in soluble form at high levels in *E. coli*, *Ct*GH5 was predominantly insoluble and only a small amount of soluble protein was generated in the enteric bacterium. All three proteins were purified by IMAC to electrophoretic homogeneity.

6.1.3.2. *Ct*Xyl5A is an arabinoxylanase

Screening the capacity of *Ct*Xyl5A to hydrolyse plant structural polysaccharides revealed that the enzyme was able to degrade rye and wheat arabinoxylan, displayed limited activity against oat spelt xylan, but was unable to act on glucuronoxylan, birch or beech xylan, Table 6.1. The enzyme displayed no activity against a range of mannans, pectins, galactans, arabinans, and β-glucans (data not shown). The individual kinetic constants of *Ct*Xyl5A against rye and wheat arabinoxylan could not be determined as the K_M was greater than the maximum concentration of soluble substrate, however, the catalytic efficiency of the enzyme was similar for both rye and wheat arabinoxylan. The high K_M may reflect weak affinity for the substrate, or the glycosidic bonds targeted by *Ct*Xyl5A occur rarely in the arabinoxylan substrates. The enzyme displayed trace activity against xylohexaose with a k_{cat}/K_M estimated to be $<10^1$ min^{-1} M^{-1}. These data indicate that *Ct*Xyl5A hydrolyses arabinoxylans but does not act on xylans that contain few arabinofuranose side chains. This is in sharp contrast to typical xylanases, located mainly in GH10 and GH11, which display a preference for the

poorly decorated xylans from birch and beechwood (Pell *et al.*, 2004). These data show that *Ct*Xyl5A displays specificity for arabinoxylans and as such is defined as an arabinoxylanase, an activity not previously reported.

Table 6.1| Catalytic activity of *Ct*Xyl5A and its variants.

Proteins	K_{cat}/K_M (min^{-1} mg^{-1} ml)		
	Rye Arabinoxylan	**Wheat Arabinoxylan**	**Oat spelt xylan**
*Ct*Xyl5A	1322 ± 357	808 ± 76	15 ± 1.2
*Ct*Xyl5A + 2 mM $CaCl_2$	2271 ± 274	1658 ± 56	ND[a]
*Ct*Xyl5A + 5 mM EDTA	858 ± 102	343 ± 32	ND
*Ct*GH5-CBM6	1012 ± 83	652 ± 21	14 ± 2.2357
W424A (*Ct*GH5-CBM6)	1656 ± 173	728 ± 127	ND
F478A (*Ct*GH5-CBM6)	983 ± 52	713 ± 66	ND
E279A (*Ct*GH5-CBM6)	NA[b]	NA± 357	ND
E171A (*Ct*GH5-CBM6)	NA	NA	ND
*Ct*GH5	1.1 ± 0.45	0.6 ± 0.07	NA

The enzymes were assayed at 60 ^{o}C in 50 mM sodium phosphate buffer, pH 7.0, containing substrate at a concentration of 1 mg ml^{1}. The reaction was monitored by the release of reducing sugar (Miller G.L., 1959). The catalytic rate could be used to determine K_{cat}/K_M as the substrate concentration was $<<K_M$ (the rate of reaction was directly proportion to substrate concentration up to 2 mg ml^{-1}).

[a]ND: Not determined.

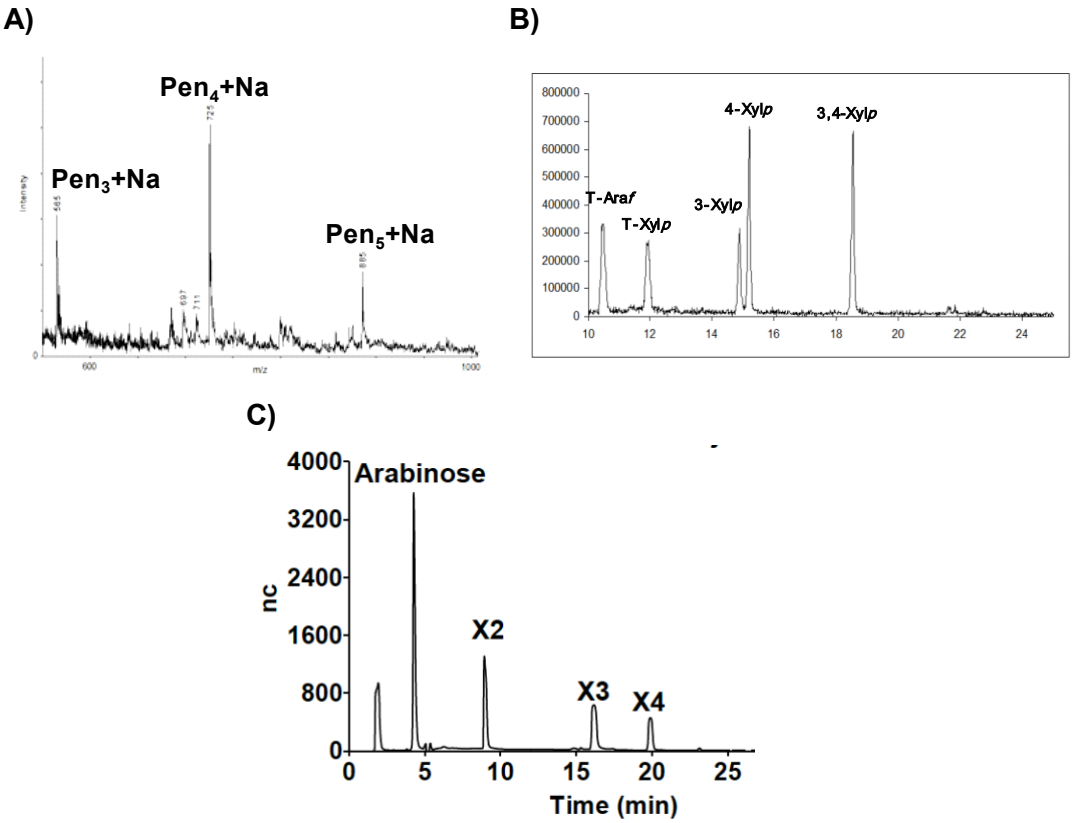
[b]NA: No activity detected.

6.1.3.3. Characterization of the reaction products generated by *Ct*Xyl5A from arabinoxylan

To explore the substrate specificity of *Ct*Xyl5A in more detail, the reaction products generated by treating rye arabinoxylan with the enzyme were partially purified by size exclusion chromatography to remove high molecular weight polymers. The fractions containing the majority of the products were pooled (designated henceforth as Fraction 1). Fraction 1 was per-O-methylated and the products were analyzed by MALDI-TOF-MS. The data revealed that the major reaction products were pentose-containing oligosaccharides with degrees of polymerization (DPs) of 3 (*m/z* 549), 4 (*m/z* 709) and 5 (*m/z* 869), respectively, Figure 6.2.A. Partially methylated alditol acetate derivates were then prepared from per-O-methylated Fraction 1 and analyzed by GC-EIMS, Figure 6.2.B. This semi-quantitative analysis revealed terminal Ara*f* (methylated at O2, O3 and O5), terminal Xyl*p* (methylated at O2, O3, and O4), 3-linked Xyl*p*, 4-linked Xyl*p*, and 3,4-linked Xyl*p*. No Xyl*p* residues decorated at O2 or at both O2 and O3 were observed. These data indicate that the oligosaccharides consist of a backbone of (1→4) linked Xyl*p* residues decorated with Ara*f* side chains at O3 of internal or reducing Xyl*p* residues (3,4-linked Xyl*p*), or at O3 of non-reducing terminal Xyl*p* residues (3-linked Xyl*p*). Fraction 1 was also treated with *Cj*Abf51A,

an arabinofuranosidase that releases Ara*f* residues from O2 or O3 of singly-branched Xyl*p* residues in the xylan backbone (Beylot, McKie, Voragen, Doeswijk-Voragen, & Gilbert, 2001). HPAEC analysis of the *Cj*Abf51A digestion products revealed the presence of arabinose, xylobiose, xylotriose and xylotetraose, Figure 6.2.C, indicating that the predominant *Ct*Xyl5A products are xylooligosaccharides in which at least one of the Xyl*p* residues bear a mono-Ara*f* side chain. By contrast, GH10 and GH11 xylanases generate predominately xylose and xylobiose from wheat arabinoxylan, reflecting a preference for undecorated regions of the polysaccharide (Pell *et al.*, 2004).

Figure 6.2| Analysis of the reaction products generated by *Ct*Xyl5A from arabinoxylan.

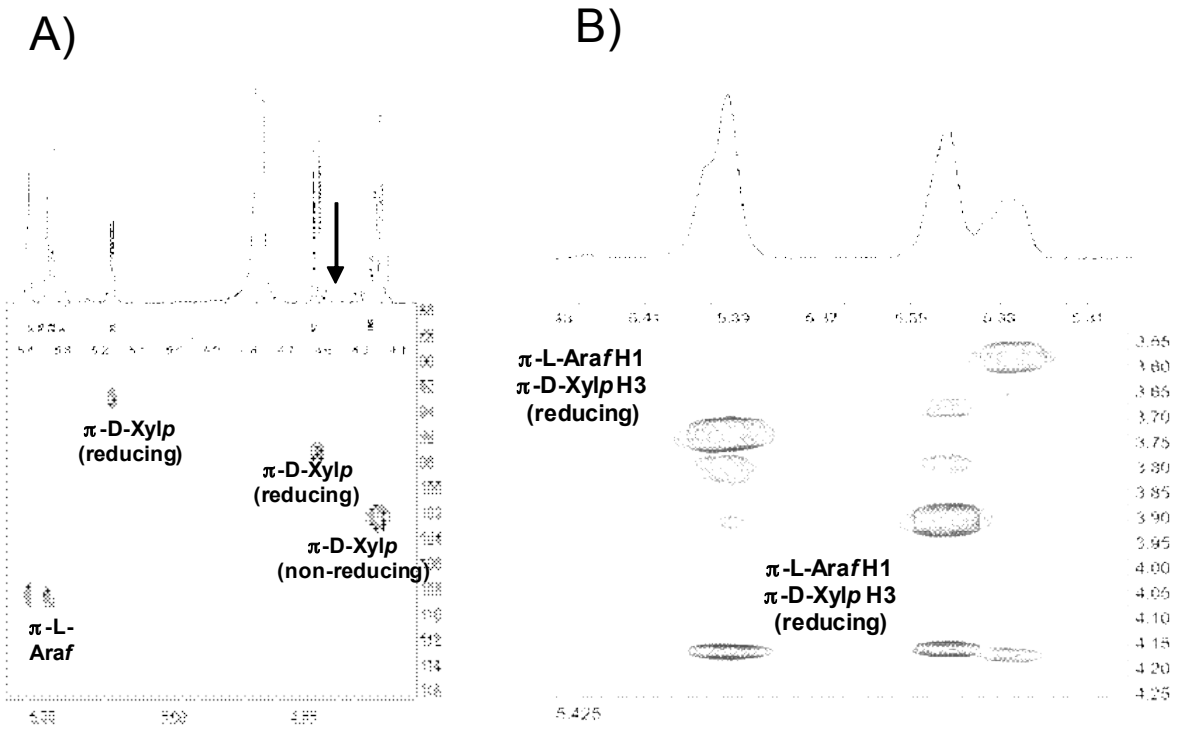


Rye arabinoxylan was incubated with *Ct*Xyl5A until the reaction was complete and the products purified by size exclusion chromatography. Fraction 1 contained the most abundant oligosaccharides. **A)** MALDI-TOF MS analysis of permethylated and $NaBH_4$ reduced oligosaccharides in Fraction 1. Molecules that contained exclusively pentaose sugars are labelled Pen with the DP in subscript. **B)** GC-EIMS analysis of Fraction 1. All the hydroxyls of T-Ara*f* and T-Xyl*p* are methylated, while 4-Xyl*p* 3,4-Xyl*p* and 3-Xyl*p* signify the positions of the hydroxyls that are not methylated and were thus involved in a linkage prior to TFA cleavage. **C)** HPAEC analysis of Fraction 1 treated with the arabinofuranosidase *Cj*Abf51A. Peaks X2, X3 and X4 co-migrate with xylobiose, xylotriose and xylotetraose.

The oligosaccharides in Fraction 1 were analyzed by several 2D NMR methods, including gCOSY, HSQC, TOCSY, and ROESY. These analyses provided scalar and dipolar correlations that allowed the resonances of the most abundant spin systems to be assigned to specific sugar residues (Table S6.3 for a more detailed description of this approach, see,

for example, Gruppen *et al.*, (1992); Hoffmann, Leeflang, de Barse, Kamerling, & Vliegenthart, (1991); Mazumder & York, (2010). Upfield shifts typical of reducing residues (Gruppen *et al.*, 1992; Hoffmann *et al.*, 1991; Mazumder & York, 2010) were observed for two C1 resonances (δ 92.4 and 96.6) in the HSQC spectrum of the *Ct*Xyl5A-generated oligosaccharides, Figure 6.3.A. In combination with other 2D NMR data, this allowed these two resonances to be assigned to α-Xyl*p* and β-Xyl*p* residues at the reducing end of the oligosaccharides. However, the exact ^{1}H and ^{13}C shifts of these reducing residues indicate that they are structurally distinct from the unbranched (4-linked) sugars at the reducing end of oligosaccharides, generated by more typical endoxylanases (Gruppen *et al.*, 1992; Hoffmann *et al.*, 1991; Mazumder & York, 2010). The data reveal the presence of an Ara*f* side chain at O3 (along with a β-Xyl*p* at O4) of the reducing Xyl*p* residues of the *Ct*Xyl5A-generated oligosaccharides. For example, the C3 resonances of the reducing α-Xyl*p* and β-Xyl*p* units exhibit diagnostic downfield glycosylation shifts (δ_C 77.7 and 77.8), relative to the corresponding unbranched reducing residues produced by more typical endoxylanases (δ_C 71.2 and 73.8). Furthermore, the ROESY spectrum of Fraction 1, Figure 6.3.B, revealed strong dipolar interactions between the two most intense α-Ara*f* H1 resonances (δ_H 5.342 and 5.391) and the reducing α-Xyl*p* and β-Xyl*p* H3 resonances (δ_H 3.906 and 3.736, respectively), indicating that most of the α-Ara*f* residues are linked to O3 of reducing Xyl*p* moieties. The identification of branched, reducing Xyl*p* residues in Fraction 1 is consistent with the detection of 3,4-linked Xyl*p* residues in the partially methylated derivatives, Figure 6.2.B. Resonances corresponding to unbranched 4-linked β-Xyl*p* residues at the reducing end of the oligosaccharides (e.g. H1 at δ 4.584, Figure 6.3.A) were not detectable in the NMR spectra. Integration of the Xyl*p* and Ara*f* H1 resonances in the 1D spectrum of the *Ct*Xyl5A-generated oligosaccharides, Figure 6.3.A, allowed the following quantitative conclusions to be drawn: the oligosaccharides have an average backbone DP of 2.76 and an average overall DP of 4.04; >99 % of the oligosaccharides have an α-L-Ara*f* side-chain on O3 of the reducing Xylp residue; approximately 30 % of the oligosaccharides have a second α-L-Ara*f* side chain.

Figure 6.3| NMR analysis of the oligosaccharides generated by *Ct*Xyl5A.



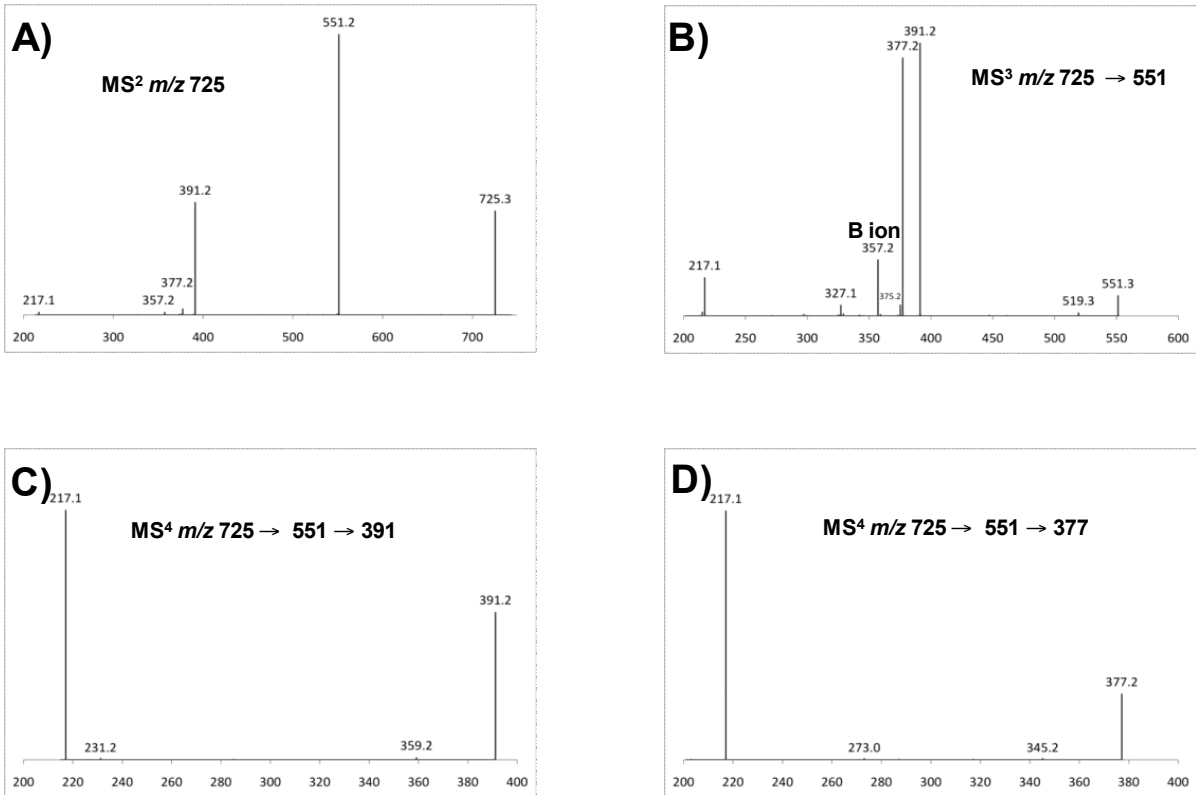
A) Partial HSQC spectrum of Fraction 1 showing upfield shifts of reducing Xyl*p* residues. The arrow indicates the barely detectable H1 resonance of (unbranched) reducing 4-linked residues.

B) Partial ROESY spectrum of Fraction 1 showing interglycosidic dipolar contacts between the Ara*f* H1 and the reducing Xyl*p* H3 resonances.

To analyze Fraction 1 by ESI-MS^n, the oligosaccharides in this sample were treated with $NaBH_4$, and the resulting oligoglycosyl alditols were methylated prior to fragmentation. This procedure imparts a distinctive mass label to the newly formed alditol end of the oligosaccharide, facilitating ESI-MS^n analysis (Mazumder & York, 2010). The data, examples of which are shown in Figure 6.4, provided unambiguous evidence supporting the presence of branched reducing residues in the oligosaccharides in Fraction 1. This conclusion is exemplified by the analysis of the possible tetrasaccharides in Fraction 1. Thus, based on the structure of the polysaccharide substrate, linkage and NMR analysis of Fraction 1, only five different tetrasaccharide structures (Ia, Ib, IIa, IIb, and III) are theoretically possible, Figure 6.5. The ESI-MS^n analysis provided information regarding the topology of the oligomers, but did not define the stereochemistry (identity) of the individual pentose residues. Therefore, the terminal pentose residues at the non-reducing end of the main chain in structures Ia, Ib, IIa, IIb, displayed in Figure 6.5, are indicated by the letter P (as the sugar can be either Ara*f* or Xyl*p* residues). However, in Figure 6.5, non-terminal backbone residues, and sugars attached to branched backbone units (backbone sugars that are linked at O4 and O3 to other sugars), are known to be Xyl*p* and Ara*f*, respectively. Thus, structure I could be (Ara*f*)-Xyl*p*-Xyl*p*-Xylol (Ia) or Xyl*p*-Xyl*p*-Xyl*p*-Xylol (Ib) in which (Ara*f*) is an arabinose decoration appended to the following xylose residue, while Xylol is the alditol form of the xylose at the

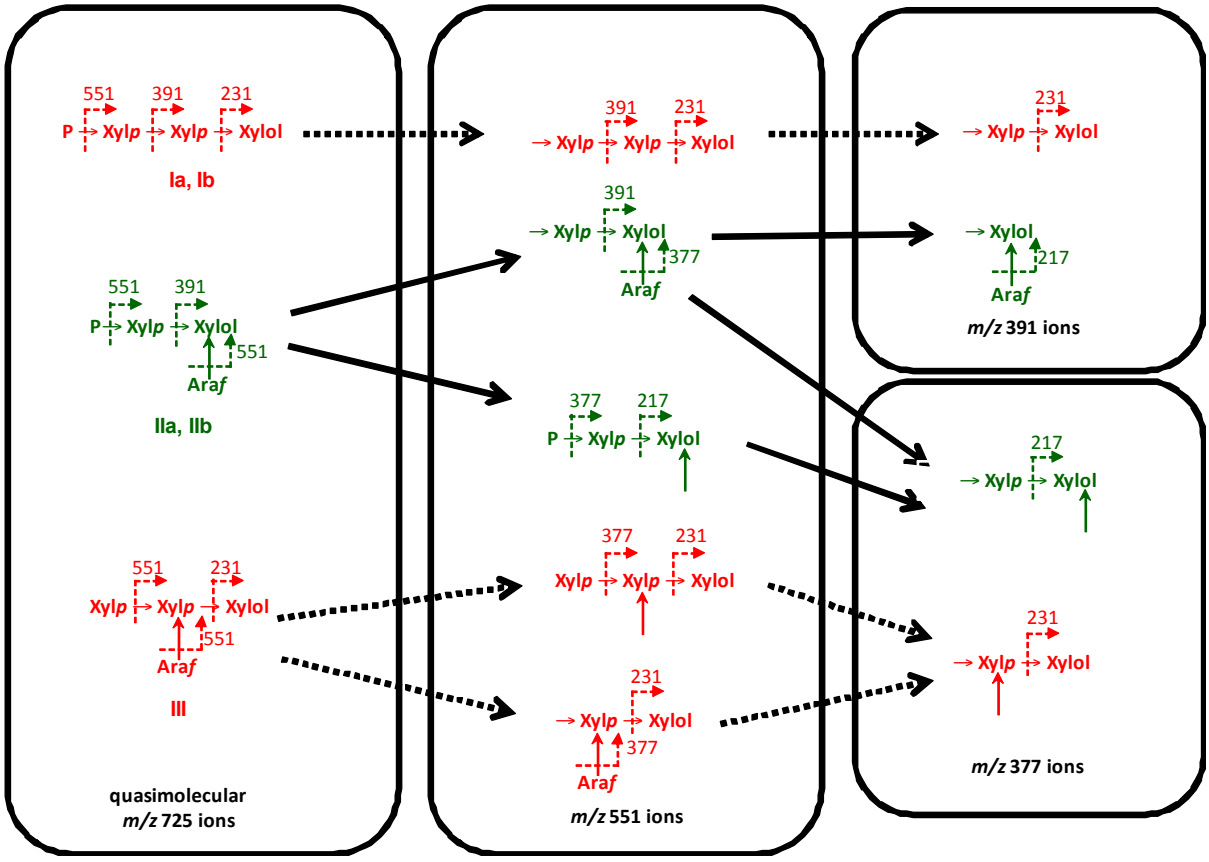
reducing end. Structure II could be Ara*f*-Xyl*p*-(Ara*f*)-Xylol (IIa) or Xyl*p*-Xyl*p*-(Ara*f*)-Xylol (IIb) and III is Xyl*p*-(Ara*f*)-Xyl*p*-Xylol. The quasimolecular ($M+Na^+$) ion at *m/z* 725, corresponding to these DP4 structures was selected for MS^2, Figure 6.4A. The fragmentation pattern is dominated by Y-ions (Domon & Costello, 1988; Mazumder & York, 2010), which contain the alditol end of the oligomer. The Y-ion (*m/z* 551) generated by loss of a single terminal pentosyl residue was selected as the precursor for MS^3 fragmentation, Figure 6.4B. Comparison of this MS^3 spectrum, Figure 6.4A-B, to the theoretical fragmentation pattern for all possible *m/z* 551 ions, Figure 6.5, indicates that structures I and III are not present, as these would fragment to form ions at *m/z* 231, which were not observed. This was confirmed by MS^4 analysis, Figure 6.4C-D, in which MS^3 fragment ions at *m/z* 391 and 377 were selected as precursors. Here, the extremely low abundance of ions at *m/z* 231 confirms the absence of significant amounts of structures I and III, Figure 6.5. However, all ions predicted for structure II were observed, notably the high-abundance ion at *m/z* 217, which consists of the alditol residue with two unmethylated hydroxyl groups that were exposed by cleavage of glycosidic bonds during MS^2 and MS^3, Figure 6.4.

Figure 6.4| ESI-MS of the tetrasaccharides in Fraction 1.



The tetrasaccharides in Fraction 1, which contains the most abundant products, were analyzed by $ESI\text{-}MS^n$. *PanelA* shows the fragmentation of the *m/z* 725 ion which comprises the tetrasaccharides. **B)** shows the fragmentation of the *m/z* 551 ion derived from the *m/z* 725 ion in **A)**, ***C)*** and **D)** depict the fragmentation pattern of the *m/z* 391 and *m/z* 377 ions, respectively, derived from the *m/z* 551 ion generated in **B)**. The masses of Y-ions are indicated unless otherwise stated.

Figure 6.5| The structure of the tetrasaccharides generated by *Ct*Xyl5A.



Based on the data displayed in Figure 6.4, the structures of the tetrasaccharides in Fraction 1 were identified. The sugars labelled P can be Ara*f* or Xyl*p*. The data showed that the oligosaccharide ions coloured green were present, while those coloured red were not evident. The solid arrows between oligosaccharides showed the conversion of one oligosaccharide into another, through ESI-MS fragmentation. Dotted arrows between oligosaccharides indentified theoretical ESI-MS-mediated oligosaccharide conversions that did not occur in these analyses. The dotted arrow between sugar linkages within the oligosaccharides shows the fragmentation site and the ion identified. Arrows pointing at sugars (but did not link two sugars together) identified hydroxyl groups that were not methylated as they comprised a glycosidic linkage in a parental ion. Xylol is the reducing end xylose that has been reduced to its alditol form by $NaBH_4$.

