



**Diana de Oliveira Ribeiro**

M.Sc in Structural and Functional Biochemistry

**Protein-carbohydrate recognition in the  
biodegradation of the plant cell wall:  
Functional and structural studies using  
carbohydrate microarrays and X-ray  
crystallography**

Thesis submitted for the degree of Doctor of Philosophy in  
Biochemistry

Supervisor: Angelina Sá Palma, PhD  
Assistant Researcher UCIBIO

Faculdade de Ciências e Tecnologia, NOVA  
Co-Supervisor: Ana Luísa Carvalho, PhD  
Assistant Researcher UCIBIO

Faculdade de Ciências e Tecnologia, NOVA  
Co-Supervisor: Ten Feizi, MD

Full Professor and Director of Glycosciences Laboratory  
Imperial College London

Examination Committee:

President: Professor Maria João Lobo de Reis Madeira Crispim Romão

Main Examiners: Doctor Serge Perez  
Doctor Isabel Maria Travassos de Almeida de Jesus Bento

Examiners: Professor Carlos Mendes Godinho de Andrade Fontes  
Doctor Maria Angelina de Sá Palma  
Doctor Filipa Margarida Barradas Morais Marcelo



FACULDADE DE  
CIÊNCIAS E TECNOLOGIA  
UNIVERSIDADE NOVA DE LISBOA

March 2020



**Universidade Nova de Lisboa**  
**Faculdade de Ciências e Tecnologia**

**Diana de Oliveira Ribeiro**

M.Sc in Structural and Functional Biochemistry

**Protein-carbohydrate Recognition in the Biodegradation  
of the Plant Cell Wall:**  
Functional and Structural Studies Using Carbohydrate  
Microarrays and X-ray Crystallography

Thesis submitted for the degree of Doctor of Philosophy in Biochemistry

**March 2020**



**Protein-carbohydrate Recognition in the Biodegradation of the Plant Cell Wall:**  
Functional and Structural Studies Using Carbohydrate Microarrays and X-ray Crystallography

“Copyright” em nome de Diana de Oliveira Ribeiro, FCT-NOVA

A Faculdade de Ciências e Tecnologia e a Universidade Nova de Lisboa têm o direito, perpétuo e sem limites geográficos, de arquivar e publicar esta dissertação através de exemplares impressos reproduzidos em papel ou de forma digital, ou por qualquer outro meio conhecido ou que venha a ser inventado, e de a divulgar através de repositórios científicos e de admitir a sua cópia e distribuição com objectivos educacionais ou de investigação, não comerciais, desde que seja dado crédito ao autor e editor.



The work presented in this Thesis was developed in the scope of the Individual Doctorate Grant SFRH/BD/100569/2014 and the projects PTDC/BIA-MIB/31730/2017, PTDC/BBB-BEP/0869/2014, PTDC/BIA-MIC/5947/2014, RECI/BBB-BEP/0124/2012, funded by Fundação para a Ciência e a Tecnologia - Ministério da Ciência, Tecnologia e Ensino Superior (FCT-MCTES), in the research unit UCIBIO, funded by FCT-MCTES, and complemented with a stay at the Glycosciences Laboratory, Imperial College London, co-funded by Wellcome Trust.

From the work developed have resulted the following publications:

1. **Ribeiro, D.O.**, Pinheiro, B.A, Carvalho, A.L., Palma, A.S., Targeting protein-carbohydrate interactions in plant cell-wall biodegradation: the power of carbohydrate microarrays in Carbohydrate Chemistry: Chemical and biological approaches, eds. A. P. Rauter, T. Lindhorst and Y. Queneau, Royal Society of Chemistry, 2018, vol. 43, pp. 159–176 (DOI: 10.1039/9781788010641-00159).
2. **Ribeiro, D.O.**, Viegas, A., Pires, V.M.R., Medeiros-Silva, J., Bule, P., Chai, W., Marcelo, F., Fontes, C.M.G.A., Cabrita, E.J., Palma, A.S., Carvalho, A.L., Molecular basis for the preferential recognition of  $\beta$ 1,3-1,4-glucans by the family 11 Carbohydrate-Binding Module from *Clostridium thermocellum*, FEBS Journal, 2019, in press (DOI: 10.1111/febs.15162).





## Acknowledgments

*This Thesis is dedicated to my partner in life, Nuno Mariani. For always being at my side, since day one, no matter what, pushing me further. For all the love, unconditional strength and support, and all the cooked meals! I wouldn't have done it without you. And the best is yet to come!*

*To my supervisors, Doctor Angelina Sá Palma and Doctor Ana Luísa Carvalho, a profound thank you for the opportunity, the guidance, encouragement and all the support even through the hardest moments, it has been 5 years of professional but also personal enrichment.*

*A big and warm thank you to Prof. Ten Feizi for receiving me with open arms in her group and sharing her expertise and knowledge, and always giving her tirelessly support.*

*A special thank you to Doctor Wengang Chai, for eagerly supervising me and for all his continued guidance, during my stay and even after from a distance.*

*I would like to express my gratitude to Prof. Maria João Romão, for opening the doors of her lab to me all those years ago, which eventually lead me to this path.*

*To our collaborators, Prof. Carlos Fontes, Doctor Joana Brás, Doctor Virgínia Pires, Doctor Natércia Brás, Prof. Manuel Coimbra, Doctor Eurico Cabrita and Doctor Filipa Marcelo.*

*To Doctor Benedita Pereira, Doctor Márcia Correia and Doctor Yibing Zhang for always being available to help and share their knowledge and guidance.*

*To Viviana Correia and Raquel Costa for all the support, companionship and help keeping my mental sanity during these years. Also, for Raquel's valuable contributions to this work, it was a pleasure teaching you, learning with and from you.*

*A thank you to my colleagues, past and present, Lisete, Francisco, Raquel S., Marino, Filipa, Filipe, Catarina and Cecília.*

*To my friends Ana, Rita, Joana, Liliana, Cláudia, and all others that have accompanied me and given their support during this journey.*

*To my mother, the biggest thank you, for always believing in me, unconditionally supporting me every step of the way. I am where I am because of you.*

*To Carlos, for contributing more than he probably realises.*

*To my father, for all his support, and my sisters and brother Patrícia, Susana and Hugo for all that I've learned from you.*

*To my dearest in-laws, Idalécio and São, for all their love and kindness.*

*Finally, to my grandfather Artur, that will always be present in my thoughts.*

*Thank you*



*“The beauty of a living thing is not the atoms that go into it,  
but the way those atoms are put together”*

Carl Sagan



## Abstract

The plant cell wall is, in its majority, constituted by complex and structurally diverse polysaccharides that are valuable resources for industrial and biotechnological applications. Anaerobic microbial organisms are highly efficient for plant cell wall polysaccharide biodegradation and have evolved a multi-enzyme complex system, the Cellulosome, where catalytic enzymes have non-catalytic Carbohydrate Binding Modules (CBMs) appended that highly potentiate the enzymes' catalytic efficiency. Deciphering at molecular level the mechanisms underlying plant cell wall carbohydrate recognition and deconstruction by different cellulolytic bacteria is crucial to elucidate these complex biological systems, as well as to further promote novel potential applications. The work developed in this Thesis focused on the unique approach of combining carbohydrate microarrays with X-ray crystallography, to uncover carbohydrate ligands for CBMs and to structurally characterize novel CBM-carbohydrate interactions of two anaerobic bacteria that reside in different ecological niches: *Clostridium thermocellum*, found in soils, and *Ruminococcus flavefaciens* FD-1, present in the rumen of herbivorous. To this end, microarrays featuring carbohydrate probes with polysaccharide and oligosaccharide sequences representative of the structural diversity found on plant cell walls, but also in fungal and bacterial cell walls, were developed and then used to screen the carbohydrate-binding and ligand-specificity of 150 CBMs of *C. thermocellum* and *R. flavefaciens* CBMs. The groups of polysaccharides that are differentially recognised were revealed for 59 CBMs and novel CBM-ligand specificities were identified for 23 modules from *C. thermocellum* and 21 from *R. flavefaciens*. Overall, the two bacteria differentially expressed CBM families with different carbohydrate-binding specificities, which may reflect adaptation to substrate availability in their specific ecological niche or the complexity of their Cellulosome. Using the information derived from the high-throughput microarray analysis, and according to their biotechnological relevance or novelty, CBMs and the respective ligands were selected for further structural studies. The novel CBM structures solved, complemented with biochemical and biophysical data, enabled the characterization of the molecular determinants for the recognition of mixed-linked  $\beta$ 1,3-1,4-glucans by *C. thermocellum* family 11 CBM, chitin and peptidoglycan-derived sequences by a novel LysM domain from *C. thermocellum* family 50 CBMs, and pectic arabinans by *R. flavefaciens* family 13 CBM. The results reported here allow to assign a functional role for these CBMs and CBM families and contribute to the classification of the novel CBMs identified in the genome of the two bacteria, particularly those from *R. flavefaciens* FD1. Furthermore, the information derived from this integrative study, can promote a better understanding of cellulolytic capabilities of these bacteria, as well as to potentiate biotechnological applications of CBMs.

**Keywords:** Carbohydrate-Binding Modules (CBMs) • Protein-carbohydrate recognition • *Clostridium thermocellum* • *Ruminococcus flavefaciens* FD-1 • Carbohydrate microarrays • X-ray crystallography



## Resumo

A parede celular das plantas é maioritariamente constituída por polissacáridos complexos e estruturalmente diversos, também designados de hidratos de carbono, que são importantes recursos para aplicações industriais e biotecnológicas. Alguns microorganismos que residem em ambiente anaeróbio são altamente especializados e eficientes na degradação de hidratos de carbono da parede celular vegetal, tendo desenvolvido para esta função um complexo multi-enzimático de tamanho megadalton à superfície celular, designado de Celulossoma. Neste complexo, os módulos catalíticos encontram-se adjacentes a módulos não-catalíticos de ligação a hidratos de carbono (CBMs), que potenciam a eficiência catalítica das enzimas. De forma a elucidar estes sistemas biológicos complexos e promover novas aplicações, é crucial decifrar ao nível molecular os mecanismos subjacentes ao reconhecimento e desconstrução de hidratos de carbono por diferentes bactérias celulolíticas. Neste sentido, esta Tese combina a tecnologia dos *microarrays* de hidratos de carbono e a cristalografia de raios-X para identificar ligandos e caracterizar estruturalmente novas interações CBM-hidratos de carbono de duas bactérias anaeróbias que residem em diferentes nichos ecológicos: *Clostridium thermocellum*, encontrada maioritariamente no solo, e *Ruminococcus flavefaciens* FD-1, presente no rúmen de herbívoros. No trabalho desenvolvido, foram aplicados *microarrays* constituídos de hidratos de carbono (polissacáridos e oligossacáridos) com sequências representativas da diversidade estrutural da parede celular das plantas, e outras de fungos ou bactérias, para investigar o reconhecimento e a especificidade de 150 CBMs identificados no genoma destas bactérias. Os resultados identificaram diferentes grupos de polissacáridos reconhecidos por 59 CBMs e novas especificidades para 23 CBMs de *C. thermocellum* e 21 CBMs de *R. flavefaciens*. Os CBMs das duas bactérias exibiram especificidades diferentes para ligação a hidratos de carbono, o que poderá reflectir a adaptação ao seu nicho ecológico e a complexidade dos seus celulossomas. Utilizando a informação derivada dos *microarrays*, e de acordo com a relevância biotecnológica ou novidade, CBMs e respectivos ligandos foram seleccionados para caracterização estrutural. As novas estruturas de CBMs resolvidas permitiram a caracterização dos determinantes moleculares para o reconhecimento de  $\beta$ 1,3-1,4-glucanos pelo CBM da família 11 de *C. thermocellum*, de quitina e peptidoglicano por um novo domínio LysM da família de CBMs 50 de *C. thermocellum*, e de arabinanos pécticos por CBMs da família 13 de *R. flavefaciens*. Os resultados permitiram atribuir um papel funcional a estes CBMs e famílias de CBMs nas duas bactérias, contribuindo para a classificação dos novos CBMs identificados, particularmente de *R. flavefaciens*. Este estudo integrativo permitiu uma melhor compreensão das capacidades celulolíticas destas bactérias, e servirá para potenciar aplicações biotecnológicas destes CBMs.

**Palavras-chave:** Módulos de ligação a hidratos de carbono (CBMs) • Reconhecimento proteína-hidrato de carbono • *Clostridium thermocellum* • *Ruminococcus flavefaciens* FD-1 • *Microarrays* de hidratos de carbono • Cristalografia de raios-X





## Table of contents

Acknowledgments.....	V
Abstract.....	IX
<i>Resumo</i> .....	XI
Table of contents.....	XIII
List of figures.....	XIX
List of tables.....	XXIII
Abbreviations and symbols.....	XXV
Thesis outline.....	XXIX
<b>Chapter 1.....</b>	<b>1</b>
<b>1 General Introduction.....</b>	<b>3</b>
1.1 <i>Structural diversity of plant cell wall polysaccharides</i> .....	3
1.2 <i>Cellulolytic microorganisms' express proteomes highly efficient in plant cell wall biodegradation</i> .....	4
1.2.1 Carbohydrate-binding modules: the non-catalytic domains associated with Carbohydrate Active enZymes.....	7
1.2.1.1 Classification of CBMs.....	7
1.2.1.2 Functional Roles of CBMs.....	8
1.2.2 Biotechnological applications of carbohydrate-binding proteins.....	10
1.2.3 <i>Clostridium thermocellum</i> and <i>Ruminococcus flavefaciens</i> FD-1.....	11
1.3 <i>Methods for characterizing protein-carbohydrate interactions</i> .....	12
1.3.1 Carbohydrate microarrays.....	12
1.3.1.1 Carbohydrate microarray platforms.....	13
1.3.1.2 Microarrays focused on plant carbohydrates for recognition studies.....	17
1.3.1.3 Combining microarray analysis with mass spectrometry.....	18
1.3.2 Protein crystallography.....	19
1.3.2.1 Protein crystallization.....	20
1.3.2.2 X-ray diffraction.....	23
1.3.2.3 3D structure determination.....	25
1.3.2.4 Model building and validation.....	26
1.4 <i>Structural characterization of protein-carbohydrate interactions</i> .....	27
1.5 <i>Thesis main objectives</i> .....	31
<b>Chapter 2.....</b>	<b>33</b>
<b>2 Development of glucan and hemicellulose oligosaccharide microarrays applied to plant cell wall carbohydrate recognition.....</b>	<b>35</b>
2.1 <i>Introduction</i> .....	35
2.2 <i>Results</i> .....	37
2.2.1 Construction of glucan and hemicellulose oligosaccharide microarrays.....	37
2.2.2 Validation of the constructed microarrays.....	40

2.2.2.1	Recognition of gluco-oligosaccharides with $\alpha$ - and $\beta$ -glycosidic linkages in linear or branched chains.....	40
2.2.2.2	Recognition of linear $\beta$ -xylans, branched arabinoxylans and $\alpha$ -arabinans.....	42
2.2.2.3	Differential recognition of linear $\beta$ 1,4-mannans and branched galactomannans.....	44
2.2.2.4	Recognition studies of xyloglucan oligosaccharides using complex oligosaccharide mixtures and a deconvolution strategy.....	46
2.3	<i>Discussion</i> .....	50
2.4	<i>Conclusions</i> .....	54
2.5	<i>Experimental procedures</i> .....	55
2.5.1	Monoclonal antibodies, CBMs and lectins used for probe structural validation and microarray quality control.....	55
2.5.2	Sources of carbohydrates.....	55
2.5.3	Preparation of oligosaccharide fractions.....	55
2.5.4	Preparation of AO-NGLs by oxime-ligation.....	56
2.5.5	Preparation of DAN-conjugated xyloglucan oligosaccharides by reductive amination.....	56
2.5.6	Preparation of DHPA-NGLs by reductive amination.....	57
2.5.7	MALDI Mass Spectrometry.....	57
2.5.8	Electrospray Mass Spectrometry.....	57
2.5.9	Carbohydrate microarrays construction and analysis.....	58
2.5.10	Microarray data analysis and presentation.....	59
2.6	<i>Work contributions</i> .....	59
<b>Chapter 3</b>	.....	<b>61</b>
<b>3 Cellulolytic bacteria express CBMomes that dictate their ecological niche polysaccharide utilization</b>	.....	<b>63</b>
3.1	<i>Introduction</i> .....	63
3.2	<i>Results and Discussion</i> .....	64
3.2.1	Multi-step strategy to assign carbohydrate-binding specificities of CBMs in a high-throughput manner.....	64
3.2.2	Bacterial CBMomes from different ecological niches.....	66
3.2.3	Carbohydrate microarray platforms for ligand discovery.....	67
3.2.4	Screening <i>C. thermocellum</i> and <i>R. flavefaciens</i> FD-1 CBMomes for carbohydrate-binding specificity.....	69
3.2.4.1	Recognition of $\alpha$ -glucans and $\beta$ -glucans with linear or branched chains.....	75
3.2.4.2	Recognition of linear $\beta$ 1,4 mannans and branched galactomannans.....	77
3.2.4.3	Recognition of $\alpha$ -arabinose- and galactose-containing sequences in different polysaccharides.....	78
3.2.4.4	Assignment of <i>C. thermocellum</i> family 50 CBMs ligand specificity towards $\beta$ 1,4-GlcNAc oligosaccharides.....	79
3.2.4.5	Assignment of ligand specificity and chain-length requirement for families 6 and 22 CBMs towards $\beta$ -xylans.....	80

3.2.5	CBM families for which carbohydrate binding was not identified in the microarray analyses.....	82
3.2.6	CBMs spectrum of carbohydrate recognition reflects the bacteria's ecological niche.....	82
3.3	<i>Conclusions</i> .....	83
3.4	<i>Experimental procedures</i> .....	84
3.4.1	Monoclonal antibodies, CBMs and lectins used for microarray quality control.....	84
3.4.2	High-throughput cloning, expression and purification of <i>C. thermocellum</i> and <i>R. flavefaciens</i> FD-1 CBMs.....	84
3.4.3	Sources of carbohydrates.....	85
3.4.4	Carbohydrate microarray analysis.....	85
3.4.5	Microarray data analysis and presentation.....	86
3.4.6	Affinity gel electrophoresis with soluble polysaccharides.....	87
3.5	<i>Work contributions</i> .....	87
<b>Chapter 4</b> .....		<b>89</b>
<b>4</b>	<b>Molecular basis for the preferential recognition of <math>\beta</math>1,3-1,4-glucans by the family 11 CBM from <i>Clostridium thermocellum</i></b> .....	<b>91</b>
4.1	Introduction.....	91
4.2	Results and Discussion.....	93
4.2.1	Specificity assignment using carbohydrate microarrays.....	93
4.2.2	Crystal structure of CtCBM11 bound to $\beta$ 1,3-1,4-gluco-oligosaccharides.....	95
4.2.3	CtCBM11 binding mode.....	98
4.2.4	The CH- $\pi$ stacking and hydrogen bonding network as determinants of the ligand-specificity.....	101
4.2.5	CtCBM11 ligand specificity in the context of CAZy CBMs.....	104
4.3	Conclusions.....	105
4.4	Experimental procedure.....	105
4.4.1	Gene cloning, mutagenesis and protein purification.....	105
4.4.2	Sources of carbohydrates.....	106
4.4.3	Mass spectrometry analysis of barley hexasaccharide.....	106
4.4.4	Carbohydrate microarray analysis.....	106
4.4.5	Crystallization and X-ray Diffraction Data Collection.....	107
4.4.6	Phasing, model building, and refinement.....	107
4.4.7	Isothermal titration calorimetry.....	108
4.5	Work contributions.....	109
<b>Chapter 5</b> .....		<b>111</b>
<b>5</b>	<b>Unravelling family 50 CBMs of <i>Clostridium thermocellum</i>: Structural and functional characterization of a new LysM domain</b> .....	<b>113</b>
5.1	Introduction.....	113
5.2	Results and Discussion.....	114

5.2.1	Oligosaccharide specificity of <i>C. thermocellum</i> family 50 CBMs.....	114
5.2.2	CtCBM50 structure in complex with $\beta$ 1,4-GlcNAc trisaccharide.....	116
5.2.3	Binding affinity of CtCBM50 to $\beta$ 1,4-GlcNAc oligosaccharides and influence of chain-length.....	120
5.2.4	Molecular determinants of CtCBM50 ligand recognition and chain-length dependency.....	122
5.2.5	CtCBM50 interaction with peptidoglycan sequences.....	124
5.2.6	<i>Clostridium thermocellum</i> family 50 CBMs in the context of LysM domains.....	128
5.3	<i>Conclusions</i> .....	130
5.4	<i>Experimental procedure</i> .....	131
5.4.1	Gene cloning, mutagenesis and protein purification.....	131
5.4.2	Sources of carbohydrates.....	131
5.4.3	Carbohydrate microarray analysis.....	132
5.4.4	Crystallization and X-ray Diffraction Data Collection.....	132
5.4.5	Phasing, Model Building, and Refinement.....	132
5.4.6	Isothermal titration calorimetry.....	133
5.4.7	Molecular modelling.....	133
5.4.8	Minimization, molecular dynamics simulations and binding energies.....	134
5.4.9	Binding to insoluble polysaccharides by co-precipitation assays.....	134
5.5	<i>Work contributions</i> .....	135
<b>Chapter 6</b> .....		<b>137</b>
<b>6 Assigning the carbohydrate specificity of <i>Ruminococcus flavefaciens</i> family 13 CBMs: Recognition of pectic arabinans by a novel CBM13</b> .....		<b>139</b>
6.1	<i>Introduction</i> .....	139
6.2	<i>Results and Discussion</i> .....	142
6.2.1	Ligand specificity of <i>R. flavefaciens</i> family 13 CBMs.....	142
6.2.2	Crystal structure of RfCBM13-1 revealing the putative binding sites.....	144
6.2.3	Characterization of RfCBM13-1-ligand interaction.....	145
6.2.4	<i>R. flavefaciens</i> family 13 CBMs in the context of plant cell wall recognition.....	150
6.3	<i>Conclusions</i> .....	153
6.4	<i>Experimental procedure</i> .....	154
6.4.1	Gene cloning, mutagenesis and protein purification.....	154
6.4.2	Sources of carbohydrates.....	155
6.4.3	Carbohydrate microarray analysis.....	155
6.4.4	Crystallization and X-ray Diffraction Data Collection.....	155
6.4.5	Phasing, Model Building, and Refinement.....	155
6.4.6	Isothermal titration calorimetry.....	156
6.5	<i>Work contributions</i> .....	156

<b>Chapter 7.....</b>	<b>157</b>
<b>7 Conclusions and future perspectives.....</b>	<b>159</b>
7.1 <i>General conclusions.....</i>	<i>159</i>
7.2 <i>Future perspectives.....</i>	<i>161</i>
<b>References.....</b>	<b>163</b>
<b>Appendix.....</b>	<b>177</b>
<b>Chapter 2 - Supplementary Information.....</b>	<b>179</b>
<i>Supplementary Figures.....</i>	<i>179</i>
<i>Supplementary Tables.....</i>	<i>181</i>
<b>Chapter 3 - Supplementary Information.....</b>	<b>199</b>
<i>Supplementary Figures.....</i>	<i>199</i>
<i>Supplementary Tables.....</i>	<i>202</i>
<b>Chapter 4 - Supplementary Information.....</b>	<b>233</b>
<i>Supplementary Figures.....</i>	<i>233</i>
<i>Supplementary Tables.....</i>	<i>234</i>
<b>Chapter 5 - Supplementary Information.....</b>	<b>237</b>
<i>Supplementary Figures.....</i>	<i>237</i>
<i>Supplementary Tables.....</i>	<i>239</i>
<b>Chapter 6 - Supplementary Information.....</b>	<b>245</b>
<i>Supplementary Figure.....</i>	<i>245</i>
<i>Supplementary Table.....</i>	<i>245</i>



## List of figures

### Chapter 1

Figure 1.1. Illustrative representation of the diversity of major polysaccharides in the plant cell wall. ....	3
Figure 1.2. Examples of structures of polysaccharides found in the plant cell wall. ....	5
Figure 1.3. The anaerobic bacterial cellulosome. ....	6
Figure 1.4. Carbohydrate recognition by CBMs. ....	9
Figure 1.5. Schematic overview of the main steps comprising the analysis using neoglycolipid (NGL)-based carbohydrate microarrays. ....	14
Figure 1.6. Graphic representation of examples of immobilization strategies used to generate carbohydrate microarrays. ....	16
Figure 1.7. Schematic overview of the main steps comprised in the determination of protein and protein-ligand structures by X-ray crystallography. ....	20
Figure 1.8. Phase diagram for protein crystallization. ....	21
Figure 1.9. Assembly of unit cells in a three-dimensional crystal lattice. ....	23
Figure 1.10. Bragg's Law. ....	24
Figure 1.11. Examples of carbohydrate-aromatic CH- $\pi$ interactions. ....	28

### Chapter 2

Figure 2.1. Binding patterns revealed by probing the glucan and hemicellulose oligosaccharide microarrays with sequence-specific carbohydrate-binding proteins. ....	39
Figure 2.2. Microarray analysis of glucan-binding proteins. ....	41
Figure 2.3. Microarray analysis of xylan- and arabinan-binding proteins. ....	43
Figure 2.4. Microarray analysis of mannan-binding proteins. ....	45
Figure 2.5. Deconvolution of the xyloglucan oligosaccharides from tamarind. ....	47
Figure 2.6. Positive-ion ESI-MS/MS product-ion spectra used for sequencing of xyloglucan-DAN DP-8. ....	48
Figure 2.7. Preparation of the xyloglucan-DAN-NGL probes from tamarind included in the new xyloglucan microarrays. ....	49
Figure 2.8. Validation and analysis of the new xyloglucan microarrays. ....	51

### Chapter 3

Figure 3.1. Schematic representation of the multi-step strategy followed in this work. ....	65
Figure 3.2. Overview of the selected CBMs of the two bacteria. ....	66

Figure 3.3. Validation of the polysaccharide microarrays with sequence-specific carbohydrate-binding proteins.....	68
Figure 3.4. Analysis of <i>C. thermocellum</i> and <i>R. flavefaciens</i> FD-1 CBMs using polysaccharide microarrays – 1 <sup>st</sup> screening for carbohydrate-binding activities. ....	70
Figure 3.5. Validation of the <i>C. thermocellum</i> CBMs microarray binding patterns using affinity gel electrophoresis. ....	71
Figure 3.6. Validation of the <i>R. flavefaciens</i> CBMs microarray binding patterns using affinity gel electrophoresis. ....	72
Figure 3.7. Analysis of <i>C. thermocellum</i> and <i>R. flavefaciens</i> FD-1 CBMs using oligosaccharide microarrays – 2 <sup>nd</sup> screening for assigning carbohydrate-binding specificities. ....	73
Figure 3.8. Comparison of carbohydrate-binding specificities of families 6 and 22 CBMs from <i>C. thermocellum</i> and <i>R. flavefaciens</i> FD-1 to xylan sequences. ....	81

## Chapter 4

Figure 4.1. Top view on the identified binding site of wild-type CtCBM11. ....	92
Figure 4.2. Analysis of carbohydrate binding specificity using a microarray of sequence-defined gluco-oligosaccharides. ....	94
Figure 4.3. Ribbon representation of the three-dimensional crystal structures of CtCBM11 complexes. ....	96
Figure 4.4. Comparison of unbound and ligand-bound CtCBM11 structures. ....	98
Figure 4.5. CtCBM11-ligand interactions. ....	99
Figure 4.6. Representative isothermal calorimetry titrations of binding of CtCBM11 and its mutants to oligosaccharides. ....	101
Figure 4.7. Alignment of CBM11 family members. ....	104

## Chapter 5

Figure 5.1. Modular architecture of proteins containing family 50 CBMs in the genome of <i>C. thermocellum</i> . ....	115
Figure 5.2. Oligosaccharide microarray analysis of <i>C. thermocellum</i> family 50 CBMs. ....	116
Figure 5.3. Comparative analysis of <i>C. thermocellum</i> family 50 CBMs binding to $\beta$ 1,4-linked GlcNAc oligosaccharides. ....	117
Figure 5.4. Ribbon representation of the three-dimensional crystal structure of the CtCBM50-GlcNAc <sub>3</sub> complex. ....	118
Figure 5.5. Isothermal calorimetry titrations of binding of CtCBM50 and its mutant derivatives to $\beta$ 1,4-linked GlcNAc oligosaccharides. ....	121
Figure 5.6. Representation of the last simulation structure of the various GlcNAc oligosaccharides bound to the CtCBM50 complex of chains A and B. ....	123
Figure 5.7. Molecular dynamics simulations of CtCBM50 chain-length dependency. ....	125



Figure 5.8. Binding of CtCBM50 to insoluble chitin and peptidoglycan. ....	126
Figure 5.9. Close-up of inter- and intra-chain hydrogen bonds involving the HO-C6 group of the GlcNAc residue in simulations with GlcNAc and MurNAc-GlcNAc pentasaccharides. ....	127
Figure 5.10. Alignment of CBM50 family members. ....	128
Figure 5.11. Superposition of CtCBM50 with <i>Thermus thermophilus</i> LysM1. ....	129
Figure 5.12. Schematic representation illustrating the hypothesized cooperative binding by CtCBM50 to short and long GlcNAc oligosaccharides. ....	130

## Chapter 6

Figure 6.1. Schematic representation of the major pectic polysaccharide structural domains. ....	140
Figure 6.2. Modular architecture of <i>R. flavefaciens</i> proteins containing family 13 CBMs. ....	141
Figure 6.3. Carbohydrate microarray analysis of <i>R. flavefaciens</i> family 13 CBMs. ....	142
Figure 6.4. Binding analysis of <i>R. flavefaciens</i> family 13 CBMs arabinan-derived oligosaccharides included in the carbohydrate microarrays. ....	143
Figure 6.5. Ribbon representation of RfCBM13-1 three-dimensional structure. ....	145
Figure 6.6. Isothermal calorimetry titrations of binding of RfCBM13-1 to $\alpha$ 1,5-linked arabinan sequences. ....	147
Figure 6.7. Isothermal calorimetry titrations of binding of RfCBM13-1 mutant derivatives to $\alpha$ 1,5-linked arabinan sequences. ....	149
Figure 6.8. Alignment of CBM13 family members. ....	151
Figure 6.9. Superposition of RfCBM13-1 with <i>Streptomyces avermitilis</i> CBM13. ....	152



## List of tables

### Chapter 2

Table 2.1. MALDI-MS analysis of AO-NGLs derived from hemicellulose oligosaccharides. ....	38
Table 2.2. MALDI-MS analysis of xyloglucan-DAN-DHPA NGLs investigated. ....	50
Table 2.3. MALDI-MS analysis of the new xyloglucan-AO-NGL probes generated. ....	52

### Chapter 3

Table 3.1. Summary of the <i>C. thermocellum</i> CBMs ligand recognition and specificity obtained in the carbohydrate microarray screenings and affinity gel electrophoresis (AGE), cross-referencing with the available literature. ....	74
Table 3.2. Summary of the <i>R. flavefaciens</i> FD-1 CBMs ligand recognition and specificity obtained in the carbohydrate microarray screenings and affinity gel electrophoresis (AGE). ....	76

### Chapter 4

Table 4.1. X-ray diffraction and structure refinement parameters and statistics for CtCBM11-G4G4G3G and CtCBM11-G4G3G4G4G3G structures. ....	97
Table 4.2. Thermodynamic parameters of the binding of CtCBM11 and its mutant derivatives to polysaccharides and oligosaccharides. ....	102

### Chapter 5

Table 5.1. X-ray diffraction and structure refinement parameters and statistics for CtCBM50-GlcNAc <sub>3</sub> . ....	119
Table 5.2. Thermodynamic parameters of the binding of CtCBM50 and its mutant derivatives to polysaccharides and oligosaccharides. ....	122
Table 5.3. Binding enthalpies and binding free-energies of the GlcNAc ligands to the CtCBM50 chains A and B. ....	124
Table 5.4. Binding enthalpies and binding free-energies of the peptidoglycan ligands to CtCBM50. ....	127

### Chapter 6

Table 6.1. X-ray diffraction and structure refinement parameters and statistics for RfCBM13-1. ....	146
Table 6.2. Thermodynamic parameters of the binding of RfCBM13-1 wild type and its mutant derivatives to polysaccharides and oligosaccharides. ....	148



## Abbreviations and symbols

















<b>ACN</b>	Acetonitrile
<b>AG-I</b>	Arabinogalactan type I
<b>AG-II</b>	Arabinogalactan type II
<b>AGE</b>	Affinity Gel Electrophoresis
<b>AOPE (AO)</b>	Aminoxy-functionalized-1,2-dihexadecyl- <i>sn</i> -glycero-3-phosphoethanolamine
<b>BSA</b>	Bovine Serum Albumin
<b>CAZymes</b>	Carbohydrate Active enZymes
<b>CBD</b>	Cellulose Binding Domain
<b>CBM</b>	Carbohydrate Binding Module
<b>CBMomes</b>	All CBM modules assigned to CAZy families up to 2015
<b>CE</b>	Carbohydrate Esterase
<b>CFG</b>	Consortium for Functional Glycomics
<b>CID</b>	Collision-Induced Dissociation
<b>CoMPP</b>	Comprehensive Microarray Polymer Profiling
<b>Ct</b>	<i>Clostridium thermocellum</i> ( <i>C. thermocellum</i> )
<b>Cy3</b>	Cyanine 3 fluorophore
<b>DAN</b>	1,5-diaminonaphthalene
<b>DHPA</b>	<i>N</i> -(4-formylbenzamide)-1,2-dihexadecyl- <i>sn</i> -glycero-3-phosphoethanolamine
<b>DHPE</b>	1,2-dihexadecyl- <i>sn</i> -glycero-3-phosphoethanolamine
<b>DOC</b>	Dockerin
<b>DP</b>	Degree of Polymerization
<b><i>E. coli</i></b>	<i>Escherichia coli</i>
<b>ESI</b>	Electrospray Ionization
<b>ESRF</b>	European Synchrotron Radiation Facility
<b>GH</b>	Glycoside Hydrolases
<b>HEC</b>	Hydroxyethyl cellulose
<b>HEPES</b>	Hydroxyethyl piperazineethanesulfonic acid
<b>HG</b>	Homogalacturonan
<b>HPLC</b>	High Performance Liquid Chromatography
<b>HPTLC</b>	High-Performance Thin Layer Chromatography
<b>IMAC</b>	Ion Metal Affinity Chromatography
<b>IPTG</b>	Isopropyl $\beta$ -D-1-thiogalactopyranoside
<b>ITC</b>	Isothermal Titration Calorimetry
<b>LB</b>	Luria-Bertani
<b>LIC</b>	Ligation Independent Cloning
<b>Lic</b>	Lichenase
<b>LPMO</b>	Lytic Polysaccharide Monooxygenases

<b>LysM</b>	Lysin Motif domain
<b>MALDI</b>	Matrix-Assisted Laser Desorption/Ionization
<b>MD</b>	Molecular Dynamics
<b>MeOH</b>	Methanol
<b>MIRAGE</b>	Minimum Information Required for A Glycomics Experiment
<b>MOPS</b>	3-morpholinopropane-1-sulfonic acid
<b>MR</b>	Molecular Replacement
<b>MS</b>	Mass Spectrometry
<b>NGL</b>	NeoGlycoLipid
<b>NHS</b>	<i>N</i> -hydroxysuccinimide
<b>NMR</b>	Nuclear magnetic resonance
<b>OD<sub>600nm</sub></b>	Optical density at 600 nm
<b>PDB</b>	Protein Data Bank
<b>PEG</b>	Polyethylene glycol
<b>PG</b>	Peptidoglycan
<b>PL</b>	Polysaccharide Lyases
<b>Rf</b>	<i>Ruminococcus flavefaciens</i> FD-1 ( <i>R. flavefaciens</i> )
<b>RG-I</b>	Rhamnogalacturonan I
<b>RG-II</b>	Rhamnogalacturonan II
<b>rmsd</b>	Root-mean-square deviation
<b>rpm</b>	Rotation per minute
<b>SDS-PAGE</b>	Sodium dodecyl sulphate - Polyacrylamide gel electrophoresis
<b>SLS</b>	Swiss Light Source
<b>SNFG</b>	Symbol Nomenclature for Glycans
<b>TBA</b>	Tetrabutylammonium cyanoborohydride
<b>UV</b>	Ultraviolet
<b>WT</b>	Wild type
<b>XGA</b>	Xylogalacturonan

## Amino acid abbreviations

Alanine	Ala	A
Arginine	Arg	R
Asparagine	Asn	N
Aspartate	Asp	D
Cysteine	Cys	C
Glutamate	Glu	E
Glutamine	Gln	Q
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I
Leucine	Leu	L
Lysine	Lys	K
Methionine	Met	M
Phenylalanine	Phe	F
Proline	Pro	P
Serine	Ser	S
Threonine	Thr	T
Tryptophan	Trp	W
Tyrosine	Tyr	Y
Valine	Val	V

## Monosaccharide abbreviations and symbols<sup>a</sup>

3-deoxy-D-lyxo-2-heptulosaric acid	Dha	
3-deoxy-D-manno-2-octulosonic acid	Kdo	
Apiose	Api	
Arabinose	Ara	
Fucose	Fuc	
Galactose	Gal	
Galacturonic acid	GalA	
Glucose	Glc	
Glucosamine	GlcN	
<i>N</i> -acetyl-glucosamine	GlcNAc	
Glucuronic acid	GlcA	
Mannose	Man	
Muramic acid	MurNAc	
Rhamnose	Rha	
Xylose	Xyl	
3-C-carboxy-5-deoxy-L-xylose	Aceric acid	

<sup>a</sup>The monosaccharide symbolic representation used was according to the updated Symbol Nomenclature for Glycans (SNFG)<sup>1</sup> from the 3<sup>rd</sup> Edition of the Essentials of Glycobiology<sup>2</sup>.





## Thesis outline

The work described in this Thesis focused on a unique approach combining carbohydrate microarrays with X-ray crystallography, to uncover the carbohydrate-binding specificity of CBMs from the cellulolytic bacteria *C. thermocellum* and *R. flavefaciens* FD-1 and to structurally characterize novel CBM-carbohydrate recognition mechanisms at a molecular level. To this end, sequential projects were followed, and the respective results and conclusions were divided into 5 chapters (Chapters 2-6) as outlined below.

Chapter 1 starts with a general introduction on the composition of plant cell wall polysaccharides and the systems that cellulolytic microorganisms employ for its biodegradation, including several biotechnological applications of CBMs. These sections are followed by an overview on the carbohydrate microarrays technology and on protein X-ray crystallography as the two main techniques applied in this Thesis, focusing on the characterization of protein-carbohydrate interactions. At the end of the chapter, the rationale and the main objectives of the Thesis work plan are presented.

Chapter 2 is the first results chapter and describes the construction and validation of an NGL-microarrays platform comprised of naturally-derived glucan- and hemicellulose-related oligosaccharides to address the need of increasing the diversity of plant-derived sequence-defined microarrays developed to this date. The application of the microarrays to assign carbohydrate ligands and the specificities of CBMs and plant carbohydrate-specific antibodies is demonstrated.

The application of these new microarrays to the high-throughput screening of *C. thermocellum* and *R. flavefaciens* CBMs and to assign novel CBM-carbohydrate specificities is described in Chapter 3. Derived from the results obtained from the microarrays screening analysis, and considering their functional, biotechnological and industrial relevance, three representative CBMs, with the respective ligands, were selected for further biochemical and biophysical characterization, and are explored in the next three chapters.

Chapter 4 details the molecular determinants for the ligand recognition of family 11 *C. thermocellum* CBM to mixed-linked  $\beta$ 1,3-1,4-glucans, describing the CBM 3D structures in complex with tetra- and hexa-saccharide ligands.

In the same trend, Chapter 5 is focused on the structural and functional characterization of a *C. thermocellum* CBM from family 50, which is identified as a novel LysM domain binding to chitin ( $\beta$ 1,4-linked GlcNAc) and peptidoglycan sequences.

In Chapter 6, the carbohydrate ligand recognition by *R. flavefaciens* family 13 CBMs is explored, and the structure of a newly identified CBM13 and its binding specificity to pectic  $\alpha$ 1,5-arabinan sequences is described.

Finally, in Chapter 7 the main conclusions of the Thesis work are drawn in an integrative manner, and future perspectives are presented.

# CHAPTER 1

---

## GENERAL INTRODUCTION<sup>1</sup>

---

<sup>1</sup>Some sections of this Introduction were reproduced and updated from Ribeiro, D.O., Pinheiro, B.A., Carvalho, A.L., Palma, A.S. (2018) Targeting protein-carbohydrate interactions in plant cell-wall biodegradation: the power of carbohydrate microarrays. In *Carbohydrate Chemistry: Chemical and biological approaches* (Rauter AP, Lindhorst T, & Queneau Y, eds), pp. 159–176. Royal Society of Chemistry (DOI: 10.1039/9781788010641-00159).

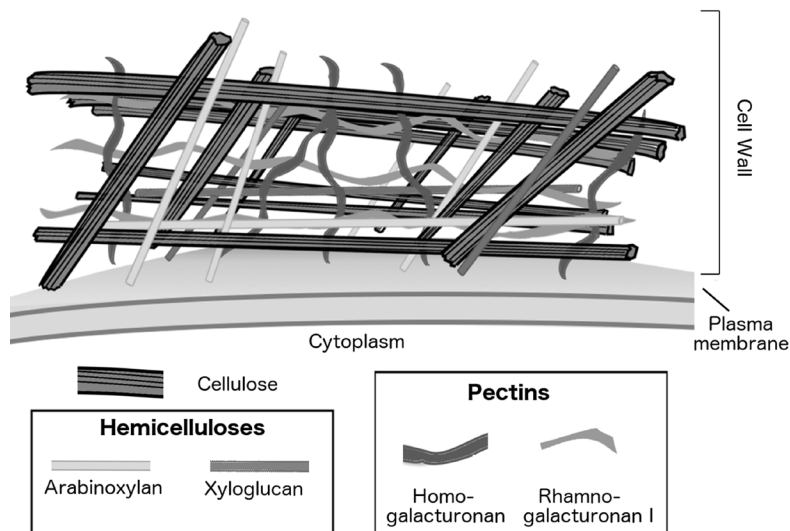


# 1 General Introduction

## 1.1 Structural diversity of plant cell wall polysaccharides

The plant cell wall is an intricate structure composed in its majority by complex polysaccharides and a smaller number of structural proteins. The composition in polysaccharides is highly variable, depending if the plant cell wall is expanding (primary cell wall) or if its role is to give additional structural support to the cell (secondary plant cell wall). It also differs between species, with distinct chemical compositions among the cell walls of grass and flowering plant species<sup>3,4</sup>. The wider molecular and functional diversity of the polysaccharides is mainly observed in the primary cell wall with their configurations changing throughout the plant cell development, expansion and division<sup>5</sup>.

In plant cell walls, microfibrils of the major polysaccharide cellulose form a network embedded in a matrix of various complex polysaccharides, such as hemicelluloses,  $\beta$ -glucans and pectins (Figure 1.1). The hemicelluloses are interconnected with cellulose reinforcing the strength and resilience of the network, while the hydrated gels composed of pectin that intercalate this network, determine the porosity and thickness of the cell wall. The entire structure is maintained by non-covalent interactions, both spontaneous physico-chemical interactions and enzymatic crosslinking, that exist between these polysaccharides<sup>4,6</sup>.



**Figure 1.1. Illustrative representation of the diversity of major polysaccharides in the plant cell wall.** In combination with hemicelluloses, cellulose microfibrils form a network interspersed by pectin polysaccharides. The main hemicellulose polysaccharides found in plant cell walls are xyloglucan (dicot species) and arabinoxylan (grasses). The major pectin polysaccharides are rhamnogalacturonan I and homogalacturonan<sup>4</sup>.

The plant cell wall polysaccharides are structurally diverse (Figure 1.2). Cellulose is composed of aligned linear homopolymers of  $\beta$ 1,4-D-linked glucosyl residues, organized in sheets packed in a

“parallel-up” fashion forming a structure that is, in its majority, crystalline<sup>7</sup>. Hemicelluloses maintain a  $\beta$ 1,4-D-linked backbone but are structurally more complex and composed of other residues, in addition to glucose, such as mannose and xylose, in linear or branched sequences.

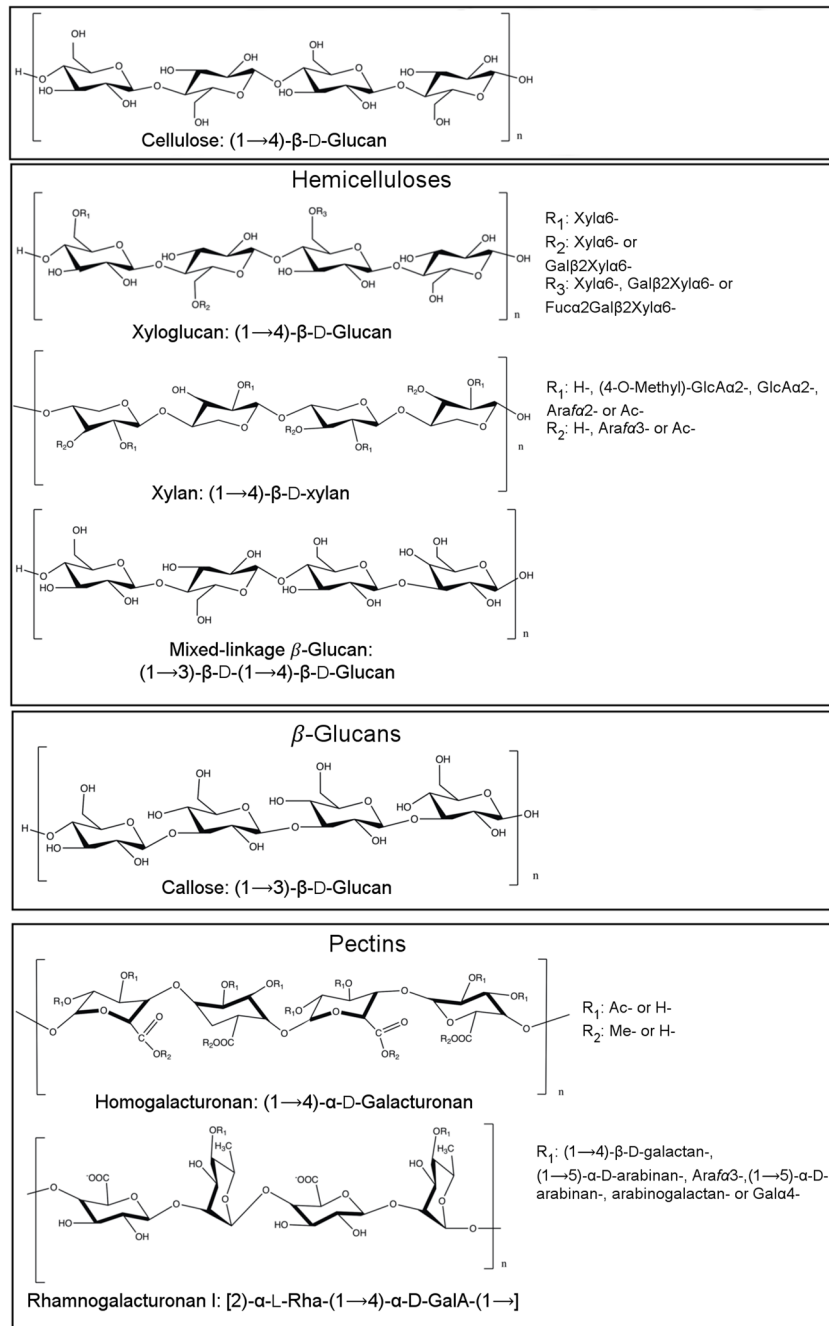
Examples of hemicelluloses include xyloglucans, xylans, mannans, glucomannans and mixed-linked  $\beta$ 1,3-1,4-D-glucans ( $\beta$ 1,4-D-linked glucans with interspersed single  $\beta$ 1,3-D-linkages). All hemicelluloses have significant structural similarity as their backbone residues share the same equatorial configuration at C1 and C4 positions. While xyloglucans are widespread in land plants, the others are more species specific: xylans (dicots and commelinid monocots), mannans (charophytes) and  $\beta$ 1,3-D- $\beta$ 1,4-D-glucans (grasses)<sup>8</sup>. Callose is a  $\beta$ 1,3-D-glucan that is widespread and occurs in specialized walls or wall-associated structures, specifically at stages of differentiation<sup>9</sup>. Pectins are a structurally diverse group of polysaccharides constituted by galacturonic acid (GalpA) in their backbone sequences. The predominant pectins in the primary plant cell wall are: homogalacturonan (HG), a polysaccharide with an unsubstituted backbone of  $\alpha$ 1,4-D-linked GalpA residues, and rhamnogalacturonan-I (RG-I), a polysaccharide with a backbone of the repeating disaccharide  $[-\alpha$ 1,2-L-Rhap- $\alpha$ 1,4-D-GalpA]<sub>n</sub>, substituted at the rhamnose residue with different structural domains, such as galactans, arabinans or arabinogalactans (Figure 1.2). Other pectins such as xylogalacturonan (XGA), which has a homogalacturonan backbone substituted by xylose, and rhamnogalacturonan-II (RG-II), a highly ramified polysaccharide with a homogalacturonan backbone comprising 7 to 9  $\alpha$ 1,4-D-linked GalpA residues are present in smaller amounts. Rhamnogalacturonan-II is the most structurally complex cell wall known polysaccharide as it has 12 different monosaccharide residues interconnected by more than 20 glycosidic types of linkages<sup>10</sup>.

As consequence of this structural diversity, cellulolytic microorganisms that are highly specialised in degrading the plant cell wall polysaccharides have developed a consortium of Carbohydrate Active Enzymes (CAZymes) appended by Carbohydrate Binding Modules (CBMs) for which the diversity of activities and specificities matches the variety of polysaccharides.

## **1.2 Cellulolytic microorganisms' express proteomes highly efficient in plant cell wall biodegradation**

The machinery for degrading plant cell wall polysaccharides differs from anaerobic to aerobic microorganisms, however the modular organization of the polysaccharide-degrading enzymes is maintained in both.

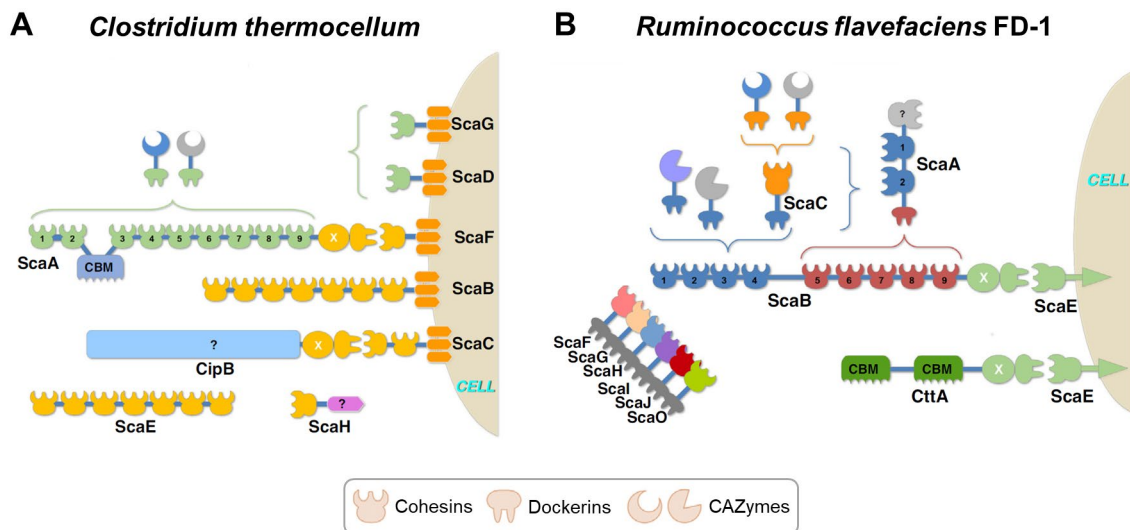
Due to energetic constraints and competition between the species found in anaerobic environments, most anaerobic cellulolytic microorganisms have arranged an efficient but rather elaborated system, where the produced and secreted CAZymes, such as endoglucanases, exoglucanases and  $\beta$ -glucosidases, are assembled in supramolecular complexes (molecular



**Figure 1.2. Examples of structures of polysaccharides found in the plant cell wall.** The Haworth conformational structure of the main chain representative tetrasaccharide sequences is represented; –R, possible ramification position; the different possible sequences of the ramifications to the main backbone chain are depicted. In the mixed-linked β-glucans, the β1,3-linkages separate segments of 2 up to 14 glucose residues linked with β1,4-linkages<sup>9</sup>.

weight >3 MDa), termed the ‘cellulosome’ (Figure 1.3) (see references<sup>11–14</sup> for a comprehensive review on cellulosomes). Cellulosomes show different levels of complexity and are mainly localized at the cell surface<sup>12</sup>. These are known as integrating systems and are composed by one or more scaffoldins, where CAZymes are integrated and brought to the vicinity of the substrate. The scaffoldin is a structural subunit composed of several cohesin modules that bind their binding

partners, the dockerin modules, present in CAZymes and other relevant proteins. Most of the scaffoldins have in their modular structure a CBM from family 3a, that specifically binds to recalcitrant cellulosic substrates<sup>13</sup>. These multi-enzyme complexes can be attached to the bacteria surface and display very complex and dynamic assemblies. One example is the cellulosome from *Clostridium thermocellum*, which has a primary scaffoldin (ScaA) comprising nine highly conserved type I cohesins, which allow the incorporation of different CAZymes and associated CBMs, through their type I dockerins (Figure 1.3A). To attach the scaffoldin subunit to the surface of the bacterial cell, membrane-associated proteins are bound to a type II cohesin<sup>12,15</sup>. Bacteria of the genera *Acetivibrio*, *Clostridium*, *Ruminococcus*, *Thermotoga*<sup>16,17</sup> and fungi of the genera *Neocallimastix*, *Piromyces* and *Orpinomyces*<sup>12</sup> are examples of anaerobic cellulolytic microorganisms. The ecosystems where anaerobes are found to degrade plant polysaccharides to soluble sugars are as diverse as soils, sediments or water bodies. Recently, much attention has been given to the polysaccharide-degrading systems from anaerobic organisms that reside in the digestive tracts of invertebrates and vertebrates<sup>18,19</sup>. These offer novel systems for studying carbohydrate recognition and are important for communication with the host, such as the human host, promoting health and nutritional benefits.



**Figure 1.3. The anaerobic bacterial cellulosome.** Schematic representation the architecture of the cellulosomal assembly of (A) *C. thermocellum* and (B) *R. flavefaciens* FD-1. The colour code represents the different specificities of cohesin-dockerin (Coh-Doc), that compose the various scaffoldins (Sca). The CAZymes modules, can additionally have appended CBMs. (Adapted from Bule *et al.*, 2018<sup>14</sup>).

Aerobic microorganisms secrete large quantities of CAZymes to the environment, organized in much simpler non-integrating systems. Bacteria from genera *Bacillus*, *Micromonospora*, *Cellvibrio* and *Pseudomonas*<sup>20</sup> and fungi from genera *Aspergillus*<sup>21</sup> are examples of aerobic cellulolytic microorganisms. The enzymatic activities of the CAZymes are still complementary, showing strong synergy in the degradation of plant cell walls.

It is worth emphasizing in this section the recently identified lytic polysaccharide monooxygenases (LPMO) as key players in the first steps of plant biomass degradation<sup>19</sup>. LPMOs are



copper-dependent enzymes which cleave crystalline substrates by oxidizing the glycosidic bonds. By introducing chain breaks in insoluble polysaccharides, such as cellulose and chitin, and also in some hemicelluloses, LPMOs have the ability to enhance the activity of the glycoside hydrolases.

CAZymes and CBMs are classified into sequence-based families in the CAZy database ([www.cazy.org](http://www.cazy.org))<sup>22</sup>. This database, with regular updates, is dedicated to classifying and analyse genomic, structural and biochemical information concerning CAZymes and associated CBMs involved in the synthesis, modification and breakdown of oligo- and polysaccharides<sup>22</sup>. Currently (as of March 2020), there are numerous different CAZymes and CBM families identified: 167 families of glycoside hydrolases (GHs), 110 families of glycosyl transferases, 40 families of polysaccharide lyases (PLs), 17 families of carbohydrate esterases (CEs), 16 families of auxiliary activities (including the 6 LPMO families), and 86 families of CBMs.

### 1.2.1 Carbohydrate-binding modules: the non-catalytic domains associated with Carbohydrate Active enZymes

CBMs are a class of carbohydrate-binding proteins, defined as non-catalytic protein domains, with amino acid sequences ranging from 30 to 200 amino acids<sup>23,24</sup>. These modules were initially defined as cellulose-binding domains (CBDs), as the first examples of CBMs mainly bound to crystalline cellulose<sup>25</sup>. However, these modules show a highly diverse range of ligand specificities, between different families and even within the same family. Several characterized CBMs recognize non-crystalline cellulose, chitin, xylan, mannan, galactan, soluble  $\alpha$ - and  $\beta$ -glucans and insoluble storage polysaccharides, such as starch and glycogen<sup>22</sup>.

The number of newly identified CBM sequences with putative carbohydrate binding is growing fast due to the exponential increase of sequence information derived from microbial genomics, metagenomics and transcriptomics data. Many of these proteins await elucidation and assignment of a carbohydrate-binding function<sup>26,27</sup>.

#### 1.2.1.1 Classification of CBMs

Based on the conservation of protein fold, CBMs are divided into 7 different fold families. Most CBMs identified to date are classified in the  $\beta$ -sandwich family of protein folds. To provide additional functional relevance to the CBM classification, these modules have been grouped into three types: A, B, and C, according to the mode of interaction with the carbohydrates and the architecture of the binding site<sup>23,24</sup>.

CBMs from type A have a planar hydrophobic surface decorated by aromatic residues that interact with flat crystalline polysaccharides, such as chitin or cellulose. This type of interaction is observed in the crystal structure of a CBM63-containing *Bacillus subtilis* expansin in complex with  $\beta$ 1,4-D-linked cellohexaose<sup>28</sup>. Some type A CBMs have been reported to bind not only to

crystalline cellulose but also to soluble polysaccharides such as CBMs from family 2 and 3, which can also bind xyloglucans<sup>29</sup>, and CBM64 from *Spirochaeta thermophila* that binds a variety of hemicelluloses<sup>30</sup>.

CBMs from type B or endo-type are classified as CBMs that bind to internal oligosaccharide sequences. These CBMs exhibit a cleft or groove, that accommodates oligosaccharide chains with four or more residues, and show higher binding affinities with the increase of the oligosaccharide chain length.

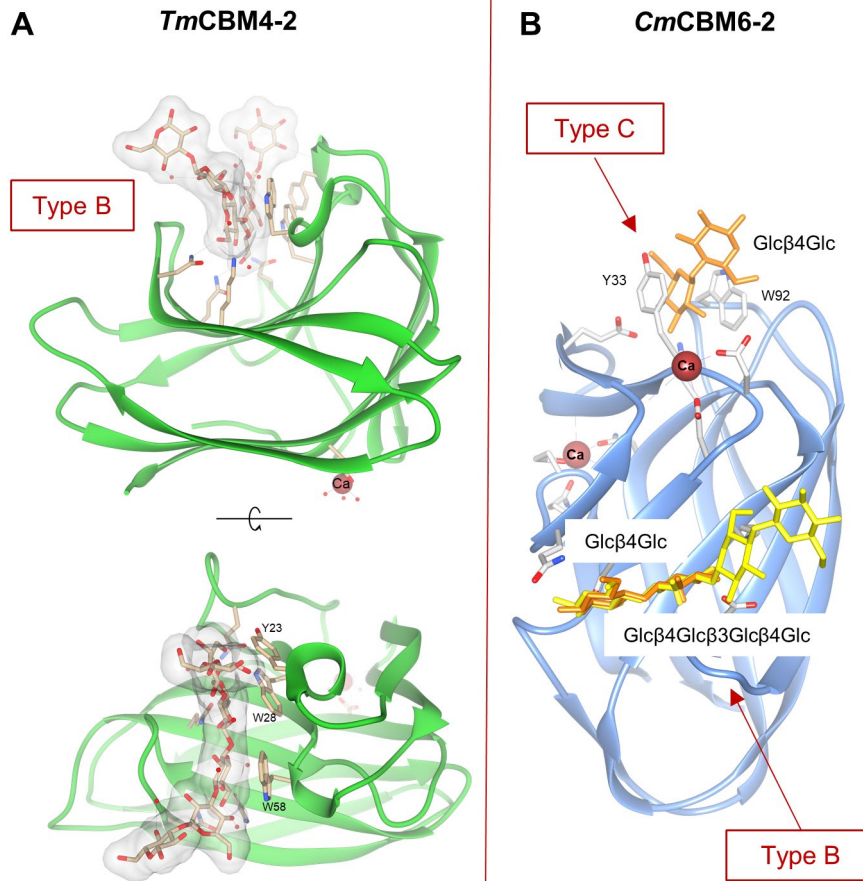
One example is the highly thermostable family 4 CBM from *Thermotoga maritima* that binds  $\beta$ 1,3-glucans and  $\beta$ 1,3-1,4-glucans<sup>31,32</sup> (Figure 1.4A). This CBM comprises the C-terminus of the putative laminarinase Lam16A and the recognition of  $\beta$ 1,3-D-linked laminarioligosaccharides involves three tryptophan residues (W28, W58, and W99) and one tyrosine residue (Y23). One other example is the family 11 CBM from *C. thermocellum*, which recognizes mixed-linked  $\beta$ 1,3-1,4-glucans<sup>32-34</sup> (Chapter 4 of this Thesis is dedicated to the understanding of the molecular determinants of this CBM specificity). This CBM is associated to the enzyme Lic26A-Cel5E, an enzyme that contains GH5 and GH26 catalytic domains that display  $\beta$ 1,4-glucan and  $\beta$ 1,3-1,4-glucan endoglucanase activity, respectively<sup>33</sup>.

CBMs from type C, or exo-type, are classified as CBMs that recognize the non-reducing end of an oligosaccharide sequence, binding in an optimal way to mono-, di- or trisaccharides, due to steric restriction in the binding site. Unlike the type B CBMs, type C do not contain the extended grooves in the binding-sites. Examples of this CBM type are the family 6 CBM from *Bacillus halodurans* in complex with laminarihexaose<sup>35</sup>, the family 42 CBM from *C. thermocellum*, a  $\beta$ -trefoil lectin that binds the arabinose side chains of complex hemicelluloses<sup>36</sup>, and family 13 CBM from *Streptomyces lividans*, another  $\beta$ -trefoil binding xylose or xylo-oligosaccharides<sup>37</sup>.

Remarkably, the family 6 CBM from *Cellvibrio mixtus* exhibits a type B cleft capable of recognizing  $\beta$ 1,4-,  $\beta$ 1,3-, and  $\beta$ 1,3-1,4-linked glucose oligosaccharides, and a type C cleft that interacts with terminal residues of  $\beta$ 1,4- and  $\beta$ 1,3-linked glucose and also  $\beta$ 1,4-xylose oligosaccharides<sup>38,39</sup> (Figure 1.4B).

### 1.2.1.2 Functional Roles of CBMs

Although, in general, CBMs display low affinity towards their target ligands (in the  $\mu$ M-mM range)<sup>23</sup>, modular enzymes increase their avidity by containing multiple copies of CBMs, enhancing the affinity for their target polysaccharide<sup>26</sup>. Although the mechanism by which CBMs potentiate catalysis remains elusive<sup>26</sup>, the following four functional major roles are recognized for CBMs<sup>23,24,26</sup>: 1) Targeting function, where CBMs target the joined catalytic modules to specific regions on the carbohydrate substrate, such as reducing end, non-reducing end or internal polysaccharide chain, bringing the enzyme into close and prolonged contact with the target ligand; 2) Proximity effect, where CBMs increase the concentration of enzyme near its substrate, leading



**Figure 1.4. Carbohydrate recognition by CBMs.** (A) Type B protein-carbohydrate interactions illustrated by 2 views of the 3D structure of family 4 CBM from *Thermotoga maritima* in complex with  $\beta$ 1,3-D-linked laminarihexaose (PDB ID: 1GUI<sup>31</sup>). This type of CBMs displays a cleft arrangement in which the binding site accommodates glycan chains with four or more monosaccharide units. (B) Family 6 CBM from *Cellvibrio mixtus* exhibits a type B cleft (in complex with  $\beta$ 1,3-1,4-linked glucose tetrasaccharide (PDB ID: 1UZ0) and a type C cleft (in complex with  $\beta$ 1,4-linked glucose disaccharide) (PDB ID: 1UYX)<sup>38</sup>. Representations (not to scale) of individual 3D structures were done with program Chimera<sup>40</sup> using the PDB atomic coordinates.

to a more rapid and efficient carbohydrate degradation; 3) Disruptive function, where CBMs act to disrupt the surface of tightly packed polysaccharides, such as crystalline cellulose fibres and starch granules, causing them to become more exposed to the catalytic module and, hence, increasing the degradation efficiency; and 4) Cell attachment, where CBMs adhere enzymes onto the surface of bacterial cell wall components, while exerting catalytic activity on an external carbohydrate substrate.

Several organisms possess CBMs that are not directly involved in plant cell wall degradation and have been found to perform other functions, acting in isolated or tandem forms. These modules can act as potential carbohydrate-sensors (or sensing domains) of the biomass availability in the extracellular medium, and examples are *C. thermocellum* CBMs from families 3 and 42 that trigger the expression of the enzymatic machinery specific for the degradation of the detected polysaccharides<sup>13,41</sup>. Additionally, certain CBMs, also known as Lysin motif domains or LysM domains, are involved in signalling between bacteria and plants, but can also be found in fungal

proteins acting as host immunity modulators, or be involved in the development of bacterial spores<sup>42</sup>.

Given the wide variety of ligand specificities, CBMs are an excellent model to study protein-carbohydrate recognition mechanisms. Additionally, these modules are interesting candidates for various applications in biotechnology, and some specific examples will be referenced in the next section.

### 1.2.2 Biotechnological applications of carbohydrate-binding proteins

Plant cell wall polysaccharides present major potential for biological and biotechnological applications, from paper and textiles industries, biotransformation of lignocellulosic materials into biofuels and other renewable products of biorefinery, and as sources of dietary fibres for both human and animal nutrition. In the past decades, the use of microbial and enzymatic systems to overcome the difficult deconstruction of such recalcitrant polysaccharides, yielding efficient and low-cost mechanisms, have gained importance<sup>43–45</sup>.

Not long after the identification of cellulosomes, their potential applications to biotechnology started to be explored<sup>46</sup>. Given their vast array of CAZymes and extreme habitat variability, cellulosomes can be engineered as a multi-functional protein complex tool<sup>12,45,47</sup>. Chimeric cellulosomes or designer cellulosomes have been used, for instance, as potential replacements or extensions of native cellulosomes to produce biofuels from cellulosic biomass<sup>48</sup>.

Additionally, each protein module in the cellulosome can be explored, individually or combined with free-acting CAZymes, for a wide spread of industrial and biotechnological applications<sup>47</sup>. Synergistic actions between individual enzymes have been exploited to produce a combination of feed enzymes with the ability to degrade *Chlorella vulgaris* cell wall, with the purpose of improving the bioavailability of its valuable nutritional compounds for monogastric animal diets and facilitate the cost-effective use of microalgae by the feed industry<sup>44</sup>.

CBMs have also been largely explored for many biotechnological applications. Given their vast diversity, the ability to function autonomously in chimeric proteins and the controllable binding specificities so that the right solution can be adapted to an existing problem, make CBMs attractive candidates for a variety of applications<sup>49,50</sup>. From enhancers in biomass degradation<sup>51</sup>; improvement of cellulose fibre properties for the paper and textile industries; functionalization of biomaterials in biomedicine; production, purification and immobilization of recombinant proteins; in food industry for the improvement of animal feed nutritional value; and as molecular probes for protein-carbohydrate interactions<sup>50</sup>.

As molecular probes, CBMs are valuable tools for the study of plant cell wall architecture. Recombinant CBMs have been used for characterizing native complex carbohydrates and engineered biomaterials<sup>50</sup>. CBMs from CAZy families 2a, 6, and 29 containing poly-histidine tags have been used for the analysis and detection of polysaccharides in maize coleoptiles cell walls,

such as crystalline cellulose, xylans, galactomannan and glucomannan<sup>52</sup>. Two recombinant fluorescent CBM-probes consisting of a green fluorescent protein fused with a CBM3, that binds to both amorphous and crystalline cellulose, and mono-cherry fluorescent protein fused with a CBM17, that only binds to amorphous cellulose, have been used to reveal the surface accessibilities of amorphous and crystalline celluloses in Avicel<sup>53</sup>.

Given the increasing repertoire of CBMs, and their various potential and valuable applications, understanding the structural basis of their ligand specificity is of the most importance.

### 1.2.3 *Clostridium thermocellum* and *Ruminococcus flavefaciens* FD-1

The work developed in this Thesis focused on CBMs from *Clostridium thermocellum* (*C. thermocellum*) and *Ruminococcus flavefaciens* FD-1 (*R. flavefaciens*), two Gram-positive, anaerobic, cellulosome-producing cellulolytic bacteria that reside in different ecological niches, and of relevance in fields such as of bioprocessing and animal nutrition. *C. thermocellum* is a thermophilic bacterium<sup>54</sup>, found mostly in soils and hot springs and it is considered the most efficient cellulolytic microorganism for the degradation of lignocellulosic biomass<sup>55</sup>. *C. thermocellum* cellulosome (Figure 1.3A) was the first to be identified and characterized<sup>15</sup> (see section 1.2), and its complex architectural components have been extensively studied<sup>12,56,57</sup>. On its turn, *R. flavefaciens* species are found in the digestive tracts of ruminants, other herbivorous animals and humans, and are among the most important ruminal cellulolytic bacteria, considered to be primarily responsible for plant cell wall biodegradation in the rumen<sup>58</sup>. Recent sequencing of *R. flavefaciens* FD-1 genome has revealed one of the most complex cellulosomes (Figure 1.3B) identified to date, with one of the largest collection of cellulosome-associated proteins among known fibre-degrading bacteria<sup>14,27</sup>. Besides the numerous modular GHs, several CBMs were identified to belong to known CAZy families and also putative CBM sequences<sup>27</sup>, some of which have recently been classified into new families 75 to 80<sup>59</sup>.

Throughout the years, efforts were made to elucidate the molecular mechanism for the assembly of cellulosomes, enabling the identification of its different structural components. As such, *C. thermocellum* CBMs have been widely studied and characterized, yielding a vast amount of available information, with up to 90 CBMs from diverse families deposited in the CAZy database. Despite this, there are still numerous *C. thermocellum* CBMs that await elucidation and experimental validation. In contrast, given the more recent *R. flavefaciens* FD-1 genome sequencing, there is not much information available yet. Given the crucial roles of CBMs in plant cell wall biodegradation, the elucidation of the carbohydrate-specificities of the numerous assigned *R. flavefaciens* FD-1 CBMs and the understanding of mechanisms of carbohydrate recognition will promote the knowledge of this bacterium cellulolytic capabilities.

### 1.3 Methods for characterizing protein-carbohydrate interactions

Understanding the molecular basis for carbohydrate recognition by proteins and the relationship between structure and function are prerequisites for the development of future innovations in biological, biotechnological and industrial fields<sup>60</sup>. To this end, several state-of-the-art techniques have been developed for the characterization of protein-carbohydrate interactions.

Characterization of protein-carbohydrate recognition often uses a combination of analytical methods, that can provide insights into the biological roles of each protein and their ligands, with structural studies, to understand the mechanisms of ligand-binding at molecular level. Carbohydrate microarrays<sup>61</sup> allow to decode carbohydrate recognition by screening proteins for ligand-binding and specificity. Other analytical methodologies, such as Isothermal Titration Calorimetry (ITC)<sup>62</sup>, MicroScale Thermophoresis (MST)<sup>60</sup>, Enzyme-Linked Immunosorbent Assay (ELISA)<sup>63</sup> and Affinity Gel Electrophoresis (AGE)<sup>62</sup> can then be used to assess protein-carbohydrate affinity to target poly- and oligosaccharides. Structural characterization of the protein-carbohydrate recognition can be achieved by determining the molecular structures of protein-carbohydrate complexes, and X-ray crystallography<sup>64</sup> and Nuclear Magnetic Resonance (NMR) spectroscopy<sup>60</sup> are the methods of choice, which can be used in a complementary manner. High resolution data can also be integrated with lower resolution results from Cryo-Electron Microscopy (Cryo-EM) and Small Angle X-ray Scattering (SAXS) when characterizing large multi-component complexes<sup>65</sup>. Additionally, computational methodologies, such as Molecular Dynamics (MD) and molecular docking<sup>66</sup> are advantageous tools to complement the experimental data.

In the next sections emphasis will be given to the methodological aspects and application of carbohydrate microarrays and X-ray crystallography, as these were the two major techniques applied to achieve the objectives of the proposed Thesis work plan.

#### 1.3.1 Carbohydrate microarrays

The development of carbohydrate microarrays in the recent decades came to revolutionize the study of carbohydrate-protein interactions, satisfying the high demand for high-throughput methods to systematically array carbohydrate libraries and identify the specificity and biological role of carbohydrate-binding proteins<sup>67-73</sup>.

The main advantage of the microarray technology is that a wide diversity of carbohydrate probes can be immobilized on a microarray surface and simultaneously assessed for binding events, using only minute amounts of samples. This miniaturization feature of the microarrays, takes the most out of precious materials, both carbohydrates and protein analytes, while generating a large amount of information on a variety of carbohydrate-recognising systems<sup>74,75</sup>. Another important feature of the microarrays is the multivalent display of the arrayed carbohydrates, which enables

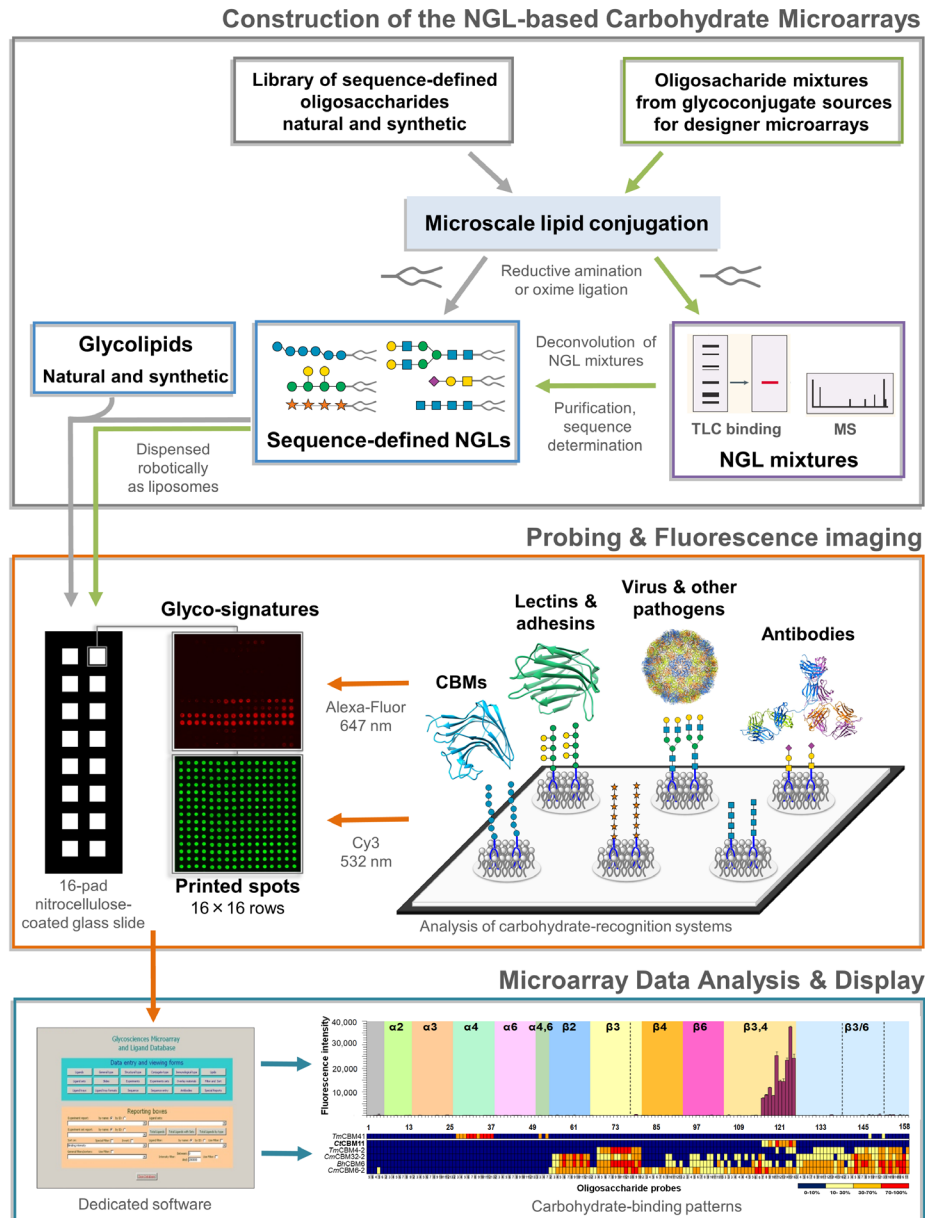
the optimisation of carbohydrate presentation for interaction with the protein and promotes detection of very low affinities of carbohydrate-protein interactions.

Carbohydrate microarrays are generally of two categories: polysaccharide or glycoprotein microarrays and oligosaccharide microarrays<sup>74</sup>. The carbohydrate samples can either be isolated from natural sources or be chemically or chemo-enzymatically synthesized. On the one hand, polysaccharide or glycoprotein microarrays can comprise the full diversity of a particular glycome and may avoid the loss of any labile or conformational determinants during the release of the oligosaccharides from their molecules of origin<sup>68,69,71,74</sup>. On the other hand, oligosaccharide microarrays are powerful tools to assess binding-specificity in carbohydrate recognition events and to identify the binding epitopes<sup>71,74</sup>. Polysaccharides and glycoproteins can be readily and randomly immobilized on solid matrices based on hydrophobic physical adsorption or charge-based interaction<sup>68,69,74</sup>. The immobilization of oligosaccharides is more challenging, given their low mass and hydrophilic nature, and chemical derivatization at the reducing monosaccharide is usually required prior to immobilization, to introduce suitable functional groups<sup>74–76</sup>. Being in equilibrium between the hemiacetal closed-ring and the aldehyde open-chain form, the reducing monosaccharide can serve as an electrophilic group for a chemoselective reaction with numerous nucleophilic amine-, hydrazide-, or oxyamine-containing reagents<sup>74–76</sup>. For oligosaccharides obtained through chemical synthesis, the functionality is generally carried out by placing a linker at the reducing terminal monosaccharide residue in a form suitable for flexible modifications<sup>74–76</sup>. These different strategies allow oligosaccharides to be immobilized on a compatible microarray surface. Polymer-based surfaces such as nitrocellulose or plastic are usually an attractive solid surface for non-covalent immobilization<sup>67–69,72</sup>, whereas gold or functionalised glass are used to covalently attach carbohydrate probes<sup>70,77</sup>.

### 1.3.1.1 Carbohydrate microarray platforms

To date, several carbohydrate microarray platforms have been developed that: 1) use alternative chemical strategies to overcome the limitation of direct immobilization of oligosaccharides onto solid matrices; 2) differ on the type of carbohydrates and how they are displayed on the array surface; and 3) are based on covalent or non-covalent immobilization to different surfaces. These are reviewed in detail in recent references<sup>73,75,76,78–81</sup>. Here, some high-throughput platforms that contain a high diversity of oligosaccharide probes and that use different strategies for their immobilization and presentation will be highlighted.

Feizi and colleagues have developed a microarray system based on the neoglycolipid (NGL) technology<sup>82</sup>, in which the oligosaccharides are linked to a lipid<sup>67,71,72,78,83,84</sup>. A summary of the key steps involved in the construction of the NGL-based microarrays and their analysis to reveal carbohydrate-binding patterns is depicted in Figure 1.5. Central to this microarray system is the microscale lipid conjugation of oligosaccharides (natural or chemically synthesized sequence-defined or as mixtures) to an aminophospholipid to produce NGLs. The generated NGL

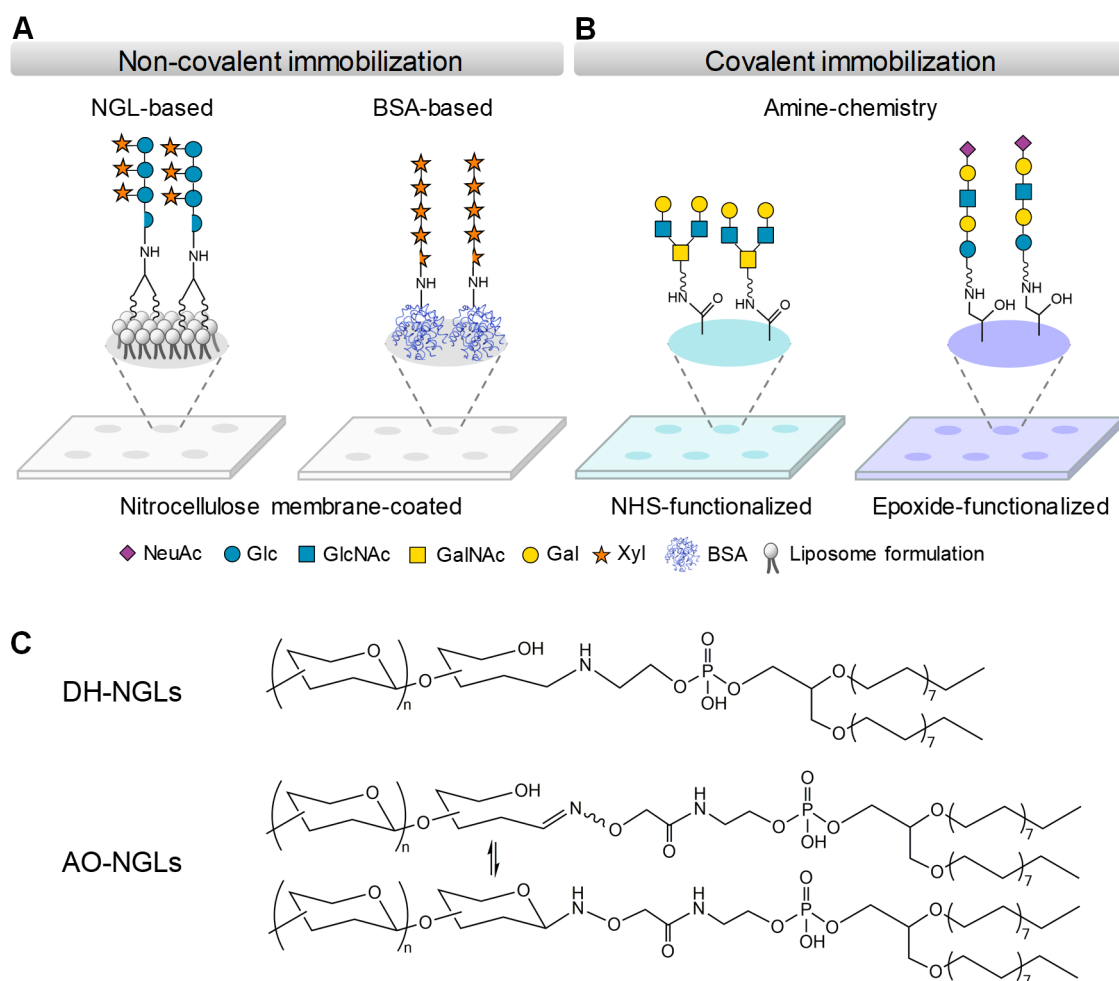


**Figure 1.5. Schematic overview of the main steps comprising the analysis using neoglycolipid (NGL)-based carbohydrate microarrays.** The key steps of this methodology are depicted in three stages: **1) Construction of the NGL-based carbohydrate microarrays:** the probes are all lipid-linked and comprise both NGLs prepared from natural or chemically synthesized oligosaccharides and glycolipids, natural or synthetic; the interface with mass spectrometry (MS) and HPTLC or HPLC, enables purification and characterization of oligosaccharides or NGL mixtures; the NGLs can also be prepared from oligosaccharide mixtures derived from ligand-bearing glycomes, to reveal and characterise the oligosaccharide ligands they harbour ('designer' microarray methodology); the NGL and glycolipid probes are robotically dispensed onto nitrocellulose-coated glass slides using a liposome formulation in the presence of carrier lipids. **2) Probing and Fluorescence imaging:** Cyanine 3 fluorophore is included in the NGLs liposome formulation, so that the immobilized probes can be visualised by fluorescence imaging at 532 nm; the microarrays can then be probed for carbohydrate-binding by monoclonal antibodies, CBMs, lectins and viruses or other pathogens; the binding signals are revealed by scanning for Alexa-fluor647 emission at 647 nm. **3) Microarray data analysis and display:** the fluorescence intensities are quantified and analysed to reveal the carbohydrate binding patterns using a dedicated software, which includes a database that holds all of the microarray data and metadata on experimental conditions and information on probes and proteins; interactive tools are then used for semi-automatic presentation of microarray data by filtering, sorting and deep mining every data point. Figure adapted from Palma *et al.* 2014<sup>78</sup> and Palma *et al.* 2015<sup>32</sup>.



probes have amphipathic properties, which enables efficient display onto nitrocellulose-coated glass slides using a liposome formulation in the presence of carrier lipids<sup>85</sup> (Figures 1.5 and 1.6A). The reducing oligosaccharides can be conjugated through reductive amination to the aminolipid 1,2-dihexadecyl-*sn*-glycero-3-phosphoethanolamine (DHPE, DH-NGLs) (Figure 1.6C). This procedure yields the ring-opening of the monosaccharide at the reducing end<sup>74</sup>. To overcome this limitation, NGLs with ring-closed monosaccharide cores have been introduced by Liu and colleagues<sup>86</sup>. These are prepared by conjugating reducing oligosaccharides to an aminoxy-functionalized DHPE by oxime ligation (without reduction) (AOPE, AO-NGLs) (Figure 1.6C). This procedure enables the efficient presentation of short oligosaccharides for direct binding assays<sup>86</sup>. The non-covalent immobilization of NGLs in a lipid environment onto a nitrocellulose surface introduces an element of mobility. This mode of presentation simulates to some extent the cell surface display of glycans and may be advantageous for detection of binding for particular recognition systems<sup>78</sup>. The NGL-based microarray system currently contains a repertoire of around 900 sequence-defined probes, with a high content of natural oligosaccharide sequences, including NGLs derived from various oligosaccharides of mammalian sources, from polysaccharides of bacterial, fungal, and plant origins, and natural and synthetic glycolipids<sup>78</sup> (accessed through the link <https://glycosciences.med.ic.ac.uk/glycanLibraryIndex.html>).

The microarray platform of the Consortium for Functional Glycomics (CFG) developed by the early work of Blixt and colleagues<sup>70,87</sup> is also based upon amine chemistry, whereby oligosaccharides linked at the reducing end with an amine-terminating linker are covalently immobilized onto *N*-hydroxysuccinimide (NHS) ester-derivatized glass slides (Figure 1.6B). Recent microarray versions are composed of around 600 mammalian-type probes (mammalian printed array version 5.2). Other strategies that also use an amino-linker involve immobilization of the amine-terminated oligosaccharides onto epoxide-derivatized slides (Figure 1.6B). Examples are by Cummings and colleagues that used this method for immobilization of naturally-derived oligosaccharide libraries<sup>88</sup> and by Varki and colleagues who developed a structurally diverse microarray of sialylated oligosaccharides<sup>89</sup>. Gildersleeve and colleagues demonstrated that oligosaccharides conjugated to bovine serum albumin (BSA) or human serum albumin (HSA) (displayed as neoglycoproteins) may also be efficiently immobilized using amine chemistry onto epoxide functionalized glass slides for binding studies<sup>90</sup>. Other groups, have developed covalent microarrays based on different chemistries, such as the early work by Shin and colleagues using the thiol chemistry, whereby maleimide-functionalized oligosaccharides are immobilized onto thiol-derivatized slides<sup>91</sup>. In these covalent oligosaccharide microarray platforms, the nature and length of the linkers between the oligosaccharide and the array surface are important for accessibility of the oligosaccharide to the protein and detection of specific binding.



**Figure 1.6. Graphic representation of examples of immobilization strategies used to generate carbohydrate microarrays.** (A) Non-covalent microarrays: immobilization onto nitrocellulose-coated glass slides of reducing oligosaccharides derivatized by reductive amination to an aminophospholipid, to prepare neoglycolipids (NGLs)<sup>78</sup>, or to BSA, to prepare neoglycoproteins<sup>92</sup>. (B) Covalent microarrays: immobilization of synthetic oligosaccharides derivatized at the reducing end to an amino-terminating linker onto *N*-hydroxysuccinimide (NHS)-functionalized glass slides<sup>70</sup> or onto epoxide-functionalized glass slides<sup>89</sup>. (C) NGL probes prepared from reducing oligosaccharides by reductive amination (DHPE, DH-NGLs)<sup>93</sup> and by oxime ligation (AOPE, AO-NGLs)<sup>86</sup>; the derivatization of the oligosaccharide by oxime ligation produces an equilibrium between the open- and closed-ring form of the reducing monosaccharide<sup>86</sup>. Examples of carbohydrate structures in the different libraries are shown using the symbol nomenclature for glycans (SNFG) according to Varki *et al.*, 2015<sup>94</sup>.

The NGL-based microarray facility and that of the Consortium of Functional Glycomics (accessed through links <https://www.imperial.ac.uk/glycosciences/> and <http://www.functionalglycomics.org/>, respectively) are the two largest platforms assembled to date that are open to the broad scientific community for microarray screening analyses of carbohydrate-binding proteins in different biological contexts.

Although the number of sequence-defined probes in carbohydrate microarrays has been expanding, the increase in diversity to date has been mainly on mammalian-type sequences. Some groups, however, have focused on development of microarrays from microbial<sup>32,95,96</sup> or plant-derived carbohydrates<sup>32,92,97–102</sup>. Sequence-defined oligosaccharides can be derived from

natural polysaccharides and the development of methods for fine-tuned depolymerisation, purification, high-sensitive sequencing and structural characterisation of the oligosaccharide fragments are crucial<sup>32,92</sup>. Methods for chemical<sup>103–105</sup> or chemo-enzymatic synthesis<sup>106</sup> of structural elements from complex microbial or plant cell wall polysaccharides offer powerful complementary approaches to develop sequence-defined microarrays. Recently, Seeberger *et al.* combining different carbohydrate synthesis approaches including automated glycan assembly, solution-phase synthesis and chemoenzymatic methods, successfully obtained a library of over 300 structures of different microbial oligosaccharides, which were used to develop the most diverse microbe-focused carbohydrate microarray platform to date<sup>95</sup>. The genetic engineering of bacterial strains with specific CAZymes gene deletions to produce oligosaccharides in the presence of a target substrate<sup>107</sup>, offer an alternative approach to achieve the much needed structural diversity.

In the context of the work developed in this thesis, some selected examples of microarray approaches to study plant-carbohydrate recognition by CBMs will be highlighted in the sections below.

### 1.3.1.2 Microarrays focused on plant carbohydrates for recognition studies

Early work by Willats and colleagues, reported on a carbohydrate-based approach for high-throughput plant polysaccharide cell wall profiling<sup>99</sup>. This Comprehensive Microarray Polymer Profiling (CoMPP) is based on the extraction of *Arabidopsis thaliana* and *Physcomitrella patens* polysaccharides and printing of the polysaccharide-rich fractions onto nitrocellulose-based arrays. These are then probed with CBMs and monoclonal antibodies of known specificities for plant cell wall polysaccharides. The CoMPP strategy enables the plant cell wall composition to be assessed in a semi-quantitative high-throughput way by revealing the relative abundance of polysaccharide epitopes<sup>99</sup>. More recently, using the same principle of arraying chemically extracted polysaccharides and the information of monoclonal antibodies, Waldron and colleagues have analysed quantitatively the abundance of different non-cellulosic polysaccharides in 331 genetically different *Brassica napus* cultivars<sup>100</sup>. These studies are providing insights to plant cell wall biosynthesis and restructuring<sup>100</sup>. This high-throughput screening of polysaccharide structures requires the use of proteins for which carbohydrate-specificity is known. Thus, development of sequence-defined plant-based carbohydrate microarrays is highly important to provide these protein tools.

Later on, Willats and colleagues have developed a suitable platform for high-throughput analysis of the specificities of CBMs and monoclonal anti-carbohydrate antibodies<sup>92</sup>. This microarray is composed of linear and branched oligosaccharides, either isolated from polysaccharides, such as glucans, xylans, mannans, galactans, xyloglucans or arabinans, using enzymatic or chemical hydrolysis or generated by chemical synthesis<sup>92</sup>. The oligosaccharides are coupled to BSA by reductive amination, producing a ring-opened monosaccharide at the reducing end of the

oligosaccharide (Figure 1.6A)<sup>92</sup>. These neoglycoproteins are arrayed non-covalently together with plant polysaccharides onto a solid matrix, such as nitrocellulose-coated glass slides. The developed microarrays were recently used to characterize unknown CBMs of *R. flavefaciens* FD-1 cellulosome, revealing six previously unidentified CBM families targeting  $\beta$ -glucans,  $\beta$ -mannans and pectic homogalacturonan<sup>59</sup>. This study was important to gain knowledge on the complexity of the *R. flavefaciens* cellulosome and its extended repertoire of CBMs for efficient plant cell wall degradation in absence of CBMs that target cellulose. More recently, the sequence diversity in these microarrays was expanded to contain linear, branched and phosphorylated  $\alpha$ 1,4-D-linked glucose maltooligosaccharides<sup>108</sup>. These were applied to characterize the starch binding domain CBM20 of *Aspergillus niger* as tool for high-throughput screening of starch structures during development and germination<sup>108</sup>.

Microarray platforms comprised of synthetic plant-based oligosaccharide sequences have also been developed in more recent years. Pfrengle and colleagues, have constructed microarrays of plant cell wall oligosaccharides obtained by solution-phase synthesis and by automated glycan assembly, comprising a range of sequences from xylans, glucans, xyloglucans, galactans, arabinogalactans and pectins<sup>109,110</sup>. The synthesized oligosaccharides equipped with an aminoalkyl-linker at the reducing end, were printed onto NHS-functionalized glass slides. These microarrays allowed to determine the binding epitopes of 79 plant cell wall carbohydrate-directed antibodies, and can also be used to identify and characterize unknown glycoside hydrolases substrate-specificities<sup>110</sup>.

### 1.3.1.3 Combining microarray analysis with mass spectrometry

A key feature of the NGL technology is its interface with Mass Spectrometry (MS) and High-Performance Thin Layer Chromatography (HPTLC) or High Performance Liquid Chromatography (HPLC)<sup>74,78</sup> (Figure 1.5). This enables bioactive oligosaccharides released from a glycome source to be resolved from heterogeneous mixtures, characterized and purified, allowing the discovery and characterization of novel ligands of biological relevance<sup>67,71,78,85</sup>. Based on their previous work, which used this 'designer' approach from ligand-bearing glucans to assign the oligosaccharide ligands for the immune receptor Dectin-1<sup>72</sup> and anti-fungal therapeutic antibodies<sup>74</sup>, Palma and colleagues have developed a sequence-defined 'glucome' microarray as a screening tool for glucan-binding proteins<sup>32</sup>. The glucome microarray comprised 153 gluco-oligosaccharide probes, with diverse sequences and chain lengths representing major sequences in glucans, including those present in plant cell wall. The oligosaccharides were prepared by depolymerization of glucans and multiple chromatographic methods at microscale or were synthesized chemically. The linkage and sequence determination of the linear and branched oligosaccharides were determined at high-sensitivity by development of a negative-ion Electrospray Ionization Collision-Induced Dissociation Mass Spectrometry (ESI-CID-MS/MS) method for gluco-oligosaccharides<sup>32</sup>. The oligosaccharides were converted to NGL probes by

oxime ligation to an aminoxy-functionalized lipid (AO-NGLs)<sup>86</sup> (Figure 1.6C). The optimization of this method enabled long chains of gluco-oligosaccharides, otherwise difficult to derivatize, to be displayed in the microarrays for interaction studies<sup>86,96</sup>.

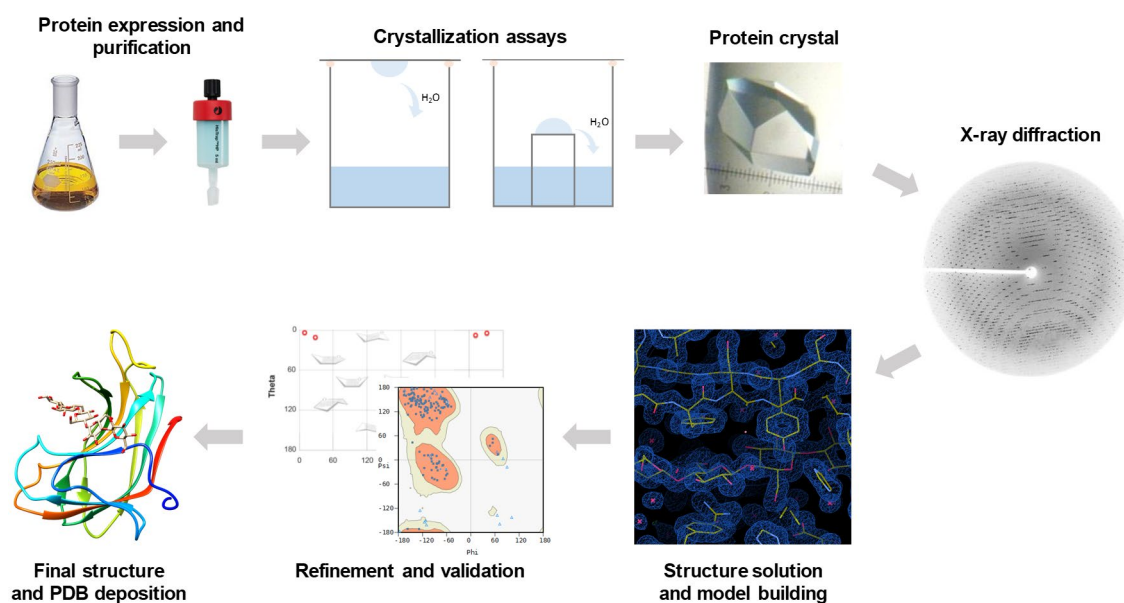
Combining the microarray analysis with MS sequencing enabled the high sensitivity detection and unambiguous assignment of specificity of glucan recognition. The purity of glucan polysaccharides of different structural types is of particular importance for assignment of specificity. The partial fragmentation of the polysaccharides, the sequencing of the oligosaccharides by the negative-ion ESI-CID-MS/MS method and their interrogation on the microarrays, not only provided detailed information on linkage, sequence and chain-length requirements of glucan-recognizing proteins, but also were a sensitive means of revealing unsuspected sequences in the polysaccharides<sup>32</sup>.

### 1.3.2 Protein crystallography

Once interactions between proteins and carbohydrates are analysed and the binding specifics are assigned, elucidation of the three-dimensional structures of their complexes is a prerequisite for a better understanding of the molecular basis underlying the recognition process and hence the relationship between structure and function. Macromolecular X-ray Crystallography is a method of choice for determining the structures of proteins and their complexes. Although consisting of a laborious and time-consuming process, with the increasing of computing power, allied to the development of modern molecular biology techniques, commercial screening solutions and dedicated instrumentation, X-ray crystallography has largely contributed to the elucidation of biologically relevant carbohydrate-mediated recognition events in the recent decades. For a more comprehensive understanding on biological crystallography methods, Rupp, 2009<sup>111</sup> is a reference textbook.

Determining the structure of proteins and protein-ligand complexes by X-ray crystallography, entails a series of methods and steps that determine the success of the resulting structure (Figure 1.7). Starting from a relatively large amount of a purified protein sample at an appropriate concentration, crystallization conditions are screened and eventually single crystals with quality suitable for X-ray diffraction are obtained. By irradiating the crystal with an X-ray beam, the resulting diffraction pattern reflects its composition and can be used to calculate an electron density map, upon calculating the phases. From the density map, an atomic model of the protein can be progressively built, refined and validated before its deposition in the Protein Data Bank (PDB)<sup>64,65,112</sup>.

As crystallization is a very time-consuming step and requires considerable amounts of protein, assessing purity, conformational stability and monodispersity of the protein sample prior to the crystallization trials is a good practice. Besides native and SDS-PAGE, techniques such as



**Figure 1.7. Schematic overview of the main steps comprised in the determination of protein and protein-ligand structures by X-ray crystallography, from the purified protein to the final three-dimensional structure.** As an example, *Cellulomonas fimi* CBM4 bound to  $\beta$ 1,4-D-glucose pentasaccharide is represented (PDB ID: 1GU3).

Differential Scanning Fluorimetry (DSF, also referred to as Thermofluor) or Dynamic Light Scattering (DLS) are routinely used<sup>64,65</sup>.

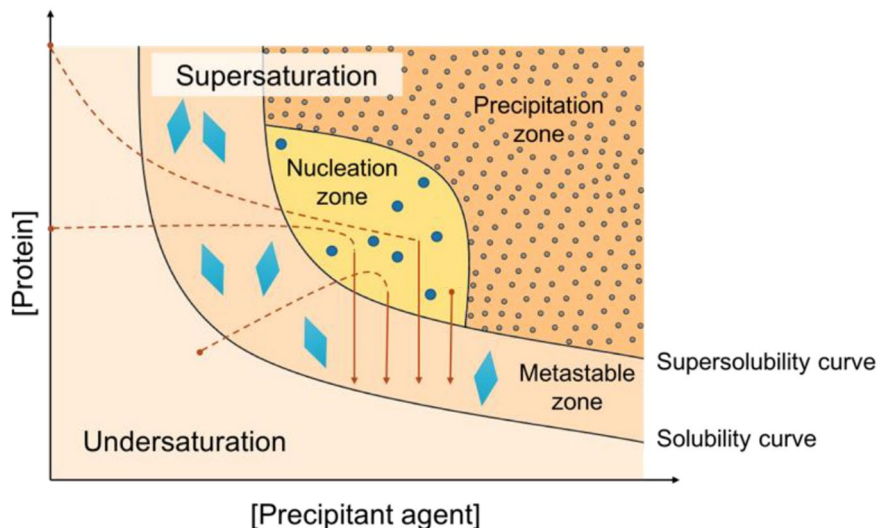
Additionally, when attempting to obtain protein-ligand complexes, the protein's capabilities to its putative ligands should be confirmed prior to the crystallization assays, especially when dealing with ligands difficult to obtain in large quantities, such as oligosaccharides. In section 1.3 above, some state-of-the-art biophysical methods that are usually employed to analyse protein-ligand binding are referred. This preliminary analysis allows to assess the experimental conditions, such as the affinity of the interaction, that will most likely lead to the complex formation and increase the chances of obtaining the desired crystals<sup>65</sup>.

One of the limiting steps in X-ray crystallography is obtaining well-ordered single crystals suitable for X-ray diffraction. This is especially true when trying to obtain protein-carbohydrate crystals. Due to the intrinsic high flexibility of carbohydrates, protein-oligosaccharide complexes can be challenging to crystallize<sup>113</sup>. Even when succeeding in obtaining crystals, quite often part or the whole ligand is not observed in the electron density map, even from high resolution diffraction data<sup>65,113</sup>.

### 1.3.2.1 Protein crystallization

In order to form crystals, protein molecules must separate from solution and self-assemble into a periodic crystal lattice structure. A comprehensive representation of the phenomenon of protein crystallization is usually depicted in terms of a phase diagram showing how the concentration of a specific protein in solution changes relative to the concentration of a precipitating agent

(Figure 1.8). The precipitating agent (or precipitant) is usually a solvent compound, such as polyethylene glycol, ammonium sulphate or sodium chloride, present in relatively high concentration in the protein solution. Initially, in the protein solution concentrated in the range of mg/mL, molecules are found in random orientation, surrounded by water and precipitant molecules. Interactions between solvent and protein are stronger than among protein molecules themselves. Once the solubility limit is exceeded, the solution becomes metastable. With concentration increase to the supersaturation phase, spontaneous or homogeneous nucleation occurs. During the nucleation stage, the solute protein molecules dispersed in the solvent start to gather into clusters and form stable nuclei. These clusters are stable only if they reach some critical size, which depends on physical conditions, such as supersaturation, temperature and pressure. Once this critical size is reached, the protein crystal will grow spontaneously as long as the solution is in the supersaturated state. This should be a slow process, to allow the protein molecules to assemble orderly in the crystal lattice and promote crystal growth. Crystal growth stops when the equilibrium is reached<sup>65,111</sup>.



**Figure 1.8. Phase diagram for protein crystallization.** The diagram contains a region of undersaturation and supersaturation. The supersolubility curve separates the condition where nucleation or precipitation spontaneously occur from the condition where the solution remains clear. Crystals can only grow from a supersaturated solution. The supersaturated region is divided in the metastable zone, where nuclei will grow into crystals; the nucleation zone, where nuclei will form; and the precipitation zone. (Adapted from Carvalho *et al.*, 2018<sup>65</sup>)

Solubility-reducing agents or precipitants are the primary component in this process. The pH of the crystallization cocktail, usually stabilized by a buffer, also determines the level of protein solubility, and shifts the local surface charge distribution of the protein. Temperature also affects protein solubility, which depending on the precipitant composition, may either increase or decrease with temperature. In addition, protein crystallization is an entropy-driven process, and the release of water molecules from across hydrophobic and polar residues during the crystal formation, contributes to the entropy gain of the system<sup>111</sup>.

Nowadays, automated crystallization nanodrop robots are used, which increase the number of conditions tested and require a smaller amount of protein and ligand, when compared to traditional manual crystallization procedures. Robots are the easiest way to screen over thousands of conditions thanks to the high diversity of crystallization formulations that are commercially available. In addition to precipitants and buffer solutions, these screenings can also include a diversity of additive compounds to help stabilizing intermolecular crosslinks in protein crystals and promote lattice formation<sup>64,65,111</sup>.

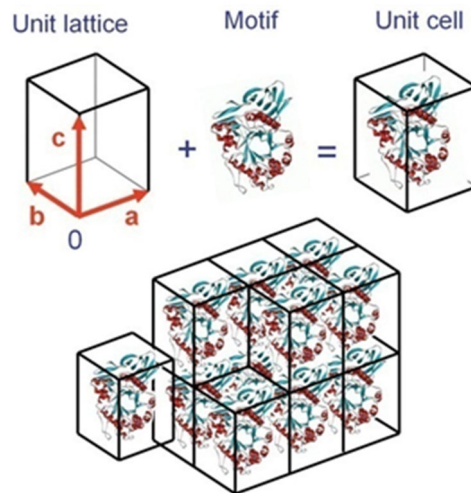
When aiming to obtain crystals of protein-ligand complexes, soaking and co-crystallization are the common methods used<sup>64,65</sup>. The soaking method involves incubation of crystals of the unliganded-protein with the ligand of interest for a time period (from seconds to days) that may require optimization. For co-crystallization, crystals are prepared starting from a solution of the protein pre-incubated with the ligand at a high molar ratio (5 to 10-fold excess). The latter is usually the method of choice when working with oligosaccharides. However, even when crystallization conditions are already well established for the unliganded-protein, obtaining crystals of the protein-ligand complex may not be as straightforward and optimization of the conditions may be required<sup>64,65</sup>.

Crystallization experiments are usually carried out through vapour diffusion using the hanging-drop or the sitting-drop method (Figure 1.7), in manual or automated setups. In this technique, a drop containing a mixture of the protein solution, previously incubated with the ligand in the case of co-crystallization, and the precipitant compound is placed in a sealed reservoir to equilibrate against the precipitant solution. Since the protein-precipitant mixture in the drop is less concentrated than in the reservoir solution, water evaporates from the drop and as a result, the concentrations of both protein and precipitant in the drop slowly increase until equilibrium is reached. As the protein solution in the drop becomes supersaturated, nucleation and phase separation occurs, and protein crystals may form<sup>64,111</sup>.

Crystals consist of periodic assemblies of fundamental building blocks – unit cells – orderly disposed in a rigid three-dimensional crystal lattice (Figure 1.9), capable of diverting X-ray photons. The unit cell is the smallest unit that can reproduce the whole crystal content by translations in the three-dimensional space, and can either be atoms, small molecules or whole proteins, forming a sparse network of weak intermolecular interactions. The smallest unit that can generate the whole unit cell, using the crystallographic symmetry operators, is defined as the asymmetric unit. Unit cells are described by three unique cell axes  $a$ ,  $b$  and  $c$ , and three unique angles between them,  $\alpha$ ,  $\beta$ , and  $\gamma$ . The specific relationship between these six parameters defines the crystal's lattice and the space group, which in turn will define the exact position of each spot (reflection) in the diffraction pattern produced by the crystal<sup>65,111</sup>.

Once crystals are formed, only a limited number of molecular interactions exist forming the network of weak intermolecular forces that keep the large protein molecules connected, mainly





**Figure 1.9. Assembly of unit cells in a three-dimensional crystal lattice.** Schematic representation of the assembly of unit cells to form a protein crystal, where each unit cell contains 1 copy of the asymmetric unit packed according to the space group's symmetry operations. **a**, **b** and **c** are the cell axes that define the angles  $\alpha$ , between **b** and **c**;  $\beta$ , between **a** and **c**; and  $\gamma$  between **a** and **b**. (Adapted from Rupp, 2009<sup>111</sup>)

ionic interactions and hydrogen bonds established between atoms from surface amino acids and, quite often, mediated by water molecules. These interactions are very specific, and inherent to each particular protein and have to take place at specific locations on its surface, in order to self-assemble the molecules into a well-formed, periodic crystal<sup>65,111</sup>. Proteins frequently present irregular shapes and dangling ends, and disordered termini or flexible loops, that are not easily stacked and assembled into a regular, periodic lattice. In addition to the weak intermolecular forces keeping the molecules together in the crystal, the substantial size of protein molecules, and low number of contacts per unit volume, also reduce the crystals' stability. Also, protein crystals contain on average 40-60% solvent, mostly disordered in large solvent channels between the stacked molecules, and also along plain rotation axes in the crystal structure<sup>65,111</sup>.

Once a single crystal is obtained, suitable for X-ray diffraction, it is harvested under the microscope and mounted on the X-ray diffractometer, or cryo-preserved to be later measured using synchrotron radiation. The minimum size of crystals needed for diffraction experiments range from 50 $\mu\text{m}$  to 0.5mm, but crystals of 20-50  $\mu\text{m}$  are considered amenable for synchrotron data collection in microfocus beamlines<sup>64,111</sup>.

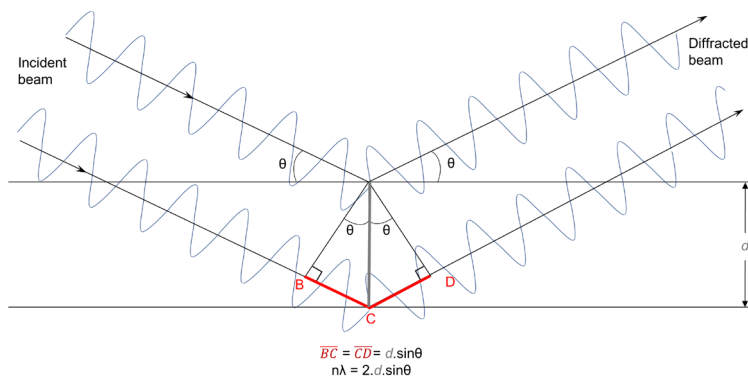
### 1.3.2.2 X-ray diffraction

X-rays are electromagnetic radiation ranging from 0.1 to 100 Å and can be produced by bombarding a metal target, most often copper, with an electron beam that is generated by a heated filament and accelerated by an electric field. As a result, a high-energy electron then collides with the metal target displacing an electron of the metal from a low-lying orbital and making an electron from a higher orbital to drop to the vacated one. This transition of the electron from an M-shell to a K-shell results in the emission of the excess electrons' energy in the form of an X-ray photon<sup>112</sup>.

By exposing a single crystal of a protein to a monochromated X-rays beam, the X-rays are scattered by the atoms present in the crystal and a set of diffraction spots, called reflections, is recorded on the detector originating the diffraction pattern. Each reflection spot in the diffraction pattern results from a monochromatic wave, constructively scattered by all equivalent lattice points that fulfil Bragg's Law (Equation 1.1 and Figure 1.10),

$$n \lambda = 2 d \cdot \sin \theta \quad (1.1)$$

that describes the relationship between the angle of the incident beam  $\theta$  and the wavelength  $\lambda$ , where  $d$  is the minimum spacing between equivalent planes in the crystal, i.e. the maximum resolution  $d_{min} = \lambda/2 \sin \theta_{max}$ .



**Figure 1.10. Bragg's Law defines the relationship between the angle of the incident radiation,  $\theta$ , the distance between the planes of a crystal,  $d$ , and the wavelength of the incident radiation,  $\lambda$ .** The waves of incident monochromatic radiation are reflected by the parallel equidistant planes of a crystal. When the difference in optical path between the scattered waves takes a multiple of the wavelength, constructive interference occurs, and a diffraction spot is produced.

The intensity of the reflection is recorded by the detector and corresponds to the intensity of this constructive wave. The diffraction data contains information from all atoms in the structure and is obtained as a list of reflection intensities with  $hkl$  positions ( $I_{hkl}$ )<sup>64,65,114,115</sup>, defined by the crystal planes that originated each reflection. At this point, the unit cell parameters and space group can be determined, and a full data collection experiment is performed. Afterwards, all the collected diffraction images are integrated and the intensity and  $hkl$  position of each reflection is extracted.

The power of the crystal to divert the X-ray photons is what dictates the high resolution limit of the diffraction data set and the global quality of the data set<sup>65</sup>. This quality is assessed by calculation of a series of important parameters: the value of  $1/\sigma(I)$ , which corresponds to the signal-to-noise ratio; the completeness, which corresponds to the percentage of the reflections relative to the total number of reflections that could be measured for that crystal, and should be greater than 90%; the redundancy or multiplicity, which is the number of observations per reflection (that is, the number of times the same reflection was measured); the  $R_{merge}$  (Equation 1.2) and  $R_{p.m.i}$  (Equation 1.3) factors, which compare the intensities measured for the various reflections, and should be as low as possible since equivalent reflections (related by symmetry) must have similar intensity values; and the  $CC_{1/2}$  (correlation coefficient between random half data sets<sup>116</sup>), that is

generally close to 1 at low resolution and falls to near zero at higher resolution as the intensities become weaker. Individually, these parameters do not indicate the resolution limits, but are used globally to assess the quality of a data set.

(1.2)

$$R_{merge} = \frac{\sum_{hkl} \sum_{i=1}^n |I_i(hkl) - \bar{I}(hkl)|}{\sum_{hkl} \sum_{i=1}^n I_i(hkl)}$$

(1.3)

$$R_{p.i.m.} = \frac{\sum_{hkl} \sqrt{1/(n-1)} \sum_{i=1}^n |I_i(hkl) - \bar{I}(hkl)|}{\sum_{hkl} \sum_{i=1}^n I_i(hkl)}$$

Nowadays, due to the brightness of X-rays from a synchrotron facility, which is a thousand times greater than that from a laboratory X-ray generator of fixed-wavelength, most structures are solved by using synchrotron radiation<sup>64</sup>. Synchrotrons have the additional advantage of providing tunable radiation, which can be modulated to different wavelengths of interest. Synchrotron facilities have contributed to the structural characterization of protein-carbohydrate interactions of most carbohydrate-recognising protein families<sup>117</sup>.

### 1.3.2.3 3D structure determination

To solve the structure of a protein or protein-ligand complex, the electron density that surrounds all the atoms of the macromolecule in the crystal must be calculated, and structure factors and phase information are necessary, as expressed by the electron density equation (Equation 1.4):

(1.4)

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

where  $|F_{hkl}|$ , known as the amplitude of the structure factor, is obtained from the intensities of each reflection, measured experimentally ( $|F_{hkl}| = \sqrt{I_{hkl}}$ ),  $\alpha$  is the phase angle of the scattered wave,  $x_j y_j z_j$  is the position of atom  $j$  in the unit cell and  $V$  is the volume of the unit cell.

However, from the dataset collected phase angle values cannot be obtained. This is what it is known as the Phase Problem in crystallography. While the intensities and the structure factors are directly measured in the diffraction experiment, phase angle values need to be correctly estimated in order to obtain the electron density map<sup>64,65</sup>.

Three methods are generally used to determine the phases. Single or Multiple Isomorphous Replacement (SIR/MIR), that uses heavy atoms such as Au, Pt, and Hg usually soaked into the native crystals. Multiwavelength Anomalous Dispersion (MAD) and Single-wavelength Anomalous Dispersion (SAD), that exploit the anomalous dispersion (or scattering) effect of specific atoms, typically selenium, recurring to selenomethionyl proteins in which methionines are

replaced by selenomethionine upon protein expression. SIRAS and MIRAS are additional methods that combine Isomorphous Replacement and anomalous scattering. Ultimately, the Molecular Replacement (MR) method can be applied if the structure of the same protein or a similar one, with at least 30% amino acid sequence identity, has already been solved<sup>64,65</sup>. MR is of particular importance for protein-ligand complexes, as the unliganded-structure phases can be combined with the diffraction intensities of the crystal with the bound ligand to calculate its electron density map<sup>65</sup>.

#### 1.3.2.4 Model building and validation

Once the phases have been estimated and the electron density map obtained, preliminary model building takes place. This process is greatly affected by the quality of the map which depends on the maximum resolution of the diffraction data<sup>64,65</sup>.

In protein-ligand complexes, the electron density should reveal the ligand so that its model can be built and integrated in the list of coordinates. The level of detail observed for the ligand depends on the diffraction data quality and but also on the affinity of the protein for that ligand. The affinity to a ligand can influence its occupancy and, in consequence, the definition of the electron density in the protein binding site. Additionally, longer ligands that do not bind entirely to the protein, can be partially disordered in the regions exposed<sup>65,113</sup>. The number of observed solvent molecules is also highly dependent on the resolution of the data and its addition can help in high resolution model building<sup>65</sup>.

Several cycles of model refinement and manual rebuilding take place in order to improve the data statistics. The accuracy of the model is defined by the  $R_{work}$  and  $R_{free}$  factors, that indicate the error between the calculated and the observed amplitudes and should not differ by more than 5%<sup>64,65,118</sup>. The stereochemistry of the model is another validation point, which is given in the form of root mean standard deviations (*rmsd*) for bond lengths and bond angles. The  $R_{work}$  and  $R_{free}$  factors together with *rmsd* values, are good indicators of how well the model fits the data. Model building aims to find the model that best explains the measured data and hence the crystal's content. The final set of atom coordinates achieved should reflect only the interpretable parts of the calculated electron density, with the best possible set of phases, which cannot be dissociated from the quality and high resolution limit of the data<sup>65</sup>.

Once refinement statistics are taken to its best possible values, model validation takes place. The Ramachandran plot reveals the distribution of amino acid residues in the energetically allowed regions. The distribution of temperature factors (B factor or Atomic Displacement Parameter), side chain torsion angles, close contacts and water network contacts are among the several parameters that are validated in order to help correcting and finalizing the best structural model<sup>65</sup>.

This is a process assisted by several available software packages and platforms dedicated to model refinement and validation<sup>65</sup>. When working with a model of a protein-carbohydrate

complex, there are specific tools that must be employed to validate the carbohydrate structure. This is the case of the recently developed Privateer tool that checks the fit of carbohydrates within electron density, validates the sugar ring conformation and can detect stereochemistry problems as well<sup>119,120</sup>.

Once the final model is achieved, the structure can be deposited in the Protein Data Bank (PDB), in the form of a list of 3D coordinates, associated to the respective list of observed structure factors (the measured X-ray diffraction data).

#### 1.4 Structural characterization of protein-carbohydrate interactions

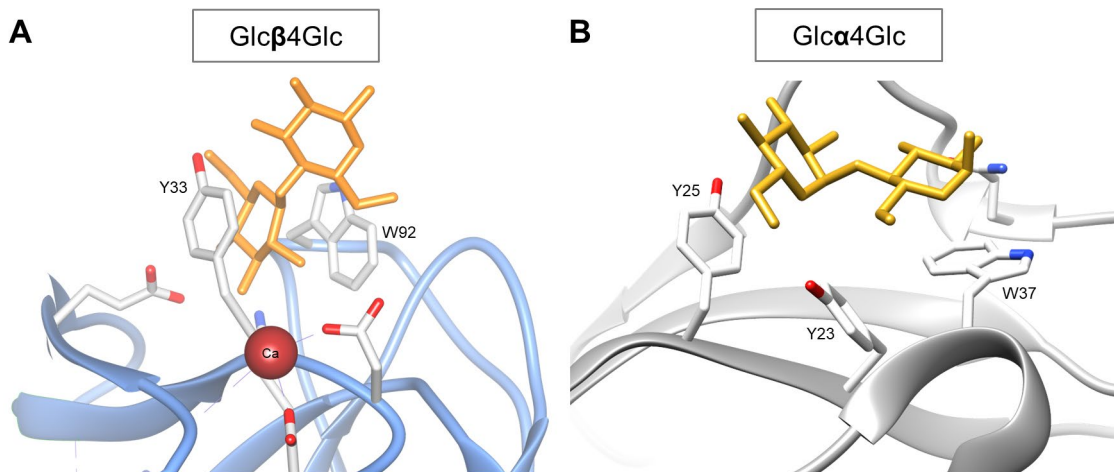
Due to the recognised importance of carbohydrates in diverse biological processes, there has been an increase of research to study protein-carbohydrate interactions and elucidate the molecular mechanisms responsible for the binding recognition. With the increasing number of protein-carbohydrate structures characterized and deposited in the PDB in recent years, ~450 structures released since 2015 from a total of ~2000 protein structure entries assigned as carbohydrate-binding, great knowledge has been gained in understanding the common molecular features that govern the formation of protein-carbohydrate complexes.

The formation of these complexes is driven by favourable changes in enthalpy ( $\Delta H$ ) and entropy ( $\Delta S$ ), accompanied by an increase of the free energy ( $\Delta G$ ) of binding from monosaccharides to longer chain-length oligosaccharides<sup>113</sup>. Entropic penalties may also occur due to restricted conformational freedoms of the protein and ligand<sup>62</sup>. Additionally, upon protein-carbohydrate interaction, there is an increase in avidity due to a multivalent effect<sup>113</sup>.

It is well known that weak molecular forces are largely responsible for protein-carbohydrate recognition<sup>113,121–123</sup>. The most important interactions are van der Waals forces, electrostatic interactions and hydrogen bonds, where the CH- $\pi$  hydrogen bonding are detrimental for the binding events<sup>113,121–123</sup>.

Although sometimes overlooked when studying protein-ligand interactions, water-mediated hydrogen bonding have also been shown to influence protein-carbohydrate affinity<sup>124</sup>. Given its unique ability to donate and accept two hydrogen bonds, water molecules mediate protein-ligand interactions. Ligands compete with a water molecule for binding to the protein binding site, where water molecules will be replaced, retained, or displaced to favour or not ligand binding<sup>125</sup>. Water molecules impact the  $\Delta G$  of binding both enthalpically, by displacing poorly ordered waters or forming newly ordered water network, and entropically, through hydrophobic effect<sup>124,125</sup>. Water networks can make considerable contributions to the protein-ligand affinity, where perturbation of these networks, even without disrupting the ligand or the protein, can substantially decrease enthalpically optimal interactions and introduce solvent mobility, hence having an impact on the ligands' binding<sup>124</sup>.

CH- $\pi$  interactions involve a type of hydrogen bond between aliphatic and aromatic CH's as the hydrogen donor, and the  $\pi$ -systems of arenes as acceptors<sup>113,121</sup>. These interactions have also been referred to as 'edge-to-face', 'T-shape', ' $\pi$ - $\pi$ ' or 'arene-arene'-interactions<sup>121</sup>. Carbohydrate-aromatic CH- $\pi$  interactions are described as stacking interactions due to the parallel orientation of the interacting carbohydrate and the aromatic rings<sup>123</sup> (Figure 1.11).



**Figure 1.11. Examples of carbohydrate-aromatic CH- $\pi$  interactions.** (A) Family 6 CBM from *Cellvibrio mixtus* in complex with cellobiose (Glc $\beta$ 4Glc) in its type C cleft (PDB ID: 1UYX)<sup>38</sup>, exhibiting CH- $\pi$  stacking between both faces of the first glucopyranose ring from a Tyr33 and a Trp92 residue; and (B) *Bacillus halodurans* family 26 CBM bound to maltose (Glc $\alpha$ 4Glc) (PDB ID: 2C3H)<sup>126</sup>, exhibiting CH- $\pi$  interactions between the top faces of the glucose rings and Tyr25 and Trp37, with Tyr23 contributing to hydrogen bonding. Representations (not to scale) of individual 3D structures were done with program Chimera<sup>40</sup> using the PDB atomic coordinates.

Although usually weaker than other interactions, the CH- $\pi$  effects define the protein-carbohydrate enthalpy of binding<sup>113</sup>. The side chains of aromatic amino acids Trp, Tyr, Phe and His, are typical  $\pi$ -systems that can act as acceptor groups. The carboxyl or carboxamide side chains in Asp, Glu, Asn and Gln may also contribute. Additionally, main chain peptide groups, which constitute  $\pi$ -systems with some degree of delocalization, may also act as  $\pi$ -acceptors. The frequent occurrence of aliphatic and aromatic CH- $\pi$  interactions, not just in proteins but as well in nucleic acids, membrane lipids and polysaccharides, suggests an important functional role<sup>121</sup>.

Analysis of protein-carbohydrate structures deposited in the PDB have revealed that many carbohydrate-binding proteins contain aromatic amino acid residues in their binding sites and that these residues interact with their carbohydrate ligands in a stacking geometry through CH- $\pi$  interactions<sup>113,122,123</sup>. Recent studies have revealed that aliphatic hydrophobic residues in the carbohydrate-binding sites are not favoured when compared to aromatic side chains, with a higher preference for Trp residues followed by Tyr<sup>122,123</sup>. Aromatic CH- $\pi$  interactions can be found in carbohydrate-binding proteins, such as CBMs, lectins and CAZymes, and are involved in a wide range of processes from carbohydrate-binding, catalytic processing and transport<sup>123</sup>.

$\beta$ -D-glucopyranose is structurally predisposed for carbohydrate-aromatic interactions due to the fact that all of its ring C-H hydrogens are oriented axially, which makes it possible to interact with

an aromatic system, like of Trp, Tyr and Phe, in a parallel stacking geometry<sup>123</sup>. This stacking interaction can happen either at the top, the bottom or both faces of the carbohydrate ring (Figure 1.11A). On the contrary,  $\alpha$ -D-glucopyranose only interacts with the carbohydrate ring top face (Figure 1.11B) because the anomeric hydroxyl group blocks the bottom face<sup>123</sup>.

These carbohydrate-aromatic CH- $\pi$  interactions have been defined as dispersion interactions, tuned by electrostatics and partially stabilized by a hydrophobic effect in solvated systems<sup>122,123</sup>. As electrostatic interactions are highly influenced by directionality and charge distribution on donor and acceptor molecules, they can further strengthen and orientate the carbohydrate-aromatic complexes<sup>123</sup>. Additionally, because the electrostatic surfaces and the electropositive characters of C-H bonds of the carbohydrates engaging in CH- $\pi$  interactions differ between carbohydrate isomers, the aromatic side chains of the protein engage with different regions of the carbohydrate. This effect provides a mechanism for discriminating between carbohydrate monomers, influencing which bind to the protein and how they are positioned within carbohydrate-binding sites<sup>122</sup>.





## 1.5 Thesis main objectives

The work described in this Thesis aimed to add knowledge about biotechnologically relevant bacterial species by identifying the carbohydrate ligands and structurally characterizing the ligand-specificity of novel CBMs from *C. thermocellum* and *R. flavefaciens* FD-1. These are highly efficient cellulolytic anaerobic bacteria, which present cellulosomes expressing different complex architectures and a high number of yet uncharacterized CBMs. Given the crucial roles of CBMs, the elucidation of their binding specificities and mechanisms of carbohydrate recognition will contribute to the characterization of the bacterial cellulosomes, promoting the knowledge of these bacteria cellulolytic capabilities. The experimental design of this Thesis will explore the potential of applying a unique approach combining carbohydrate microarrays and X-ray crystallography, aiming to contribute to the classification and elucidation of those CBMs biological roles, as well as to their potential biotechnological applications. To achieve these aims, this Thesis will contemplate the following major objectives:

- 1) To construct carbohydrate microarray platforms of polysaccharides and sequence-defined oligosaccharides, representative of those found in plant cell walls, putative ligands for the CBMs of *C. thermocellum* and *R. flavefaciens* FD-1.
- 2) To perform initial screening analysis of carbohydrate binding for *C. thermocellum* and *R. flavefaciens* FD-1 CBMs with predicted or unknown specificities, using the validated microarrays platform of plant cell wall polysaccharides, to identify its carbohydrate binding patterns.
- 3) To conduct a second screening using the newly constructed oligosaccharides microarrays to identify the oligosaccharide ligands and assign the specificity of *C. thermocellum* and *R. flavefaciens* FD-1 CBMs for which polysaccharide binding patterns were obtained in the first screening.
- 4) To determine the structural basis of the carbohydrate recognition mechanisms of selected novel CBMs and CBM-ligand complexes from *C. thermocellum* and *R. flavefaciens* FD-1.



# CHAPTER 2

---

**DEVELOPMENT OF GLUCAN AND HEMICELLULOSE  
OLIGOSACCHARIDE MICROARRAYS APPLIED TO PLANT  
CELL WALL CARBOHYDRATE RECOGNITION**



## 2 Development of glucan and hemicellulose oligosaccharide microarrays applied to plant cell wall carbohydrate recognition

### 2.1 Introduction

The plant cell wall is constituted by structurally diverse and complex polysaccharides (Figures 1.1 and 1.2, Chapter 1), comprising cellulose, glucans, hemicelluloses and pectins, which provide valuable resources for industrial and biotechnological applications<sup>43,61,127,128</sup>. The composition in polysaccharides is highly variable, depending on the growth stage, tissue type and phylogenetic groups and plant species<sup>3,4,61</sup>. The understanding of plant cell wall polysaccharide molecular structures, functions, and biosynthesis, as well as the biological mechanisms underlying carbohydrate recognition and deconstruction by microorganisms, is required to further promote their industrial and biotechnological use.

Given the continuously increasing amount of information derived from microbial genomic sequencing, there is a high demand for high-throughput and sensitive micro-methods to interrogate and characterize the high complexity of newly identified carbohydrate recognition systems. The carbohydrate microarray technology has emerged as a powerful high-throughput screening tool for ligand discovery and characterization of carbohydrate-protein interactions. Its application to many biological systems has led to rapid advances in the decoding of glycomes<sup>7-10</sup> and the development of oligosaccharide microarrays has played a major role in unravelling carbohydrate-binding specificities for proteins<sup>71,92</sup>.

In the above context, the development of sequence-defined plant-based carbohydrate microarrays has been highly important to provide proteins with characterised specificities for plant cell wall research. Carbohydrate-binding proteins, such as lectins and CBMs, and carbohydrate-directed monoclonal antibodies, are serving as tools in the detailed characterization of the diversity of plant cell wall carbohydrate structures<sup>129,130</sup>. These proteins can be used in quantitative and high-throughput assays, giving measurements about specific carbohydrate epitopes present in plant polysaccharides<sup>129</sup>. However, detailed characterization of epitope requires the availability of focused plant carbohydrate microarray platforms with purified and well characterized plant oligosaccharides probes.

As reviewed in Chapter 1 (section 1.3.1), some laboratories have developed microarrays from plant-derived carbohydrates<sup>32,92,97-102</sup>, and the number of sequence-defined probes has been expanding. Willats and colleagues have established microarrays of oligosaccharides coupled to BSA at the reducing end by reductive amination for arraying non-covalently onto nitrocellulose-coated glass slides<sup>92</sup>. The oligosaccharides comprised linear and branched sequences of glucans, xylans, mannans, galactans, xyloglucans or arabinans, either isolated from

polysaccharides or generated by chemical synthesis<sup>92</sup>. Pfrengle and colleagues, have constructed microarrays of plant cell wall oligosaccharides obtained by solution-phase or automated glycan assembly synthesis, covalently immobilized onto NHS ester-derivatized glass slides through an aminoalkyl-linker at the reducing end, comprising a range of sequences from xylans, glucans, xyloglucans, galactans, arabinogalactans and pectins<sup>109,110</sup>. Palma and colleagues have developed sequence-defined glucome microarrays<sup>32</sup>, comprised of 153 gluco-oligosaccharides with diverse sequences and chain lengths representing major sequences in glucans, including those present in plant cell walls, converted to NGL probes by oxime ligation to an aminoxy-functionalized lipid (AO-NGLs) and non-covalently immobilized onto nitrocellulose-coated glass slides<sup>86</sup>. A key feature of the NGL technology is its interface with mass spectrometry (MS) and high-performance thin layer chromatography (HPTLC) or high performance liquid chromatography (HPLC), enabling bioactive oligosaccharides released from a glycome source to be resolved from heterogeneous mixtures, characterized and purified<sup>32,74,78</sup> (Figure 1.5, Chapter 1). The partial depolymerisation of the polysaccharides, the sequencing of the oligosaccharides by the negative-ion electrospray ionization mass spectrometry with collision-induced dissociation (ESI-CID-MS/MS) method and their interrogation on the microarrays, not only provides detailed information on linkage, sequence and chain-length requirement of carbohydrate-recognizing proteins, but also provides a sensitive means of revealing unsuspected sequences in the polysaccharides<sup>32</sup>.

The work presented in this chapter aimed at extending the glucome NGL-microarray platform with naturally-derived linear and branched hemicellulose oligosaccharides to address the need of increasing oligosaccharide structural diversity and chain-length range, including galactomannan and fucosylated-xyloglucan sequences that are under-represented in microarrays, for studies of plant cell wall carbohydrate recognition. To this end, novel microarrays with AO-NGL probes were constructed that comprised oligosaccharide fragments derived from xylans, arabinoxylans, arabinans, mannans, galactomannans and xyloglucans. The development of plant oligosaccharide microarrays from naturally-derived heterogeneous mixtures poses considerable challenges in obtaining size homogenous and structurally-defined probes.

The main difficulty for construction of microarrays from plant-derived oligosaccharides is the lack of any chromophore in their hexose and pentose constituents, as the commonly used low wavelength UV detection (e.g. 195 nm or 206 nm) for mammalian glycans is not applicable to these plant oligosaccharides. In this work, this was addressed by: 1) using size exclusion fractionation of oligosaccharide mixtures or HPTLC upon lipid derivatization and MALDI-MS analysis; 2) probing the microarrays with CBMs and lectins, for which specificity is known, and monoclonal carbohydrate-directed antibodies to validate and characterize the oligosaccharide sequences of the NGL-probes; and 3) devising a new strategy to conjugate oligosaccharides with a bifunctional UV/fluorescence tag for sensitive detection in HPLC, allowing detailed fine separation/purification of structurally similar oligosaccharides before its conversion into NGL

probes. In this work, we aimed to extend the use of negative-ion ESI-CID-MS/MS for homo gluco-oligosaccharides (Palma et al., 2015<sup>32</sup>) to hetero hexo-/pento-oligosaccharides.

The application of the developed microarrays was valuable to assign the carbohydrate-binding specificities for *Clostridium thermocellum* CBMs from families 25 and 35 and for the galactomannan-directed monoclonal antibody CCRC-M70. These validated glucan and hemicellulose oligosaccharide microarrays will be essential tools to unravel the carbohydrate-binding for CBMs of cellulolytic microorganisms that exhibit highly diverse specificities. This will be a topic further developed in the following chapters of this Thesis.

## 2.2 Results

### 2.2.1 Construction of glucan and hemicellulose oligosaccharide microarrays

The glucan microarrays comprised 153 gluco-oligosaccharides derived from partial depolymerisation of glucans or from chemical synthesis and prepared as NGLs probes (Table S2.1, probes 1 to 153). Among these, were linear oligosaccharide sequences with homo linkages in  $\alpha$  or  $\beta$  configurations: 1,2-, 1,3-, 1,4-, or 1,6-linked, with degree of polymerization (DP) ranging from DP-2 to DP-13; linear oligosaccharide sequences with hetero linkages:  $\alpha$ 1,4-1,6, ranging from DP-3 to DP-7 and  $\beta$ 1,3-1,4 ranging from DP-3 to DP-16; branched oligosaccharide sequences with  $\beta$ 1,3( $\beta$ 1,6) linkages, DP-2 to DP-13; and synthetic branched oligosaccharides of  $\beta$ 1,3-linked linear backbones (DP-8 or DP-9), with a  $\beta$ 1,6-linked mono-glucosyl branches. These NGL probes were previously described by Palma and colleagues to generate the sequence-defined “glucome” microarray platform<sup>32</sup>.

In this study, the strategy used for the glucome microarray was followed to prepare hemicellulose oligosaccharide NGL probes. To this end, reducing oligosaccharide fragments of discrete chain lengths were isolated from oligosaccharide mixtures of depolymerised xylan, arabinoxylan, arabinan, mannan, galactomannan and xyloglucan (Table S2.2) by size exclusion chromatography with off-line MALDI-MS analysis. The oligosaccharides thus prepared, were conjugated to an aminoxy-functionalized-lipid via oxime-ligation to generate AO-NGL probes (Figure 1.6, Chapter 1), following standard procedures<sup>32,86,131</sup>. The purified NGL products were analysed by MALDI-MS for molecular mass determination and therefore assignment of degree of polymerisation (DP) for the major components in each NGL probe (Table 2.1). The NGLs prepared comprised oligosaccharide sequences derived from: linear  $\beta$ 1,4- or linear mixed-linked  $\beta$ 1,3-1,4-xylans (DP-3 to DP-13); arabinoxylans ( $\beta$ 1,4-xylans with  $\alpha$ 1,2- and/or  $\alpha$ 1,3-linked arabinose branches; DP-3 to DP-6); linear  $\alpha$ 1,5-arabinans (DP-2 to DP-9);  $\alpha$ 1,5-arabinans with  $\alpha$ -arabinose branches at position O3 (DP-5 and DP-6) or at both O2 and O3 (DP-6); linear  $\beta$ 1,4-mannans (up to DP-8); galactomannan ( $\beta$ 1,4-mannan with  $\alpha$ 1,6-linked galactose branches; DP-2 to DP-11); xyloglucan from tamarind (DP-7 to DP-9); and fucosylated-xyloglucan from apple (DP-7 and DP-10). Carbohydrate sequence information on these probes is in Table S2.1 (probes 154 to 204).