When the DP5 oligoglycosyl alditols in Fraction 1 were analyzed by MS^n, virtually all of the alditol moieties were branched, Figures S6.1 and S6.2 (see annex). ESI-MS^n data for the DP5 oligoglycosyl alditols also provide further insight into the extent to which Ara*f* side chains can decorate the xylooligosacchrides produced by *Ct*Xyl5A. Notably, MS^4 of the *m/z* 537 ion (derived from the alditol pentasaccharide) generates an *m/z* 363 Y-ion that yields an *m/z* 217 ion at MS^5. As shown in the schematic, Figure S6.2, these species can only be generated if the xylosyl alditol and the adjacent Xyl*p* are both branched. The detection of a *m/z* 377 ion at MS^3, however, demonstrates that the structure Xyl*p*–Xyl*p*-(Ara*f*)-Xylol is also present. Fragmentation of DP3 oligoglycosyl alditols yields an *m/z* 217 Y-ion at MS^3, while only trace amounts of a *m/z* 231 ion were evident, Figures S6.3 and S6.4 (annex). This again demonstrates that the xylosyl alditol contains a branch and thus the structure of the trisaccharide is predicted to be Xyl*p*-(Ara*f*)-Xylol.

6.1.3.4. Binding of *Ct*Xyl5A to arabinoxylan

The terminal reaction products produced by endo-acting glycoside hydrolases reflects an iterative process in which the products from initial hydrolytic reactions serve as substrates in subsequent rounds of catalysis. Analysis of the structure of the terminal reaction products (which are unable to be further hydrolysed) provides insight into the possible modes of substrate binding to both the negative and positive subsites (see below). The subsite nomenclature of glycoside hydrolases were defined previously by Davies and colleagues (Davies *et al.*, 1997). Briefly, the scissile bond is positioned between subsites -1 and +1, and subsites that extend towards the non-reducing and reducing ends of the substrate are assigned increasing negative and positive numbers, respectively. The Xyl*p* at the reducing and the non-reducing end of the oligosaccharide products are derived from substrate bound at the -1 and +1 subsites, respectively. As ~ 99 % of the reducing end Xyl*p* residues contain an O3 Ara*f* branch, it is evident that the arabinose decoration of the xylose bound at the -1 subsite is a key specificity determinant of the enzyme. The detection of terminal Xyl*p* (in which O2, O3 and O4 are methylated) and 3-linked Xyl*p* residues, both of which occur at the non-reducing end of the oligosaccharide backbone, indicates that a Xyl*p* with an Ara*f* side chain at O3 can be accommodated in the +1 subsite of *Ct*Xyl5A, but a side chain in this position is not a specificity determinant. As both (Ara*f*)-Xyl*p*-(Ara*f*)-Xylol and Xyl*p*-Xyl*p*-(Ara*f*)-Xylol were identified in the tetrasaccharide, an O3-Ara*f* side chain is present on some, but not all, of the Xyl*p* residues bound in the -2 subsite. Thus, while an O3-Ara*f* side chain can be accommodated at the -2 subsite, the arabinose decoration does not define enzyme specificity. The identification of Xyl*p*-(Ara*f*)-Xyl*p*-(Ara*f*)-Xylol in the pentasaccharide reaction products not only confirms that Ara*f* can be present at the -1 and -2 subsites, but also demonstrates that the +2 and +3 (if it exists) subsites can accommodate Xyl*p* residues bearing arabinose side chains. It should be noted, however, that Xyl*p*-(Ara*f*)-Xyl*p*-(Ara*f*)-Xyl*p* is a potential substrate for the enzyme (binding from subsites -2 to +1), suggesting that this molecule is only hydrolyzed very slowly by the enzyme, possibly because it is unable to access the +2 subsites. This is consistent with the absence of Xyl*p* or (Ara*f*)-Xyl*p* in the reaction products; xylose or decorated xylose can only be generated if the substrate is hydrolyzed when it occupies only +1 of the positive subsites of the enzyme. Thus, to summarize, subsites -2 to +2 of *Ct*Xyl5A can accommodate Xyl*p* residues that contain an O3-Ara*f* side chain, however, only at the -1 subsite does the arabinose decoration act as an essential specificity determinant.

6.1.3.5. *Ct*CBM6 specificity

To investigate whether *Ct*CBM6 is a functional CBM, the capacity of *Ct*GH5-CBM6 to bind to various carbohydrates was assessed by ITC. The data showed that *Ct*GH5-CBM6 bound to cellohexaose and cellobiose with similar affinity, Table 6.2 (example titrations in Figure 6.6).

By contrast, binding to glucose was too low to quantify. The protein also displayed affinity for the reaction products generated by *Ct*Xyl5A and for undecorated xylooligosaccharides. The protein did not appear to bind to various xylans or to β-1,3-β-1,4-glucans. This indicates that *Ct*GH5-CBM6 recognises the terminal region of these polysaccharides, as the concentration of ligand available to the protein in these polymers, which have DPs >300, would be very low and thus binding would not be detected. It is possible that the catalytic module, rather than CBM6, mediates binding to the xylo- and cello-oligosaccharides. To test this hypothesis, the ligand binding profile of variants of *Ct*GH5-CBM6, in which either Trp424 or Phe478 had been substituted with Ala, was assessed. As discussed below these two aromatic residues are highly conserved in the CBM6 family and comprise the primary binding site in this protein family (Czjzek *et al.*, 2001). Both *Ct*GH5-CBM6:W424A and *Ct*GH5-CBM6:F478A, although catalytically active, Table 6.1, displayed no binding to the xylan- and cellulose-derived oligosaccharides, Table 6.2 It is evident, therefore, that the CBM6 component of *Ct*GH5-CBM6 mediates the observed binding to oligosaccharides.

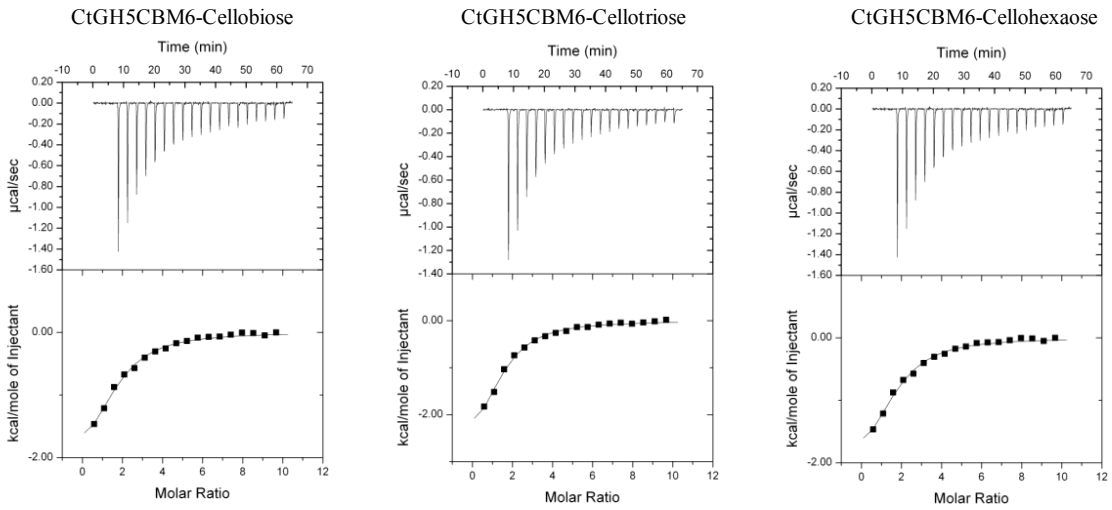
Table 6.2| Binding of *Ct*Xyl5A derivatives to polysaccharides and oligosaccharides.

*Ct*Xyl5A derivative	Ligand	$K_a \times 10^3$ (M^{-1})	ΔG (kcal mol^{-1})	ΔH (kcal mol^{-1})	$T\Delta S$ (kcal mol^{-1})	n
*Ct*GH5-CBM6	Cellobiose	13.7±1.0	-5.6	-3.8±0.11	1.8	1.03±0.13
*Ct*GH5-CBM6	Cellotriose	17.4±3.8	-5.8	-4.7±0.75	1.1	1.02±0.13
*Ct*GH5-CBM6	Cellohexaose	21.0±4.9	-5.9	-3.2±0.44	2.7	1.12±0.12
*Ct*GH5-CBM6	Xylotetraose	3.4±0.2	-4.8	-1.2±0.13	3.6	1.24±0.15
*Ct*GH5-CBM6	Xylotriose	2.5±0.6	-4.6	-1.3±0.03	3.3	1.26±0.13
*Ct*GH5-CBM6	Xylobiose	2.4±0.4	-4.6	-1.2±0.06	3.4	1.1±0.10
*Ct*GH5-CBM6	WAX[a]_treated with *Ct*Xyl5A	9.1±2.1	-5.4	-2.4±0.5	3.0	1.01±0.12
*Ct*GH5-CBM6	WAX treated with *Ct*Xyl5A and *Cj*AbF51A[b]	15.1±2.3	-5.7	-4.9±0.45	0.8	1.08±0.06
*Ct*GH5-CBM6:E279A	Cellohexaose	13.4±2.1	-5.6	-4.2±0.98	1.4	1.12±0.14
*Ct*GH5-CBM6:E279A	WAX treated with *Cj*Xyn10A[c]	14.2±1.2	-5.7	-3.6±0.83	2.1	1.12±0.12
*Ct*GH5CBM6E279A	No binding to WAX					
*Ct*GH5-CBM6:W424A	No binding to cellohexaose, xylotetraose, WAX treated with *Cj*Xyn10A, *Ct*Xyl5A or both *Ct*Xyl5A and *Cj*Abf51A.					
*Ct*GH5-CBM6:F478A	No binding to cellohexaose, WAX treated with *Cj*Xyn10A, *Ct*Xyl5A or both *Ct*Xyl5A and *Cj*Abf51A.					
*Ct*CBM6	No binding to cellohexaose, xylotetraose, xylotriose, xylobiose, WAX treated with *Ct*Xyn5A or *Cj*Xyn10A					

The binding of derivatives of *Ct*Xyl5A to ligands was measured by ITC. The protein was at the 145 μM in the cell and polysaccharide (3-5 mg/ml) or oligosaccharide (5-15 mM) was in the syringe. ITC was carried out in 50 mM Na/HEPES buffer, pH 7.5, at 25 ^{o}C. The concentration of the oligosaccharides generated by the digestion of wheat arabinoxylan (WAX) was fitted to give an n value close to 1.

[a]WAX; wheat arabinoxylans; [b]*Cj*Xyn10A; GH10 xylanase from *Cellvibrio japonicus;* [c]*Cj*Abf51A; GH51 arabinofuranosidase from *C. japonicus.*

Figure 6.6| Representative ITC data of *Ct*GH5-CBM6 to oligosaccharides.



The ligands (10 mM arabinose) in the syringe was titrated into *Ct*GH5-CBM6 (100 µM) in the cell. The top half of each panel shows the raw ITC heats; the bottom half, the integrated peak areas fitted using a one single binding model by MicroCal Origin software. ITC was carried out in 50 mM Na/HEPES, pH 7.5 at 37 °C.

6.1.3.6. Crystal structure of *Ct*GH5-CBM6

The structure of *Ct*GH5-CBM6 was solved by selenomethionine SAD and the resulting structure used as a starting model for refinement against native data extending to 1.5 Å resolution, (PDB code 2y8k) (see Table S6.2 in annex). The polypeptide chain is visible from Ser37 to Ile516.

6.1.3.6.1. *Ct*GH5

As expected, the N-terminal *Ct*GH5 module displays a $(\beta/\alpha)_8$ barrel architecture, although α-helix-8 points away from the barrel and towards *Ct*CBM6 module (discussed below), Figure 7. GH5 enzymes are members of clan GH-A in which the two catalytic residues are invariant glutamates presented at the end of β-strands 4 and 7 (Henrissat *et al.*, 1995; Jenkins, Leggio, Harris, & Pickersgill, 1995). From the structure of *Ct*GH5-CBM6, the catalytic acid-base is likely to be Glu171 (end of β-strand 4) and the catalytic nucleophile Glu279 (end of β-strand 7). The catalytic role of these two residues is confirmed by the observation that the mutants E171A and E279A are inactive, Table 6.1. A narrow V-shaped cleft, approximately 25 Å in length, extends along the full length of the protein and sits over the top of the β-barrel. The dimensions of the cleft, in the centre of which is the catalytic apparatus, suggest that the protein contains ~5 subsites extending from -3 to +2.

An analysis of structural homologues of the *Ct*GH5 component of *Ct*GH5-CBM6 by the DaliLit webserver (http://ekhidna.biocenter.helsinki.fi/dali_server) identified a large number of GH5 and Clan GH-A enzymes that displayed significant structural similarity to *Ct*GH5. The *Pseudoalteromonas haloplanktis* cellulase Cel5G (PDB 1tvn) with a root mean square deviation (rmsd) of 2.8 Å over 253 Cα atoms and a Z-score of 24.1, and the *Bacillus*

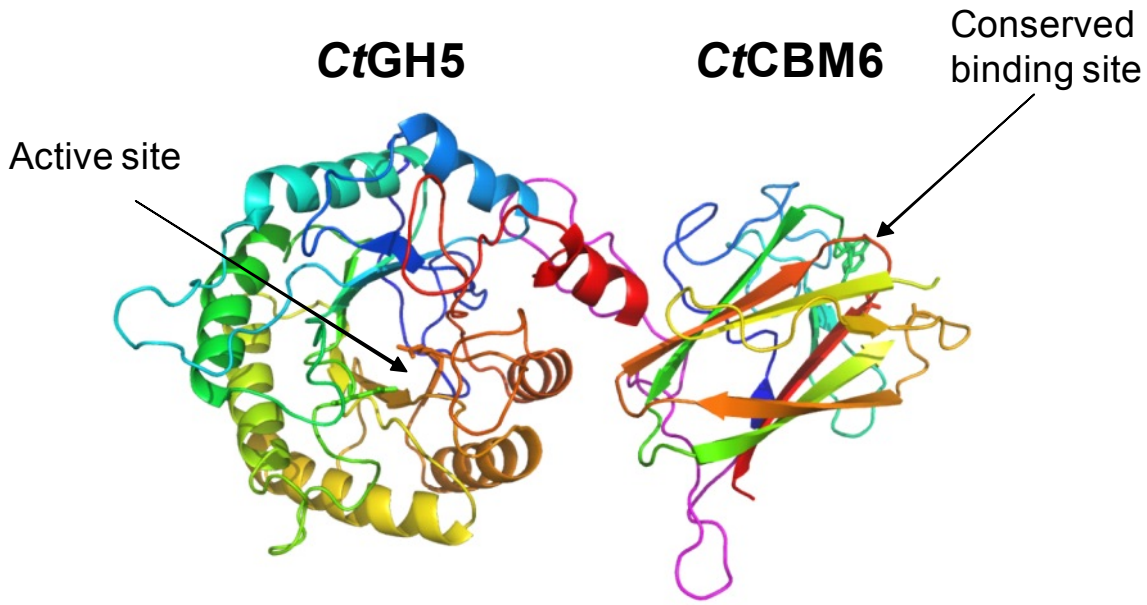
agaradhaerens cellulase *Ba*Cel5A (PDB 1qi2） with an rmsd of 2.9 Å over 254 Cα atoms and a Z-score of 23.6, are representative, close structural homologs. The critical -1 subsite, where the transition state is formed, is similar in the arabinoxylanase and the GH5 cellulases. In addition to the two catalytic glutamates *Ct*GH5 contains several key residues that have been identified as “strictly conserved” in family GH5 enzymes (Hilge *et al.*, 1998). These residues in the *Ct*GH5 module, which superimpose with amino acids in the active site of *Ba*Cel5A (the cellulase residues are shown in parentheses) are as follows: Asn170 (Asn138), Glu171 (Glu139), Tyr255 (Tyr202), Glu279 (Glu228), and Phe310 (Trp262), Figure 6.8A. The catalytic acid-base, Glu171, makes hydrogen bonds with Asn139 and His253, and these interactions likely contribute to both the position and ionization state of this critical amino acid. Asn170 is highly conserved in clan GH-A glycoside hydrolases and plays an important role in transition state stabilization by making a hydrogen bond with the O2 of the sugar at the -1 subsite (Williams, *et al.* 2000). The position of the catalytic nucleophile, Glu279, is stabilized through a hydrogen bond with Tyr255, whereas Phe310, based upon comparison with other related hydrolases, is likely to form the sugar-binding hydrophobic platform in subsite -1.

Despite numerous attempts, no structure of *Ct*GH5-CBM6 in complex with its substrate or reaction products has been obtained, in part due to the preference of this protein to crystallize with the N-terminal residues of a symmetry related molecule positioned in the substrate binding cleft, and because co-crystallization experiments did not yield diffracting crystals. Consequently, it is difficult to define precisely the structural basis for the unusual substrate specificity displayed by the arabinoxylanase. Superimposing *Ba*Cel5A in complex with 2-deoxy-2-fluoro cellotriose with *Ct*GH5 provides some insight into the specificity displayed by the arabinoxylanase. As discussed above, the catalytic apparatus, the residues that interact with O2 and the endocyclic oxygen of the -1 sugar, and the hydrophobic platform are conserved in *Ct*GH5, Figure 6.8A. It is evident, however, that the arabinoxylanase lacks the residues that, in other GH5 enzymes, hydrogen bond with O3 of the active site sugar. For instance, His101 and Tyr66 in *Ba*Cel5A hydrogen bond with O3 of the -1 Glc, whereas the equivalent residues in *Ct*GH5 are Gly134 and Cys95, respectively, Figure 6.8A. Indeed, in the -1 subsite of the arabinoxylanase there is a large pocket around the O3 of the superimposed Glc that could accommodate a sugar decoration such as Ara*f,* Figure 6.8B. The pocket contains a tyrosine (Tyr92) that may make hydrophobic interactions with the arabinose, and several polar residues, Glu68, Asn135, Asn139 and Asn170 that could make polar contacts with the sugar. Based on the presence of glycerol and water molecules within this region of the enzyme, an Ara*f* molecule was modeled into the pocket, Figure 6.8C.

6.1.3.6.2. *Ct*CBM6

The structure of the *Ct*CBM6 module displays a β-sandwich fold typical of other family CBM6 members (Abbott *et al.*, 2009; van Bueren, Morland, Gilbert, & Boraston, 2005; Czjzek *et al.*, 2001), Figure 6.7. The twisted pair of β-sheets, which can be viewed as forming an extended barrel, consist of five and four anti-parallel β-strands, respectively. The structure of *Ct*CBM6 shows strong similarity with numerous CBM6 members. The closest homolog is the CBM6 module (designated *Cm*CBM6) from the *Cellvibrio mixtus* lichenase *Cm*Lic5A (PDB 1uz0; rmsd 1.5 Å over 123 Cα atoms and a Z-score of 18.1). The major binding site in the CBM6 family is in the loops connecting the two β-sheets. This region, referred to as site A (Abbott *et al.*, 2009; Czjzek *et al.*, 2001), may comprise a pocket if terminal sugars are recognized (Henshaw *et al.*, 2004), or a cleft for the binding of internal regions of polysaccharides (Czjzek *et al.*, 2001). A central feature of site A is a pair of aromatic residues, which bind to the α and β face, respectively, of the terminal sugar (or central sugar in the case of xylan binding modules) and an asparagine, located at the base of the site that makes critical hydrogen bonds with O2, O3 or O4. Specificity is conferred by additional polar and hydrophobic interactions (Abbott *et al.*, 2009). Site A in *Ct*CBM6 displays a pocket-like topology and contains all the key ligand binding residues present in *Cm*CBM6 (Pires *et al.*, 2004), Figure 6.9. The pair of aromatic residues in *Cm*CBM6, Trp92 and Tyr33, which straddle the non-reducing, terminal sugar correspond to Phe478 and Trp424, respectively, in *Ct*CBM6. Furthermore, Glu20 and Asn121 in *Cm*CBM6, which make polar contacts with O3 and O4 of the non-reducing terminal sugar in cello- and xylooligosaccharides, superimpose with Glu411 and Asn507, respectively, in *Ct*CBM6. Finally, the amide nitrogen of Tyr33 in *Cm*CBM6 makes a polar contact with O2 and O3 of the terminal sugar, a contact that is likely to be replicated by that of Trp424 in *Ct*CBM6. The structural conservation between site A in *Ct*CBM6 and *Cm*CBM6 is consistent with the similar ligand specificities displayed by this binding site in the two proteins, Table 6.2. and (Henshaw *et al.*, 2004). Thus, both proteins bind to *xylo*- and *gluco*-configured oligosaccharides but do not display affinity for the corresponding polysaccharides. Thus, the structural similarity between *Cm*CBM6 and *Ct*CBM6 is consistent with the view that the *Clostridium* module targets the terminal regions of oligosaccharides. In *Cm*CBM6 cellooligosaccharides can bind to site A in both orientations, consistent with the targeting of O1/O4, O2 and O3, but not the endocyclic oxygen or O6, which would adopt different positions in the two orientations. Given that the key interactions with the ligand at site A is with the terminal sugar, it is perhaps surprising that *Ct*CBM6 does not display measurable binding to xylose or glucose. It is possible that the entropic cost of locking the sugar into a pyranose ring conformation may contribute to the weak binding, although it is also possible that the protein makes indirect, water-mediated interactions to the penultimate sugar in the oligosaccharides, as observed in *Cm*CBM6-ligand complexes (Pires *et al.*, 2004).

Figure 6.7| Crystal structure of *Ct*GH5-CBM6.

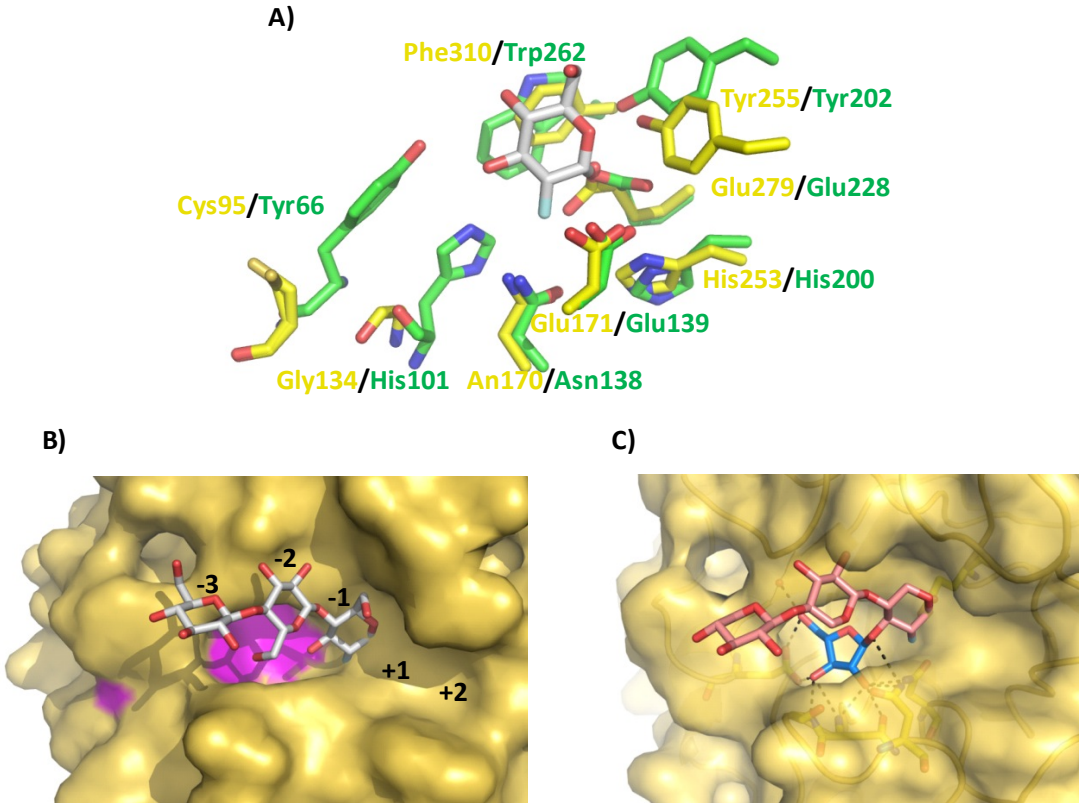


In the protein cartoon of *Ct*GH5-CBM6 both modules are colour ramped from N-terminus (blue) to C-terminus (red). The loop connecting the two modules is coloured magenta. The two catalytic residues (Glu171 and Glu279) in *Ct*GH5 and the two aromatic amino acids that are conserved in the ligand binding site of family 6 CBMs (Trp424, Phe478) are shown in stick format. The figure, and the other structural figures, was drawn with PyMol (DeLano Scientific; http://pymol.sourceforge.net/).

6.1.3.6.3. The linker connecting *Ct*GH5 with *Ct*CBM6

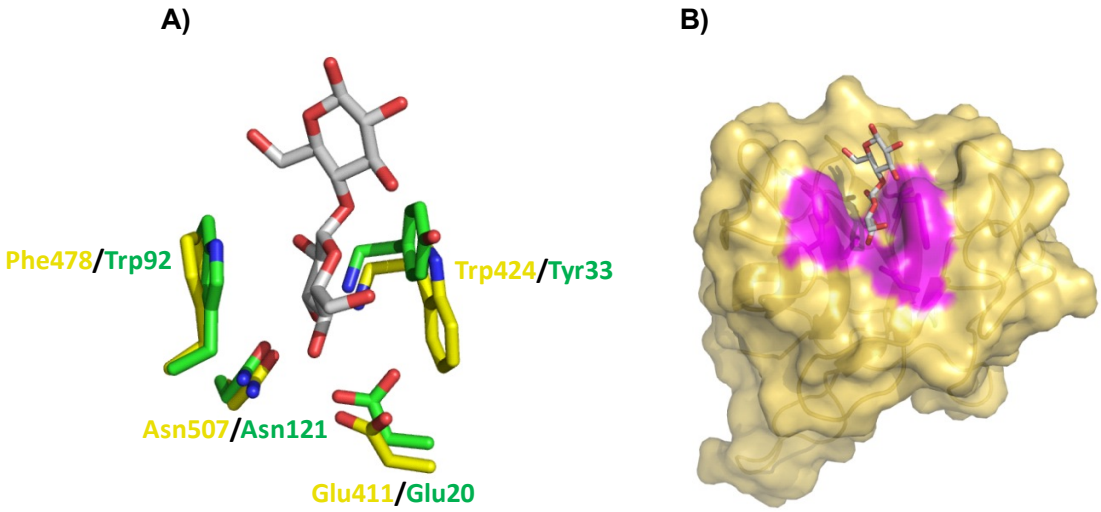
*Ct*CBM6 is connected to *Ct*GH5 by a sequence extending from residues Gly336 to Thr373. This linker, which adopts a stable conformation based on its B-factor, makes numerous internal polar contacts and forms hydrogen bonds with β-strand 3 and the loop connecting β-strands 3 and 4 of *Ct*CBM6, and α-helices 7 and 8 of *Ct*GH5. Furthermore, the C-terminal region of α-helix 8 and the internal region of α-helix 7 make hydrogen bonds with β-strands 3 and 7 of *Ct*CBM6. The polar contacts between the two modules are augmented by a large number of apolar interactions mediated by the linker sequence. The resultant burial of a significant hydrophobic surface, at the interface between *Ct*GH5 and *Ct*CBM6, likely explains why these two modules (or domains) do not fold independently, as occurs in other glycoside hydrolases that contain catalytic modules and CBMs (Boraston *et al.*, 2004). This view is consistent with the observation that *Ct*CBM6, when expressed as a discrete entity (Thr373 to Ile516), does not bind to cellohexoase or xylohexaose, and *Ct*GH5 (Asn32 to Thr373) exhibits very low catalytic activity and is considerably more thermolabile than *Ct*GH5-CBM6, Figure S6.5 (annex).

Figure 6.8| Superimposition of *Ct*GH5 and the cellulase *Ba*Cel5A.



A) Superimposition of the residues in the active site (-1 subsite) of *Ba*Cel5A (PDB 1qi2; coloured green), which interact with the substrate, with the equivalent amino acids (coloured yellow) in *Ct*GH5. **B)** Solvent accessible surface of *Ct*GH5 in which 2-deoxy-2-fluoro-cellotriose, derived from *Ba*Cel5A, has been superimposed. **C)** Model of xylotriose which contains Ara*f* appended to O3 of Xyl*p*-1, bound to *Ct*GH5. The tetrasaccharide ligand is modelled on the superimposed structure of 2-deoxy-2-fluoro-cellotriose and the glycerol and water molecules in the putative arabinose binding pocket. In **A)** and **B)** bound ligand is coloured silver (carbons), while the Xyl*p* and Ara*f* residues in **C)** are coloured salmon pink and blue (carbons), respectively.

Figure 6.9| Superimposition of *Ct*CBM6 and *Cm*CBM6.



A) Superimposition of the residues in the ligand binding site of *Cm*CBM6 (PDB 1uz0; coloured green)) with the equivalent amino acids (coloured yellow) in *Ct*CBM6. **B)** Solvent accessible surface of *Ct*CBM6 in complex with cellotriose (superimposed from *Cm*CBM6). Amino acids whose side chains are predicted to contribute to ligand recognition are coloured magenta. In both panels ligand is shown in silver (carbon) stick representation.

6.1.4. Discussion

This study reveals a *C. thermocellum* protein that displays arabinoxylanase activity, an activity not previously reported. The vast majority of xylanases are derived from GH10 and GH11 and target the β-1,4-D-xylose polymeric backbone. These enzymes do not generally distinguish between different xylans, although highly decorated forms of the polysaccharide, such as rye arabinoxylan, are poorly degraded as steric constraints restrict enzyme access (Pell *et al.*, 2004). Indeed, the only other examples of xylanases that utilize side chains as essential specificity determinants are glucuronoxylan specific enzymes from GH30. These enzymes make critical interactions with the 4-O-methyl glucuronic acid (linked α-1,2 to the xylan backbone) that decorates the xylose at the -2 subsite (Vrsanská, Kolenová, Puchart, & Biely, 2007). *Ct*Xyl5A is highly unusual in that its essential Ara*f* decoration is attached to the xylose positioned in the active site. The only other example of an active site side chain specificity determinant is the α-1,6-Xyl*p* that decorates the -1 Glc in the xyloglucan cellobiohydrolase, OXG-RCBH, from *Geotrichum* sp. (Yaoi *et al.*, 2007).

The function of *Ct*Xyl5A within the context of *C. thermocellum*, which has the genetic capacity to recruit 72 different enzymes into the cellulosome, including seven GH10 and GH11 xylanases, is intriguing. It is likely that the GH10 and GH11 enzymes target xylans that are sparsely decorated with arabinose side chains. By contrast, *Ct*Xyl5A most likely hydrolyses xylans where tandem Xyl*p*s contain Ara*f* decorations. The recognition of the termini of *xylo*- and *gluco*- configured polymers by *Ct*CBM6, suggests that the arabinoxylanase is targeted to regions of the plant cell wall that is undergoing degradation and is therefore accessible to enzyme attack. Although the primary function of CBMs is to bring their cognate enzymes into close contact with appropriate substrates (Fontes & Gilbert, 2010), there is increasing evidence that a subset of these modules, from CBM families 6, 9 and 35, target the termini of polysaccharides and thus may play a similar function to *Ct*CBM6 (Abbott *et al.*, 2009; Bolam *et al.*, 1998; Boraston *et al.*, 2001). In conclusion, *Ct*Xyl5A displays a specificity that is complementary to endoxylanases from GH10, GH11 and GH30. As such the enzyme will make a contribution to the toolbox of biocatalysts required to degrade plant cell walls to their constituent sugars, which can then be used in the biofuel and bioprocessing industries.

6.2. Purification, crystallization and preliminary X-ray characterization of the penta-modular arabinoxylanase *Ct*Xyl5A from *Clostridium thermocellum*[∞]

Joana L. A. Brás*, Márcia A. S. Correia[†], Maria J. Romão[†] José A. M. Prates*, Carlos M. G. A. Fontes* and Shabir Najmudin*

* CIISA-Faculdade de Medicina Veterinária, Pólo Universitário do Alto da Ajuda, Avenida da Universidade Técnica, 1300-477 Lisboa, Portugal

[†] REQUIMTE/CQFB, Departamento de Química, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, 2829-516 Caparica, Portugal

Adapted from Brás *et al.,* (2011) *Acta Crystallogr Sect F Struct Biol Cryst Commun*, 67(Pt 7):833-6.

Abstract

The cellulosome, a highly elaborate extracellular multi-enzyme complex of cellulases and hemicellulases, is responsible for the degradation of plant cell walls. Xylanase *Ct*Xyl5A (Cthe_2193) is a multi-modular arabinoxylanase which is one of the largest components of *Clostridium thermocellum* cellulosome. *Ct*Xyl5A N-terminal catalytic domain, a glycoside hydrolase family 5 (GH5) member, is responsible for the hydrolysis of arabinoxylans. Appended after it are three non-catalytic Carbohydrate Binding Modules (CBMs), which belong to families 6 (CBM6), 13 (CBM13) and 62 (CBM62). In addition, *Ct*Xyl5A has a fibronectin type III-like (Fn3) module preceding the CBM62 and, following it, a type I dockerin (DOC) module which allows the enzyme to be integrated into the cellulosome through the binding to a cohesin module of the protein scaffold, CipA. We have obtained crystals of the penta-modular enzyme, barring the DOC module at the C-terminal, with the domain architecture: *Ct*GH5-CBM6-CBM13-Fn3-CBM62. The structure of this penta-modular xylanase has been determined by Molecular Replacement to a resolution of 2.64 Å using the *Ct*GH5-CBM6, Fn3 and CBM62 pdb coordinates as search models.

∞ The student contributed in the following methodologies: cloning, protein expression, purification and crystallization.

6.2.1. Introduction

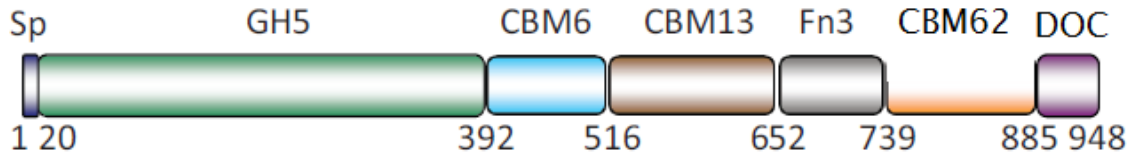
The plant cell wall is one of largest repository of intractable and fixed carbon biosource on earth. It comprises myriads of interlocking polysaccharides displaying a high physical and chemical complexity. Thus, a very large repertoire of enzymes is required to obtain its total degradation. Certain microorganisms have evolved a highly elaborate, megadalton, extracellular multi-enzyme complex of cellulases and hemicellulases, termed the cellulosome, to carry out efficiently this biological conversion of complex polysaccharides to simple monosaccharides (for reviews see Bayer *et al.* (2004); Fontes & Gilbert (2010)). The cellulosomal enzymes are multi-modular with a variable architecture and size. However, each has a dockerin (DOC) module, which allows their integration into the cellulosome through the binding of a cohesin module (COH) of the protein scaffold. The *Clostridium thermocellum* protein scaffold, CipA, has nine COH modules (Carvalho *et al.*, 2004) and the genome sequence of the bacterium revealed the presence of 72 DOC-containing proteins. One such enzyme is xylanase *Ct*Xyl5A (Cthe_2193), a multimodular arabinoxylanase that is one of the largest components of *C. thermocellum* cellulosome, comprising 948 amino acid residues (Mr 103 kDa). *Ct*Xyl5A N-terminal catalytic domain, a glycoside hydrolase family 5 (GH5) member, is responsible for the hydrolysis of arabinoxylans (chemically and structurally complex polysaccharides comprising a backbone of β-1,4-xylose residues decorated with arabinofuranose (Araf) moieties). Appended after the enzyme catalytic domain are three non-catalytic Carbohydrate Binding Modules (CBMs), which belong to families 6 (CBM6), 13 (CBM13) and 62 (CBM62). The structure of the N-terminal bi-modular *Ct*GH5-CBM6 component showed that *Ct*GH5 displays a canonical (α/β)8-barrel fold with the enzyme catalytic domain establishing a tight hydrophobic interaction with the CBM6 (Correia *et al.*, 2011). CBM62 binds to D-galactose and L-arabinopyranose and mediates calcium-dependent enzyme oligomerisation (Montanier *et al.*, 2011). In addition, *Ct*Xyl5A has a fibronectin type III-like (Fn3) module preceding the CBM62 (Alahuhta *et al.*, 2010) and following it, a type I dockerin (DOC) module. We have obtained crystals of the penta-modular derivative of *Ct*Xyl5A, excluding the C-terminal DOC module. The molecular architecture of the crystallized enzyme is: *Ct*GH5-CBM6-CBM13-Fn3-CBM62. In order to gain insights into the structural properties that govern inter-domain interactions in multi-modular enzymes, we aim to determine the crystal structure of *C. thermocellum* arabinoxylanase, *Ct*Xyl5A. In the present communication we describe the overproduction, purification, crystallization and preliminary X-ray analysis of *Ct*Xyl5A excluding the C-terminal dockerin.

6.2.2. Materials and Methods

6.2.2.1. Protein Expression and Purification

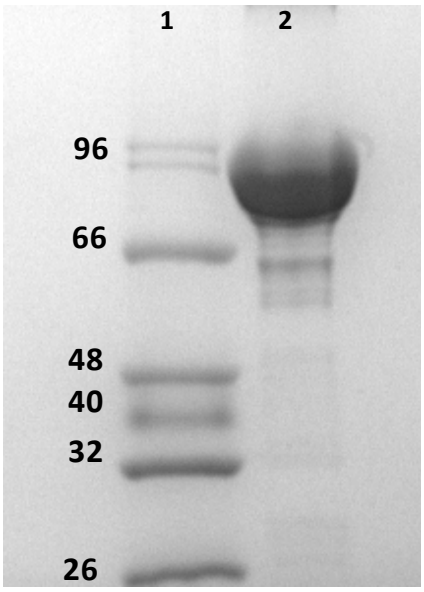
*Ct*Xyl5A is a modular enzyme containing an N-terminal GH5 catalytic domain followed by a CBM6 and CBM13, a fibronectin type III-like (Fn3), a CBM62 module and a type I dockerin module (Figure 6.10). The gene encoding the N-terminal of *Ct*Xyl5A (residues 21 to 885, lacking the DOC module, Mr 91 kDa) was amplified by PCR from *C. thermocellum* genomic DNA, using the NZYlong DNA polymerase (NZYTech Ltd, Portugal) and primers 5'CTC GCT AGC AGC CCG CAA CGT GGC CGG and 3'CAC CTC GAG ATG CAC ATC ATC ATT CTC C that contain *NheI* and *XhoI* restriction sites, respectively. The DNA product was cloned into the *NheI/XhoI* sites of the *Escherichia coli* expressing vector pet21a (Novagen) to generate *Ct*Xyl5A. Recombinant *Ct*Xyl5A contains a C-terminal His6-tag. *Escherichia coli* Tuner DE3 cells harboring pXyl5A were cultured in Luria-Bertani broth at 310 K to mid-exponential phase (A600nm 0.6) and recombinant protein expression was induced by the addition of 0.2 mM isopropyl-1-thio-β-D-galactopyranoside and incubation for a further sixteen hours at 292 K. The His6-tagged recombinant protein was purified from cell-free extracts by immobilized metal ion affinity chromatography (IMAC) as described previously (Najmudin *et al.*, 2005). Purified *Ct*Xyl5A was buffer exchanged into 50 mM NaHepes buffer, pH 7.5, containing 200 mM NaCl and 5 mM $CaCl_2$ and then subjected to gel filtration using a HiLoad 16/60 Superdex 75 column (GE Healthcare) at a flow rate of 1 ml/min. Preparation of *E. coli* to generate seleno-methionine *Ct*Xyl5A was performed as described in (Carvalho *et al.*, 2004) and the protein was purified using the same procedures employed for the native *Ct*Xyl5A. Purified *Ct*Xyl5A (Figure 6.11) was concentrated using an Amicon-10 kDa molecular weight centrifugal concentrator and washed three times with 5 mM DTT, 2mM $CaCl_2$ (for the SeMet protein) or 2mM $CaCl_2$ (for the native protein).

Figure 6.10| Domain architecture of *Ct*Xyl5A.



The abbreviated modules not included in our construct are as follows: Sp, signal peptide; DOC, dockerin type I module (adapted from Montanier *et al.* (2011).

Figure 6.11| Coomasie Brilliant Blue-stained 14%PAGE gel evaluation of protein purity.



The protein sample used here had been stored at 277 K for more than a year. **Lane 1:** Low Molecular Weight (LMW) Protein Marker (NZYtech Lda.), **Lane 2:** *Ct*Xyl5A**.**

6.2.2.2. Crystallization

The crystallization conditions were screened by the hanging-drop vapour-phase diffusion method using the commercial kits Crystal Screen, Crystal Screen 2 and PEG Ion Screen I and II from Hampton Research (California, USA). Drops of 1 μl of 20 mg/ml *Ct*Xyl5A and 1 μl of reservoir solution were prepared at 292K for both the native and seleno-L-methionine-containing protein. No hits were seen in any condition after a year of countless trials. However, after a breakdown in the air conditioning of 292K crystallization room (which caused a fluctuation of temperatures of anything between 292 K and 310 K over a few days), a year after the drops were set up, six crystals of seleno-L-methionine-containing protein appeared in a single drop in the condition: Clear Strategy Screen II - MD-15 n.13 (40% v/v 2-methyl-2,4-pentanediol (MPD)). These crystals (maximum dimension ~200 x 50 x 50 μm) were immediately cryo-cooled in liquid nitrogen and taken to the ESRF for data collection. Numerous attempts were made to optimize this condition in order to get better crystals. However, the best crystals of *Ct*Xyl5A obtained subsequently were in the presence 40 % (v/v) MPD and 10-20 % isopropanol. The largest crystals in the needle clusters are approximately 200 x 10 x 10 μm in size (Figure 6.12). None of these diffracted as well as the crystals from the one-off anomaly.

Figure 6.12| Crystals of *Ct*Xyl5A obtained by hanging-drop vapour diffusion in the presence 40 % (v/v) 2-methyl-2,4-pentandiol and 10-20 % isopropanol.

The largest crystals in the needle clusters are approximately 200 x 10 x 10 μm in size.

6.2.2.3. Data collection and processing

Initially, a SeMet-labeled dataset was collected on beamline ID14-1 at the ESRF (Grenoble, France) using a Quantum 315r charge-coupled device detector (ADSC) with the crystal cooled at 100 K using a Cryostream (Oxford Cryosystems Ltd.). The crystal diffracted to a resolution of 2.64 Å (Figure 6.13). A second MAD data (at the selenium-edge) was also collected on beamline ID14-4 at the ESRF (Grenoble, France) from another crystal, diffracting to 2.98 Å. The data were collected at wavelengths of 0.9795 Å (inflection point, f′ = -10.09, f′′ = 3.44), 0.9790 Å (peak, f′ = -5.78, f′′ = 6.56) and 0.9770 Å (remote). These crystals were delicate and suffered radiation damage after peak data collection. All datasets were processed using the programs iMOSFLM (Leslie, 1992) and SCALA (Evans, 2006) from the CCP4 suite (Collaborative Computational Project, Number 4, 1994). The crystal belongs to the orthorhombic space group ($P2_12_12$) from POINTLESS (Evans, 2006). Data-collection statistics are given in Table 6.3 The Matthews coefficient (V_m = 3.66 $Å^3Da^{-1}$) indicated the presence of one molecule in the asymmetric unit and a solvent content of 68 % (Matthews, 1968). Attempts to solve the structure using all the available phasing programs proved unfruitful. It was difficult to locate any of the five expected Se sites, with the best figure of merits less than 0.2. The structure was eventually solved by Molecular Replacement (MR) using independently solved structures of some of the modules of the *Ct*Xyl5A: *Ct*GH5-CBM6 (pdb code: 2y8k, (Correia *et al.*, 2011)), Fn3 (pdb code: 3mpc, (Alahuhta *et al.*, 2010)) and *Ct*CBM62 (pdb codes: 2y8m, 2y9i and 2y9s, (Montanier *et al.*, 2011)) when they became available. PHASER was used for carrying out MR in a stepwise manner (McCoy *et al.*, 2007). The structure of the *Ct*GH5-CBM6 (2y8k) was used initially to find a solution which gave RFZ and TFZ scores of 27.7 and 55.8, respectively, with an LLG of 2081. This solution was fixed and a second round of PHASER was performed using the Fn3 (3mpc) structure. The RFZ and TFZ scores were 3.3 and 22.6, respectively, with an LLG of 2391. In the third round of

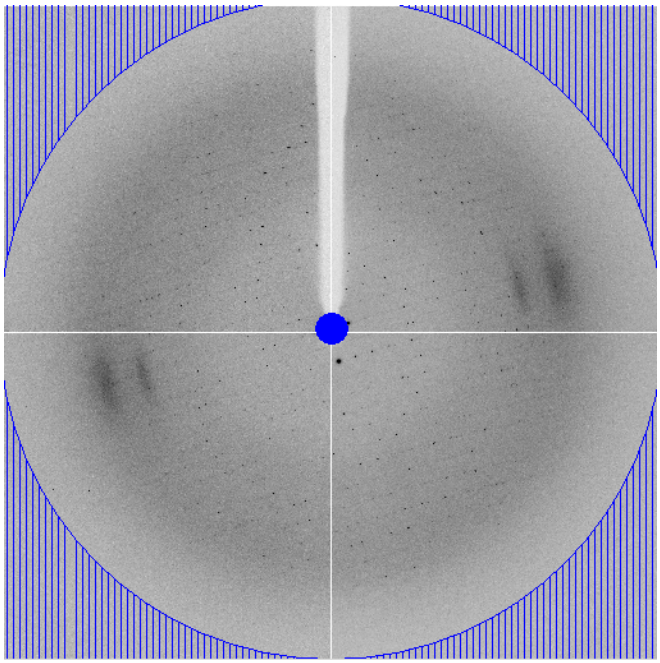
PHASER the solution with both *Ct*GH5-CBM6 and Fn3 was fixed and *Ct*CBM62 was used as a search model. However, the LLG decreased to 1984 and a sensible solution could not be obtained for the *Ct*CBM62 domain. A blast search (Altschul, Gish, Miller, Myers, & Lipman, 1990) showed that the best structural homologues for the *Ct*CBM13 module were the four Ricin B-type lectin modules from the mosquitocidal toxin from *Bacillus sphaericus* (pdb code: 2vsa, (Treiber, Reinert, Carpusca, Aktories, & Schulz, 2008)) with sequence identities in the range 22-25 %. These four domains were superposed on top of each other and used as a search ensemble also in the third round. Twenty three solutions were found with LLG scores ranging from 2434 to 2489. Fixing this result and searching for the *Ct*CBM62 domain once again gave no solutions. So at this stage the *Ct*CBM62 domain could not be located even though our SDS-PAGE analysis clearly shows that the protein did not suffer proteolysis (Figure 6.11). Separate autobuilding runs were carried out using BUCCANEER (Cowtan, 2006) and PHENIX (Terwilliger *et al.*, 2008) and the phases from the third round of PHASER with the *Ct*GH5-CBM6 and Fn3 domains as starting models in their fixed position, but the 2vsa models was not used, thus removing model bias. PHENIX built 637 amino acid residues in 11 fragments with R_{work} of 28.8 % and R_{free} of 34.0 %, whereas BUCCANEER located 657 residues in 10 fragments with R_{work} of 32.5 % and R_{free} of 37.8 %. The two models were superimposed in COOT (Emsley & Cowtan, 2004) using SSM (Krissinel & Henrick, 2004) and used for further manual rebuilding. Structure completion and analysis are ongoing.

Table 6.3| Data collection statistics.

Beamline	**ESRF ID14-EH1**	**ESRF ID14-EH4**	**ESRF ID14-EH4**
Dataset	Fixed	Peak	Inflection point
Space Group	P 2_1 2_1 2	P 2_1 2_1 2	P 2_1 2_1 2
Wavelength (Å)	0.9334	0.979	0.9795
Unit-cell parameters			
a (Å)	147.4	147.3	147.6
b (Å)	191.7	191.1	191.3
c (Å)	50.7	50.7	50.6
Resolution limits (Å)	50.7 – 2.64	80.15-2.98	80.26-2.98
No. of observations	244,475 (9,421/29,324)	135,258 (3,685/19,377)	139,396 (3,604/20,027)
No. of unique observations	42,246 (1,525/5,920)	29,061 (945/4,124)	29,500 (906/4,227)
Multiplicity	5.8 (6.2/5.0)	4.7 (3.9/4.7)	4.7 (4.0/4.7)
Completeness (%)	98.5 (99.6/96.4)	96.7 (88.8/95.9)	98.0 (85.3/98.0)
<I/σ(I)>	8.0 (21.6/2.0)	9.9 (14.1/4)	11.6 (15.2/6.2)
R_{merge} [1]	16.5 (4.5/69.5)	9.4 (5.0/23.5)	8.5 (5.8/17.9)

[1] $R_{merge} = \Sigma_h \Sigma_i |I(h,i) - \langle I(j)\rangle| / \Sigma_h \Sigma_i I(h,i)$, where I(h,i) is the intensity of the measurement of reflection h and <I(h)> is the mean value of I(h,i) for all i measurements. Values in parentheses are for the lowest/highest resolution shells, with the range being 50.7-8.75/2.78– 2.64, 80.15-9.42/3.14-2.98 and 80.26-9.42/3.14-2.98 for each dataset respectively.

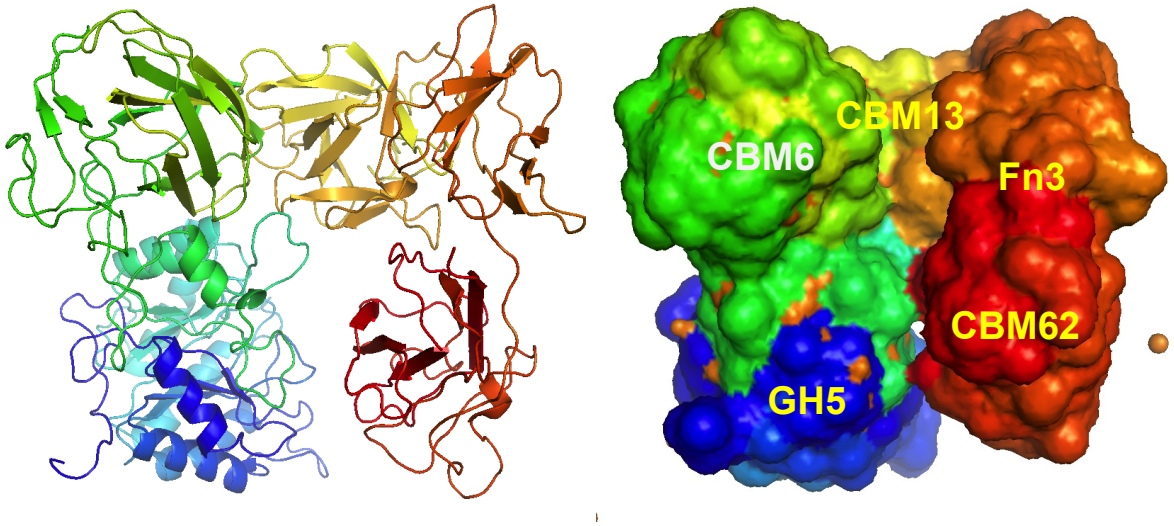
Figure 6.13| Representative diffraction pattern of a *Ct*Xyl5A crystal (the outer circle corresponds to 2.64 Å resolution).



6.2.3. Brief description of the penta-modular cellulosomal arabinoxylanase

As explained above, the crystal structure of this arabinoxylanase has been determined by Molecular Replacement using the GH5-CBM6, Fn3 and CBM62 pdb coordinates to a resolution of 2.64 Å. Overall this 103 KDa penta modular protein displays a compact structure for the first four domains GH5-CBM6-CBM13-FN with a huge cleft in the center (Figure 6.14). Furthermore, it reveals a great flexibility for the CBM62 domain, which is located in a huge solvent channel in the crystal. As for the CBM13 module, it was built *de novo* and it seems to display a classic β-trefoil fold with an unusual track of eight close tryptophan residues in one motif. Four of the tryptophan residues form a flat surface -the binding site - and the other four are internal. A SSM superpositioning (Secondary structure matching) of CBM13 shows that it is homologous with other β-trefoils, CBM13 and CBM42.