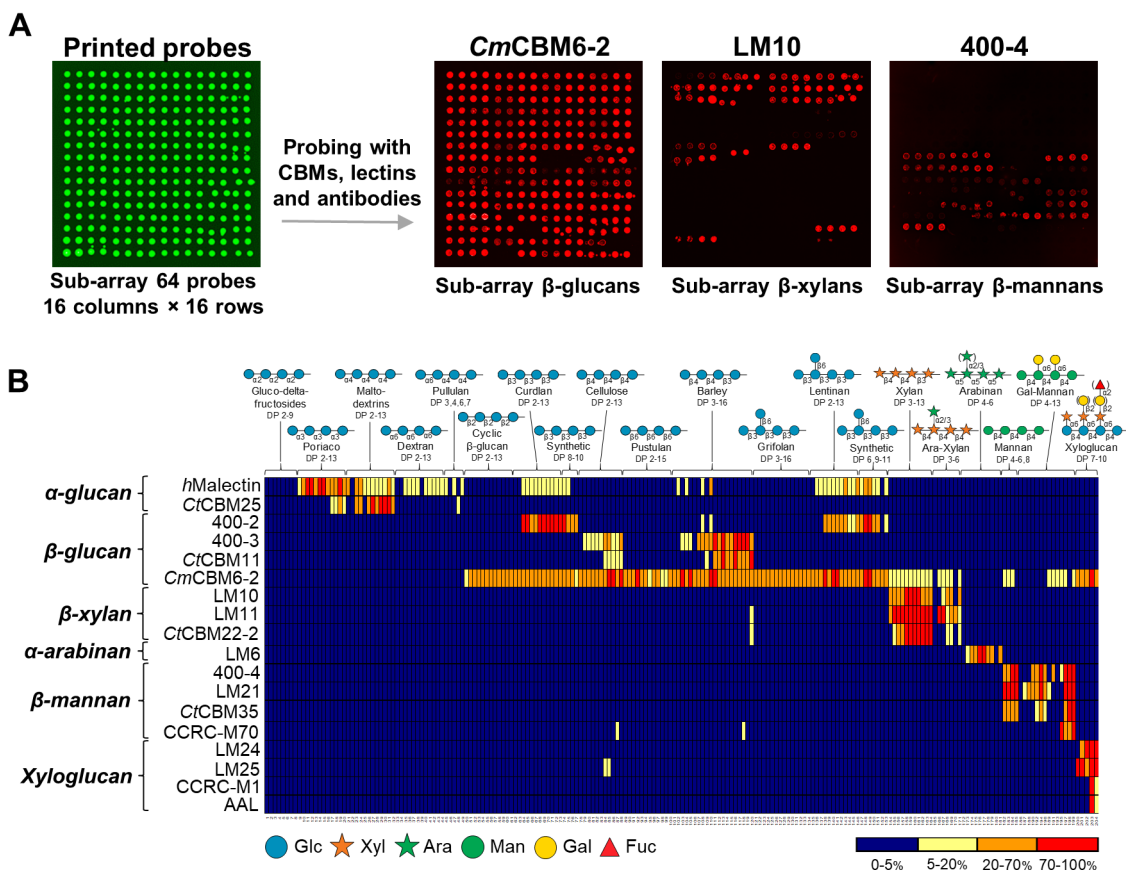
**Table 2.1. MALDI-MS analysis of AO-NGLs derived from hemicellulose oligosaccharide fractions.**

Oligosaccharide series <sup>a</sup>	NGL probe designation	DP <sup>b</sup>	[M-H] <sup>-</sup> calculated <sup>c</sup>	[M-H] <sup>-</sup> detected <sup>d</sup>
<b>Xylan</b>	Xyl-3( $\beta$ 3-4)	3	1131.69	1131.8
	Xyl-4( $\beta$ 3-4)	4	1263.73	1263.8
	Xyl-5( $\beta$ 3-4)	5	1395.78	1395.9
	Xyl-6( $\beta$ 3-4)	6	1527.82	1527.9
	Xyl-7( $\beta$ 3-4)	7	1659.86	1660.0
	Xyl-8( $\beta$ 3-4)	8	1791.90	1792.0 (1660.0, 1924.1) <sup>e</sup>
	Xyl-9( $\beta$ 3-4)	9	1923.95	1924.1 (1792.0, 2056.2)
	Xyl-10( $\beta$ 3-4)	10	2055.99	2056.1 (1924.0, 2188.2)
	Xyl-11( $\beta$ 3-4)	11	2188.03	2188.2 (2056.1, 2320.2)
	Xyl-12( $\beta$ 3-4)	12	2320.07	2320.2 (2188.1, 2452.2)
Xyl-13( $\beta$ 3-4)	13	2452.11	2452.2 (2320.1, 2584.2)	
<b>Arabinoxylan (Ara-Xylan)</b>	Ara-Xylan-3	3	1131.69	1131.6
	Ara-Xylan-4a	4	1263.73	1263.7
	Ara-Xylan-4b	4	1263.73	1263.7
	Ara-Xylan-5a	5	1395.78	1395.8
	Ara-Xylan-5b	5	1395.78	1395.8
	Ara-Xylan-5c	5	1395.78	1395.8
	Ara-Xylan-6	6	1527.82	1527.8
<b>Arabinan</b>	Ara-2( $\alpha$ 5)	2	999.65	999.6
	Ara-3( $\alpha$ 5)	3	1131.69	1131.6
	Ara-4( $\alpha$ 5)	4	1263.73	1263.8
	Ara-5( $\alpha$ 5)	5	1395.78	1396.0
	Ara-6( $\alpha$ 5)	6	1527.82	1527.9
	Ara-7( $\alpha$ 5)	7	1659.86	1659.9
	Ara-8( $\alpha$ 5)	8	1791.90	1791.9
	Ara-9( $\alpha$ 5)	9	1923.95	1924.0 (1791.9)
	Ara-4B3	4	1263.73	1263.7
	Ara-5B	5	1395.78	1395.9
<b>Mannan</b>	Man-4( $\beta$ 4)	4	1383.78	1383.8
	Man-5( $\beta$ 4)	5	1545.83	1545.7
	Man-6( $\beta$ 4)	6	1707.88	1707.9
	Man-8( $\beta$ 4)	8	2031.99	2031.9
<b>Galactomannan (Gal-Mannan)</b>	Gal-Mannan-2e	2	1059.67	1059.7
	Gal-Mannan-3e	3	1221.72	1221.7
	Gal-Mannan-4e	4	1383.78	1383.8
	Gal-Mannan-5e	5	1545.83	1545.9
	Gal-Mannan-6e	6	1707.88	1708.0 (1545.9)
	Gal-Mannan-7e	7	1869.93	1870.0
	Gal-Mannan-8e	8	2031.99	2032.1
	Gal-Mannan-5m	5	1545.83	1545.9
	Gal-Mannan-6m	6	1707.88	1707.9
	Gal-Mannan-7m	7	1869.93	1870.0
	Gal-Mannan-8m	8	2031.99	2032.0
	Gal-Mannan-9e	9	2194.04	2194.2
	Gal-Mannan-10e	10	2356.09	2356.2 (2194.2)
Gal-Mannan-11e	11	2518.15	2518.3 (2356.2)	
<b>Xyloglucan</b>	Xyl-Glucan-7	7	1779.90	1779.9
	Xyl-Glucan-8	8	1941.96	1942.0 (1779.9)
	Xyl-Glucan-9	9	2104.01	2104.0
	Fuc-Xyl-Glucan-6	6	1647.86	1647.8 (1793.8)
	Fuc-Xyl-Glucan-9	9	2104.01	2104.0 (1941.9, 2250.0)

<sup>a</sup>The designation of the linear and branched oligosaccharide moieties corresponds to the sources of the oligosaccharide fragments to prepare the NGLs (depolymerised polysaccharides or oligosaccharides); <sup>b</sup>Degree of polymerization (DP) for the major components in each fraction; <sup>c</sup>Calculated masses for major components are given; <sup>d</sup>Negative-ion MALDI-MS was used for the analysis of the AO-NGLs and [M-H]<sup>-</sup> were detected; <sup>e</sup>Where multiple components were detected, relative intensities of molecular ions greater than 20% are shown in brackets. The molecular ions detected for Fuc-Xyl-Glucan-6 at 1647.8 m/z (DP-6) correspond to the 1793.8 m/z sequence (DP-7) without the fucose monomer. For Fuc-Xyl-Glucan-9, the molecular ions detected at 2104.0 m/z (DP-9) correspond to the 2250.0 m/z sequence (DP-10) without the fucose monomer. Both fucosylated and non-fucosylated components are present in these NGL-probes.



To construct the microarrays, a total of 204 AO-NGLs from glucan and hemicellulose oligosaccharides were printed by non-covalent immobilization using a liposome formulation onto 16-pad nitrocellulose-coated glass slides. Figure 2.1A shows the typical 16-pad subarray layout used, which featured 64 probes, each printed at 2 levels at 2 and 5 fmol/spot in duplicate (four spots for one probe in a row); up to 256 spots (16×16) in total in each subarray. The analysis of different carbohydrate-binding proteins in these microarrays highlighted the structural diversity of oligosaccharide probes included (Figure 2.1A-B). The binding signals observed were probe-dose dependent, and the results of the analysis with the probes at 5 fmol/spot will be described in the sections below.



**Figure 2.1. Binding patterns revealed by probing the glucan and hemicellulose oligosaccharide microarrays with sequence-specific carbohydrate-binding proteins. (A)** Microarray fluorescence imaging showing a typical subarray of immobilized 64 probes upon visualisation using Cyanine-3 (Cy3) fluorophore (532 nm) included in the printing solution (panel at the left) and subarrays showing binding spots observed upon probing with *CmCBM6-2* (sub-array featuring  $\beta$ -glucan sequences) and monoclonal antibodies LM10 (sub-array featuring  $\beta$ -xylan sequences) and 400-4 (sub-array featuring  $\beta$ -mannan sequences). **(B)** Heat map representing the relative binding intensities, calculated as the percentage of the fluorescence signal intensity given by the probe (printed at 5 fmol/spot) most strongly bound by each protein (normalized as 100%). Numerical fluorescence intensity signals are given in Table S2.4. The microarray comprises 204 neoglycolipid probes with a wide degree of polymerization (DP) range of linear and branched oligosaccharide sequences of  $\alpha$ - and  $\beta$ -glucans,  $\beta$ -xylans,  $\alpha$ -arabinans,  $\beta$ -mannans and xyloglucans. The major representative structural domain for each probe series is depicted at the top of the panel using a tetrasaccharide backbone sequence as a reference. Carbohydrate sequence information on these probes is in Table S2.1. The carbohydrate-specific monoclonal antibodies, lectins and CBMs investigated in the microarrays targeting the different carbohydrate groups are depicted at the left. The monosaccharide symbolic representation used was according to the updated SNFG<sup>1</sup>.

## 2.2.2 Validation of the constructed microarrays

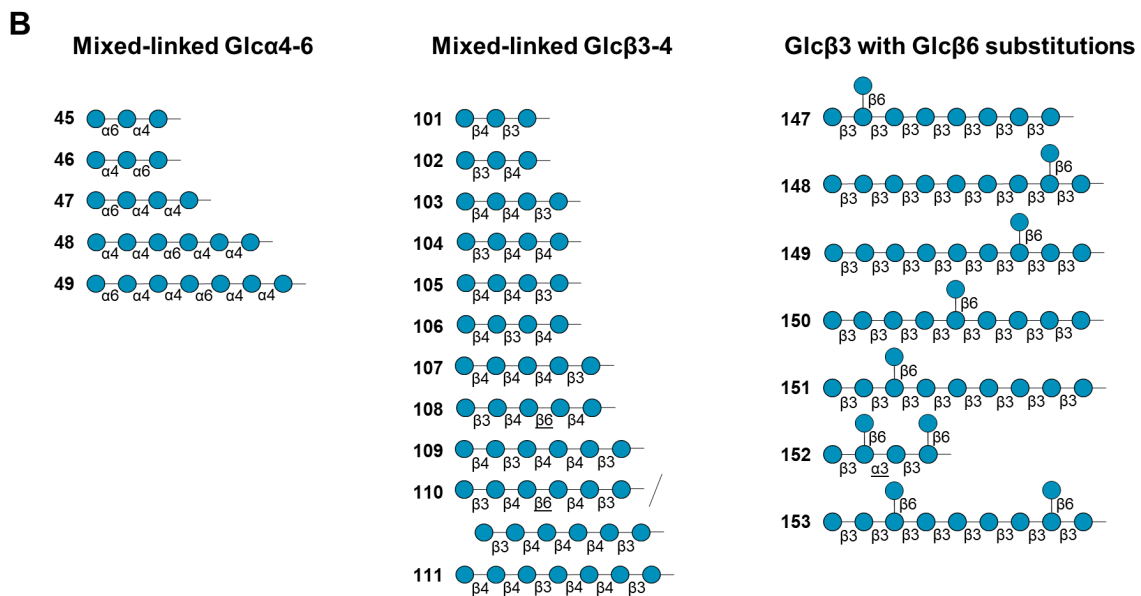
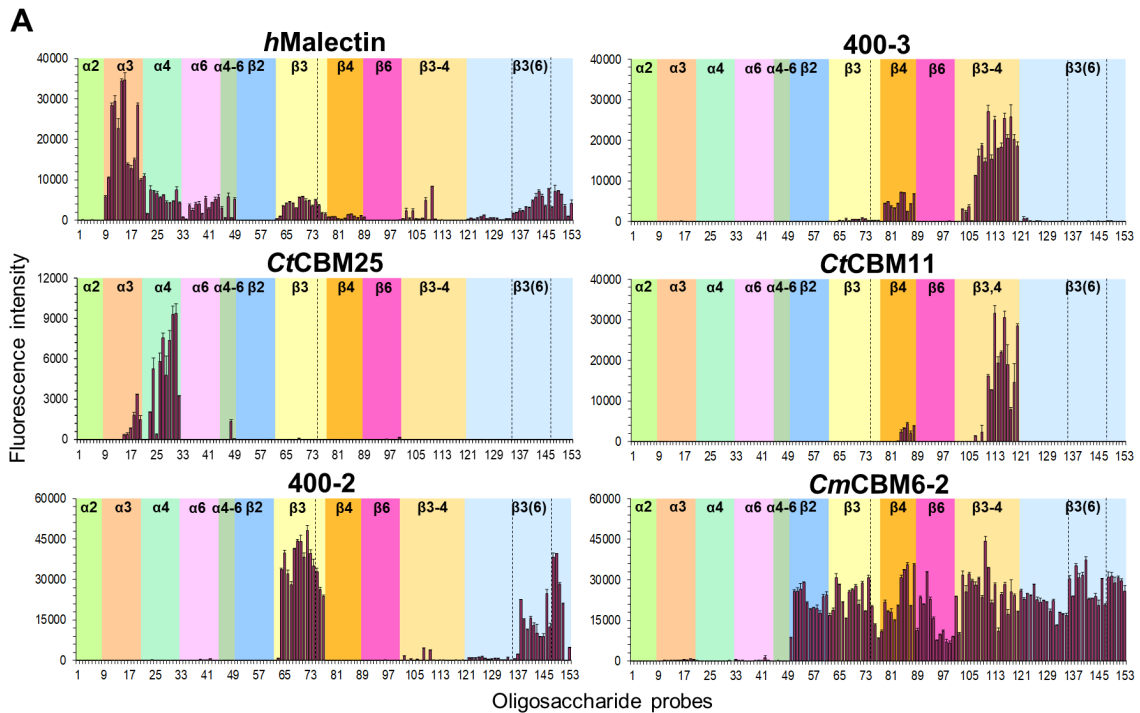
To validate the microarray for carbohydrate recognition studies, a total of 18 carbohydrate-binding proteins were used to probe the microarray, including carbohydrate-specific monoclonal antibodies, lectins and microbial CBMs. The criteria to select the proteins for analysis were two-fold: proteins that have their binding specificities characterised by different methods, and proteins for which the microarrays would serve as important tools to assign novel binding specificities. The binding features obtained with the characterised proteins were used for microarray quality control and structural validation of the generated probes. As shown in Figure 2.1A-B, discrete clusters of binding could be identified with the grouping of probes by main carbohydrate structural domain. Overall, these were in agreement with the reported carbohydrate-binding specificities of the proteins investigated (Table S2.3). That is, there was a clear pattern differentiation of  $\alpha$ - or  $\beta$ -glucans,  $\beta$ -xylans (linear or ramified with arabinose),  $\alpha$ -arabinans,  $\beta$ -mannans (linear or ramified with galactose), or xyloglucans. Additional specificities were also revealed using the newly constructed microarrays. The prominent findings of the microarray analysis will be described in the sections below.

### 2.2.2.1 Recognition of gluco-oligosaccharides with $\alpha$ - and $\beta$ -glycosidic linkages in linear or branched chains

The binding profiles observed with the glucan-recognizing proteins are highlighted in Figure 2.2A, which shows a snapshot of the glucan oligosaccharide probes in the arrays (Table S2.1, probes 1 to 153).

The  $\alpha$ -glucan-derived oligosaccharide probes have been previously validated using anti- $\alpha$ 1,4-,  $\alpha$ 1,3- and  $\alpha$ 1,6-glucan specific monoclonal antibodies, and a *Thermotoga maritima*  $\alpha$ 1,4-glucan binding CBM from family 41<sup>32</sup>. In the current study, the microarrays were probed with human malectin, which is a highly conserved lectin in the endoplasmic reticulum of mammals with a CBM-type fold (CBM family 57, as classified in the CAZy database)<sup>132,133</sup>. This lectin exhibits a unique specificity towards  $\alpha$ 1,3-di-glucosylated high-mannose *N*-glycans, but similarly to its CBM structural homologues also recognises glucan sequences, with higher affinity for  $\alpha$ -glucans<sup>132,133</sup>. The current microarray analysis showed the glucan-binding property of malectin to linear  $\alpha$ 1,3-,  $\alpha$ 1,4-,  $\alpha$ 1,6-linked glucose sequences and also that malectin exhibited a chain-length dependency up to the DP-4/DP-5, only in the case of  $\alpha$ 1,3-glucan sequences. Malectin also bound to mixed-linked  $\alpha$ 1,4-1,6-glucan oligosaccharides that have an  $\alpha$ 1,6-linked glucose at the non-reducing end (Figure 2.2B, probes 45, 47 and 49). Although this lectin was more avid towards  $\alpha$ 1,3-linked glucose, it also showed binding to linear  $\beta$ 1,3-linked glucose oligosaccharides, as previously reported<sup>132,133</sup>, and to branched oligosaccharides containing  $\beta$ 1,3-linked glucose sequences.

The microarrays were also applied to assign the  $\alpha$ -glucan binding specificity for an uncharacterised putative starch-binding domain from *Clostridium thermocellum* assigned to CAZy



**Figure 2.2. Microarray analysis of glucan-binding proteins. (A)** Binding with human malectin and *CtCBM25*, and  $\beta$ -glucan-specific monoclonal antibodies 400-2, 400-3 and CBMs *CtCBM11* and *CmCBM6-2*. The binding signals are depicted as means of fluorescence intensities of duplicate spots at 5 fmol of oligosaccharide probe arrayed (with error bars) and are representative of at least two independent experiments. Numerical fluorescence intensity signals are given in Table S2.4. The microarray comprise 153 gluco-oligosaccharide-NGLs<sup>32</sup> for which the glucose linkages are indicated in the coloured panels. Carbohydrate sequence information on these probes is in Table S2.1. **(B)** The sequences of the mixed-linked linear  $\alpha$ 4,6- and  $\beta$ 3,4-glucan and branched  $\beta$ 3(6)-glucan probes are depicted indicating the position in the binding charts.

family 25 (*CtCBM25*<sub>Cthe\_0956</sub>). The microarray analysis showed that this CBM exhibits a different binding profile when compared to malectin, showing exclusive binding to linear sequences of  $\alpha$ 1,4-linked glucose with a chain-length dependency higher than DP-4. This CBM also bound, albeit weakly, to linear  $\alpha$ 1,3-glucan sequences with DP-10 to DP-13, which might be explained

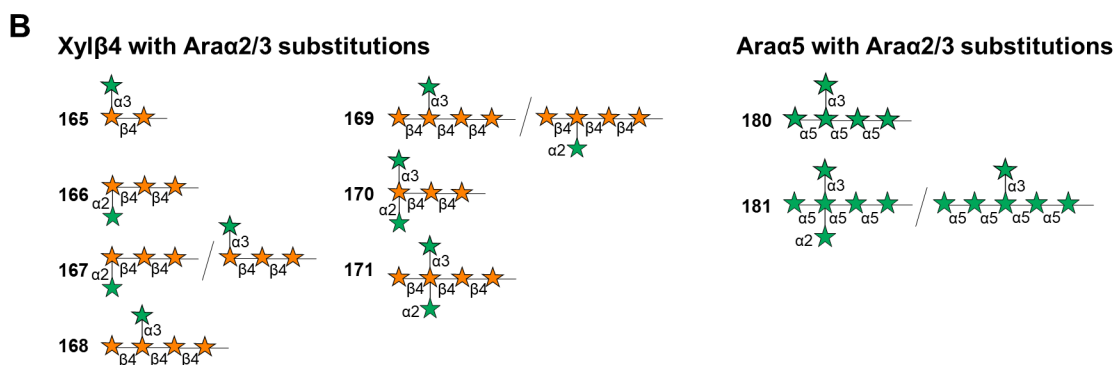
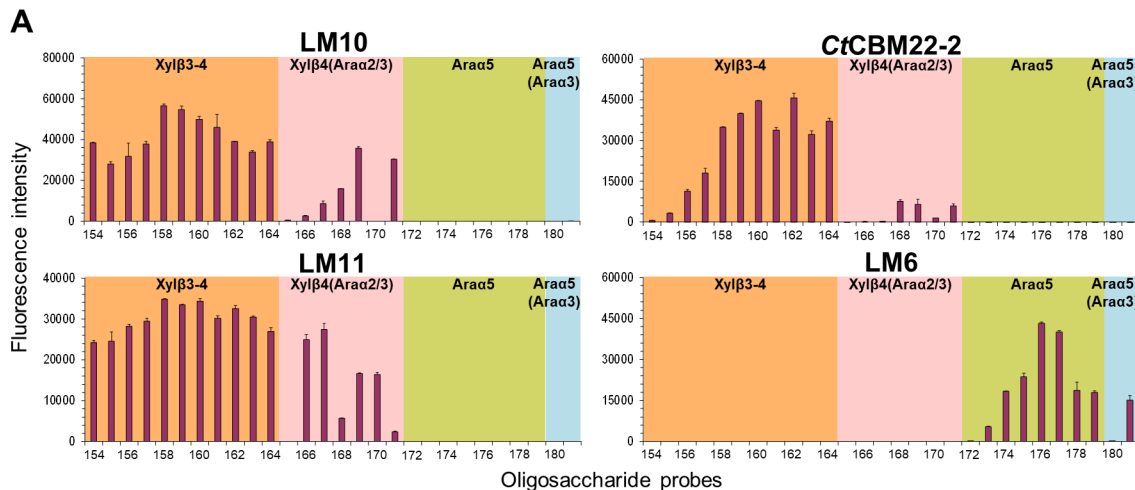
by the presence of  $\alpha$ 1,4-glucan sequences in these fractions (Table S2.1)<sup>32</sup>, and to the mixed-linked  $\alpha$ 1,4-1,6-glucan hexasaccharide with a  $\alpha$ 1,4-trisaccharide sequence at the non-reducing end (probe 48, Figure 2.2B).

The  $\beta$ -glucan-derived microarrays have been previously validated using antibodies and CBMs with specificity towards linear  $\beta$ 1,3-,  $\beta$ 1,4-,  $\beta$ 1,6-glucans, as well as to mixed-linked  $\beta$ 1,3- $\beta$ 1,4-glucans<sup>32</sup>. In the current microarray analysis, the *Clostridium thermocellum* (*C. thermocellum*) family 11 CBM (*Ct*CBM11<sub>Cthe\_1472</sub>) and the *Cellvibrio mixtus* family 6 CBM (*Cm*CBM6-2) were also used in parallel with the  $\beta$ 1,3-glucan-specific (400-2) and  $\beta$ 1,3- $\beta$ 1,4-glucan-specific (400-3) monoclonal antibodies (Table S2.3). The antibody 400-2 exhibited the predicted binding specificity towards  $\beta$ 1,3 oligosaccharide sequences, with a chain-length dependency of DP-4. The binding observed with the 400-3 and *Ct*CBM11<sub>Cthe\_1472</sub> corroborated the reported mixed-linked  $\beta$ 1,3-1,4-glucan binding specificity of these proteins<sup>32,134</sup>, with DP ranges from DP-4 to DP-13 and DP-7 to DP-16, respectively, and a chain-length dependency more pronounced for *Ct*CBM11<sub>Cthe\_1472</sub>. The tetrasaccharide with an internal  $\beta$ 1,3-linkage didn't elicit any binding signals with both proteins (Figure 2.2B, probe 106), indicating that the spacing and the position of the  $\beta$ 1,3-linkages within the chain is important for recognition. The capability of binding with weak affinity to linear  $\beta$ 1,4-linked, but not to  $\beta$ 1,3-linked glucose sequences, was also observed for both proteins. *Cm*CBM6-2 showed the broadest binding profile, and bound to all of the  $\beta$ -linked glucose oligosaccharides with DP-2 and longer as previously observed<sup>32</sup>, including the ability to bind to xyloglucan oligosaccharides, which have a  $\beta$ 1,4-linked glucose backbone (Figure 2.1). This CBM also bound to the oligosaccharide series of other  $\beta$ -linked hexoses, including xylose and mannose, although with weaker intensities.

### 2.2.2.2 Recognition of linear $\beta$ -xylans, branched arabinoxylans and $\alpha$ -arabinans

The binding recognition profiles observed with the xylan- and arabinan-binding proteins are highlighted in Figure 2.3A, which shows a snapshot of the linear  $\beta$ 1,4- and mixed-linked  $\beta$ 1,3-1,4-xylans, branched arabinoxylans and  $\alpha$ 1,5-arabinan oligosaccharide probes comprised in the arrays (Table S2.1, probes 154 to 181).

The  $\beta$ 1,4-xylan-specific monoclonal antibodies LM10 and LM11<sup>135</sup> showed binding to linear  $\beta$ 1,4-linked xylose tetra- and pentasaccharides and to the mixed-linked  $\beta$ 1,3-1,4-xylan oligosaccharides derived from *Palmaria palmate* xylan (DP-3 to DP-13). LM11, having a higher tolerance for backbone substitution, showed also strong binding to  $\beta$ -xylose sequences with  $\alpha$ 1,2- or  $\alpha$ 1,3-arabinose substitutions at the non-reducing end (probes 166 and 167 Figure 2.3B). The weak binding observed with LM10 to these probes, and the ability to bind to probe 171, with an internal  $\alpha$ 1,2- or  $\alpha$ 1,3-arabinose substitution, shows evidence that this antibody has specificity towards the non-reducing end of  $\beta$ 1,4-linked xylans<sup>109</sup>. The weak binding of LM11 to the latter indicates that this antibody binds internally and requires longer epitopes of at least 3 xylose monomers.



**Figure 2.3. Microarray analysis of xylan- and arabinan-binding proteins.** (A) The binding signals of  $\beta$ 1,4-xylan-specific monoclonal antibodies LM10, LM11 and CtCBM22-2, and anti- $\alpha$ 1,5-arabinan antibody LM6 are depicted as means of fluorescence intensities of duplicate spots at 5 fmol of oligosaccharide probe arrayed (with error bars) and are representative of at least two independent experiments. Numerical scores are given in Table S2.4. The different carbohydrate groups are indicated in the coloured panels. The microarrays included 10 linear  $\beta$ 1,3-1,4-xylan-NGLs from DP-3 to DP-13, 7 branched  $\beta$ 1,4-xylan( $\alpha$ 1,2/1,3-arabinose) ranging from DP-3 to DP-6, 8 linear  $\alpha$ 1,5-arabinan of DP range from 2 to 9 and 2 branched  $\alpha$ 1,5-arabinan( $\alpha$ 1,2/1,3-arabinose) of DP-5 and DP-6. Carbohydrate sequence information on these probes is in Table S2.1. (B) The sequences of the mixed-linked linear and branched probes are depicted by microarrays position.

The analysis of the *C. thermocellum* CBM from family 22 (CtCBM22-2<sub>Cthe\_0912</sub>), which has a reported binding specificity for  $\beta$ 1,4-linked xylose oligosaccharides<sup>136</sup>, showed binding to the linear  $\beta$ -xylan probes, exhibiting a chain-length dependency from DP-3 up to DP-8. CtCBM22-2<sub>Cthe\_0912</sub> also bound, albeit weakly, to the branched arabinoxylan probes that present a tetrasaccharide backbone and internal  $\alpha$ 1,2- or  $\alpha$ 1,3-arabinose substitutions (probes 168, 169 and 171, Figure 2.3B).

The  $\alpha$ 1,5-arabinose-specific monoclonal antibody LM6<sup>137</sup> was also used to validate the arabinose containing oligosaccharide probes. The antibody showed strong binding to the linear  $\alpha$ 1,5-arabinose sequences with a chain-length dependency from DP-3 up to DP-6. Binding was also observed to the branched arabinan probe presenting a mixture of  $\alpha$ 1,2- and  $\alpha$ 1,3-arabinose substituted tetrasaccharide and  $\alpha$ 1,3-arabinose substituted pentasaccharide (probe 181,

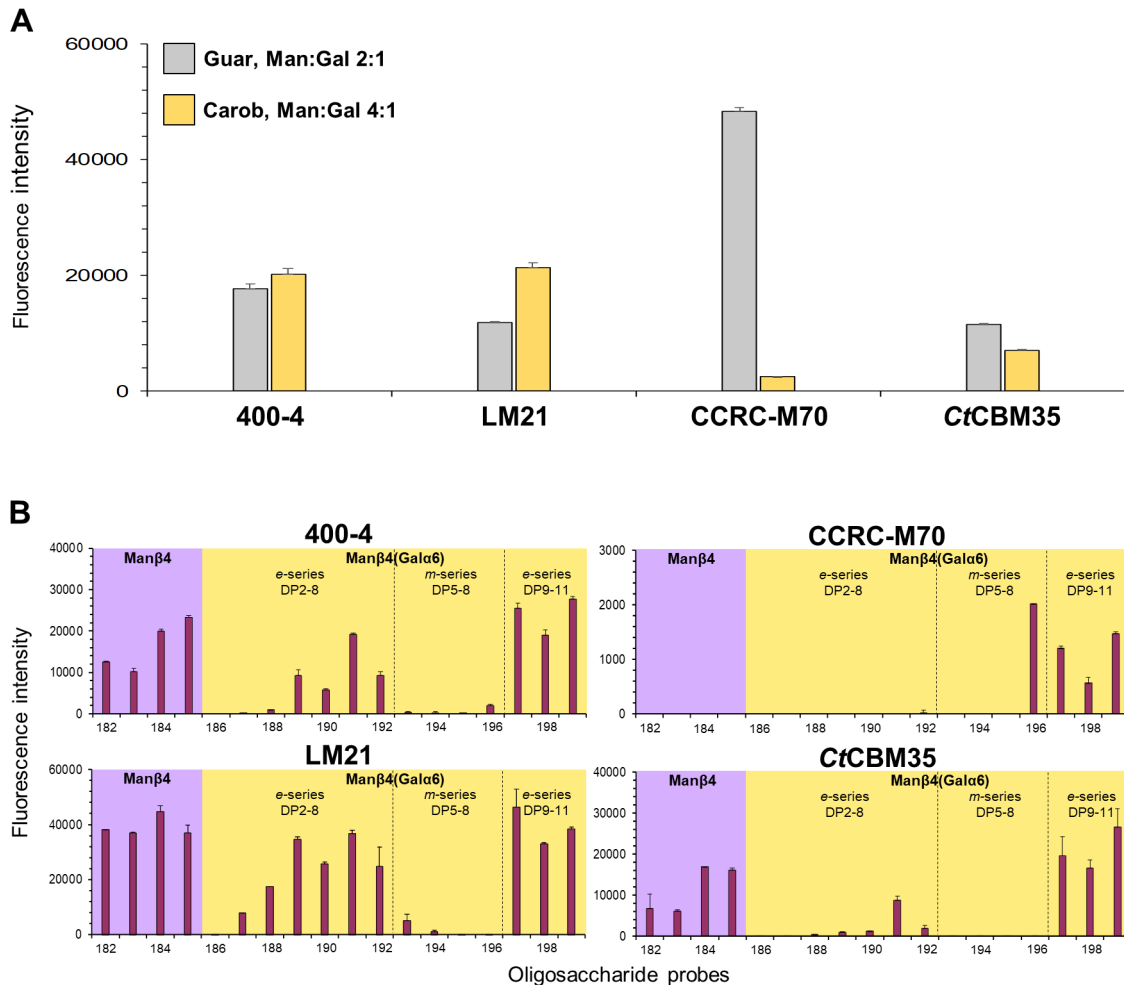
Figure 2.3B). As this antibody showed a chain-length dependency, the binding observed is likely to the latter.

### 2.2.2.3 Differential recognition of linear $\beta$ 1,4-mannans and branched galactomannans

The  $\beta$ 1,4-mannose containing NGL probes were prepared from oligosaccharide fractions obtained by partial hydrolysis of  $\beta$ 1,4-mannans and  $\alpha$ 1,6-galactose substituted carob galactomannans (Tables S2.1 and S2.2, probes 182 to 199). The latter comprised oligosaccharides obtained from two different sources: 1) after size exclusion fractionation of a carob galactomannan hydrolysate mixture (Elicityl preparation), designated as *e-series* (DP-2 to DP-11, probes 186-192 and 197-199); 2) after HPTLC purification of the NGLs derived from a carob di-galactosyl mannose pentasaccharide (Megazyme preparation), designated as *m-series* (DP-5 to DP-8, probes 193 to 196).

To corroborate their backbone sequences, the microarrays were probed with the well characterised monoclonal antibodies 400-4 and LM21, which are specific for linear  $\beta$ 1,4-mannose oligosaccharides<sup>138,139</sup>. These antibodies were initially analysed for their binding to guar and carob galactomannan polysaccharides, which present different degrees of  $\alpha$ 1,6-galactose substitution of the  $\beta$ 1,4-mannan backbone, that is, Man:Gal ratios of 2:1 and 4:1, respectively (Figure 2.4A). The 400-4 and LM21 antibodies bound to both polysaccharides, but the lower binding intensity showed by LM21 to guar galactomannan, revealed a lower tolerance of this antibody for Gal substitutions. In the NGL-microarrays, the 400-4 and LM21 exhibited the predicted binding to the linear  $\beta$ 1,4-manno-oligosaccharides (Figure 2.4B), with 400-4 showing a chain-length dependency up to DP-8 and LM21 binding with similar intensities to these probes. The antibodies also bound to the galactomannan-derived NGLs of the *e-series*, but with different binding profiles: 400-4 only bound to probes with DP-5 and higher and LM21 bound to probes with DP-3 up to DP-11. This in accord with the  $\beta$ 1,4-mannose chain-length requirements of these antibodies. Remarkably, none of the antibodies bound to the NGLs of the *m-series*. These results suggest a different ratio of  $\alpha$ 1,6-galactose substitutions of the  $\beta$ 1,4-mannan backbone in the fractions from *e-series* and *m-series* of the same DP.

Aiming at gaining more information about the possible differential recognition of the galactose-substituted oligosaccharide fractions of these series, the microarrays were also probed with the guar galactomannan-directed antibody CCRC-M70<sup>140</sup>, which to our knowledge has no oligosaccharide-specificity reported to date. In contrast with the other anti-mannan antibodies, CCRC-M70 showed a highly restricted binding profile and bound only to the galactomannan DP-8 probe of the *m-series* (probe 196), and DP-9 to DP-11 probes of the *e-series* (probes 197-199). This showed evidence for a lower Man:Gal ratios in their oligosaccharide sequences and recognition of a galactosylated epitope by CCRC-M70 that is only present in these NGL probes. In accordance to this, CCRC-M70 antibody showed a strong binding to guar galactomannan, but



**Figure 2.4. Microarray analysis of mannan-binding proteins. (A)** Microarray binding results of mannan- and galactomannan-specific antibodies LM21, 400-4, CCRC-M70 and CtCBM35 to guar and carob galactomannan polysaccharides. **(B)** The binding signals of  $\beta$ 1,4-mannan-specific monoclonal antibodies 400-4, LM21 and CtCBM35, and anti-galactomannan CCRC-M70 are depicted as means of fluorescence intensities of duplicate spots at 5 fmol of oligosaccharide probe arrayed (with error bars) and are representative of at least two independent experiments. Numerical scores are given in Table S2.4. The different carbohydrate groups are indicated in the coloured panels. The microarrays included 4 linear  $\beta$ 1,4-mannan-NGLs of DP-4, to DP-6 and DP-8, 14  $\beta$ 1,4-xylan( $\alpha$ 1,6-galactose)-NGLs from two different sources, *e*-series with sizes ranging from DP-2 to DP-11, and *m*-series with range from DP-5 to DP-8. Carbohydrate sequence information on these probes is in Table S2.1.

only marginal binding to carob galactomannan (Figure 2.4A), which emphasize its requirement for highly galactose-substituted sequences.

The *C. thermocellum* CBM of family 35 (CtCBM35<sub>Cthe\_2811</sub>) was also selected for analysis, as this CBM binds to  $\beta$ 1,4-linked mannose sequences in galactomannan and glucomannan polysaccharides, with higher affinity for the latter<sup>141</sup>, but no oligosaccharide-specificity has been reported to our knowledge. In this analysis, CtCBM35<sub>Cthe\_2811</sub> was shown to bind to linear  $\beta$ 1,4-linked mannose oligosaccharides, exhibiting a chain-length dependency up to DP-6. CtCBM35<sub>Cthe\_2811</sub> also bound to the *e*-series DP-7 galactomannan oligosaccharide (probe 191), and with stronger intensities to the DP-9 to DP-11 sequences (probes 197 to 199), but not to the galactomannan probes of the *m*-series.

In sum, the results of the microarray analysis corroborated the presence of  $\beta$ 1,4-linked mannose oligosaccharides in the isolated fractions, and the higher degree of Gal substitutions in the galactomannan from the *m*-series and in the fractions of longer DPs from the *e*-series.

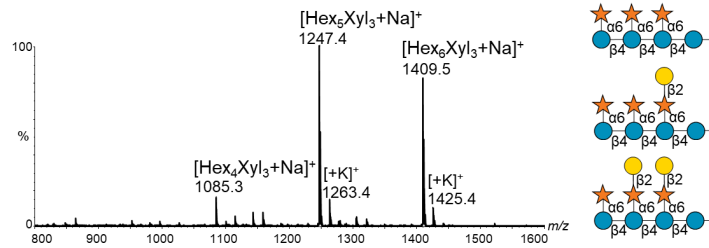
#### 2.2.2.4 Recognition studies of xyloglucan oligosaccharides using complex oligosaccharide mixtures and a deconvolution strategy

The xyloglucan NGL probes included in the glucan and hemicellulose oligosaccharide microarrays comprised xyloglucan oligosaccharide sequences from two different sources: 1) tamarind xyloglucans, which are  $\beta$ 1,4-glucans with  $\alpha$ 1,6-linked xylose branches that can also be substituted with a  $\beta$ 1,2-linked galactose; and 2) apple fucosylated-xyloglucans, in which the branch terminal  $\beta$ 1,2-galactose is substituted with an  $\alpha$ 1,2-linked fucose (probes 200-204, Figure 2.1 and S2.1 and Tables S2.1 and S2.2). The NGLs were prepared from oligosaccharide mixtures with multiple components as shown by the MALDI-MS analysis in Figures 2.5A and S2.2A. Although with the derivatization to the lipid a second level of purification was possible using HPTLC or silica columns, multiple components were still observed in some of the AO-NGLs prepared (Table 2.1): e.g. for probe 202 a major component with DP-8 was observed, therefore was designated Xyl-Glucan-8, but a component with DP-7 was also present; for the fucosylated-xyloglucan NGL probes, Fuc-Xyl-Glucan-6 (probe 203) and Fuc-Xyl-Glucan-9 (probe 204) were also mixtures and both fucosylated and non-fucosylated components were present. Despite the fact that NGL probes were arrayed as multiple components, the binding they elicited with the well characterised anti-xyloglucan monoclonal antibodies LM24 and LM25 were in accordance with their reported specificities<sup>92</sup> (Figure S2.1 and Table S2.3). Both antibodies required the substitution of the  $\beta$ 1,4-glucose backbone chain: LM25 bound to all the xyloglucan NGLs probes that are substituted with  $\alpha$ 1,6-linked xylose, whereas LM24 showed more restricted binding requiring at least substitution with one  $\beta$ 1,2-linked galactose (probes 200-204). Furthermore, the analysis with the anti-xyloglucan antibody CCRC-M1 that requires the  $\alpha$ -linked fucose<sup>142</sup>, and the  $\alpha$ -fucose-specific *Aleuria aurantia* lectin (AAL)<sup>143</sup>, could distinguish for the presence of these epitopes as they bound only to the probes derived from the fucosylated-xyloglucan (probes 203-204).

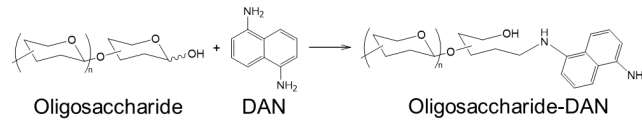
Using the xyloglucan oligosaccharide mixture as an example, a deconvolution strategy was applied to obtain more homogeneous and sequence-defined NGL probes for investigation of their interaction with proteins. The method involved conjugation of the oligosaccharides with a bi-functional UV/fluorescence tag 1,5-diaminonaphthalene (DAN, Figure 2.5B) aiming at sensitive detection in HPLC to allow detailed fine separation/purification of the structurally similar oligosaccharides, before its conversion into NGL probes for microarray construction using the second functional group. To this end, the reducing tamarind xyloglucan oligosaccharide mixture was conjugated to DAN. HPLC analysis showed that each DP component of the mixtures could be resolved as discrete individual peaks (Figure 2.5C). MS analysis prior (Figure 2.5A) and after



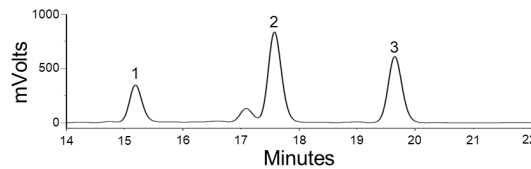
**A MALDI-MS of Xyloglucan DP7-9**



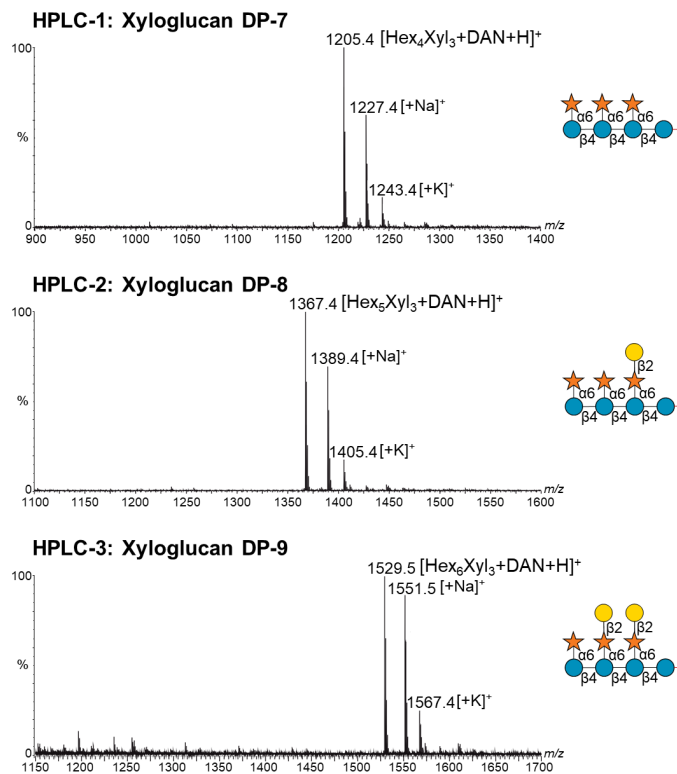
**B Fluorescence tagging strategy**



**C HPLC of Xyloglucan-DAN DP7-9**

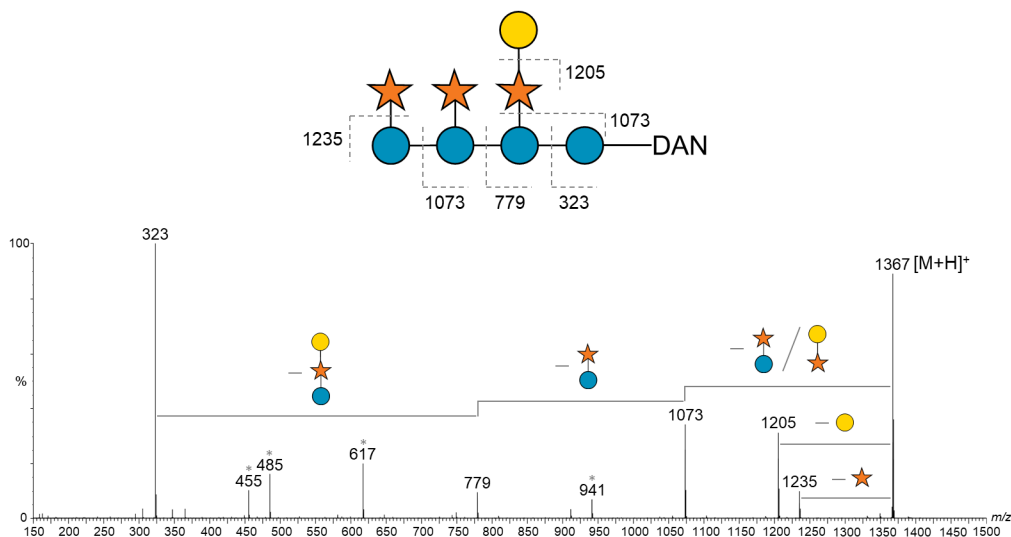


**D ESI-MS of HPLC fractions**



**Figure 2.5. Deconvolution of the xyloglucan oligosaccharides from tamarind. (A)** MALDI-MS spectrum of xyloglucan oligosaccharides investigated from tamarind comprising DP-7 to DP-9. **(B)** Schematic representation of bi-functional derivatization of by 1,5-diaminonaphthalene (DAN) conjugation used for fluorescence tagging of xyloglucan oligosaccharides. **(C)** HPLC separation of DAN-conjugated fractions of xyloglucan (DP-7, DP-8 and DP-9). **(D)** Positive-ion ESI-MS spectra of the HPLC fractions obtained. Xyloglucan oligosaccharides are depicted. The red link represents the oligosaccharides reducing end conjugated to DAN.

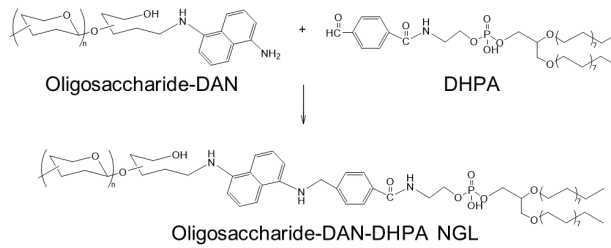
DAN-conjugation (Figure 2.5D) allowed to assess the decrease on sample heterogeneity upon purification and confirmed the molecular mass of each of the individual components. Xyloglucan-DAN DP-8 fraction (Figure 2.5D, HPLC-2) was selected for sequencing by positive-ion ESI-MS/MS (Figure 2.6), prior to NGL probes preparation. The purified xyloglucan-DAN oligosaccharides were then used for preparation of NGLs by reductive amination with the aldehyde functionalized lipid *N*-(4-formylbenzamide)-1,2-dihexadecyl-*sn*-glycero-3-phosphoethanolamine (DHPA) (Figure 2.7A). The analysis of the NGL products by HPTLC showed a single NGL product (Figure 2.7B). MALDI-MS of the purified NGL products corroborated the assignment of their molecular masses, as shown in Table 2.2 and Figure 2.7C. The same strategy was then applied for fucosylated-xyloglucan oligosaccharides from apple (Figures S2.2 and S2.3), however, is still under optimization for obtaining more homogeneously enriched NGL fractions.



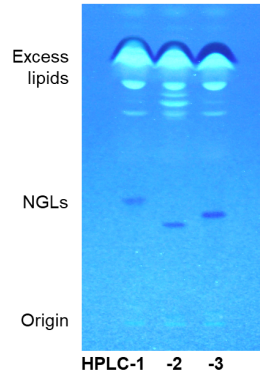
**Figure 2.6. Positive-ion ESI-MS/MS product-ion spectra used for sequencing of xyloglucan-DAN DP-8.** Xyloglucan-DAN DP-8 fraction (Figure 2.5D, HPLC-2) was used for sequencing and branching points are represented. The symbolic oligosaccharide structure is depicted representing the fragmentation ions obtained. (\*) represent double cleavage product-ions.

To interrogate the recognition of the xyloglucan-DAN-DHPA NGLs, microarrays were constructed of 10 xyloglucan probes (Table S2.5): newly xyloglucan-DAN-DHPA-NGLs (probes 1 to 3, 7 and 8) and the AO-NGLs arrayed in the hemicellulose platform as controls (probes 4 to 6, 9 and 7, corresponding to probes 200-204 in Table S2.1). The microarrays were probed with the monoclonal antibodies LM24, LM25 and CCRC-M1 and the lectin ALL, as shown in Figure 2.8A. Remarkably, the analysis showed in overall similar binding profiles of the proteins to the xyloglucan-DAN-DHPA-NGLs and AO-NGL controls. LM24 bound only to the NGL probes that were substituted with  $\beta$ 1,2-linked galactose, showing stronger binding to the di-galactosylated probe. This antibody also bound to the Fuc-Xyloglucan probes, suggesting that substitution of the galactose residue with  $\alpha$ 1,2-linked fucose could be tolerated. LM25 showed a different tendency and bound preferentially to the probes substituted with  $\alpha$ 1,6-linked xylose, and exhibited less

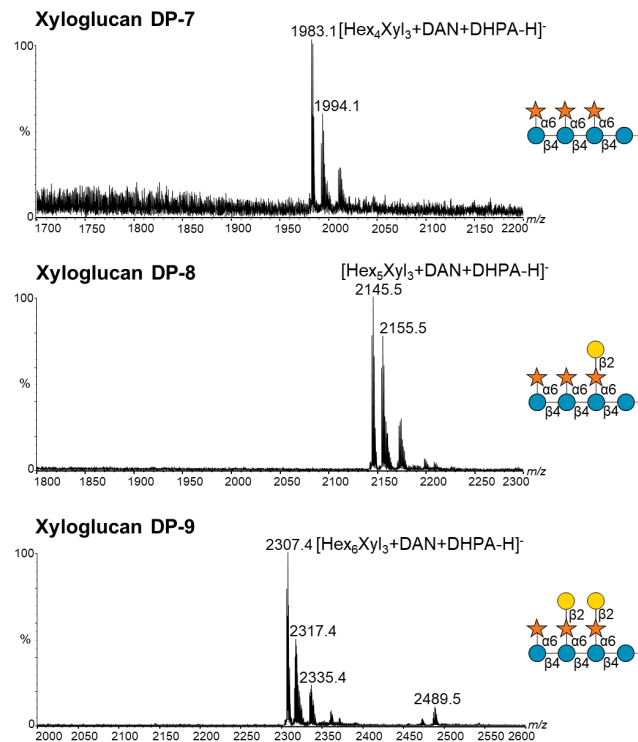
**A Preparation DAN-DHPA NGLs**



**B HPTLC of Xyloglucan-DAN-DHPA NGLs**



**C MALDI-MS of NGL fractions**



**Figure 2.7. Preparation of the xyloglucan-DAN-NGL probes from tamarind included in the new xyloglucan microarrays.** (A) Schematic representation of preparation of xyloglucan-DAN-DHPA-NGLs after HPLC separation of DAN-conjugated DP-7, DP-8 and DP-9 fractions (Figure 2.5, HPLC-1, -2 and -3). (B) HPTLC analysis of conjugation mixtures DAN-DHPA-NGLs of xyloglucan fractions revealed by primulin-staining. (C) MALDI-MS analysis of the xyloglucan-DAN-DHPA NGL probes printed and validated in the xyloglucan microarrays. Xyloglucan oligosaccharides are depicted. The red link represents the oligosaccharides reducing end conjugated to DAN.

**Table 2.2. MALDI-MS analysis of xyloglucan-DAN-DHPA NGLs investigated.**

Xyloglucan probes	DP <sup>a</sup>	[M-H] <sup>-</sup> calculated <sup>b</sup>	[M-H] <sup>-</sup> detected <sup>c</sup>
Xyl-Glucan-7-DAN-DHPA	7	1983.12	1983.1
Xyl-Glucan-8-DAN-DHPA	8	2145.17	2145.5
Xyl-Glucan-9-DAN-DHPA	9	2307.23	2307.4
Fuc-Xyl-Glucan-7-DAN-DHPA	7	1997.14	1997.9 (1869.9, 2007.9, 2178.9) <sup>d</sup>
Fuc-Xyl-Glucan-10-DAN-DHPA	10	2453.28	2453.2 (2327.2)

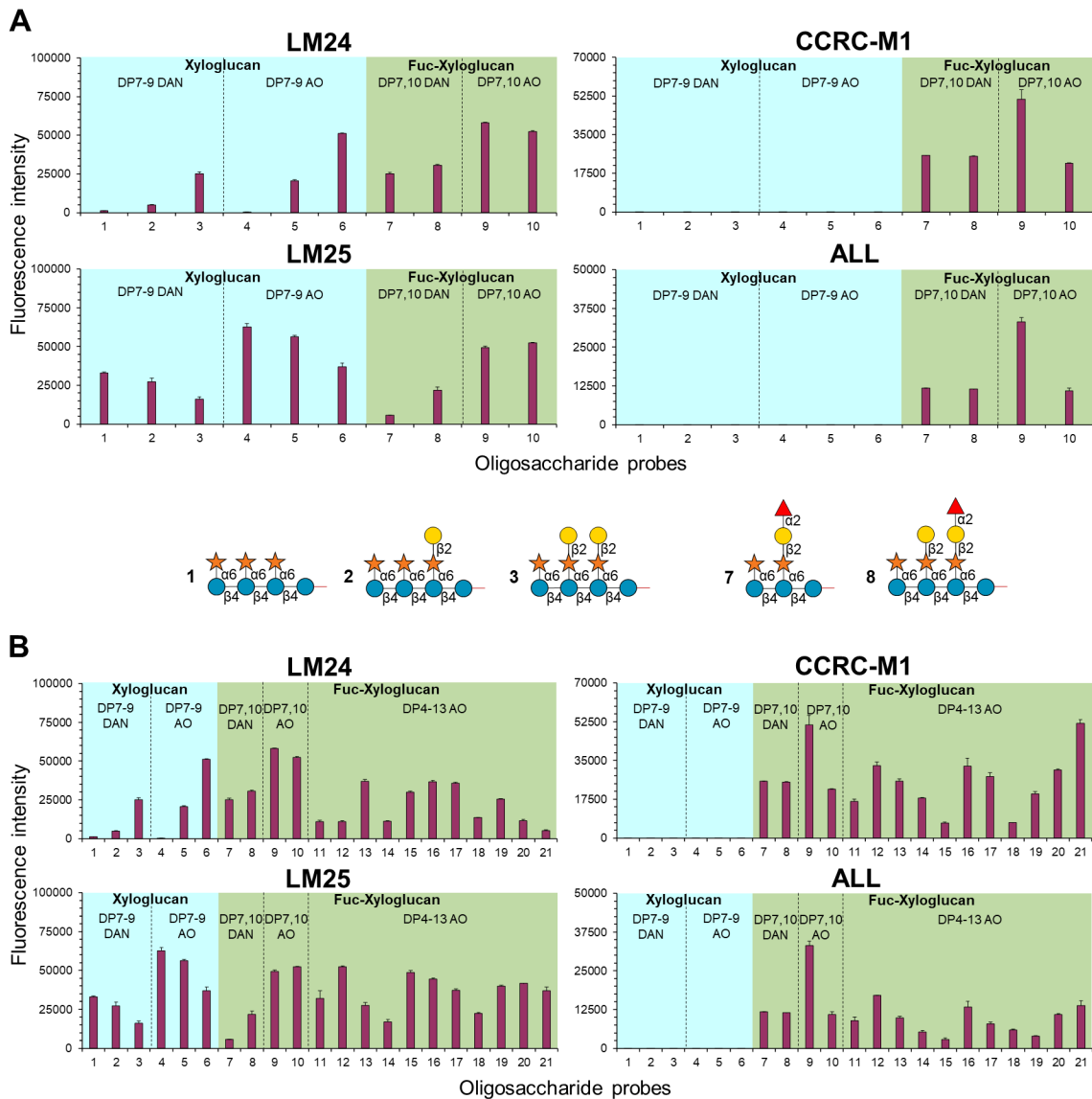
<sup>a</sup>Degree of polymerization (DP) for the major components in each fraction; <sup>b</sup>Calculated masses for major components are given; <sup>c</sup>Negative-ion MALDI-MS was used for the analysis of DAN-DHPA-NGLs and [M-H]<sup>-</sup> were detected; <sup>d</sup>Where multiple components were detected, relative intensities of ions greater than 20% are shown in brackets.

tolerance for galactose and fucose substituents near the non-reducing end (probes 3 and 7). CCRC-M1 bound equally well to the Fuc-Xyloglucan probes 7 to 10, pointing to a single  $\alpha$ -Fuc-(1,2)- $\beta$ -Gal epitope as the minimum requirement for carbohydrate recognition by this antibody.

Aiming to increase the diversity of xyloglucan sequences included in microarray platforms and to gain information about the fucosylated-xyloglucan fractions used as start materials, a highly heterogeneous xyloglucan oligosaccharide mixture from apple was investigated. After deconvolution by size exclusion chromatography, the obtained fractions were converted to AO-NGLs (DP range 4 to 13) for preparation of microarrays and their analysis with the anti-xyloglucan antibodies. Upon purification of the NGL products and MALDI-MS analysis, 10 new xyloglucan-AO-NGL fractions were obtained and printed on nitrocellulose-coated glass slides at 2 and 5 fmol/spot (probes 11 to 21, Table S2.5). The compositions of the major components in these fractions, including acetyl groups, are proposed in Table 2.3, based on MALDI-MS analysis of the NGLs and the microarray results. While probing with the xyloglucan-directed antibodies and ALL was successful, the binding profiles showed high heterogeneity of the new AO-NGLs (Figure 2.8B), as all the probes were recognized by the antibodies and AAL. Given these results, further deconvolution steps will be required prior to DAN-conjugation to obtain less heterogeneous mixtures.

## 2.3 Discussion

Aiming at addressing the need for increased structural diversity and to cover a wide range of plant-oligosaccharide chain-length in microarray platforms developed to date, we report here an oligosaccharide-NGL microarray platform, comprised of naturally-derived hemicellulose-related oligosaccharides. The microarray of these sources further extended and complemented the previously constructed glucan microarray platform<sup>32</sup>. The microarray analysis with the different carbohydrate-binding proteins investigated validated the glucan and hemicellulose oligosaccharide microarrays for their application to plant cell wall carbohydrate recognition. The



**Figure 2.8. Validation and analysis of the new xyloglucan microarrays.** Binding signals of non-fucosylated- and fucosylated-xyloglucan-specific monoclonal antibodies LM24, LM25 and CCRC-M1 and  $\alpha$ -fucose-specific lectin ALL are depicted as means of fluorescence intensities of duplicate spots at 5 fmol of oligosaccharide probe arrayed (with error bars) and are representative of at least two independent experiments. The different xyloglucan sources are indicated in the coloured panels. **(A)** The microarrays included the xyloglucan-DAN-DHPA-NGLs probes 1 to 3, 7 and 8, as controls and the initially arrayed AO-NGLs 4 to 6, 9 and 10 (probes 1 to 10). **(B)** The new xyloglucan-AO-NGLs derived from the deconvolution of xyloglucan oligosaccharide mixture from apple were also included for analysis (probes 11 to 21). Carbohydrate sequence information of the DAN-DHPA- and AO-NGL probes is in Table S2.5.

results showed robust binding signals and differing patterns with respect to oligosaccharide structural type, linkage and sources by the anti-plant carbohydrate antibodies, lectins and bacterial CBMs. The results were in overall consistent with data obtained previously on the specificity of these proteins using different techniques. Additionally, the wide range of oligosaccharide sequences covered in the microarrays allowed to broaden the knowledge on their carbohydrate binding and specificities.

**Table 2.3. MALDI-MS analysis of the new xyloglucan-AO-NGL probes generated.**

<b>Xyloglucan probes</b>	<b>DP<sup>a</sup></b>	<b>[M-H]<sup>-</sup> calculated<sup>b</sup></b>	<b>[M-H]<sup>-</sup> detected<sup>c,d</sup></b>	<b>Main composition<sup>e</sup></b>
Xyl-Glucan DP4-AO	4	1353.77	1354.0 (1690.2)	Hex <sub>3</sub> Xyl <sub>1</sub> (Hex <sub>4</sub> Xyl <sub>2</sub> (Ac))
Xyl-Glucan DP5-AO	5	1485.81	1486.1 (1354.0)	Hex <sub>3</sub> Xyl <sub>2</sub> (Hex <sub>3</sub> Xyl <sub>1</sub> )
Xyl-Glucan DP6a-AO	6	1647.86	1648.1	Hex <sub>4</sub> Xyl <sub>2</sub>
Xyl-Glucan DP6b-AO	6	1689.86	1690.1 (1836.2)	Hex <sub>4</sub> Xyl <sub>2</sub> (Ac) (Hex <sub>4</sub> Xyl <sub>2</sub> Fuc(Ac))
Xyl-Glucan DP7-AO	7	1779.90	1780.1	Hex <sub>4</sub> Xyl <sub>3</sub>
Xyl-Glucan DP8a-AO	8	1941.96	1942.3 (2089.4, 2147.4, 2293.5)	Hex <sub>5</sub> Xyl <sub>3</sub> (Hex <sub>5</sub> Xyl <sub>3</sub> Fuc, Hex <sub>6</sub> .Xyl <sub>3</sub> (Ac), Hex <sub>6</sub> Xyl <sub>3</sub> Fuc(Ac))
Xyl-Glucan DP8b-AO	8	1941.96	1942.2 (2105.3, 2251.4)	Hex <sub>5</sub> Xyl <sub>3</sub> (Hex <sub>6</sub> Xyl <sub>3</sub> , Hex <sub>5</sub> Xyl <sub>3</sub> Fuc)
Xyl-Glucan DP9-AO	9	2104.01	2105.2 (2251.4)	Hex <sub>6</sub> Xyl <sub>3</sub> (Hex <sub>6</sub> Xyl <sub>3</sub> Fuc)
Xyl-Glucan DP11/12-AO	11	2454.12	2454.6 (2586.6)	Hex <sub>7</sub> Xyl <sub>3</sub> Fuc(Ac) (Hex <sub>7</sub> Xyl <sub>4</sub> Fuc(Ac))
Xyl-Glucan DP13-AO	13	2706.21	2707.7	Hex <sub>8</sub> Xyl <sub>4</sub> Fuc

<sup>a</sup>Degree of polymerization (DP) for the major components in each fraction; <sup>b</sup>Calculated masses for major components are given; <sup>c</sup>Negative-ion MALDI-MS was used for the analysis of the AO-NGLs and [M-H]<sup>-</sup> were detected; <sup>d</sup>Where multiple components were detected, relative intensities of ions greater than 20% are shown in brackets; <sup>e</sup>Proposed composition of the major components as detected by negative-ion MALDI-MS.

The binding recognition of linear  $\alpha$ 1,2-,  $\alpha$ 1,3-,  $\alpha$ 1,4- and  $\alpha$ 1,6-glucan-oligosaccharides was differentiated for  $\alpha$ -glucan recognising proteins, where the glucan-binding property of human malectin was shown predominantly to linear  $\alpha$ 1,3-glucan oligosaccharides, exhibiting a chain-length dependency up to the tetrasaccharide. The additional binding to  $\alpha$ 1,4-,  $\alpha$ 1,6- and  $\beta$ 1,3-linked glucose oligosaccharides highlights the plasticity of the malectin binding site to accommodate other glucose linkages dependent of their conformation and linkage. The ligand-specificity of *CtCBM25*<sub>Cthe\_0956</sub> was assigned to  $\alpha$ 1,4-linked glucose epitopes in linear or mixed-linked glucan-oligosaccharides. The chain-length dependency observed for sequences longer than DP-4 points to a type B topology of this CBM's binding site, able to accommodate a minimum of 4  $\alpha$ 1,4-glucose units. The microarray analysis also enabled to discriminate specific binding towards linear  $\beta$ 1,3-,  $\beta$ 1,4- or mixed-linked  $\beta$ 1,3- $\beta$ 1,4-glucans and the chain-length requirements of  $\beta$ -glucan recognising proteins. The broader binding patterns exhibited by *CmCBM6-2* emphasize the plasticity of its type B and C binding sites to a wide range of  $\beta$ -linked hexoses. The analysis also showed the influence of the spacing and positioning of  $\beta$ 1,3-linkages within mixed-linked  $\beta$ 1,3- $\beta$ 1,4-glucan chains for binding by the antibody 400-3 and *CtCBM11*<sub>Cthe\_1472</sub>.

The carbohydrate-binding proteins analysed on the new hemicellulose-related microarrays developed, allowed not only to increase knowledge on the carbohydrate-specificities for the proteins analysed, but also to shed light into the sequence of some of the NGL-oligosaccharides derived from heterogeneous mixtures, for which sequence is still under full assignment.

The binding patterns of  $\beta$ 1,4-linked xylan-specific antibodies LM10 and LM11 and *CBM22-2*<sub>Cthe\_0912</sub> highlighted the differences in the binding mode by these proteins, which can

relate to their binding sites topology. LM10 showed a specificity towards the non-reducing end of  $\beta$ -linked xylose oligosaccharides, although being able to accommodate to some extent  $\alpha$ 1,2-/1,3-arabinose terminal branches at this position, which points to a cavity-type antibody. LM11 on its turn required the internal  $\beta$ -xylose backbone accommodating longer epitopes of at least 3 xylose monomers, suggesting a groove-type antibody. The binding of CtCBM22-2<sub>Cthe\_0912</sub> to linear  $\beta$ -linked xylan oligosaccharides with a chain-length dependency from DP-4 up to DP-8 depicts the typical binding profile observed for a groove-type B CBM. Remarkably, the fact that both the anti- $\beta$ 1,4-xylan antibodies and CtCBM22-2<sub>Cthe\_0912</sub> bound strongly to the extent of the mixed-linked  $\beta$ 1,3- $\beta$ 1,4-xylose probes (probes 153 to 164), allows to infer that the  $\beta$ 1,3-linkage, if present, might be positioned at the reducing end of the smaller DP probes, and hence not exposed to the proteins. On the one hand, in longer DP fractions,  $\beta$ 1,3-linkages could be present every 4 to 5 xylose monomers<sup>144</sup>, exposing  $\beta$ 1,4-linked xylose stretches available for binding. On the other hand, binding to these sequences might also not be influenced by the presence of a  $\beta$ 1,3-linkage. While the microarray analysis allows to infer about these oligosaccharide sequences, their defined sequence needs to be confirmed, recurring for example to diagnostic fragmentation method by negative-Ion ESI-CID-MS/MS<sup>32</sup>.

The microarray analysis with mannan-directed proteins showed the specific and different chain-length requirement for binding to linear  $\beta$ 1,4-mannose sequences or the requirement of the  $\alpha$ -galactose substitutions, highlighting proteins that selectively bind to linear mannans or galactomannans. While 400-4 showed a chain-length dependency of the  $\beta$ 1,4-linked mannan backbone, pointing to a groove-type antibody, LM21 bound to shorter chains, pointing to recognition of the non-reducing end of the  $\beta$ 1,4-linked mannans and to a cavity-type antibody. The assignment of the specificity for CtCBM35<sub>Cthe\_2811</sub> to the  $\beta$ 1,4-linked mannan backbone of mannans with a chain-length of at least 4 residues is in agreement with a type B CBM, as generally observed for family 35<sup>141</sup>. In contrast with these proteins, the antibody CCRC-M7 requires the  $\alpha$ -galactose substitutions of  $\beta$ 1,4-mannose backbone for binding. The lack of binding to the di-galactosyl-mannopentaose probe (probe 195) and the higher binding intensity to probe 196 (DP-8), may indicate that, for recognition to occur, CCRC-M70 requires either more than 2  $\alpha$ -galactose residues or a  $\beta$ 1,4-mannose backbone of more than 5 residues. The integration of the microarray data showed that probes from the *e*-series are composed mainly by linear  $\beta$ 1,4-mannose up to the DP-9 probe, and that the *m*-series comprise sequences with a higher ratio of  $\alpha$ -galactose substitutions. Using the same rationale, the binding profiles to the longer *e*-series probes of DP-9 to DP-11, points to longer DPs of the  $\beta$ 1,4-mannose backbone and the presence of at least 2 or higher  $\alpha$ -galactose substitutions in these sequences. Sequencing of these galactomannan-derived oligosaccharide sequences is under way using the established ESI-MS/MS method and NMR analysis<sup>32</sup>.

The analysis with the xyloglucan microarrays revealed the different binding features and epitopes for 3 anti-xyloglucan antibodies: LM24 towards  $\beta$ 1,2-linked galactose substituted xyloglucans;

LM25 towards  $\alpha$ 1,6-linked xylose substituted xyloglucans, exhibiting less tolerance for highly substituted  $\beta$ 1,2-linked galactose or  $\alpha$ 1,2-linked fucose xyloglucans; and CCRC-M1 towards fucosylated xyloglucans requiring a single  $\alpha$ -Fuc-(1,2)- $\beta$ -Gal as the minimum recognition epitope. Deconvolution of the detailed oligosaccharide epitopes recognized by these antibodies is detrimental for their application as research tools for detection and characterization of specific carbohydrate sequences present in plant polysaccharides.

Aiming at diversifying the microarrays with sequence-defined NGL-oligosaccharides, the deconvolution method with conjugation to the bi-functional DAN, followed by derivatization to an aldehyde-functionalized lipid, was attempted to separate the neutral xyloglucan oligosaccharides and to achieve a relatively homogenous population of DAN-DHPA-NGL probes. This strategy showed potential to enrich each of the tamarind xyloglucan oligosaccharide fractions. However, application of the method to more heterogeneous oligosaccharide mixtures, like the fucosylated-xyloglucans from apple, requires further optimisation in order to improve the conjugation yields and chromatographic resolution. Additionally, as the purified oligosaccharides from complex plant polysaccharides are frequently obtained in limited amounts, this precludes the use of conventional NMR method for sequencing, which is currently ongoing for xyloglucan fractions, as well as for galactomannans from the *m*-series and *e*-series.

## 2.4 Conclusions

This work demonstrated the effectiveness and versatility of the NGL oligosaccharide microarrays to assess the binding profiles of glucan- and hemicellulose-recognition systems with wide specificities for oligosaccharide sequences, linkages, and anomeric configurations. Additionally, the development of the method of bi-functional conjugation allowing for oligosaccharide purification and NGL probe derivatization, proved its potential for the deconvolution of neutral oligosaccharide mixtures and is under optimisation to prepare sequence-defined plant oligosaccharides derived from natural sources. This can be achieved by combining methods of depolymerisation of the polysaccharides and multistage purification of the oligosaccharides with the use of specific enzymes to reduce heterogeneity. Furthermore, mass spectrometry and NMR can be combined to perform detailed structural analysis of these oligosaccharides.

While the goal of the work developed in this Chapter was to construct naturally-derived plant oligosaccharide microarrays, new specificities and epitope-level information was also obtained for antibodies and CBMs. These microarrays will be a valuable tool for functional analyses of proteins encoded by plant cell wall polysaccharide-degrading genes, while assisting the classification of newly identified CBMs or CBMs assigned to known families in the CAZy database. Towards this aim, the CBMs described here will be analysed in parallel with other *C. thermocellum* and *Ruminococcus flavefaciens* CBMs in the comparative CBM screening analyses presented in Chapter 3.



## 2.5 Experimental procedures

### 2.5.1 Monoclonal antibodies, CBMs and lectins used for probe structural validation and microarray quality control

The information on the plant cell wall carbohydrate-directed monoclonal antibodies, CBMs and lectins used for microarray quality control is given in Table S2.3. Carbohydrate-directed monoclonal antibodies 400-2, 400-3, and 400-4 were purchased from Biosupplies (Yagoona, Australia); LM5, LM6, LM10, LM11, LM21, LM24 and LM25 were purchased from Plant probes (Leeds, UK); and CCRC-M1 and CCRC-M70 were purchased from Agrisera (Vännäs, Sweden). Biotinylated lectin ALL was purchased from Vector Laboratories (Burlingame, California, US). Human malectin and CBMs were produced as recombinant proteins in *Escherichia coli*. Experimental procedure on the production of family 22, 25 and 35 CBMs from *C. thermocellum* (CtCBM22-2<sub>Cthe\_0912</sub>, CtCBM25<sub>Cthe\_0956</sub> and CtCBM35<sub>Cthe\_2811</sub>, respectively) will be detailed in Chapter 3, and Family 11 CBM from *C. thermocellum* (CtCBM11<sub>Cthe\_1472</sub>) in Chapter 4. Family 6 CBM from the *Cellvibrio mixtus* (CmCBM6-2) was provided by Harry Gilbert (University of Newcastle, UK). The recombinant CtCBM22-2<sub>Cthe\_0912</sub>, CtCBM25<sub>Cthe\_0956</sub>, CtCBM35<sub>Cthe\_2811</sub> and CmCBM6-2 contained N-terminal hexa-histidine tags (His-tag), and CtCBM11<sub>Cthe\_1472</sub> a C-terminal His-tag.

### 2.5.2 Sources of carbohydrates

The sources of the glucan oligosaccharides to construct the glucan oligosaccharide microarrays were described in Palma *et al.*<sup>32</sup> and are indicated in Table S2.1. A range of plant-related hexose and pentose oligosaccharides or oligosaccharide mixtures with different DPs and glycosidic linkages on linear or branched chains were used to construct the hemicellulose microarrays. These were obtained from the commercial suppliers Megazyme (Bray, Ireland) and Elicityl (Elicityl, Crolles, France) and were prepared by chemical or enzymatic partial depolymerisation of polysaccharides. The sources and analysis performed for the preparation of the AO-NGL probes are given in Table S2.1 and Table S2.2.

### 2.5.3 Preparation of oligosaccharide fractions

To prepare the oligosaccharide fractions of different degrees of polymerization (DP), the oligosaccharide mixtures of *Palmaria palmata* xylan, Ivory nut mannan, carob galactomannan, and tamarind xyloglucan were reconstituted in the smallest volume possible of deionized water (20 mg in 2 mL as reference) and fractionated by size-exclusion chromatography on a Bio-Gel P4 column (1.6×90 cm) eluted with deionized water at a flow rate of 15 mL/h. The column was previously calibrated with a standard dextran hydrolysate mixture (4 mg in 1mL water) containing 0.1 mg dextran polysaccharide (as the marker of the exclusion volume). The mixture was eluted with deionized water at a flow rate of 15 mL/h. The eluates were monitored on-line by refractive index and pooled according to their average DP, as determined by MALDI-MS. Quantitation of

oligosaccharide fractions was by dot-orcinol hexose assay<sup>72</sup> using glucose as standard. The solutions (1  $\mu\text{L}$ ) of oligosaccharides and standards were spotted on to a TLC plate using an accurate syringe. Visualization of hexose on TLC was made by orcinol staining after colour development followed heating at 105 °C in a vented oven. The hexose content in samples is determined from the calibration curve of density vs ( $\mu\text{g}$ ) hexose derived from the glucose standard samples.

#### 2.5.4 Preparation of AO-NGLs by oxime-ligation

The reducing oligosaccharides were converted to AO-NGLs by oxime-ligation using the aminoxy-functionalized (AO) lipid reagent 1,2-dihexadecyl-*sn*-glycero-3-phosphoethanolamine (DHPE), resulting in the AOPE lipid, following the Liu *et al* method<sup>86</sup>, and its modified procedure for the oligosaccharides with DP > 7 described in Zhang *et al.* 2016<sup>96</sup>. In brief, 100 nmol of oligosaccharides with DP < 7 were dried and incubated with 200 nmol AOPE in a solvent of  $\text{CHCl}_3/\text{MeOH}/\text{H}_2\text{O}$  (10:10:1, by volume) at ambient temperature for 16 h before slow solvent evaporation (over the course of 1 h) at 60 °C. Purification of the NGLs from reaction mixtures was carried out by semi-preparative HPTLC<sup>145</sup> for oligosaccharides in the range of DP-2 to DP-6, and silica gel cartridge for DP > 7. The purified NGLs were analyzed by MALDI-MS and HPTLC before quantitation by primulin staining on HPTLC plates<sup>145</sup>. Molecular masses of the NGLs were determined by MALDI-TOF-MS. The sizes and compositions of the oligosaccharide fractions obtained are indicated in Tables 2.1 and 2.3, which give the results of MALDI-MS analyses of the NGL probes derived from each of the oligosaccharide fractions.

#### 2.5.5 Preparation of DAN-conjugated xyloglucan oligosaccharides by reductive amination

Xyloglucan oligosaccharide mixtures were resolved through bi-functional derivatization by 1,5-diaminonaphthalene (DAN) conjugation, followed by HPLC separation. Briefly, 280  $\mu\text{L}$  of DAN at 100 nmol/ $\mu\text{L}$  in methanol (MeOH) was added to 400 nmol of dried oligosaccharide mixture and then dried under a nitrogen stream. Reduction was carried out by addition of 120  $\mu\text{L}$  of tetrabutylammonium cyanoborohydride (TBA) at 20  $\mu\text{g}/\mu\text{L}$ , freshly prepared in MeOH, followed by addition of 20  $\mu\text{L}$  glacial acetic acid and 60  $\mu\text{L}$  of MeOH. Incubation of the reaction mixture was at 80 °C for 3 h, upon which the mixture was dried under a nitrogen stream. Removal of excess DAN was done by chloroform extraction after addition of  $\text{CHCl}_3/\text{H}_2\text{O}$  1:1. The water phase was washed up to 6 times using 400  $\mu\text{L}$  of  $\text{CHCl}_3$  to ensure complete removal of reaction agents. The water phase was freeze-dried, and the DAN-conjugated product was then resuspended in 100  $\mu\text{L}$  of  $\text{ACN}/\text{H}_2\text{O}$  1:1 for HPLC loading. HPLC was carried out using a normal-phase X-Bridge amide column (3.5  $\mu\text{m}$ , 4.6 $\times$ 250 mm, Waters, Elstree, UK) for purification of the DAN-conjugated oligosaccharides prepared. Elution was performed by a gradient of  $\text{ACN}/\text{H}_2\text{O}$  (from 80:20 to 20:80) at a flow rate of 0.7 mL/min and detection by UV absorption at 328 nm. Each fraction

collected was analysed by ESI-MS or MALDI-MS (Figures 2.5A and S2.2). Given the purple colour of the DAN solution, DAN-conjugated oligosaccharides could not be quantified by the standard dot-ornicol assay<sup>72</sup>. To this end, quantitation of the purified xyloglucan-DAN fractions was carried out by UV spectrophotometry at 328 nm, using a calibration curve of DAN serial dilutions in ACN/H<sub>2</sub>O 1:1 (Figure S2.4).

### 2.5.6 Preparation of DHPA-NGLs by reductive amination

DAN-conjugated amino-terminating xyloglucan oligosaccharides were converted to NGLs by reductive amination using an aldehyde-functionalized phospholipid reagent *N*-(4-formylbenzamide)-1,2-dihexadecyl-*sn*-glycero-3-phosphoethanolamine (DHPA) (Chunxia Li, Angelina S. Palma, Pengtao Zhang *et al.*, 2020, submitted). 40  $\mu$ L of DHPA at 10 nmol/ $\mu$ L in CHCl<sub>3</sub>/MeOH/ 1:1 was added to 100 nmol of dried oligosaccharides, followed by 10  $\mu$ L of TBA at 20  $\mu$ g/ $\mu$ L, freshly prepared in MeOH, as a reducing agent, and 60  $\mu$ L of CHCl<sub>3</sub>/MeOH/ 1:1. The reaction mixture was incubated for 16 h at 60 °C, upon which 200  $\mu$ L of CHCl<sub>3</sub>/MeOH/H<sub>2</sub>O 25:25:8 was added. Purification of the NGLs from reaction mixtures was carried out using silica cartridges (Waters, 3CC, 100 mg). The purified NGL fractions were analyzed by HPTLC<sup>145</sup>, followed by MALDI-MS ( Figures 2.7C and S2.3B). Quantitation of the purified DAN-DHPA-NGLs was performed by UV spectroscopy at 328 nm, using DAN standard solutions prepared as series dilutions in CHCl<sub>3</sub>/MeOH/H<sub>2</sub>O 25:25:8 (Figure S2.4). The sizes and compositions of the oligosaccharide fractions obtained are indicated in Table 2.2, which give the results of MALDI-MS analyses of the DAN-DHPA-NGL probes derived from each of the oligosaccharide fractions.

### 2.5.7 MALDI Mass Spectrometry

MALDI-MS was carried out for molecular mass determination of oligosaccharides and NGLs on an Axima Resonance mass spectrometer with a QIT-TOF configuration (Shimadzu, Manchester, UK). A nitrogen laser (with a power setting at between 80-100 V) was used to irradiate samples at 337 nm, with an average of 200 shots accumulated. The oligosaccharides were dissolved in H<sub>2</sub>O and NGLs in CHCl<sub>3</sub>/MeOH/H<sub>2</sub>O 25:25:8, at a concentration of ~10 pmol/ $\mu$ L, and 0.5  $\mu$ L was deposited on the sample target together with a matrix of 2',4',6',-trihydroxyaceto-phenone for analysis.

### 2.5.8 Electrospray Mass Spectrometry

For oligosaccharide sequence analyses, negative- and positive-ion ESI-MS and ESI-CID-MS/MS were carried out on a Waters Q-TOF mass spectrometer Synapt G2-S (Manchester, UK). Nitrogen was used as desolvation and nebulizer gas at a flow rate of 500 L/h and 150 L/h, respectively. Source temperature was 80 °C, and the desolvation temperature 150 °C. A cone voltage of 80 V was used for negative-ion detection and the capillary voltage was maintained at 3 kV. MS/MS product-ion spectra were obtained from CID using argon as the collision gas at a

pressure of 0.17 MPa. The collision energy was adjusted between 17 and 28 V for optimal fragmentation. For quasi-MS<sup>3</sup> to obtain a further product-ion spectrum from a selected fragment ion as the precursor, the cone voltage was raised to 180 V to encourage the primary fragmentation. A scan rate of 1.0 s/scan and data acquisition of ~1 minute were used for both ES-MS and ES-CID-MS/MS experiments, and the acquired spectra were summed for presentation. For analysis, oligosaccharides were dissolved in ACN/H<sub>2</sub>O 1:1 (v/v), typically at a concentration of 15 pmol/μL, of which 2 μL was loop injected. Solvent (ACN/2 mM NH<sub>4</sub>HCO<sub>3</sub>, 1:1, v/v) was delivered by a Harvard syringe pump at a flow rate of 10 μL/min.

### 2.5.9 Carbohydrate microarrays construction and analysis

The microarray data and metadata provided here is described according to the MIRAGE (Minimum Information Required for A Glycomics Experiment) glycan microarray guidelines, as described by Liu *et al.* 2016<sup>146</sup>.

The information on the probe ID, sequence or monosaccharide composition analysis of the carbohydrate probes featuring in the microarrays is shown in Tables S2.1 and S2.5. For the preparation of the microarray, the probes were immobilized non-covalently onto 16-pad nitrocellulose-coated UniSart® 3D Microarray Slide from Sartorius (Goettingen, Germany), using a non-contact arrayer robot Nano-Plotter 2.1 (GeSiM, Radeberg, Germany), with a spot delivery volume of approximately 330 pL, following established protocols<sup>32</sup>. In brief, each probe was printed in duplicate at two levels 5 and 15 μM (2 and 5 fmol/spot, respectively). For the non-covalent immobilization, the NGLs were formulated as liposomes by adding lipid carriers, 1,2-dihexanoyl-*sn*-glycero-3-phosphocholine (DHPC) and cholesterol (both from Sigma-Aldrich, St. Louis, Missouri, US) at 100 pmol/μL<sup>85</sup>. To prepare the liposomes, the lipid carriers were mixed with the 5 pmol/mL NGL working solution and the samples are allowed to evaporate to complete dryness at 37 °C in an incubator for ~16 h. The dried mixture was then reconstituted in a solution of Cy3-water by vortex and brief centrifugation, followed by sonication for 15 minutes in a sonic water bath at 30 °C, and centrifugation at 8000×g. The Cy3 NHS ester fluorophore (GE Healthcare, Chicago, Illinois, US) Cy3 was included in the printing solution at 20 ng/mL (26 fmol/μL) as a marker for quality control of sample delivery while arraying and spot visualization, as well as for quantitation analysis.

Microarray binding analysis was performed using AlexaFluor-647-labeled Streptavidin for readout, essentially as described by Palma *et al.* 2015<sup>32</sup>. His-tagged CBMs were tested at 5 to 10 μg/mL, and human Malectin at 15 μg/mL precomplexed with mouse monoclonal anti-poly-histidine (Ab1) (Sigma-Aldrich, H1029) and biotinylated anti-mouse IgG (Ab2) (Sigma-Aldrich, B7264) antibodies, at a ratio of 1:3:3 (by weight). The protein-antibody complexes were prepared by preincubating Ab1 with Ab2 for 15 minutes at room temperature, followed by addition of CBMs and incubation for further 15 minutes, after which the final concentration of the proteins was achieved by dilution in the blocking solution made of 1% (w/v) Casein (Thermo

Scientific, 37583) 1:50 1% BSA (Sigma-Aldrich, A8577) in HBS (Sigma-Aldrich, H0887) (5 mM HEPES buffer pH 7.4, 150 mM NaCl) with 5 mM CaCl<sub>2</sub>. Monoclonal antibodies LM5, LM6, LM10, LM11, LM21, LM24, LM25, CCRC-M1 and CCRC-M70 were probed at 1:10 ratio, as described by Moller *et al*, 2008<sup>147</sup>, and antibodies 400-2, 400-3, and 400-4 at 10 µg/mL, diluted in the same blocker, followed by the biotinylated anti-mouse-IgG (Sigma-Aldrich, B7264), anti-rat-IgG (Sigma-Aldrich, B7139) or anti-rat-IgM (Rockland, Gilbertsville, Pennsylvania, US, 612-4607) as appropriate, at 10 µg/mL in the same blocker. Biotinylated lectin AAL was analyzed using a single step overlay at a final concentration of 2 µg/mL in blocker 3% BSA in HBS with 5 mM CaCl<sub>2</sub>. Biotinylated anti-rat and anti-mouse IgG and IgM antibodies were analysed in separate as a negative control. Slides were scanned using GenePix® 4300A microarray scanner (Molecular Devices, San Jose, California, US), at 532 nm for Cy3 spot visualisation, prior to binding assays, and at 647 nm for detection of the binding. Imaging analysis and quantitation was carried out using GenePixPro7 Software (Molecular Devices).

### 2.5.10 Microarray data analysis and presentation

Microarray data analysis was performed using a dedicated software<sup>148</sup>, developed by Mark Stoll of the Glycosciences Laboratory (Imperial College London, UK), that comprises a suite of modules to store, retrieve and display carbohydrate microarray data. In brief, microarray results were entered into an in-house database that holds all of the microarray data and metadata on experimental conditions and information on probes and proteins. A software for retrieval and display, that has a comprehensive system of sorters, filters and arrangers, was then used allowing to customize the data presentation as charts, tables and 2D matrices (heatmaps). No particular normalization method or statistical analysis was used for data processing.

After the scrutiny of all the microarray results the following decisions were made as regards presentation of data: 1) to modify the original printed microarray set layout excluding repeated probes and sorting the probes according to the nature of the sample and predominant oligosaccharide sequence (resulting arrangement of probes is in Tables S2.1 and S2.5); 2) present microarray data in the form of a matrix represented as a heatmap of the relative binding intensities (Figure 2.1B), in order to highlight the different binding patterns obtained for the proteins and antibodies analysed; 3) in order to accurately depict the binding patterns for each protein and antibody, the printed high level of NGLs (5 fmol/spot) were selected to generate the matrices and graphics (Figures 2.1 to 2.4 and 2.8 and Figure S2.1).

## 2.6 Work contributions

The preparation of the NGL probes and construction of the oligosaccharide microarrays resulted from the long-standing collaborative work of the Supervisor, Dr. Angelina Palma, with the group of Prof. Ten Feizi (Glycosciences Laboratory, Imperial College London), and the colleagues Dr. Hongtao Zhang, Dr. Yibing Zhang, Dr. Lisete M. Silva, Dr. Yan Liu and Dr. Wengang Chai are

acknowledged for their contribution to this work. In particular, the hemicellulose microarrays were planned and performed together with Dr. Wengang Chai, with a major contribution of Dr. Yibing Zhang for oligosaccharide fractionation and analysis, NGL probe preparation, MALDI-MS and ESI-MS analysis. The experimental planning and work reported here related to the carbohydrate microarray validation, binding and data analysis and interpretation, were performed by the author of the Thesis. Mass spectrometry analysis, oligosaccharide purification and preparation of the xyloglucan DAN-DHPA-NGL and xyloglucan AO-NGL probes included in the xyloglucan carbohydrate microarrays, as well as the microarray construction, validation, binding and data analysis, were performed by the author of the thesis at the Glycosciences Laboratory, Imperial College London, under the supervision of Prof. Ten Feizi and Dr. Wengang Chai. Dr. Lisete M. Silva is acknowledged for the planning and guidance in microarray printing. Protein expression and purification of CtCBM11, was performed by the author of the Thesis (as detailed in Chapter 4). CtCBM22-2<sub>Cthe\_0912</sub>, CtCBM25<sub>Cthe\_0956</sub> and CtCBM35<sub>Cthe\_2811</sub>, were prepared using a high-throughput platform at NZYTEch (Lisbon, Portugal) (as detailed in Chapter 3), as result from a long-standing collaborative work of the Supervisor and Co-supervisor, Dr. Ana Luísa Carvalho, with Prof. Carlos Fontes (CIISA-FMV, ULisboa). Human malectin was kindly provided by Dr. Benedita Pinheiro (UCIBIO, NOVA).

# CHAPTER 3

---

**CELLULOLYTIC BACTERIA EXPRESS CBMOMES THAT  
DICTATE THEIR ECOLOGICAL NICHE  
POLYSACCHARIDE UTILIZATION**





## 3 Cellulolytic bacteria express CBMomes that dictate their ecological niche polysaccharide utilization

### 3.1 Introduction

Anaerobic microbial organisms are highly efficient for plant cell wall polysaccharide biodegradation and have evolved a multi-enzyme complex system, the Cellulosome, where catalytic carbohydrate-active enzymes (CAZymes) have non-catalytic Carbohydrate Binding Modules (CBMs) appended. The latter exhibit different functional roles in highly potentiating the enzymes' catalytic efficiency (reviewed in Chapter 1, section 1.2.1.2). Deciphering at molecular level the mechanisms underlying plant cell wall carbohydrate recognition and deconstruction by different cellulolytic bacteria is crucial to elucidate these complex biological systems, as well as to further promote novel potential applications.

In the past years, genomic sequencing has promoted an exponential increase in the identification of CAZymes and CBM sequences, leading to a substantial number of proteins for which the carbohydrate binding specificities and mechanisms of ligand recognition are awaiting elucidation<sup>26,27</sup>. The implementation of the CAZy database<sup>22</sup> provides a valuable tool to access updated lists of CBM-containing proteins, displaying characterised (with structural or biochemical data) and uncharacterised putative CBM domains grouped into sequence-related CBM families.

The genome sequencing of the cellulolytic bacterium *Ruminococcus flavefaciens* FD-1, a rumen bacterium, has revealed numerous modular glycoside hydrolases (GH) and several CBMs that were grouped into known CAZy families by sequence homology<sup>27</sup>, but for which assignment of carbohydrate-binding specificity is still required; also putative protein modules, some of which have been classified into previously un-identified CBMs families 75 to 80<sup>59</sup>. Additionally, *R. flavefaciens* FD-1 has been reported to have one of the largest collection of cellulosome-associated proteins among known fibre-degrading bacteria<sup>27</sup>. A large variety of GH families found within *R. flavefaciens* FD-1 genome, can also be found in *Clostridium thermocellum*<sup>27</sup>, a thermophilic anaerobic bacterium for which the first cellulosome was characterized<sup>12,54</sup>. Although more extensively studied, *C. thermocellum* also possess various CBMs that await elucidation.

With development of bioinformatics tools for annotating the available genomes, structural biology methods and accessible databases, predictions can be made about the probable, or possible, carbohydrate-binding activities of proteins on the basis of protein-sequence homologies and structures<sup>22,149</sup>. Nevertheless, direct binding experiments are required for validation and characterization of these predictions and to discover novel mechanisms of ligand binding. The detection and characterization of carbohydrate-protein interactions have been challenging due to the complexity of carbohydrate structures, availability of the sample (both proteins and

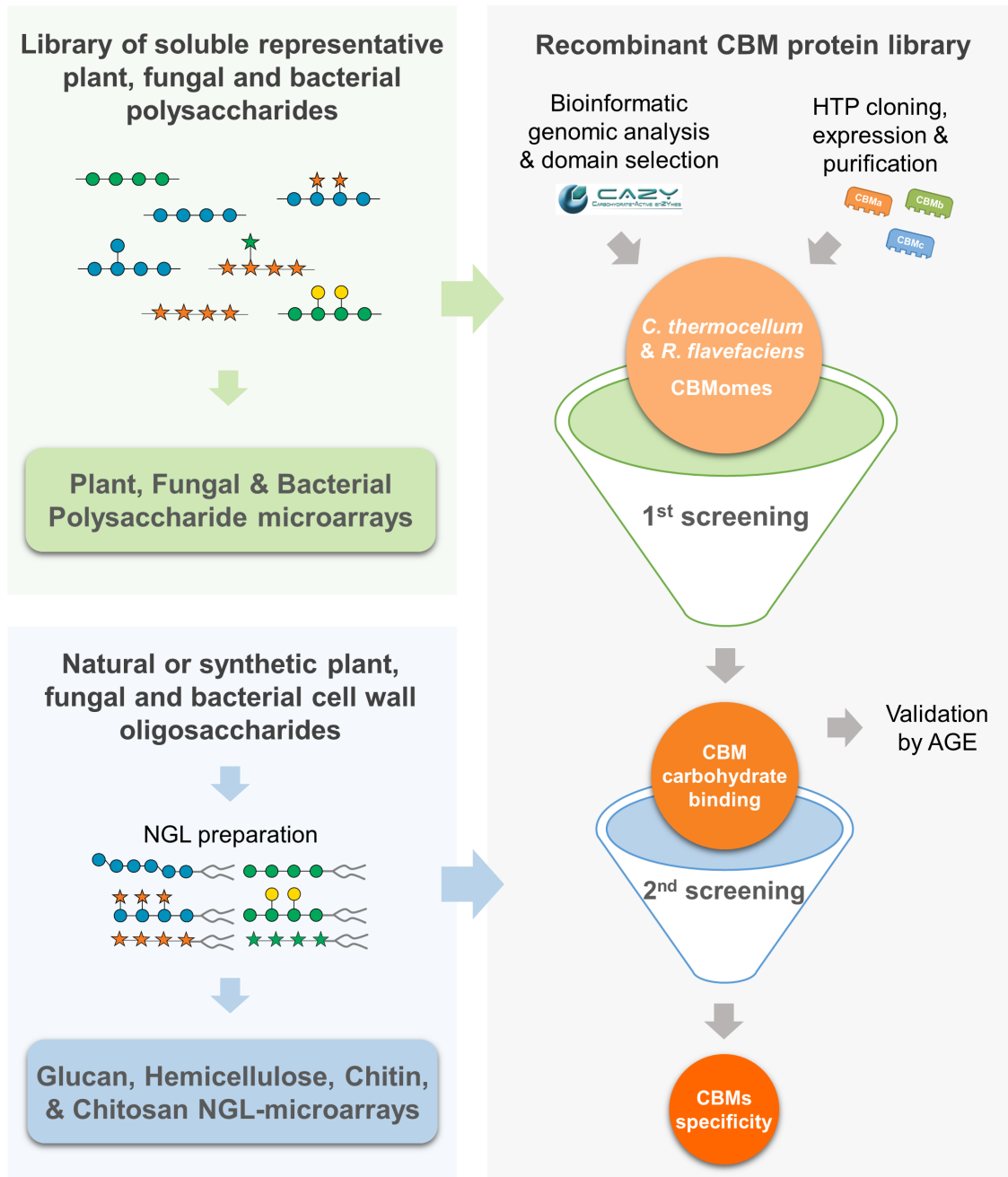
sequence-defined carbohydrates) and the need to develop miniaturized tools to study these. The development of the high-throughput carbohydrate microarray technology came to address some of these challenges and since the first proof-of-principle papers in 2002, has revolutionized the studies of carbohydrate-recognising systems<sup>61,74,78</sup>.

In the present work, to broaden the understanding on the functionality of non-catalytic CBMs of cellulolytic organisms our approach was two-fold: 1) to implement a multi-step strategy combining high-throughput methods for cloning, expression and purification of proteins with carbohydrate microarrays, which was applied to interrogate and assign carbohydrate-binding specificities for a representative set of CBMs assigned into different CAZy families (here referred to as CBMomes); and 2) to compare the carbohydrate-binding specificities of the CBMomes from two cellulolytic bacteria, *C. thermocellum* and *R. flavefaciens* FD-1, which reside in different, highly dynamic and populated ecological niches, the soil and the rumen of mammals, respectively. To this end, polysaccharide microarrays and neoglycolipid (NGL)-based oligosaccharide microarrays, comprising mainly carbohydrate sequences present on plant cell walls, but also on fungal and bacterial cell walls, were used to screen the carbohydrate-binding and ligand-specificity of up to 105 *R. flavefaciens* FD-1 and *C. thermocellum* CBMs. The groups of polysaccharides that are differentially recognised by the CBMs were revealed and novel CBM-ligand specificities were identified.

## 3.2 Results and Discussion

### 3.2.1 Multi-step strategy to assign carbohydrate-binding specificities of CBMs in a high-throughput manner

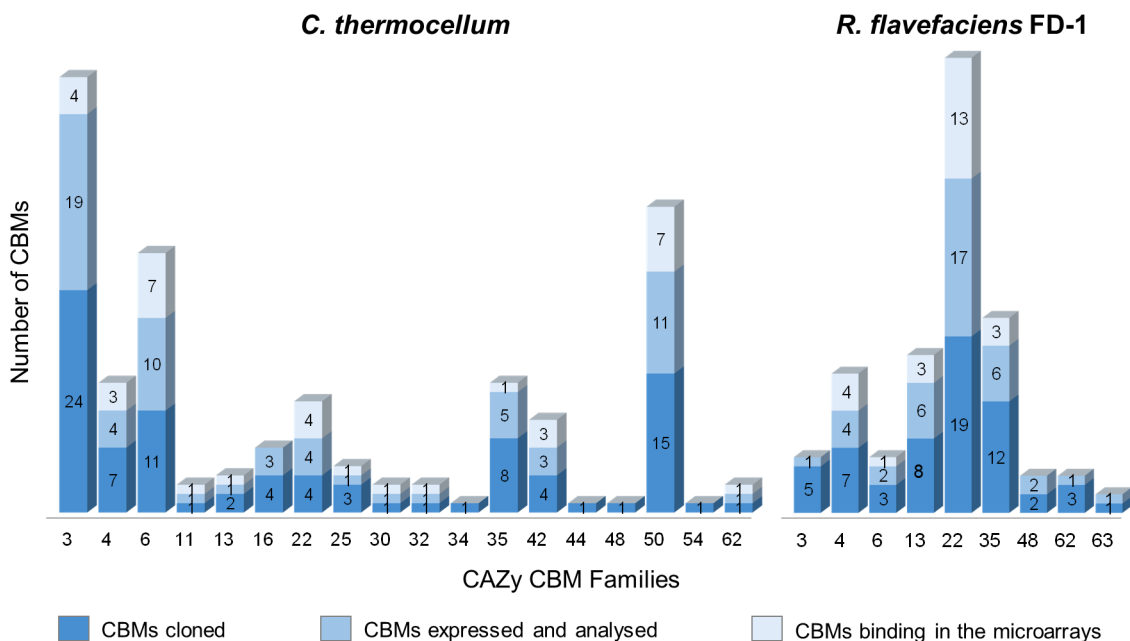
Aiming at deciphering the repertoire of carbohydrate-binding specificities of CBMs from *C. thermocellum* and *R. flavefaciens* FD-1 (henceforward referred to only as *R. flavefaciens*), a multi-step strategy integrating high-throughput platforms was followed, as illustrated in Figure 3.1. All *C. thermocellum* and *R. flavefaciens* CBM sequences assigned into different families in the CAZy database (at the start of the Thesis work in 2015) – here designated the ‘CBMomes’ – were amplified from the genome of the bacteria and prepared as recombinant proteins using high-throughput methods for cloning, expression and purification. These recombinant ‘CBMomes’ were screened for carbohydrate binding using a microarray comprising soluble polysaccharides with sequences found on plant, fungal and bacterial cell walls (1<sup>st</sup> screening). The 1<sup>st</sup> screening microarray data obtained for each CBM was validated using semi-quantitative affinity gel electrophoresis (AGE). The ‘CBMomes’ were screened only for their recognition profile to soluble polysaccharides. Carbohydrate-binding to insoluble substrates was not assessed in the present study. All CBMs that gave binding patterns in the polysaccharide microarrays, were subjected to a 2<sup>nd</sup> screening using an NGL-microarray comprised of oligosaccharide sequences found on plant cell wall, fungal and bacterial polysaccharides, to assign carbohydrate binding-specificity.



**Figure 3.1. Schematic representation of the multi-step strategy followed in this work.** Bioinformatics analysis was carried out on annotated sequences in the bacterial genomes for selection of DNA sequences coding for putative CBMs assigned to different families in the CAZY database ('CBMomes'). These were cloned, expressed and purified using a high-throughput platform. Polysaccharide microarrays comprised of plant-, fungal- and bacterial-related sequences were developed to screen *C. thermocellum* and *R. flavefaciens* FD-1 CBMomes for carbohydrate binding (1<sup>st</sup> screening). The microarray data was cross-validated using affinity gel electrophoresis (AGE). All CBMs that gave binding patterns in the 1<sup>st</sup> screening, were screened for ligand-specificity assignment using NGL-microarray platform composed of oligosaccharides sequences representative of plant and fungal cell walls (2<sup>nd</sup> screening). The monosaccharide symbolic representation used was according to the updated SNFG<sup>1</sup>.

### 3.2.2 Bacterial CBMomes from different ecological niches

A total of 150 genes were amplified and cloned in an *Escherichia coli* expression vector, coding for 90 CBMs from *C. thermocellum* and 60 CBMs from *R. flavefaciens*, assigned to different CAZy families (Figure 3.2). The majority of the CBMs were readily expressed and purified through homogeneity by affinity chromatography (Figure S3.1). Modularity and sequence information of the CBMs for which binding patterns were obtained and for the ones that were poorly expressed or did not bound in the microarrays are presented in Tables S3.1 and S3.2, respectively. The modular architectures of both bacteria CBM's repertoire show that while most of the CBMs are found associated with polysaccharide degrading CAZymes and to Dockerin (DOC) modules, indicating a probable association with the cellulosome, many are found as single domains or associated to other proteins, hinting other possible functions. Overall, *C. thermocellum* contains a broader diversity of CAZy CBMs families, showing a predominance of those described to target crystalline cellulose (family 3),  $\beta$ -xylans,  $\beta$ 1,4-glucans,  $\beta$ 1,3-glucans, mixed-linked  $\beta$ 1,3-1,4-glucans,  $\beta$ 1,3-glucans (family 6) and chitin or peptidoglycan (family 50). For *R. flavefaciens* a more restricted number of CBMs assigned into existing CAZy families are identified, with prevalence for hemicellulose-recognising families and in higher number from families 22 and 35, which are known to be specific for  $\beta$ -xylans and pectic polysaccharides or  $\beta$ 1,4-mannans, respectively (Figure 3.2).



**Figure 3.2. Overview of the selected CBMomes of the two bacteria.** The chart shows the numbers of *C. thermocellum* and *R. flavefaciens* FD-1 CBMs from each CAZy family that were cloned, analysed and for which binding in the microarrays was obtained. A total of 150 CBMs were cloned (bottom bar), 90 from *C. thermocellum* and 60 from *R. flavefaciens* FD-1, 105 CBMs from both bacteria were successfully expressed as recombinant domains in *E. coli* and analysed in the carbohydrate microarrays (middle bar). Binding patterns were obtained for 35 *C. thermocellum* CBMs and 24 *R. flavefaciens* FD-1 CBMs (top bar). Modularity and sequence information of the CBMs are presented in Tables S3.1 and S3.2.

### 3.2.3 Carbohydrate microarray platforms for ligand discovery

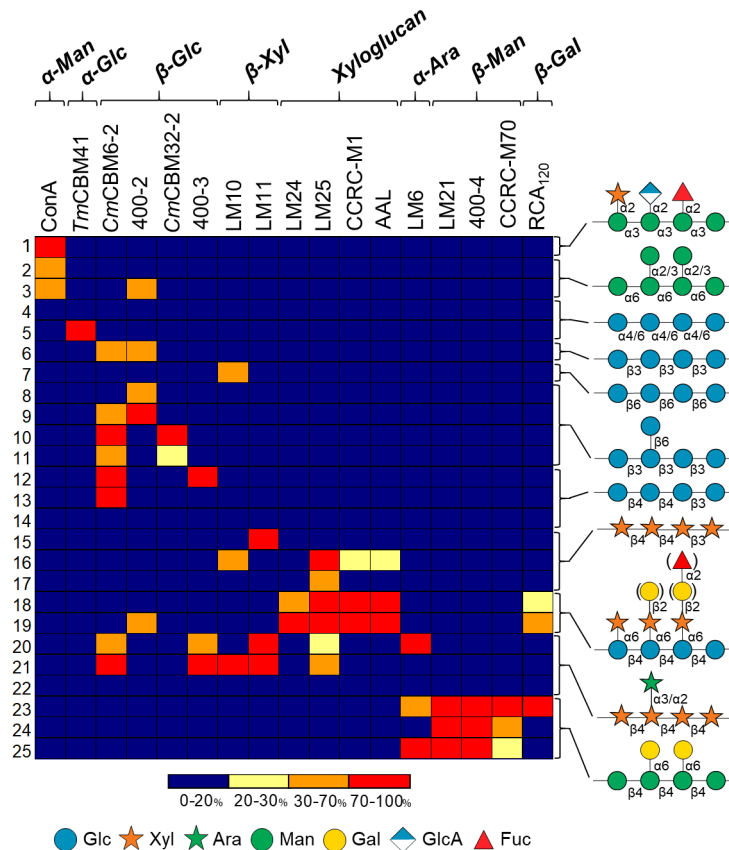
The polysaccharide microarray platform used in the 1<sup>st</sup> screening to investigate carbohydrate-binding activities for the expressed CBMs, designated *Plant, Fungal and Bacterial polysaccharide (PS) set 1*, comprised 25 structurally diverse carbohydrate probes (Table S3.3). These are representative of major sequences found in plant cell wall  $\beta$ -glucans and hemicelluloses, such as xylans, xyloglucans, arabinoxylans and galactomannans, in fungal  $\alpha$ -glucans,  $\beta$ -glucans and  $\alpha$ -mannans, and in bacterial  $\alpha$ -glucans. The validation of the polysaccharide microarrays was carried out using 11 carbohydrate-directed monoclonal antibodies, 3 CBMs of different microorganisms and 3 plant lectins, for which the carbohydrate binding specificities have been characterized (Chapter 2, Table S2.3). Overall, the binding patterns observed to the different polysaccharide samples were in agreement with the reported carbohydrate binding for the antibodies and proteins analysed (Figure 3.3 and Table S3.4), as well as the specificity observed in the NGL-microarrays presented in Chapter 2, section 2.2.2.

*$\alpha$ -mannans* – The  $\alpha$ -linked mannose-specific lectin Concanavalin A (ConA)<sup>150</sup> exhibited restricted binding to the polysaccharides with an  $\alpha$ 1,3-Man backbone or with an  $\alpha$ 1,6-linked mannose backbone with  $\alpha$ 1,2- or 1,3-mannose branches (probes 1-3).

*$\alpha$ -glucans and  $\beta$ -glucans* – The *Thermotoga maritima*  $\alpha$ 1,4-linked glucose specific CBM from family 41 (*TmCBM41*)<sup>32</sup>, bound specifically to the mixed-linked  $\alpha$ 1,6-1,4-linked glucan (probe 5).

The *Cellvibrio mixtus* family 6 and 32 CBMs (*CmCBM6-2* and *CmCBM32-2*, respectively) were used in parallel with the  $\beta$ 1,3-glucan-specific (400-2) and  $\beta$ 1,3- $\beta$ 1,4-linked glucan-specific (400-3) monoclonal antibodies to validate the different  $\beta$ -glucan polysaccharides. *CmCBM6-2* showed strong binding to most of the  $\beta$ -glucans<sup>32</sup> (probes 6 and 9 to 13), whereas 400-2 and 400-3 antibodies showed their preferential binding to  $\beta$ 1,3-glucans<sup>151</sup> (probes 6, 8 and 9) or mixed-linked  $\beta$ 1,3-1,4-glucans<sup>134</sup> (probe 12 and 14), respectively. These results were in agreement with binding patterns observed to related oligosaccharides in the NGL-microarrays (Chapter 2, Figure 2.2). The strong binding observed with *CmCBM6-2* and 400-3 antibody to the arabinoxylan fractions (probes 20-21) showed the presence of a  $\beta$ 1,4- or mixed-linked  $\beta$ 1,3-1,4-glucan component in these heterogeneous fractions. *CmCBM32-2* showed highly restricted binding to the  $\beta$ 1,3-glucans with higher  $\beta$ 1,6-glucose branching (probes 10-11), in accordance with the predicted involvement of the  $\beta$ 1,6-branching for preferential recognition by this CBM<sup>32</sup>.

*Linear and branched  $\beta$ -xylans* - The  $\beta$ 1,4-xylan-specific monoclonal antibodies LM10 and LM11<sup>135</sup>, showed binding to polysaccharides of linear  $\beta$ 1,3-1,4- and  $\beta$ 1,4-linked xylose isolated from *Palmaria palmata* and plum (probes 15 and 16, respectively) and  $\beta$ 1,4-linked xylose with  $\alpha$ 1,2- or  $\alpha$ 1,3-arabinose substitutions isolated from brewer's spent grain (probes 20 and 21), in line with the specificities observed in NGL-microarrays (Chapter 2, Figure 2.3A). Remarkably, the  $\alpha$ 1,5-arabinose-specific monoclonal antibody LM6<sup>137</sup> showed binding, albeit weak, to the arabinoxylan fraction with degree of polymerization 41 (DP41) (probe 20), which might be due to



**Figure 3.3. Validation of the polysaccharide microarrays with sequence-specific carbohydrate-binding proteins.** The microarrays included 25 polysaccharides, for which sequences of major structural domains are depicted at the left. Information available for these polysaccharide samples is in Table S3.3. The monoclonal antibodies, plant lectins and CBMs of characterized carbohydrate-binding specificity used for the microarray validation are exhibited at the top. The major representative structural domain for each polysaccharide is depicted at the left using a tetrasaccharide backbone sequence as a reference. The binding results are shown as a heatmap of the relative binding intensities calculated as the percentage of the fluorescence signal intensity at 150 pg/spot given by the probe most strongly bound by each protein (normalized as 100%). Numerical scores are given in Table S3.4.

the presence of  $\alpha$ 1,2- and  $\alpha$ 1,3-arabinose, as indicated by the binding of this antibody to the branched arabinan probes in NGL-microarrays (Chapter 2, Figure 2.3).

**Xyloglucans** - The monoclonal antibodies LM24, LM25 directed to xyloglucans<sup>92</sup>, the CCRC-M1 antibody that requires the  $\alpha$ -linked fucose, and the  $\alpha$ -fucose-specific *Aleuria aurantia* lectin (AAL)<sup>143</sup>, all bound to the plum xyloglucan polysaccharides (probes 18-19). Considering the specificities of these antibodies, observed in the NGL-microarrays (Chapter 2, Figure S2.1), the results indicated the presence of  $\alpha$ 1,6-linked xylose,  $\beta$ 1,2-linked galactose and  $\alpha$ 1,2-linked fucose substitutions of the xyloglucans in these fractions. In addition, the binding observed with the galactose-specific *Ricinus communis* agglutinin I (RCA<sub>120</sub>)<sup>152</sup> also indicated the presence of  $\beta$ 1,2-linked galactose.

**Galactomannans** - In accord with the binding results from the NGL-microarrays (Chapter 2, Figure 2.4), the  $\beta$ 1,4-linked mannose-specific monoclonal antibodies 400-4 and LM21<sup>138,139</sup>, the guar galactomannan-directed antibody CCRC-M70<sup>140</sup>, and RCA<sub>120</sub><sup>152</sup> bound differentially to the

guar and carob galactomannans (probes 23-25), with the last two showing preferential binding to the highly  $\alpha$ 1,6-galactose-substituted guar galactomannan (probe 23). Of note was that LM6 also bound to the guar galactomannans, indicating a possible  $\alpha$ -arabinose component in these polysaccharide fractions.

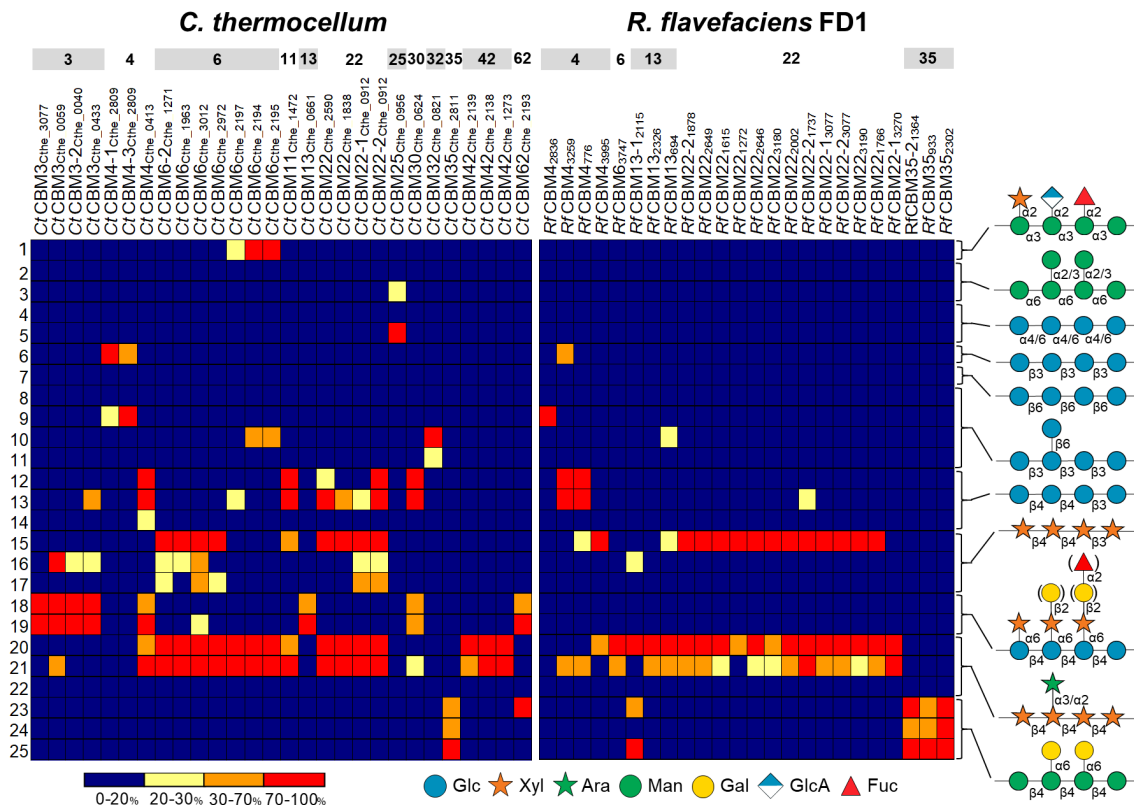
Overall, the results showed that the polysaccharide microarray presented the different structural domains as highlighted in Figure 3.3 and was functional for carbohydrate-binding screening analysis of *C. thermocellum* and *R. flavefaciens* CBMomes, which are described in section 3.2.4 below.

For the 2<sup>nd</sup> screening to assign carbohydrate-binding specificities for the CBMs, the microarray platform used comprised glucan, hemicellulose, chitin and chitosan oligosaccharides prepared as NGL probes (Chapter 2, Table S2.1). The glucan and hemicellulose oligosaccharide microarrays were described in Chapter 2 (probes 1-204). The additional probes investigated in this Chapter included sequence-defined  $\beta$ 1,4-linked-*N*-acetylglucosamine (GlcNAc, chitin) oligosaccharides ranging from DP-2 to DP-8 (probes 205-211) and  $\beta$ 1,4-linked glucosamine (GlcN, chitosan) oligosaccharides from DP-4 to DP-6 (probes 213-215). Also included were 4 miscellaneous probes (disaccharides and trisaccharides) (probes 215-218).

### 3.2.4 Screening *C. thermocellum* and *R. flavefaciens* FD-1 CBMomes for carbohydrate-binding specificity

Of the 150 CBMs clones, a total of 105 proteins from both bacteria were successfully expressed and screened for carbohydrate binding using the polysaccharide microarrays (1<sup>st</sup> screening). The heatmap in Figure 3.4 highlights the different patterns of polysaccharide recognition observed for the CBMs of the two bacteria (binding scores in Tables S3.5 and S3.6). The microarray results were representative of the analysis of two protein batches and two microarray platforms of similar carbohydrate composition, prepared independently. The carbohydrate recognition of each CBM was validated using AGE with the respective polysaccharides (Figures 3.5 and 3.6). The results are from at least two experiments carried out using two different batches of CBMs. For the majority of the CBMs, the interaction data obtained by AGE was in agreement with the screening results obtained in the polysaccharide microarrays. However, for the following CBM interactions this could not be observed: CtCBM3<sub>Cthe\_3077</sub> with xyloglucans; CtCBM25<sub>Cthe\_0956</sub> with pullulan; RfCBM6<sub>3747</sub>, RfCBM13<sub>2326</sub>, RfCBM13<sub>694</sub>, RfCBM22<sub>2649</sub> and RfCBM22-1<sub>3270</sub> with arabinoxylans. This could relate with a weak affinity of the CBMs for the polysaccharides tested, as microarrays are a highly sensitive technique, and weak binding could be detected and not in the AGE analysis. Additionally, the electrophoretic mobility of the CBMs may also influence the migration in the AGE, hindering discrete protein bands to be observed when the binding to the polysaccharide is weak.

Upon validation of the polysaccharide binding patterns, the CBMs were screened for their oligosaccharide ligand-specificity using the oligosaccharide NGL-microarray platform (2<sup>nd</sup>

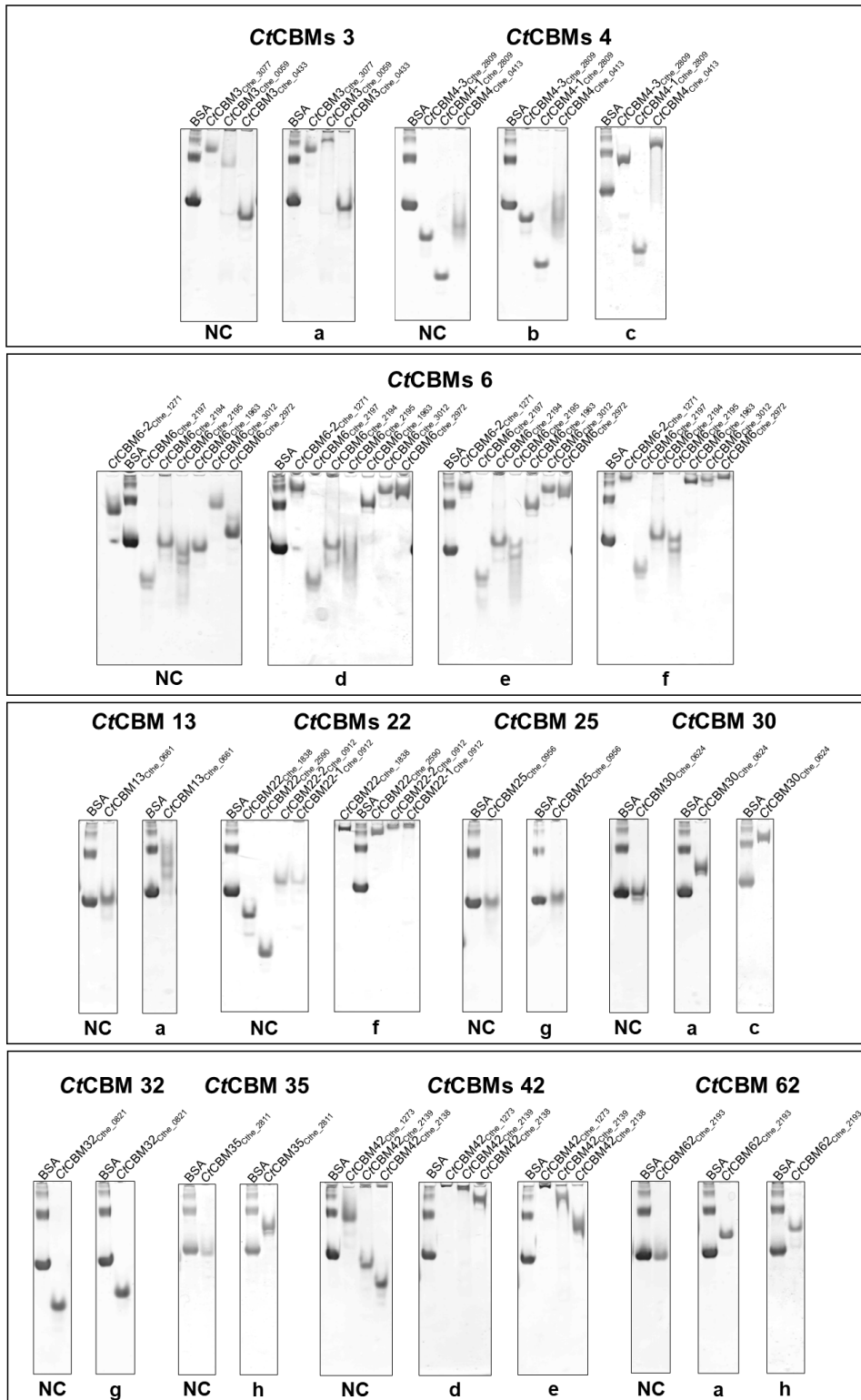


**Figure 3.4. Analysis of *C. thermocellum* and *R. flavefaciens* FD-1 CBMs using polysaccharide microarrays – 1<sup>st</sup> screening for carbohydrate-binding activities.** CBMs for which binding was obtained are depicted at the top for each bacterium and organised by CAZy family. The heatmap representation highlights the different polysaccharide binding patterns revealed by the microarray analysis. The relative binding intensities were calculated as the percentage of the fluorescence signal intensity at 150 pg/spot, with exception of *CtCBM4-3<sub>Cthe\_2809</sub>*, *CtCBM25<sub>Cthe\_0956</sub>* and *CtCBM62<sub>Cthe\_2193</sub>* at 30 pg/spot, given by the probe most strongly bound by each protein (normalized as 100%). Numerical scores are given in Tables S3.5 and S3.6.

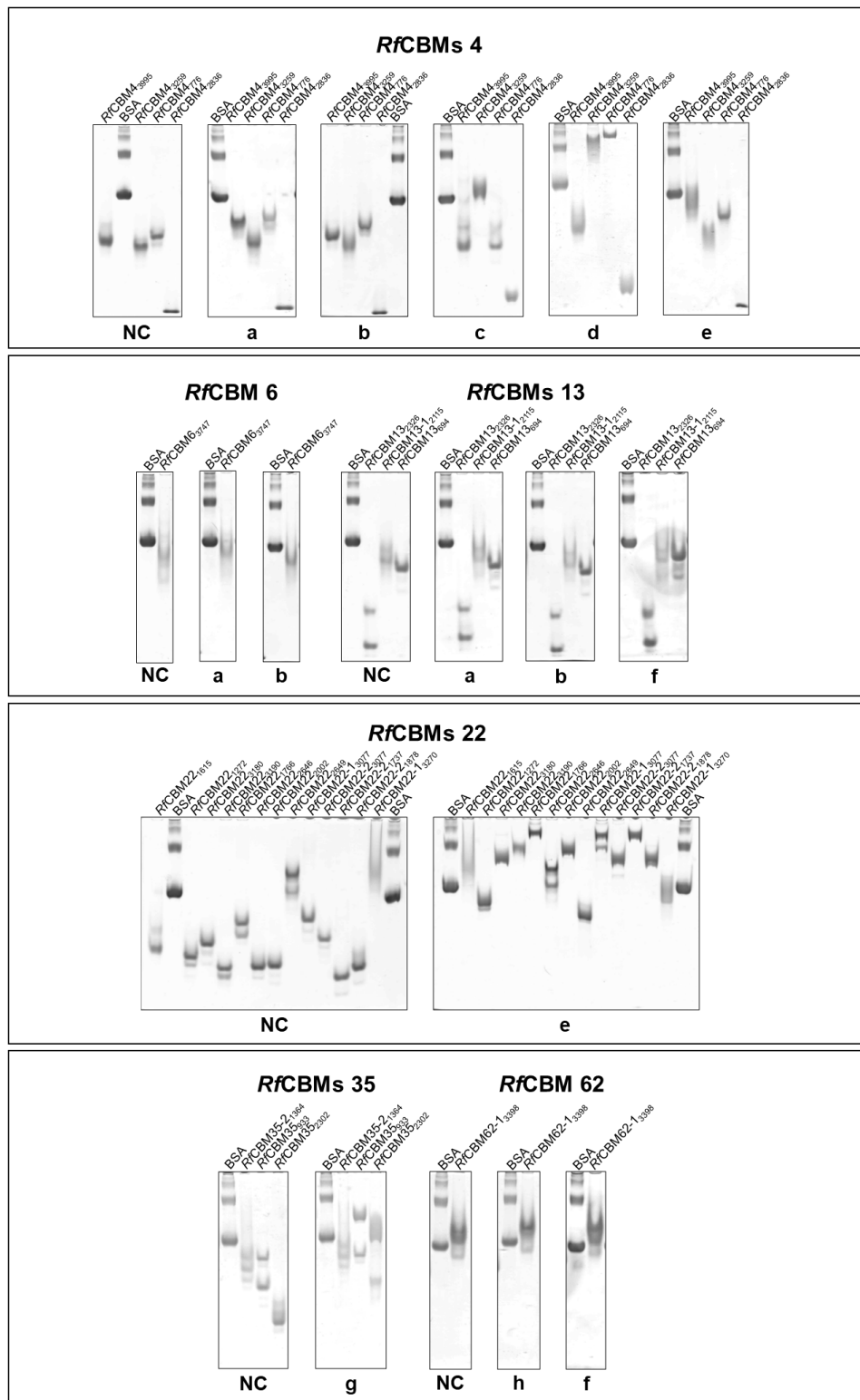
screening). The heatmap in Figure 3.7 highlights the different patterns of oligosaccharide recognition observed for the CBMs of the two bacteria (binding scores in Tables S3.7 and S3.8). From both microarray screening analysis, carbohydrate binding was identified for 59 CBMs, 35 from *C. thermocellum* and 24 from *R. flavefaciens* (Figures 3.2).

Overall, the polysaccharide and oligosaccharide binding patterns observed were in accordance and are supported by the specificities reported in the literature for the respective CBMs and CBM families. *C. thermocellum* CBMs showed broader binding patterns, more specific for  $\beta$ -glucans, while also exhibiting binding to  $\beta$ -xylans,  $\alpha$ -arabinans,  $\beta$ -mannans and  $\beta$ 1,4-linked GlcNAc. *R. flavefaciens* on its turn, showed more restricted carbohydrate binding patterns, with a greater number of CBMs targeting the hemicelluloses  $\beta$ -xylans, and fewer CBMs recognising  $\beta$ -glucans,  $\beta$ -mannans and  $\alpha$ -arabinans. A summary of the CBM's carbohydrate recognition and specificity obtained from both microarray screenings, its validation through electrophoretic methods and cross-referenced with available literature, is presented in Tables 3.1 and 3.2. The main binding patterns validated for each CBM family is interpreted in the following sections.

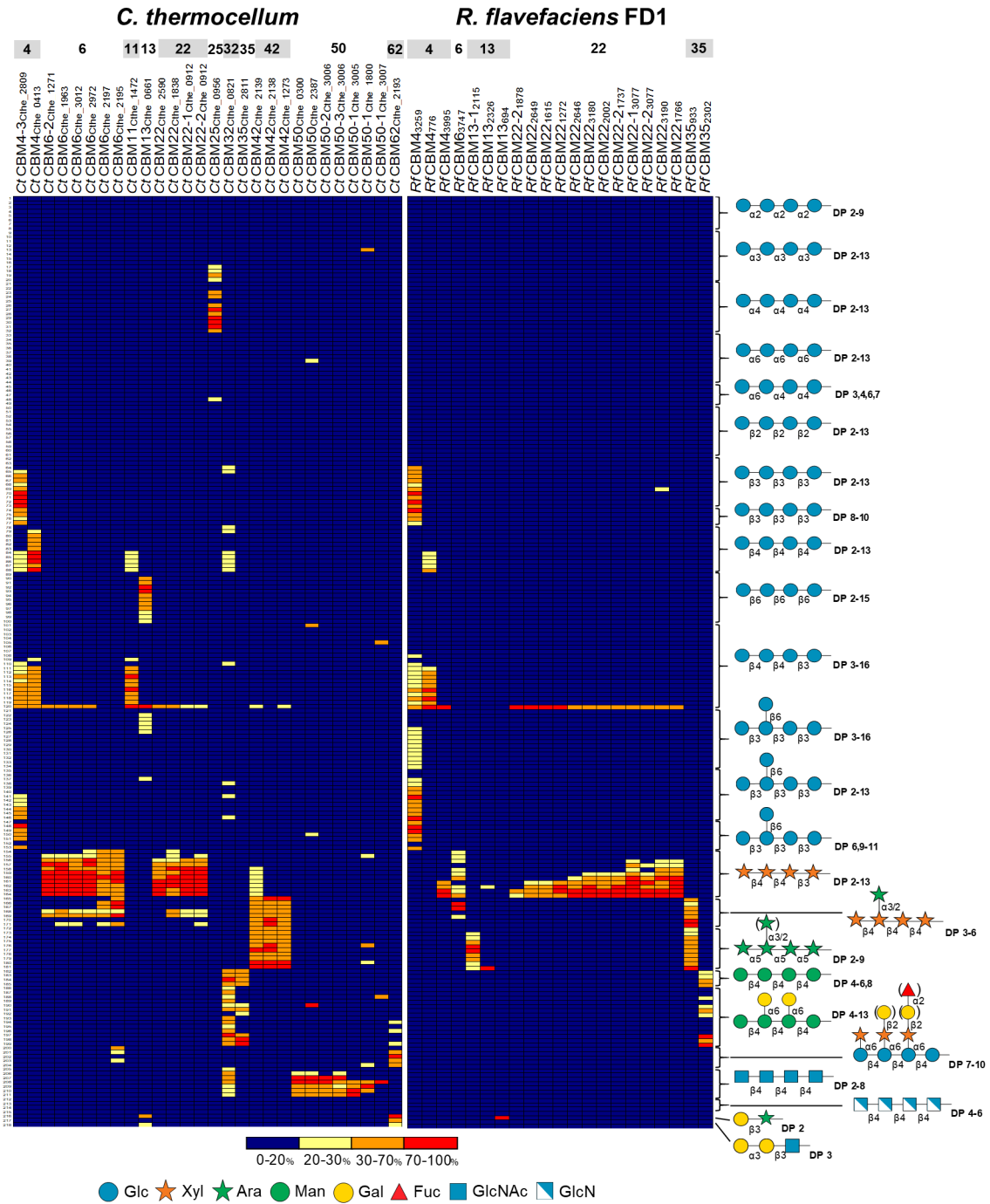




**Figure 3.5. Validation of the *C. thermocellum* CBMs microarray binding patterns using affinity gel electrophoresis.** The AGE analysis is presented in a rational order grouping the CBMs by CAZy family. CBMs were subjected to non-denaturing electrophoresis in a gel containing 0.1% (w/v) the soluble polysaccharide: **a)** Xyloglucan (boiled plum), **b)** PGG- $\beta$ -glucan, **c)** Barley  $\beta$ -glucan, **d)** Arabinoxylan (DP41), **e)** Arabinoxylan (DP24), **f)** Xylan (*Palmaria p.*), **g)** Pullulan, **h)** Galactomannan (guar); **NC)** control non-denaturing electrophoresis gel without the ligand was ran simultaneously. Bovine serum albumin (BSA lanes) was used as a marker.



**Figure 3.6. Validation of the *R. flavefaciens* CBMs microarray binding patterns using affinity gel electrophoresis.** The AGE analysis is presented in a rational order grouping the CBMs by CAZy family. CBMs were subjected to non-denaturing electrophoresis in a gel containing 0.1% (w/v) the soluble polysaccharide: **a)** Arabinoxylan (DP41), **b)** Arabinoxylan (DP24), **c)** PGG- $\beta$ -glucan, **d)** Barley  $\beta$ -glucan, **e)** Xylan (*Palmaria p.*), **f)** Pectic galactan (lupin), **g)** Galactomannan (guar), **h)** Pectin (apple); **NC)** control non-denaturing electrophoresis gel without the ligand was ran simultaneously. Bovine serum albumin (BSA lanes) was used as a marker.



**Figure 3.7. Analysis of *C. thermocellum* and *R. flavefaciens* FD-1 CBMs using oligosaccharide microarrays – 2<sup>nd</sup> screening for assigning carbohydrate-binding specificities.** The microarray comprises 218 NGL probes with a wide degree of polymerization (DP) range of linear and branched oligosaccharide sequences of  $\alpha$ - and  $\beta$  glucans<sup>32</sup>,  $\beta$ -xylans,  $\alpha$ -arabinans,  $\beta$ -mannans, xyloglucans, chitin and chitosan. The major representative structural domain for each probe series is depicted at the left using a tetrasaccharide backbone sequence as a reference. Carbohydrate sequence information on these probes is in Chapter 2, Table S2.1. CBMs for which binding was obtained are depicted at the top for each bacterium and organised by CAZy family. The heatmap representation highlights the different oligosaccharide binding patterns revealed by the microarray analysis. The relative binding intensities were calculated as the percentage of the fluorescence signal intensity at 5 fmol given by the probe most strongly bound by each protein (normalized as 100%). Numerical scores are given in Tables S3.7 and S3.8.

**Table 3.1. Summary of the *C. thermocellum* CBMs ligand recognition and specificity obtained in the carbohydrate microarray screenings and affinity gel electrophoresis (AGE), cross-referencing with the available literature.**

Family	CBM	1 <sup>st</sup> screening		2 <sup>nd</sup> screening		Reported specificity <sup>3</sup> (PDB/Reference)
		Polysaccharide microarrays <sup>1</sup>	AGE <sup>2</sup>	NGL-microarrays		
3	<i>Ct</i> CBM3 <sub>Cthe_3077</sub>	Xyloglucans	-			Cellulose (Tormo <i>et al.</i> 1996 <sup>153</sup> )
	<i>Ct</i> CBM3 <sub>Cthe_0059</sub>	Xyloglucans**	++			Cellulose (Yaniv <i>et al.</i> 2014 <sup>41</sup> )
		Xylans**	NT		-	
		Arabinoxylans	NT			
	<i>Ct</i> CBM3-2 <sub>Cthe_0040</sub>	Xyloglucans	NT			Cellulose (Petkun <i>et al.</i> 2015 <sup>154</sup> )
<i>Ct</i> CBM3 <sub>Cthe_0433</sub>	Xyloglucans** Mixed-linked $\beta$ -glucans	-/+ NT			NA	
4	<i>Ct</i> CBM4-1 <sub>Cthe_2809</sub>	Curdlan**	NT			NA
		PGG- $\beta$ -glucan*	+		-	
	<i>Ct</i> CBM4-3 <sub>Cthe_2809</sub>	PGG- $\beta$ -glucan**	+			NA
		Curdlan Mixed-linked $\beta$ -glucans*	NT ++		$\beta$ 1,3-Glc	
<i>Ct</i> CBM4 <sub>Cthe_0413</sub>	Mixed-linked $\beta$ -glucans** Xyloglucans Arabinoxylans	+++ NT NT		$\beta$ 1,4-Glc	Cellobiose (3K4Z) (Alahuhta <i>et al.</i> 2010 <sup>155</sup> )	
6	<i>Ct</i> CBM6-2 <sub>Cthe_1271</sub>		+++ ++			NA
	<i>Ct</i> CBM6 <sub>Cthe_1963</sub>	Xylans Arabinoxylans	+++ ++			NA
	<i>Ct</i> CBM6 <sub>Cthe_3012</sub>		+++ ++	$\beta$ 1,4-Xyl/ $\beta$ 1,3-1,4-Xyl	NA	
	<i>Ct</i> CBM6 <sub>Cthe_2972</sub>		+++ ++			Xylopentaose (1UXX) (Pires <i>et al.</i> 2004 <sup>38</sup> )
	<i>Ct</i> CBM6 <sub>Cthe_2197</sub>	Arabinoxylans** Glucurono-xyloMannan*	-/+ -			NA
	<i>Ct</i> CBM6 <sub>Cthe_2194</sub>	Arabinoxylans** Glucurono- XyloMannan**	-/+ -		-	NA
	<i>Ct</i> CBM6 <sub>Cthe_2195</sub>	Arabinoxylans	-/+		$\beta$ 1,4-Xyl/ $\beta$ 1,3-1,4-Xyl	NA
Glucurono-XyloMannan		-/+				
11	<i>Ct</i> CBM11 <sub>Cthe_1472</sub>	Mixed-linked $\beta$ -glucans	NT <sup>#</sup>		$\beta$ 1,3-1,4-Glc	$\beta$ 1,3-1,4-gluco- oligosaccharides (6R31, 6R3M) (Palma <i>et al.</i> 2015 <sup>32</sup> , Ribeiro <i>et al.</i> , 2019 <sup>34</sup> )
13	<i>Ct</i> CBM13 <sub>Cthe_0661</sub>	Xyloglucan**	+		Gal- $\beta$ 1,3-Ara; $\beta$ 1,6-Glc <sup>#</sup>	Galactose (3VSZ, 3VT1) (Jiang <i>et al.</i> 2012 <sup>156</sup> )
22	<i>Ct</i> CBM22 <sub>Cthe_2590</sub>		+++ NT			NA
	<i>Ct</i> CBM22 <sub>Cthe_1838</sub>	Xylans Arabinoxylans	+++ NT		$\beta$ 1,4-Xyl/ $\beta$ 1,3-1,4-Xyl	NA
			+++ NT			NA
	<i>Ct</i> CBM22-1 <sub>Cthe_0912</sub>		+++ NT			Xylan (Charnock <i>et al.</i> 2000 <sup>136</sup> )
25	<i>Ct</i> CBM25 <sub>Cthe_0956</sub>	Pullulan	-		$\alpha$ 1,4-Glc	NA
30	<i>Ct</i> CBM30 <sub>Cthe_0624</sub>	Mixed-linked $\beta$ -glucans** Xyloglucans	+++ ++		-	Mixed-linked $\beta$ -glucans, Xyloglucan (Najmudin <i>et al.</i> 2006 <sup>157</sup> )

Table 3.1. (cont.)

Family	CBM	1 <sup>st</sup> screening		2 <sup>nd</sup> screening		Reported specificity <sup>3</sup> (PDB/Reference)
		Polysaccharide microarrays <sup>1</sup>	AGE <sup>2</sup>	NGL-microarrays		
32	<i>Ct</i> CBM32 <sub>Cthe_0821</sub>	Lentinan** Pullulan*	NT -/+	β1,4-Man	β1,4-mannose oligosaccharides (Mizutani <i>et al.</i> 2012 <sup>158</sup> )	
35	<i>Ct</i> CBM35 <sub>Cthe_2811</sub>	Galactomannans	++	β1,4-Man	Galactomannan (Ghosh <i>et al.</i> 2014 <sup>141</sup> )	
42	<i>Ct</i> CBM42 <sub>Cthe_2139</sub>	Arabinoxylans	+++	α1,5-Ara;	NA	
	<i>Ct</i> CBM42 <sub>Cthe_2138</sub>		+++	α1,5-Ara-		
	<i>Ct</i> CBM42 <sub>Cthe_1273</sub>		+++	(α1,2/α1,3-Ara)		
50	<i>Ct</i> CBM50 <sub>Cthe_0300</sub>	-	-	β1,4-GlcNAc <sup>#</sup>	NA	
	<i>Ct</i> CBM50 <sub>Cthe_2387</sub>					
	<i>Ct</i> CBM50-2 <sub>Cthe_3006</sub>					
	<i>Ct</i> CBM50-3 <sub>Cthe_3006</sub>					
	<i>Ct</i> CBM50-1 <sub>Cthe_3005</sub>					
	<i>Ct</i> CBM50-1 <sub>Cthe_1800</sub>					
	<i>Ct</i> CBM50-1 <sub>Cthe_3007</sub>					
62	<i>Ct</i> CBM62 <sub>Cthe_2193</sub>	Galactomannans** Xyloglucans	++ +	Gal-β1,3-Ara; Gal-Xyloglucan; Gal-α1,3-Gal- β1,3-GlcNAc	Galactosyl-mannotriose, Xyloglucan XLXG (2YB7, 2YFZ) (Montanier <i>et al.</i> 2011 <sup>159</sup> )	

<sup>1</sup>\*\*Major binding; \*Weak binding (below 30%); <sup>2</sup>+++ , strong binding; ++, significant binding; +, weak binding; -/+, very weak binding; -, no binding; NC, not conclusive; NT, not tested; NA, not available; <sup>3</sup>CBMs for which carbohydrate-binding and structural characterization was already reported are referenced; <sup>#</sup>Binding specificity of *Ct*CBM11<sub>Cthe\_1472</sub> to mixed-linked β1,3-1,4 glucans is detailed in Chapter 4; Binding of *Ct*CBM13<sub>Cthe\_0661</sub> to β1,6-linked-glucose was assessed by ITC with pustulan polysaccharide (Figure S3.2); Binding of *Ct*CBM50<sub>Cthe\_0300</sub> to insoluble chitin polysaccharide was assessed by co-precipitation assay by SDS-PAGE and to β1,4-GlcNAc oligosaccharides by ITC (data shown in Chapter 5).

### 3.2.4.1 Recognition of α-glucans and β-glucans with linear or branched chains

Among the CBM families investigated of both bacteria, only one CBM showed binding to α-glucans, the *C. thermocellum* family 25 *Ct*CBM25<sub>Cthe\_0956</sub>. This CBM showed a moderate but restricted binding to the pullulan polysaccharide (mixed-linked α1,6-1,4-glucose). This protein was described in Chapter 2 for validation of the NGL-screening microarrays, and was shown to be highly specific for linear α1,4-linked glucose, exhibiting a chain-length dependency for sequences longer than DP-4 (Figures 2.2 and Table S2.1, probes 23-32). This CBM is a new member of starch-binding family CBM25 characterised as a α1,4-glucan binding domain, in accordance with CBM25 from *Bacillus halodurans* which has also been reported to bind α1,4-linked maltose or amylose oligosaccharides<sup>126</sup>.

The specific recognition of β-glucans with different linear or branched sequences was identified for CBMs of both bacteria, which is exemplified in *C. thermocellum* by CBMs of families 3, 4, 11 and 30 and in *R. flavefaciens* by CBMs of family 4.

Four out of the nineteen family 3 *Ct*CBMs analysed, showed binding to xyloglucan polysaccharide fractions (probes 18-19). These included the characterised CBM3 from cellulosomal scaffoldin CipA *Ct*CBM3<sub>Cthe\_3077</sub>, the anti-σ-cell surface cellulose sensor Rsg11 *Ct*CBM3<sub>Cthe\_0059</sub>, the *Ct*CBM3-2<sub>Cthe\_0040</sub> from endoglucanase 9I (Cel9I) and the uncharacterised *Ct*CBM3<sub>Cthe\_0433</sub>. Of

**Table 3.2. Summary of the *R. flavefaciens* FD-1 CBMs ligand recognition and specificity obtained in the carbohydrate microarray screenings and affinity gel electrophoresis (AGE).**

Family	CBM	1 <sup>st</sup> screening		2 <sup>nd</sup> screening
		Polysaccharide microarrays <sup>1</sup>	AGE <sup>2</sup>	NGL-microarrays
4	<i>RfCBM4</i> <sub>2836</sub>	PGG- $\beta$ -glucan	+	-
	<i>RfCBM4</i> <sub>3259</sub>	Mixed-linked $\beta$ -glucans**	+++	$\beta$ 1,3-Glc
		Arabinoxylans	-	
	<i>RfCBM4</i> <sub>776</sub>	Curdlan	NT	$\beta$ 1,3-1,4-Glc
		Mixed-linked $\beta$ -glucans**	+++	
<i>RfCBM4</i> <sub>3995</sub>	Arabinoxylans	-/+	$\beta$ 1,4-Xyl/ $\beta$ 1,3-1,4-Xyl	
6	<i>RfCBM6</i> <sub>3747</sub>	Arabinoxylans	NC	$\beta$ 1,4-Xyl/ $\beta$ 1,3-1,4-Xyl; $\beta$ 1,4-Xyl( $\alpha$ 1,3-Ara)
13	<i>RfCBM13-1</i> <sub>2115</sub>	Arabinoxylans*	-/+	$\alpha$ 1,5-Ara
		Galactomannans*	NT	
	<i>RfCBM13</i> <sub>2326</sub>	Arabinoxylans	NC	$\alpha$ 1,5-Ara-( $\alpha$ 1,2/ $\alpha$ 1,3-Ara)
	<i>RfCBM13</i> <sub>694</sub>	Arabinoxylans***	-#	Gal- $\beta$ 1,3-Ara
22	<i>RfCBM22-2</i> <sub>1878</sub>		+++	
	<i>RfCBM22</i> <sub>2649</sub>		NC	
	<i>RfCBM22</i> <sub>1615</sub>		++	
	<i>RfCBM22</i> <sub>1272</sub>		++	
	<i>RfCBM22</i> <sub>2646</sub>		++	
	<i>RfCBM22</i> <sub>3180</sub>	Xylans	+++	$\beta$ 1,4-Xyl/ $\beta$ 1,3-1,4-Xyl
		Arabinoxylans	+++	
	<i>RfCBM22-1</i> <sub>2002</sub>		+++	
	<i>RfCBM22-1</i> <sub>3077</sub>		+++	
	<i>RfCBM22-2</i> <sub>3077</sub>		+++	
	<i>RfCBM22</i> <sub>3190</sub>		+++	
	<i>RfCBM22</i> <sub>1766</sub>		+++	
	<i>RfCBM22-2</i> <sub>1737</sub>		+++	
<i>RfCBM22-1</i> <sub>3270</sub>		NC <sup>#</sup>	-	
35	<i>RfCBM35-2</i> <sub>1364</sub>		-/+	-
	<i>RfCBM35</i> <sub>2302</sub>	Galactomannans	+	$\beta$ 1,4-Man
			++	$\alpha$ 1,5-Ara; $\beta$ 1,4-Xyl( $\alpha$ 1,3-Ara)

1\*\*Major binding; \*Weak binding (below 30%); <sup>2</sup>+++ , strong binding; ++, significant binding; +, weak binding; -/+, very weak binding; -, no binding; NC, not conclusive; NT, not tested; NA, not available; #Binding of *RfCBM13*<sub>694</sub> was observed to pectin polysaccharides using a different microarray (Figure S3.3) and those results were validated by AGE (Figure 3.6); *RfCBM22-1*<sub>3270</sub> presented a stability issue which reflected in a non-conclusive AGE analysis.

these, only *CtCBM3*<sub>Cthe\_0433</sub> showed binding to mixed-linked  $\beta$ 1,3- $\beta$ 1,4-glucans (probes 12-13). The main binding reported for the characterised family 3 CBMs is to crystalline cellulose<sup>160-162</sup>. These CBMs have a planar hydrophobic binding surface that makes apolar interactions with stretches of  $\beta$ 1,4-Glc sequences, typical of type A CBMs. But these CBMs could also bind to xyloglucan or other plant cell wall  $\beta$ -glucans with weaker affinity. This explains why only four of the nineteen family 3 CBMs analysed were able to bind to the soluble polysaccharide fractions that share a  $\beta$ 1,4-linked glucose backbone, while no binding was observed to oligosaccharides in the 2<sup>nd</sup> screening NGL-microarrays.