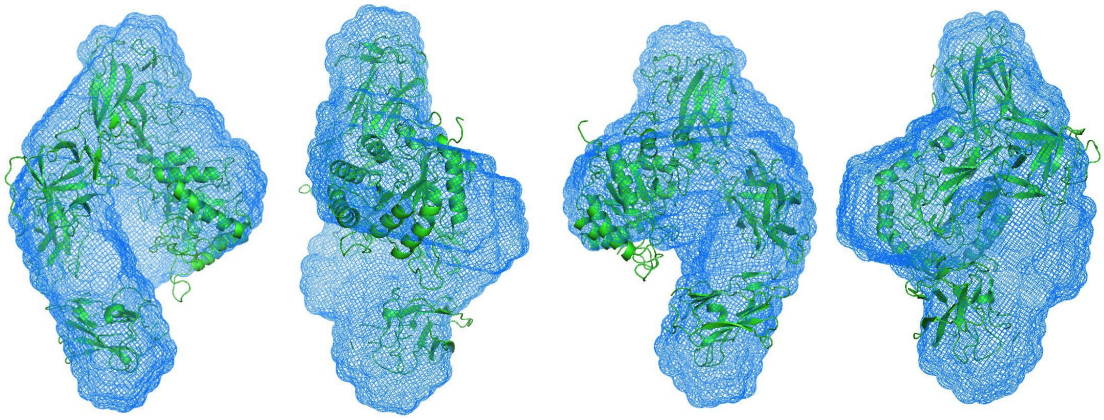
Figure 6.14| Crystal structure of the penta-modular cellulosomal arabinoxylanase.



Overall this 91 kDa penta-modular protein displays a compact structure for the first four modules with greater flexibility for the CBM62 and a huge cleft in the centre.

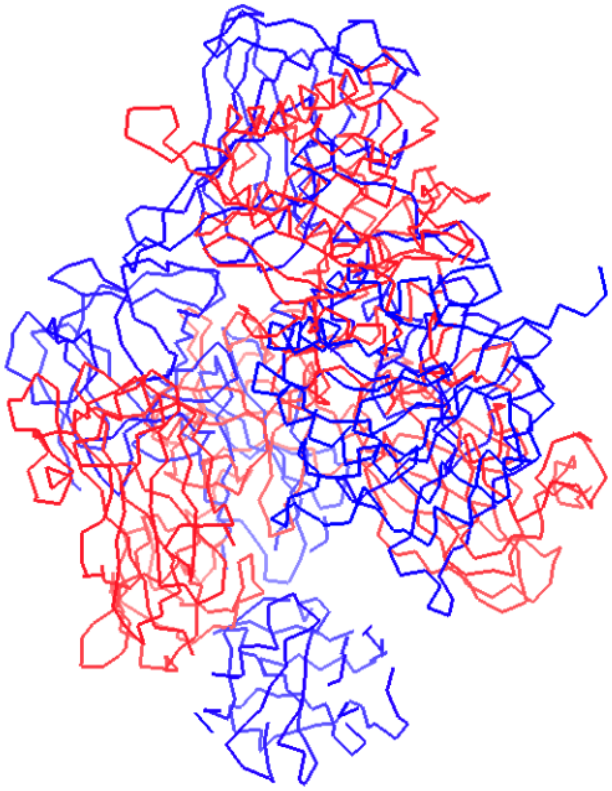
Although Small -angle X-ray scattering and X-ray crystallography are fundamentally similar techniques, SAXS studies were performed in order to complement the crystallographic structure data and obtain more information about flexibility between domains of the *Ct*Xyl5A protein (Figure 6.15). Indeed, SAXS analysis corroborated that all five domains are present, but in a more extended and flexible form.

Figure 6.15| Small Angle X-ray Scattering (SAXS).



The shape of *Ct*Xyl5A in solution obtained from the average of 20 independent simulations produced from DAMMIN (blue). The fit of the five domains (χ=2.2 Å) is shown as cartoon.

Figure 6.16| The SAXS model of the *Ct*Xyl5A.



The SAXS model (blue) shows that *Ct*Xyl5A is more extended and more flexible in solution than *in crystal* (red). The data was collected at the bioSAXS beamline ID14-3, ESRF and processed with the ATSAS suite.

A more detailed analysis of both crystallographic structure and SAXS data is required in order to obtain more information about specific orientations of the five modules within the all protein.

7. GENERAL DISCUSSION AND FUTURE PERSPECTIVES

In the early 1980s, Raffi Lamed and Ed Bayer initiated a work which led to the discovery of the cellulosome (Bayer, Morag, & Lamed, 1994; Lamed, Setter, & Bayer, 1983). Quoting Ed Bayer, “The discovery of the cellulosome is a story of serendipity, a story of imagination, persistence, and a triumph over dogma.” (Bayer & Lamed, 2006). In fact, in the following years much work was performed and only a combination of molecular biology, biochemical, biophysical, immunochemical and ultra-structural techniques led to the elucidation of the cellulosome structure. However, in spite of the tremendous amount of knowledge gained, several questions concerning cellulosome structural organization and function remain unknown. Thus, the main goal of this work was to contribute to the elucidation of some unclear structural and functional issues regarding, not only novel type I and type II cohesin-dockerin interactions, but also the role of cellulosomal GHs and CBMs in plant cell wall hydrolysis.

As stated above, the development of innovative molecular biology and biochemical methods to study the cellulosome is crucial for the scientific advance in this field. Therefore, the second chapter of this thesis reviewed different methods that may be applied to solve the structure of cohesin-dockerin complexes. The initial cohesin-dockerin structures started to reveal the molecular determinants responsible for the high affinity and tight specificity displayed by these protein:protein interactions. Among the methodologies described in chapter 2, perhaps the most critical one is the co-expression of both cohesin and dockerin encoding genes in the same *E. coli* cells under the control of different promoters. Dockerins are highly unstable and very susceptible to proteolysis when expressed individually and co-expression leads to dockerin stabilization. The fusion of cohesins with His tags allowed the use of IMAC as an initial step for complex purification since dockerins bind their cognate partners *in vivo*. Unbound cohesin is then removed through ion exchange chromatography, thus allowing the purification of the protein complex. The application of these methodologies resulted in the production of high levels of stable cohesin-dockerin complexes, an initial pre-requisite for obtaining crystals of the protein complexes. Usually, His-tags are fused to cohesins that normally are expressed at higher levels than dockerins. A recent work of K. Cameron (unpublished data) revealed that integration of a His-tag at the dockerin sequence might result in higher levels of expression of this small peptide. Thus, more rational approaches for the production of protein complexes may be developed which could contemplate the cloning of the two genes in the same vector such that the tag may be inserted either at the cohesin or at the dockerin molecule, both at the C- or N-terminus. This

strategy would allow testing the best expression conditions for each specific protein:protein complex.

Another issue that needed to be considered while producing cohesin-dockerin complexes for crystallization was the dual binding mode expressed by dockerins which can considerably inhibit cohesin-dockerin complex crystallization. Thus, to improve the chances of obtaining crystals of the protein complexes, an inactivation of one of the cohesin binding interfaces is highly recommended, which can be achieved through the use of site-directed mutagenesis to change the residues at positions 11 and 12 of one of the binding faces. The development of molecular biology strategies that allowed 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. For instance, this strategy could be used to clone, in the same vector, different cellulolytic enzymes containing their endogenous dockerins, and also a mini-scaffoldin containing a series of cohesin modules. This approach would allow assembling of mini-cellulosomes *in vivo* and through a quick IMAC method the purification of the complex if the scaffoldin contained an appropriate tag. Given that the dual binding mode expressed by cohesin-dockerin partners is thought to introduce plasticity in the quaternary organization of multi-enzyme complexes, potentiating the hydrolysis of plant cell wall polysaccharides, it is possible to use these methods to test the hydrolytic efficiency of different combinations of catalytic and non-catalytic cellulolytic components. Several applications, such as the using of mini-cellulosomes for the production of biofuels (Cha *et al.*, 2007) or for the improvement of the nutritive value of cereal-based diets for poultry (Ribeiro *et al.*, 2008) could potentially benefit by the implementation of these methods for the generation of protein complexes.

Cohesin-dockerin interactions dictate the overall architecture of cellulosomes. Thus, as stated in chapter 3, the previous methods were applied to determine the crystal structures of two novel type I cohesin-dockerin complexes, here termed OlpA-Doc918 and OlpC-Doc124. *C. thermocellum* cell surface proteins, OlpA and OlpC, were found to have exclusively one type I cohesin each. Thus, dockerin containing enzymes may also adhere directly, and individually, onto the bacterial cell wall (Fontes & Gilbert, 2010; Salamitou *et al.*, 1994). Recently, Pinheiro *et al.,* (2009) revealed that Doc124A appears to bind preferentially to the OlpC type I cohesin, suggesting that a particular set of enzymes might preferentially bind directly to the bacterium cell surface rather than to CipA cohesins. As for the Doc918, the same study argued that this protein displays higher affinity for CipA cohesins, although it also binds OlpA cohesin. Interestingly, both dockerins deviate from the canonical *C. thermocellum* motifs at least in one of the cohesin binding interfaces, which mean that both dockerins lack a distinctive symmetry at the binding interfaces. As such, one of the main goals of this study was to determine the structural determinants of the exquisite specificity revealed by these

novel dockerin modules. The achievement of high quality crystals for both complexes using wild type dockerins was an initial strong indication that these dockerins present a single binding mode. The structures of the two complexes revealed that the critical positions 11 and 12 of the dockerin non-interacting interface are occupied predominantly with acidic residues, which are unable to bind the cohesin hydrophobic cavity that holds the usually highly conserved Ser-Thr dyad. Therefore, Doc124A was shown to bind to the OlpC cohesin through its helix α1, although some amino acid residues of helix α3 are also involved in the binding. ITC data obtained with Doc124 mutant derivatives, in which the main interacting residues were changed, revealed a dramatic reduction in cohesin affinity further supporting the view that this dockerin only contains a single-cohesin binding interface. Considering that this dockerin binds preferentially to the OlpC cell surface protein through the interaction established with its single type I cohesin, it is possible to suggest that there is an evolutionary pressure for a single binding mode for dockerins presented at the cell surface. In fact, the increased flexibility resulting from the dockerin's dual binding mode does not seem to be particularly beneficial when enzymes, such as Doc124A, are predicted to bind to cell surface proteins that only contain a single cohesin. As stated in chapter 5, Doc124A, is an endo-acting cellulase with unique features. Interestingly, cellulose degradation is known to be more efficient when the cellulosome is appended to the bacterium surface rather than when it is released into the culture media (Lu *et al.*, 2006). Therefore, increased cellulose degradation might result from synergistic interactions established between enzymes that are appended to the cellulosome and enzymes, such as *Ct*Cel124, which are predicted to be located at the bacterium cell surface. By targeting the interface between crystalline and amorphous regions of cellulose, *Ct*Cel124 might generate novel substrates for the action of exo-acting cellulases, which are generally attached to cellulosome. Similarly to what was described above for the type I cohesin-dockerin complex OlpC-Doc124A, the OlpA-Doc918 complex revealed an asymmetric interaction that is performed primarily by the dockerin helix α3. The Ser-Thr dyad, only conserved in the C-terminal helix, is responsible for the hydrogen bond network with the cohesin. Mutation of the amino acids Ser49, Thr50 and Lys57 in helix α3 caused knockout of the binding, which further supports that this dockerin only has a single binding mode. The biological significance of this feature is not completely clear, since this dockerin was shown to bind to both cohesins of CipA scaffoldin or OlpA cohesin (Pinheiro *et al.*, 2009). Moreover, the function of protein Cthe_0918 remains unknown, which hinders the interpretation of these results. Nevertheless, OlpA and OlpC might also function by transiently retaining cellulosomal enzymes at the bacterium cell surface before they are assembled into the multi-enzyme complex. This mechanism would also promote the synergistic interactions between catalytic components of the cellulosome and enzymes directly appended to the *C. thermocellum* cell surface. In addition, the dual binding mode, which seems to be a key feature of the majority of *C. thermocellum* dockerins (68 out of 72),

may reduce the steric constraints that are likely to be imposed in the assembling of a large number of different catalytic modules into a single cellulosome and may also introduce the required quaternary flexibility into multi-enzyme complexes. These mechanisms should promote substrate targeting and the synergistic interactions between cellulosomal enzymes.

In chapter 4, the crystal structure of a novel type II cohesin-dockerin complex from *C. thermocellum* cellulosome was described. Once more, the methods revised in the second chapter of this thesis were applied to solve the structure of this cohesin-dockerin complex. The type II cohesin of this complex is the second domain from the anchoring scaffoldin, previously named Orf2 and here renamed ScaC. As for the dockerin, it belongs to a large extracellular protein, named CipB, and is fused to an N-terminal X module. In contrast to what is observed for type I interactions, the type II structure presented in chapter 4 shows that both dockerin helices $\alpha 1$ and $\alpha 3$ interact with the ScaC2 cohesin. However, the number of contacts established by helix $\alpha 3$ predominate and are dominated by amino acids Asn146, Leu147 and Phe148. In addition, the dockerin has two calcium ions that are fundamental for the folding and stabilization of the dockerin and for cohesin recognition. Furthermore, the X module, which presents a β-sheet topology, forms three water mediated hydrogen bonds with the cohesin. The structure of the cohesin displays a high similarity to the previously described type II cohesin from ScaA scaffoldin, at the bound and unbound state. Nevertheless, in this novel complex there is no significant hydrogen bonding present at the complex interface. In fact, the type II complex interface presented in chapter 4 is much more hydrophobic than the one from the previously described CohScaA-XDocCipA complex, indicating that the two type II complex interfaces are relatively different. Alignment of the X-Doc regions of CipA and CipB revealed significant amino acid substitutions at key positions involved in type II cohesin recognition. However, the alignment of CipA dockerin with CipB dockerin rotated 180º shows a higher conservation at the amino acid residues that are responsible for the main contacts with the cohesin ScaA, suggesting that CipB dockerin might interact with ScaA cohesin through its N-terminal helix. In contrast, there is a higher conservation of ScaC2 cohesin-contact residues when the structure of CipB dockerin is superimposed with that of CipA dockerin rotated by 180º. Taken together, these observations suggest that although *C. thermocellum* dockerins bind the cohesins through a single binding mode, they might contain two different cohesin binding interfaces that will discriminate from the different type II cohesins. Further structural work is on-going in order to investigate this hypothesis.

The functional importance of specific amino acid residues identified at the CohScaC2-XDocCipB binding surface was assessed through ITC using ScaC2 cohesin and CipB dockerin. Since the affinity between the wild-type proteins was above the detection limits for this technique it was not possible to derive appropriate affinity constants for this interaction.

Interestingly, recent genome sequencing revealed the presence of two novel proteins, ScaD and ScaE, containing type II cohesins. ScaD contains one type II cohesin and a module with unknown function. ScaE contains seven type II cohesins but lacks the SLH domain required for cell surface anchoring. This circumstance suggests that this protein may be exclusively extracellular, allowing the presence of free non-cell attached cellulosomes in *C. thermocellum*. Native gel electrophoresis revealed that ScaE6 is the cohesin to which XDocCipB binds with the lowest affinity. Thus, it was possible to use ITC to determine the affinity constants of the type II dockerin of CipB against ScaE6 cohesin. Site-directed mutagenesis data revealed that substitution of residues Phe124, Leu147 and Phe148 for alanine in the XDocCipB resulted in the complete abolition of cohesin binding. As for the affinities between these three XDocCipB mutant derivatives and the type II cohesin ScaC2, the values were also shown to be lower than the detection limit.

Qualitative competition experiments suggested that type II cohesins do not have a preference for either CipA or CipB dockerins. This observation indicates that both CipA and CipB can bind equally well to all anchoring scaffoldins described so far, which means that poly-cellulosomes may contain these two proteins. However, both dockerins display a clear preference for specific cohesins. The ScaE6 was found to be the less preferred, which is in agreement with the lowest affinity obtained in the ITC assays. Besides, as stated in the introduction section, the abundance of anchor proteins is inversely proportional to the number of type II cohesin modules borne by them (Raman *et al.*, 2009). Indeed, ScaE contains seven cohesin modules and is one of the least abundant anchoring proteins. This suggest that ScaE scaffoldin may have an auxiliary function in the later stages of the cell wall degradation and only when all the cell surface proteins are already saturated with CipA or CipB. Moreover, ScaC2 was the preferred protein partner for both dockerins tested. ScaC is likely to be the preferred anchoring scaffoldin since it has two type II cohesin modules. This way, it can incorporate two CipA molecules at the same time, thus obtaining a poly-cellulosome which can contain a total of 18 enzymes. According to Raman *et al.* (2009), ScaC is one of the most abundant anchoring scaffoldins, which supports a major role for this protein in poly-cellulosome assembly.

Although the large multi-modular 2177-residue extracellular protein CipB has been extensively investigated for its capacity to act either as a CBM or as a carbohydrate active enzyme, its function remains unknown. Nevertheless, recent proteomic studies revealed that CipB is up-regulated when *C. thermocellum* is grown on cellulose (Raman *et al.*, 2009). Additional studies, such as Small Angle X-ray Scattering assays might provide some vital data to elucidate the role of this unusual multi-modular cellulosomal protein in carbohydrate hydrolysis.

In chapter 5, the crystal structure and the biochemical properties of the cellulosomal protein *Ct*Cel14 previously of unknown function were determined. The *Ct*Cel124 was described here as a cellulase. A previous proteomic study revealed that this protein is highly up-regulated when *C. thermocellum* is cultured on crystalline cellulose (Raman *et al.*, 2009). This fact is consistent with the results obtained here for *Ct*Cel124 role in cellulose metabolism. In the presence of phosphoric acid swollen cellulose (PASC) the reaction products released by the enzyme ranged from cellotriose to cellohexaose, which is typical of endo-acting enzymes. Thus, *Ct*Cel124 was classified as an endo-β-1,4-glucanase that likely plays a vital role in the degradation of plant cell wall polysaccharides by *C. thermocellum*.

As described in the bibliographic review (chapter 1), exo- and endo-acting cellulases act in synergy to hydrolyse cellulose. Thus, *Ct*Cel48S, the most abundant exo-acting cellulase in *C. thermocellum*, was shown to act cooperatively with *Ct*Cel124 in order to hydrolyse highly crystalline cellulose. In addition, a more extensive synergy was observed between the two enzymes when both were appended to CBM3a modules, which bind to crystalline cellulose.

The crystal structure of *Ct*Cel124 was determined in complex with two cellotriose molecules. *Ct*Cel124 displays a α8 superhelical fold, which has not been previously observed in other cellulase families. It is highly likely that convergent evolution may explain the existence of so many different folds in cellulases. Since there are four more proteins which display significant sequence identity with *Ct*Cel124, it was proposed that this new enzyme comprises the founding member of a new CAZy family designated by GH124.The aforementioned superhelical fold provides a platform that supports the substrate binding cleft, which was crystallized in the presence of two cellotriose molecules. The bound cellooligosaccharides adopt two distinct conformations, thus providing vital insights into the likely topological features of cellulose bound to *Ct*Cel124. In addition, the linear conformation adopted by one of the cellotriose molecules is similar to the structure of the glucan chains in crystalline cellulose. Also, the helical structure observed in the other cellotriose molecule is usually adopted by cellooligosaccharides in solution. As such, it is likely that the substrate binding cleft of *Ct*Cel124 is tailored to target the interface between crystalline and amorphous regions of cellulose. Therefore, some endoglucanases expressed by cellulose-degrading bacteria might be tailored to recognize specific regions of the cellulose structure, providing novel opportunities to generate different substrates for the exo-acting cellulases.

In chapter 6, the crystal structure of the multi-modular arabinoxylanase *Ct*Xyl5A from *C. thermocellum* cellulosome was described. This protein is one of the largest components of the cellulosome and has a molecular weight of 103 kDa. As described before, *Ct*Xyl5A comprises a N-terminal GH5, followed by two CBMs from families 6 and 13, a fibronectin type III-like module, a CBM from family 62 and, finally, a type I dockerin. Initially, the crystal structure and biochemical properties of the *Ct*GH5 catalytic module appended to the CBM6

were assessed. The structure revealed that *Ct*GH5 displays a canonical $(\alpha/\beta)_8$-barrel fold with the substrate binding cleft running along the surface of the protein. The enzyme displays specificity for arabinoxylans and thus is defined as an arabinoxylanase, a catalytic activity not previously reported. Xylanases known so far do not generally distinguish between different xylans, although they are unable to access highly decorated forms of xylan, such as arabinoxylans. As for the *Ct*CBM6, it adopts a β-sandwich fold and recognizes the termini of xylo- and gluco-configured oligosaccharides. *Ct*CBM6 displays affinity for the reaction products generated by *Ct*GH5 and for undecorated xylooligosaccharides. This suggests that this arabinoxylanase is targeted to regions of the plant cell wall that are undergoing active degradation. Interestingly, there is an extensive hydrophobic interface between *Ct*GH5 and *Ct*CBM6, which is in contrast with the typical modular GHs. Consequently, the two modules cannot fold as independent entities. Indeed, when expressed independently, *Ct*CBM6 does not bind to cellohexaose or xylohexaose and *Ct*GH5 exhibits very low catalytic activity. Recently, the structure of the pentamodular arabinoxylanase *Ct*Xyl5A was determined by molecular replacement. *Ct*Xyl5A displays a compact structure for the first four domains GH5, CBM6, CBM13 and FN with a huge cleft in the centre. In addition, the structure revealed a great flexibility for the *Ct*CBM62 domain. The SAXS model showed that *Ct*Xyl5A is more extended and more flexible in solution than in crystal. A previous study revealed that *Ct*CBM62 binds tightly to xyloglucan, arabinogalactan and galactomannan, but does not recognize arabinoxylans hydrolysed by *Ct*Xyl5A. Furthermore, *Ct*CBM62 is capable of calcium induced dimerization (Montanier *et al.*, 2011). Thus, one can predict that the dimerization may lead to the formation of poly-cellulosomes, through a crosslinking process between two *Ct*Xyl5A, which would be bound to different CipA molecules. On the other hand, it is also possible that two *Ct*CBM62 molecules recruit two molecules of *Ct*Xyl5A onto the same cellulosome. The *Ct*CBM13 module was built *de novo* and seems to display a classic β-trefoil fold. Binding assays (data not shown) revealed that *Ct*CBM13 interacts more weakly to insoluble wheat arabinoxylans than *Ct*CBM6. As for the FN, the function of this domain remains unclear, but previous studies revealed that it might function as ligand-binding module, as a compact form of peptide linker or spacer between other domains, as cellulose-disrupting module or even as a protein that help large enzyme complexes to remain soluble (Alahuhta *et al.*, 2010). Further studies are required in order to clarify FN function in the pentamodular structure. The biological significance of the five modules with different functions, together in the same protein remains to be elucidated. A more detailed analysis of all results obtained so far is needed in order to clarify the possible synergy between all *Ct*Xyl5A modules or between its modules and other cellulosomal enzymes.

Plant cell wall polysaccharides offer an extraordinary source of carbon and energy that can be used by many microorganisms. The enzymatic degradation of plant cell wall plays a

central role in the carbon cycle and is of increasing environmental and industrial significance. Several applications of cellulases or hemicellulases are being developed for several industries. Total hydrolysis of cellulose into glucose, which could be fermented into ethanol, isopropanol or butanol, is not yet economically feasible. However, the need to reduce emissions of greenhouse gases provides an additional incentive for the development of processes generating fuels from cellulose. Indeed, understanding the mechanisms by which cellulases and hemicellulases interact synergistically to hydrolyse the plant cell wall, as well as elucidating the main structure and function relationships between all cellulosomal components, are important biological issues that, although partially addressed here, need further research.

8. BIBLIOGRAPHIC REFERENCES

Abbott, D. W., Ficko-Blean, E., van Bueren, A. L., Rogowski, A., Cartmell, A., Coutinho, P. M., Henrissat, B., *et al.* (2009). Analysis of the structural and functional diversity of plant cell wall specific family 6 carbohydrate binding modules. *Biochemistry*, *48*(43), 10395-10404.

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.

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., 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.

Adams, P. D., Afonine, P. V., Bunkóczi, G., Chen, V. B., Davis, I. W., Echols, N., Headd, J. J., *et al.* (2010). PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta crystallographica. Section D, Biological crystallography*, *66*(Pt 2), 213-221.

Alahuhta, M., Xu, Q., Brunecky, R., Adney, W. S., Ding, S.-Y., Himmel, M. E., & Lunin, V. V. (2010). Structure of a fibronectin type III-like module from *Clostridium thermocellum*. *Acta Crystallographica. Section F, Structural Biology and Crystallization Communications*, *66*(Pt 8), 878-880.

Ali, B. R., Zhou, L., Graves, F. M., Freedman, R. B., Black, G. W., Gilbert, H. J., & Hazelwood, G. P. (1995). Cellulases and hemicellulases of the anaerobic fungus *Piromyces* constitute a multiprotein cellulose-binding complex and are encoded by multigene families. *FEMS Microbiology Letters*, *125*(1), 15-21.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403-410.

Araki, R., Ali, M. K., Sakka, M., Kimura, T., Sakka, K., & Ohmiya, K. (2004). Essential role of the family-22 carbohydrate-binding modules for beta-1,3-1,4-glucanase activity of *Clostridium stercorarium* Xyn10B. *FEBS Letters*, *561*(1-3), 155-158.

Bagnara-Tardif, C., Gaudin, C., Belaich, A., Hoest, P., Citard, T., & Belaich, J. P. (1992). Sequence analysis of a gene cluster encoding cellulases from *Clostridium cellulolyticum. Gene*, *119*(1), 17-28.

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.

Barral, P., Suárez, C., Batanero, E., Alfonso, C., Alché, J. de D., Rodríguez-García, M. I., Villalba, M., *et al.* (2005). An olive pollen protein with allergenic activity, Ole e 10, defines a novel family of carbohydrate-binding modules and is potentially implicated in pollen germination. *The Biochemical Journal*, *390*(Pt 1), 77-84.

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, Morag, E., & Lamed, R. (1994). The cellulosome--a treasure-trove for biotechnology. *Trends in Biotechnology*, *12*(9), 379-386.

Bayer, E A, Setter, E., & Lamed, R. (1985). Organization and distribution of the cellulosome in *Clostridium thermocellum*. *Journal of Bacteriology*, *163*(2), 552-559.

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

Bayer, Edward A, Lamed, R., White, B. A., & Flint, H. J. (2008). From cellulosomes to cellulosomics. *Chemical record (New York, N.Y.)*, *8*(6), 364-377.

Bayer, Edward 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*(1), 521-554.

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

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.

Béguin, P., & Alzari, P. M. (1998). The cellulosome of *Clostridium thermocellum*. *Biochemical Society transactions*, *26*(2), 178-185.

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

Béguin, 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.

Berg Miller, M. E., Antonopoulos, D. A., Rincon, M. T., Band, M., Bari, A., Akraiko, T., Hernandez, A., *et al.* (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.

Beylot, M. H., McKie, V. A., Voragen, A. G., Doeswijk-Voragen, C. H., & Gilbert, H. J. (2001). The *Pseudomonas cellulosa* glycoside hydrolase family 51 arabinofuranosidase exhibits wide substrate specificity. *The Biochemical Journal*, *358*(Pt 3), 607-614.

Blake, A. W., McCartney, L., Flint, J. E., Bolam, D. N., Boraston, A. B., Gilbert, H. J., & Knox, J. P. (2006). Understanding the biological rationale for the diversity of cellulose-directed carbohydrate-binding modules in prokaryotic enzymes. *The Journal of Biological Chemistry*, *281*(39), 29321-29329.

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

Bolam, David N, Xie, H., Pell, G., Hogg, D., Galbraith, G., Henrissat, B., & Gilbert, H. J. (2004). X4 modules represent a new family of carbohydrate-binding modules that display novel properties. *The Journal of Biological Chemistry*, *279*(22), 22953-22963.

Boraston, A B, Creagh, A. L., Alam, M. M., Kormos, J. M., Tomme, P., Haynes, C. A., Warren, R. A., *et al.* (2001). Binding specificity and thermodynamics of a family 9 carbohydrate-binding module from *Thermotoga maritima* xylanase 10A. *Biochemistry*, *40*(21), 6240-6247.

Boraston, Alisdair 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.

Boraston, Alisdair 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.

Brás, J. L. A., Alves, V. D., Carvalho, A. L., Najmudin, S., Prates, J. A. M., Ferreira, L. M. A., Bolam, D. N., et al. (2012). Novel Clostridium thermocellum type I cohesin-dockerin complexes reveal a single binding mode. *The Journal of biological chemistry*. doi:10.1074/jbc.M112.407700

Brás, J. L. A., Cartmell, A., Carvalho, A. L. M., Verzé, G., Bayer, E. A., Vazana, Y., Correia, M. A. S., *et al.* (2011). Structural insights into a unique cellulase fold and mechanism of cellulose hydrolysis. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(13), 5237-5242.

Brás, J. L. A., Carvalho, A. L., Viegas, A., Najmudin, S., Alves, V. D., Prates, J. A. M., Ferreira, L. M. A., *et al.* (2012). *Escherichia coli* Expression, Purification, Crystallization, and Structure Determination of Bacterial Cohesin-Dockerin Complexes. *Methods in enzymology*, *510*, 395-415.

Brás, J. L. A., Correia, M. A. S., Romão, M. J., Prates, J. A. M., Fontes, C. M. G. A., & Najmudin, S. (2011). Purification, crystallization and preliminary X-ray characterization of the pentamodular arabinoxylanase *Ct*Xyl5A from *Clostridium thermocellum*. *Acta Crystallographica. Section F, Structural Biology and Crystallization Communications*, *67*(Pt 7), 833-836.

Braun, C., Meinke, A., Ziser, L., & Withers, S. G. (1993). Simultaneous high-performance liquid chromatographic determination of both the cleavage pattern and the stereochemical outcome of the hydrolysis reactions catalyzed by various glycosidases. *Analytical Biochemistry*, *212*(1), 259-262.

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

van Bueren, A. L., Morland, C., Gilbert, H. J., & Boraston, A. B. (2005). Family 6 carbohydrate binding modules recognize the non-reducing end of beta-1,3-linked glucans by presenting a unique ligand binding surface. *The Journal of Biological Chemistry*, *280*(1), 530-537.

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.

Carvalho, Ana L, Dias, F. M. V., Prates, J. A. M., Nagy, T., Gilbert, H. J., Davies, G. J., Ferreira, L. M. A., *et al.* (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.

Carvalho, Ana L, Goyal, A., Prates, J. A. M., Bolam, D. N., Gilbert, H. J., Pires, V. M. R., Ferreira, L. M. A., *et al.* (2004). The family 11 carbohydrate-binding module of *Clostridium thermocellum* Lic26A-Cel5E accommodates beta-1,4- and beta-1,3-1,4-mixed linked glucans at a single binding site. *The Journal of Biological Chemistry*, *279*(33), 34785-34793.

Carvalho, Ana L, Pires, V. M. R., Gloster, T. M., Turkenburg, J. P., Prates, J. A. M., Ferreira, L. M. A., Romão, M. J., *et al.* (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.

Carvalho, Ana Luísa, Dias, F. M. V., Nagy, T., Prates, J. A. M., Proctor, M. R., Smith, N., Bayer, E. A., *et al.* (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.

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.

Charnock, S J, Lakey, J. H., Virden, R., Hughes, N., Sinnott, M. L., Hazlewood, G. P., Pickersgill, R., *et al.* (1997). Key residues in subsite F play a critical role in the activity of *Pseudomonas fluorescens* subspecies cellulosa xylanase A against xylooligosaccharides but not against highly polymeric substrates such as xylan. *The Journal of Biological Chemistry*, *272*(5), 2942-2951.

Charnock, S J, Bolam, D. N., Turkenburg, J. P., Gilbert, H. J., Ferreira, L. M., Davies, G. J., & Fontes, C. M. (2000). The X6 «thermostabilizing» domains of xylanases are

carbohydrate-binding modules: structure and biochemistry of the *Clostridium thermocellum* X6b domain. *Biochemistry*, *39*(17), 5013-5021.

Charnock, Simon J, Bolam, D. N., Nurizzo, D., Szabó, L., McKie, V. A., Gilbert, H. J., & Davies, G. J. (2002). Promiscuity in ligand-binding: The three-dimensional structure of a *Piromyces* carbohydrate-binding module, CBM29-2, in complex with cello- and mannohexaose. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(22), 14077-14082.

Chen, V. B., Arendall, W. B., Headd, J. J., Keedy, D. A., Immormino, R. M., Kapral, G. J., Murray, L. W., *et al.* (2010). MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallographica Section D: Biological Crystallography*, *66*(Pt 1), 12-21.

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.

Ciucanu, I., & Kerek, F. (1984). A simple and rapid method for the permethylation of carbohydrates. *Carbohydrate Research*, *131*(2), 209-217.

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

Correia, M. A. S., Mazumder, K., Brás, J. L. A., Firbank, S. J., Zhu, Y., Lewis, R. J., York, W. S., *et al.* (2011). Structure and function of an arabinoxylan-specific xylanase. *The Journal of Biological Chemistry*, *286*(25), 22510-22520.

Cosgrove, D J. (1997). Assembly and enlargement of the primary cell wall in plants. *Annual Review of Cell and Developmental Biology*, *13*, 171-201.

Cosgrove, D J. (2001). Wall structure and wall loosening. A look backwards and forwards. *Plant Physiology*, *125*(1), 131-134.

Cosgrove, Daniel J. (2005). Growth of the plant cell wall. *Nature Reviews. Molecular Cell Biology*, *6*(11), 850-861.

Coutinho, P. M., & Reilly, P. J. (1994). Structure-function relationships in the catalytic and starch binding domains of glucoamylase. *Protein Engineering*, *7*(3), 393-400.

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.

Cowtan, K. D., & Main, P. (1993). Improvement of macromolecular electron-density maps by the simultaneous application of real and reciprocal space constraints. *Acta Crystallographica. Section D, Biological Crystallography*, *49*(Pt 1), 148-157.

Czjzek, M., Bolam, D. N., Mosbah, A., Allouch, J., Fontes, C. M., Ferreira, L. M., Bornet, O., *et al.* (2001). The location of the ligand-binding site of carbohydrate-binding modules that have evolved from a common sequence is not conserved. *The Journal of Biological Chemistry*, *276*(51), 48580-48587.

Damude, H. G., Withers, S. G., Kilburn, D. G., Miller, R. C., Jr, & Warren, R. A. (1995). Site-directed mutation of the putative catalytic residues of endoglucanase CenA from *Cellulomonas fimi*. *Biochemistry*, *34*(7), 2220-2224.

Dashtban, M., Schraft, H., & Qin, W. (2009). Fungal bioconversion of lignocellulosic residues; opportunities & perspectives. *International Journal of Biological Sciences*, *5*(6), 578-595.

Davies, G. J., Wilson, K. S., & Henrissat, B. (1997). Nomenclature for sugar-binding subsites in glycosyl hydrolases. *The Biochemical Journal*, *321 (Pt 2)*, 557-559.

DeBoy, R. T., Mongodin, E. F., Fouts, D. E., Tailford, L. E., Khouri, H., Emerson, J. B., Mohamoud, Y., *et al.* (2008). Insights into plant cell wall degradation from the genome sequence of the soil bacterium *Cellvibrio japonicus*. *Journal of Bacteriology*, *190*(15), 5455-5463.

Demain, A. L., Newcomb, M., & Wu, J. H. D. (2005). Cellulase, clostridia, and ethanol. *Microbiology and Molecular Biology Reviews: MMBR*, *69*(1), 124-154.

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.

Dijkerman, R., Op den Camp, H. J., Van der Drift, C., & Vogels, G. D. (1997). The role of the cellulolytic high molecular mass (HMM) complex of the anaerobic fungus *Piromyces* sp. strain E2 in the hydrolysis of microcrystalline cellulose. *Archives of Microbiology*, *167*(2-3), 137-142.

Ding, S. Y., Rincon, M. T., Lamed, R., Martin, J. C., McCrae, S. I., Aurilia, V., Shoham, Y., *et al.* (2001). Cellulosomal scaffoldin-like proteins from *Ruminococcus flavefaciens. Journal of Bacteriology*, *183*(6), 1945-1953.

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

Dolinsky, T. J., Nielsen, J. E., McCammon, J. A., & Baker, N. A. (2004). PDB2PQR: an automated pipeline for the setup of Poisson-Boltzmann electrostatics calculations. *Nucleic acids research*, *32*(Web Server issue), W665-667.

Domon, B., & Costello, C. E. (1988). Structure elucidation of glycosphingolipids and gangliosides using high-performance tandem mass spectrometry. *Biochemistry*, *27*(5), 1534-1543.

Emsley, P, Lohkamp, B., Scott, W. G., & Cowtan, K. (2010). Features and development of Coot. *Acta crystallographica. Section D, Biological crystallography*, *66*(Pt 4), 486-501.

Emsley, Paul, & Cowtan, K. (2004). Coot: model-building tools for molecular graphics. *Acta Crystallographica. Section D, Biological Crystallography*, *60*(Pt 12 Pt 1), 2126-2132.

Evans, P. (1993). Data reduction, CCP4 Daresbury Study Weekend: Data Collection and Processing. *Daresbury Laboratory, Warrington, UK*, *DL/SCI/R34*, 114-122.

Evans, Philip. (2006). Scaling and assessment of data quality. *Acta crystallographica. Section D, Biological crystallography*, *62*(Pt 1), 72-82.

Felix, C. R., & Ljungdahl, L. G. (1993). The cellulosome: the exocellular organelle of *Clostridium*. *Annual review of microbiology*, *47*, 791-819.

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., Bayer, E. A., Tardif, C., Czjzek, M., Mechaly, A., Bélaïch, A., Lamed, R., *et al.* (2002). Degradation of cellulose substrates by cellulosome chimeras. Substrate targeting versus proximity of enzyme components. *The Journal of Biological Chemistry*, *277*(51), 49621-49630.

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*, 655-681.

La Fortelle L., & G. Bricogne. (1997). Maximum-Likelihood Heavy-Atom Parameter Refinement for Multiple Isomorphous Replacement and Multiwavelength Anomalous Diffraction Methods. *Methods Enzymol.*, *276*, 472-494.

Fry, S. (1989). Cellulases, Hemicellulases and Auxin-stimulated Growth - A possible relationship. *Physiologia Plantarum*, *75*(4), 532-536.

Gao, P.-J., Chen, G.-J., Wang, T.-H., Zhang, Y.-S., & Liu, J. (2001). Non-hydrolytic Disruption of Crystalline Structure of Cellulose by Cellulose Binding Domain and Linker Sequence of Cellobiohydrolase I from *Penicillium janthinellum*. *Sheng Wu Hua Xue Yu Sheng Wu Wu Li Xue Bao Acta Biochimica Et Biophysica Sinica*, *33*(1), 13-18.

García-Alvarez, B., Melero, R., Dias, F. M. V., Prates, J. A. M., Fontes, C. M. G. A., Smith, S. P., Romão, M. J., *et al.* (2011). Molecular architecture and structural transitions of a *Clostridium thermocellum* mini-cellulosome. *Journal of Molecular Biology*, *407*(4), 571-580.

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.

Gilbert, H. J. (2007). Cellulosomes: microbial nanomachines that display plasticity in quaternary structure. *Molecular Microbiology*, *63*(6), 1568-1576.

Gilbert, H. J. (2010). The Biochemistry and Structural Biology of Plant Cell Wall Deconstruction. *Plant Physiology*, *153*(2), 444 -455.

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.

Gilkes, N. R., Warren, R. A., Miller, R. C., Jr, & 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.

Gold, N. D., & Martin, V. J. J. (2007). Global view of the *Clostridium thermocellum* cellulosome revealed by quantitative proteomic analysis. *Journal of bacteriology*, *189*(19), 6787-6795.

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*, 89.

Gruppen, H., Hoffmann, R. A., Kormelink, F. J., Voragen, A. G., Kamerling, J. P., & Vliegenthart, J. F. (1992). Characterisation by 1H NMR spectroscopy of enzymically derived oligosaccharides from alkali-extractable wheat-flour arabinoxylan. *Carbohydrate Research*, *233*, 45-64.

Guillén, D., Sánchez, S., & Rodríguez-Sanoja, R. (2010). Carbohydrate-binding domains: multiplicity of biological roles. *Applied Microbiology and Biotechnology*, *85*(5), 1241-1249.

Guimarães, B. G., Souchon, H., Lytle, B. L., David Wu, J. H., & Alzari, P. M. (2002). The crystal structure and catalytic mechanism of cellobiohydrolase CelS, the major enzymatic component of the *Clostridium thermocellum* Cellulosome. *Journal of molecular biology*, *320*(3), 587-596.

Ha, M., Apperley, D., & Jarvis, M. (1997). Molecular rigidity in dry and hydrated onion cell walls RID F-7858-2011. *Plant Physiology*, *115*(2), 593-598.

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.

Hall, J., Black, G. W., Ferreira, L. M., Millward-Sadler, S. J., Ali, B. R., Hazlewood, G. P., & Gilbert, H. J. (1995). The non-catalytic cellulose-binding domain of a novel cellulase from *Pseudomonas fluorescens* subsp. cellulosa is important for the efficient hydrolysis of Avicel. *The Biochemical Journal*, *309 (Pt 3)*, 749-756.

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.

Hammel, M., Fierobe, H.-P., Czjzek, M., Kurkal, V., Smith, J. C., Bayer, E. A., Finet, S., *et al.* (2005). Structural basis of cellulosome efficiency explored by small angle X-ray scattering. *The Journal of Biological Chemistry*, *280*(46), 38562-38568.

Han, S. O., Cho, H.-Y., Yukawa, H., Inui, M., & Doi, R. H. (2004). Regulation of expression of cellulosomes and noncellulosomal (hemi)cellulolytic enzymes in *Clostridium cellulovorans* during growth on different carbon sources. *Journal of Bacteriology*, *186*(13), 4218-4227.

Han, S. O., Yukawa, H., Inui, M., & Doi, R. H. (2003). Transcription of *Clostridium cellulovorans* cellulosomal cellulase and hemicellulase genes. *Journal of Bacteriology*, *185*(8), 2520-2527.

Harholt, J., Suttangkakul, A., & Vibe Scheller, H. (2010). Biosynthesis of pectin. *Plant Physiology*, *153*(2), 384-395.

Hashimoto, H. (2006). Recent structural studies of carbohydrate-binding modules. *Cellular and Molecular Life Sciences: CMLS*, *63*(24), 2954-2967.

Hayashi. (1989). Xyloglucans in the Primary-Cell Wall. *Annual review of plant physiology and plant molecular biology*, *40*(1), 139-168.

Hazlewood, G. P., Davidson, K., Laurie, J. I., Romaniec, M. P., & Gilbert, H. J. (1990). Cloning and sequencing of the celA gene encoding endoglucanase A of *Butyrivibrio fibrisolvens* strain A46. *Journal of General Microbiology*, *136*(10), 2089-2097.

Heinze, T. (Ed.). (2005). *Polysaccharides I: Structure, Characterisation and Use* (1st ed.). Springer.

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

Henrissat, B., Callebaut, I., Fabrega, S., Lehn, P., Mornon, J. P., & Davies, G. (1995). Conserved catalytic machinery and the prediction of a common fold for several families of glycosyl hydrolases. *Proceedings of the National Academy of Sciences of the United States of America*, *92*(15), 7090-7094.

Henshaw, J. L., Bolam, D. N., Pires, V. M. R., Czjzek, M., Henrissat, B., Ferreira, L. M. A., Fontes, C. M. G. A., *et al.* (2004). The family 6 carbohydrate binding module *Cm*CBM6-2 contains two ligand-binding sites with distinct specificities. *The Journal of Biological Chemistry*, *279*(20), 21552-21559.

Hilge, M., Gloor, S. M., Rypniewski, W., Sauer, O., Heightman, T. D., Zimmermann, W., Winterhalter, K., *et al.* (1998). High-resolution native and complex structures of thermostable beta-mannanase from *Thermomonospora fusca* - substrate specificity in glycosyl hydrolase family 5. *Structure (London, England: 1993)*, *6*(11), 1433-1444.

Himmel, M. E., & Bayer, E. A. (2009). Lignocellulose conversion to biofuels: current challenges, global perspectives. *Current Opinion in Biotechnology*, *20*(3), 316-317.

Hoffmann, R. A., Leeflang, B. R., de Barse, M. M., Kamerling, J. P., & Vliegenthart, J. F. (1991). Characterisation by 1H-n.m.r. spectroscopy of oligosaccharides, derived from arabinoxylans of white endosperm of wheat, that contain the elements ----4)[alpha-L-Araf-(1----3)]-beta-D-Xylp-(1---- or ----4)[alpha- L-Araf-(1--2)][alpha-L-Araf-(1----3)]-beta-D-Xylp. *Carbohydrate Research*, *221*, 63-81.

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.

Jamal-Talabani, S., Boraston, A. B., Turkenburg, J. P., Tarbouriech, N., Ducros, V. M.-A., & Davies, G. J. (2004). Ab initio structure determination and functional characterization of CBM36; a new family of calcium-dependent carbohydrate binding modules. *Structure (London, England: 1993)*, *12*(7), 1177-1187.

Jenkins, J., Lo Leggio, L., Harris, G., & Pickersgill, R. (1995). Beta-glucosidase, beta-galactosidase, family A cellulases, family F xylanases and two barley glycanases form a superfamily of enzymes with 8-fold beta/alpha architecture and with two conserved glutamates near the carboxy-terminal ends of beta-strands four and seven. *FEBS Letters*, *362*(3), 281-285.

Jindou, S., Borovok, I., Rincon, M. T., Flint, H. J., Antonopoulos, D. A., Berg, M. E., White, B. A., *et al.* (2006). Conservation and divergence in cellulosome architecture between two strains of *Ruminococcus flavefaciens. Journal of Bacteriology*, *188*(22), 7971-7976.

Jindou, S., Brulc, J. M., Levy-Assaraf, M., Rincon, M. T., Flint, H. J., Berg, M. E., Wilson, M. K., *et al.* (2008). Cellulosome gene cluster analysis for gauging the diversity of the ruminal cellulolytic bacterium *Ruminococcus flavefaciens*. *FEMS Microbiology Letters*, *285*(2), 188-194.

Jindou, S., Soda, A., Karita, S., Kajino, T., Béguin, P., Wu, J. H. D., Inagaki, M., *et al.* (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.

Kakiuchi, M., Isui, A., Suzuki, K., Fujino, T., Fujino, E., Kimura, T., Karita, S., *et al.* (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.

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.

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.

Keegstra, K, Talmadge, K. W., Bauer, W. D., & Albersheim, P. (1973). The Structure of Plant Cell Walls: III. A Model of the Walls of Suspension-cultured Sycamore Cells Based on the Interconnections of the Macromolecular Components. *Plant Physiology*, *51*(1), 188-197.

Keegstra, Kenneth. (2010). Plant cell walls. *Plant Physiology*, *154*(2), 483-486.

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.

Koivula, A., Ruohonen, L., Wohlfahrt, G., Reinikainen, T., Teeri, T. T., Piens, K., Claeyssens, M., *et al.* (2002). The active site of cellobiohydrolase Cel6A from *Trichoderma reesei:* the roles of aspartic acids D221 and D175. *Journal of the American Chemical Society*, *124*(34), 10015-10024.

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.

Krissinel, Evgeny, & Henrick, K. (2007). Inference of macromolecular assemblies from crystalline state. *Journal of molecular biology*, *372*(3), 774-797.

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.

Lamed, R., & Zeikus, J. G. (1980). Ethanol production by thermophilic bacteria: relationship between fermentation product yields of and catabolic enzyme activities in *Clostridium thermocellum* and *Thermoanaerobium brockii*. *Journal of Bacteriology*, *144*(2), 569-578.

Lamzin, V. S., & Wilson, K. S. (1993). Automated refinement of protein models. *Acta Crystallographica. Section D, Biological Crystallography*, *49*(Pt 1), 129-147.

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.

Laskowski, R. A., MacArthur, M. W., Moss, D. S., & Thornton, J. M. (1993). PROCHECK: a program to check the stereochemical quality of protein structures. *Journal of Applied Crystallography*, *26*(2), 283-291.

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., 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.

Leslie, A. G. W. (1992). Joint CCP4/ESF-EAMCB. *Newsl. Protein Crystallogr.*, *26*, 27-33.

Leslie, Andrew G W. (2006). The integration of macromolecular diffraction data. *Acta Crystallographica. Section D, Biological Crystallography*, *62*(Pt 1), 48-57.

Leslie, A. G. W., & Powel, H. R. (2007). Processing diffraction data with mosflm. *Evolving Methods for Macromolecular Crystallography*, *245*, 41-51 ISBN 978-1-4020-6314-5.

Ljungdahl, L. G. (2008). The cellulase/hemicellulase system of the anaerobic fungus *Orpinomyces* PC-2 and aspects of its applied use. *Annals of the New York Academy of Sciences*, *1125*, 308-321.

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.

Long, F., Vagin, A. A., Young, P., & Murshudov, G. N. (2008). BALBES: a molecular-replacement pipeline. *Acta crystallographica. Section D, Biological crystallography*, *64*(Pt 1), 125-132.

Lu, Y., Zhang, Y.-H. P., & Lynd, L. R. (2006). Enzyme-microbe synergy during cellulose hydrolysis by *Clostridium thermocellum*. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(44), 16165-16169.

Lykidis, A., Mavromatis, K., Ivanova, N., Anderson, I., Land, M., DiBartolo, G., Martinez, M., *et al.* (2007). Genome sequence and analysis of the soil cellulolytic actinomycete *Thermobifida fusca* YX. *Journal of Bacteriology*, *189*(6), 2477-2486.

Lynd, L. R., Cushman, J. H., Nichols, R. J., & Wyman, C. E. (1991). Fuel ethanol from cellulosic biomass. *Science (New York, N.Y.)*, *251*(4999), 1318-1323.

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.

Lytle, B. L., Volkman, B. F., Westler, W. M., & Wu, J. H. (2000). Secondary structure and calcium-induced folding of the *Clostridium thermocellum* dockerin domain determined by NMR spectroscopy. *Archives of Biochemistry and Biophysics*, *379*(2), 237-244.

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.

Matthews, B. W. (1968). Solvent content of protein crystals. *Journal of molecular biology*, *33*(2), 491-497.

Mazumder, K., & York, W. S. (2010). Structural analysis of arabinoxylans isolated from ball-milled switchgrass biomass. *Carbohydrate Research*, *345*(15), 2183-2193.

McCarter, J. D., & Withers, S. G. (1994). Mechanisms of enzymatic glycoside hydrolysis. *Current Opinion in Structural Biology*, *4*(6), 885-892.

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.

McGreal, E. P., Martinez-Pomares, L., & Gordon, S. (2004). Divergent roles for C-type lectins expressed by cells of the innate immune system. *Molecular Immunology*, *41*(11), 1109-1121.

McLean, B W, Bray, M. R., Boraston, A. B., Gilkes, N. R., Haynes, C. A., & Kilburn, D. G. (2000). Analysis of binding of the family 2a carbohydrate-binding module from *Cellulomonas fimi* xylanase 10A to cellulose: specificity and identification of functionally important amino acid residues. *Protein Engineering*, *13*(11), 801-809.

McLean, Bradley W, Boraston, A. B., Brouwer, D., Sanaie, N., Fyfe, C. A., Warren, R. A. J., Kilburn, D. G., *et al.* (2002). Carbohydrate-binding modules recognize fine substructures of cellulose. *The Journal of Biological Chemistry*, *277*(52), 50245-50254.

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.

Miller, G. L. (1959). The use of dinitrosalicylic acid reagent for the determination of reducing sugar. *Analytical Chemistry*, *31*, 426-428.

Mingardon, F., Chanal, A., López-Contreras, A. M., Dray, C., Bayer, E. A., & Fierobe, H.-P. (2007). Incorporation of fungal cellulases in bacterial minicellulosomes yields viable, synergistically acting cellulolytic complexes. *Applied and environmental microbiology*, *73*(12), 3822-3832.

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.

Montanier, C. Y., Correia, M. A. S., Flint, J. E., Zhu, Y., Baslé, A., McKee, L. S., Prates, J. A. M., *et al.* (2011). A novel, noncatalytic carbohydrate-binding module displays specificity for galactose-containing polysaccharides through calcium-mediated oligomerization. *The Journal of Biological Chemistry*, *286*(25), 22499-22509.

Moraïs, S., Barak, Y., Hadar, Y., Wilson, D. B., Shoham, Y., Lamed, R., & Bayer, E. A. (2011). Assembly of xylanases into designer cellulosomes promotes efficient hydrolysis of the xylan component of a natural recalcitrant cellulosic substrate. *mBio*, *2*(6).