The family 4 CBMs showed a broad binding to  $\beta$ 1,3-,  $\beta$ 1,4- or mixed-linked  $\beta$ 1,3- $\beta$ 1,4-glucans, and to  $\beta$ 1,3- $\beta$ 1,4-xylans or arabinoxylans. These CBMs are associated with glycoside hydrolases

of family 16 (GH16), which is reported to be active on  $\beta$ 1,4- or  $\beta$ 1,3-glycosidic linkages of glucans<sup>163</sup>, or family 9 (GH9) with main cellulase activity, but also can exhibit xylanase activity<sup>164</sup> (proteins modular architecture in Table S3.1). The analysis in the 2<sup>nd</sup> screening NGL-microarrays could differentiate the carbohydrate-binding specificities of the CBMs from both bacteria. The *C. thermocellum* CtCBM4-1<sub>Cthe\_2809</sub> and CtCBM4-3<sub>Cthe\_2809</sub> associated in tandem with a GH16, exhibited a restricted specificity towards  $\beta$ 1,3-glucans, but only the latter could bind to oligosaccharides, whereas CtCBM4<sub>Cthe\_0413</sub> of a GH9 bound only to  $\beta$ 1,4-glucose oligosaccharides. The *R. flavefaciens* RfCBM4<sub>776</sub> and RfCBM4<sub>3955</sub>, which are associated to GH9 enzymes, bound specifically to  $\beta$ 1,4-glucose oligosaccharides or to  $\beta$ 1,3- $\beta$ 1,4-xylan oligosaccharides, respectively. On its turn, the RfCBM4<sub>3259</sub> associated with a GH16 showed restricted binding to  $\beta$ 1,3-linked glucose oligosaccharides. Thus, the screening microarray results shows that the specificity identified for these CBMs may resemble the substrate specificity of the associated enzymes and agree with what has been reported for CBM4 family, which targets primarily  $\beta$ 1,3-glucans,  $\beta$ 1,4-glucans or xylans<sup>155</sup>. In addition, the analysis in the NGL-microarrays showed increased binding intensities with the oligosaccharide chain-length for all the CBMs, indicating their endo-binding mode as type B CBMs.

CtCBM11<sub>Cthe\_1472</sub> is the archetypal member of family 11 and the only *C. thermocellum* CBM assigned to family 11. This CBM exhibited binding to mixed-linked  $\beta$ 1,3-1,4-glucans, and to the range of  $\beta$ -xylans and xyloglucan polysaccharide fractions included in the microarrays. But in the 2<sup>nd</sup> screening NGL-microarrays, this CBM exhibited preferential binding to mixed-linked  $\beta$ 1,3-1,4-glucose oligosaccharides, as previously reported by Palma *et al.* 2015<sup>32</sup>, showing only a weak binding to  $\beta$ 1,4-glucose oligosaccharides. The molecular basis for the unique binding specificity of this CBM towards mixed-linked  $\beta$ 1,3-1,4-glucans will be further explored in Chapter 4. The family 30 CtCBM30<sub>Cthe\_0624</sub> also showed strong binding to mixed-linked  $\beta$ 1-3,1-4-glucans, but bound in addition to xyloglucan polysaccharide fractions and no binding was detected to oligosaccharides. The results are in accord with the data reported for this CBM by Najmudin *et al.* 2006<sup>165</sup>, which showed binding to barley  $\beta$ -glucan and xyloglucan sequences.

#### 3.2.4.2 Recognition of linear $\beta$ 1,4 mannans and branched galactomannans

The recognition of mannans was identified for family 35 CBMs of both bacteria, which showed main binding to galactomannan polysaccharides. CtCBM35<sub>Cthe\_2811</sub> and RfCBM35<sub>2302</sub> exhibited a similar binding pattern in the 2<sup>nd</sup> screening NGL-microarrays, which was restricted to  $\beta$ 1,4-linked mannose oligosaccharides and dependent on the oligosaccharide chain-length up to DP-8. The substitution of the backbone with an  $\alpha$ 1,6-galactose prevented the binding (Chapter 2, Table S2.1, probes 193-196). Evidence for CtCBM35<sub>Cthe\_2811</sub> affinity for carob galactomannan and konjac glucomannan has been demonstrated previously by Ghosh *et al.* 2014<sup>141</sup>, and reflects the binding recognition of  $\beta$ 1,4-mannans, galactomannans and glucomannans, reported for other CBM members of family 35<sup>166,167</sup>. Unexpectedly, in the NGL-microarrays RfCBM35<sub>933</sub> showed binding

to oligosaccharides containing  $\alpha$ -linked arabinose, including arabinoxylan, linear and branched arabinan oligosaccharides (probes 165-181). Thus, the microarray results reported here for this CBM are not conclusive and further analyses will be required to clarify the carbohydrate binding specificity of this CBM.

The family 32 *Ct*CBM32<sub>Cthe\_0821</sub> showed main binding to  $\beta$ 1,4-linked-mannose oligosaccharides, in accordance with the specificity reported by Mizutani *et al.* 2014<sup>168</sup> using isothermal titration calorimetry (ITC). However, for this CBM divergent polysaccharide binding patterns to both  $\alpha$ - and  $\beta$ -glucans were observed: binding to the branched  $\beta$ 1,3(1,6)-branched glucans lentinan and grifolan, to mixed-linked  $\alpha$ 1,6-1,4 glucan pullulan, and no binding to the galactomannan polysaccharides. Given these results, the specificity of this CBM needs to be further explored in order to fully understand its binding capabilities.

### 3.2.4.3 Recognition of $\alpha$ -arabinose- and galactose-containing sequences in different polysaccharides

Remarkably, the family 13 CBMs of the two bacteria exhibited distinct binding profiles. The *Rf*CBM13-<sub>12115</sub> showed main binding to linear  $\alpha$ 1,5-linked arabinose oligosaccharides (probes 173-179), whereas *Rf*CBM13<sub>2326</sub> bound exclusively to the  $\alpha$ 1,2(1,3)-branched arabinose oligosaccharide (probe 181) and *Rf*CBM13<sub>694</sub> didn't show binding to any of the  $\alpha$ -arabinose containing oligosaccharides, but bound, albeit weakly, to the Gal $\beta$ 1,3Ara disaccharide (probe 216). These results explain why only *Rf*CBM13<sub>2326</sub> showed a stronger binding to arabinoxylan polysaccharide fractions (arabinans were not included in polysaccharide microarrays) but results with AGE were inconclusive. The *Ct*CBM13<sub>Cthe\_0661</sub> showed main binding to xyloglucan polysaccharides, while in NGL-microarray this CBM bound, albeit weakly to the Gal $\beta$ 1,3Ara disaccharide (probe 216). Results reported by Jiang *et al.*, 2012 showed that *Ct*CBM13<sub>Cthe\_0661</sub> bound to galactose and  $\beta$ -galactose-containing oligosaccharides. Thus, the binding pattern for this CBM in the polysaccharide microarray may be explained by the presence of  $\beta$ -galactose in the xyloglucans. Overall, family 13 CBMs have been described to have multivalent carbohydrate-binding ability, being found in distinct GHs such as  $\beta$ -xylanases,  $\alpha$ -galactosidases and endo- $\beta$ 1,3-1,4-glucanases<sup>169</sup>, which seem to be supported by the results reported here.

The restricted binding of both *Ct*CBM13<sub>Cthe\_0661</sub> and *Rf*CBM13<sub>694</sub> to the non-reducing  $\beta$ -galactose sequences, raised the possibility for recognition of a  $\beta$ -galactose epitope in pectic polysaccharides, which were not included in the screening microarrays. Indeed, in a more recent analysis using a microarray of pectin polysaccharides, these CBMs showed differential and possible specific binding to pectic  $\beta$ -galactans (Figure S3.3 and Tables S3.9). *Ct*CBM13<sub>Cthe\_0661</sub> showed restricted binding to pectic  $\beta$ -galactans from Lupin, whereas *Rf*CBM13<sub>694</sub> showing a broader binding to both lupin and potato  $\beta$ -galactans (also observed by AGE in Figure 3.6), and also to Soybean rhamnogalacturonan and other pectin fractions. The quality control of these microarrays is ongoing and the specificity of these CBMs towards these types of structures require



further investigation. The binding specificity of *R. flavefaciens* family 13 will be further explored in Chapter 6 with the 3-D structure determination of *RfCBM13-1*<sub>2115</sub>.

The *C. thermocellum* CBMs from family 42, *CtCBM42*<sub>Cthe\_2139</sub>, *CtCBM42*<sub>Cthe\_2138</sub> and *CtCBM42*<sub>Cthe\_1273</sub>, showed a restricted strong binding to arabinoxylan polysaccharide fractions and a binding-specificity towards  $\alpha$ -linked arabinose in branched arabinoxylan, arabinan or linear  $\alpha$ 1,5-linked arabinose oligosaccharide sequences. *CtCBM42*<sub>Cthe\_2139</sub> could also bind to albeit weakly to unsubstituted  $\beta$ -xylans. These results reflect the ligand-specificities reported for CBMs of this family and activities of their associated GH43 catalytic modules towards arabinoxylans and arabinans<sup>36</sup>.

*C. thermocellum* family 62 *CtCBM62*<sub>Cthe\_2193</sub> showed binding to galactomannan and xyloglucan polysaccharides. This binding may be explained by a common galactose epitope in the galactomannan and xyloglucan that is being recognised. In line with this, the CBM exhibited a restricted binding to the Gal $\beta$ 1,3Ara disaccharide (probe 216) and to the  $\beta$ -galactose substituted xyloglucan oligosaccharides (probes 202-205). Indeed, when analysed in the pectin polysaccharide microarrays *CtCBM62*<sub>Cthe\_2193</sub> showed restricted binding to the pectic  $\beta$ 1,4-galactans from Lupin. These binding patterns are in accordance with the data previously reported by Montanier *et al.*, 2011 for this CBM binding to the terminal D-galactose residues in xyloglucan, galactomannan and arabinogalactan polysaccharides<sup>159</sup>. Although carbohydrate binding could not be detected for *R. flavefaciens* family 62 *RfCBM62-1*<sub>3398</sub>, this CBM was analysed in parallel with the other proteins in the pectin polysaccharide microarrays (Figure S3.3 and Tables S3.9). In this analysis, the CBM showed a restricted binding to the pectin fractions from *Vernonia kotschyana* (Vk100-Fr.I) and *Sambucus nigra* (100WSnFI-S2) and weak binding to the pectic galactans from lupin. AGE analysis with pectin from apple and pectic galactan from lupin, evidenced a slight reduction of the CBM's electrophoretic migration (Figure 3.6). The binding patterns exhibited by the proteins analysed in the pectin microarrays, suggest a galactose-containing epitope that is being differentially recognised by *RfCBM62-1* in these pectin fractions. Further work will be needed to clarify the binding specificity of family 62 CBMs in *R. flavefaciens*.

#### **3.2.4.4 Assignment of *C. thermocellum* family 50 CBMs ligand specificity towards $\beta$ 1,4-GlcNAc oligosaccharides**

For *C. thermocellum* family 50 CBMs, binding was not observed in the 1<sup>st</sup> microarray screening. Given the high number of these CBMs in the genome of *C. thermocellum* and the predicted binding of this CBM family to  $\beta$ 1,4-GlcNAc residues in bacterial peptidoglycans and in chitin<sup>42</sup>, all the 11 expressed CBMs were analysed in the 2<sup>nd</sup> screening NGL-microarrays, which included  $\beta$ 1,4-GlcNAc and  $\beta$ 1,4-GlcN sequence-defined oligosaccharides (Table S2.1, probes 205-214). The validation of these NGL probes in the microarrays was carried out using the plant lectins Wheat Germ agglutinin (WGA) and *Datura stramonium* lectin (DSL), which showed the reported

specificity and chain-length requirements towards  $\beta$ 1,4-GlcNAc oligosaccharides (Figure S3.4). The microarray analyses identified binding for 7 CBMs, which showed restricted binding to  $\beta$ 1,4-GlcNAc oligosaccharides. The binding intensities observed increased with the oligosaccharide chain-length, indicating an endo-binding mode of a type B CBM. The binding specificity of CtCBM50s, chain-length requirement and the molecular determinants that govern the carbohydrate recognition, with insights into their function will be further explored in Chapter 5.

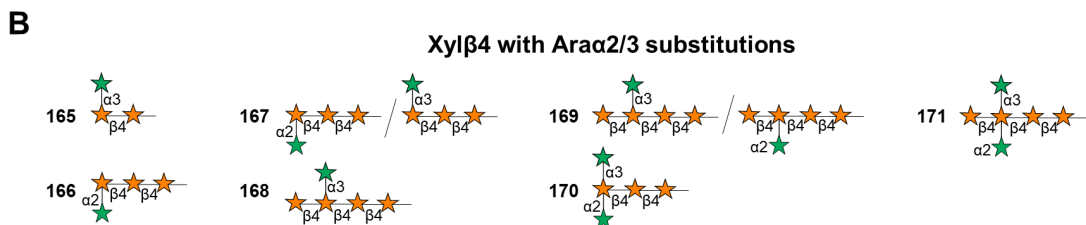
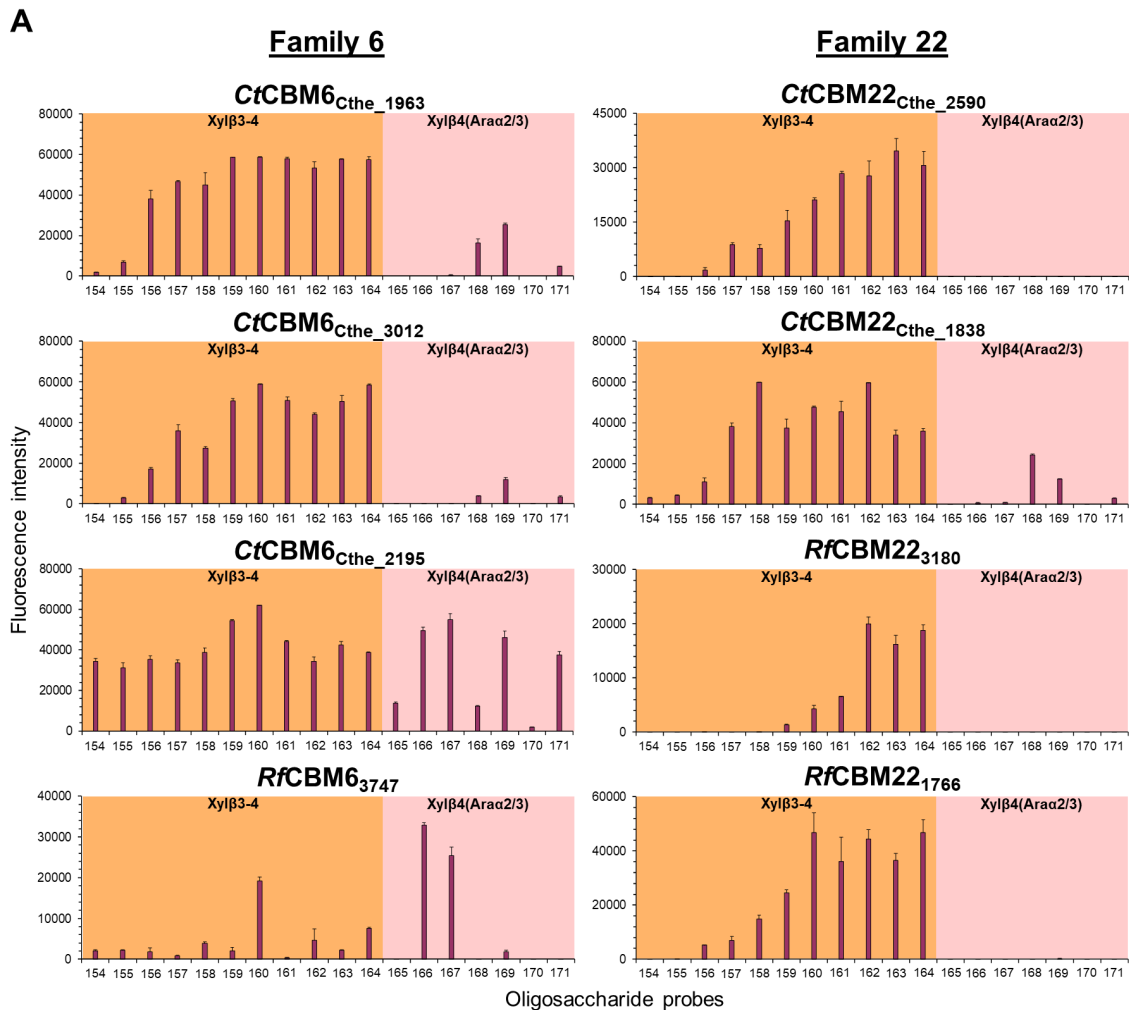
### 3.2.4.5 Assignment of ligand specificity and chain-length requirement for families 6 and 22 CBMs towards $\beta$ -xylans

The main binding obtained to  $\beta$ -xylans and arabinoxylans is exemplified by families 6 and 22 CBMs from both bacteria (Figure 3.8A). Remarkably *C. thermocellum* showed a higher number of family 6 CBMs, which exhibited binding to  $\beta$ -xylans and arabinoxylans, whereas *R. flavefaciens* showed a higher number of family 22 CBMs highly specific to  $\beta$ -xylans and arabinoxylans. In overall, the results reflect the reported specificity for members of family 6 to non-reducing termini of xylo-oligosaccharides, xylans and arabinoxylans<sup>38,170–172</sup>, as well as the preferential binding of family 22 CBMs to  $\beta$ -xylans and xylo-oligosaccharides<sup>136</sup>.

Comparing the oligosaccharide binding patterns, family 22 CBMs exhibited a binding specificity restricted to linear  $\beta$ 1,4- or  $\beta$ 1,3- $\beta$ 1,4-xylose sequences, with a chain-length dependency from DP-5 up to DP-13, displaying a mono-specific carbohydrate recognition in both bacteria. Additionally, CtCBM22<sub>Cthe\_1838</sub> showed binding to  $\beta$ 1,4-xylose tetrasaccharides with a single internal  $\alpha$ 1,2- or  $\alpha$ 1,3-arabinose branching (Figure 3.8B, probes 168-169). CBMs from family 6 however, presented distinct binding patterns between the two bacteria. *C. thermocellum* family 6 CBMs showed similar binding specificities to  $\beta$ 1,4- or  $\beta$ 1,3- $\beta$ 1,4-xylose sequences in the range of DP-3 to DP-13, also recognising  $\beta$ 1,4-xylose oligosaccharides with  $\alpha$ -arabinose substitutions. *R. flavefaciens* RfCBM6<sub>3747</sub> bound predominantly to arabinoxylan-derived oligosaccharides with  $\alpha$ 1,2-arabinose substitutions in the non-reducing terminal xylose (probes 166 and 167).

The binding patterns of family 6 CBMs to the arabinoxylan-derived oligosaccharides evidenced the importance of the free non-reducing  $\beta$ 1,4-xylose terminal for recognition, but also the influence of the  $\alpha$ -arabinose branching. Although the majority of the CBMs bound only sequences exhibiting the free non-reducing xylose terminal, CtCBM6<sub>Cthe\_2195</sub> and CtCBM6<sub>Cthe\_2197</sub> were able to accommodate sequences with  $\alpha$ -arabinose branches in the non-reducing terminal xylose (Figure 3.8A, probes 165 to 167). When comparing the binding intensities, it becomes evident that these two CBMs were able to bind sequences exhibiting  $\alpha$ 1,2-arabinose substitutions in the non-reducing terminal xylose, whereas the  $\alpha$ 1,3 configuration was disfavoured. The same binding trend could be observed for probes 168 and 169.

Interestingly, out of the three family 6 CBMs expressed by *R. flavefaciens*, two are found in large modular proteins associated with family 22 CBMs, RfCBM6<sub>2649</sub> which did not bind in the microarrays and RfCBM6<sub>1737</sub> which did not express. These CBMs are also associated with GH43



**Figure 3.8. Comparison of carbohydrate-binding specificities of families 6 and 22 CBMs from *C. thermocellum* and *R. flavefaciens* FD-1 to xylan sequences. (A)** The binding signals of representative CBMs from each bacterium are depicted as means of fluorescence intensities of duplicate spots at 5 fmol of oligosaccharide probe arrayed (with error bars) and are representative of at least two independent experiments (corresponding to the binding patterns shown in Figure 3.7). Numerical scores are given in Tables S3.7 and S3.8. The different carbohydrate groups are indicated in the coloured panels. **(B)** The sequences of the branched  $\beta$ 1,4-xylan( $\alpha$ 1,2-arabinose) probes are depicted by microarrays position. Carbohydrate sequence information on these probes is in Chapter 2, Table S2.1.

catalytic modules, which are reported to have  $\alpha$ -arabinofuranosidase,  $\beta$ -xylosidase,  $\alpha$ -arabinanase and  $\beta$ -galactosidase activity in the degradation of hemicelluloses and pectins, and are frequently found in association with family 6 CBMs<sup>173</sup>. Given the poly-specificity of CBMs from family 6 and family 43 GH's, it is not surprising that *R. flavefaciens* family 6 CBMs might exhibit

distinct binding specificities than those from *C. thermocellum*, and for which the target sequences were not included in the microarrays.

In the context of *R. flavefaciens* cellulosome, the small number of family 6 CBMs and its association with family 22 CBMs, of distinct binding specificities, might point to a complementary function of these modules, evidencing a crucial role of family 6 CBMs in *R. flavefaciens*.

### 3.2.5 CBM families for which carbohydrate binding was not identified in the microarray analyses

*C. thermocellum* CBMs from families 34, 44, 48 and 54 were not successfully expressed using the high-throughput strategy and were not analysed in the screening microarrays. Of those CBMs that were expressed and analysed, none of the *C. thermocellum* CBMs from family 16 or *R. flavefaciens* CBMs from families 3, 48 or 63 showed binding in the microarrays (Table S3.2). The family 16 CBMs from *Caldanaerobius polysaccharolyticus* (formerly *Thermoanaerobacterium polysaccharolyticum*) were reported to bind both  $\beta$ 1,4-linked glucose and  $\beta$ 1,4-linked mannose sequences, suggesting the linear  $\beta$ 1,4-glucomannan as natural substrate<sup>174</sup>. The *C. thermocellum* CBMs from this family may share similar carbohydrate-binding specificity, explaining why no binding was observed, as glucomannan polysaccharides or oligosaccharides were not included in the microarrays. Although binding was observed with 4 out of the 19 *C. thermocellum* family 3 CBMs to soluble glucans with a  $\beta$ 1,4-linked backbone, these CBMs are characterized to bind to crystalline cellulose with higher affinity<sup>160–162</sup>. The analysed *RfCBM3*<sub>929</sub>, which is associated to a GH9 cellulase, could be involved in the recognition of insoluble  $\beta$ 1,4-glucans<sup>154</sup>. Family 48 CBMs have reported activities towards starch and glycogen, binding to linear and cyclic sequences of  $\alpha$ 1,4- and  $\alpha$ 1,6-linked glucose<sup>175</sup>. As only linear  $\alpha$ -glucans were included in the microarrays, might be possible that the *RfCBM48*s require branched or cyclic sequences for binding recognition. The *Bacillus subtilis* CBM63 is reported to be associated with expansin module EXLX1 and to mediate the expansin binding to cell wall cellulose<sup>176</sup>. The analysed *RfCBM63*<sub>2821</sub>, being also linked to an expansin module, may have similar activity towards different forms of cellulose.

### 3.2.6 CBMs spectrum of carbohydrate recognition reflects the bacteria's ecological niche

*C. thermocellum* CBMs showed broader binding patterns, while having a larger repertoire of CAZy families that include a higher number of CBMs specific for recognition of  $\beta$ -glucans. This larger cohort of *C. thermocellum* CBMs, not only from different CAZy families but within the same family (such as the high numbers of family 3 and 6 CBMs), may contribute for its high efficiency in degradation of a wide range of plant cell wall polysaccharides. In addition, the elevated number of family 50 CBMs (LysMs) may confer to *C. thermocellum* an advantage for survival in the extreme conditions of the ecological niche it resides, as LysMs have also been reported to play a role in the development of spores in other spore-forming bacteria, such as *Bacillus subtilis*<sup>42,177</sup>.

*R. flavefaciens* exhibited a more restricted carbohydrate binding recognition and, although apparently expressing less CAZy families, holds a greater number of CBMs targeting hemicelluloses  $\beta$ -xylans,  $\beta$ -mannans and pectic  $\alpha$ -arabinans and galactans. Noteworthy, is the evidence that *R. flavefaciens* cellulosome contains a large number protein modules of unknown sequence homology to assigned CAZy families, six of which were in the course of this thesis reported by Venditto and colleagues<sup>59</sup> to exhibit carbohydrate binding and were assigned to the new CBM families 75 to 80. These CBMs target as major substrates the hemicelluloses xyloglucan and  $\beta$ -mannans,  $\beta$ 1,4- and mixed-linked  $\beta$ 1,3-1,4-glucans, and pectins. In this context, the data reported here complements this study in that the complex *R. flavefaciens* cellulosome seems to incorporate an extended CBM repertoire that promotes the efficient plant cell wall hemicellulose and pectin degradation, even though expressing a small number of CBMs that specifically target crystalline cellulose. Considering its highly dynamic and populated ecological niche, *R. flavefaciens* may also benefit from a cooperative relationship with other members of the mammalian rumen microbiome, such as *Ruminococcus albus* and *Fibrobacter succinogenes* responsible for the breakdown of the recalcitrant structure of cellulose<sup>178</sup>, ensuring its substrate acquisition and survival.

### 3.3 Conclusions

In the present work, different patterns of polysaccharide and oligosaccharide binding by *C. thermocellum* and *R. flavefaciens* CBMs were revealed and novel specificities were assigned. Although *C. thermocellum* CBMs have been more extensively studied and characterized, new CBM carbohydrate binding specificities were identified for CBMs families 25, 42 and 50. For *R. flavefaciens*, ligand-specificities were obtained for 21 CBMs from families 4, 6, 13, 22 and 35. Aiming to decipher the complete CBMome of *R. flavefaciens*, analysis of the remaining CBM families in the oligosaccharide microarrays described in this Chapter and Chapter 2, is required.

Overall, the combined use of high-throughput methodologies allowed to explore the function of *C. thermocellum* and *R. flavefaciens* CBMomes, revealing that the two bacteria present CBMs expressing different carbohydrate-binding specificities, which reflect at some extent the different polysaccharides that each bacterium may encounter in its ecological niche. This comparative study of two bacteria residing in different ecological niches, provides experimental evidence supporting that substrate availability in different habitats may modulate the evolutionary selection of CAZymes to present modules with distinct carbohydrate ligand specificities.

This study also highlights the importance of developing high-throughput methodologies to study these complex systems and unravel carbohydrate recognition. The approach of using in parallel polysaccharide and oligosaccharide microarrays, allows detailed characterization of the specificities of CBMs. While polysaccharide microarrays enable carbohydrate-binding to be assigned, oligosaccharide microarrays can reveal subtle differences in binding profiles and chain-length dependencies, which enables to differentiate between the different topologies of

CBMs binding sites and their functional types. The information obtained from the carbohydrate microarray analyses is crucial to assess the structural characterization of the interactions of CBMs with their oligosaccharide ligands. These integrative studies will be important to elucidate cellulolytic capabilities of these bacteria at the molecular level. To this end, up to 13 CBMs belonging to different CAZy families of both bacteria were selected for large-scale protein expression and purification. Preliminary conditions were obtained for CBMs of families 25 and 50 from *C. thermocellum* and families 6, 13, and 62 from *R. flavefaciens*. Structural characterization of the carbohydrate-binding specificity of CBMs from *C. thermocellum* family 11 and 50 and *R. flavefaciens* family 13 will be explored in Chapters 4, 5 and 6, respectively.

### 3.4 Experimental procedures

#### 3.4.1 Monoclonal antibodies, CBMs and lectins used for microarray quality control

Details on the plant cell wall carbohydrate-directed monoclonal antibodies, CBMs with characterised carbohydrate-binding specificities and plant lectins used for microarray quality control are given in Table S2.3 and section 2.5.1 in Chapter 2.

#### 3.4.2 High-throughput cloning, expression and purification of *C. thermocellum* and *R. flavefaciens* FD-1 CBMs

The bioinformatics sequence analysis of the bacterial genomes for CBM domain selection and the high-throughput gene cloning, protein expression and purification was performed through collaboration with NZYTech Ltd (Lisbon, Portugal), following their established or proprietary protocols. Information on the CBMs protein sequences and protein modularity is described in Tables S3.1 and S3.2. Briefly, the selected genes encoding the CBMs sequences were amplified by PCR from *C. thermocellum* ATCC 27405 (NCBI:txid203119) and *R. flavefaciens* FD-1 (NCBI:txid641112) genomic DNA, using specific primers for ligation independent cloning (LIC) into pHTP1-A57, a pET24a derived vector containing a kanamycin resistance cassette for selection<sup>179</sup>. For recombinant protein expression, *E. coli* BL21 harbouring each CBM encoding gene, containing an *N*-terminal hexa-histidine tag, were cultured in NZY AutoInduction Luria-Bertani (LB) medium (NZYTech, Portugal) at 37 °C until OD<sub>600nm</sub> reached 1.5, at which point temperature was lowered to 25 °C for overnight incubation. Protein purification was achieved by ion metal affinity chromatography (IMAC) using a high-throughput column system. Purified CBMs were in a 50 mM sodium HEPES buffer, pH 7.5, containing 1 M NaCl, 5mM CaCl<sub>2</sub> and 300 mM imidazole. In order to lower the concentration of imidazole for the analysis in the microarrays, a dilution of each CBM solution was performed using the same buffer without imidazole, reaching a final concentration of approximately 170 mM imidazole.

For quality control, CBMs were subjected to SDS-PAGE on 13% (w/v) acrylamide gels, stained with Coomassie Brilliant Blue, in order to assess the purity of recombinant proteins (Figure S3.1).

All proteins were assessed as >95% pure as judged by SDS-PAGE and their concentrations were determined from their calculated molar extinction coefficient using the ProtParam tool (<http://www.expasy.org/tools/protparam.html>) at 280 nm using a SpectraDrop Micro-Volume Microplate (Molecular Devices, USA). From the purified, 105 were selected for microarray screening analysis.

### 3.4.3 Sources of carbohydrates

Hemicellulose polysaccharide fractions included in the polysaccharide microarrays isolated from different sources were obtained through collaboration with Prof. Manuel Coimbra (University of Aveiro, Portugal). These included xylans and xyloglucans from plum; arabinoxylans from brewer's spent grain; arabinogalactan from spent coffee grounds; and mannoprotein isolated from brewer's spent yeast. Pectin fractions isolated from medicinal plants found in Africa included in the pectin microarrays were obtained through collaboration with Prof. Berit Paulsen (University of Oslo, Norway). Pectin from apple was purchased from Sigma-Aldrich (St. Louis, MO, USA). The remaining polysaccharides included in the microarrays and used for AGE analysis, some of which had been previously analysed<sup>32</sup>, were purchased from Megazyme (Bray, Ireland) and Elicytal (Crolles, France). The sources of the soluble polysaccharides its monosaccharide composition can be found in Tables S3.3 and S3.9. Information on the oligosaccharides and sources included in the NGL-microarrays are given in Tables S2.1 and S2.2 and section 2.5.2, in Chapter 2.

### 3.4.4 Carbohydrate microarray analysis

The microarray data and metadata provided here is described according to the MIRAGE glycan microarray guidelines, as described by Liu *et al.* 2016<sup>146</sup>.

The polysaccharide microarray constructed for the 1<sup>st</sup> screening analysis of the CBMs was designated Plant, Fungal and Bacterial Polysaccharide (PS) set 1 and was comprised of a total of 25 structurally diverse polysaccharide samples with major sequences found in plant cell walls  $\beta$ -glucans and hemicelluloses, in fungal  $\alpha$ -glucans,  $\beta$ -glucans and  $\alpha$ -mannans, and in bacterial  $\alpha$ -glucans. The 2<sup>nd</sup> microarray screening was performed using the glucan and hemicellulose oligosaccharide microarray platform comprising 204 neoglycolipid (NGL) probes described in Chapter 2, to which were added 7  $\beta$ 1,4-linked-*N*-acetylglucosamine (chitin) and 3  $\beta$ 1,4-linked glucosamine (chitosan) sequence-defined oligosaccharides, and 4 miscellaneous disaccharides and trisaccharides prepared as NGL probes. Some of the CBMs were also analysed in a pectin polysaccharide microarray designated *Pectin PS set 1*, comprised of 26 pectic polysaccharide fractions.

The information on the probe ID, sequence or monosaccharide composition of the carbohydrate probes featuring in the different types of microarray platforms is shown in Table S2.1 (Glucan, hemicellulose, chitin and chitosan NGL microarray), Table S3.3 (*Plant, Fungal and Bacterial PS set 1*) and Table S3.9 (*Pectin PS set 1*).

For the preparation of the microarrays, the carbohydrate probes were immobilized non-covalently onto 16-pad nitrocellulose-coated FASTTM glass slides (Z721158, Sigma), using a non-contact arrayer robot (Piezorray, Perkin Elmer, Sear Green, UK), with a spot delivery volume of approximately 330  $\mu\text{L}$ , following established protocols<sup>32</sup>. In brief, each carbohydrate probe was printed in duplicate at two levels: polysaccharides at 0.1 and 0.5 mg (dry weight) /mL (30 and 150  $\mu\text{g}/\text{spot}$ ) and NGLs at 5 and 15  $\mu\text{M}$  (2 and 5  $\text{fmol}/\text{spot}$ ). The Cy3 fluorophore was included in the printing solution as a marker for quality control of sample delivery while arraying and spot visualization, as well as for quantitation analysis.

Microarray binding analyses were performed using AlexaFluor-647-labeled Streptavidin for readout, essentially as described by Palma *et al.* 2015<sup>32</sup>. His-tagged CBMs were tested at 5  $\mu\text{g}/\text{mL}$  for the 1<sup>st</sup> screening and at 20  $\mu\text{g}/\text{mL}$  for the 2<sup>nd</sup> screening, precomplexed with mouse monoclonal anti-poly-histidine (Ab1) (Sigma, H1029) and biotinylated anti-mouse IgG (Ab2) (Sigma, B7264) antibodies, at a ratio of 1:3:3 (by weight). The protein–antibody complexes were prepared by preincubating Ab1 with Ab2 for 15 minutes at room temperature, followed by addition of CBMs and incubation further for 15 min, after which the final concentration of the proteins was achieved by dilution in the blocking solution made of 1% (w/v) Casein (Thermo Scientific, 37583) 1:50 1% BSA (Sigma, A8577) in HBS (Sigma, H0887) (5 mM HEPES buffer pH 7.4, 150 mM NaCl) with 5 mM  $\text{CaCl}_2$ . Monoclonal antibodies from Plant probes and Agrisera were probed at 1:10 ratio, as described by Moller *et al.*, 2008<sup>147</sup>, and antibodies from Biosupplies at 10  $\mu\text{g}/\text{mL}$ , diluted in the same blocker, followed by the biotinylated anti-mouse-IgG (Sigma, B7264), anti-rat-IgG (Sigma, B7139) or anti-rat-IgM (Rockland, 612-4607) as appropriate, at 10  $\mu\text{g}/\text{mL}$  in the same blocker. Biotinylated lectins AAL (Vector, B-1395) was analyzed using a single step overlay at a final concentration of 2  $\mu\text{g}/\text{mL}$  in blocker 3% BSA in HBS with 5 mM  $\text{CaCl}_2$ . DSL (Vector B-1185) was analyzed at a final concentration of 25  $\mu\text{g}/\text{mL}$  in blocker 3% BSA and 0.5% Casein in HBS with 5 mM  $\text{CaCl}_2$ . Biotinylated anti-rat and anti-mouse IgG and IgM antibodies were analysed in separate as a negative control. Slides were scanned using GenePix® 4300A microarray scanner (Molecular Devices), at 532 nm for Cy3 spot visualisation, prior to binding assays, and at 647 nm for detection of the binding. Imaging analysis and quantitation was carried out using GenePixPro7 Software (Molecular Devices).

### 3.4.5 Microarray data analysis and presentation

Microarray data analysis was performed using a dedicated software<sup>148</sup>, developed by Mark Stoll of the Glycosciences Laboratory.

After the scrutiny of all the microarray results the following decisions were made as regards presentation of data: 1) to modify the original printed microarray set excluding repeated probes and sorting the probes according to the nature of the sample and predominant oligosaccharide sequence (resulting arrangement of probes is in Tables S2.1, S3.3 and S3.9.); 2) present microarray data in the form of a matrix represented as a heatmap of the relative binding intensities



(Figures 3.3, 3.4 and 3.7 and Figure S3.3), in order to highlight the different binding patterns obtained for the proteins and antibodies analysed; 3) in order to accurately depict the binding patterns for each protein and antibody, the probes printed high level (5 fmol/spot) were selected to generate the matrices and graphics (Figures 3.3, 3.4, 3.7 and 3.8, and Figures S3.3 and S3.4).

#### **3.4.6 Affinity gel electrophoresis with soluble polysaccharides**

All CBMs from each family that gave binding patterns in the 1<sup>st</sup> screening were used to validate the microarray results through affinity gel electrophoresis (AGE) with the respective ligand for which major binding was observed, following the method described by Abbott *et al*, 2000<sup>62</sup>. Soluble polysaccharides were prepared in Mili-Q water at 1% (w/v). CBMs at 5 µg were subjected to non-denaturing electrophoresis in gels containing 13% (w/v) acrylamide and the soluble polysaccharide at 0.1% (w/v). Bovine serum albumin (BSA) was used as a non-binding negative control. A control non-denaturing electrophoresis gel without the ligand was ran simultaneously.

### **3.5 Work contributions**

The printing of the polysaccharide microarrays was carried out the the Glycosciences Laboratory (Imperial College London) with the help of Dr. Lisete M. Silva. Mannoprotein, xylan, arabinoxylan and arabinogalactan polysaccharide fractions were obtained through a collaboration with Prof. Manuel Coimbra (Universidade de Aveiro, Portugal). Pectin fractions were obtained from Prof. Berit Paulsen (Oslo University, Norway). Protein expression and purification of the CBMs using the high-throughput platform was performed by Dr. Joana Brás at NZYTech (Lisbon, Portugal), as result from a long-standing collaborative work of the Supervisors, with Prof. Carlos Fontes. The experimental planning and work reported here related to the, carbohydrate microarray validation, binding and data analysis and interpretation, protein and AGE analysis analyses were performed by the author of the Thesis.



# CHAPTER 4

---

## **MOLECULAR BASIS FOR THE PREFERENTIAL RECOGNITION OF $\beta$ 1,3-1,4-GLUCANS BY THE FAMILY 11 CBM FROM *CLOSTRIDIUM THERMOCELLUM*<sup>2</sup>**

---

<sup>2</sup>Partially reproduced from Ribeiro, D.O., Viegas, A., Pires, V.M.R., Medeiros-Silva, J., Bule, P., Chai, W., Marcelo, F., Fontes, C.M.G.A., Cabrita, E.J., Palma, A.S., Carvalho, A.L., Molecular basis for the preferential recognition of  $\beta$ 1,3-1,4-glucans by the family 11 Carbohydrate-Binding Module from *Clostridium thermocellum*. FEBS J. (2019) (DOI: 10.1111/febs.15162).



## 4 Molecular basis for the preferential recognition of $\beta$ 1,3-1,4-glucans by the family 11 CBM from *Clostridium thermocellum*

### 4.1 Introduction

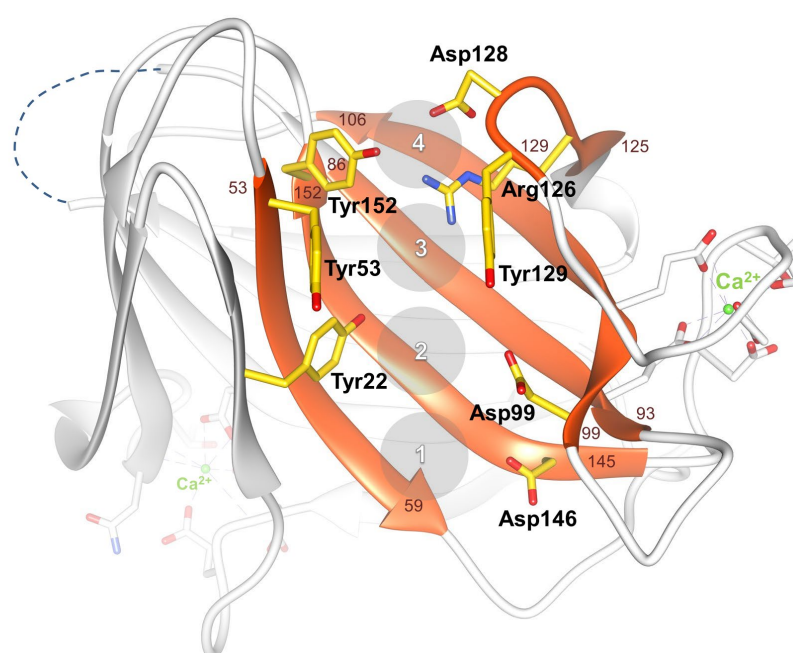
Plant cell walls are composed of structurally diverse and complex polysaccharides presenting many biological and biotechnological applications<sup>43,61,127,128</sup>. The mixed-linked  $\beta$ 1,3-1,4-glucan polysaccharides (or mixed-linked glucans) are unevenly distributed across the plant kingdom but are abundant in the cell walls of most *Poaceae* members. These include the endosperm of cereals and grasses, which are of considerable economic importance as storage tissues<sup>180–182</sup>. Mixed-linked glucans are also found in the walls of algae, pathogenic fungi, and lichen-forming ascomycete symbionts. These glucans have several commercial and biotechnological applications and are of particular interest for the malting and brewing processes and bioenergy production<sup>43,183</sup>, as well as sources of dietary fibres with major health benefits<sup>184</sup>. These properties of  $\beta$ 1,3-1,4-glucans make the study of their recognition by proteins of fundamental importance.

Mixed-linked glucans are composed by a linear chain of 2 to 5  $\beta$ 1,4-linked D-glucopyranose residues separated by single  $\beta$ 1,3 linkages (Figure 1.2, Chapter 1)<sup>185</sup>. The  $\beta$ 1,4-linked residues form rigid regions while the  $\beta$ 1,3-linkages are flexible, creating kinks within the linear backbone chain<sup>180,186</sup>. This results in an extended twisted conformation of the polysaccharide, which presents a unique binding surface for recognition by proteins<sup>185,186</sup>, including CBMs<sup>33</sup> and GHs<sup>187</sup>. In addition, the backbone incorporation of  $\beta$ 1,3-linkages renders the polysaccharide much more soluble than cellulose.

In recent years, enzymatic systems employed by cellulolytic microorganisms to efficiently hydrolyse the plant cell wall polysaccharides have been gaining interest to reduce energy costs and avoid the usage of environmentally harmful chemical processes. One of these microorganisms is the thermophilic anaerobic bacterium *Clostridium thermocellum*<sup>12</sup> that assembles its enzymatic machinery at the cell surface in a multi-protein complex termed the cellulosome (Figure 1.3, Chapter 1). Several CBMs are involved in the recognition of  $\beta$ 1,3-1,4-glucans and, due to their diversity in the cellulosome, are excellent case-studies to rationalize molecular recognition mechanisms that determine the specificity of mixed-linked glucans recognition in general<sup>43,128,188–190</sup>. An archetypal example is the family 11 CBM (CtCBM11<sub>Cthe\_1472</sub>) of the *C. thermocellum* Lic26A-Cel5E, an enzyme that contains GH5 and GH26 catalytic domains that display  $\beta$ 1,4- and  $\beta$ 1,3-1,4-mixed-linked endoglucanase activity<sup>33</sup>.

Previous work has demonstrated that CtCBM11<sub>Cthe\_1472</sub>, henceforward designated as CtCBM11, exhibited a preference for mixed-linked  $\beta$ 1,3-1,4-glucans and lower affinity for  $\beta$ 1,4-linked glucans<sup>32,33</sup>. The three-dimensional structure of CtCBM11 (PDB ID 1V0A), in harness with

mutagenesis studies, revealed a typical type B CBM with a  $\beta$ -sandwich fold with a concave side forming a putative single binding cleft that could accommodate  $\beta$ 1,3-1,4- and  $\beta$ 1,4-linked glucans<sup>33</sup>. Aminoacid residues Tyr22, Tyr53, and Tyr129, located in the putative binding cleft, were identified as playing a central role in the recognition of the ligands<sup>33,191</sup>. Interaction studies using STD-NMR with the  $\beta$ 1,4-linked cellohexasaccharide, showed that CtCBM11 interacted preferably with the central four glucose-units, mainly through interactions with internal positions 2 and 5 of the glucose rings<sup>191,192</sup>. Overall, these studies suggested that CtCBM11 contained four binding subsites (Figure 4.1), with the carbohydrate reducing end always facing the same side of the protein (subsite 1). The approximately four times higher affinity for the mixed-linked tetrasaccharide G4G4G3G, when compared to  $\beta$ 1,4-linked cellotetrasaccharide, suggested that CtCBM11 displays a preference for a  $\beta$ 1,3-linked glucose in at least one of the four subsites.



**Figure 4.1. Top view on the identified binding site of wild-type CtCBM11.** Analysis of the crystal structure of unbound CtCBM11 (PDB ID 1V0A)<sup>33</sup>, together with mutagenesis and interaction studies using ITC, NMR and molecular docking allowed to pinpoint the protein's binding site (englobed by the  $\beta$ -strands in orange) and identify some key residues involved in ligand recognition (e.g.: Tyr22, Tyr53, Asp99, Arg126, Tyr129, Asp146 or Tyr152, here represented as sticks and depicted with yellow carbon atoms)<sup>33,191,192</sup>. The polypeptide chain of CtCBM11 is depicted in white ribbon, with stretches Tyr53-Ser59, Arg86-Ser93, Asp99-Ser106, Arg125-Tyr129, Asn144-Tyr152 coloured in orange and numbered. The individual glucose binding subsites are schematized as transparent grey circles, numbered from 1 to 4. Subsite 1 accommodates the carbohydrate reducing end<sup>191</sup>. Calcium atoms are represented as green spheres. Images generated using UCF Chimera<sup>40</sup>.

In recent studies, the ability of CtCBM11 to bind to  $\beta$ 1,4- and with higher affinity  $\beta$ 1,3-1,4-linked glucans, has been exploited for its use as a tool for the biotransformation of lignocellulosic materials. Fonseca-Maldonado *et al.* investigated the  $\beta$ 1,3-1,4 glucanase activity of a chimeric *Bacillus subtilis* endo- $\beta$ 1,4-glucanase (*BsCel5A*), after exchanging its CBM3 domain by CtCBM11, which resulted in an increase of the hydrolytic efficiency of the enzyme towards  $\beta$ 1,3-1,4-glucans<sup>193</sup>. Cattaneo *et al.*, have designed a chimeric protein by adding a CtCBM11

module to the C-terminus of a hyperthermostable endoglucanase from *Dictyoglomus turgidum* (*Dtur* CelA). The resulting chimeric enzyme displayed enhanced stability at extreme pHs, with higher affinity and activity on insoluble cellulose<sup>194</sup>. Furthermore, Furtado *et al.*, combined directed protein evolution and phage display approaches to obtain engineered CtCBM11 mutants that would exhibit high affinity to xyloglucans<sup>195</sup>.

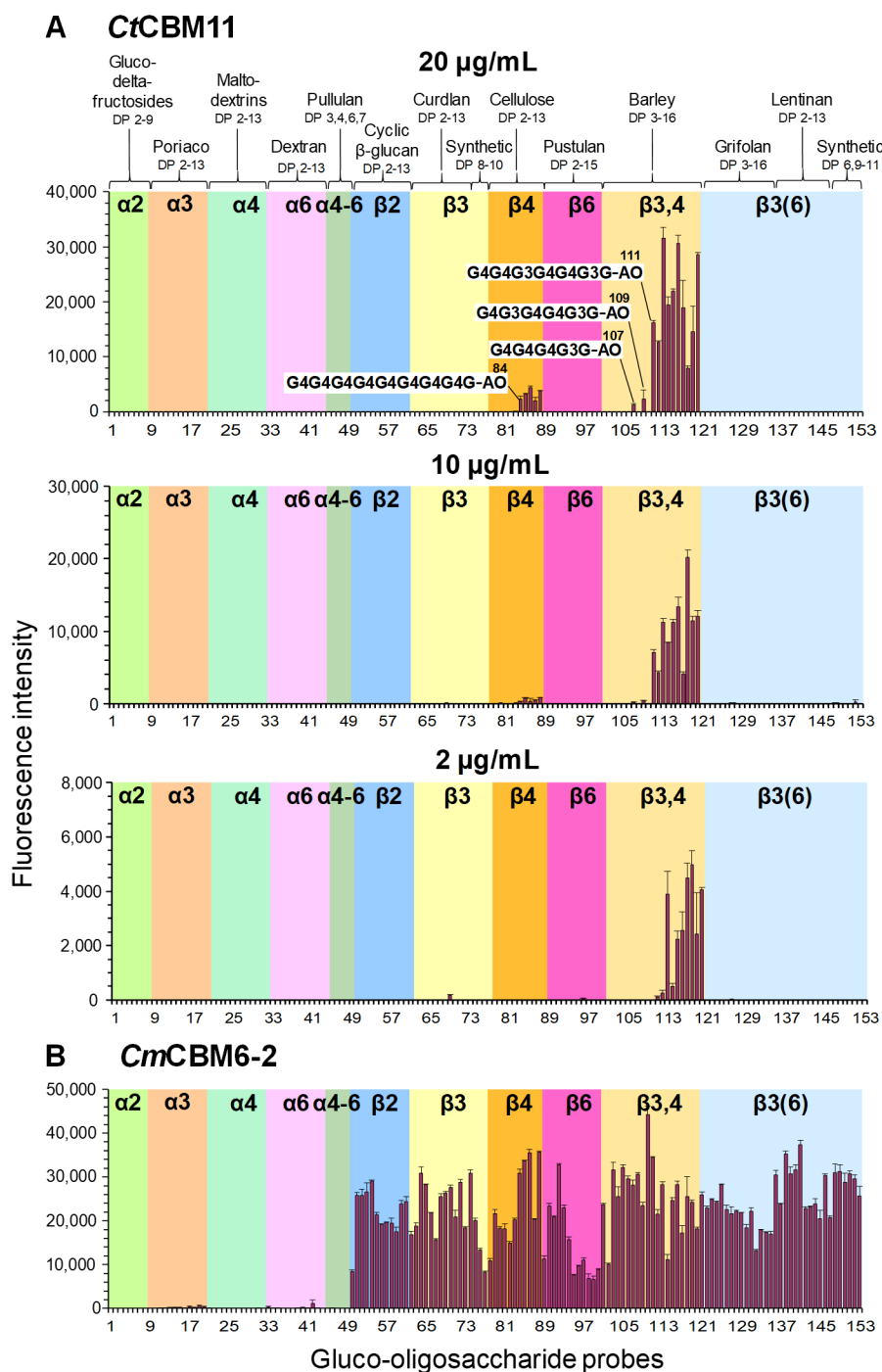
In the present work, and following the microarrays analysis data presented in Chapter 3, an integrated approach combining carbohydrate microarrays, NMR, X-ray crystallography, site-directed mutagenesis and ITC was conducted to extend the knowledge on the molecular determinants that enable CtCBM11 to distinguish between linear and mixed-linked  $\beta$ -glucans. The results now reported demonstrate the preference of CtCBM11 for mixed-linked glucans via a conformation-selection mechanism, in which CH- $\pi$  stacking and hydrogen bonding interactions contribute for the specific ligand chain conformation and orientation in the binding clef. The optimal conformation is achieved by having a  $\beta$ 1,3-linkage at the reducing end of the saccharide, while the central units are linked by  $\beta$ 1,4 glycosidic linkages. Ultimately, the structural and affinity data confirmed the sequence G4G4G3G as the minimum binding epitope and evidenced that recognition by CtCBM11 is not only dependent on the ligand chain-length and the  $\beta$ 1,3-linked glucose in the reducing end, but also on its specific position between the  $\beta$ 1,4-linked glucose units.

## 4.2 Results and Discussion

Previous studies identified Tyr22, Tyr53, Asp129, Arg126, Asp128, Tyr129 and Asp146 as key residues in ligand recognition by CtCBM11<sup>33,191,192</sup> and suggested that the binding cleft contained four binding subsites (Figure 4.1), with a preference for a  $\beta$ 1,3-linked glucose residue in at least one of those subsites. In the present work, complementary studies were carried to characterize CtCBM11's selectivity for the  $\beta$ 1,3-linked glucose and to elucidate the molecular determinants of the specificity towards mixed-linked glucans. In the combined approach employed, NMR experiments were also performed. These were carried out by collaborators from the laboratory of (Bio)molecular Structure and Interactions by NMR (UCIBIO, FCT-NOVA). These studies are not presented here but can be consulted in the peer-reviewed publication from which this chapter is adapted.

### 4.2.1 Specificity assignment using carbohydrate microarrays

To resume carbohydrate binding specificity at oligosaccharide level, the CtCBM11 was first analysed using carbohydrate microarrays comprising diverse sequence-defined gluco-oligosaccharides prepared as NGL probes<sup>32</sup>. These microarrays were validated and applied for analysis of protein binding in Chapters 2 and 3. The gluco-oligosaccharides highlighted here (positions 1 to 153, Table S2.1 in Chapter 2) encompassed different chain lengths (from DP-2 up to DP-16) and linear or branched sequences with  $\alpha$ - or  $\beta$ -configurations (Figure 4.2).



**Figure 4.2. Analysis of carbohydrate binding specificity using a microarray of sequence-defined gluco-oligosaccharides. (A)** CtCBM11 was analysed using serial dilutions at the indicated concentrations; **(B)** CmCBM6-2 was analysed as a positive control. The validated microarray encompassed 153 gluco-oligosaccharide probes prepared as NGLs<sup>32</sup>. The DP and glucose linkages are indicated on top of the coloured panels. Some relevant carbohydrate probe sequences are highlighted for binding to CtCBM11 in panel A; G, Glucose; AO, NGLs prepared from reducing oligosaccharides by oxime ligation with an aminoxy (AO) functionalized lipid DHPE (1,2-dihexadecyl-sn-glycero-3-phosphoethanolamine)<sup>8632</sup>. The sequence information on the oligosaccharide probes is depicted in Chapter 2, Table S2.1. The binding signals are means of fluorescence intensities of duplicate spots at 5 fmol of oligosaccharide probe arrayed (with error bars) and are representative of at least two independent experiments.



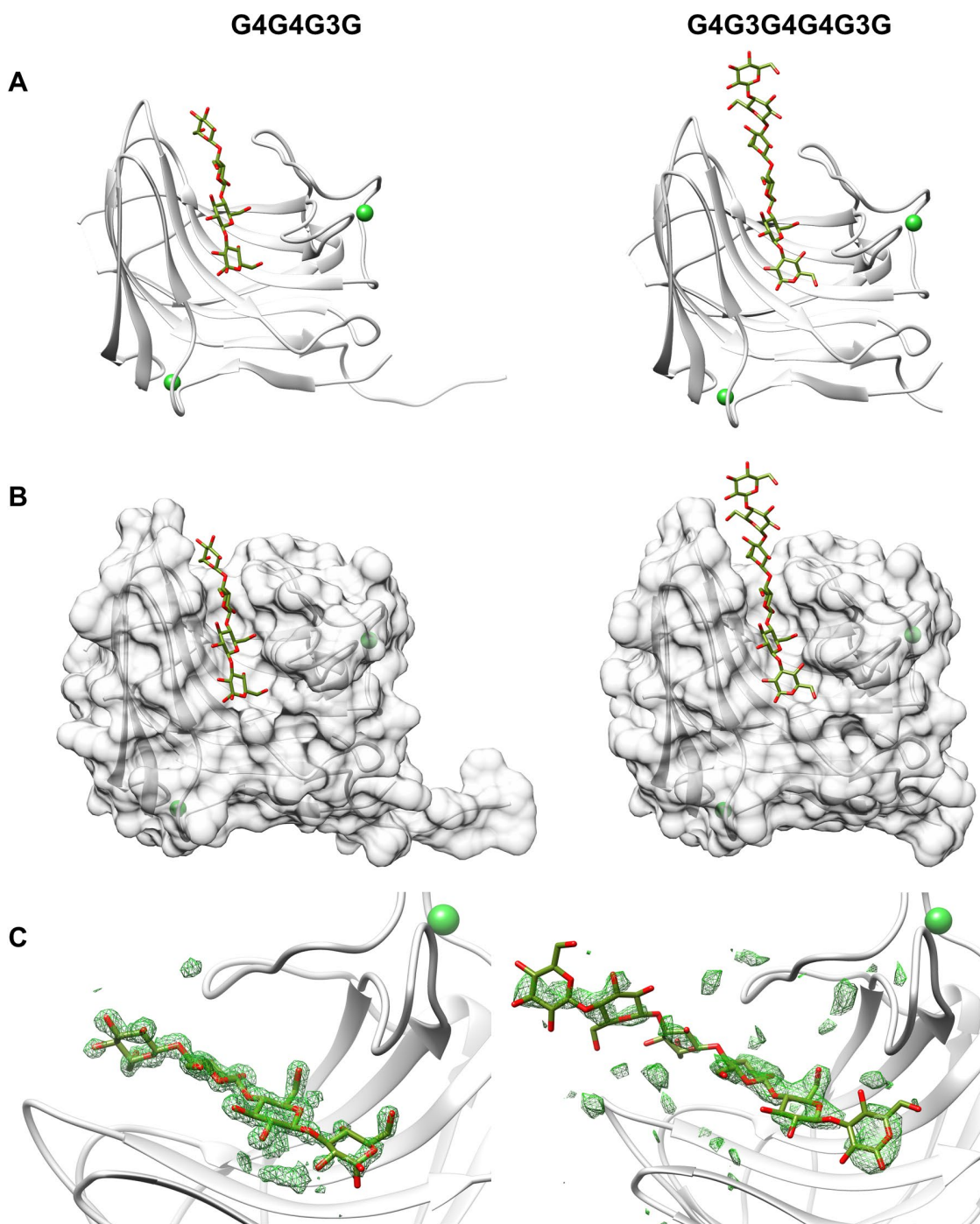
CtCBM11 showed a narrow binding profile, exhibiting strong binding to barley-derived  $\beta$ 1,3-1,4-mixed-linked oligosaccharides (DP-7 to DP-16, probes 111-120, Figure 4.2A) and displaying only a weak binding to  $\beta$ 1,4-linked cello-oligosaccharides (DP-9 to DP-13, probes 84-88). The binding of CtCBM11 contrasted with the broad  $\beta$ -glucan binding profile observed with CBM6-2 of *Cellvibrio mixtus* (CmCBM6-2) used as a control protein (Figure 4.2B), in accord with the reported specificity attributed to its two binding clefts<sup>32,39</sup>. The serial dilution of CtCBM11 concentration highlighted its specificity for  $\beta$ 1,3-1,4-mixed-linked glucose sequences (Figure 4.2A), in accord with previous carbohydrate microarray data<sup>32</sup>. The observed oligosaccharide chain-length dependency for CtCBM11 binding is in agreement with the current knowledge on type B CBMs. These CBMs bind the carbohydrate chains internally (endo-type), hence requiring a minimum chain-length for the recognition event to take place.

For immobilization on the array surface the oligosaccharides were conjugated via the reducing end glucose to an aminoxy-functionalised lipid by oxime-ligation (Figure 1.6C, Chapter 1)<sup>32</sup>. Although, the oxime-linked NGLs have a significant proportion of the lipid-linked monosaccharide core in a ring closed form, the conjugation and presentation in the microarray may have hindered access of the CBM to the binding epitope presented in short mixed-linked oligosaccharides with a 3-linkage at the reducing end. This would explain the lack of binding to the mixed-linked tetrasaccharide G4G4G3G (probe 103, Figure 4.2A), for which high affinity was previously reported<sup>33</sup>, and the weak binding observed to the pentasaccharide G4G4G4G3G (probe 107) and to the hexasaccharide G4G3G4G4G3G (probe 109). These results, together with the binding to the cellooligosaccharides, where binding was not detected to probes shorter than DP-9, hinted that both the sequence of  $\beta$ 1,4-linkages adjacent to a  $\beta$ 1,3-linked glucose and the chain-length are important for ligand recognition by this CBM. The higher binding intensities observed to the mixed-linked heptasaccharide G4G4G3G4G4G3G and longer chain probes (probes 111-120), where the sequence G4G4G3G is preserved for interaction, suggested this tetrasaccharide as the minimum epitope recognised by CtCBM11.

#### 4.2.2 Crystal structure of CtCBM11 bound to $\beta$ 1,3-1,4-gluco-oligosaccharides

To obtain atomic detail on the CtCBM11-ligand interactions that promote the recognition of mixed-linked  $\beta$ -glucans and the preference for the  $\beta$ 1,3-linked glucose, the crystal structures of CtCBM11 were determined in complex with mixed-linked oligosaccharides featuring a  $\beta$ 1,3-linkage at the reducing end (tetrasaccharide G4G4G3G) and both at the reducing end and at an internal position (hexasaccharide G4G3G4G4G3G). The linkages and sequence were determined by negative-ion ESI-CID-MS/MS sequencing<sup>32</sup> (Figure S4.1).

The bound structures of CtCBM11 were solved at a resolution of 1.45 Å and 2.6 Å for complexes with G4G4G3G (PDB ID 6R3M) and G4G3G4G4G3G (PDB ID 6R31), respectively (Figure 4.3). Statistics of data processing and model refinement and validation are presented in Tables 4.1 and S4.1. Both structures presented a classical distorted  $\beta$ -jelly roll fold already revealed for the



**Figure 4.3. Ribbon representation of the three-dimensional crystal structures of CtCBM11 complexes.** Representation of the overall structure of CtCBM11 in complex with the ligands, exhibiting the typical distorted  $\beta$ -barrel conformation. Left panel – CtCBM11-G4G4G3G (PDB ID 6R3M). Right panel - CtCBM11-G4G3G4G4G3G (PDB ID 6R31); **(A)** and **(B)** Cartoon and surface representation of the CtCBM11 complexes; The concave side of CtCBM11 forms the binding cleft where the ligands are accommodated; **(C)** Initial  $mF_o-DF_c$  electron density maps for the complexes of CtCBM11 with G4G4G3G and G4G3G4G4G3G calculated in the absence of ligand and with resolutions of 1.45 Å and 2.60 Å, respectively. The ligands are overlaid in the picture for reference. The electron density maps are shown in green mesh, contoured at 2.5  $\sigma$ ; Calcium ions are indicated as green spheres. Images generated using UCF Chimera<sup>40</sup>.

**Table 4.1. X-ray diffraction and structure refinement parameters and statistics for CtCBM11-G4G4G3G and CtCBM11-G4G3G4G4G3G structures.**

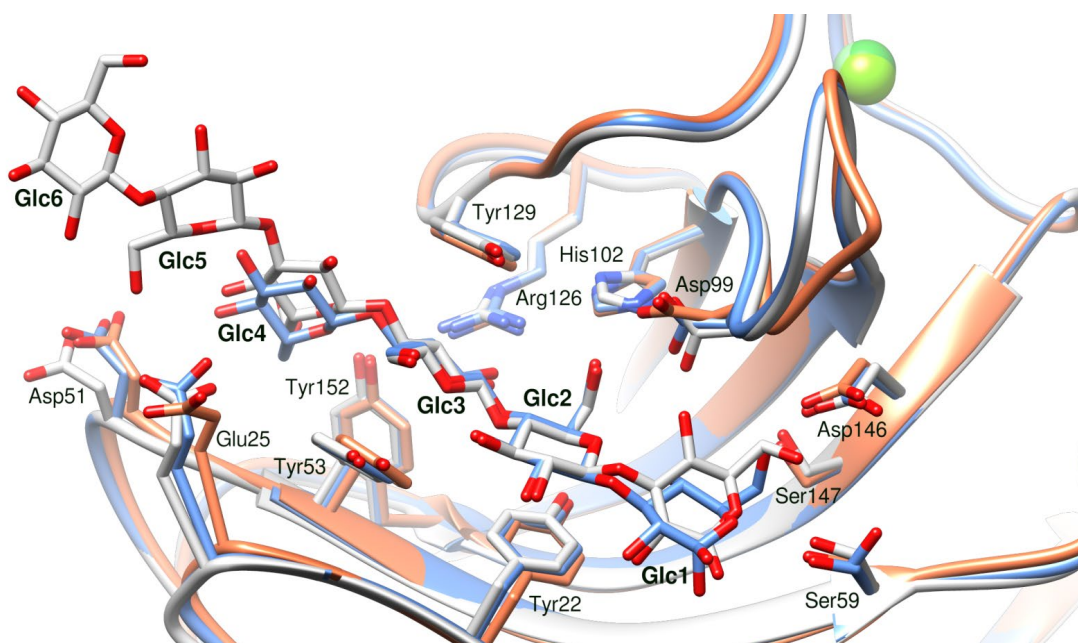
	CtCBM11-G4G4G3G	CtCBM11-G4G3G4G4G3G
<b>Data collection</b>		
Beamline	Diamond Light Source, I02	ESRF, ID23-2
Space Group	<i>H</i> 3	<i>H</i> 3
<b>Cell parameters</b>		
<i>a</i> , <i>b</i> (Å)	103.2	104.9
<i>c</i> (Å)	39.6	39.5
Wavelength, Å	0.9763	0.8729
Resolution of data (outer shell), Å	51.62-1.45 (1.48-1.45)	29.83-2.60 (2.72-2.60)
Total number of reflections (outer shell)	101660 (4271)	11893 (1320)
Number of unique reflections	27701 (1373)	4803 (590)
<i>R</i> <sub>pim</sub> (outer shell), % <sup>a</sup>	4.6 (22.0)	13.7 (32.0)
<i>R</i> <sub>merge</sub> (outer shell), % <sup>b</sup>	7.0 (30.0)	18.9 (40.8)
Mean <i>I</i> / $\sigma$ ( <i>I</i> ) (outer shell)	10.0 (3.5)	3.7 (2.0)
CC(1/2)	0.996 (0.637)	0.927 (0.697)
Completeness (outer shell), %	99.8 (97.2)	96.2 (96.9)
Redundancy (outer shell)	3.7 (3.1)	2.5 (2.5)
<b>Structure refinement</b>		
No. of protein atoms	1414	1451
No. of solvent waters	212	57
Resolution used in refinement, Å	51.62-1.45	29.83-2.60
No. of reflections	26308	4290
<i>R</i> <sub>work</sub> / <i>R</i> <sub>free</sub> <sup>c</sup>	0.177 / 0.208	0.190 / 0.254
rms deviation bonds (Å)	0.013	0.010
rms deviation angles (degrees)	1.662	1.624
rms deviation chiral volume (Å <sup>3</sup> )	0.133	0.083
<b>Avg B factors (Å<sup>2</sup>)</b>		
Main chain	6.5	18.8
Side chain	10.1	18.9
Calcium 1	7.9	27.5
Calcium 2	9.5	31.4
	<b>G4G4G3G</b>	<b>G4G3G4G4G3G</b>
<i>Glucose 1</i>	17.8	22.0
<i>Glucose 2</i>	10.0	21.1
<i>Glucose 3</i>	11.0	21.1
<i>Glucose 4</i>	21.3	22.2
<i>Glucose 5</i>	-	23.3
<i>Glucose 6</i>	-	23.6
Phosphate ion 1	9.4	30.0
Phosphate ion 2	16.1	38.6
Phosphate ion 3	17.5	-
Phosphate ion 4	21.7	-
Acetate ion 1	15.6	-
Water molecules	23.5	27.3
<b>Ramachandran statistics</b>		
<i>favored</i>	99.4	98.2
<i>allowed</i>	0.6	1.8
<i>generously allowed</i>	0	0
<i>forbidden</i>	0	0
<b>PDB deposition code</b>		
	<b>6R3M</b>	<b>6R31</b>

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

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

<sup>c</sup>  $R_{work} = \left( \frac{\sum_{hkl} |F_{hkl}^{obs} - F_{hkl}^{calc}|}{\sum_{hkl} F_{hkl}^{obs}} \right) \times 100$ , 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.