Murshudov, G. N., Vagin, A. A., & Dodson, E. J. (1997). Refinement of macromolecular structures by the maximum-likelihood method. *Acta crystallographica. Section D, Biological crystallography*, *53*(Pt 3), 240-255.

Nagy, T., Tunnicliffe, R. B., Higgins, L. D., Walters, C., Gilbert, H. J., & Williamson, M. P. (2007). Characterization of a double dockerin from the cellulosome of the anaerobic fungus *Piromyces equi. Journal of Molecular Biology*, *373*(3), 612-622.

Nahálka, J., & Gemeiner, P. (2006). Thermoswitched immobilization-a novel approach in reversible immobilization. *Journal of biotechnology*, *123*(4), 478-482.

Najmudin, S., Guerreiro, C. I. P. D., Ferreira, L. M. A., Romão, M. J. C., Fontes, C. M. G. A., & Prates, J. A. M. (2005). Overexpression, purification and crystallization of the two C-terminal domains of the bifunctional cellulase *Ct*Cel9D-Cel44A from *Clostridium*

thermocellum. *Acta Crystallographica. Section F, Structural Biology and Crystallization Communications*, *61*(Pt 12), 1043-1045.

Nishiyama, Y., Johnson, G. P., French, A. D., Forsyth, V. T., & Langan, P. (2008). Neutron crystallography, molecular dynamics, and quantum mechanics studies of the nature of hydrogen bonding in cellulose Ibeta. *Biomacromolecules*, *9*(11), 3133-3140.

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.

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.

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.

Nölling, J., Breton, G., Omelchenko, M. V., Makarova, K. S., Zeng, Q., Gibson, R., Lee, H. M., *et al.* (2001). Genome sequence and comparative analysis of the solvent-producing bacterium *Clostridium acetobutylicum*. *Journal of bacteriology*, *183*(16), 4823-4838.

Nordon, R. E., Craig, S. J., & Foong, F. C. (2009). Molecular engineering of the cellulosome complex for affinity and bioenergy applications. *Biotechnology letters*, *31*(4), 465-476.

Olson, D. G., Tripathi, S. A., Giannone, R. J., Lo, J., Caiazza, N. C., Hogsett, D. A., Hettich, R. L., *et al.* (2010). Deletion of the Cel48S cellulase from *Clostridium thermocellum*. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(41), 17727-17732.

OSullivan, A. (1997). Cellulose: the structure slowly unravels. *Cellulose*, *4*(3), 173-207.

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.

Pell, G., Szabo, L., Charnock, S. J., Xie, H., Gloster, T. M., Davies, G. J., & Gilbert, H. J. (2004). Structural and biochemical analysis of *Cellvibrio japonicus* xylanase 10C: how variation in substrate-binding cleft influences the catalytic profile of family GH-10 xylanases. *The Journal of Biological Chemistry*, *279*(12), 11777-11788.

Pell, G., Williamson, M. P., Walters, C., Du, H., Gilbert, H. J., & Bolam, D. N. (2003). Importance of hydrophobic and polar residues in ligand binding in the family 15 carbohydrate-binding module from *Cellvibrio japonicus* Xyn10C. *Biochemistry*, *42*(31), 9316-9323.

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.

Pflugrath, J. W. (2004). Macromolecular cryocrystallography--methods for cooling and mounting protein crystals at cryogenic temperatures. *Methods (San Diego, Calif.)*, *34*(3), 415-423.

Pinheiro, Benedita A, Gilbert, H. J., Sakka, K., Sakka, K., Fernandes, V. O., Prates, J. A. M., Alves, V. D., *et al.* (2009). Functional insights into the role of novel type I cohesin and dockerin domains from *Clostridium thermocellum*. *The Biochemical Journal*, *424*(3), 375-384.

Pinheiro, Benedita A, Proctor, M. R., Martinez-Fleites, C., Prates, J. A. M., Money, V. A., Davies, G. J., Bayer, E. A., *et al.* (2008). The *Clostridium cellulolyticum* dockerin displays a dual binding mode for its cohesin partner. *The Journal of Biological Chemistry*, *283*(26), 18422-18430.

Pinheiro, Benedita Andrade, 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.

Pires, V. M. R., Henshaw, J. L., Prates, J. A. M., Bolam, D. N., Ferreira, L. M. A., Fontes, C. M. G. A., Henrissat, B., *et al.* (2004). The crystal structure of the family 6 carbohydrate binding module from *Cellvibrio mixtus* endoglucanase 5a in complex with oligosaccharides reveals two distinct binding sites with different ligand specificities. *The Journal of Biological Chemistry*, *279*(20), 21560-21568.

Popper, Z. A. (2008). Evolution and diversity of green plant cell walls. *Current Opinion in Plant Biology*, *11*(3), 286-292.

Popper, Z. A., Michel, G., Hervé, C., Domozych, D. S., Willats, W. G. T., Tuohy, M. G., Kloareg, B., *et al.* (2011). Evolution and diversity of plant cell walls: from algae to flowering plants. *Annual Review of Plant Biology*, *62*, 567-590.

Proctor, M. R., Taylor, E. J., Nurizzo, D., Turkenburg, J. P., Lloyd, R. M., Vardakou, M., Davies, G. J., *et al.* (2005). Tailored catalysts for plant cell-wall degradation: redesigning the exo/endo preference of *Cellvibrio japonicus* arabinanase 43A. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(8), 2697-2702.

Quiocho, F. A. (1986). Carbohydrate-binding proteins: tertiary structures and protein-sugar interactions. *Annual review of biochemistry*, *55*, 287-315.

Ragauskas, A. J., Williams, C. K., Davison, B. H., Britovsek, G., Cairney, J., Eckert, C. A., Frederick, W. J., Jr, *et al.* (2006). The path forward for biofuels and biomaterials. *Science (New York, N.Y.)*, *311*(5760), 484-489.

Raman, B., Pan, C., Hurst, G. B., Rodriguez, M., Jr, McKeown, C. K., Lankford, P. K., Samatova, N. F., *et al.* (2009). Impact of pretreated Switchgrass and biomass carbohydrates on *Clostridium thermocellum* ATCC 27405 cellulosome composition: a quantitative proteomic analysis. *PloS One*, *4*(4), e5271.

Ribeiro, T, Ponte, P. I. P., Guerreiro, C. I. P. D., Santos, H. M., Falcão, L., Freire, J. P. B., Ferreira, L. M. A., *et al.* (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.

Ribeiro, Teresa, Santos-Silva, T., Alves, V. D., Dias, F. M. V., Luís, A. S., Prates, J. A. M., Ferreira, L. M. A., *et al.* (2010). Family 42 carbohydrate-binding modules display multiple arabinoxylan-binding interfaces presenting different ligand affinities. *Biochimica Et Biophysica Acta*, *1804*(10), 2054-2062.

Riederer, A., Takasuka, T. E., Makino, S., Stevenson, D. M., Bukhman, Y. V., Elsen, N. L., & Fox, B. G. (2011). Global gene expression patterns in *Clostridium thermocellum* as determined by microarray analysis of chemostat cultures on cellulose or cellobiose. *Applied and environmental microbiology*, *77*(4), 1243-1253.

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.

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. *Journal of Bacteriology*, *187*(22), 7569-7578.

Rincón, M. T., Martin, J. C., Aurilia, V., McCrae, S. I., Rucklidge, G. J., Reid, M. D., Bayer, E. A., *et al.* (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.

Rouillard, J.-M., Lee, W., Truan, G., Gao, X., Zhou, X., & Gulari, E. (2004). Gene2Oligo: oligonucleotide design for in vitro gene synthesis. *Nucleic Acids Research*, *32*(Web Server issue), W176-180.

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., Raynaud, O., Lemaire, M., Coughlan, M., Béguin, P., & Aubert, J. P. (1994). Recognition specificity of the duplicated segments present in *Clostridium thermocellum* endoglucanase CelD and in the cellulosome-integrating protein CipA. *Journal of Bacteriology*, *176*(10), 2822-2827.

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*(1), 263-289.

Sheldrick, G. M. (1990). Phase annealing in SHELX-90: direct methods for larger structures. *Acta Crystallographica Section A Foundations of Crystallography*, *46*(6), 467-473.

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.

Shimon, L. J., Pagès, S., Belaich, A., Belaich, J. P., Bayer, E. A., Lamed, R., Shoham, Y., *et al.* (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.

Skubák, P., Murshudov, G. N., & Pannu, N. S. (2004). Direct incorporation of experimental phase information in model refinement. *Acta Crystallographica. Section D, Biological Crystallography*, *60*(Pt 12 Pt 1), 2196-2201.

Somerville, C. (2006). Cellulose Synthesis in Higher Plants. *Annual Review of Cell and Developmental Biology*, *22*(1), 53-78.

Somerville, C., Bauer, S., Brininstool, G., Facette, M., Hamann, T., Milne, J., Osborne, E., *et al.* (2004). Toward a systems approach to understanding plant cell walls. *Science (New York, N.Y.)*, *306*(5705), 2206-2211.

Sugiyama, H., Nitta, T., Horii, M., Motohashi, K., Sakai, J., Usui, T., Hisamichi, K., *et al.* (2000). The conformation of alpha-(1-->4)-linked glucose oligomers from maltose to maltoheptaose and short-chain amylose in solution. *Carbohydrate Research*, *325*(3), 177-182.

Talbott, L. D., & Ray, P. M. (1992). Molecular size and separability features of pea cell wall polysaccharides : implications for models of primary wall structure. *Plant Physiology*, *98*(1), 357-368.

Tamaru, Y., Karita, S., Ibrahim, A., Chan, H., & Doi, R. H. (2000). A large gene cluster for the *Clostridium cellulovorans* cellulosome. *Journal of bacteriology*, *182*(20), 5906-5910.

Tanaka T, Huang WC, Noguchi M, Kobayashi A, Shoda S-IT, T. (2009). Direct synthesis of 1,6-anhydro sugars from unprotected glycopyranoses by using 2-chloro-1,3-dimethylimidazoliniumchloride. *Tetrahedron Lett*, *50*, 2154–2157.

Terwilliger, T. C., Grosse-Kunstleve, R. W., Afonine, P. V., Moriarty, N. W., Zwart, P. H., Hung, L. W., Read, R. J., *et al.* (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.

The CCP4 suite: programs for protein crystallography. (1994). *Acta Crystallographica. Section D, Biological Crystallography*, *50*(Pt 5), 760-763.

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.

Tolonen, A. C., Chilaka, A. C., & Church, G. M. (2009). Targeted gene inactivation in Clostridium phytofermentans shows that cellulose degradation requires the family 9 hydrolase Cphy3367. *Molecular Microbiology*, *74*(6), 1300-1313.

Tomme, P., Boraston, A., McLean, B., Kormos, J., Creagh, A. L., Sturch, K., Gilkes, N. R., *et al.* (1998). Characterization and affinity applications of cellulose-binding domains. *Journal of Chromatography. B, Biomedical Sciences and Applications*, *715*(1), 283-296.

Tomme, P., Warren, R. A., & Gilkes, N. R. (1995). Cellulose hydrolysis by bacteria and fungi. *Advances in Microbial Physiology*, *37*, 1-81.

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.

Treiber, N., Reinert, D. J., Carpusca, I., Aktories, K., & Schulz, G. E. (2008). Structure and mode of action of a mosquitocidal holotoxin. *Journal of Molecular Biology*, *381*(1), 150-159.

Tsai, S.-L., Goyal, G., & Chen, W. (2010). Surface display of a functional minicellulosome by intracellular complementation using a synthetic yeast consortium and its application to cellulose hydrolysis and ethanol production. *Applied and environmental microbiology*, *76*(22), 7514-7520.

Urzhumtseva, L., Afonine, P. V., Adams, P. D., & Urzhumtsev, A. (2009). Crystallographic model quality at a glance. *Acta crystallographica. Section D, Biological crystallography*, *65*(Pt 3), 297-300.

Varrot, A., Schülein, M., & Davies, G. J. (1999). Structural changes of the active site tunnel of *Humicola insolens* cellobiohydrolase, Cel6A, upon oligosaccharide binding. *Biochemistry*, *38*(28), 8884-8891.

Villalobos, A., Ness, J. E., Gustafsson, C., Minshull, J., & Govindarajan, S. (2006). Gene Designer: a synthetic biology tool for constructing artificial DNA segments. *BMC Bioinformatics*, *7*, 285.

Vrsanská, M., Kolenová, K., Puchart, V., & Biely, P. (2007). Mode of action of glycoside hydrolase family 5 glucuronoxylan xylanohydrolase from *Erwinia chrysanthemi*. *The FEBS Journal*, *274*(7), 1666-1677.

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.

Wang, L., Zhang, Y., & Gao, P. (2008). A novel function for the cellulose binding module of cellobiohydrolase I. *Science in China. Series C, Life Sciences / Chinese Academy of Sciences*, *51*(7), 620-629.

Warren, R. A. J. (1996). Microbial hydrolysis of polysaccharides. *Annual Review of Microbiology*, *50*(1), 183-212.

Weiner, R. M., Taylor, L. E., 2nd, Henrissat, B., Hauser, L., Land, M., Coutinho, P. M., Rancurel, C., *et al.* (2008). Complete genome sequence of the complex carbohydrate-degrading marine bacterium, *Saccharophagus degradans* strain 2-40 T. *PLoS Genetics*, *4*(5), e1000087.

Weiss, M. (2001). Global indicators of X-ray data quality. *J. Appl. Crystallogr.*, *34*, 130-135.

Williams, S. J., A Notenboom V,, Wicki J, Rose DR, & Withers. (2000). A new, simple, high-affinity glycosidase inhibitor: analysis of binding through X-ray crystallography, mutagenesis, and kinetic analysis. *Journal of the American Chemical Society*, *122*, 4229-4230.

Winn, M D, Isupov, M. N., & Murshudov, G. N. (2001). Use of TLS parameters to model anisotropic displacements in macromolecular refinement. *Acta Crystallographica. Section D, Biological Crystallography*, *57*(Pt 1), 122-133.

Winn, Martyn D, Ballard, C. C., Cowtan, K. D., Dodson, E. J., Emsley, P., Evans, P. R., Keegan, R. M., *et al.* (2011). Overview of the CCP4 suite and current developments. *Acta crystallographica. Section D, Biological crystallography*, *67*(Pt 4), 235-242.

Wood, B. E., & Ingram, L. O. (1992). Ethanol production from cellobiose, amorphous cellulose, and crystalline cellulose by recombinant *Klebsiella oxytoca* containing chromosomally integrated *Zymomonas mobilis* genes for ethanol production and plasmids expressing thermostable cellulase genes from *Clostridium thermocellum*. *Applied and Environmental Microbiology*, *58*(7), 2103-2110.

Xie, H., Gilbert, H. J., Charnock, S. J., Davies, G. J., Williamson, M. P., Simpson, P. J., Raghothama, S., *et al.* (2001). *Clostridium thermocellum* Xyn10B carbohydrate-binding module 22-2: the role of conserved amino acids in ligand binding. *Biochemistry*, *40*(31), 9167-9176.

Xu, J., Bjursell, M. K., Himrod, J., Deng, S., Carmichael, L. K., Chiang, H. C., Hooper, L. V., *et al.* (2003). A genomic view of the human-*Bacteroides thetaiotaomicron* symbiosis. *Science (New York, N.Y.)*, *299*(5615), 2074-2076.

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.

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.

Yaoi, K., Kondo, H., Hiyoshi, A., Noro, N., Sugimoto, H., Tsuda, S., Mitsuishi, Y., *et al.* (2007). The structural basis for the exo-mode of action in GH74 oligoxyloglucan reducing end-specific cellobiohydrolase. *Journal of Molecular Biology*, *370*(1), 53-62.

Yaron, S., Morag, E., Bayer, E. A., Lamed, R., & Shoham, Y. (1995). Expression, purification and subunit-binding properties of cohesins 2 and 3 of the *Clostridium thermocellum* cellulosome. *FEBS letters*, *360*(2), 121-124.

Zhang, K. Y., Cowtan, K., & Main, P. (1997). Combining constraints for electron-density modification. *Methods in enzymology*, *277*, 53-64.

Zhang, X.-Z., Sathitsuksanoh, N., & Zhang, Y.-H. P. (2010). Glycoside hydrolase family 9 processive endoglucanase from *Clostridium phytofermentans:* heterologous expression, characterization, and synergy with family 48 cellobiohydrolase. *Bioresource Technology*, *101*(14), 5534-5538.

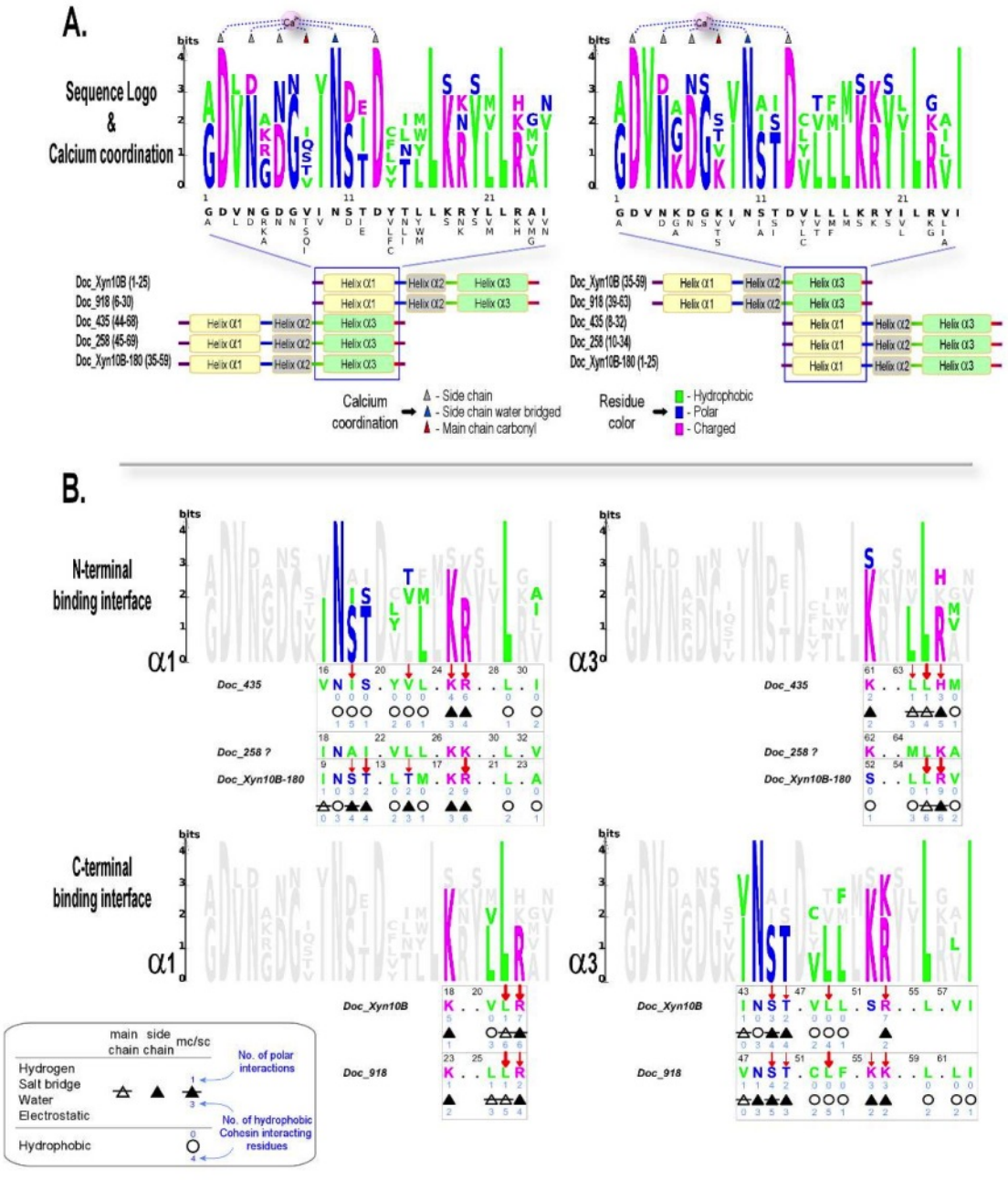
Zhang, Y.-H. P., & Lynd, L. R. (2005). Cellulose utilization by *Clostridium thermocellum*: bioenergetics and hydrolysis product assimilation. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(20), 7321-7325.

Zverlov, V. V., Schantz, N., & Schwarz, W. H. (2005). A major new component in the cellulosome of *Clostridium thermocellum* is a processive endo-beta-1,4-glucanase producing cellotetraose. *FEMS Microbiology Letters*, *249*(2), 353-358.

ANNEXES

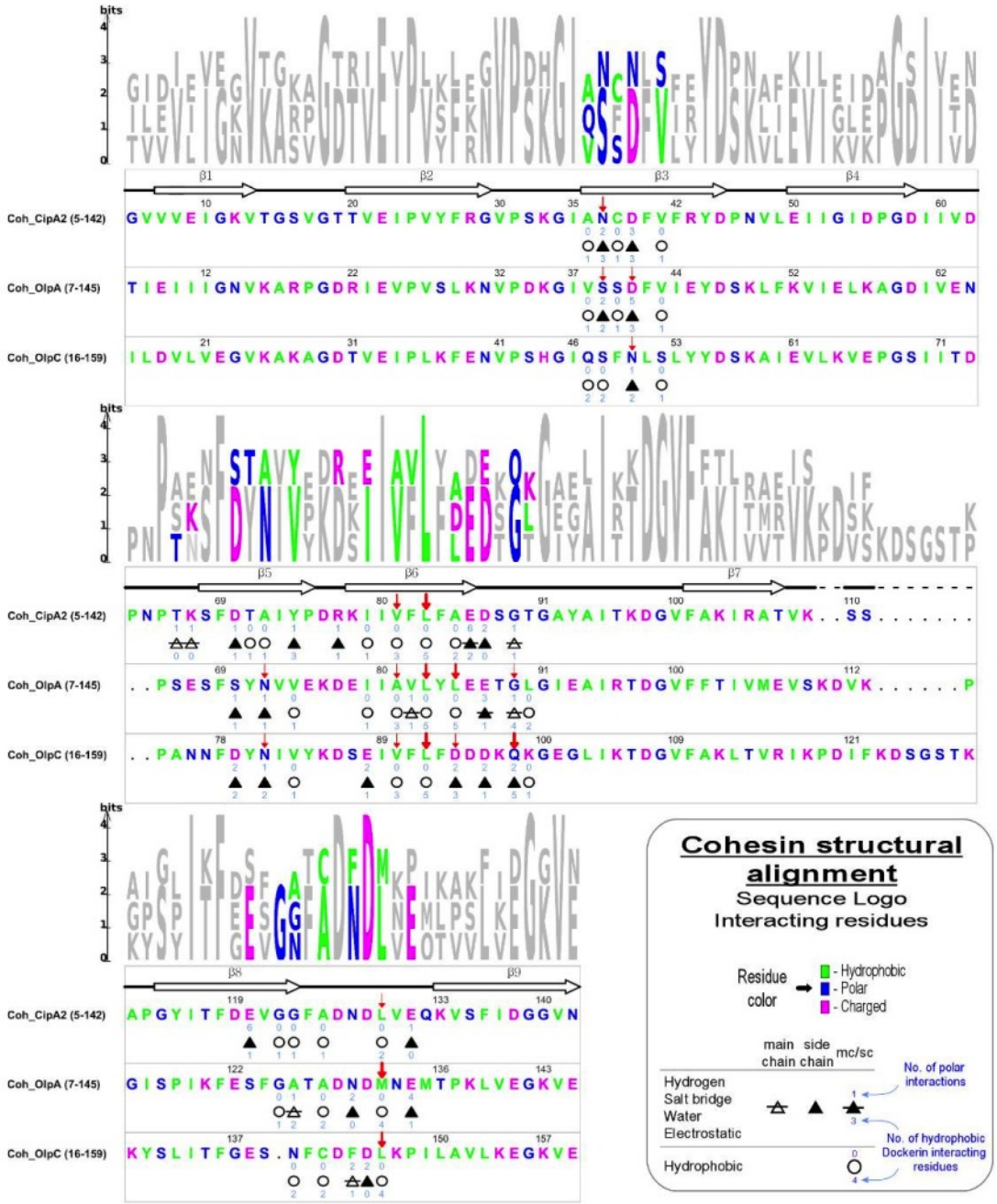
Supplemental information – Chapter 3

Figure S3.1| Dockerin sequence alignment and interacting residues.



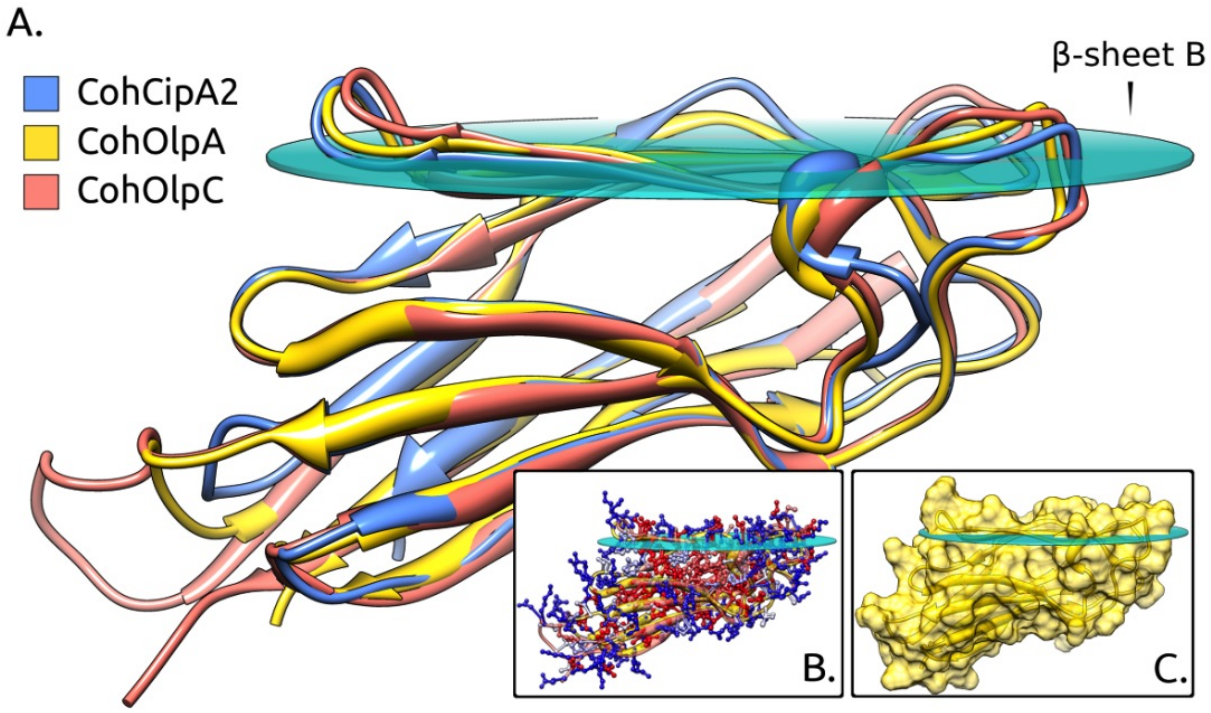
A) Sequence Logo and dockerin residues involved in Calcium coordination. **B)** Dockerin residues involved in N-terminal and C-terminal binding interfaces. Residues with a significant contribution to the Coh-Doc contact surface area are identified (top variable-width small red arrow) and show a quantification of polar and hydrophobic interactions as per legend. Doc435 = Doc124A.

Figure S3.2| Cohesin sequence alignment and interacting residues.



Cohesin residues with a significant contribution to the Coh-Doc contact surface area are marked with a top variable-width small red arrow. A quantification of the polar and hydrophobic interactions is shown according to the legend.

Figure S3.3| Cohesins structure superposition.



A) Structure superposition of cohesin domains from CipA2, OlpA and OlpC. The cyan transparent circle marks the plane defined by β-sheet B, whose β-strands form a distinctive dockerin interacting plateau. **B)** Hydrophobic core of CohOlpA and CohOlpC, with residues color rendered according to the hydrophobicity scale of Kyte and Doolittle, from red (+4.5, Ile) to blue (-4.5, Arg). **C)** Overall curved cylindrical shape of CohOlpA, defined by its molecular surface with the location of the β-sheet B plane.

Supplemental information – Chapter 4

Table S4. 1| CipB XDoc interface hydrogen bonds and salt bridges

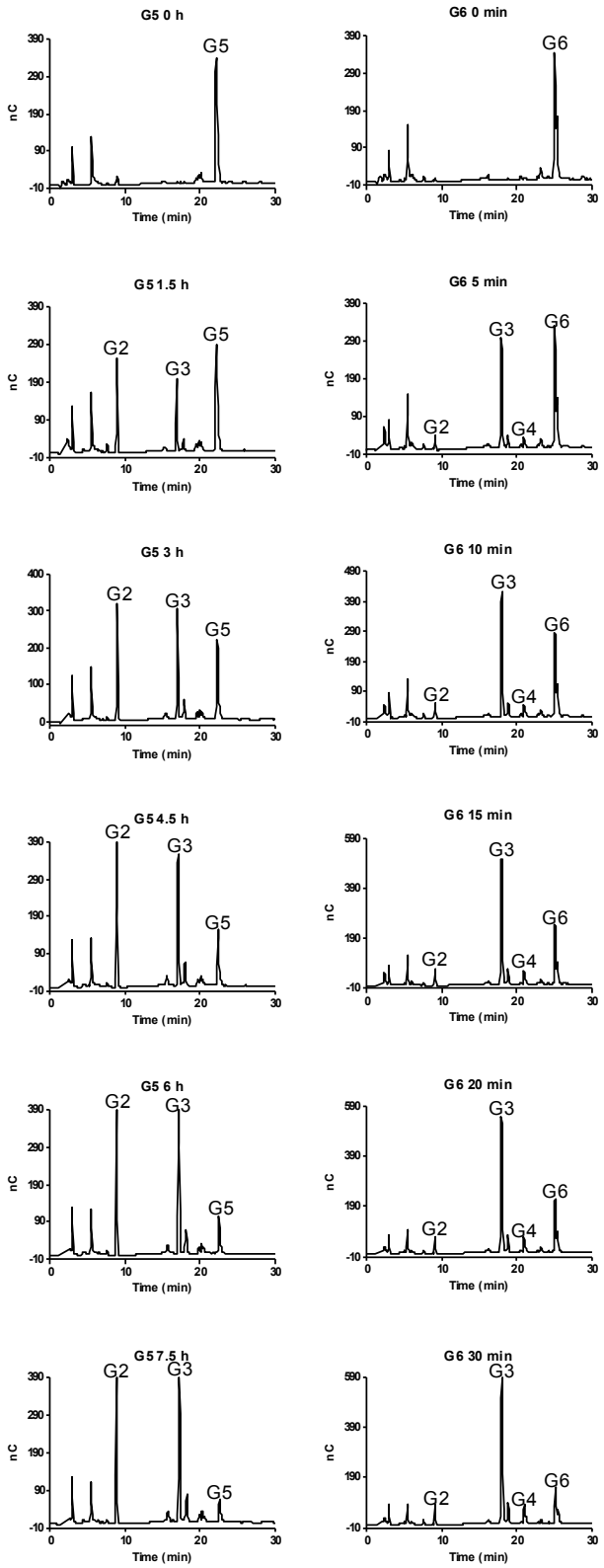
Direct hydrogen bonds					
#	**Module X**		**Distance (Å)**	**Dockerin**	
	Residue	Atom		Residue	Atom
1	Ser25	Oγ	2.62	Met141	O
2	Lys70	O	3.76	Gln108	Nε1
3	Arg71	Nε	2.98	Glu103	Oε1
4	Asn72	O	2.88	Glu103	Oε1
5	Asn72	O	3.14	Asp101	N
6	Asn72	O	2.76	Gly100	N
7	Asn72	Oδ1	3.00	Glu103	N
8	Asn72	Nδ2	3.05	Glu136	O
9	Tyr73	N	3.83	Asp101	O
10	Tyr73	N	3.85	Gln108	Oε1

Water-mediated hydrogen bonds								
#	**Module X**		**Distance (Å)**	**H_2O**		**Distance (Å)**	**Dockerin**	
	Residue	Atom		Residue	Atom		Residue	Atom
1	Asp18	Oδ1	2.81	H_2O270	O	3.13	Ala99	N
2	Lys70	O	3.58	H_2O282	O	2.74	Gln108	Nε2
3	Lys70	O	3.58	H_2O282	O	3.51	Gln108	Oε1
4	Asn72	Oδ1	2.92	H_2O244	O	2.14	Glu103	O

Salt bridges					
#	**Module X**		**Distance (Å)**	**Dockerin**	
	Residue	Atom		Residue	Atom
1	Asp18	Oδ1	2.96	His156	Nε2
2	Asp18	Oδ2	3.61	His156	Nδ1
3	Asp18	Oδ2	3.34	His156	Nε2
4	Arg71	Nε	2.98	Glu103	Oε1
5	Arg71	Nη1	3.40	Glu103	Oε1

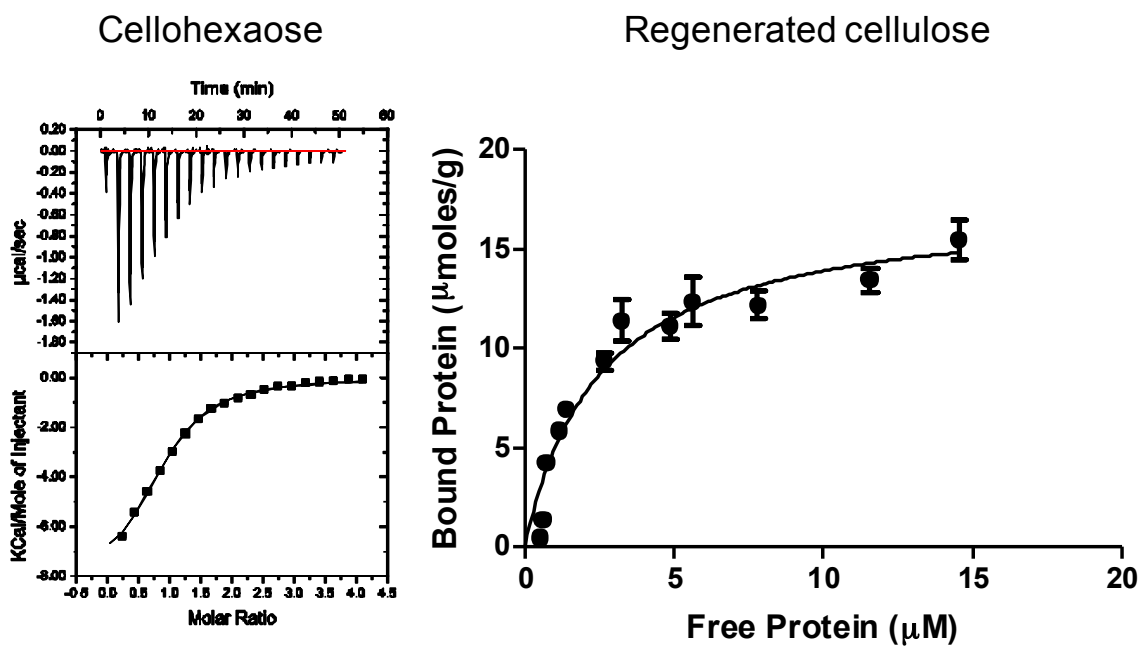
Supplemental information – Chapter 5

Figure S5.1| HPLC analysis of cellopentaose and cellohexaose hydrolysis by *Ct*Cel124



Cellopentaose (G5) and cellohexaose (G6), both at 15 μM, were incubated with 10 μM and 3 μM of *Ct*Cel124, respectively, in MES buffer pH 5.5 at 60 ^{o}C and the rate of hydrolysis and the nature of the reaction products were assessed by HPLC. G2, cellobiose; G3, cellotriose; G4, cellotetraose.

Figure S5.2| Binding of the *Ct*Cel124 derivative E96A to regenerated cellulose and cellohexaose



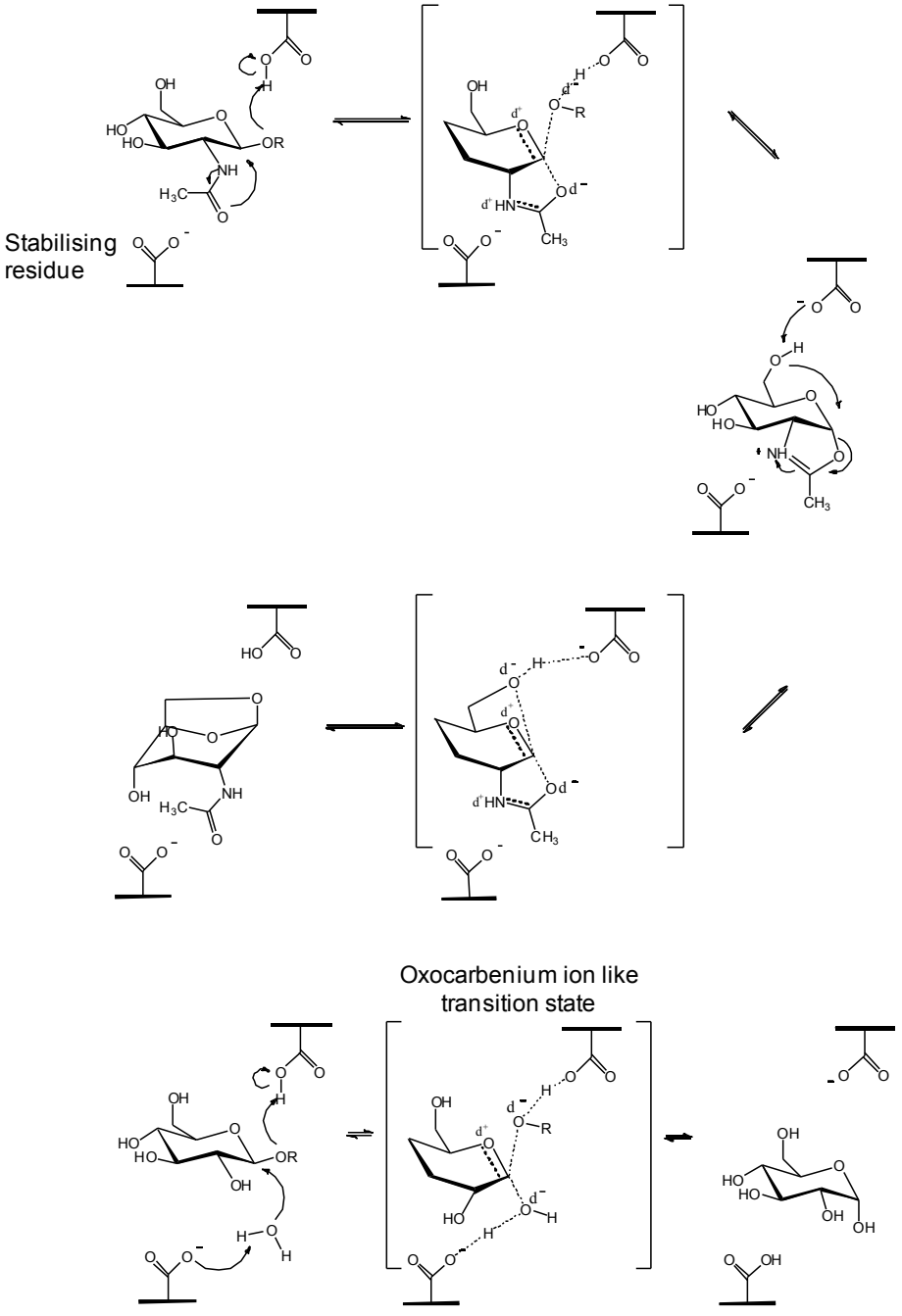
A) Typical ITC data. The top half of the panel shows the raw ITC heats; the bottom half, the integrated peak areas fitted using a one single model by MicroCal Origin software. During the titration experiment, the protein at 88 µM, stirred at 1000 rpm in a 200 µl reaction cell, was injected with 20 successive 2 µl aliquots of 2 mM cellohexaose.
B) Depletion isotherms carried out in 50 mM MES buffer, pH 5.5, at 40 ^{o}C. Protein was added to 1 mg of RC in a final volume of 1 ml and incubated for 1 h with gentle mixing at 40 ^{o}C. The data were analyzed by nonlinear regression using a standard one-site binding model (GraphPad Prism, v2.01), and the N_0 and K_a values were obtained from the regressed isotherm data.

Figure S5.3| Alignment of GH124 sequences

```
CtCel124          DGYFPPGTSKHELIARASSLKVSEVKAIIKKQVDEHWDVIRDVCGFKNKEVAYAFFFGMA 60
EFL97259          ----PANASTAEIRAINHNMTVQEVKYALYKEIDEHWDLISTRLGQPEREKVYALFLGLG 56
ZP_06145404       ----PADASTEEIRAINHKMTVQEVKYALYKEIDEHWDLISTRLGQPDREKMYALFLGLG 56
EFL63545          DRYFSEGTSKQVMIQKASELSASQVKALIKQHVDEHYDLLAQQFGFTSKDDVYAFCFGWA 60
                            *                ** *** *      *       **   *

CtCel124          TRESTFR-----AATET------GSGASHAFGPLQTAETAYANANPNYMPEHNVPEMHQY 109
EFL97259          TRESTLGEGKIGSDHETAYEDGWGVNSAHAYGTLQTAVTAFADCNPKFMPEDNVPEMFQY 116
ZP_06145404       SRESTLGEGKIGSDHETAYEDGWGVNSAHAYGTLQTAVTAFKDCDPKFMPEDNVPEMFQY 116
EFL63545          TRESTLN-----AELETAIA-GWGENSAHAYGPFQTAETAFKNNYKDFMTEYDVPEMTQF 114
                   ****          **      *    ** *  *** **        * *  **** *

CtCel124          DFTEYNFYDVGISVHMGIRHFLHFARL-AKEK-YSGRDIARHGLMGYNTGWIDGADES-W 166
EFL97259          SFTESNFYDCIISNHMGIRKILHFAEICINEEGMHGYQVIRNSLKGFNTGWCYMAEDDGA 176
ZP_06145404       SFTPSNFYDCIISNHMGIRKILHFAEICINEEGMHGYQVVRNSLKGFNTGWCYMAEDDGA 176
EFL63545          DLNESNFYDVGISIHMGIRKLVHFSKD-AKAKGYSGKDVLRRSVLAFNTGWVEGASES-W 172
                       ****  ** *****   **           *    *      ****   *

CtCel124          IVRYADETAALGAWYLRNNHMSDDEFTWDTDPRVDRSNPWEIYY 210
EFL97259          YKTYADEICSMAQFYYNEGHLYDNVFTWTSAGGNMEKYRTADRW 224
ZP_06145404       YATYADEICAMAQFYYNEGHLYDNVFTWTSANGEKINKYRTANR 224
EFL63545          INEYSDEIASLAGWYL-NGHLSDDQWTWTGDSRVDRSNPWSWY- 214
                     * **       *    *  *   **
```

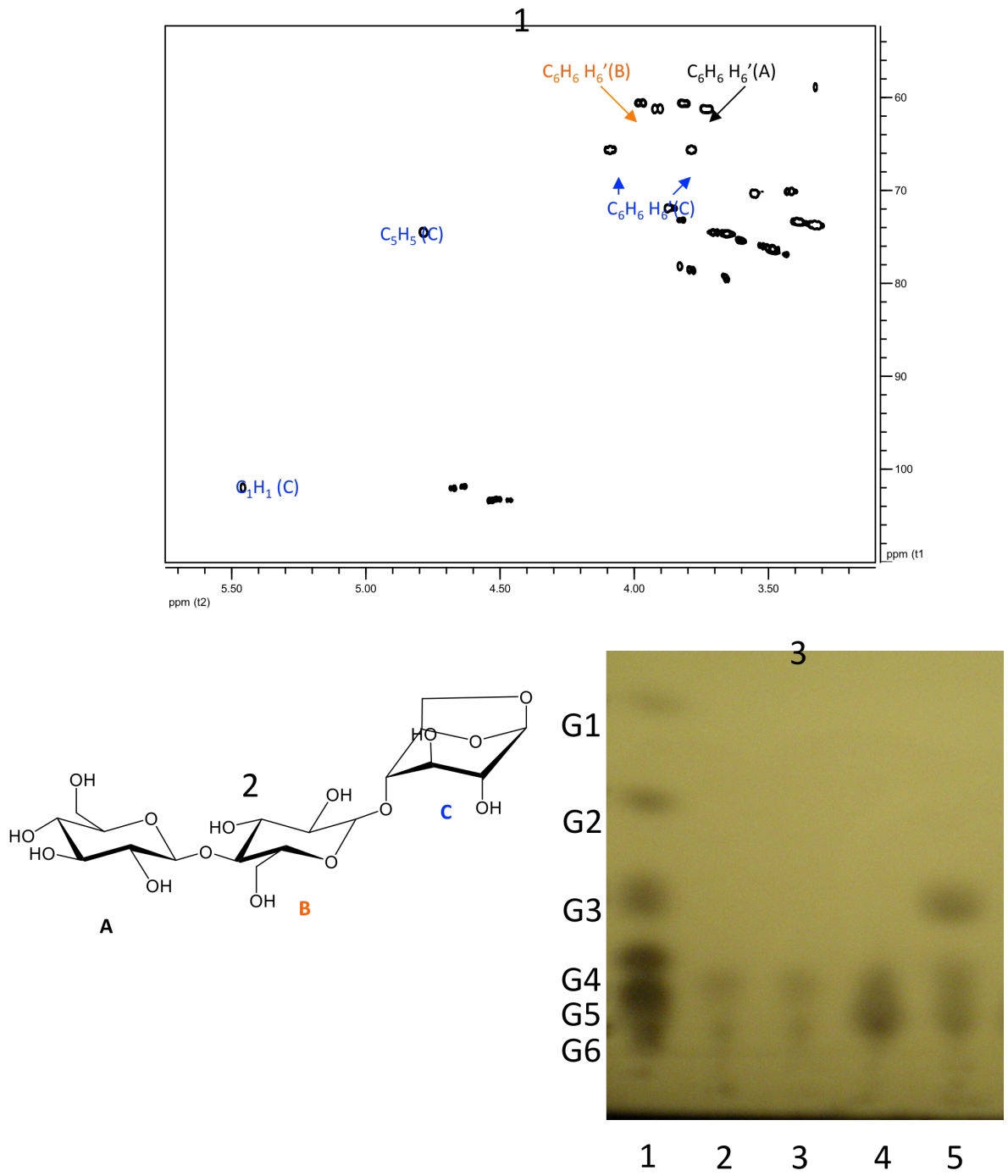
The sequence alignment was derived from a search of the UNIPROT dataset using *Ct*Cel124$_{CD}$ as the query sequence and the BLASTw search engine. All proteins have a value of $<e^{-37}$. Residues that are invariant within the family are indicated by an asterisk.

Figure S5.4| Schematic of the acid base single displacement and lytic transglycosylase mechanisms.



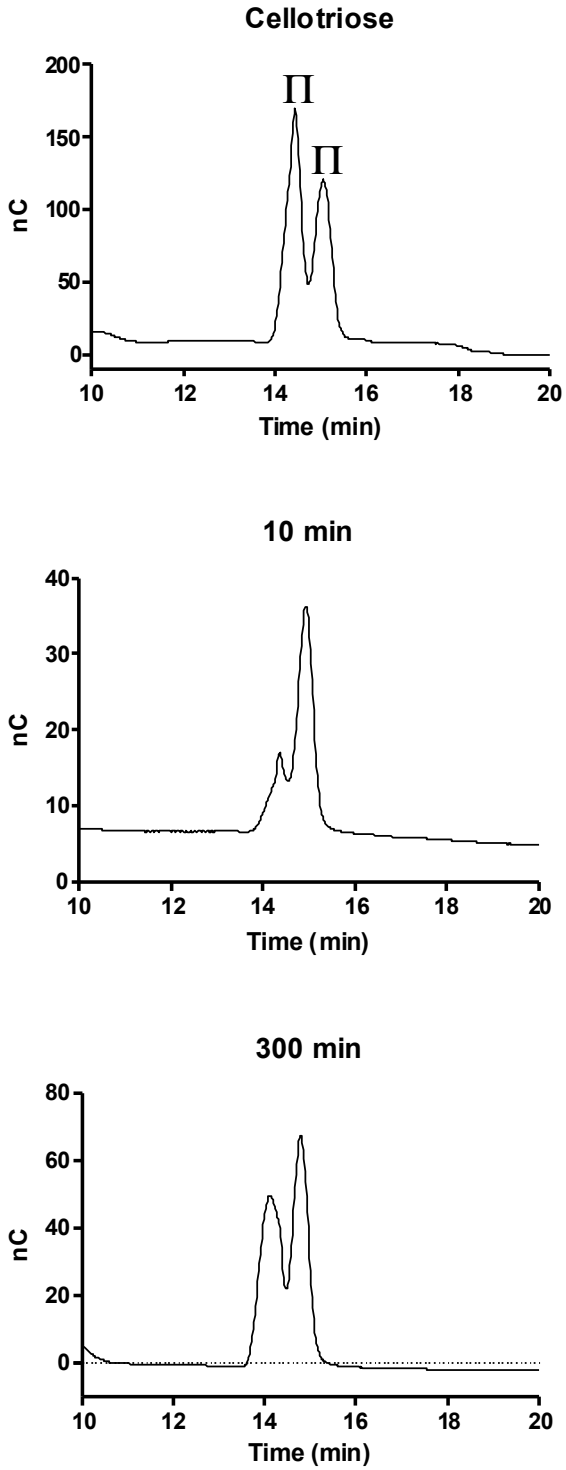
Schematic of the inverting mechanism.

Figure S5.5| Synthesis and use of 1,6-anhydro cellotriose.



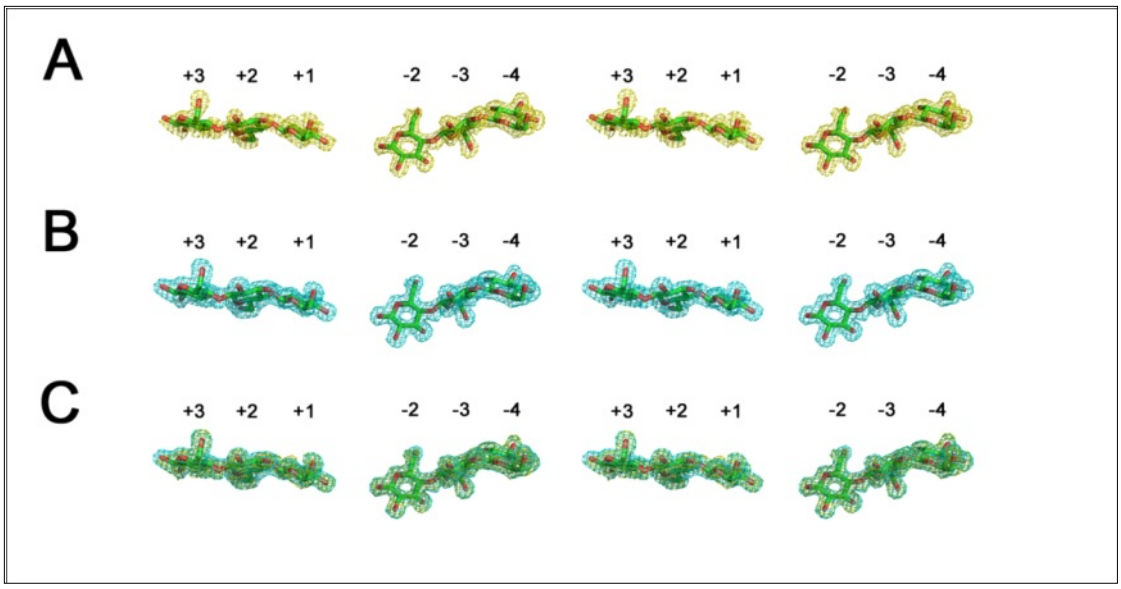
Panel 1 shows that gHSQC spectrum of 1,6 - anhydro cellotriose synthesized by the method of Tanaka *et al.*, (2009). Black and orange arrows identify signals from β-1,4 glucose and blue identifies the presence of 1,6 anhydro glucose (see *Panel 2* for sugar identification). *Panel 3* shows TLC of oligosaccharides incubated with *Ct*Cel124. Lane 1, standards; lane 2, 1,6 anhydrocellotriose; lane 3, 1,6 anhydrocellotriose incubated with *Ct*Cel124 for 4 h; lane 4, cellohexaose; lane 5, cellohexaose incubated with *Ct*Cel124 for 4 h. Cellohexaose and 1,6 anhydrocellotriose, at 1 mM, were incubated with 7 μM *Ct*Cel124 in 50 mM MES buffer, pH 5.5, containing 20 mM NaCl, at 60 °C. G1 = Glucose; G2 = Cellobiose; G3 = Cellotriose; G4 = Cellotetraose; G5 = Cellopentaose and G6 = Cellohexaose.