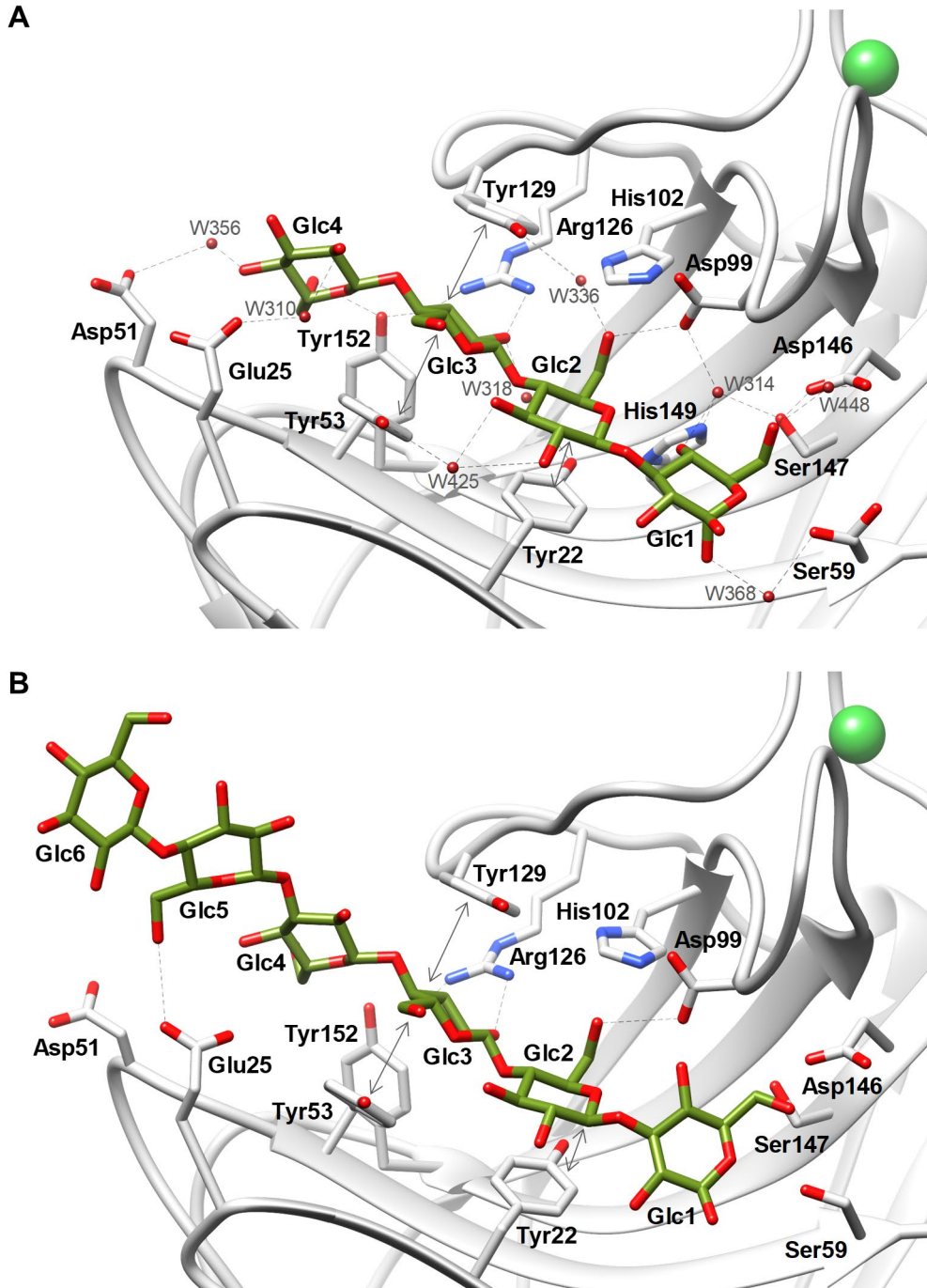
unbound CtCBM11<sup>33</sup>, consisting of two six-stranded anti-parallel  $\beta$ -sheets, which form a convex side, and a concave side that constitutes the binding cleft where each ligand is accommodated (Figure 4.3A and B). In the CtCBM11-G4G4G3G structure, the anomeric carbon of Glc1 was observed in both  $\alpha$  and  $\beta$  configuration, as supported by residual  $mF_o - DF_c$  electron density map and was modelled in both anomers (Figure 4.3C). The overall fold of the two bound structures was similar to the fold of native CtCBM11 (PDB ID 1V0A)<sup>33</sup>, with a root-mean-square deviation (rmsd) value of 0.435 Å over 141 C $\alpha$  atoms, for the tetrasaccharide-bound structure, and 0.509 Å over 150 C $\alpha$  atoms, for the hexasaccharide-bound structure (Figure 4.4). The high similarity between the free and bound conformations of CtCBM11, is in good agreement with the relaxation and internal mobility data obtained previously by NMR<sup>191</sup> that showed only minor dynamical variations upon cello-tetrasaccharide binding. This is consistent with a rigid protein backbone that selects a defined oligosaccharide conformation, i.e., CtCBM11 recognizes its ligands by a conformation-selection mechanism.



**Figure 4.4. Comparison of unbound and ligand-bound CtCBM11 structures.** Superposition of CtCBM11 unbound structure (PDB ID 1V0A) (orange) with the bound structures of CtCBM11-G4G4G3G (PDB ID 6R3M) (grey) and CtCBM11-G4G3G4G4G3G (PDB ID 6R31) (blue), with a root-mean-square deviation (rmsd) value of 0.435 and 0.509, respectively. Images generated using UCF Chimera<sup>40</sup>.

### 4.2.3 CtCBM11 binding mode

The identified residues that constitute the binding cleft of CtCBM11 are solvent-exposed and interact with the ligands through hydrophobic CH- $\pi$  stacking interactions, hydrogen bonds and van der Waals contacts (Figure 4.5, and Tables S4.2 and S4.3). The ligand G4G4G3G interact with the CBM by direct hydrogen bonds of the equatorial OH groups of all the 4 glucose monomers with residues Tyr152, Arg126, Asp99 and Asp146, as well as water-mediated contacts with residues Asp51, Glu25, Tyr22, Tyr53, Tyr129, His149, Ser147 and Ser59 (Figure 4.5A and Table



**Figure 4.5. CtCBM11-ligand interactions.** Close-up view on the CtCBM11 binding site, evidencing the protein-ligand contacts between the CBM and (A) the tetrasaccharide G4G4G3G and (B) the hexasaccharide G4G3G4G4G3G, as listed in Tables S4.2 and S4.3. The carbohydrate chains and the side chains of the amino acid residues inside the binding cleft that interact with the ligands are shown as stick models. Water molecules are represented by red spheres and calcium ion as green sphere. Hydrogen bonding is indicated by dashed lines and CH- $\pi$  stacking interactions are represented as double arrows. Images generated using UCF Chimera<sup>40</sup>.

S4.2). For the hexasaccharide G4G3G4G4G3G ligand, the same direct hydrogen bonds were observed, although, due to the lower resolution of the hexasaccharide complex, no water-mediated hydrogen bonds were identified as the water molecules could not be

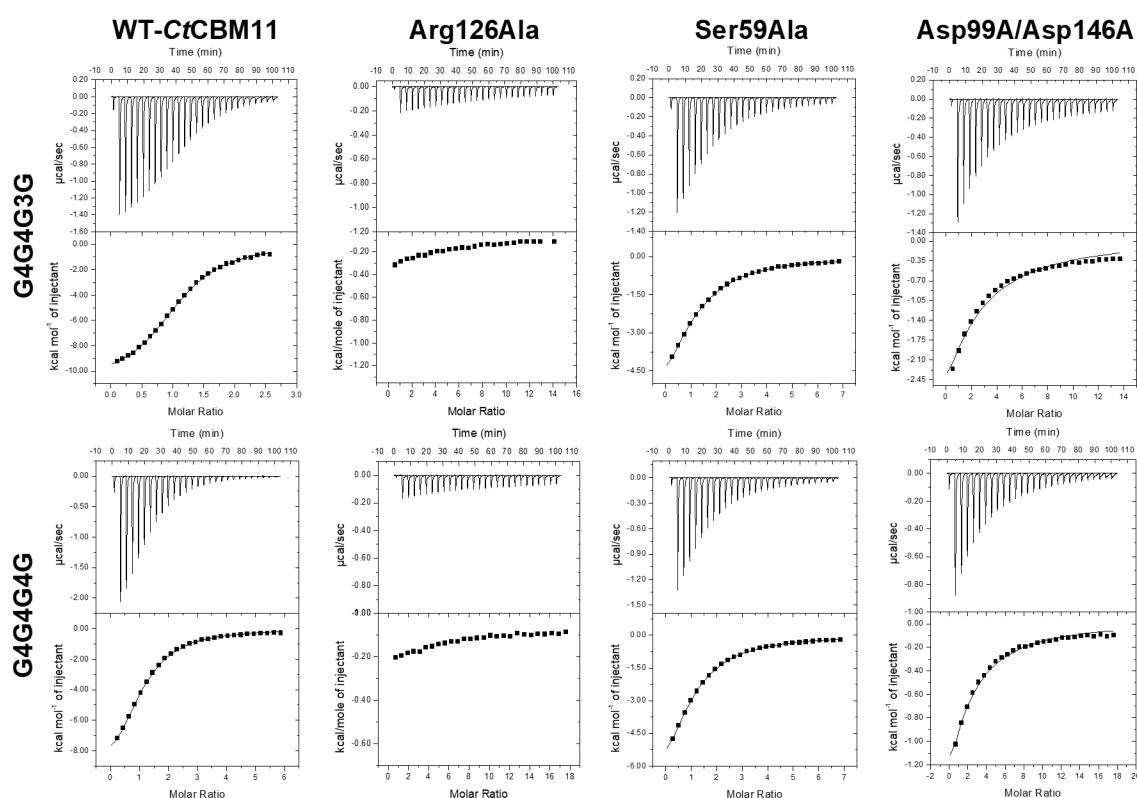
unequivocally modelled (Figure 4.5B and Table S4.3). The CH- $\pi$  stacking interactions between residues Tyr22, Tyr53 and Tyr129 with the glucose rings at the centre of the cleft (Glc2 and Glc3) were evident, which validated our previous models using computational studies (molecular docking and molecular dynamics)<sup>192</sup> and confirmed these residues to play a key role by guiding and stabilizing the ligand chain for recognition by CtCBM11.

The structures of the bound CtCBM11 provided clear evidence for a conformation-selection mechanism. While the central  $\beta$ 1,4-linked glucose units appear to be pivotal for CtCBM11 recognition through CH- $\pi$  stacking with the tyrosine residues, the flanking  $\beta$ 1,3-linked glucose at the reducing end (Glc1) seems to impose a specific ligand chain conformation and, consequently, its orientation in the binding cleft. This is probably due to the hydrogen bond between Asp146 and the OH of the Glc1 methylene group. If, at this position, a  $\beta$ 1,4 glycosidic bond was present instead of the  $\beta$ 1,3 glycosidic bond (as in the case of cellotetrasaccharide), the glucose ring would be in a different orientation, with the CH<sub>2</sub>OH group rotated by about 180° (Figure S4.2). Although in this conformation a hydrogen bond is still possible between the OH group of carbon 2 and Asp146, the OH group would sit further away from Asp146 and the hydrogen bond would be weaker, thus explaining the lower affinity to  $\beta$ 1,4-linked oligosaccharides. This is in very good agreement with the specificity observed in carbohydrate microarrays (Figure 4.2) as well as with the STD-NMR data (please see Ribeiro *et al.*, 2019<sup>34</sup>) that showed that for the cellotetrasaccharide the most affected proton of Glc1 was H2, whereas for the mixed-linked tetrasaccharide G4G4G3G the methylene protons were the ones showing more saturation.

The superposition of the bound structures highlighted that the Glc2 and Glc3 stacked by the tyrosine residues were almost completely coincident (Figure 4.4), which provides further evidence for the importance of the positioning of these two  $\beta$ 1,4-linked monosaccharides at the central subsites 2 and 3 (Figure 4.1). Comparing the two bound structures, the Glc5 and Glc6 of the hexasaccharide were mostly exposed to the solvent, not establishing significant contacts with the protein residues (Figure 4.4), other than the direct hydrogen bond between Glu25 and the CH<sub>2</sub>OH group of the  $\beta$ 1,3-linked Glc5 (Figure 4.5B). This observation provides evidence for the major contribution of subsites 1-3 in CtCBM11 binding and confirms the sequence G4G4G3G as a minimum binding epitope, whereas a second  $\beta$ 1,3-linked glucose (putative subsite 5) may affect affinity or ligand specificity. Superimposing also the unbound structure (PDB ID 1V0A)<sup>33</sup> (which exhibited the *C-terminus* residues of a symmetry-related molecule in the binding cleft), showed that the residues previously identified in the binding cleft to interact with the *C-terminus* tail were coincident with the ones now identified to be responsible for the ligand stabilization (Figure 4.4). The majority of these residues suffered only minimal changes in the bound CtCBM11 structures, in accordance with a conformation-selection model mechanism.

#### 4.2.4 The CH- $\pi$ stacking and hydrogen bonding network as determinants of the ligand-specificity

The structural data allowed not only the identification of key residues involved in CtCBM11 binding, but also structural features of the oligosaccharide ligands that were able to modulate binding. With this structure-based rationale, mutant alanine derivatives of residues involved in direct hydrogen bonds with the ligand (Ser59, Asp99, Arg126, Asp146) were produced to analyse influence of hydrogen bonding on CtCBM11 binding affinity towards different carbohydrates (polysaccharides and oligosaccharides), for which chain-length as well as the presence and position of  $\beta$ 1,3 linkages varied (Figure 4.6 and Table 4.2).



**Figure 4.6. Representative isothermal calorimetry titrations of binding of CtCBM11 and its mutants to oligosaccharides.** The top portion of each panel shows the raw power data while the bottom parts show the integrated and heat of dilution corrected data. The solid lines show the non-linear curve fits to a one site binding model with the stoichiometry fixed at 1. Thermodynamic parameters are given in Table 4.2.

In agreement with the structural data, the comparison of the binding affinity of CtCBM11 wild-type (WT-CtCBM11) to the three tetrasaccharides analysed (G4G4G4G, G4G3G4G and G4G4G3G) showed that the marked affinity effect occurred when introducing the  $\beta$ 1,3 glycosidic bond at the central part of the ligand (i.e., for ligand G4G3G4G). When compared with the cellotetrasaccharide (G4G4G4G), this modification caused a decrease in the affinity of about 4.4-fold. Inversely, when placing the  $\beta$ 1,3 bond at subsite 1 (G4G4G3G) the increase in the affinity was about 2.4-fold, supporting the preference CtCBM11 for a  $\beta$ 1,3 linkage at the reducing end.

**Table 4.2. Thermodynamic parameters of the binding of CtCBM11 and its mutant derivatives to polysaccharides and oligosaccharides.**

CtCBM11 variant	Ligand	$K_a$ ( $M^{-1}$ )	$\Delta G$ ( $kcal.mole^{-1}$ )	$\Delta H$ ( $kcal.mole^{-1}$ )	$TAS$ ( $kcal.mole^{-1}$ )	$n$
WT	$\beta$ -Glucan	$4.94 (\pm 0.23) \times 10^5$	-7.77	$-11.23 \pm 0.09$	-3.46	$1.02 \pm 0.00$
	Lichenan	$3.08 (\pm 0.36) \times 10^5$	-7.49	$-8.41 \pm 0.23$	-0.92	$1.04 \pm 0.02$
	G4G4G4G4G4G	$1.11 (\pm 0.04) \times 10^5$	-6.88	$-12.30 \pm 0.15$	-5.43	$1.11 \pm 0.01$
	G4G4G4G	$5.86 (\pm 0.16) \times 10^4$	-6.49	$-11.18 \pm 0.18$	-4.68	$1.08 \pm 0.01$
	G4G4G3G	$1.41 (\pm 0.03) \times 10^5$	-7.01	$-10.95 \pm 0.07$	-3.94	$1.11 \pm 0.00$
	G4G3G4G	$1.32 (\pm 0.09) \times 10^4$	-5.61	$-12.62 \pm 0.04$	-7.01	$1.00 \pm 0.00$
	HEC	$4.45 (\pm 0.07) \times 10^3$	-4.98	$-6.20 \pm 0.06$	-1.23	$1.00 \pm 0.00$
Asp99Ala	$\beta$ -Glucan	$3.06 (\pm 0.20) \times 10^4$	-6.11	$-10.97 \pm 0.49$	-4.86	$1.11 \pm 0.04$
	G4G4G4G4G4G	$1.19 (\pm 0.09) \times 10^4$	-5.56	$-7.76 \pm 0.30$	-2.19	$1.00 \pm 0.00$
	G4G4G4G	$2.19 (\pm 0.13) \times 10^3$	-4.55	$-8.25 \pm 0.27$	-3.69	$1.00 \pm 0.00$
	G4G4G3G	$3.88 (\pm 0.45) \times 10^4$	-6.26	$-5.18 \pm 0.22$	-1.09	$1.00 \pm 0.06$
	G4G3G4G	$2.67 (\pm 0.05) \times 10^3$	-4.68	$-9.16 \pm 0.11$	-4.47	$1.00 \pm 0.00$
Arg126Ala	$\beta$ -Glucan	$8.46 (\pm 0.10) \times 10^3$	-5.35	$-8.48 \pm 0.06$	-3.13	$1.00 \pm 0.00$
	G4G4G4G4G4G			Weak binding		
	G4G4G4G			No binding		
	G4G4G3G			No binding		
	G4G3G4G			No binding		
Asp146Ala	$\beta$ -Glucan	$3.31 (\pm 0.23) \times 10^4$	-6.15	$-10.89 \pm 0.57$	-4.74	$1.10 \pm 0.04$
	G4G4G4G4G4G	$5.08 (\pm 0.20) \times 10^4$	-6.41	$-9.12 \pm 0.27$	-2.71	$0.93 \pm 0.02$
	G4G4G4G	$1.36 (\pm 0.12) \times 10^4$	-5.64	$-9.99 \pm 0.41$	-4.35	$1.00 \pm 0.00$
	G4G4G3G	$2.63 (\pm 0.08) \times 10^4$	-6.04	$-10.54 \pm 0.52$	-4.50	$0.95 \pm 0.04$
	G4G3G4G	$2.53 (\pm 0.14) \times 10^3$	-4.59	$-6.81 \pm 0.24$	-2.21	$1.00 \pm 0.00$
Asp99Ala/ Asp146Ala	$\beta$ -Glucan	$3.09 (\pm 0.09) \times 10^4$	-6.12	$-6.77 \pm 0.17$	-0.66	$1.14 \pm 0.02$
	G4G4G4G4G4G	$1.14 (\pm 0.02) \times 10^4$	-5.53	$-7.59 \pm 0.40$	-2.06	$0.90 \pm 0.04$
	G4G4G4G	$3.56 (\pm 0.27) \times 10^3$	-4.86	$-10.23 \pm 0.46$	-5.37	$1.00 \pm 0.00$
	G4G4G3G	$6.22 (\pm 0.32) \times 10^3$	-5.17	$-14.98 \pm 0.04$	-9.81	$1.00 \pm 0.00$
	G4G3G4G			No binding		
Val57Ala	$\beta$ -Glucan	$3.12 (\pm 0.24) \times 10^5$	-7.49	$-10.15 \pm 0.16$	-2.65	$1.13 \pm 0.01$
	G4G4G4G4G4G	$1.01 (\pm 0.02) \times 10^5$	-6.82	$-10.04 \pm 0.09$	-3.22	$0.94 \pm 0.00$
	G4G4G4G	$5.62 (\pm 0.16) \times 10^4$	-6.48	$-10.06 \pm 0.17$	-3.58	$1.12 \pm 0.01$
	G4G4G3G	$1.12 (\pm 0.03) \times 10^5$	-6.89	$-10.03 \pm 0.10$	-3.13	$0.99 \pm 0.00$
	G4G3G4G	$1.33 (\pm 0.05) \times 10^4$	-5.62	$-8.66 \pm 0.14$	-3.04	$1.10 \pm 0.00$
Ser59Ala	$\beta$ -Glucan	$4.48 (\pm 0.20) \times 10^4$	-6.35	$-7.43 \pm 0.25$	-1.09	$1.06 \pm 0.02$
	G4G4G4G4G4G	$5.09 (\pm 0.28) \times 10^4$	-6.43	$-9.71 \pm 0.46$	-3.28	$0.91 \pm 0.02$
	G4G4G4G	$2.74 (\pm 0.11) \times 10^4$	-5.99	$-12.55 \pm 0.24$	-6.56	$1.00 \pm 0.01$
	G4G4G3G	$3.08 (\pm 0.07) \times 10^4$	-6.12	$-8.66 \pm 0.22$	-2.54	$1.08 \pm 0.02$
	G4G3G4G			Weak binding		
Glu25Ala	$\beta$ -Glucan	$2.75 (\pm 0.50) \times 10^5$	-7.32	$-8.18 \pm 0.36$	-0.86	$1.00 \pm 0.03$
	HEC			Weak binding		
Asp51Ala	$\beta$ -Glucan	$2.31 (\pm 0.36) \times 10^5$	-7.43	$-7.35 \pm 0.34$	0.08	$1.01 \pm 0.03$
	HEC			Weak binding		
Ser59Ala/ Asp146Ala	$\beta$ -Glucan			Very weak binding		
	HEC			Very weak binding		

The CtCBM11 Arg126Ala mutant bound to  $\beta$ -glucans with a 100-fold lower affinity than wild type CtCBM11. This corroborated what was observed in the CtCBM11-G4G4G3G structure, where atoms N $\eta$ 1 and N $\eta$ 2 of Arg126 are hydrogen bonded to the O3 and O2 atoms, respectively, of the glucose residue located at subsite 3. Thus, the two hydrogen bonding contacts of Arg126, together with the CH- $\pi$  stacking with the tyrosines, are fundamental for holding the ligand. The affinity of the Asp99Ala and Asp146Ala mutants for  $\beta$ -glucan and the oligosaccharides tested was reduced by approximately 4 to 10-fold. While Asp146 is hydrogen bonded to the OH group of the methylene group from the glucose residue at subsite 1, Asp99 established polar contacts with O6



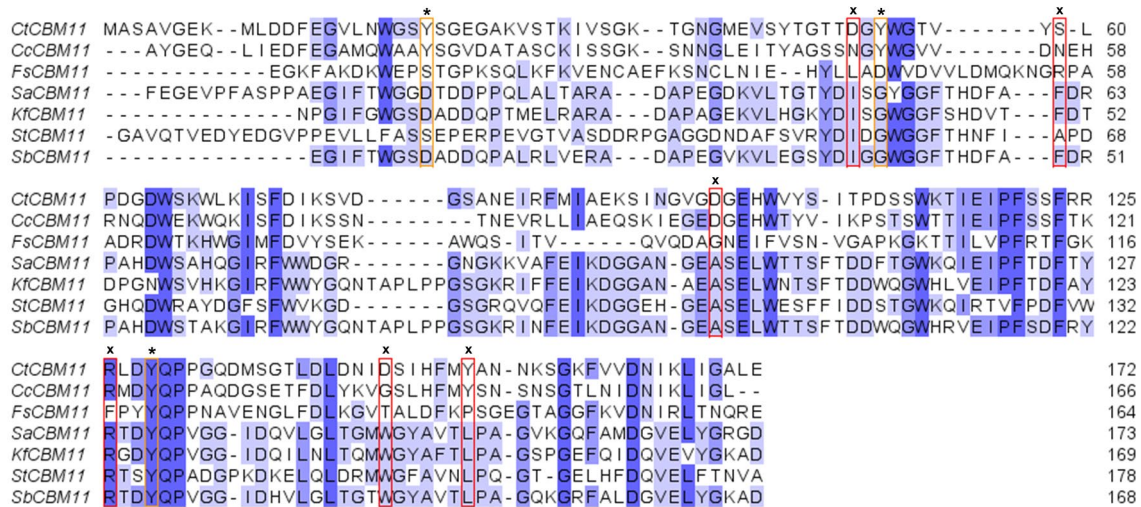
of the glucose residue located at subsites 2 and O4 of the glucose residue located at subsite 1 and 2 (Figure 4.5). The cumulative effect of the double mutation Asp99Ala/Asp146Ala resulted in a similar trend, although leading to an overall lower affinity. This suggested that these residues are equally important for binding both  $\beta$ 1,3-1,4-mixed-linked and  $\beta$ 1,4-linked glucans. Furthermore, the hydrogen bond interactions of Asp146 may also contribute to the higher affinity observed towards G4G4G3G. The  $\beta$ 1,3 glycosidic linkage brings the CH<sub>2</sub>OH group of Glc1 in closer proximity to the sidechain of Asp146 when compared with a  $\beta$ 1,4 bond in the same position as observed in the structure of the complex, making a stronger hydrogen bond. As such, a  $\beta$ 1,3 glycosidic linkage towards the reducing end of the oligosaccharide is preferred. Replacement of Ser59 by an Ala led also to a significant loss in binding affinity, which corroborated the disruption of an important hydrogen bond established with the endocyclic oxygen of Glc1. The cumulative effect of Ser59Ala/Asp146Ala results in an almost complete loss of binding to both  $\beta$ -glucan and HEC, highlighting the importance of subsite 1 for substrate recognition. The CtCBM11 Val57Ala mutation, which was produced to assess the influence of hydrophobic interactions at subsite 1, showed no significant effect in the binding affinity to the mixed-linked ligands. This result highlights that the major contributions of subsite 1 for CtCBM11 binding are mediated by hydrogen bonding interactions.

The hexasaccharide complex showed Glu25 making a hydrogen bond to the CH<sub>2</sub>OH group of the  $\beta$ 1,3-linked Glc5 (Figure 4.5B). As a  $\beta$ 1,4-linked Glc5 would have its CH<sub>2</sub>OH group facing away from Glu25 and Asp51, these two residues could play an important role in CtCBM11 preference for  $\beta$ 1,3-1,4-mixed-linked over  $\beta$ 1,4-linked glucans. Glu25Ala and Asp51Ala mutants were produced to test this hypothesis. Although there was a slight decrease in the ability of both mutants to bind  $\beta$ -glucan, the affinity for HEC was also affected. This means that although important for binding to the ligand, this putative subsite 5 does not seem to be key for substrate specificity.

In summary, the structural and affinity data demonstrate the contribution of CH- $\pi$  stacking and hydrogen bonding interactions for specific ligand chain conformation and orientation in the binding cleft are determinant for the specificity of CtCBM11 towards mixed-linked  $\beta$ -glucans. The conformational change in the orientation of the glucose residues by the introduction of a  $\beta$ 1,3 glycosidic bond, leads to key hydrogen bonds with Asp 146 and Ser 59 (subsite 1) and Asp99 (subsite 1 and 2), which have a direct impact on CtCBM11 specificity and on the affinity displayed towards the different ligands. The data also show evidence that the central part of the oligosaccharide (the residues that bind at subsites 2 and 3) must be planar ( $\beta$ 1,4-linked), in order to take full advantage of the CH- $\pi$  stacking interactions with tyrosine residues 22, 53 and 129, and hydrogen bonding with Arg126 (subsite 3). The hydrogen bonding mediated by Glu25 and Asp51 (subsite 4 and 5) contribute to increase the affinity to the ligands, but not to the specificity towards the mixed-linked glucans.

#### 4.2.5 CtCBM11 ligand specificity in the context of CAZy CBMs

The analysis of conservation of the interacting protein residues by sequence alignment of six family 11 CAZy CBMs, revealed that only Tyr129 was invariant, whereas Arg126 was conserved in five out of the six sequences (Figure 4.7), which highlights the critical role of these residues in the ligand recognition by CAZY family 11 CBMs. In its turn, Tyr22, Tyr53, Tyr152 and Asp99 were conserved in only two of the six CBMs. However, the lack of conservation of other key residues involved in the ligand recognition by CtCBM11 is not totally unexpected as plasticity of specificities is often observed within type B CBM families.



**Figure 4.7. Alignment of CBM11 family members.** Primary sequences aligned from *Clostridium thermocellum* (CtCBM11, P16218), *Clostridium cellulolyticum* (CcCBM11, P25472), *Fibrobacter succinogenes* (FsCBM11, C9RQE4), *Streptomyces avermitilis* (SaCBM11, Q82JP6), *Kribbella flavida* (KfCBM11, D2PWV9), *Salinispora tropica* (StCBM11, A4X7P1) and *Streptomyces bingchenggensis* (SbCBM11, D7BY98). Identity to CtCBM11 is indicated with blue boxes. Residue numbers refer to the corresponding CBM11 sequence. The (\*) identifies the CtCBM11 residues involved in the CH- $\pi$  stacking of the oligosaccharide ligands and the (x) identifies the residues that establish hydrogen bonds with the ligand. The sequence alignment was calculated with the program Clustal Omega<sup>196</sup>, and the picture was produced with the program Jalview<sup>197</sup>.

In general, the ligand specificity of type B CBMs reflects the substrate specificity of the associated catalytic modules. CtCBM11 is comprised in the *celH* gene, which also encodes two functional catalytic domains, a GH from family 5 (GH5, Cel5E) and a second from family 26 (GH26, Lic26A). While Cel5E is a bifunctional  $\beta$ -1,4-endoglucanase/xylanase<sup>198</sup>, Lic26A has lichenase activity, specific for  $\beta$ 1,3-1,4 mixed-linked glucans, accommodating in its binding cleft substrates that comprise the G4G4G3G sequence<sup>199</sup>. The observed preference of CtCBM11 for a  $\beta$ 1,4 glycosidic bond in the central part of the ligand and  $\beta$ 1,3-linked glycosidic bond at a reducing end provides evidence that this CBM mimics the specificity of the associated GH26 mixed-linked endoglucanase.

CBMs that bind  $\beta$ -glucan chains often display broad specificity recognizing  $\beta$ 1,4-glucans, mixed-linked  $\beta$ 1,3-1,4-glucans and xyloglucan, by targeting the  $\beta$ 1,4-glucan backbone common to these polysaccharides. According to the information deposited in the CAZY database, besides

family 11, CBMs from families 4, 6, 22, 28<sup>22</sup>, 46<sup>200</sup> and 65<sup>201</sup> have been reported to bind mixed-linked glucans. However, to our knowledge, only CtCBM11 has been described to have a more restricted binding specificity and affinity to mixed-linked glucans. Noteworthy, CtCBM11 is the only CBM from family 11 found in *C. thermocellum*, which might point to a crucial role played by CtCBM11 in the metabolism of mixed-linked glucans of this cellulosome-expressing bacterium.

### 4.3 Conclusions

In this work, a combined approach of methodologies was used to unravel, at a molecular level, the ligand recognition of CtCBM11. The analysis of the interaction by carbohydrate microarrays and NMR and the crystal structures of CtCBM11 bound to  $\beta$ 1,3-1,4-linked glucose oligosaccharides, showed that both the chain-length and the position of the  $\beta$ 1,3-linkage are important for recognition, and identified the tetrasaccharide Glc $\beta$ 1,4Glc $\beta$ 1,4Glc $\beta$ 1,3Glc sequence as a minimum epitope required for binding. The structural data, along with site-directed mutagenesis and ITC studies, demonstrated the specificity of CtCBM11 for the twisted conformation of mixed-linked  $\beta$ 1,3-1,4-glucans. This is mediated by a conformation-selection mechanism of the ligand in the binding cleft through CH- $\pi$  stacking and a hydrogen bonding network, which is dependent not only on ligand chain length, but also on the presence of a  $\beta$ 1,3-linkage at the reducing end and at specific positions along the  $\beta$ 1,4-linked glucan chain.

In the context of the cellulosome, the structural details here revealed on the CtCBM11 ligand-recognition site may influence the planning and development of efficient and low-cost mechanisms for the conversion of biomass into usable sources of energy, as well as, into nutrients for animal feedstock. Additionally, the understanding, at the molecular level, of the detailed mechanism by which CtCBM11 can distinguish between linear and mixed-linked  $\beta$ -glucans, may inspire the design of new biomolecules with improved capabilities to be explored in health and agriculture applications.

### 4.4 Experimental procedure

#### 4.4.1 Gene cloning, mutagenesis and protein purification

Plasmid pAG1, a pET21a (Novagen, Darmstadt, Germany) derivative encoding CtCBM11, was selected for these experiments<sup>33</sup>. Recombinant CtCBM11 generated by pAG1 contains a C-terminal hexa-histidine tag. Site-directed mutants were generated using the NZYMutagenesis kit (NZYTech, Lisbon, Portugal) according to the manufacturer's instructions using pAG1 as template. Primers used to generate the mutant DNA sequences are listed in Table S4.4. Recombinant sequences of all mutant plasmid derivatives were verified by sequencing to ensure that only the appropriate mutations were incorporated into the nucleic acids.

To express CtCBM11 in *Escherichia coli*, the CtCBM11 encoding gene was constructed as described previously<sup>33</sup>. *E. coli* BL21 harbouring the CtCBM11 encoding gene containing a

C-terminal His<sub>6</sub> tag was cultured in LB containing 100 µg/mL ampicillin at 37 °C until mid-exponential phase (OD<sub>600nm</sub> = 0.6), at which point isopropyl-β-D-thiogalactopyranoside (IPTG) was added to a final concentration of 1 mM. Cultures were then further incubated overnight at 30 °C. Cells were collected by centrifugation and the cell pellet resuspended in a 50 mM sodium HEPES buffer, pH 7.5, containing 1 M NaCl and 10 mM imidazole. CtCBM11 was purified by Ni<sup>2+</sup>-immobilized ion metal affinity chromatography (IMAC). Fractions containing the purified complex were buffer-exchanged into Milli-Q water containing 2 mM CaCl<sub>2</sub> and concentrated with Amicon 10-kDa molecular-mass centrifugal membranes to a final protein concentration of 40 mg/mL.

#### 4.4.2 Sources of carbohydrates

The soluble barley β-glucan, the celooligosaccharides and the β1,3-1,4-mixed-linked tetrasaccharides were purchased from Megazyme international (Bray, Ireland). The hydroxyethyl cellulose (HEC) and lichenan were purchased from Sigma-Aldrich (St. Louis, MO, USA). For the NMR studies, the cellotetrasaccharide was obtained from Seikagaku Corporation (Tokyo, Japan). The barley hexasaccharide fraction used for X-ray crystallography was obtained as described<sup>32</sup> by enzymatic hydrolysis of barley β-glucan with a cellulase (Novozymes, Copenhagen, Denmark) and purified by repeated gel filtration chromatography on a Bio-Gel P4 column.

#### 4.4.3 Mass spectrometry analysis of barley hexasaccharide

Sequence analysis of β1,3-1,4-mixed-linked tetrasaccharides (G4G4G3G and G4G3G4G) and of the barley-derived hexasaccharide fraction used in the co-crystallization studies was carried out by negative-ion electrospray tandem mass spectrometry with collision induced dissociation (ESI-CID-MS/MS) on a Synapt G2-S instrument (Waters, Manchester, U.K.), essentially as described<sup>32</sup>. Cone voltage was kept at 80 eV for MS and CID-MS/MS. For pseudo-MS<sup>3</sup> to encourage in-source fragmentation, the cone voltage was increased to 180 eV. Collision gas (Ar) at a pressure of 7.3 x 10<sup>-3</sup> mbar. The collision energy was between 15-17 eV for optimal fragmentation. The ESI-CID-MS/MS confirmed the sequences of the tetrasaccharides as reported previously<sup>32</sup>, and showed that the barley-derived hexasaccharide fraction contains mainly the sequence Glcβ1,4Glcβ1,3Glcβ1,4Glcβ1,4Glcβ1,3Glc (G4G3G4G4G3G) (Figure S4.1).

#### 4.4.4 Carbohydrate microarray analysis

The binding specificity of CtCBM11 was analysed using carbohydrate microarrays that included 153 gluco-oligosaccharide-NGL probes prepared as previously described in Chapter 2. Carbohydrate sequence information of these probes is in Table S2.1 (positions 1 to 153). The quality control of these microarrays was described in Chapter 2.

Microarray binding analyses were performed essentially as described in Chapter 2 (section 2.5.10). CtCBM11 was analysed at a final concentration of 2, 10 or 20 µg/mL. The

CmCBM6-2 was included as a protein control and analysed at final concentration of 2 µg/mL. The microarray data and metadata, including details of the gluco-oligosaccharide probe library, the generation of the microarrays, imaging, and data analysis are in accordance with the MIRAGE guidelines for reporting glycan microarray-based data<sup>146</sup>.

#### 4.4.5 Crystallization and X-ray Diffraction Data Collection

The complexes of CtCBM11 were produced by overnight incubation of the protein (15-20 mg/mL) with β1,3-1,4 mixed-linked tetrasaccharide (G4G4G3G) and hexasaccharide (G4G3G4G4G3G) ligands at 1:10 molar ratio, respectively. Crystals of each complex were grown in hanging drops, using the vapor diffusion method. Crystals grew from precipitant solutions containing 20-28% (m/v) polyethyleneglycol (PEG) 3350 and 0.2 M potassium phosphate in 0.1 M sodium acetate buffer, pH 4.6. For the CtCBM11-G4G4G3G complex, although sea urchin-like crystals appeared in the drops in one or two days, hexagonal crystals grew later over a period of three weeks. Crystals of the CtCBM11-G4G3G4G4G3G complex appeared after a period of two weeks, although affected by significant multiplicity. All crystals were harvested using a 0.1 M sodium acetate-buffered solution (pH 4.6) containing 30% (m/v) PEG 3350 and 0.2 M potassium phosphate. Crystals grown in 20-24% (m/v) PEG 3350 were flash-cooled frozen in liquid nitrogen using 30% (v/v) glycerol as cryoprotectant added to the harvesting solution, while crystals grown with 28% (m/v) PEG 3350 were flash-cooled using paratone oil.

X-ray diffraction data from a single crystal of the CtCBM11-G4G4G3G complex was collected under a nitrogen stream at 100K in I02 beamline at Diamond Light Source (Oxfordshire, UK), to a maximum resolution of 1.45 Å and using radiation of 0.9763 Å wavelength. The CtCBM11-G4G4G3G crystal indexed in space group *H3* (*R3:H*), with cell constants  $a = b = 103.2$  Å, and  $c = 39.6$  Å, corresponding to a calculated Matthews coefficient of 2.05 Å<sup>3</sup>/Da and a solvent content of 40%, suggesting the presence of one molecule of CtCBM11 in the asymmetric unit. Data for the CtCBM11-G4G3G4G4G3G complex were collected, from a crystal protected with paratone oil and flash-cooled in nitrogen stream at 100 K, in ID23-2 beamline at the ESRF (Grenoble, France) to a maximum resolution of 2.6 Å and using X-ray radiation at a fixed wavelength of 0.8729 Å. The CtCBM11-G4G3G4G4G3G crystals indexed in space group *H3* (*R3:H*), with cell constants  $a = b = 104.9$  Å, and  $c = 39.5$  Å. Data collection, processing, model building and validation statistics are shown in Table 4.1.

#### 4.4.6 Phasing, model building, and refinement

Data sets were processed using MOSFLM<sup>202</sup> and SCALA<sup>203</sup> from the CCP4 suite<sup>204</sup>. Phasing for the CtCBM11-G4G4G3G complex was performed by molecular replacement with Phaser MR<sup>205</sup> from CCP4 using the CtCBM11 polypeptide chain of the PDB ID 1V0A structure<sup>33</sup> to position the protein model in the indexed *H3* space group. After model building and refinement, the polypeptide chain of this new structure (PDB ID 6R3M) was used, in a similar procedure, to solve

the structure of the CtCBM11-G4G3G4G4G3G complex (PDB ID 6R31). Models completion, editing, and initial validation were carried out in COOT<sup>206</sup>. Automatic addition of water molecules and restrained refinement of the full models were done using REFMAC5<sup>207</sup>. Phenix.elBOW from the PHENIX suite<sup>208</sup> was used to generate restraints for  $\beta$ -D-glucose monomers used in refinement of G4G4G3G.

Structure validation was performed using MolProbity<sup>209</sup> and PDB-REDO<sup>210</sup> was used to generate the final models. PRIVATEER<sup>120</sup> was used for the validation of the stereochemistry and conformation of the carbohydrate ligands (Table S4.1). The CtCBM11-G4G4G3G complex, with  $R = 15.7\%$  ( $R_{free} = 18.8\%$ ), consists of 178 amino acid residues, two calcium ions, one acetate and four phosphate ions, 212 water molecules, and one G4G4G3G ligand. The side chain of Leu172 was omitted due to disorder and consequent absence of meaningful electron density. For the CtCBM11- G4G3G4G4G3G complex a final  $R = 18.8\%$  ( $R_{free} = 24.6\%$ ), consisting of 173 amino acid residues, two calcium ions, two phosphate ions, 57 water molecules, and one G4G3G4G4G3G ligand. Residues Asp79 to Ser81 were omitted from the model due to poorly-defined electron density.

In the CtCBM11-G4G4G3G structure, the anomeric carbon of Glc1 could be observed in both  $\alpha$  and  $\beta$  conformation, as supported by the  $mF_o - DF_c$  electron density map (Figure 4.3C). As such, the hydroxyl group was hence modelled in both positions, with partial occupancy.

#### 4.4.7 Isothermal titration calorimetry

Isothermal titration calorimetry (ITC) was performed essentially as described previously<sup>33</sup>, using a MicroCal VP-ITC calorimeter (Northampton, MA, USA) at 25 °C. Before the experiment, purified proteins were buffer-exchanged against 50 mM phosphate buffer, pH 7.0, containing 0.1 mM CaCl<sub>2</sub>. The reaction cell contained protein at 35-50  $\mu$ M, while the syringe contained either the oligosaccharides at 0.5-10 mM or the soluble polysaccharides at 1-6 mg/mL. The ligands were dissolved in the dialysis buffer (separately) to minimize heats of dilution. Titrations were performed by a first injection of 2  $\mu$ L followed by 28 subsequent injections of 10  $\mu$ L aliquots of either polysaccharide or oligosaccharide at 220-s intervals into ITC sample cell (volume 1.4467 mL) containing different enzyme samples. The stirring speed and reference power were set at 307 rpm and 15  $\mu$ cal/s, respectively. The heat background was measured under the same conditions by serial injections of buffer into protein. The molar concentration of CBM binding sites present in polysaccharide ligands was determined as described previously<sup>211</sup>. Data analysis was performed by non-linear regression using a single binding model (MicroCal Origin 7.0 software), and thermodynamic parameters, such as the association constant ( $K_a$ ), number of binding sites in the protein ( $n$ ) and the binding enthalpy change ( $\Delta H$ ) were determined (Table 4.2). Gibbs free energy change ( $\Delta G$ ) and the entropy change ( $\Delta S$ ) were calculated according to Equation 4.1:

$$-RT\ln K_a = \Delta G = \Delta H - T\Delta S \quad (4.1)$$

where  $R$  is the gas constant and  $T$  represents the absolute temperature.

#### 4.5 Work contributions

Experimental planning and work here reported related to the carbohydrate microarrays validation, binding and data analysis, crystallographic structure determination and sequence similarity analysis, as well as protein expression and purification, were performed by the author of the thesis. Mutagenesis experiments and ITC assays were performed by Dr. Virgínia Pires and Dr. Pedro Bule (CIISA-FMV, ULisboa), upon discussion and planning with Prof. Carlos Fontes, as well as data analysis by the author. Barley hexasaccharide preparation and mass spectrometry analysis was performed by Dr. Wengang Chai (Glycosciences Laboratory, Imperial College London). The preparation of the NGL probes and construction of the microarrays resulted from the long-standing collaborative work of the Supervisor, Dr. Angelina Palma, with the group of Prof. Ten Feizi (Glycosciences Laboratory, Imperial College London), and Dr. Hongtao Zhang, Dr. Yibing Zhang, Dr. Lisete M. Silva, Dr. Yan Liu and Dr. Wengang Chai are acknowledged for their contribution to this work. Various barley hydrolysates as sources of oligosaccharides to prepare the NGL probes were provided by Barry V. McCleary (Megazyme International, Ireland) NMR work mentioned in this chapter were executed and analysed by Dr. Aldino Viegas, MSc João Silva, Dr. Filipa Marcelo and Prof. Eurico Cabrita (UCIBIO, NOVA).





# CHAPTER 5

---

**UNRAVELLING FAMILY 50 CBMS OF *CLOSTRIDIUM*  
*THERMOCELLUM*: STRUCTURAL AND FUNCTIONAL  
CHARACTERIZATION OF A NEW LYSM DOMAIN**



## 5 Unravelling family 50 CBMs of *Clostridium thermocellum*: Structural and functional characterization of a new LysM domain

### 5.1 Introduction

Members of CAZy family 50 are also known as Lysin Motif domains (LysMs)<sup>212</sup>. LysMs are widespread protein modules found in prokaryotes and eukaryotes that are highly conserved across all kingdoms of life. They were first identified in the lysozyme of Bacillus phage  $\phi$ 29 by Garvey *et al.* in 1986<sup>213</sup>. These domains are approximately 40 amino acid residues long and present a canonical three-dimensional structure consisting of a  $\beta\alpha\beta$ -fold, in which the two  $\alpha$ -helices are packed against one side of the two-stranded antiparallel  $\beta$ -sheet<sup>42</sup>. LysMs were first classified as a CBM in 2008, upon demonstration that an *N*-terminal LysM domain from *Pteris ryukyuensis chitinase-A* bound to  $\beta$ 1,4-linked *N*-acetylglucosamine (GlcNAc) residues<sup>42,214–216</sup> present in chitin, a polysaccharide that is the main constituent of fungal cell walls. LysMs also recognise different types of bacterial cell wall peptidoglycan (PG), an alternating polymer of GlcNAc and  $\beta$ 1,4-linked-*N*-acetyl-muramic acid (MurNAc)<sup>217</sup>. As chitin and PG display an helical structure<sup>217</sup>, LysMs are suggested to be type B CBMs<sup>215</sup> (for CBM classification see section 1.2.1.1 in Chapter 1).

LysM domains are found individually or in multiple tandem copies (up to 12), mainly at the *N*- or *C*-terminal, in modular proteins<sup>42,214</sup>. In bacteria, LysMs have been reported to mediate recognition of chitin and PG sequences, where multiple LysM domains act additively to increase the binding affinity<sup>215,218</sup>. While GlcNAc seems to be the common monosaccharide bound by all the characterised LysM domains, these modules are present in proteins involved in diverse biological functions<sup>42</sup>. LysMs are present in bacterial extracellular proteins, such as hydrolases and adhesins, acting in bacterial cell wall degradation, but also in bacteriophage lysins, peptidases, chitinases, esterases, reductases and nucleotidases. Numerous LysM domain-containing proteins are virulence factors of human bacterial pathogens, such as *Staphylococcus aureus* (*S. aureus*) that expresses five LysM domains<sup>42,214,215</sup>. They are also found in proteins produced by fungal pathogens as modulators of host immunity, but also in plants involved in defence against pathogens and in symbiotic signalling between bacteria and plants, such as in Nod factors secreted by *Rhizobium* species recognized by the LysMs of plant receptors<sup>42,215</sup>. Furthermore, LysMs also play a role in the development of spores in sporulating bacteria, such as *Bacillus subtilis*<sup>42,177</sup>.

Given their binding properties, LysM domains have been explored for various medical and industrial applications. These domains have been applied for binding to and detection of microbial cells, using LysM-containing proteins or chimeric fusions with other proteins to visualize the cell

wall architecture of Gram-positive bacteria and for display of heterologous proteins on bacterial cell surfaces<sup>219</sup>. LysMs are also used in cell immobilization, to entrap industrially relevant microorganisms in an inert matrix for the production of enzymes, proteins, antibiotics and chemical compounds<sup>219</sup>. The potential to use non-genetically modified bacteria as vectors in vaccine development have also been explored, where LysMs allow the immobilization of purified fusion proteins to Gram-positive bacteria<sup>219,220</sup>. The LysM domain of an *N*-acetylglucosaminidase from *Lactococcus lactis* has been used to bind antigens to non-genetically modified Gram-positive bacteria for immunization purposes<sup>220</sup>.

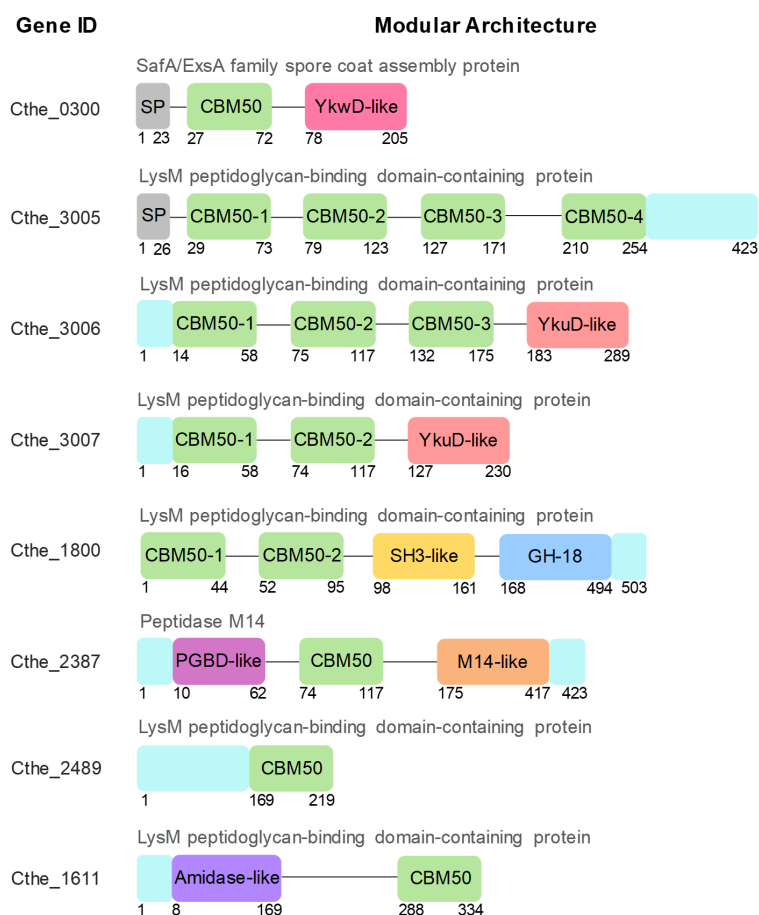
The *Clostridium thermocellum* genome expresses a high number of family 50 CBMs for which a carbohydrate-binding function is yet to be assigned. These CtCBMs are found in tandem or isolated in LysM-containing proteins, associated with putative catalytic modules, including a family 18 glycoside hydrolase (GH18), and some unidentified proteins (Figure 5.1). Given the high number of family 50 CBMs in this bacterium (15 CBMs), the second highest after family 3 CBMs (Figure 3.2, Chapter 3), and the potential biotechnological applications of LysM domains, we sought to determine the structure and ligand-binding specificity of these modules.

The carbohydrate microarray analysis presented in Chapter 3 revealed the carbohydrate binding for these CBMs and showed that these are highly specific for GlcNAc oligosaccharides exhibiting a chain-length dependency. In this chapter, the binding specificity of *C. thermocellum* family 50 CBMs was further explored along with the structural characterization of one LysM domain in complex with its GlcNAc trisaccharide ligand. Binding capability to insoluble chitin and to PGs from different bacteria was also assessed using co-precipitation assays. Mutagenesis, ITC and molecular dynamics simulation studies allowed to identify the molecular determinants of carbohydrate recognition to chitin and PG sequences. The understanding of the carbohydrate recognition mechanism by these modules to chitin and peptidoglycan, will contribute to elucidating their role in *C. thermocellum* and will also potentiate the development of novel strategies using LysM domains in industrial and therapeutic applications.

## 5.2 Results and Discussion

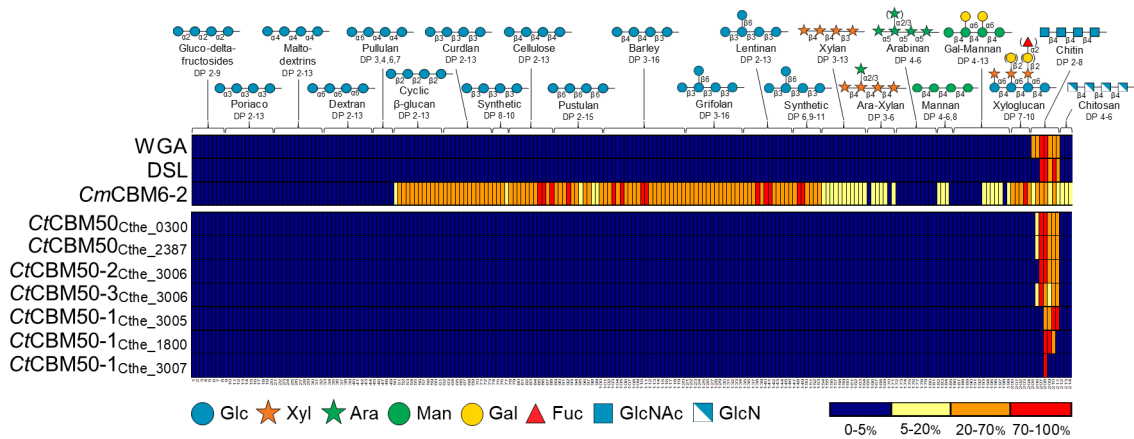
### 5.2.1 Oligosaccharide specificity of *C. thermocellum* family 50 CBMs

To assign the carbohydrate binding specificity at oligosaccharide level, 11 out of the 15 family 50 CtCBMs (Figure 3.2 and Tables S3.1 and S3.2 in Chapter 3) were analysed using a NGL-microarray comprised of diverse sequence-defined oligosaccharides, which included  $\beta$ 1,4-GlcNAc or  $\beta$ 1,4-glucosamine (GlcN) oligosaccharides with different chain lengths (Table S2.1 in Chapter 2, probes 205 to 214). Binding patterns were obtained for 7 CtCBMs 50, which showed these to be highly specific for  $\beta$ 1,4-GlcNAc sequences with increased binding intensities with the oligosaccharide chain-length (Figures 5.2 and 5.3, and Table S3.7 in Chapter 3). The narrow binding of CtCBMs 50 was supported by the binding patterns of GlcNAc-specific plant lectins *Datura stramonium* (DSL) and wheat germ agglutinin (WGA) and contrasted



**Figure 5.1. Modular architecture of proteins containing family 50 CBMs in the genome of *C. thermocellum*.** Signal peptides (SP) are coloured grey, unknown amino acid sequences are coloured light blue and family 50 CBMs are coloured green. The associated enzymes and domains are coloured according to sequence homology: YkwD - subgroup of cysteine-rich secretory proteins, antigen 5, and pathogenesis-related 1 proteins (CAP); YkuD - L,D-transpeptidase/carboxypeptidase; SH3 - SRC homology 3 domain; GH18 - family 18 glycoside hydrolase; PGBD - Peptidoglycan binding domain; M14 - Peptidase M14. The predicted linker sequences are depicted by a line. The modular proteins are identified by gene ID (left panel) and the annotated protein names are shown (right panel). Sequence homology search was performed using Basic Local Alignment Search Tool from NCBI<sup>221</sup>, Uniprot<sup>222</sup> and InterProScan<sup>223</sup>.

with the broad binding to  $\beta$ -linked sequences by *Cm*CBM-6-2 (Figure 5.2 and Table S2.4 in Chapter 2). While *Ct*CBM50<sub>Cthe\_0300</sub>, *Ct*CBM50<sub>Cthe\_2387</sub> and *Ct*CBM50-3<sub>Cthe\_3006</sub> exhibited a minimum chain-length requirement of 3 GlcNAc residues, *Ct*CBM50-2<sub>Cthe\_3006</sub> and *Ct*CBM50-1<sub>Cthe\_3005</sub> and *Ct*CBM50-1<sub>Cthe\_1800</sub> seem to require longer epitopes for GlcNAc recognition, binding from DP-4 and DP-5 onwards, respectively (note that for *Ct*CBM50-1<sub>Cthe\_1800</sub>, both low and high levels are represented, as there was a misprint of DP-8 high level). *Ct*CBM50-1<sub>Cthe\_3007</sub>, showed weak binding signal and bound only to the GlcNAc probe with DP-5. Although all the *Ct*CBMs 50 seem to require the *N*-acetyl group for binding, as none bound to the  $\beta$ 1,4-GlcN probes, the fact that some showed different chain-length requirements may suggest subtle differences in their ligand recognition mechanisms.



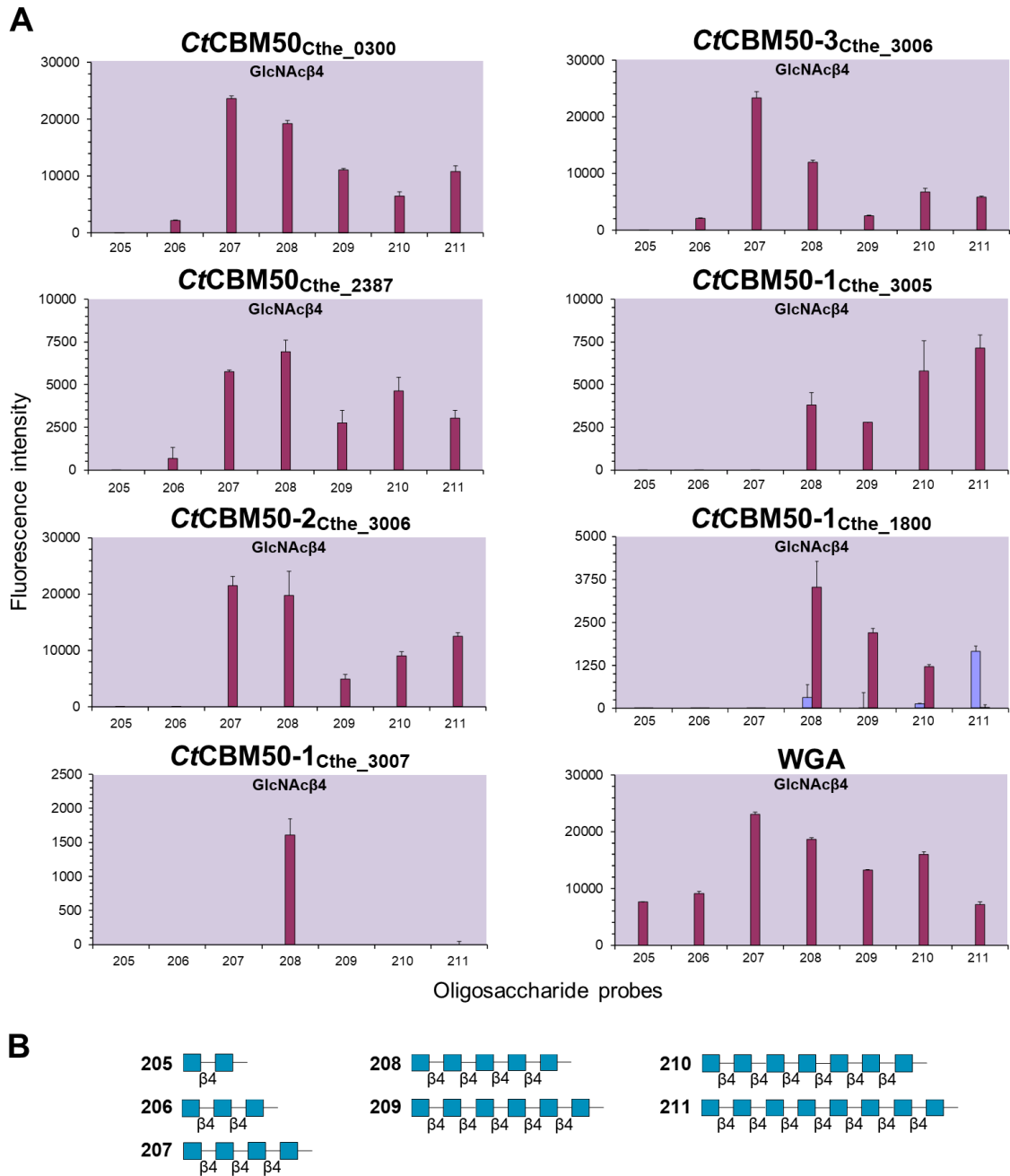
**Figure 5.2. Oligosaccharide microarray analysis of *C. thermocellum* family 50 CBMs.** The microarray highlighted comprises 214 NGL probes with a wide degree of polymerization (DP) range of linear and branched oligosaccharide sequences of  $\alpha$ - and  $\beta$ -glucans<sup>32</sup>,  $\beta$ -xyllans,  $\alpha$ -arabinans,  $\beta$ -mannans, xyloglucans, chitin and chitosan. Carbohydrate sequence information on these probes is shown in Chapter 2, Table S2.1. The proteins analysed are depicted at the left: plant lectins WGA and DSL, and *CmCBM6-2*, used in the comparative validation of the microarrays (upper panel); and *CtCBMs 50* (bottom panel). The relative binding intensities were calculated as the percentage of the fluorescence signal intensity at 5 fmol given by the probe most strongly bound by each protein (normalized as 100%). Numerical scores are given in Chapter 3, Table S3.7. The monosaccharide symbolic representation used was according to the updated SNFG<sup>1</sup>.

Given that *CtCBM50*<sub>Cthe\_0300</sub> (henceforward designated as *CtCBM50*) exhibited a higher binding avidity to  $\beta$ 1,4-GlcNAc sequences in the microarrays, and is associated as a single LysM domain on a putative spore coat assembly protein, interest arose in the structural characterization of its carbohydrate recognition interface and associated mechanisms, which will be explored in the following sections.

### 5.2.2 *CtCBM50* structure in complex with $\beta$ 1,4-GlcNAc trisaccharide

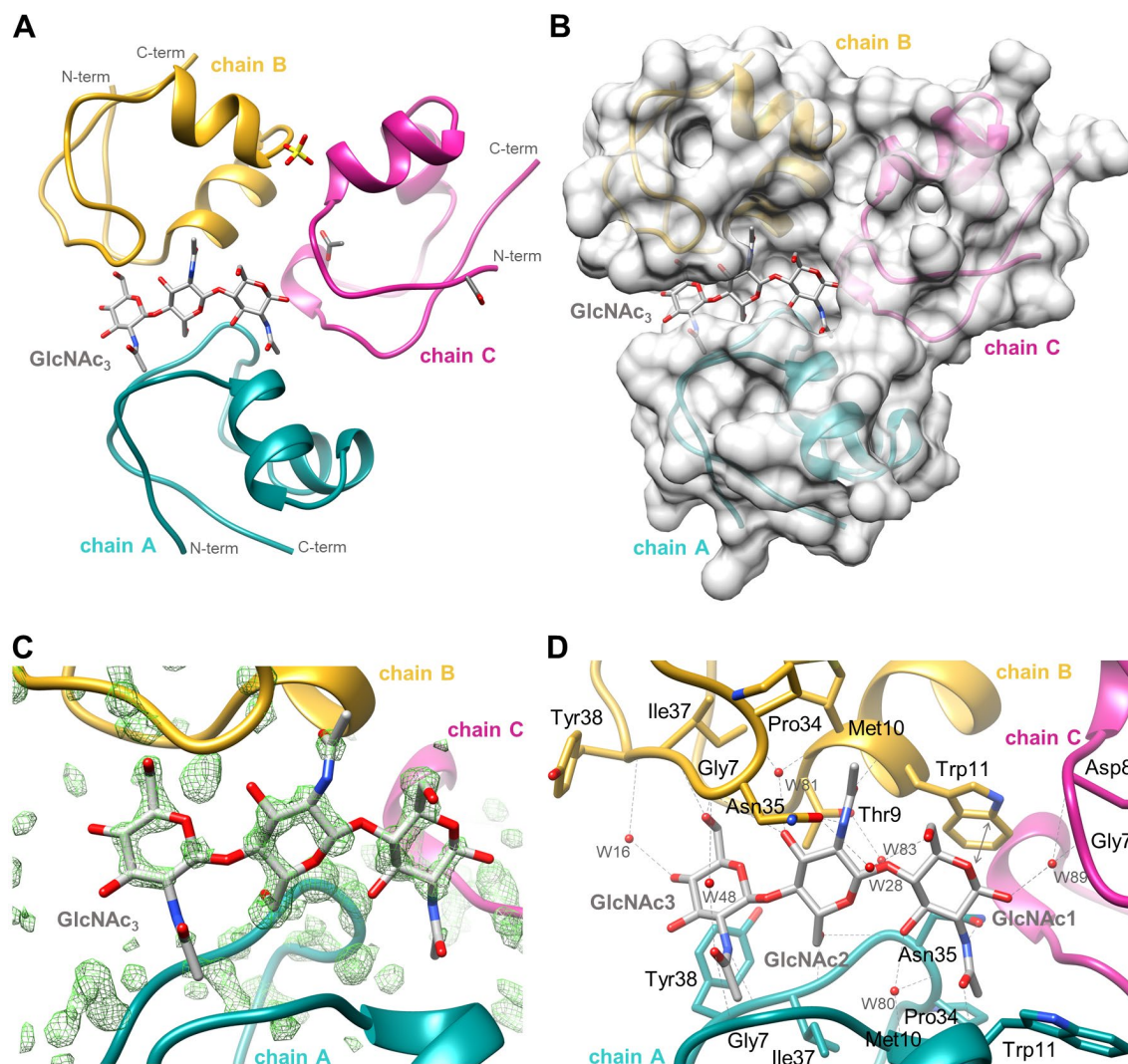
Crystallization experiments were carried out for *CtCBM50* both as isolated and after incubation with  $\beta$ 1,4-linked *N*-acetyl GlcNAc trisaccharide (GlcNAc<sub>3</sub>) for protein-ligand complex formation, as this was the minimum epitope recognised by the CBM in the microarrays.

The *CtCBM50* crystal structure could only be solved in the presence of GlcNAc<sub>3</sub>, at a resolution of 1.45 Å (Figure 5.4), as crystallization was unsuccessful for the isolated CBM. Statistics of X-ray diffraction data processing, model building, refinement and validation are presented in Tables 5.1, S5.1 and S5.2. *CtCBM50* presented the typical  $\beta\alpha\alpha\beta$ -fold of LysM domains, with 3 molecules of the CBM (chains A, B and C) and one GlcNAc<sub>3</sub> ligand (Figure 5.4A) in the asymmetric unit. GlcNAc<sub>3</sub> was accommodated at the interface between chains A and B, with minor contact of chain C with the reducing end of GlcNAc<sub>3</sub>. The interface between the *CtCBM50* chains A and B (*CtCBM50*<sub>AB</sub>) formed a binding cleft-like site (Figure 5.4B). The unbiased  $mF_o-DF_c$  electron density map calculated in the absence of the GlcNAc<sub>3</sub> atom coordinates (Figure 5.4C) supported the ligand location.



**Figure 5.3. Comparative analysis of *C. thermocellum* family 50 CBMs binding to  $\beta$ 1,4-linked GlcNAc oligosaccharides. (A)** The binding signals of each CtCBM50 are depicted as means of fluorescence intensities of duplicate spots at 5 fmol (and also at 2 fmol for CtCBM50-1<sub>cthe\_1800</sub>) of oligosaccharide probe arrayed (with error bars) and are representative of at least two independent experiments. **(B)** The microarrays included 7  $\beta$ 1,4-linked GlcNAc NGL-oligosaccharides with DP-2 to DP-8.

The identified residues that constitute the binding site of CtCBM50<sub>AB</sub> interacted with the ligand mostly through hydrogen bonding (Figure 5.4D). As such, GlcNAc<sub>3</sub> established direct hydrogen bonds with chains A and/or B residues Trp11, Asn35, Ile37, and Gly7, and water-mediated hydrogen bonds with Pro34, Met10, Thr9, Asn35 and Ile37 (Table S5.3). These interactions likely contribute to define the specific conformation of GlcNAc<sub>3</sub> in CtCBM50<sub>AB</sub>'s binding site. Although most contacts were established with the polypeptide's main chain atoms, the direct hydrogen



**Figure 5.4. Ribbon representation of the three-dimensional crystal structure of the CtCBM50-GlcNAc<sub>3</sub> complex.** (A) Representation of the overall structure of CtCBM50 exhibiting the typical  $\beta\alpha\beta$ -fold of LysM domains, with 3 molecules of the CBM and 1 GlcNAc<sub>3</sub> trisaccharide ligand in the asymmetric unit, chain A-B-C. (B) Cartoon and surface representation of the CtCBM50-GlcNAc<sub>3</sub> complex. The 3 CBM chains form a binding cleft where the ligand is accommodated. (C) Initial mF<sub>o</sub>-DF<sub>c</sub> electron density map, calculated in the absence of GlcNAc<sub>3</sub>, at a maximum resolution of 1.45 Å. GlcNAc<sub>3</sub> is overlaid in the picture for reference. The electron density map is shown in green mesh, contoured at 2.5  $\sigma$ ; (D) Close-up view on the binding site of CtCBM50 evidencing the protein-ligand contacts established between the CBM chains and GlcNAc<sub>3</sub> as listed in Table S5.3. Chain A is represented in cyan, chain B in yellow and chain C in magenta. GlcNAc<sub>3</sub> is represented as stick model in grey and by atom type. The carbohydrate chains and the side chains of the amino acid residues that interact with the ligand are shown as sticks coloured by atom type. Water molecules are indicated as red spheres. Hydrogen bonding is indicated by dashed lines and CH- $\pi$  stacking interactions are represented as double arrows.

bonding between the O $\delta$  atom in the carboxyl group of Asn35 side chain and the amine of GlcNAc1 and GlcNAc2 *N*-acetyl groups, in chain A and B respectively, may point to a key role of this residue. This highlights as well the importance of the *N*-acetyl group for CtCBM50 binding recognition, as indicated in the carbohydrate microarray analysis (Figure 5.2). Key interactions with GlcNAc *N*-acetyl groups have also been reported for *Enterococcus faecalis* AtIA, in which the methyl groups fit into hydrophobic pockets while the carbonyl groups form hydrogen bonds<sup>215</sup>.



**Table 5.1. X-ray diffraction and structure refinement parameters and statistics for C<sub>t</sub>CBM50-GlcNAc<sub>3</sub>.**

<b>Data collection</b>	
Beamline	ESRF, ID29
Space Group	C2
<b>Cell parameters</b>	
<i>a</i> , <i>b</i> , <i>c</i> (Å)	99.39, 41.77, 42.87
$\alpha$ , $\beta$ , $\gamma$ (°)	90.00, 96.89, 90.00
Wavelength, Å	0.9677
Resolution of data (outer shell), Å	42.56-1.45 (1.48-1.45)
Total number of reflections (outer shell)	260560 (9664)
Number of unique reflections (outer shell)	30139 (1427)
$R_{\text{pim}}$ (outer shell), % <sup>a</sup>	0.025 (0.468)
$R_{\text{merge}}$ (outer shell), % <sup>b</sup>	0.048 (0.735)
Mean $I/\sigma(I)$ (outer shell)	20.9 (2.20)
CC(1/2) (outer shell)	1.00 (0.81)
Completeness (outer shell), %	97.1 (91.7)
Redundancy (outer shell)	8.6 (6.8)
<b>Structure refinement</b>	
No. of protein atoms	
Chain A	368
Chain B	373
Chain C	371
No. of solvent waters	125
Resolution used in refinement, Å	1.45
No. of reflections	28591
$R_{\text{work}} / R_{\text{free}}$ <sup>c</sup>	0.176 / 0.194
rms deviation bonds (Å)	0.014
rms deviation angles (°)	2.277
rms deviation chiral volume (Å <sup>3</sup> )	0.151
<b>Avg B factors (Å<sup>2</sup>)</b>	
Main chain A	17.8
Main chain B	22.7
Main chain C	18.9
Side chain A	26.8
Side chain B	23.1
Side chain C	24.8
GlcNAc 1	14.9
GlcNAc 2	13.6
GlcNAc 3	14.5
Acetate ion 1	21.2
Acetate ion 2	42.9
Sulphate ion 1	62.1
Water molecules	35.5
<b>Ramachandran statistics</b>	
<i>favored</i>	128
<i>allowed</i>	2
<i>generously allowed</i>	0
<i>forbidden</i>	0

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

<sup>b</sup>  $R_{\text{sym}} = \left( \frac{\sum_{hkl} \sum_j |I_{hkl,j} - \langle I_{hkl} \rangle|}{\sum_{hkl} \sum_j I_{hkl,j}} \right)$ , where  $\langle I_{hkl} \rangle$  is the average of symmetry-related observations of a unique reflection.