Figure S5.6| Anomeric configuration of cellotriose generated by *Ct*Cel124.



Cellopentaose (30 μM) was incubated with 1 μM *Ct*Cel124 for 10 min and 300 min, and the products were separated by HPLC. The retention time of the β and α anomer of cellotriose was 14.2 min and 15 min, respectively.

Figure S5.7| Conformation of cellotriose bound to *Ct*Cel124



The figure provides a wall-eyed stereo view of the cellohexaose cleaved product complex; **A)** displays the weighted maximum-likelihood F_{obs} - F_{calc} (yellow; produced from phases generated by refinements with the cellooligosaccharides omitted), **B)** shows the $2F_{obs}$ - F_{calc} maps (cyan; produced from phases generated by refinements with the cellooligosaccharides included). Both electron density maps are contoured to 1.0 σ (0.43 $e^-/Å^3$). **C)** depicts an overlay of the two maps contoured as in **A)** and **B)**.

Table S5.1| Binding of the *Ct*Cel124$_{CD}$ mutant E96A to cellohexaose and regenerated cellulose (RC)

	Isothermal titration calorimetry						Depletion binding isotherms	
Ligand	**Temperature (K)**	**K_A (M^{-1})**	**ΔG (kcal mol^{-1})**	**ΔH (kcal mol^{-1})**	**$T\Delta S$ (kcal mol^{-1})**	**N**	**K_A (M^{-1})**	**N_0 (µmol g^{-1})**
RC	303	2.5 (±0.3) x 10^5	-7.5	-20.0 (±3.3)	-12.5 (±3.3)	1.04 ± 0.04	-	-
RC	313	1.4 (±0.4) x 10^5	-6.9	-40.2 (±6.8)	-33.3 (±7.5)	1.02 ± 0.05	3.9 (±0.5) x 10^5	17.4 (±1.0)
RC	323	1.0 (±0.4) x 10^5	-7.5	-83.1(±5.5)	-75.6 (±31)	0.96 ± 0.02	-	-
RC	333	1.0 (±0.5) x 10^5	-7.6	-79.5 (±5.4)	-71.9 (±9.1)	0.99 ± 0.01	-	-
Cellohexaose	283	5.5 (±0.4) x 10^4	-6.1	-7.4 (±0.3)	-1.3 (±0.3)	0.93 ± 0.03	-	-
Cellohexaose	293	4.2 (±0.1) x 10^4	-6.2	-8.2 (±0.1)	-2.0 (±0.4)	0.91 ± 0.02	-	-
Cellohexaose	303	3.1 (±0.1) x 10^4	-6.2	-8.1 (±0.2)	-1.9 (±0.2)	1.01 ± 0.04	-	-
Cellohexaose	313	1.5 (±0.1) x 10^4	-6.0	-8.9 (±0.7)	-2.9 (±0.7)	1.07 ± 0.11	-	-

Cellohexaose (2 mM) was titrated into E96A (88.2 µM). Conversely E96A (835 µM) was titrated into RC (29.4 mg/ml). Depletion isotherms were performed by using a concentration of 1 mg/ml and 1–30 µM E96A. All reactions were carried out in 50 mM MES pH 5.5, containing 20 mM NaCl at the temperature indicated.

Table S5.2| Primers used to generate DNA constructs encoding native and variants of *Ct*Cel124$_{CD}$.

Clone	Sequence (5'→3')	Direction
Cloning cel124$_{cd}$	ctc gct agc cct gca aat aca caa tcc	Forward
	cac ctc gag tta gta ata aat ctc cc	Reverse
E96A	gga atg gct acc aga gcc tcc act ttt aga gct	Forward
	agc tct aaa agt gga ggc tct ggt agc cat tcc	Reverse
N188A	ggc ttg atg gga tac gcg aca ggt tgg att gac ggt gcg g	Forward
	c cgc acc gtc aat cca acc tgt cgc gta tcc cat caa gcc	Reverse
N188D	ggc ttg atg gga tac gac aca ggt tgg att gac ggt gcg g	Forward
	c cgc acc gtc aat cca acc tgt gtc gta tcc cat caa gcc	Reverse
N188E	ggc ttg atg gga tac gag aca ggt tgg att	Forward
	aat cca acc tgt ctc gta tcc cat caa gcc	Reverse
S110A	acc gga agc ggg gct gcc cac gct ttt ggc cct	Forward
	agg gcc aaa agc gtg ggc agc ccc gct tcc ggt	Reverse
S110D	acc gga agc ggg gct gat cac gct ttt ggc cct	Forward
	agg gcc aaa agc gtg atc agc ccc gct tcc ggt	Reverse
S110E	acc gga agc ggg gct gag cac gct ttt ggc cct	Forward
	agg gcc aaa agc gtg ctc agc ccc gct tcc ggt	Reverse
E96A	gga atg gct acc aga gcc tcc act ttt aga gct	Forward
	agc tct aaa agt gga ggc tct ggt agc cat tcc	Reverse
N188A	ggc ttg atg gga tac gcg aca ggt tgg att gac ggt gcg g	Forward
	c cgc acc gtc aat cca acc tgt cgc gta tcc cat caa gcc	Reverse

Table S5.3| Crystal data and refinement Statistics for *Ct*Cel124

	***Ct*Cel124$_{CD}$-cellohexaose**	***Ct*Cel124$_{CD}$-cellotetraose**
Data collection		
Beamline	ESRF, ID14-EH4	Diamond light source, I02
Cell parameters		
a (Å)=*b* (Å)	74.1	73.9
c (Å)	74.8	74.8
Space Group	$P3_221$	$P3_221$
Wavelength, Å	0.954	0.98
Resolution of data (outer shell), Å	26.20 – 1.5	48.64-1.06
R_{pim} (outer shell), % [a]	4.1 (15.7)	1.6 (15)
R_{sym} (outer shell), % [b]	7.3 (26.5)	6.2 (48)
Mean I/σ(I) (outer shell)	18.7 (5.0)	21.4 (5.2)
Completeness (outer shell), %	96.3 (98.2)	98.4 (96.5)
Anomalous Completeness (outer shell), %	96.4 (97.0)	-
Redundancy	7.3 (6.9)	12.1 (10.7)
Anomalous Redundancy	3.8 (3.6)	-
FOM for 6 Se-sites (before / after solvent flattening)	0.32 / 0.91	-
Structure refinement		
No. of protein atoms	1709	1801
No. of solvent waters	272	389
Resolution used in refinement, Å	24.30 – 1.5	48.64-1.06
No. of reflections	20516	1270736
R_{work} / R_{free} (%) [c]	15.0 / 19.7	12.0/14.0
rms deviation 1-2 bonds (Å)	0.026	0.024
rms deviation 1-3 bonds (degrees)	2.059	2.090
rms deviation chiral volume (Å^3)	0.159	0.15
Avg B factors (Å^2)		
main-chain	13.8	10.3
side-chain	16.6	12.9
Mg^{2+}	12.3	7.3
CT3 (1)	10.6	8.1
CT3 (2)	15.0	10.9
water molecules	28.0	24.4

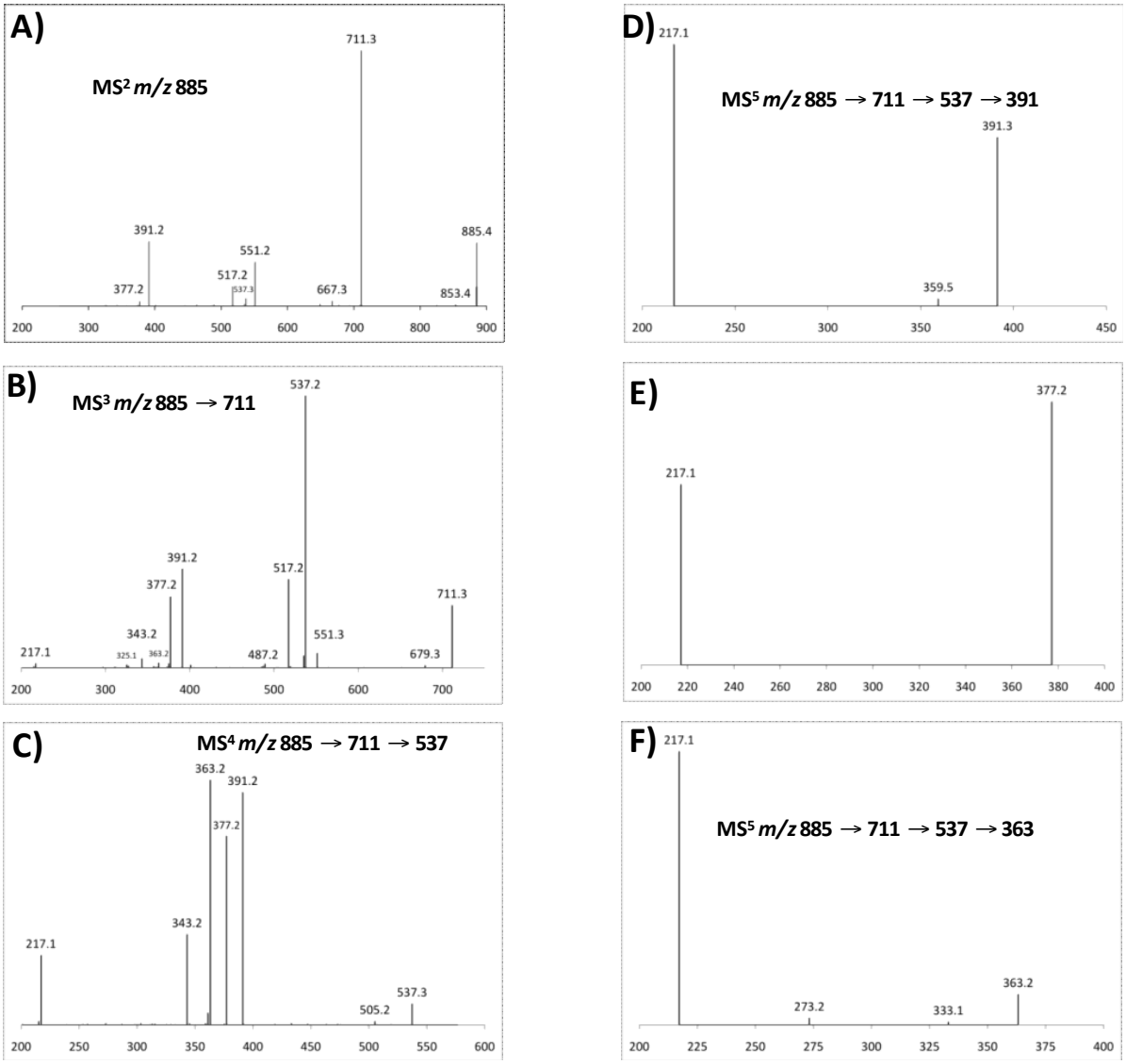
[a] $R_{p.i.m.} = \left(\sum_{hkl}\sqrt{\frac{n}{n-1}}\sum_{j=1}^{n}\left|I_{hkl,j} - \langle I_{hkl}\rangle\right|\right) \Big/ \left(\sum_{hkl}\ \sum_{j} I_{hkl,j}\right)$, where $\langle I_{hkl}\rangle$ is the average of symmetry related observations of a unique reflection.

[b] $R_{sym} = \left(\sum_{hkl}\sum_{j}\left|I_{hkl,j} - \langle I_{hkl}\rangle\right|\right) \Big/ \left(\sum_{hkl}\ \sum_{j} I_{hkl,j}\right)$, where $\langle I_{hkl}\rangle$ is the average of symmetry-related observations of a unique reflection.

[c] $R_{work} = \left(\sum_{hkl}\left|F_{hkl}^{obs} - F_{hkl}^{calc}\right|\right) \Big/ \left(\sum_{hkl} F_{hkl}^{obs}\right) \times 100$, where F^{calc} and F^{obs} are the calculated and observed structure factor amplitudes, respectively. R_{free} is calculated for a randomly chosen 10% of the reflections.

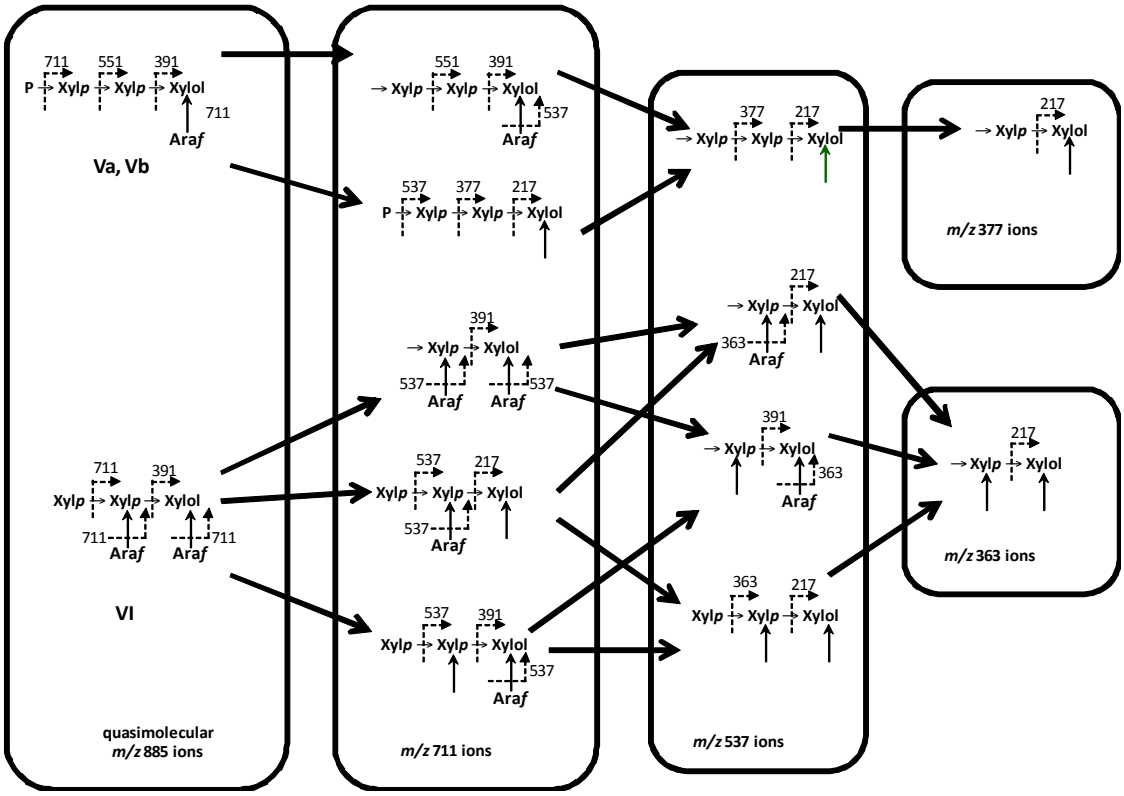
Supplemental information – Chapter 6

Figure S6.1| ESI-MS of the pentasaccharides in Fraction 1.



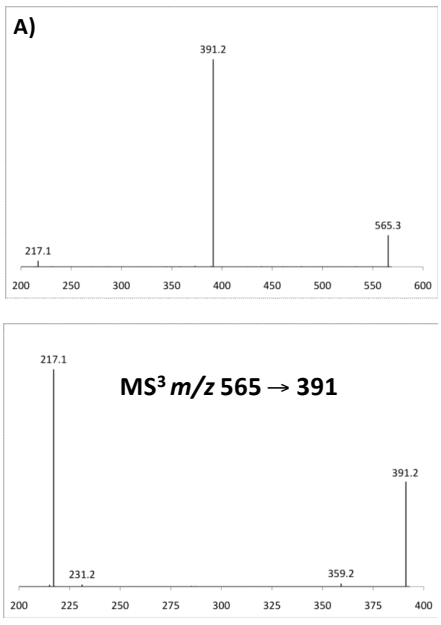
The pentasaccharides in Fraction 1 were analyzed by ESI-MS^n. **A)** Fragmentation of the *m/z* 725 ion which comprises the pentasaccharides. **B)**, **C)**, **D)**, **E)** and **F)** show the fragmentation patterns of the 711 (MS^3), 537 (MS^4), 391 (MS^5), 377 (MS^5) and 363 (MS^5) ions, respectively. The masses of Y-ions are indicated unless otherwise stated.

Figure S6.2| The structure of the pentasaccharides generated by *Ct*Xyl5A.



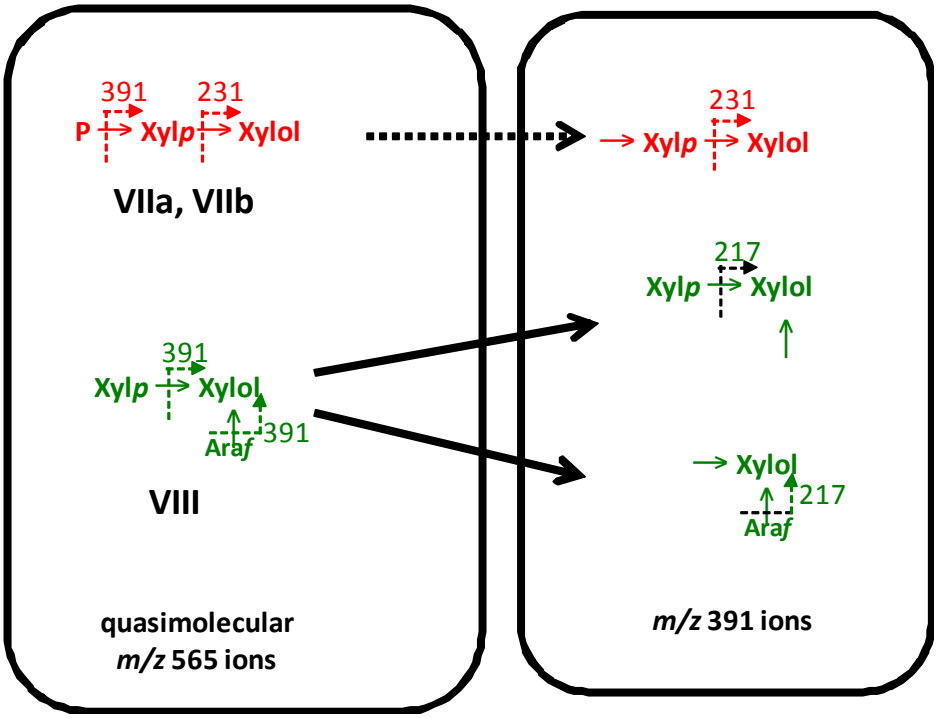
Based on the data displayed in Figure S6.1., the structures of the pentasaccharides in Fraction 1 were identified. The sugars labelled P can be Ara*f* or Xyl*p*. The dotted arrow between sugar linkages shows the fragmentation site and the ion identified. Arrows pointing at sugars (but did not link two sugars together) identified hydroxyl groups that were not methylated as they comprised a glycosidic linkage in a parental ion. Xylol is the reducing end xylose that has been reduced to its alditol form by $NaBH_4$.

Figure S6.3| ESI-MS of the trisaccharides in Fraction 1.



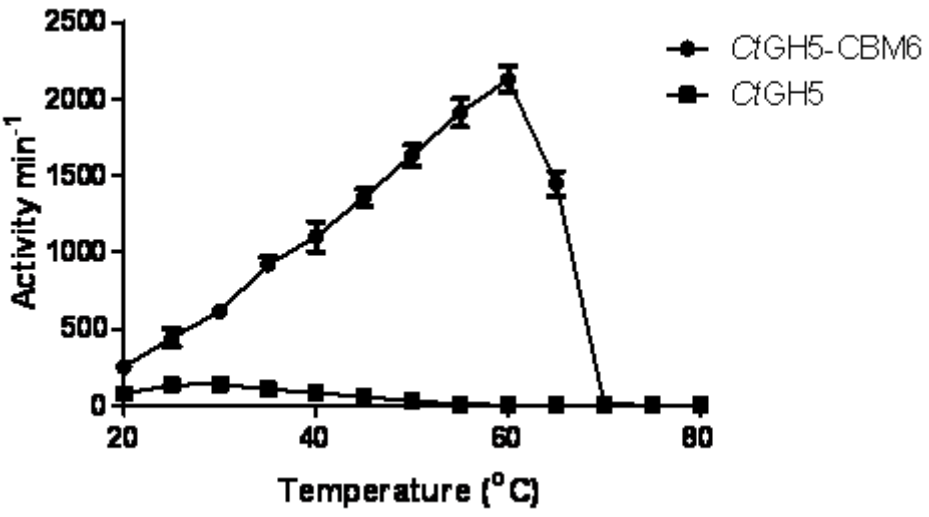
The trisaccharides in Fraction 1 were analyzed by ESI-MS^n. **A)** Fragmentation of the *m/z* 565 ion (MS^2), which comprises the trisaccharides. **B)** Fragmentation pattern of the *m/z* 391 ion (MS^3) generated from the *m/z* 565 ion.

Figure S6.4| The structure of the trisaccharides generated by *Ct*Xyl5A.



Based on the data displayed in Figure S6.3, the structures of the trisaccharides in Fraction 1 were identified. The sugars labelled P can be Ara*f* or Xyl*p*. The data showed that the oligosaccharide ions coloured green were present, while those coloured red were not evident. The solid arrows between oligosaccharides showed the conversion of one oligosaccharide into another, through ESI-MS fragmentation. Dotted arrows indentified theoretical MS-mediated oligosaccharide conversions that did not occur in these analyses. The dotted arrows between sugar linkages in the oligosaccharides show the fragmentation site and the ion identified. Arrows pointing at sugars (but did not link two sugars together) within oligosaccharides identified hydroxyl groups that were not methylated as they comprised a glycosidic linkage in a parental ion. Xylol is the reducing end xylose that has been reduced to its alditol form by $NaBH_4$.

Figure S6.5| Temperature optimum of *Ct*GH5 and *Ct*GH5-CBM6.



The two enzymes were assayed in 50 mM sodium phosphate buffer, pH 7.0, containing 50 nM enzyme and 1 mg/ml wheat arabinoxylan. The reactions were monitored for 10 min by assaying for reducing sugar release.

Table S6.1| Primers used for cloning components of *Ct*Xyl5A and for constructing mutants.

Clone	Sequence (5'-3')	Direction
*Ct*GH5	CTCGCTAGCAGCCCGCAACGTGGCCGG	Forward
	CACCTCGAGGTCATAATACGAAACCCC	Reverse
*Ct*GH5-CBM6	CTCGCTAGCCCGCAACGTGGCCGG	Forward
	CACCTCGAGTATCGGAGAAAGTTC	Reverse
*Ct*CBM6	CTCGCTAGCACGGATTCGGTGAATG	Forward
	CACCTCGAGTATCGGAGAAAGTTC	Reverse
E171	TTGTATGAAATACACAATGCGCCTGTGGCATGGGGA	Forward
	TCCCCATGCCACAGGCGCATTGTGTATTTCATACAA	Reverse
E279	CCTGCTTTATGACTGCGTATGCCGGAGGTGC	Forward
	GCACCTCCGGCATACGCAGTCATAAAGCAGG	Reverse
W242A	GGCGGTTACAATGTCGGAGCGATTTCGGAAGGAGAATC	Forward
	CATTCTCCTTCCGAAATCGCTCCGACATTGTAACCGCC	Reverse
W242A	CTGCCTGCTACCGGAGGTGCTCAGACTTGGATACA	Forward
	TGTAGTCCAAGTCTGAGCACCTCCGGTAGCAGGCAG	Reverse

Table S6.2| Crystal and structure resolution statistics.

	***Ct*GH5-CBM6 SeMet**	***Ct*GH5-CBM6 *Wt* native**
Data collection		
Space group	$P2_12_12_1$	$P2_12_12_1$
Cell dimensions		
a, *b*, *c* (Å)	69.26, 75.82, 106.11	69.12, 75.55, 105.97
Resolution (Å)	61.66-2.20 (2.32-2.20)	36.74-1.47 (1.55-1.47)
R_{merge}	0.098(0.372)	0.081 (0.639)
$I / \sigma I$	16.9 (5.8)	13.9 (2.8)
Completeness (%)	99.0 (98.3)	100 (99.9)
Redundancy	7.2 (7.4)	7.0 (6.6)
Anomalous completeness (%)	99.2 (98.5)	
Anomalous redundancy	3.8 (3.8)	
Refinement		
No. reflections	-	90100 (4744)
R_{work} / R_{free}	-	0.146 / 0.166
No. Atoms		
Protein	-	3782
Ligand / Ion	-	41
Water	-	636
B-factor		
Protein	-	16
Ligand / Ion	-	32
Water	-	33
R.m.s.[b] deviations		
Bond lengths (Å)		0.014
Bond angles (°)		1.40

Table S6.3| 1H and 13C chemical shifts of the NMR resonances of Fraction 1.

Sugar residues	H-1/C-1	H-2/C-2	H-3/C-3	H-4/C-4	H-5/C-5[a]	H-5'/C-5[a]
β-D-Xylp	4.618	3.392	3.736	3.809	4.073	3.387
(reducing)	96.6	74.8	77.8	73.8	63.0	
α-D-Xylp	5.169	3.680	3.906	nd	3.840	3.770
(reducing)	92.4	71.8	77.7		59.4	
α-L-Araf (α)	5.342	4.164	3.911	4.273	3.799	3.718
	107.9	80.8		84.9	61.5	
α-L-Araf (β)	5.391	4.169	3.911	4.283	3.799	3.718
	107.9	80.8		84.9	61.5	
α-L-Araf	5.327	4.181	3.961	4.182	3.815	3.706
(nonreducing)	108.4	80.8		84.1	61.5	
3-linked-β-D-Xylp	4.465	3.401	nd	nd	nd	nd
(nonreducing)	101.7	75.3				
4-linked-β-D-Xylp	4.453	3.280	3.54	3.761	nd	nd
(nonreducing)	101.7	73.0	73.8	76.5		
terminal–β-D-Xylp	4.438	3.241	3.42	3.597	nd	nd
(nonreducing)	101.7	73.0	75.8	69.3		

a. The axial proton on C5 of Xylp residues is designated H5 and the equatorial proton is designated H5'. The stereochemistry (*pro*-R *vs. pro*-S) of H5 and H5' of Ara*f* residues is not specified.