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

The interactions created with the *N*-acetyl groups also explain the deviation between chains A and B in the ligand interface, as the alternating orientation of the GlcNAc monomers causes one of the CBM molecules to shift towards the *N*-acetyl to establish the contact with its Asn35 carboxyl group. Chain C established only a direct contact through Gly7 NH group and the HO-C1 of

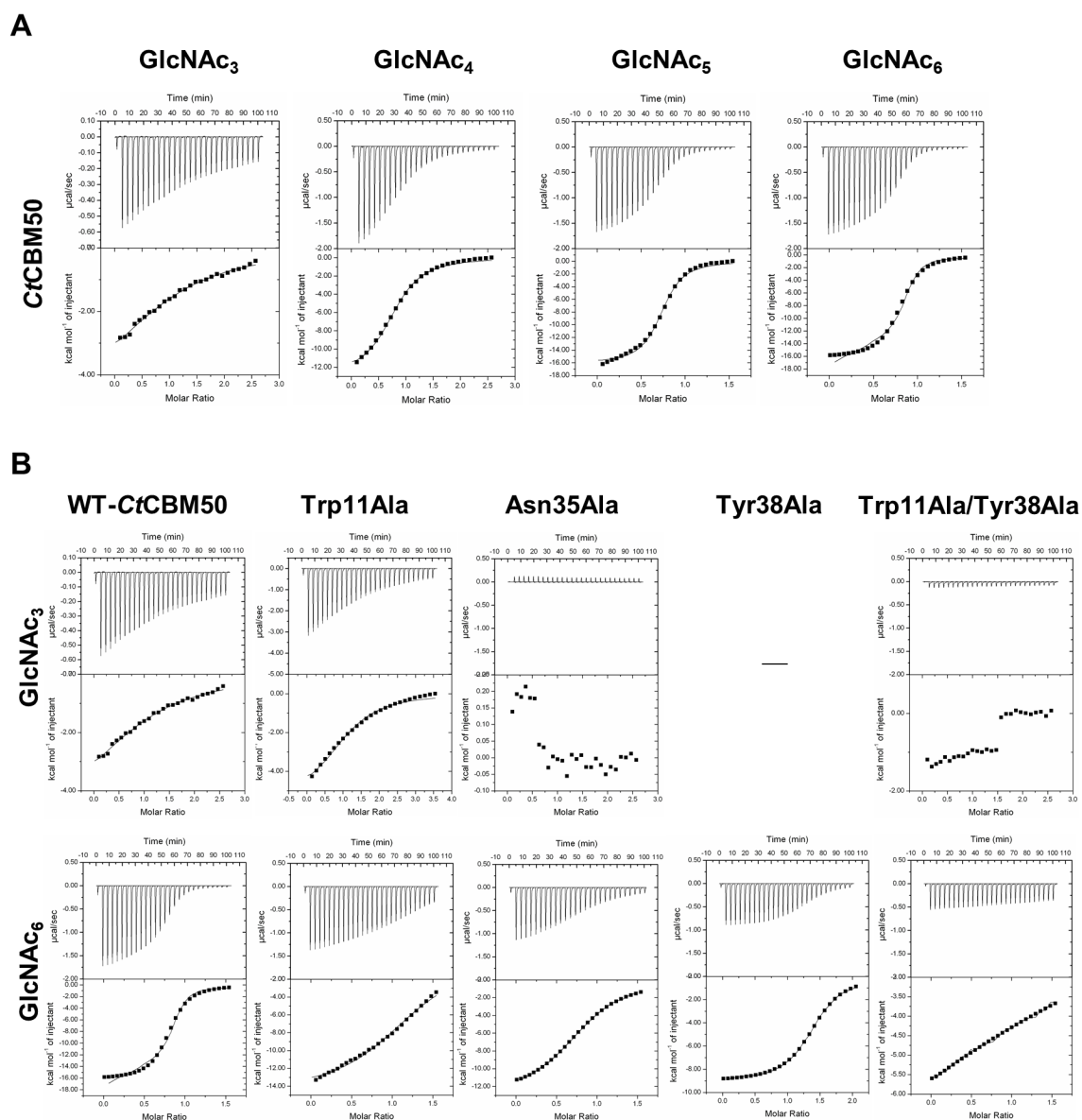
GlcNAc1, and a water-mediated hydrogen bond of Asp8 carbonyl group with HO-C1 of GlcNAc1 monomer. In the presence of a longer chain-length sequence, chain C might however re-orientate in order to accommodate also part of the ligand in its binding site. Additionally, a single hydrophobic CH- $\pi$  stacking interaction was identified, between the side chain of Trp11 from chain B and the glucose ring in GlcNAc1. This observation was also reported for *E. faecalis* AtlA, where only one CH- $\pi$  stacking interaction involving the aromatic ring was identified<sup>215</sup>. Although in a rotamer conformation pointing in the ligand direction, the aromatic ring of chain A Trp11 was placed too far from GlcNAc<sub>3</sub> to establish an interaction. Hypothetically, when binding to longer chain-length GlcNAc sequences, this residue would likely interact with the ligand, reinforcing the important role of Trp11 for CtCBM50's ligand recognition. The recognition as revealed by the CtCBM50<sub>AB</sub>-GlcNAc<sub>3</sub> structure, seems to result from a ligand-induced interchain LysM multivalent assembly event, as also reported for *Thermus thermophilus* P60\_2LysM<sup>224</sup>, evidencing how single LysMs may cooperate to increase the binding affinity<sup>218</sup>.

Aiming to a better understanding of CtCBM50 binding mechanism, co-crystallization assays were also performed with longer chain-length GlcNAc ligands, with DP-5 and DP-6, however crystals were not possible to obtain to date.

### 5.2.3 Binding affinity of CtCBM50 to $\beta$ 1,4-GlcNAc oligosaccharides and influence of chain-length

The information obtained from the oligosaccharide microarrays and the CtCBM50-GlcNAc<sub>3</sub> crystal structure allowed the identification of CtCBM50 ligand-specificity and chain-length requirement, as well as key amino acid residues involved in its binding to GlcNAc. In order to determine the contribution of additional GlcNAc monomers to the interaction with CtCBM50, ITC measurements were performed with varying DPs (GlcNAc<sub>3</sub> to GlcNAc<sub>6</sub>).

The ITC results corroborated that the affinity of CtCBM50 to GlcNAc oligosaccharides is chain-length dependent (Figure 5.5A and Table 5.2). CtCBM50 exhibited a chain-length dependency up to DP-5, with a 100-fold increase of the  $K_a$  from DP-3 to DP-5 ( $3.10 \times 10^4$  to  $1.21 \times 10^6 \text{ M}^{-1}$ ). The affinity seemed to stabilise, not increasing significantly from DP-5 to DP-6 (with a  $K_a$  of  $1.31 \times 10^6 \text{ M}^{-1}$  for DP-6). These results point to a binding site comprised of at least five binding subsites, each accommodating an individual monosaccharide, in accordance with what was previously reported for *P. ryukyuensis* LysM domain (PrLysM2)<sup>225</sup>. Additionally, the increase in the enthalpy observed is indication that new interactions are being established with the addition of the fourth and the fifth monosaccharide to GlcNAc<sub>3</sub>. Interestingly, from GlcNAc<sub>4</sub> to GlcNAc<sub>6</sub> it is observed a decrease in the calculated  $n$  value from 1 to 0.8/0.7, suggesting that a quarter of the oligosaccharide molecules may be involved in bivalent binding by the CBM. This suggests the existence of two distinct binding epitopes in GlcNAc oligosaccharides for DP greater than 4, and that the increase in affinity observed from GlcNAc<sub>3</sub> to GlcNAc<sub>5</sub>, is possibly a result of positive cooperativity between them. This assumption is supported by the substantial decrease in entropy



**Figure 5.5. Isothermal calorimetry titrations of binding of CtCBM50 and its mutant derivatives to  $\beta$ 1,4-linked GlcNAc oligosaccharides. (A) Analysis of chain-length dependency of CtCBM50 binding to GlcNAc tri- to hexaccharides; (B) Analysis of the binding of CtCBM50 wild type and its mutants to trisaccharide GlcNAc<sub>3</sub> and hexasaccharide GlcNAc<sub>6</sub>. The top portion of each panel shows the raw power data while the bottom parts show the integrated and heat of dilution corrected data. The solid lines show the non-linear curve fits to a one site binding model with the stoichiometry fixed at 1. Thermodynamic parameters are given in Table 5.2.**

that is observed from GlcNAc<sub>3</sub> to GlcNAc<sub>4</sub>, and more pronounced to GlcNAc<sub>5</sub>. The decrease in entropy is an indication of an increase in the rigidity of the system, suggesting that conformational rearrangement may be occurring upon binding, as already had been reported for the *E. faecalis* LysM domains<sup>215</sup>. This could point to an interchain multivalent assembly of LysM modules induced by the longer GlcNAc sequences, where the number of binding sites (CBM molecules) measured increase with the longer chain-length ligands. This is in line with what has been suggested, that individual LysM domains bind in a cooperative manner to long ligand chains, but not to short

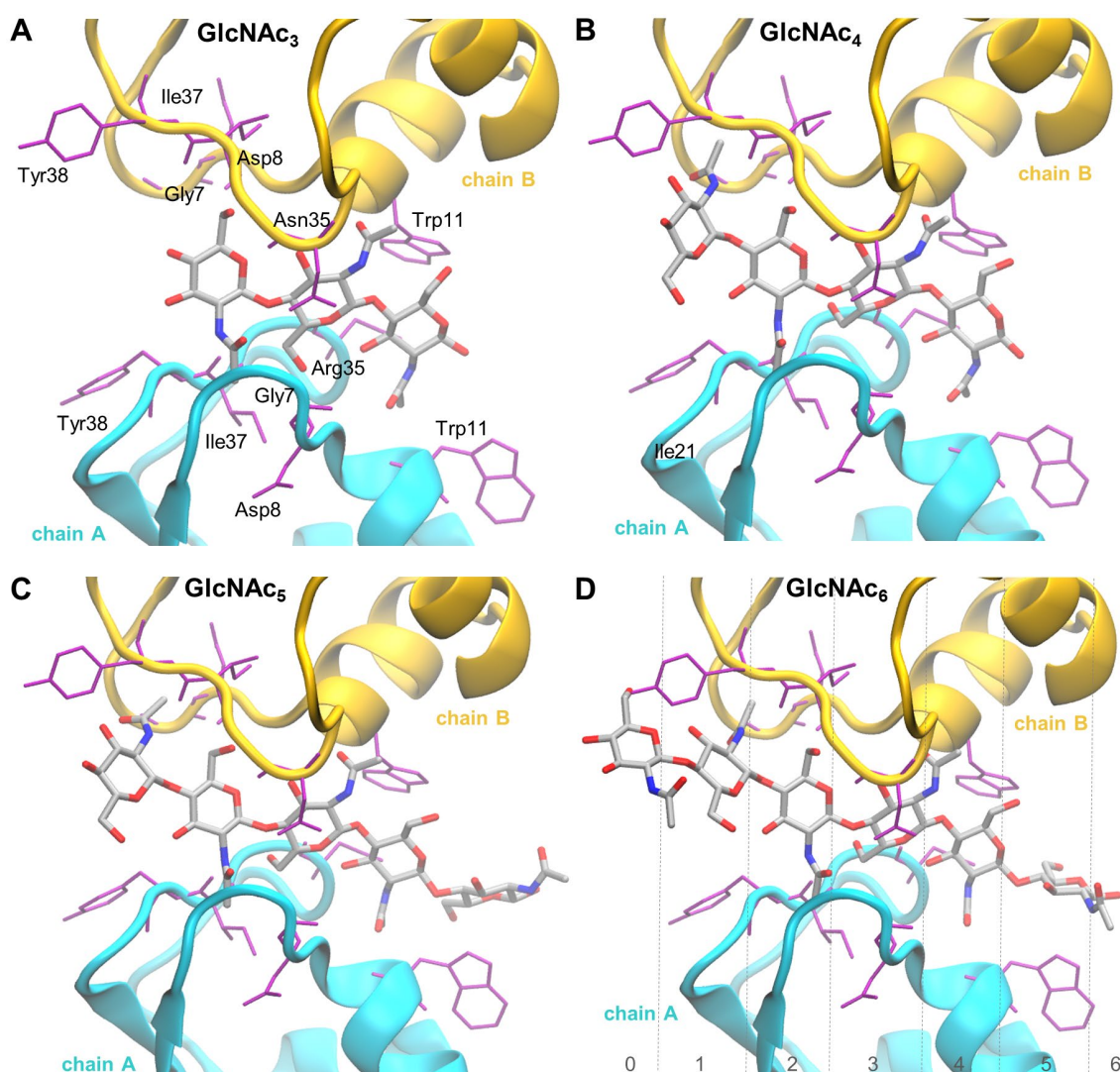
**Table 5.2. Thermodynamic parameters of the binding of CtCBM50 and its mutant derivatives to polysaccharides and oligosaccharides.**

CtCBM50 variant	Ligand	$K_a$ ( $M^{-1}$ )	$\Delta G$ ( $kcal.mol^{-1}$ )	$\Delta H$ ( $kcal.mol^{-1}$ )	$T\Delta S$ ( $kcal.mol^{-1}$ )	$n$
WT	GlcNAc <sub>3</sub>	$3.10 (\pm 0.36) \times 10^4$	-6.13	$-5.22 \pm 0.43$	0.91	$1.08 \pm 0.06$
	GlcNAc <sub>4</sub>	$2.48 (\pm 0.15) \times 10^5$	-7.35	$-12.84 \pm 0.19$	-5.49	$0.81 \pm 0.01$
	GlcNAc <sub>5</sub>	$1.21 (\pm 0.10) \times 10^6$	-8.30	$-16.05 \pm 0.16$	-7.75	$0.73 \pm 0.01$
	GlcNAc <sub>6</sub>	$1.31 (\pm 0.15) \times 10^6$	-8.35	$-16.22 \pm 0.25$	-7.87	$0.80 \pm 0.01$
Trp11Ala	GlcNAc <sub>3</sub>	$9.90 (\pm 0.35) \times 10^3$	-5.44	$-9.22 \pm 0.28$	-3.79	$1.11 \pm 0.03$
	GlcNAc <sub>6</sub>	$1.74 (\pm 0.14) \times 10^5$	-7.16	$-14.49 \pm 0.11$	-7.33	$1.25 \pm 0.01$
Asn35Ala	GlcNAc <sub>3</sub>			No binding		
	GlcNAc <sub>6</sub>	$2.03 (\pm 0.05) \times 10^5$	-7.26	$-12.92 \pm 0.78$	-5.66	$0.84 \pm 0.00$
Tyr38Ala	GlcNAc <sub>3</sub>			Not tested		
	GlcNAc <sub>6</sub>	$8.62 (\pm 0.19) \times 10^5$	-8.10	$-9.07 \pm 0.22$	-0.97	$1.40 \pm 0.00$
Trp11Ala/ Tyr38Ala	GlcNAc <sub>3</sub>			No binding		
	GlcNAc <sub>6</sub>	$8.59 (\pm 4.42) \times 10^3$	-5.36	$-10.04 \pm 0.19$	-4.68	$3.66 \pm 0.38$

oligosaccharides<sup>215</sup>. These results suggest that the interchain assembly already observed in the CtCBM50-GlcNAc<sub>3</sub> structure, although possibly induced by the packing in the crystal, could be observed in solution for the longer ligands.

#### 5.2.4 Molecular determinants of CtCBM50 ligand recognition and chain-length dependency

Failing to produce a crystal structure of CtCBM50 complexed with longer GlcNAc oligosaccharides, simulation approaches using Molecular Dynamics (MD) calculations were used to study the molecular interactions underlying CtCBM50 ligand recognition to these ligands. To validate the approach, simulated complexes of CtCBM50 with GlcNAc oligosaccharides with DP-3 were produced and studied using MD calculations in parallel with complexes of DP-4 to DP-6. The various CtCBM50<sub>AB</sub>-GlcNAc complexes resulting from the simulations are illustrated in Figure 5.6, and the respective calculated binding energies are presented in Table 5.3. The CBM-carbohydrate interactions obtained for the various oligosaccharides simulated are listed in Table S5.4. Only those complexes showing the most favorable binding energies for each oligosaccharide are shown and discussed (the other tested poses can be found in Figure S5.1 and Tables S5.5 to S5.6). For these calculations, only chains A and B of CtCBM50 were considered based on the structural evidence that chain C has no significant contacts with the GlcNAc<sub>3</sub> ligand. In agreement with the ITC results, and as foreseen in the microarray analysis, CtCBM50<sub>AB</sub> bound with higher affinity to GlcNAc oligosaccharides with chain-lengths higher than DP-3, with a more pronounced effect observed when the monosaccharide is added to the non-reducing end of GlcNAc<sub>3</sub> (Figure 5.7A). The additional monosaccharide of GlcNAc<sub>4</sub>, was recognized by chain B's NH of Gly7 and the carbonyl of Ile37 (designated subsite 2). The binding of the additional unit of GlcNAc<sub>5</sub>, was stabilized by a hydrophobic contact with the Trp11 of chain A (subsite 6), as well as via additional water-mediated hydrogen bonds with the CBM (Table S5.4). The side chain of Tyr38 from chain B interacted with the *N*-acetyl group of the extra



**Figure 5.6. Representation of the last simulation structure of the various GlcNAc oligosaccharides bound to the CtCBM50 complex of chains A and B. (A) GlcNAc<sub>3</sub>; (B) GlcNAc<sub>4</sub>; (C) GlcNAc<sub>5</sub>; and (D) GlcNAc<sub>6</sub>.** Chain A is represented in blue cartoon and chain B in yellow. GlcNAc oligosaccharides are represented as sticks and coloured by atom type. All residues involved in the binding interface and that established hydrogen bonds or hydrophobic contacts with the various oligosaccharides are represented as sticks and coloured purple. The binding subsites (1-6) are also represented in (D).

unit of GlcNAc<sub>6</sub> (subsite 1, Figure 5.7B). From this analysis, 6 binding subsites (1 to 6) could be suggested (Figure 5.7A). The study of the hydrogen bonds established between the protein chains A and B and the GlcNAc sequences, along the last 30 ns of MD simulations, revealed that the *N*-acetyl and the HO-C6 groups are the ones that interact more with the CtCBM50 chains (Table S5.4). Furthermore, the *N*-acetyl groups also mediate dispersive contacts through the methyl group that fit into hydrophobic pockets of both chains, Val4, Ile37 and Pro39 in subsites 1 and 2, and Ile21, Ile25 and Pro34 in subsites 4 and 5.

To assess if a single CtCBM50 chain is sufficient to bind the GlcNAc oligosaccharides, several MD simulations with the various ligands bound to CtCBM50 chains A or B were also carried out (Table S5.6). The lower binding free-energies obtained for the CtCBM50<sub>AB</sub>, in relation with each

**Table 5.3. Binding enthalpies and binding free-energies of the GlcNAc ligands to the CtCBM50 chains A and B.** The results are shown relative to the CtCBM50<sub>AB</sub>-GlcNAc<sub>3</sub> complex.

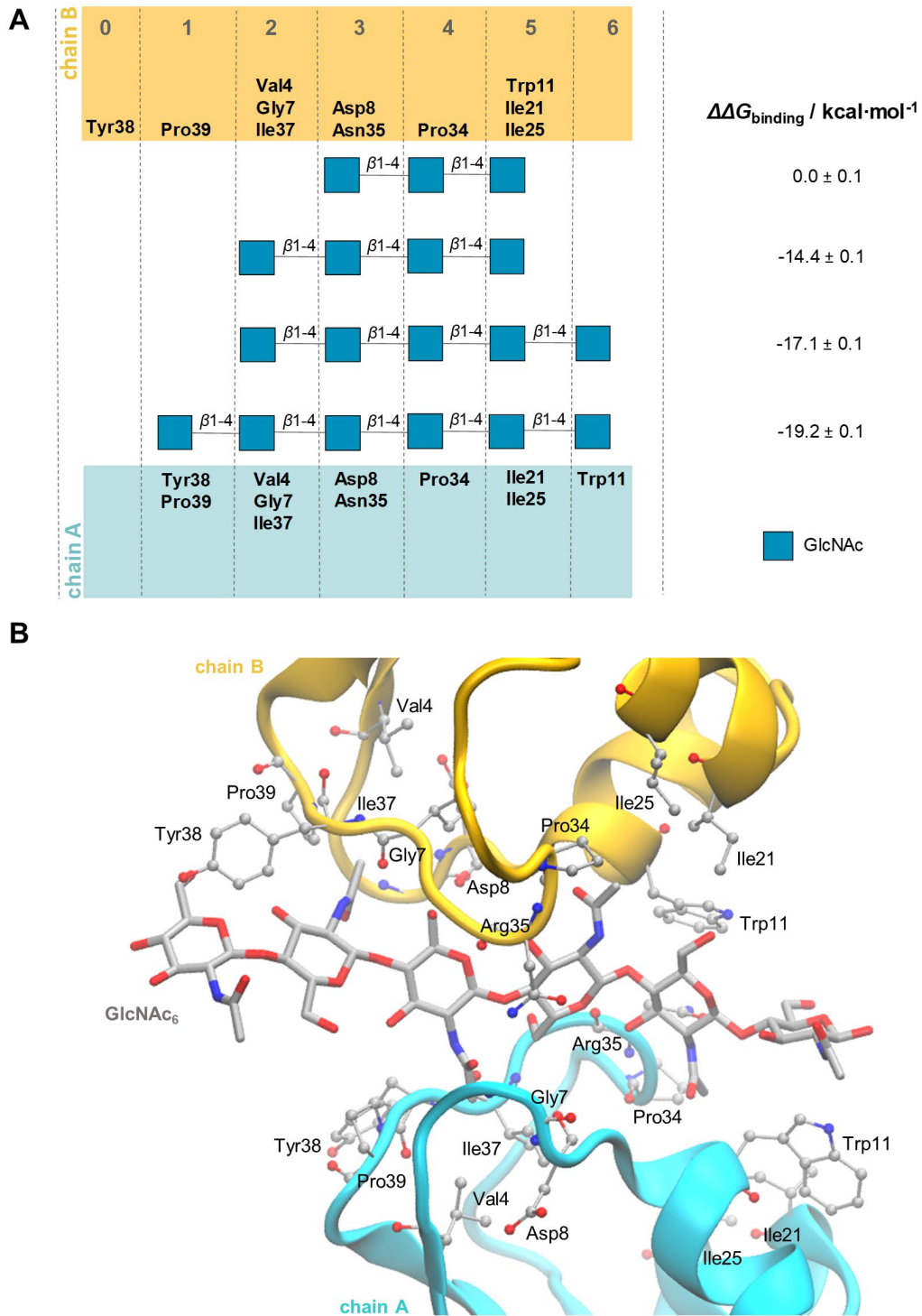
Ligand	$\Delta\Delta H_{\text{binding}}$ (kcal·mol <sup>-1</sup> )	$\Delta\Delta G_{\text{binding}}$ (kcal·mol <sup>-1</sup> )
GlcNAc <sub>3</sub>	0.0 ± 0.7	0.0 ± 0.1
GlcNAc <sub>4</sub>	-21.0 ± 0.7	-14.4 ± 0.1
GlcNAc <sub>5</sub>	-27.3 ± 0.7	-17.1 ± 0.1
GlcNAc <sub>6</sub>	-30.2 ± 0.7	-19.2 ± 0.1

individual chain, suggested that the intermolecular assembly with symmetry-related chains of CtCBM50 facilitated the carbohydrate binding. This fact is in accordance with what has been suggested for LysM domains, in which multiple domains cooperate to enhance binding to GlcNAc polymers and with the ITC results discussed above. Although chain C was not considered for the present simulations, as it did not seem to establish any direct contacts with GlcNAc<sub>3</sub> in its binding site, it would be of interest to evaluate the contribution of this chain for the binding of longer chain-length ligands.

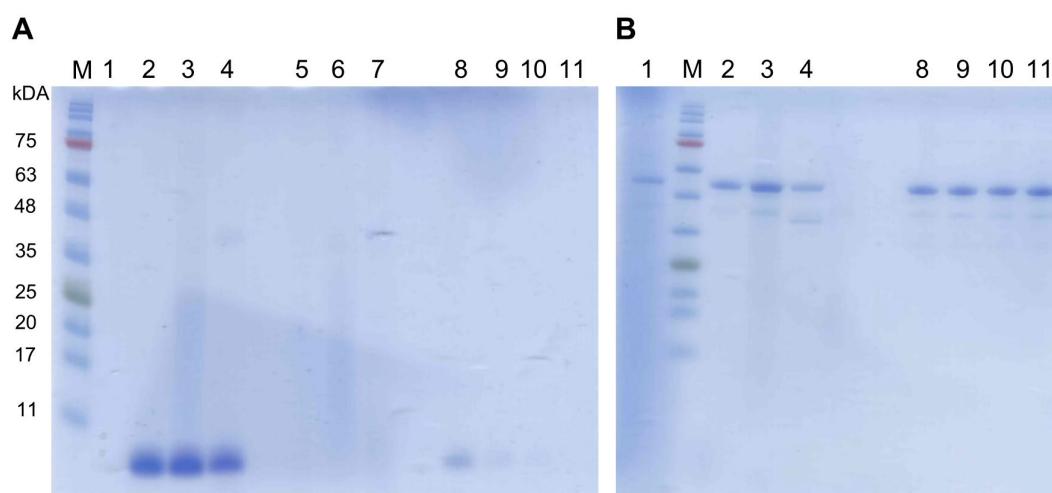
Considering the simulations results and the structure-based rationale, mutant alanine derivatives of residues involved in direct hydrogen bonds and CH- $\pi$  interactions with the ligand (Trp11, and Asn35 and Tyr38) were produced to analyse the role of the interacting amino acid residues for CtCBM50 ligand recognition and chain-length dependency (Figure 5.5B and Table 5.2). The wild type CBM and mutants were analysed by ITC against GlcNAc<sub>3</sub> and GlcNAc<sub>6</sub>. Asn35Ala mutant, produced to study the influence of the hydrogen bonding it established with GlcNAc ligands, abolished binding to GlcNAc<sub>3</sub>, while decreasing the affinity for GlcNAc<sub>6</sub> by 10-fold. The mutants Trp11Ala and Tyr38Ala were produced to assess the influence of longer chain ligands to the interaction. Trp11Ala led to a significant decrease in the affinity for GlcNAc<sub>6</sub> of 10-fold, while having only a small effect in the affinity for GlcNAc<sub>3</sub> (with a  $K_a$  of  $9.90 \times 10^3$  from  $3.10 \times 10^4$  M<sup>-1</sup>). Tyr38Ala mutant, led only to a small decrease of CtCBM50 affinity to GlcNAc<sub>6</sub> (with a  $K_a$  of  $8.62 \times 10^5$  from  $1.31 \times 10^6$  M<sup>-1</sup>). However, the double mutant Trp11Ala/Tyr38Ala showed a cumulative effect in the affinity for both ligands, by abolishing binding to GlcNAc<sub>3</sub> and significantly decreasing the affinity for GlcNAc<sub>6</sub> by over 100-fold. These results corroborate that these residues and the interactions they are mediating play a key role in CtCBM50 ligand recognition. While Trp11 and Tyr38 by itself seem not to dictate GlcNAc binding, their combined effect appears to be essential for the stabilization of the ligand in the binding site. Asn35 on its turn seems to be particularly crucial for the binding of smaller chain-length GlcNAc sequences.

### 5.2.5 CtCBM50 interaction with peptidoglycan sequences

Given that LysM domains are known to recognise PG polysaccharides, studies were carried out to address the binding capabilities of CtCBM50 to such sequences. Co-precipitation assays were performed with PG fractions isolated from *S. aureus* and *Escherichia coli* (*E. coli*) (Figure 5.8). The binding to the insoluble polysaccharide chitin was also analysed. CtCBM50 exhibited strong



**Figure 5.7. Molecular dynamics simulations of CtCBM50 chain-length dependency.** (A) Schematic representation of the position of the GlcNAc<sub>3</sub>, GlcNAc<sub>4</sub>, GlcNAc<sub>5</sub> and GlcNAc<sub>6</sub> ligands into the CtCBM50<sub>AB</sub> structure as well as their binding affinities. The interacting protein residues are represented. The binding subsites (1-6) are also indicated. (B) Close-up of the ligand binding site accommodating GlcNAc<sub>6</sub>. Chain A is depicted in blue and B in yellow. GlcNAc<sub>6</sub> is represented by sticks and coloured by atom type. All residues involved in the complex interface and that established hydrogen bonds or hydrophobic interactions with GlcNAc<sub>6</sub> are represented by balls-and-sticks, coloured by atom type.

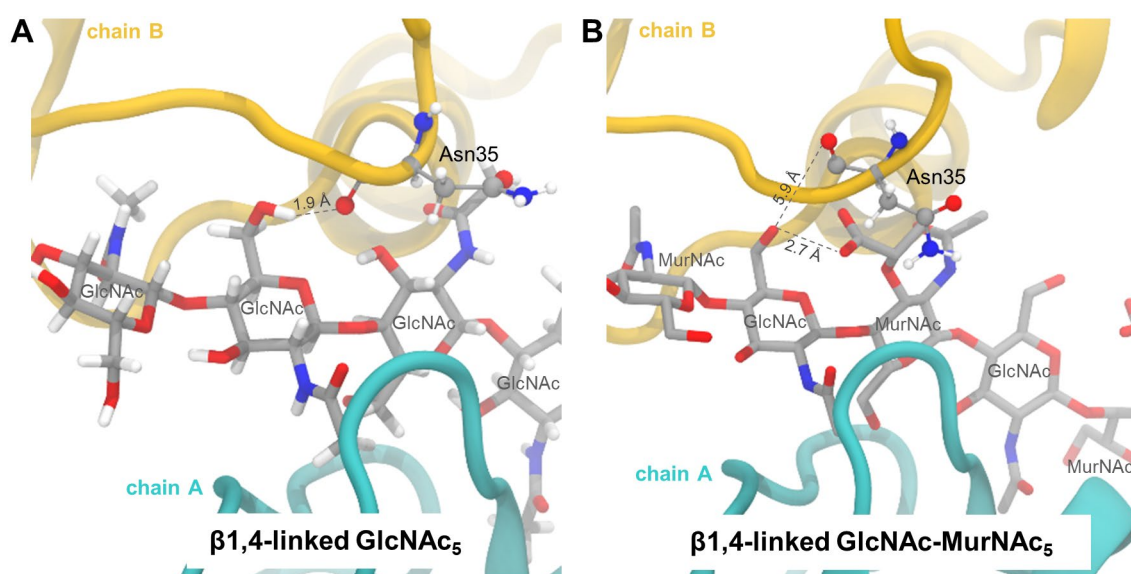


**Figure 5.8. Binding of CtCBM50 to insoluble chitin and peptidoglycan.** Qualitative co-precipitation assays of **(A)** CtCBM50 and **(B)** Peptidoglycan Recognition Protein (PGRP-SA), used as control protein, with insoluble chitin and with peptidoglycan from two different sources, *Staphylococcus aureus* (*S. aureus*) and *Escherichia coli* (*E. coli*). Bound fractions corresponding to the precipitated material: 1) protein control with no ligand; 2) chitin; 3) *S. aureus* peptidoglycan and 4) *E. coli* peptidoglycan; Ligand controls without protein: 5) chitin; 6) *S. aureus* peptidoglycan and 7) *E. coli* peptidoglycan; Unbound fractions corresponding to the supernatants: 8) protein control with no ligand; 9) chitin; 10) *S. aureus* peptidoglycan and 11) *E. coli* peptidoglycan; M) Nzytech protein marker II.

binding to both PGs, thus the CBM does not seem to distinguish between PGs derived from Gram-positive or Gram-negative bacteria (Figure 5.8A). Binding to insoluble chitin was also confirmed. The binding of *S. aureus* Peptidoglycan Recognition Protein (PGRP-SA), used as positive control, to both PGs (Figure 5.8B) as previously described<sup>226</sup>, supported the results observed for CtCBM50.

The molecular determinants for PG binding were also assessed by MD simulations. Oligosaccharide sequences with alternating GlcNAc and MurNAc units, from DP-3 to DP-6, were investigated and the ones showing the most favorable binding energies for each oligosaccharide are discussed (the other tested poses can be found in the supplementary information Figures S5.2 and S5.3 and Tables S5.7 to S5.9). Only the coordinates of CtCBM50 chains A and B were considered for this analysis. As previously shown for GlcNAc ligands, CtCBM50<sub>AB</sub> binding affinities also increased with PG oligosaccharide chain-length (Table 5.4 and Figure S5.2), where five monomers seem to have the optimal binding (Table S5.7). The results also showed that the substitution of GlcNAc residues by MurNAc decreased CtCBM50<sub>AB</sub> affinity. This was also highlighted by the reduced number of hydrogen bonds between the ligands and the CBM along the last 30 ns of MD simulations (Table S5.8). On the one hand, some HO-C6 groups of GlcNAc monomers lost interaction with CtCBM50 chain B residues Asn35 and/or Ile37 and established intra-molecular hydrogen bonds with the carboxylic groups of MurNAc instead (Figure 5.9, Table S5.8). On the other hand, the presence of the bulky lactyl groups at MurNAc units reduces the interfacial contacts with the adjacent chain of CtCBM50 due to steric hindrance. This induces conformational rearrangements of the protein as well as modifications in CtCBM50 chains





**Figure 5.9.** Close-up of inter- and intra-chain hydrogen bonds involving the HO-C6 group of the GlcNAc residue in simulations with GlcNAc and MurNAc-GlcNAc pentasaccharides. (A) CtCBM50-GlcNAc<sub>5</sub> and (B) CtCBM50-MurNAc-GlcNAc<sub>5</sub>. Chain A is depicted in blue and B in yellow. GlcNAc<sub>5</sub> is represented by sticks and coloured by atom type. Asp35 is represented by balls-and-sticks, coloured by atom type.

positioning (Figure S5.3). These structural movements were also evident in the simulations with the PG sequences bound to a single CtCBM50 chain (Table S5.9). For example, the lactyl groups of the GlcNAc-[MurNAc-GlcNAc-MurNAc]-GlcNAc ligand are directed to CtCBM50 chain A and have a lower number of protein-ligand interactions, which supports the reduced affinity with this domain ( $\Delta\Delta G_{\text{binding}}$  of  $13.8 \pm 0.1$  kcal·mol<sup>-1</sup> in relation to chain B). This higher binding affinity for GlcNAc sequences than for PG, has also been reported for *E. faecalis* AtIA LysM<sup>215</sup>. In addition to the amino acid residues mentioned in the previous section, Thr9 of chain A also contributed to the binding to PG oligosaccharides, pointing to an important role in PG recognition by CtCBM50.

The influence of PG's peptide stems on the binding to CtCBM50<sub>AB</sub> was also evaluated. MD simulations showed that the peptides do not interact with the interfacial residues of CtCBM50<sub>AB</sub>

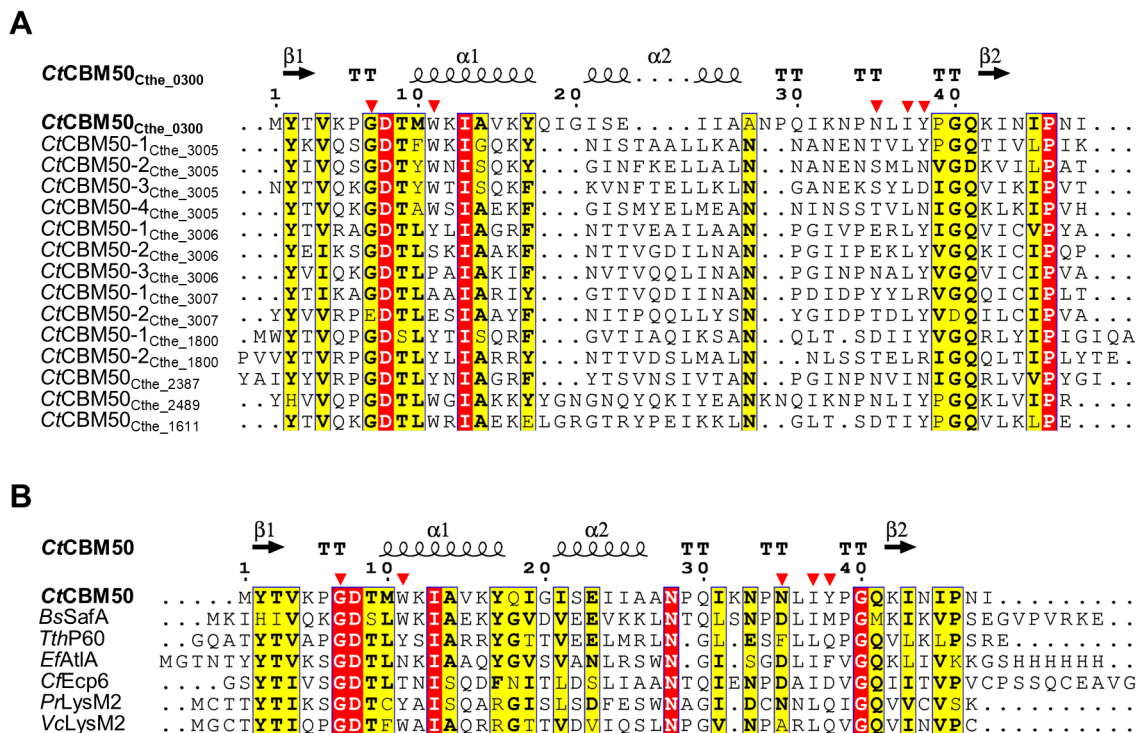
**Table 5.4.** Binding enthalpies and binding free-energies of the peptidoglycan ligands to CtCBM50. Results shown are relative to the corresponding CtCBM50<sub>AB</sub>-GlcNAc complex with the same number of units.

Ligands	$\Delta\Delta H_{\text{binding}}$ (kcal·mol <sup>-1</sup> )	$\Delta\Delta G_{\text{binding}}$ (kcal·mol <sup>-1</sup> )
GlcNAc3	0.00 ± 0.7	0.0 ± 0.1
[GlcNAc-Mur2Ac-GlcNAc]	8.9 ± 0.7	12.9 ± 0.1
[Mur2Ac-GlcNAc-Mur2Ac]	27.8 ± 0.9	28.5 ± 0.1
GlcNAc4	0.0 ± 0.7	0.0 ± 0.1
Mur2Ac-[GlcNAc-Mur2Ac-GlcNAc]	10.2 ± 0.8	11.2 ± 0.1
GlcNAc-[Mur2Ac-GlcNAc-Mur2Ac]	14.8 ± 0.8	14.6 ± 0.1
GlcNAc5	0.0 ± 0.8	0.0 ± 0.1
Mur2Ac-[GlcNAc-Mur2Ac-GlcNAc]-Mur2Ac	4.9 ± 1.0	7.9 ± 0.1
GlcNAc-[Mur2Ac-GlcNAc-Mur2Ac]-GlcNAc	17.3 ± 1.0	16.8 ± 0.1
GlcNAc6	0.0 ± 0.8	0.0 ± 0.1
GlcNAc-Mur2Ac-[GlcNAc-Mur2Ac-GlcNAc]-Mur2Ac	17.9 ± 0.8	15.5 ± 0.1
Mur2Ac-GlcNAc-[Mur2Ac-GlcNAc-Mur2Ac]-GlcNAc	18.7 ± 0.8	18.5 ± 0.1

binding site (Figure S5.4). This observation is in agreement with previous studies that indicated that the monosaccharide residues are essential for LysM recognition, whereas the peptide stems might not be recognized by the protein residues<sup>215,224</sup>.

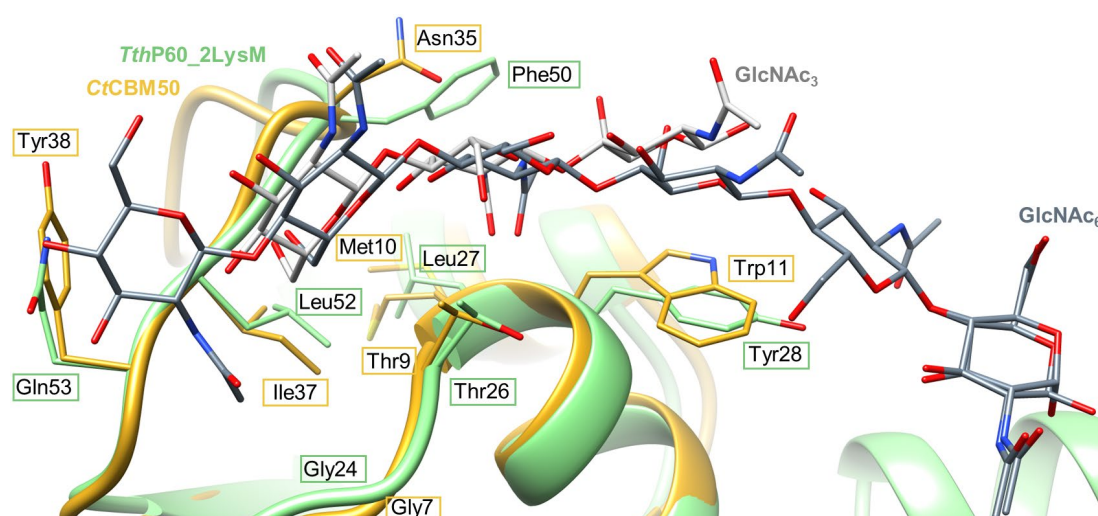
### 5.2.6 *Clostridium thermocellum* family 50 CBMs in the context of LysM domains

Analysis of protein residue conservation of *C. thermocellum* family 50 CBMs (Figure 5.10A) revealed that the interacting residues identified in CtCBM50 (CtCBM50<sub>Cthe\_0300</sub>) were poorly conserved except for Gly7 and Asp8. Trp11 and Tyr38 are only present in 7 and 9 of the 15 CtCBM50s, respectively. Asn35 only occurs in 4 CBMs, while being replaced by other polar or charged amino acids that could still be involved in hydrogen bonding. Ile37, although only found in 5 CBMs, is substituted by a Leu in the remaining proteins, retaining the hydrophobic effect at this position. These observations are not completely unexpected, as the consensus sequence of LysM domains shows that, while the motif is well conserved over the first 16 amino acid residues and slightly less over the last 10, the central region is poorly conserved except for an Asn residue<sup>214</sup>.



**Figure 5.10. Alignment of CBM50 family members.** Primary sequence alignment of (A) *C. thermocellum* family 50 CBMs and (B) CtCBM50<sub>Cthe\_0300</sub> (CtCBM50) with LysMs from other microorganisms: *Bacillus subtilis* (BsSafA)<sup>177</sup>, *Thermus thermophilus* (TthP60)<sup>224</sup>, *Enterococcus faecalis* (EfAtIA)<sup>215</sup>, *Cladosporium fulvum* (CfECP6)<sup>227</sup>, *Pteris rykyuensis* (PrLysM2)<sup>225</sup> and *Volvox carteri* (VcLysM2)<sup>228</sup>. Identity to CtCBM50<sub>Cthe\_0300</sub> is indicated with red and yellow boxes. Residue numbers refer to the corresponding CBM sequence. CtCBM50<sub>Cthe\_0300</sub> secondary structure prediction is presented above. Red triangles identify CtCBM50<sub>Cthe\_0300</sub> residues involved in the interaction with GicNAc<sub>3</sub> ligand. The sequence alignment was generated with Clustal Omega<sup>196</sup> and rendered using Esript server<sup>229</sup>.

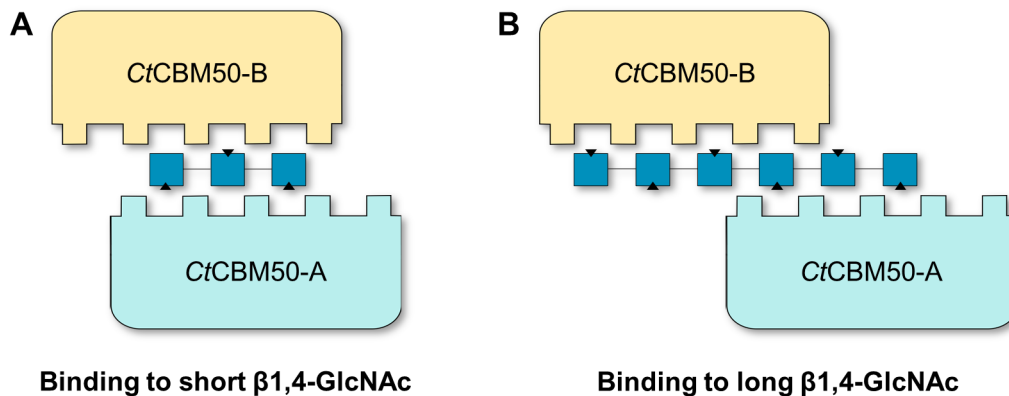
When in comparison with LysMs from other microorganisms (Figure 5.10B), the same trend is observed, with only *Ct*CBM50 interacting residues Gly7 and Asp8 being conserved. By superimposing *Ct*CBM50-GlcNAc<sub>3</sub> (chain B) structure with *Tth*P60\_2LysM bound to a GlcNAc<sub>6</sub>, the interacting residues identified for *Tth*P60 LysM1 are found at the same positions as those of *Ct*CBM50 (Figure 5.11). With a primary sequence identity of 37%, the only identified interacting residues conserved are Thr9 and Gly7. However, the remaining residues that interact with the ligands seem to establish the same type of contacts, with most of the hydrogen bonding also strongly established with the protein's main chain at the same positions, and with the aromatic ring of a Tyr, which corresponds to Trp11 in the *Ct*CBM50 structure.



**Figure 5.11. Superposition of *Ct*CBM50 with *Thermus thermophilus* LysM1.** Chain B of *Ct*CBM50 bound to GlcNAc<sub>3</sub> was superposed with LysM1 of *Tth*P60\_2LysM bound to a GlcNAc<sub>6</sub> (PDB ID 4UZ3)<sup>224</sup>. *Ct*CBM50 is represented as cartoon in yellow and *Tth*P60\_2LysM in green. GlcNAc<sub>3</sub> and GlcNAc<sub>6</sub> are shown as stick models in light grey and dark grey, respectively, and by atom type. Residues of each protein involved in the interactions with its ligand are represented by sticks and coloured by atom type. Alignment was performed using MatchMaker tool from UCF Chimera<sup>40</sup>, with an rmsd value of 0.786.

Similar with what was observed for *Tth*P60\_2LysM, *Ct*CBM50 seems to adopt an interchain assembly behaviour where multiple CBM modules bind to the same GlcNAc oligosaccharide. The results reported here, point to an interchain multivalent assembly induced by longer GlcNAc sequences, where the individual CBM modules bind in a cooperative manner to long ligand chains, but not to short oligosaccharides (Figure 5.12). We hypothesize that individual *Ct*CBM50 molecules bind to longer DP GlcNAc oligosaccharides by rearranging their position, so that each module binds an optimal binding epitope that contains at least two interacting *N*-acetyl groups. Based on the MD calculations with PG oligosaccharides, a similar behaviour could be predicted upon the binding of *Ct*CBM50 to PG sequences, where the CBM chains dislocate to better accommodate the ligand (Figure S5.3).

The modular protein containing *Ct*CBM50 (Figure 5.1) shares 48% of sequence identity with *B. subtilis* SafA, a LysM-containing spore coat assembly protein involved in the formation of the multiprotein coat that encases bacterial spores<sup>177</sup>. Given that *C. thermocellum* also produces



**Figure 5.12. Schematic representation illustrating the hypothesized cooperative binding by CtCBM50 to (A) short and (B) long GlcNAc oligosaccharides.** Experimental data points to an interchain CBM multivalent assembly induced by longer GlcNAc sequences, where individual CBM molecules bind in a cooperative manner to long ligand chains. CtCBM50 modules would bind to longer DP GlcNAc oligosaccharides by rearranging their position, so that each module binds an optimal binding epitope that contains at least two interacting *N*-acetyl groups. Black triangles represent the alternating *N*-acetyl groups.

spores, conferring for instance its elevated resistance to heat and other unfavourable growth conditions<sup>230</sup>, this can point to a possible role of family 50 CBMs in this bacterium, as well as a reasoning for expressing such high number of these CBMs.

### 5.3 Conclusions

With the present work the carbohydrate specificity of *C. thermocellum* family 50 CBMs was assigned to  $\beta$ 1,4-linked GlcNAc sequences, revealing a chain-length dependency with a trisaccharide as a minimum epitope for recognition. Additionally, the first structure of a CtCBM50 was solved and in complex with a GlcNAc trisaccharide, revealing an intermolecular interaction of two CBM molecules with the GlcNAc ligand. Besides binding to chitin and chitin-derived oligosaccharide sequences, peptidoglycan binding was also attested for CtCBM50, although with less affinity. The present results suggest that CtCBM50, acting in a multimodular way, is able to form a ligand binding site comprised of up to 6 binding subsites. Key residues were identified to mediate both chitin and peptidoglycan oligosaccharide recognition by CtCBM50, with Gly7, Asp8, Asn35 and Ile37 residues providing important hydrogen bonding network mediated by main chain atoms; aromatic residues Trp11 and Tyr38 contributing to binding by stacking interactions; and relevant dispersive contacts mediated by Val4, Ile21, Ile25, Pro34 and Pro39 residues. Given the identified residues responsible for CtCBM50 ligand recognition are poorly conserved among LysM domains, our observations point out to a coherent yet adaptable recognition mechanism, dictated by the protein's structural motifs through a critical hydrogen bonding network which results from interactions with main chain atoms and provide a contact surface with the ligand monomers. Furthermore, our results also suggest that ligand binding is favored by the multivalent assembly of CtCBM50 modules, supporting the notion of LysM domains cooperative binding.

The integrative information derived from this work will allow to understand mechanisms of carbohydrate recognition to chitin and peptidoglycan by other members of family 50 CtCBMs, contributing to elucidating their role in *C. thermocellum*. Moreover, the characterization of the carbohydrate recognition by these LysMs, opens the possibility of their potential biotechnological applications.

## 5.4 Experimental procedure

### 5.4.1 Gene cloning, mutagenesis and protein purification

Family 50 CBMs were cloned, expressed and purified using the same procedure as described in section 3.4.2 of Chapter 3. For the structural studies, CtCBM50<sub>Cthe\_0300</sub> was cloned in a pET28a plasmid (Novagen), in which the recombinant protein was generated containing a C-terminal hexa-histidine tag (His-tag). Site-directed mutants were generated using the NZYMutagenesis kit (NZYTech Ltd) according to the manufacturer's instructions using pET28a as template. Primers used to generate the mutant DNA sequences are listed in Table S5.10. Recombinant sequences of all mutant plasmid derivatives were confirmed by sequencing to ensure that only the appropriate mutations were incorporated.

*E. coli* BL21 harbouring the CtCBM50<sub>Cthe\_0300</sub> encoding gene was cultured in LB containing 50 µg/mL kanamycin at 37 °C until mid-exponential phase ( $OD_{600nm} = 0.6$ ), at which point IPTG was added to a final concentration of 1 mM. Cultures were then further incubated for 5h at 37 °C, at 150 rpm in a *Gallenkamp* Orbital *Shaker*. Cells were collected by centrifugation at 5000×g for 15 minutes at 4 °C and the cell pellet resuspended in a 50 mM sodium HEPES buffer, pH 7.5, containing 1 M NaCl, 2 mM CaCl<sub>2</sub> and 10 mM imidazole. CtCBM50<sub>Cthe\_0300</sub> was purified from the cleared cell-lysate by Ni<sup>2+</sup>-immobilized IMAC. The eluted protein fractions were subjected to SDS-PAGE on 13% (w/v) acrylamide gels, stained with Coomassie Brilliant Blue, in order to assess the purity of recombinant proteins. The fractions containing pure protein were pooled and buffer-exchanged into 50 mM MOPS buffer, pH 6, containing 50 mM NaCl and 2 mM CaCl<sub>2</sub>, for protein stability. Amicon 3-kDa molecular-mass centrifugal membranes were used to achieve higher protein concentration.

All proteins were >95% pure as judged by SDS-PAGE and their concentrations determined from their calculated molar extinction coefficient using the ProtParam tool (<http://www.expasy.org/tools/protparam.html>) at 280 nm using a SpectraDrop Micro-Volume Microplate (Molecular Devices, USA).

### 5.4.2 Sources of carbohydrates

Information on the GlcNAc oligosaccharides and sources included in the NGL-microarrays are given in Table S2.1. Insoluble chitin polysaccharide from shrimp shell was purchased from

Sigma-Aldrich. Peptidoglycan samples from *E. coli*<sup>231</sup> and *S. aureus*<sup>232</sup> were kindly provided by Professor Sérgio Filipe (UCIBIO, NOVA).

### 5.4.3 Carbohydrate microarray analysis

The NGL-microarrays results exhibited correspond to the experiments presented in Chapter 3, performed as described in section 3.2.4. The results reported here, correspond to at least two independent experiments, performed with different batches of CBMs.

### 5.4.4 Crystallization and X-ray Diffraction Data Collection

CtCBM50<sub>Cthe\_0300</sub> complex with GlcNAc was produced by overnight incubation of the protein (3.5 mg/mL) with  $\beta$ 1,4-linked GlcNAc trisaccharide (GlcNAc<sub>3</sub>) at 1:2 molar ratio. Crystallization assays were performed using an automated nano-drop dispenser Oryx8 (Douglas Instruments) and commercial screenings JBScreen Classic 2-5 (Jena Bioscience) and Structure 1 & 2 (Molecular Dimensions). 192 conditions with and without ligand were tested using the sitting-drop vapor diffusion method (SWISSCI 'MRC' 2-Drop Crystallization Plates – 96 wells, Douglas Instruments), in a 2  $\mu$ L drop (containing 50% protein). Crystals of the CtCBM50-GlcNAc<sub>3</sub> complex grew through the course of three weeks, at 20 °C, in a crystallization condition composed of 0.1 M sodium acetate buffer, pH 4.6, and 2 M ammonium sulphate. Crystals were harvested using a solution of 0.1 M sodium acetate buffer, pH 4.6, and 2.5 M ammonium sulphate, and then flash-cooled in liquid nitrogen using 30% (v/v) glycerol as cryoprotectant added to the harvesting solution.

X-ray diffraction data from a single crystal of the CtCBM50-GlcNAc<sub>3</sub> complex was collected under a nitrogen stream at 100 K in ID29 beamline at the ESRF (Grenoble, France) to a maximum resolution of 1.45 Å and using X-ray radiation at a fixed wavelength of 0.9677 Å. The CtCBM50-GlcNAc<sub>3</sub> crystal indexed in space group C2, with cell constants  $a = 99.39$ ,  $b = 41.77$ , and  $c = 42.87$  Å and  $\beta = 96.89^\circ$ , corresponding to a calculated Matthews coefficient of 2.28 Å<sup>3</sup>/Da and a solvent content of 46%. Data collection, processing, model building and validation statistics are shown in Table 5.1.

### 5.4.5 Phasing, Model Building, and Refinement

CtCBM50-GlcNAc<sub>3</sub> complex X-ray data sets were processed using MOSFLM<sup>202</sup> and SCALA<sup>203</sup> from the CCP4 suite<sup>204</sup>. Phasing was performed by molecular replacement with Phaser MR<sup>205</sup> from CCP4 using the polypeptide chain of *Volvox carteri* LysM2 structure (PDB ID 5K2L)<sup>228</sup>. Models completion, iterative building, and initial validation were carried out in COOT<sup>206</sup>. Automatic addition of water molecules and restrained refinement of the full models were done using REFMAC5<sup>207</sup>. Structure validation was performed using ProCheck<sup>233</sup> and SfCheck<sup>234</sup>. PRIVATEER<sup>120</sup> was used for the validation of the stereochemistry and conformation of the carbohydrate ligands (Table S5.1). The CtCBM50-GlcNAc<sub>3</sub> asymmetric unit, obtained with final

$R = 17.6\%$  ( $R_{free} = 19.4\%$ ), consists of 3 CBM chains of 46 amino acid residues in chains A and B and 47 in chain C, 2 acetate ions and 1 sulphate, 125 water molecules, and 1 GlcNAc<sub>3</sub> ligand.

Molecular graphics images corresponding to the crystallographic structure were produced using the UCSF Chimera package from the Computer Graphics Laboratory, University of California, San Francisco<sup>40</sup>.

#### 5.4.6 Isothermal titration calorimetry

ITC assays were performed as described previously in Chapter 4, section 4.4.7. Before the experiments, purified CBMs were buffer-exchanged into 50 mM MOPS buffer, pH 6, containing 50 mM NaCl and 2 mM CaCl<sub>2</sub>. Thermodynamic parameters are shown in Table 5.2.

#### 5.4.7 Molecular modelling

The X-ray structure of the CtCBM50 in complex with GlcNAc<sub>3</sub> derived from this work (1.45 Å resolution) was used as a starting geometry for the subsequent modelling studies. For the present analysis, chain C was not considered, since it did not establish significant contacts with GlcNAc<sub>3</sub>. Three different CBM50 systems were considered: 1) the assembly of coordinates for chains A and B, 2) chain A only, and 3) chain B only. The tetrasaccharide (GlcNAc<sub>4</sub>), pentasaccharide (GlcNAc<sub>5</sub>) and hexasaccharide (GlcNAc<sub>6</sub>) ligands were modelled by superposition with a *T. thermophilus* LysM domain structure co-crystallized with a β1,4-linked hexasaccharide (PDB ID 4UZ3, 1.75 Å resolution)<sup>224</sup>. Starting from the GlcNAc<sub>3</sub>, two, three, and four different poses of GlcNAc<sub>4</sub>, GlcNAc<sub>5</sub> and GlcNAc<sub>6</sub> were modelled, respectively. These poses differed in terms of the position of the oligosaccharides in the binding cleft. Figure S5.1 schematizes the position of the different ligands in the CtCBM50<sub>AB</sub> binding interface as well as their computed relative binding energies (see computational details below).

For each of the modelled oligosaccharides, the poses with the highest binding affinities were chosen to model the corresponding ligands composed of alternating MurNAc and GlcNAc units. Initially, only the lactyl group was considered in the MurNAc residues (*i.e.* without the peptide stem). The various molecular systems created were: *i*) CtCBM50<sub>AB</sub> complexed with GlcNAc<sub>3</sub>, GlcNAc<sub>4</sub>, GlcNAc<sub>5</sub>, GlcNAc<sub>6</sub>, MurNAc-GlcNAc-MurNAc, GlcNAc-MurNAc-GlcNAc-MurNAc, GlcNAc-MurNAc-GlcNAc-MurNAc-GlcNAc and MurNAc-GlcNAc-MurNAc-GlcNAc-MurNAc-GlcNAc; and *ii*) CtCBM50-A or CtCBM50-B chains complexed with GlcNAc<sub>3</sub>, GlcNAc<sub>4</sub>, GlcNAc<sub>5</sub>, GlcNAc<sub>6</sub>, MurNAc-GlcNAc-MurNAc, GlcNAc-MurNAc-GlcNAc-MurNAc, GlcNAc-MurNAc-GlcNAc-MurNAc-GlcNAc and MurNAc-GlcNAc-MurNAc-GlcNAc-MurNAc-GlcNAc. For the peptidoglycan fragments we also tested the complementary sequences considering the CtCBM50<sub>AB</sub>: GlcNAc-MurNAc-GlcNAc, MurNAc-GlcNAc-MurNAc-GlcNAc, MurNAc-GlcNAc-MurNAc-GlcNAc-MurNAc and GlcNAc-MurNAc-GlcNAc-MurNAc-GlcNAc-MurNAc.

The physiological protonation state of all protein residues was considered. Each protein:sugar

complex was inserted in a 15 Å rectangular TIP3P water periodic box. Three and six Cl<sup>-</sup> counter-ions were added to neutralize the charge of the monomeric and dual-chain systems, respectively. LEAP program was used to assemble the systems.

#### 5.4.8 Minimization, molecular dynamics simulations and binding energies

The Amber 12 simulation package<sup>235</sup> was used to carry out a two-step minimization and MD simulations. AMBER force field parameters set were used to describe the protein, peptidoglycan-peptide and oligosaccharides (ff99SB, GAFF and Glycam06)<sup>236,237</sup>. Firstly, only the solvent and counter-ions positions were optimized (500 cycles of steepest descent algorithm and 1500 cycles of conjugate gradient algorithm). Secondly, the position of all atoms was optimized (4000 cycles of steepest descent algorithm and 6000 cycles of conjugate gradient algorithm). The system was then equilibrated with an MD simulation of 100 ps in the NVT ensemble and using periodic boundaries conditions. This was followed by 40 ns of production MD simulation in the NPT ensemble. To control the pressure and the temperature of the systems, Berendsen barostat and the Langevin thermostat were used<sup>238</sup>. The systems were simulated at 1 atm and at 328 K (optimal growth temperature of *C. thermocellum*)<sup>239</sup>. Non-bonded interaction pairs were calculated within a 10 Å. Beyond that, Coulomb interactions were treated with the Particle-Mesh Ewald (PME) method<sup>240</sup> and vdW interactions were truncated. The SHAKE algorithm<sup>241</sup> was employed to constrain the bond lengths involving hydrogen atoms, and the equations of motion were integrated with a 2 fs time step using the Verlet leapfrog algorithm. The MD trajectories were saved every 10 ps and analysed with the CPPTRAJ module<sup>242</sup> of Amber 12, allied to the visual molecular dynamics (VMD 1.9.2) program for visualization and image rendering<sup>243</sup>.

The Molecular Mechanics/Poisson Boltzmann Surface Area (MM/PBSA) approach<sup>244</sup> was employed to determine the binding energy of each oligosaccharide and the CtCBM50 models (dual-chain or monomers). A total of 120 structures extracted from the last 30 ns of each MD simulation were used for the analysis. Entropic effects were also determined using normal mode analysis. We present the results as the relative binding energies ( $\Delta\Delta H_{\text{binding}}$  or  $\Delta\Delta G_{\text{binding}}$ ) in respect to the ligand with the smallest number of sugar units. The binding energies of each oligosaccharide and CtCBM50 chain were determined by two different strategies: *i*) using the trajectory of the CtCBM50<sub>AB</sub> complexed to the GlcNAc ligands and deleting the atoms from one of the chains (A or B); and *ii*) using the trajectories from the simulations of CtCBM50-A or CtCBM50-B chains complexed with the GlcNAc oligosaccharides.

#### 5.4.9 Binding to insoluble polysaccharides by co-precipitation assays

The co-precipitation assays were performed essentially as described by Vaz *et al.* 2019<sup>226</sup>. CtCBM50 (0.17 mg/mL) in 50 mM MOPS buffer, pH 6, 50 mM NaCl and 2 mM CaCl<sub>2</sub> were mixed with the polysaccharides suspensions, at 0.2% (w/v) chitin and 0.1% (w/v) peptidoglycans, to a final volume of 200 µL. *S. aureus* Peptidoglycan Recognition Protein (PGRP-SA)<sup>232</sup> (0.3 mg/mL),



kindly provided by Professor Sérgio Filipe (UCIBIO, NOVA), was used as a positive control for the assay. Negative controls of the proteins without the polysaccharides and the polysaccharides without the proteins were also prepared. The mixtures were incubated for 30 minutes at 25 °C at 1000 rpm, followed by 10 minutes centrifugation at 3000 rpm, upon which the supernatants (unbound fractions) were carefully removed. The pellets were washed with 200  $\mu$ L of buffer and centrifuged for 5 minutes at 6000 rpm, followed by a second wash with 200  $\mu$ L of buffer and centrifuged for 2 minutes at 13,200 rpm. The pellets (bound fractions) and 30  $\mu$ L of the unbound fractions were then mixed with 30  $\mu$ L of 2x SDS loading buffer (10% (w/v) SDS containing 10% (v/v)  $\beta$ -mercaptoethanol) and boiled for 5 minutes, after which the samples were centrifuged for 3 minutes at room temperature, at 13,200 rpm, and the supernatants were recovered (20  $\mu$ L) into a fresh tube. The bound and unbound fractions full supernatant volume was loaded on a 13% SDS-PAGE acrylamide gel and the resulting bands were visualized by Coomassie Blue Staining.

## 5.5 Work contributions

All work related to the results reported here, were executed by the author of this thesis, except for the molecular modelling and dynamics simulations which were performed by Dr. Natércia Brás (UCIBIO, REQUIMTE, Porto), upon discussion and planning with the author. PhD student Raquel Costa has also contributed for the experimental work, as part of her Master thesis work. Mutagenesis and ITC studies were planned and performed with the assistance of Dr. Benedita Pinheiro (UCIBIO, NOVA). Co-precipitation assays with insoluble PG and chitin were performed upon discussion and planning with Professor Sérgio Filipe (UCIBIO, NOVA) and PhD student Gonçalo Covas.



# CHAPTER 6

---

**ASSIGNING THE CARBOHYDRATE SPECIFICITY OF  
*RUMINOCOCCUS FLAVEFACIENS* FAMILY 13 CBMs:  
RECOGNITION OF PECTIC ARABINANS  
BY A NOVEL CBM13**



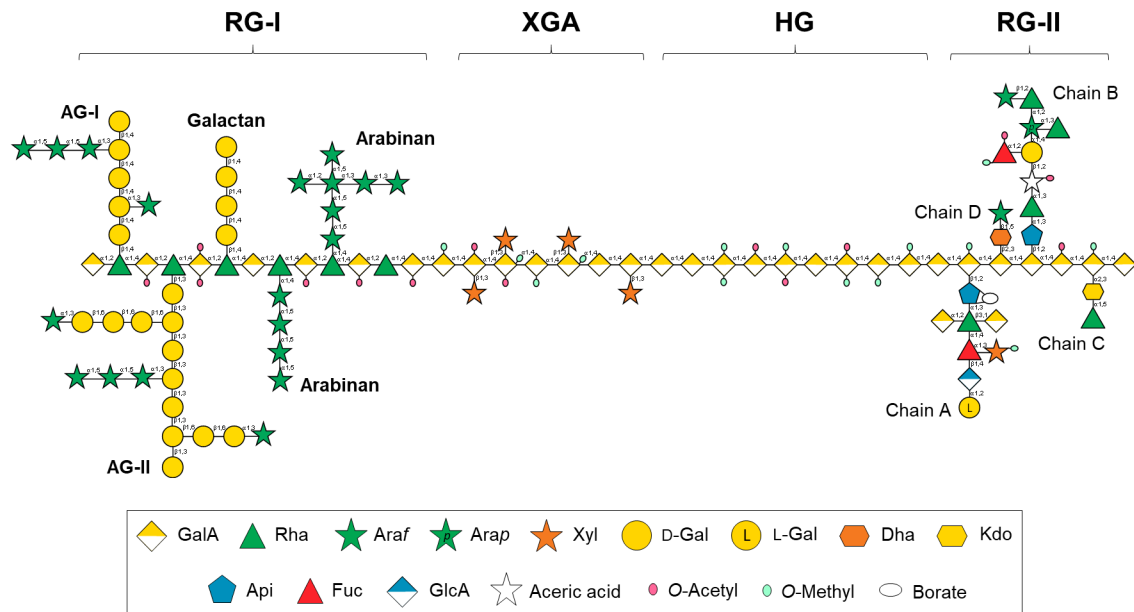
## 6 Assigning the carbohydrate specificity of *Ruminococcus flavefaciens* family 13 CBMs: Recognition of pectic arabinans by a novel CBM13

### 6.1 Introduction

Plant cell walls are composed in its majority by highly diverse and complex polysaccharides, including cellulose, glucans, hemicelluloses and pectins<sup>61</sup> (Chapter 1, section 1.1 and Figure 1.1). Pectins are one of the major plant cell wall components and the most structurally complex and heterogeneous group of polysaccharides<sup>4,10,245</sup>. While its composition and fine structure vary depending on the plant source and tissue, and even on the extraction conditions applied, these galacturonic acid-rich polysaccharides are structurally divided into three major groups: homogalacturonan (HG), rhamnogalacturonan I (RG-I), rhamnogalacturonan II (RG-II) (Figure 6.1). The HG backbone can be substituted by  $\beta$ 1,3-linked xylose comprising the xylogalacturonan (XGA) domain<sup>4,10,245,246</sup>. The RG-I, is the second most predominant group of pectic polysaccharides after  $\alpha$ 1,4-linked HG, and is composed of a backbone of alternating units of  $\alpha$ 1,2-linked rhamnose and  $\alpha$ 1,4-linked galacturonic acid units substituted with neutral side chains, such as arabinans and galactans<sup>10,61,245,247</sup>. Galactans are comprised of  $\beta$ -linked galactopyranose units, which can be partially branched with galactopyranose and arabinofuranose units, comprising arabinogalactan domains (AG-I or AG-II)<sup>245</sup>. Pectic arabinans are comprised of an  $\alpha$ 1,5-linked arabinofuranose backbone, that can be ramified at position O3 and/or O2 by single arabinosyl residues or short side chains<sup>10,245,248</sup>.

Several cellulolytic bacteria evolved to degrade the recalcitrant plant cell wall polysaccharides by employing an extracellular multi-protein complex machinery, the Cellulosome, where the catalytic modules (CAZymes) have non-catalytic CBMs appended. CBMs play a crucial role in enhancing the catalytic efficiency of the enzymes, hence contributing for the biodegradation of plant polysaccharides by the bacteria (reviewed in more detail in Chapter 1, section 1.2.1)<sup>27,58</sup>. With sequencing of bacterial genomes, information on newly identified CBMs deposited and organized by sequence similarity into different families in the CAZy database<sup>22</sup> is continually growing, opening new research for their characterization and structure-function analysis.

CBMs assigned to family 13 belong to the  $\beta$ -trefoil fold family 2, also designated ricin B-like domains after being first identified in several plant lectins, such as ricin toxin B-chain<sup>23,249</sup>. These modules are comprised of approximately 150 amino acid residues, in which 40-52 residues appear as a 3-fold internal repeat, resulting in a pseudo-3-fold axis that confers a globular structure to the protein<sup>23,169</sup>. These structural repeats comprise 3 subdomains, termed  $\alpha$ ,  $\beta$  and  $\gamma$ , each one containing a putative pocket-like ligand binding site. Family 13 CBMs generally bind



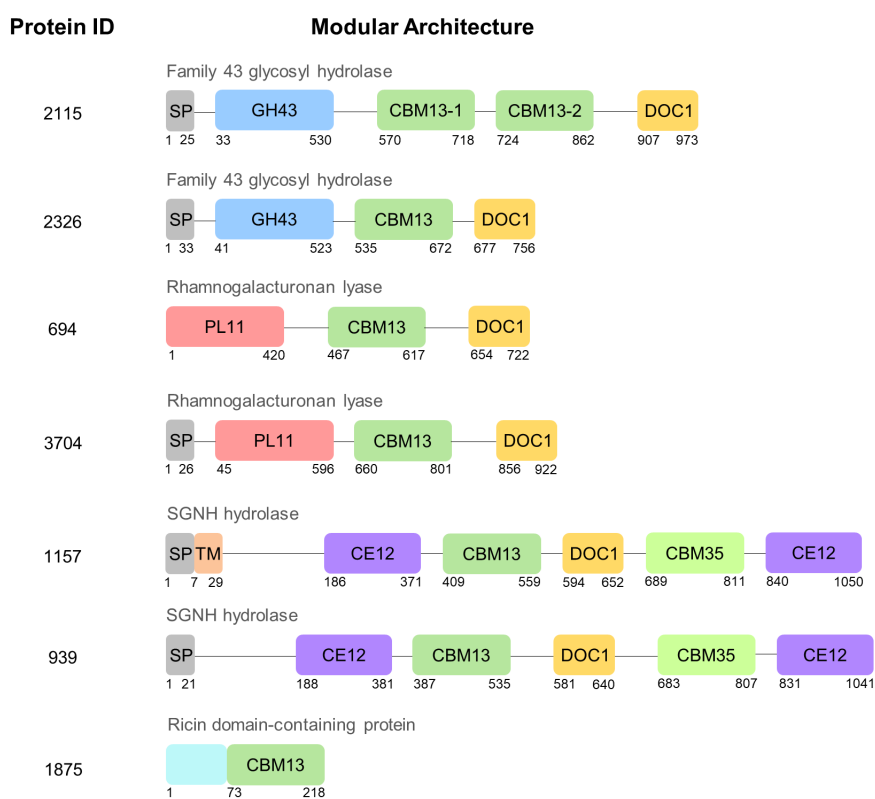
**Figure 6.1. Schematic representation of the major pectic polysaccharide structural domains.** The different types of polysaccharide structures that are present in pectins are represented: RG-I, rhamnogalacturonan I, including arabinan, galactan and arabinogalactans type I and II (AG-I and AG-II, respectively); XGA, xylogalacturonan; HG, homogalacturonan; and RG-II, rhamnogalacturonan II. The polysaccharide structures are based on Mohnen *et al.* 2008<sup>10</sup>. Structures of RG-I side chains are representative and not comprehensive. The monosaccharide symbolic representation used was according to the updated SNFG<sup>1</sup>.

one or two monosaccharide residues within a polysaccharide, hence they are usually classified as Type C CBMs<sup>169</sup>.

CBMs from family 13 exhibit a variety of carbohydrate-binding specificities and are found in several CAZymes, such as  $\beta$ -xylanases,  $\beta$ -glucanases,  $\alpha$ -galactosidases and  $\alpha$ -arabinofuranosidases, from GHs families 10, 11 and 43, but also in polysaccharide lyases (PL) such as rhamnogalacturonan lyases, carbohydrate esterases (CE) and glycosyltransferases<sup>22,169</sup>. *Streptomyces olivaceoviridis* E-86 endo- $\beta$ 1,4-xylanase SoXyl10 possesses a GH10 and CBM13 (SoCBM13) that binds  $\beta$ 1,4-xylose oligosaccharides<sup>250</sup>. *Streptomyces avermitilis*  $\beta$ -L-arabinopyranosidase SaArap27, has a GH from family 27 and a CBM13 (SaCBM13) which binds arabinopyranose monomers<sup>251</sup>. *Clostridium thermocellum* exo- $\beta$ 1,3-galactanase 1,3Gal43A, on its turn possesses a GH43 and a CBM13 (CtCBM13<sub>Cthe\_0661</sub>) that bind  $\beta$ -galactose oligosaccharides<sup>156</sup> (Chapter 3, section 3.2.3.4).

The recent genome sequencing of *Ruminococcus flavefaciens* FD-1 (henceforward referred to only as *R. flavefaciens*), a cellulolytic ruminal bacteria found in the digestive tract of bovines, has revealed one of the largest collection of celulosome-associated proteins among known fibre-degrading bacteria<sup>14,27</sup>, including a significant number of family 13 CBM sequences (Figure 3.2, Chapter 3) for which carbohydrate binding specificities and mechanisms of ligand recognition are awaiting elucidation. *R. flavefaciens* possesses 8 family 13 CBM sequences,

found associated with different CAZymes, from GHs from family 43 (GH43), PL from family 11 (PL11) and a CE from family 12 (CE12) (Figure 6.2).



**Figure 6.2. Modular architecture of *R. flavefaciens* proteins containing family 13 CBMs.** Schematic representation of the assigned CBMs, CAZymes and dockerins is shown. Family 13 and family 35 CBMs are coloured in shades of green and the associated enzymes and domains are coloured according to sequence identity: GH43, family 43 glycoside hydrolases; PL11, family 11 polysaccharide lyase; CE12, family 12 carbohydrate esterases; and DOC1, type 1 dockerin. The modular proteins are identified by an in-house protein ID (left panel). Sequence similarity search was performed using Basic Local Alignment Search Tool<sup>221</sup> and Conserved Domain Database<sup>252</sup> from NCBI, Uniprot<sup>222</sup> and InterProScan<sup>223</sup>.

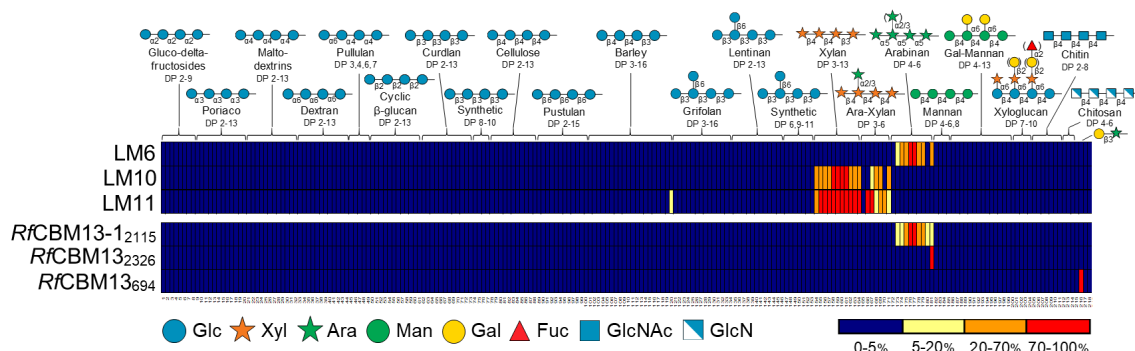
The carbohydrate microarray analysis of family 13 CBMs of *R. flavefaciens* described in Chapter 3, revealed the binding specificity for 3 of these *Rf*CBMs (Figure 3.7), to  $\alpha$ -arabinofuranose sequences in pectic arabinan-derived oligosaccharides and to  $\beta$ -galactose containing oligosaccharides. To our knowledge, binding to arabinofuranose oligosaccharides is yet to be described for family 13 CBMs.

In the work presented in this chapter, the binding specificity of *R. flavefaciens* family 13 CBMs was further explored, with determination of the structure of *Rf*CBM13-1<sub>2115</sub> along with identification of the potential molecular determinants of arabinan recognition using site-directed mutagenesis and interaction studies by isothermal titration calorimetry (ITC). The results reported here will promote the characterization of *R. flavefaciens* family 13 CBMs and their functional role, as well as to contribute for the information deposited into the CAZy database.

## 6.2 Results and Discussion

### 6.2.1 Ligand specificity of *R. flavefaciens* family 13 CBMs

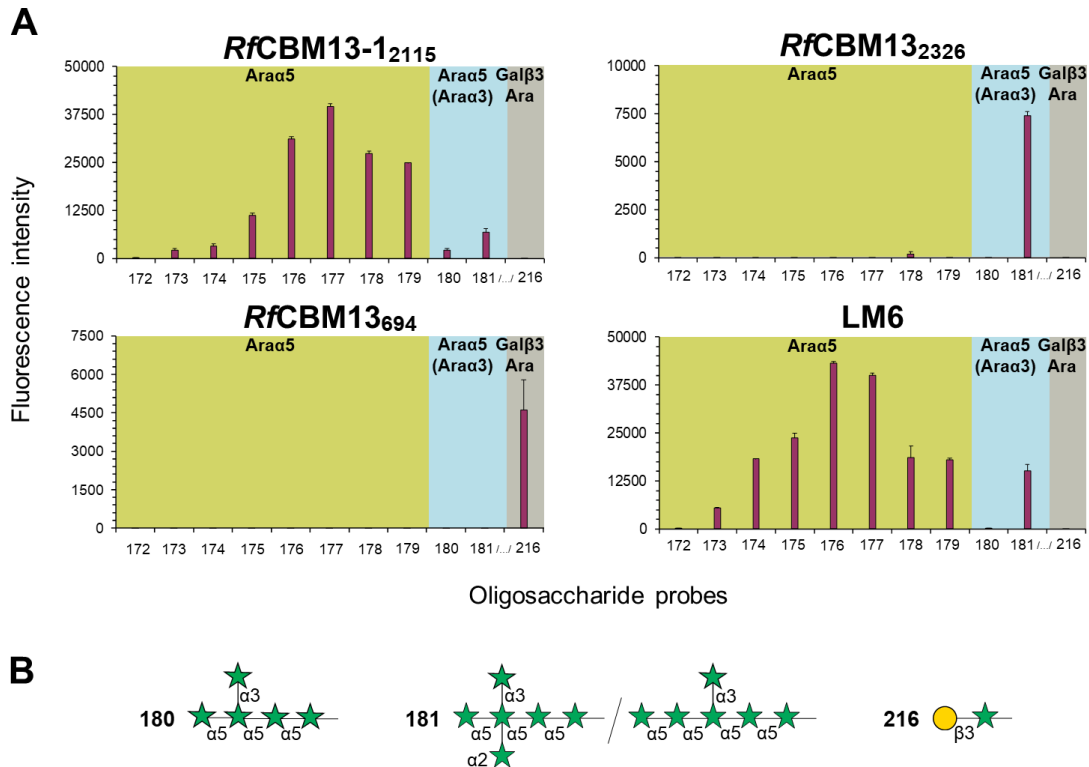
Aiming to assign the carbohydrate binding specificity at oligosaccharide level, family 13 of *RfCBMs* (Figure 6.2 and Tables S3.1 and S3.2 in Chapter 3) were analysed using the NGL-microarray comprised of diverse sequence-defined oligosaccharides, including  $\beta$ -xylans,  $\beta$ -arabinoxylans,  $\beta$ -mannans,  $\beta$ -galactomannans,  $\alpha$ -arabinans, xyloglucans and a pectin-related disaccharide Gal $\beta$ 1,3Ara (Tables S2.1, Chapter 2), as presented in Chapter 3. Binding patterns were obtained for 3 *RfCBMs* 13, which revealed the carbohydrate binding to  $\alpha$ -arabinofuranose (Ara $f$ ) and  $\beta$ -galactopyranose oligosaccharide sequences (Figures 6.3 and 6.4 and Table S3.8 in Chapter 3). The binding patterns of *RfCBMs* 13 were supported by the specific binding of the anti- $\alpha$ 1,5-arabinan monoclonal antibody LM6 used as a protein control of the microarray analysis (Figures 6.3 and 2.3 and Table S2.3 in Chapter 2).



**Figure 6.3. Carbohydrate microarray analysis of *R. flavefaciens* family 13 CBMs.** The microarrays included 219 NGL-oligosaccharides of a wide DP range of linear and branched oligosaccharide-NGL probes of  $\alpha$ - and  $\beta$ -glucans<sup>32</sup>,  $\beta$ -xylans,  $\alpha$ -arabinans,  $\beta$ -mannans, xyloglucans, chitin and chitosan (top panel). Carbohydrate sequence information on these probes is shown in Chapter 2, Table S2.1. Proteins for which binding was obtained are presented at the left: monoclonal antibodies LM6, LM10 and LM11, used in the validation of the microarrays (upper panel); and *RfCBMs* 13 (bottom panel). The relative binding intensities were calculated as the percentage of the fluorescence signal intensity at 5 fmol given by the probe most strongly bound by each protein (normalized as 100%). Numerical scores are given in Chapter 3, Table S3.8.

The 3 *R. flavefaciens* family 13 CBMs exhibited distinct binding patterns. While *RfCBM13-12115* showed main binding to linear  $\alpha$ 1,5-linked Ara $f$  sequences, *RfCBM132326* bound exclusively to the probe presenting a mixture of branched  $\alpha$ 1,2(1,3) Ara $f$  sequences with DP-6 (probe 181). Interestingly, *RfCBM13694* didn't show binding to any of the arabinofuranose probes, but bound, albeit weakly, to the NGL probe of the disaccharide Gal $\beta$ 1,3Ara (probe 216). Although it is described that CBMs from family 13 generally recognize one or two monosaccharide units within its ligands, *RfCBM13-12115* binding pattern hints a chain-length dependency to linear  $\alpha$ 1,5 Ara $f$  sequences from DP-3 up to DP-7. The restricted binding of *RfCBM132326* to the branched arabinan probe presenting both  $\alpha$ 1,2 and  $\alpha$ 1,3 Ara $f$  branches (probe 181), and not the sequence with only  $\alpha$ 1,3 Ara $f$  branch at the penultimate non-reducing end arabinose (probe 180), points to a crucial role of the  $\alpha$ 1,2-linked Ara $f$  or the  $\alpha$ 1,3 Ara $f$  branch at a more internal position, on the specificity of this CBM. *RfCBM13-12115* and *RfCBM132326* binding specificity to Ara $f$  sequences is in





**Figure 6.4. Binding analysis of *R. flavefaciens* family 13 CBMs arabinan-derived oligosaccharides included in the carbohydrate microarrays. (A)** The binding signals of each *RfCBM13-1* is depicted as means of fluorescence intensities of duplicate spots at 5 fmol of oligosaccharide probe arrayed (with error bars) and are representative of at least two independent experiments. The  $\alpha$ 1,5 arabinan-specific monoclonal antibody LM6 used in the validation of the microarrays was included as a control. The microarrays here represented included 8 linear  $\alpha$ 1,5-linked Ara NGL-oligosaccharides from DP-2 to DP-9, 2  $\alpha$ 1,5-linked Ara sequences with  $\alpha$ 1,2(1,3) branches of DP-5 and DP-6, and the Gal $\beta$ 1,3Ara disaccharide. **(B)** The sequences of the branched arabinose probes and the Gal $\beta$ 1,3Ara disaccharide are depicted indicating the position in the binding charts.

accordance with the associated family 43 GHs (Figure 6.2), for which an  $\alpha$ -arabinofuranosidase is a major activity reported on CAZy database<sup>22,26</sup>. The binding detected with *RfCBM13*<sub>694</sub> to the disaccharide with a  $\beta$ 1,3-linked Gal residue at the non-reducing end (probe 216) points to recognition of a  $\beta$ -Gal epitope in pectic polysaccharides, possibly found in galactan or arabinogalactan branches of RG-I (Figure 6.1). This is in accordance with the strong binding of this CBM to lupin and potato pectic galactans and to soy bean rhamnogalacturonan in the pectin-related polysaccharide microarrays presented in Chapter 3 (Figure S3.3). The recognition of pectic sequences by this CBM is not unexpected as it is associated with a family 11 PL (Figure 6.2), annotated to have activity on rhamnogalacturonan<sup>22</sup>.

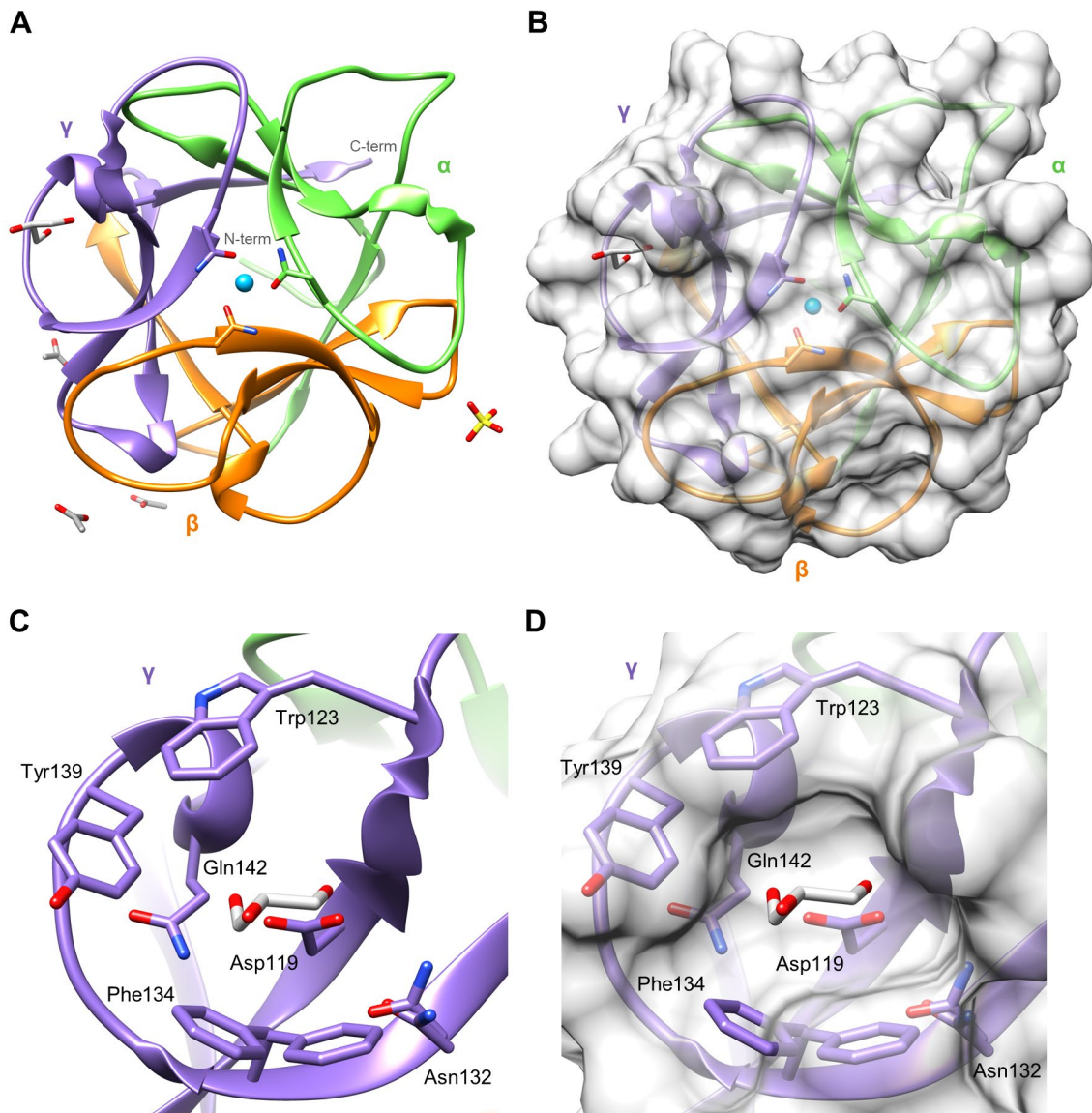
Given that *RfCBM13-1*<sub>2115</sub> (henceforward designated as *RfCBM13-1*) showed strong binding intensity in the microarrays exhibiting an unpredicted chain-length dependency, interest arose to structurally characterize its carbohydrate recognition, which will be explored in the following sections.

## 6.2.2 Crystal structure of *RfCBM13-1* revealing the putative binding sites

Crystallization assays were first carried out for *RfCBM13-1* in its unbound state and its structure was then solved at a resolution of 1.80 Å (Figure 6.5). Statistics of data processing and model refinement and validation are presented in Table 6.1. *RfCBM13-1* 3D structure presented the typical  $\beta$ -trefoil fold of family 13 CBMs, composed by the subdomains  $\alpha$ ,  $\beta$ , and  $\gamma$  (Figure 6.5A). At the centre of the trefoil a magnesium atom is coordinated by the side chains of Asn34, Asn82 and Asn130, and main chain atoms of Ile35, Val83 and Val131, of subdomain  $\alpha$ ,  $\beta$ , and  $\gamma$ , respectively. As observed for other CBMs of this family, the protein surface revealed 3 putative binding sites, one in each subdomain, with a glycerol molecule observed in the putative binding site  $\gamma$  (Figure 6.5B). The side chains of residues Asp119, Trp123, Asn132, Phe134, Tyr139 and Gln42 were identified as potential key residues for ligand recognition, forming a putative binding pocket in subdomain  $\gamma$  where the glycerol molecule was accommodated (Figure 6.5C and D).

Aiming to structurally characterize the ligand recognition by *RfCBM13-1*, co-crystallization assays were initially carried out for the CBM in complex with *Araf* trisaccharide ( $Ara_3$ ), as this was the minimum epitope recognised by the CBM in the microarrays. Although crystals were obtained and several data sets were collected, the structures solved evidenced unexplained residual electron density in up to 2 of the 3 putative binding sites that could not be attributed to the  $Ara_3$  trisaccharide or to individual *Araf* monomers (Figure S6.1). As the putative binding sites are exposed to the solvent channels, the ligand could be disordered, impairing unequivocal positioning in the electron density. Co-crystallization using *Araf* disaccharide  $Ara_2$ , a smaller ligand, was also attempted, however, the structures solved revealed the same untraceable electron density segments in the putative binding sites. Although family 13 CBMs generally recognize one or two monosaccharide units, co-crystallization and soaking experiments were also pursued with a longer chain-length ligand using the hexasaccharide  $Ara_6$ , as foreseen in the microarrays results. Although overall up to 576 conditions were tested, either in manual set ups or using the automated nanodrop dispenser (crystallization robot), and complete X-ray diffraction data were collected from several *RfCBM13-1* crystals grown in different conditions, the unidentified electron density segments were always present at the putative binding sites and could not be unequivocally attributed to the ligands used for crystallization.

The unexplained and unmodelled residual electron density segments observed in the putative binding sites of *RfCBM13-1* could result from simple ligand disorder (due to solvent exposition) or might be attributed to different binding modes of recognition by this CBM, as seen previously for other CBMs from this family<sup>156,253</sup>. Multiple binding modes would enhance the activity of the associated enzymes by recruiting a variety of potential ligands or accommodating their structures in different manners. For instance, *RfCBM13-1* could be recognising  $Ara_3$  in a non-reducing end-in manner and simultaneously binding to the middle *Araf* unit, as observed for *CtCBM13*<sub>Cthe\_0661</sub><sup>156</sup>, resulting in variable partial occupancy in the crystal lattice that culminates in disordered electron density.



**Figure 6.5. Ribbon representation of *RfCBM13-1* three-dimensional structure.** (A) and (B) cartoon and surface representation of the overall structure of *RfCBM13-1* exhibiting the typical  $\beta$ -trefoil fold of family 13 CBMs, exhibiting the 3 subdomains  $\alpha$  (green),  $\beta$  (orange) and  $\gamma$  (purple). A magnesium ion (blue sphere) is coordinated at the centre of the trefoil. A sulphate ion (yellow), 3 acetate ions (grey) and a glycerol molecule (white) present in the structure are represented; (C) and (D) close-up view on subdomain  $\gamma$  putative binding site, showing the bound glycerol molecule and the CBM's amino acid side chains that might be determinant in ligand recognition. The surface representation evidences the putative binding pocket where the glycerol has bound. Molecules and amino acid residues' side chains are shown as sticks coloured by atom type. The magnesium ion is represented by a blue sphere.

### 6.2.3 Characterization of *RfCBM13-1*-ligand interaction

Although the crystal structure of the *RfCBM13-1*-ligand complex couldn't be accomplished, the unliganded structure allowed the identification of potential binding sites and key amino acid residues for ligand recognition. This structural information was used to design and generate by site-directed mutagenesis mutant derivatives of potentially relevant amino acids. The analysis of the mutations effect was carried out by thermodynamic characterization of *RfCBM13-1* carbohydrate interaction using ITC.

**Table 6.1. X-ray diffraction and structure refinement parameters and statistics for *RfCBM13-1*.**

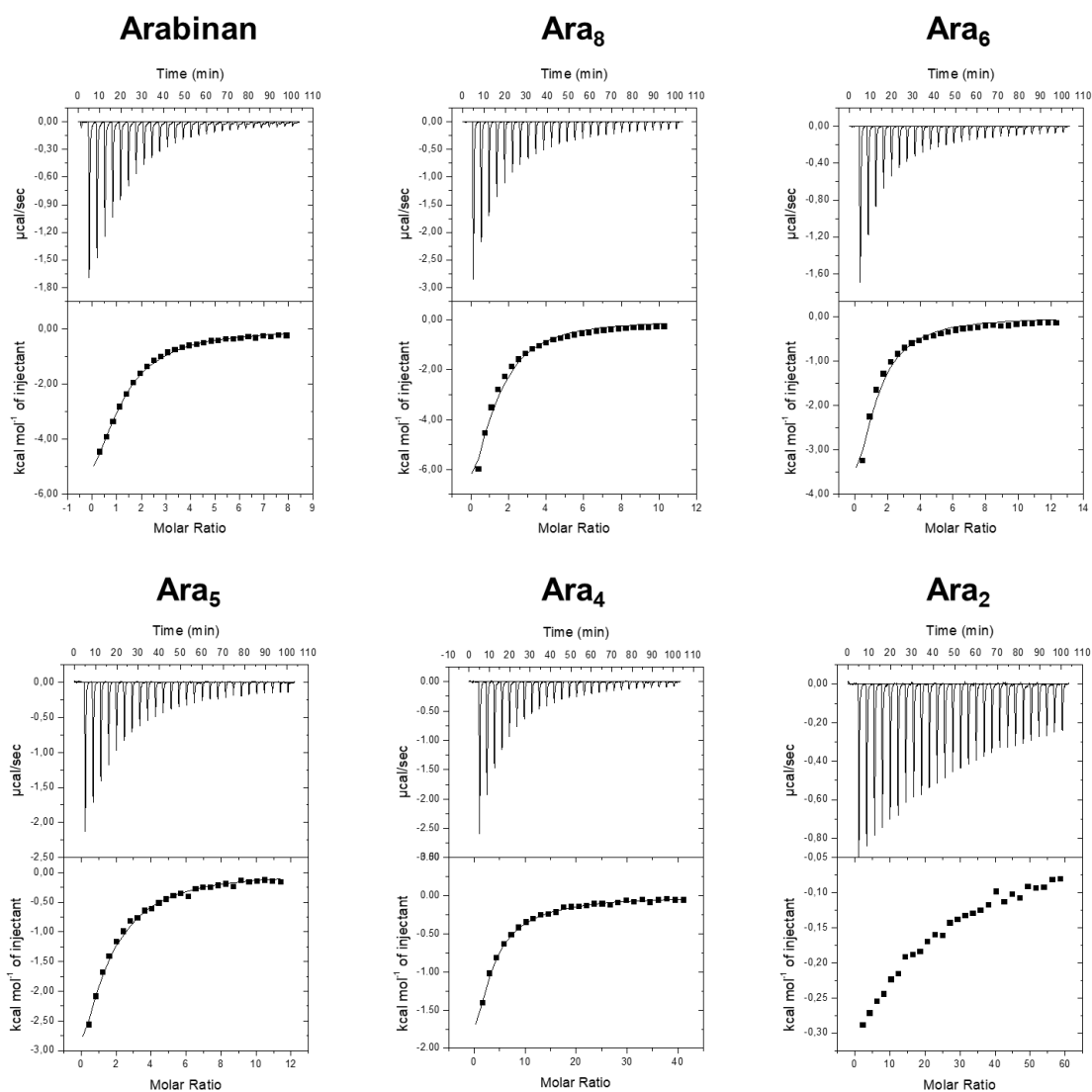
<b>Data collection</b>	
Beamline	SLS, X06DA - PXIII
Space Group	$P4_12_12$
<b>Cell parameters</b>	
$a, b, c$ (Å)	65.43, 65.43, 102.79
$\alpha, \beta, \gamma$ (°)	90.00, 90.00, 90.00
Wavelength, Å	0.9795
Resolution of data (outer shell), Å	21.34-1.80 (1.84-1.80)
Total number of reflections (outer shell)	213481 (13237)
Number of unique reflections (outer shell)	21423 (1241)
$R_{\text{pim}}$ (outer shell), <sup>a</sup>	0.031 (0.202)
$R_{\text{merge}}$ (outer shell), <sup>b</sup>	0.072 (0.480)
Mean $I/\sigma(I)$ (outer shell)	23.5 (5.5)
CC(1/2) (outer shell)	0.99 (0.95)
Completeness (outer shell), %	99.9 (100.0)
Redundancy (outer shell)	10.0 (10.7)
<b>Structure refinement</b>	
No. of protein atoms	1173
No. of solvent waters	237
Resolution used in refinement, Å	21.34-1.80
No. of reflections	19251
$R_{\text{work}} / R_{\text{free}}$ <sup>c</sup>	0.159 / 0.197
rms deviation bonds (Å)	0.012
rms deviation angles (°)	8.435
rms deviation chiral volume (Å <sup>3</sup> )	0.086
<b>Avg B factors (Å<sup>2</sup>)</b>	
Main chain	18.49
Side chain	22.96
Magnesium ion	14.43
Sulphate ion	76.31
Acetate ion 1	64.50
Acetate ion 2	51.70
Acetate ion 3	53.50
Glycerol	44.03
Water molecules	36.87
<b>Ramachandran statistics</b>	
<i>favored</i>	128
<i>allowed</i>	7
<i>generously allowed</i>	0
<i>forbidden</i>	0

<sup>a</sup>  $R_{\text{p.i.m.}} = \left( \frac{\sum_{hkl} \sqrt{\frac{n}{n-1}} \sum_{j=1}^n |I_{hkl,j} - \langle I_{hkl} \rangle|^2}{\sum_{hkl} \sum_j I_{hkl,j}} \right)^{1/2}$ , where  $\langle I_{hkl} \rangle$  is the average of symmetry-related observations of a unique reflection.

<sup>b</sup>  $R_{\text{sym}} = \left( \frac{\sum_{hkl} \sum_j |I_{hkl,j} - \langle I_{hkl} \rangle|^2}{\sum_{hkl} \sum_j I_{hkl,j}} \right)^{1/2}$ , where  $\langle I_{hkl} \rangle$  is the average of symmetry-related observations of a unique reflection.

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

ITC measurements with arabinan polysaccharide and *Araf* oligosaccharides of varying DPs (*Ara*<sub>2</sub> to *Ara*<sub>8</sub>), corroborated the binding specificity of *RfCBM13-1* to pectic arabinan and the increase of affinity with the chain-length of *Araf* sequences (Figure 6.6 and Table 6.2), as observed in the carbohydrate microarrays (Figure 6.4). The affinity displayed by the CBM to *Araf* sequences below DP-4 was relatively weak (with a  $K_a$  of  $0.47 \times 10^4 \text{ M}^{-1}$  for *Ara*<sub>4</sub>), although still detectable for DP-2 (below  $10^3 \text{ M}^{-1}$ ). The  $K_a$  of the interaction increased significantly to DP-6 ( $2.64 \times 10^4 \text{ M}^{-1}$ ) and stabilised up to DP-8 ( $K_a$  of  $2.87 \times 10^4 \text{ M}^{-1}$ ), with similar affinities as to the arabinan polysaccharide (with a  $K_a$  of  $2.77 \times 10^4 \text{ M}^{-1}$ ). The ability of the CBM to bind, albeit weakly, to *Ara*<sub>2</sub> is in accordance



**Figure 6.6. Isothermal calorimetry titrations of binding of *RfCBM13-1* to  $\alpha$ 1,5-linked arabinan sequences.** Chain-length dependency analysis of *RfCBM13-1* to arabinan polysaccharide and linear  $\alpha$ 1,5-arabinose oligosaccharide sequences with DP-2 to DP-8; The top portion of each panel shows the raw power data while the bottom parts show the integrated and heat of dilution corrected data. The solid lines show the non-linear curve fits to a one site binding model with the stoichiometry fixed at 1. Thermodynamic parameters are given in Table 6.2.

with the *RfCBM13-1* structure that shows a putative binding cleft that could accommodate up to 2 ligand monomers (Figure 6.5B). However, the higher affinity to longer chain-length oligosaccharides points to important additional interactions being established between the surface of the CBM and the remaining sequence of the ligand that is not accommodated in the binding cleft. These results suggest that up to 6 *Araf* units might be required for the CBM function in accordance with the lowest entropic contribution for the *RfCBM13-1* binding to *Ara*<sub>6</sub>, reflecting the most favourable interaction.

Given that aromatic and polar amino acid residues are known to play important roles in the binding recognition by CBMs, a set of amino acid residues present in *RfCBM13-1* predicted binding sites

**Table 6.2. Thermodynamic parameters of the binding of *RfCBM13-1* wild type and its mutant derivatives to polysaccharides and oligosaccharides.**

<i>RfCBM13-1</i> variant	Ligand	$K_a \times 10^4$ (M <sup>-1</sup> )	$\Delta G$ (kcal.mol <sup>-1</sup> )	$\Delta H$ (kcal.mol <sup>-1</sup> )	$T\Delta S$ (kcal.mol <sup>-1</sup> )	$n$
WT	Arabinan	2.77 ± 0.24	-6.01	-14.55 ± 1.27	-8.44	1.00 ± 0.07
	Arabinoxylan (Rye)					
	Galactomannan (Carob)			No binding		
	Galactomannan (Guar)					
	Ara <sub>8</sub>	2.87 ± 0.22	-6.08	-12.73 ± 0.39	-6.65	1.12 ± 0.00
	Ara <sub>6</sub>	2.64 ± 0.25	-6.03	-7.67 ± 0.29	-1.64	1.03 ± 0.00
	Ara <sub>5</sub>	1.07 ± 0.39	-5.50	-8.46 ± 0.14	-2.96	1.04 ± 0.00
	Ara <sub>4</sub>	0.47 ± 0.01	-5.02	-13.93 ± 0.17	-8.91	1.00 ± 0.00
			Weak Interaction (<10E3)			
Trp38Ala	Arabinan	1.02 ± 0.07	-5.47	-5.75 ± 0.85	-0.28	1.12 ± 0.15
	Ara <sub>8</sub>	2.85 ± 0.22	-5.74	-5.52 ± 0.28	-0.56	1.03 ± 0.10
	Ara <sub>6</sub>	2.26 ± 0.26	-5.94	-6.30 ± 0.28	-0.36	1.09 ± 0.00
Gln86Ala	Arabinan	2.48 ± 0.07	-6.01	-13.94 ± 0.26	-7.93	1.18 ± 0.02
	Ara <sub>8</sub>	1.21 ± 0.27	-5.56	-17.13 ± 0.19	-11.57	1.09 ± 0.00
	Ara <sub>6</sub>	2.87 ± 0.85	-6.40	-5.14 ± 0.54	1.26	0.93 ± 0.08
Phe134Ala	Arabinan			Weak Interaction (<10E3)		
	Ara <sub>8</sub>			No binding		
	Ara <sub>6</sub>					
Asp119Ala	Arabinan			No binding		
	Ara <sub>8</sub>					
Phe121Ala	Arabinan	2.14 ± 0.26	-5.91	-6.84 ± 0.94	-0.93	1.02 ± 0.12
	Ara <sub>8</sub>			Not tested		
Glu122Ala	Arabinan	4.03 ± 0.09	-6.28	-6.47 ± 0.11	-0.19	1.01 ± 0.01
	Ara <sub>8</sub>			Not tested		
Trp123Ala	Arabinan	0.23 ± 0.02	-4.58	-13.61 ± 5.61	-9.03	1.04 ± 0.41
	Ara <sub>8</sub>			Not tested		
Asn132Ala	Arabinan			Weak Interaction (<10E3)		
	Ara <sub>8</sub>			Not tested		
Glu138Ala	Arabinan	5.51 ± 0.92	-6.47	-22.81 ± 1.89	-16.34	1.03 ± 0.07
	Ara <sub>8</sub>			Not tested		
Tyr139Ala	Arabinan	0.55 ± 0.32	-5.10	-3.2 ± 0.75	1.87	0.99 ± 0.22
	Ara <sub>8</sub>			No binding		
Gln142Ala	Arabinan	1.52 ± 0.06	-5.70	-11.01 ± 0.56	-5.31	1.08 ± 0.05
	Ara <sub>8</sub>	0.58 ± 0.02	-5.13	-6.55 ± 0.12	-1.42	1.00 ± 0.00

that could contribute to the ligand binding, either by providing hydrophobic stacking interactions or by direct hydrogen bonding, were selected to produce mutant alanine derivatives: Trp38, Gln86, Asp119, Phe121, Glu122, Trp123, Asn132, Phe134, Glu138, Tyr139 and Gln142. The binding affinities, determined by ITC, of these mutant forms to arabinan and Ara<sub>8</sub> were compared with those from the wild type form. Representative results are shown in Figure 6.7. From all the mutants analysed, only residues located in subdomain  $\gamma$  had impact in *RfCBM13-1* binding. Asp119Ala abolished the binding to both arabinan and Ara<sub>8</sub>, while Trp123Ala, Asn132Ala, Phe134Ala and Tyr139Ala led to a significant decrease in the affinity to arabinan (from a  $K_a$  of  $2.77 \times 10^4$  M<sup>-1</sup> to  $0.23 \times 10^4$  and  $0.55 \times 10^4$  M<sup>-1</sup> for Trp123Ala and Tyr139Ala, and below  $10^3$  M<sup>-1</sup> for Asn132Ala and Phe134Ala) (Table 6.2). Additionally, Phe134Ala and Tyr139Ala also abolished binding to Ara<sub>8</sub>, while Gln142Ala only led to a decrease in the binding capability to the oligosaccharide (from a  $K_a$  of  $2.87 \times 10^4$  to  $0.58 \times 10^4$  M<sup>-1</sup>), not influencing the affinity to the polysaccharide, which corroborates the importance of the ligand chain-length for *RfCBM13-1* binding. Given these observations, it can be inferred that *RfCBM13-1* subdomain  $\gamma$  comprises a functional binding site, where Asp119, Trp123, Asn132, Phe134 and Tyr139 seem

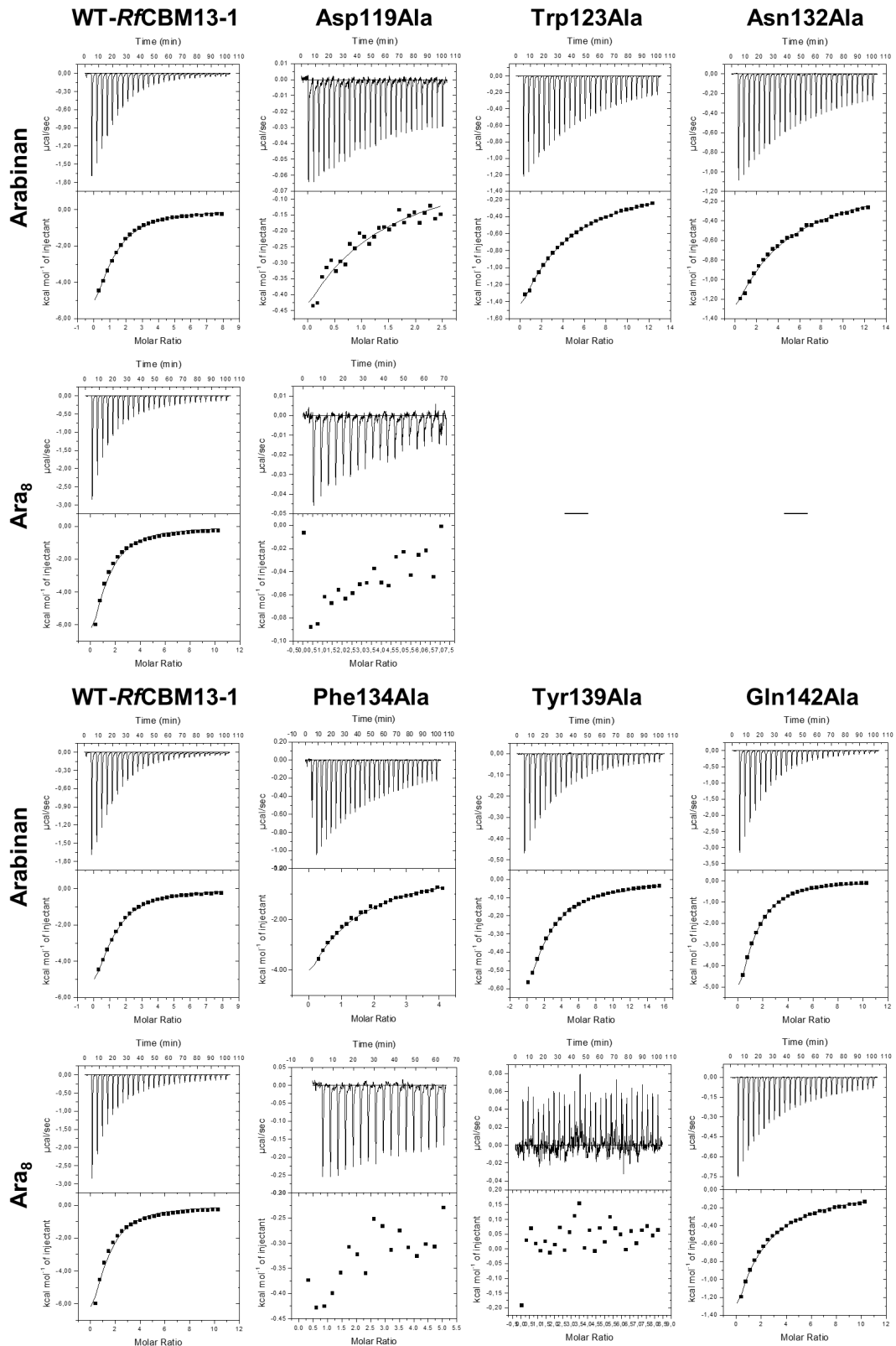


Figure 6.7. Isothermal calorimetry titrations of binding of *RfCBM13-1* mutant derivatives to  $\alpha$ 1,5-linked arabinan sequences. Binding analysis of *RfCBM13-1* wild type and mutants to arabinan polysaccharide and octasaccharide Ara<sub>8</sub>. The top portion of each panel shows the raw power data while the bottom parts show the integrated and heat of dilution corrected data. The solid lines show the non-linear curve fits to a one site binding model with the stoichiometry fixed at 1. Thermodynamic parameters are given in Table 6.2.

to be critical for the ligand recognition by this CBM.

The mutant derivatives of the amino acid residues selected from subdomains  $\alpha$  and  $\beta$ , however, mostly retained their affinity to arabinan. On the one hand, the decrease in binding affinity to arabinan observed from subdomain  $\gamma$  mutants Trp123Ala, Asn132Ala, Phe134Ala and Tyr139Ala, might indicate that binding is still occurring in the putative binding sites of subdomains  $\alpha$  and  $\beta$ . On the other hand, the lack of binding to Ara<sub>8</sub> from subdomain  $\gamma$  mutants Asp119Ala, Phe134Ala and Tyr139Ala, leaves the question of whether the putative binding sites of subdomains  $\alpha$  and  $\beta$  are in fact functional, or can be recognising a different type of carbohydrate present as a contaminant in the arabinan solution used. Nonetheless, different residues of subdomains  $\alpha$  and  $\beta$ , and multiple mutants from at least 2 binding sites, should be selected for mutation in order to have a better understanding of *RfCBM13-1* binding sites and its ligand recognition mechanisms.

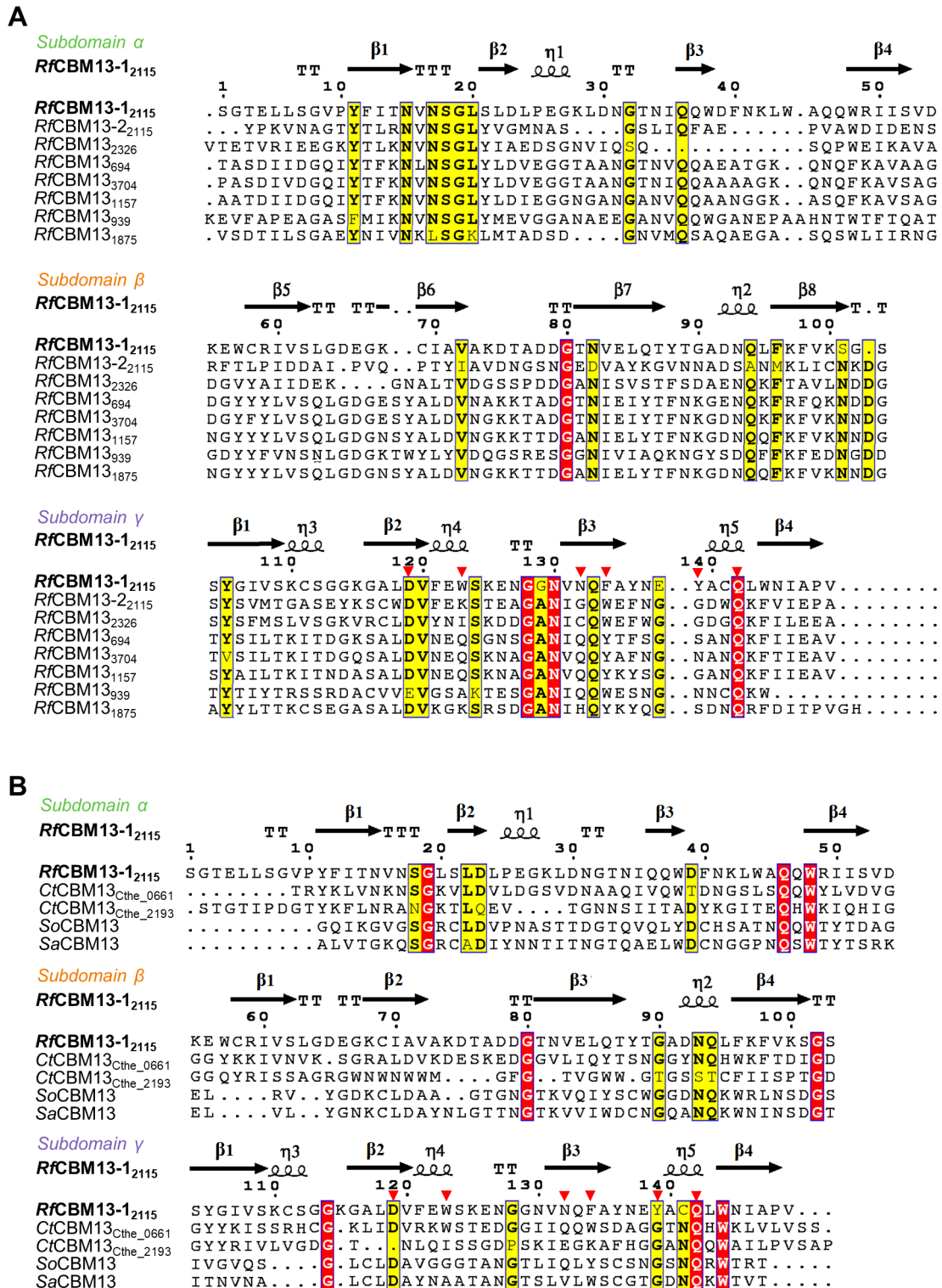
The analysis of protein residue conservation among *R. flavefaciens* family 13 CBMs (Figure 6.8A) using primary sequence alignment revealed that several residues of subdomain  $\gamma$  are highly conserved, including the interacting residues Asp119 and Gln146 identified in the binding site  $\gamma$  of *RfCBM13-1* (*RfCBM13-1*<sub>2115</sub>). The comparison with CBMs 13 from other microorganisms (Figure 6.8B), which show different binding specificities, showed the same trend, with only Asp119 and Gln146 being conserved. Looking at the sequence identities of subdomains  $\alpha$  and  $\beta$ , amino acid residues Asp23, Glu26, Gln36, Lys74, Asn93 and Gln94 could also be good candidates for site-directed mutagenesis and further ITC analysis, in order to validate these putative binding sites and identify their ligand-interacting residues.

In order to better perceive the  $\alpha$  and  $\beta$  putative binding sites of *RfCBM13-1*, a secondary structure matching superposition of the 3D structures of *RfCBM13-1* and *Streptomyces avermitilis* SaCBM13 (with Ara<sub>p</sub> monomers bound in the binding pocket) was carried out (Figure 6.9). With a primary sequence identity of 28%, only 2 of the interacting residues identified in *RfCBM13-1*'s binding site in subdomain  $\gamma$  are found in the same positions as those of SaCBM13, Asp119 and Phe134. Other residues displayed in the same positions were found to be not relevant for *RfCBM13-1* ligand recognition. This difference in the interacting residues identified in both structures, points to different binding recognition mechanisms, which might also be dependent of the sugar ring conformation of furanose for *RfCBM13-1* to pyranose for SaCBM13.

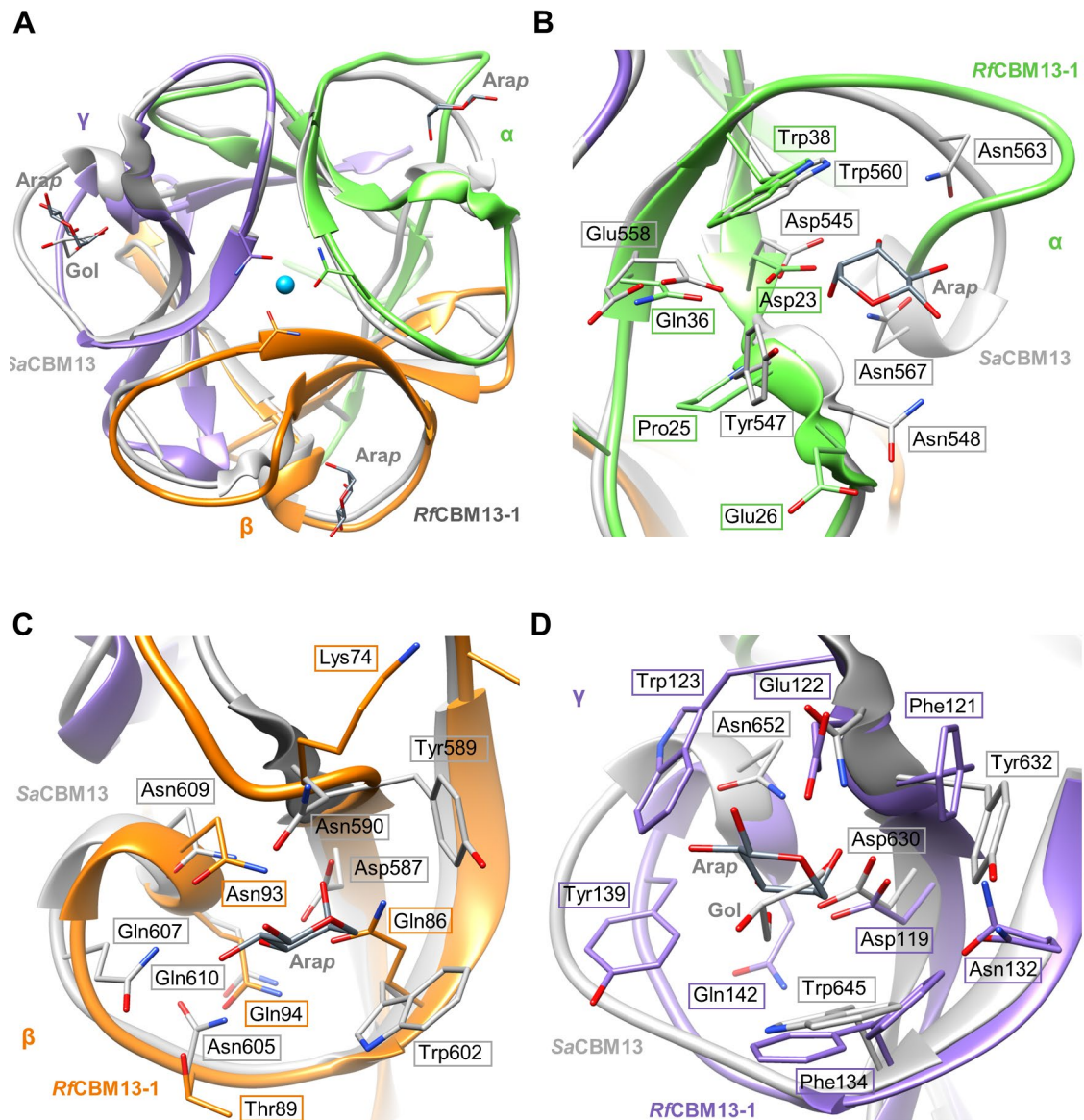
#### 6.2.4 *R. flavefaciens* family 13 CBMs in the context of plant cell wall recognition

Considering the CAZymes that are known to be associated with *R. flavefaciens* family 13 CBMs (Figure 6.2), the binding of *RfCBM13-1* (*RfCBM13-1*<sub>2115</sub>), *RfCBM13*<sub>2326</sub> and *RfCBM13*<sub>694</sub> to pectic arabinans and  $\beta$ -Gal containing sequences is not unexpected, as discussed above (section 6.2.1). *RfCBM13-1* and *RfCBM13*<sub>2326</sub> may be reflecting the associated family 43 GHs  $\alpha$ -arabinofuranosidase activity<sup>22,26</sup>, hence binding to Ara<sub>f</sub> sequences. *RfCBM13*<sub>694</sub>, associated





**Figure 6.8. Alignment of CBM13 family members.** Primary sequence alignment of (A) *R. flavefaciens* family 13 CBMs and (B) *RfCBM13-1<sub>2115</sub>* (*RfCBM13-1*) with CBMs from other microorganisms: *C. thermocellum* (*CtCBM13<sub>Cthe\_0661</sub>* and *CtCBM13<sub>Cthe\_2193</sub>*)<sup>156,254</sup>, *Streptomyces avermitilis* (*SaCBM13*)<sup>251</sup> and *Streptomyces olivaceoviridis* (*SoCBM13*)<sup>250</sup>. Identity to *RfCBM13-1* is indicated with red and yellow boxes. Residue numbers refer to the corresponding CBM sequence. *RfCBM13-1* secondary structure prediction is presented above. Red triangles identify *RfCBM13-1* residues identified to be involved in ligand recognition. The sequence alignment was generated with ClustalO<sup>196</sup> and rendered using ESPript server<sup>229</sup>.



**Figure 6.9. Superposition of *RfCBM13-1* with *Streptomyces avermitilis* CBM13.** *RfCBM13-1* structure was superimposed with SaCBM13 bound to arabinopyranose monomers (PDB ID 3V22)<sup>251</sup>. (A) The overall structures superposition evidences the putative binding sites of *RfCBM13-1*; Close-up view on subdomains (B)  $\alpha$ , (C)  $\beta$  and (D)  $\gamma$ , showing SaCBM13 amino acid residues that interact with Arap and evidencing *RfCBM13-1* putative binding sites and its potential residues for recognition of Arap ligands. *RfCBM13-1* is represented as cartoon coloured by subdomain  $\alpha$  (green),  $\beta$  (orange) and  $\gamma$  (purple) and SaCBM13 in light grey. Arabinopyranose monomers (Arap) and glycerol molecule (GOL) are shown as stick models in light grey and dark grey, respectively, and by atom type. Magnesium ion in the centre of *RfCBM13-1* trefoil is represented by a blue sphere. Amino acid residues of each CBM are represented by sticks and coloured by atom type. Structural alignment was done using MatchMaker tool from UCF Chimera<sup>40</sup>, with an rmsd value of 0.827.

with a PL11 annotated as a family of rhamnogalacturonan lyases<sup>22</sup>, may recognise non-reducing  $\beta$ -Gal sequences as those found in RG-I galactan and arabinogalactan branches (Figure 6.1).

*RfCBM13*<sub>3704</sub>, like *RfCBM13*<sub>694</sub>, is associated to a PL11. *RfCBM13*<sub>1157</sub> on its turn, is associated with a CE12, reported to comprise pectin acetyl esterases, rhamnogalacturonan acetyl esterases and acetyl xylan esterases<sup>22</sup>. The lack of binding of *RfCBM13*<sub>3704</sub> and *RfCBM13*<sub>1157</sub> in the

polysaccharide microarrays presented in Chapter 3, is probably due to absence of the relevant pectic epitopes in the microarrays.

*RfCBM13*<sub>1875</sub>, for which binding was also not observed in the microarrays, it is not associated with an assigned CAZyme or other cellulosomal protein, hence it can have a distinct function and binding specificity or even not be a functional CBM.

Although *RfCBM13*-2<sub>2115</sub> and *RfCBM13*<sub>939</sub> were not possible to analyse, their modular organization also points to pectin recognition. *RfCBM13*-2<sub>2115</sub>, like its counterpart *RfCBM13*-1, could eventually recognise *Araf* sequences as well or bind a different yet complementary epitope in RG-I, promoting GH43 activity. Interestingly, these are the only two family 13 CBMs found in tandem in *R. flavefaciens*. *RfCBM13*<sub>939</sub>, like *RfCBM13*<sub>1157</sub>, is associated to a CE12, and hence might bind to epitopes in the pectin main chain or branches.

Overall, the modular architecture of *R. flavefaciens* family 13 CBMs points to a role of the cellulosome of this bacterium directed at pectin degradation.

### 6.3 Conclusions

With the present work the carbohydrate specificity of *R. flavefaciens* family 13 CBMs was explored for 3 CBMs and assigned to distinct  $\alpha$ 1,5-linked *Araf*-containing sequences in pectic arabinans and to a yet uncharacterised  $\beta$ -Gal epitope in pectic galactans. Additionally, the first 3D structure of a *RfCBM13*, *RfCBM13*-1, was solved revealing the  $\beta$ -trefoil fold typical of family 13 CBMs and 3 putative binding clefts, one in each subdomain. Ligand binding analysis allowed to establish *RfCBM13*-1 specificity to linear  $\alpha$ 1,5 arabinan sequences and a dependency of chain-length from DP-2 to DP-6 for affinity. Although the structure of *RfCBM13*-1 exhibits putative binding clefts that could accommodate up to 2 ligand monomers, as usually described for CBMs 13, the higher affinity to longer chain-length ligands might point to a recognition mode where important interactions are established between the surface of the CBM and the remaining ligand sequence that is not accommodated in the binding cleft. Key residues that mediate arabinan recognition by *RfCBM13*-1 were identified in the binding cleft of subdomain  $\gamma$ , where Asp119, Trp123, Asn132, Phe134 and Tyr139 seem to be critical for the ligand recognition. However, further studies will be necessary to characterize the putative binding sites of subdomains  $\alpha$  and  $\beta$  and completely elucidate *RfCBM13*-1 carbohydrate recognition mechanism.

In order to complete assignment of the carbohydrate ligand specificities for *R. flavefaciens* family 13 CBMs, further carbohydrate microarray analysis should be performed using pectin polysaccharide or oligosaccharide microarrays. To achieve this, the sequence diversity of the pectin-related oligosaccharide library needs to be much extended, either with naturally-derived or synthetic sequences, in order to construct sequence-defined microarrays.

In summary, the information derived from this work points to a possible role of family 13 CBMs in the cellulosome of *R. flavefaciens* directed to pectin degradation. These results here reported will

also contribute for the information deposited into the CAZy database, promoting the knowledge of *R. flavefaciens*' cellulolytic activity in its ecological niche.

## 6.4 Experimental procedure

### 6.4.1 Gene cloning, mutagenesis and protein purification

*R. flavefaciens* family 13 CBMs analysed in the carbohydrate microarrays were cloned, expressed and purified according with the procedure described in Chapter 3, section 3.4.2.

For the structural studies of *RfCBM13-1<sub>2115</sub>*, *Escherichia coli* BL21 (DE3) cells were transformed with the desired plasmid and grown at 37 °C in 400 mL LB medium supplemented with 50 µg/mL kanamycin (Sigma-Aldrich Chemical, St Louis, Missouri) up to the mid-exponential phase ( $OD_{600nm} = 0.6$ ). Gene expression and protein production was induced by addition of 1 mM IPTG, followed by 16 h culturing at 19 °C at 150 rpm in a *Gallenkamp* Orbital *Shaker*. The cells were then harvested by centrifugation at 5000xg for 15 minutes at 4 °C and stored at -20 °C. Immediately before purification, cell pellets were resuspended in lysis buffer (50 mM NaHepes buffer, pH 7.5, supplemented with 1 M NaCl and 10 mM Imidazole), and then disrupted by sonication. Non-solubilized cell debris was removed by centrifugation (140000 rpm, 30 min, 4 °C). *RfCBM13-1<sub>2115</sub>* was purified from the cleared cell-lysate by Ni<sup>2+</sup>-immobilized IMAC using buffers of 50 mM HEPES (pH 7.5), 1 M NaCl, with 5, 50, and 500 mM imidazole for binding, washing, and elution, respectively. The eluted protein fractions were subjected to SDS-PAGE on 14% (w/v) acrylamide gels, stained with Coomassie Brilliant Blue, in order to assess the purity of recombinant proteins. The fractions containing pure protein were pooled and buffer-exchanged, using PD-10 Sephadex G-25M gel-filtration columns (GE Healthcare), into 50 mM NaHepes buffer, pH 7.5, containing 200 mM NaCl and 5 mM CaCl<sub>2</sub>. Purified proteins were concentrated using an Amicon 10-kDa molecular mass centrifugal concentrator.

For the X-ray crystallography studies, *RfCBM13-1<sub>2115</sub>* was further purified by gel filtration (size exclusion chromatography) using an AKTAexpress FPLC equipped with a HiLoad 16/60 Superdex75 column (GE Healthcare Life Sciences) at a flow rate of 1 mL/min. Purified *RfCBM13-1<sub>2115</sub>* was concentrated and exchanged into 50 mM HEPES buffer, pH 7.5, containing 1 mM CaCl<sub>2</sub> using an Amicon 10-kDa molecular mass centrifugal concentrator.

All proteins were >95% pure as judged by SDS-PAGE and their concentrations determined from their calculated molar extinction coefficient using the ProtParam tool (<http://www.expasy.org/tools/protparam.html>) at 280 nm using a NanoDrop 2000c (ThermoScientific).

For site-directed mutagenesis, single mutants of *RfCBM13-1<sub>2115</sub>* were generated using a PCR-based NZYMutagenesis kit (NZYtech Ltd) following the manufacturer's instructions. The list of primers used to generate these mutants is provide in Table S6.1. The generated nucleic acids were sequenced to ensure that only the right mutations had been incorporated in the nucleic

acids. Expression and purification of the mutants were similar with that of wild type *RfCBM13-1<sub>2115</sub>*.

#### 6.4.2 Sources of carbohydrates

Information on the oligosaccharide sequences and sources included in the NGL-microarrays are given in Table S2.1, in Chapter 2. The soluble polysaccharides and arabinose oligosaccharides used for crystallization assays or ITC were purchased from Megazyme International (Bray, County Wicklow, Ireland). All reagents, chemicals and other carbohydrates were purchased from Sigma-Aldrich (St. Louis, MO, USA) unless otherwise specified.

#### 6.4.3 Carbohydrate microarray analysis

The NGL-microarray results here reported correspond to the experiments presented in Chapter 3 and were performed as described in section 3.2.4. The results correspond to at least two independent experiments, performed with different batches of CBMs.

#### 6.4.4 Crystallization and X-ray Diffraction Data Collection

*RfCBM13-1<sub>2115</sub>* crystallization assays were performed by means of an automated nano-drop dispenser Oryx8 (Douglas Instruments) using commercial screenings, namely PEG/Ion HT (Hampton Research), JCSG-plus (Molecular Dimensions) and an in-house prepared sparse matrix screen (based on the screen of Jancarik *et al*<sup>255</sup>). The sitting-drop vapor diffusion method was used (SWISSCI 'MRC' 2-Drop Crystallization Plates – 96 wells, Douglas Instruments), in a 2  $\mu$ L drop containing 50% protein. Crystals of *RfCBM13-1* at 15 mg/mL grew at 20 °C in a crystallization condition composed of 0.1 M sodium acetate buffer, pH 4.5, and 2 M ammonium sulphate. Crystals were harvested using a solution of 0.1 M sodium acetate buffer, pH 4.5, and 2.5 M ammonium sulphate, and then flash-cooled in liquid nitrogen using 30% (v/v) glycerol as cryoprotectant added to the harvesting solution.

X-ray diffraction data from a single crystal were collected under a nitrogen stream at 100 K in X06DA - PXIII beamline at the SLS (Villigen, Switzerland) to a maximum resolution of 1.80 Å and using X-ray radiation at a fixed wavelength of 0.9795 Å. The *RfCBM13-1* crystal indexed in space group  $P4_12_12$ , with unit cell constants  $a = b = 65.43$  and  $c = 102.79$  Å, corresponding to a calculated Matthews coefficient of 3.42 Å<sup>3</sup>/Da and a solvent content of 64%. Statistics from data collection and processing, model building and validation are shown in Table 6.1.

#### 6.4.5 Phasing, Model Building, and Refinement

*RfCBM13-1* X-ray diffraction data were processed using MOSFLM<sup>202</sup> and SCALA<sup>203</sup> from the CCP4 suite<sup>204</sup>. Phasing was performed by molecular replacement with Phaser MR<sup>205</sup> from CCP4 using as MR model the polypeptide chain of *C. thermocellum* CtCBM13<sub>Cthe\_0661</sub> structure (PDB ID

3VSF)<sup>156</sup>. Models completion, editing, and initial validation were carried out in COOT<sup>206</sup>. Automatic addition of water molecules and restrained refinement of the full models were done using REFMAC5<sup>207</sup>. Structure validation was performed using ProCheck<sup>233</sup> and SFCHECK<sup>234</sup>. *RfCBM13-1* structure, with  $R = 15.9\%$  ( $R_{free} = 19.7\%$ ), consists of 1 CBM chain of 145 amino acid residues, 1 magnesium ion, 1 sulphate ion, 3 acetate ions, 1 glycerol molecule and 237 water molecules.

Molecular graphics images corresponding to the crystal structure were produced using the UCSF Chimera package from the Computer Graphics Laboratory, University of California, San Francisco<sup>40</sup>.

#### 6.4.6 Isothermal titration calorimetry

ITC assays were performed as described previously in Chapter 4, section 4.4.7. Before the experiments, purified CBMs were buffer-exchanged against 50 mM Na-HEPES buffer, pH 7.5, containing 1 mM CaCl<sub>2</sub>. Thermodynamic parameters are shown in Table 6.2.

#### 6.5 Work contributions

Experimental work and data interpretation here reported related to the carbohydrate microarrays, crystal structure determination and sequence similarity analysis, were executed by the author of this thesis. Site directed mutagenesis, expression and purification of wild type *RfCBM13-1*<sub>2115</sub> and its mutant derivatives, as well as the initial crystallization screenings of wild type *RfCBM13-1*<sub>2115</sub>, unliganded and with Ara<sub>3</sub>, and ITC assays were performed by Dr. Virgínia Pires at Prof. Carlos Fontes laboratory (CIISA-FMV-ULisboa), with planning and discussion with the author of this thesis.

# CHAPTER 7

---

**CONCLUSIONS AND FUTURE PERSPECTIVES**





## 7 Conclusions and future perspectives

### 7.1 General conclusions

In the past years, genomic sequencing of bacterial genomes has promoted an exponential increase in information available, leading to a substantial number of CAZymes and CBM sequences identified that await structural and functional characterization. The development of miniaturized high-throughput technologies, such as carbohydrate microarrays, to systematically interrogate carbohydrate libraries and its combination with structural characterization methodologies are crucial to identify the specificity and explain, at atomic level, the biological roles of these carbohydrate-binding proteins. With this major goal, this Thesis describes the application of such integrative approach by combining high-throughput methodologies of protein expression and purification and carbohydrate microarrays with X-ray crystallography, contributing to identification and structural characterization of the carbohydrate-binding specificity of the CBMs from two cellulolytic bacteria, *C. thermocellum* and *R. flavefaciens* FD-1.

At the start of this Thesis work the carbohydrate binding for the great majority of the assigned CBMs encoded in the sequenced genome of *R. flavefaciens* was still to be identified. Despite *C. thermocellum* CBMs had been more extensively studied as its cellulosome was the first to be identified, the binding specificity for several of its CBMs was also not characterized. Therefore, there was a clear need of molecular information regarding the function of these CAZy-classified modules. Increasing the microarray platforms' diversity in naturally-derived plant cell wall carbohydrate probes was determinant for the characterization of these CBMs. The first major achievement of this Thesis was the development of novel carbohydrate microarray platforms constructed and validated in Chapters 2 and 3, which included polysaccharide microarrays and neoglycolipid (NGL)-based oligosaccharide microarrays of hemicellulose-related sequences representative of those found in plants cell walls, but also sequences present in fungal and bacterial cell walls. These microarrays were then applied to screen 150 CBMs from both bacteria for carbohydrate-binding. The second major achievement was the identification of carbohydrate ligands for up to 59 CBMs, including novel CBM-ligand specificities for 21 CBMs from *R. flavefaciens* and 23 from *C. thermocellum*. Overall, it was revealed that the two bacteria present CBMs with different carbohydrate-binding specificities, which may reflect the different polysaccharide sources available in their specific ecological niches, but also the complexity and specialization of their cellulosomes.

Of the 59 CBMs that revealed binding, and for their representativeness in the CAZy pool of CBM families, several deserve further biochemical and biophysical characterization. Considering the timeframe of this Thesis, and their potential industrial and biotechnological relevance, 3 CBMs were selected for structural characterization and constitute the focus of Chapters 4 to 6. The

information derived from the microarrays in combination with X-ray crystallography studies culminated in novel 3D structures and the characterization of the binding specificities of CBMs from families 11, 13 and 50. The third major achievement, was the demonstration of the specificity of *C. thermocellum* family 11 CBM (*CtCBM11*<sub>Cthe\_1472</sub>) for the twisted conformation of mixed-linked  $\beta$ 1,3-1,4-glucans. This is mediated by CH- $\pi$  stacking and a hydrogen bonding network, which is dependent not only on ligand chain length, but also on the presence of a  $\beta$ 1,3-linkage at the reducing end and at specific positions along the  $\beta$ 1,4-linked glucan chain. The fourth major achievement was the assignment of carbohydrate-binding specificity for 7 CBMs of *C. thermocellum* family 50 towards  $\beta$ 1,4-linked GlcNAc sequences, which led to solving the first structure of a *CtCBM50* (*CtCBM50*<sub>Cthe\_0300</sub>) in complex with a GlcNAc trisaccharide. Key residues were identified to mediate both chitin and peptidoglycan oligosaccharide recognition through an hydrogen bonding network and CH- $\pi$  stacking interactions. *CtCBM50*<sub>Cthe\_0300</sub> binding was shown to be favored by an interchain multivalent assembly induced by the GlcNAc oligosaccharide chain-length, where the individual CBM molecules bind in a cooperative manner to longer ligand chains, supporting the evidence of LysM domain cooperative binding. The fifth major achievement was the assignment of carbohydrate specificity for 3 CBMs of *R. flavefaciens* family 13 towards distinct  $\alpha$ 1,5-linked *Araf*-containing sequences in pectic arabinans and to a yet uncharacterised  $\beta$ -galactose epitope in pectic galactans, which points to a possible role of family 13 CBMs in the cellulosome of *R. flavefaciens* directed to pectin degradation. These results led to solving first structure of a *RfCBM13* (*RfCBM13-1*<sub>2115</sub>), which revealed 3 putative binding clefts and to identify at least one subdomain in *RfCBM13-1* (subdomain  $\gamma$ ) comprising a functional binding site, critical for binding to linear  $\alpha$ 1,5-arabinan sequences.

Overall, the work developed through this Thesis allowed to elucidate the role of CBMs and CBM families in their microorganisms, which lead to a better understanding of these bacteria cellulolytic capabilities. The combined high-throughput approach of using carbohydrate microarrays and X-ray crystallography, proved to be an effective strategy to attain the 3D structures of novel CBMs, isolated or in complex with their biologically relevant oligosaccharide ligands, which will in turn potentiate their biotechnological applications in diverse fields. Furthermore, it adds crucial information to the classification of the novel CBMs identified, in particular those from *R. flavefaciens*, eventually contributing for important resources such as the CAZy ([www.cazy.org](http://www.cazy.org)), CAZyedia ([www.cazypedia.org](http://www.cazypedia.org)), GlycoPedia ([www.glycopedia.eu](http://www.glycopedia.eu)), GlyGen ([www.glygen.org](http://www.glygen.org)) and ProCarbDB ([www.procarbdb.science](http://www.procarbdb.science)) databases.

## 7.2 Future perspectives

The work presented in this Thesis represents an important step towards the thorough characterization of carbohydrate-binding proteins, nonetheless some challenges have arisen that are worth following-up.

Increasing the microarrays platforms' diversity in naturally-derived oligosaccharide probes from plant, fungal and bacterial cell walls is important for characterizing microbial polysaccharide-recognising systems and detailed characterization of the specificity of their interactions. Oligosaccharide microarrays (both homogenous and sequence-defined or as mixtures) can reveal subtle differences in binding profiles which may not be discriminated with analysis using polysaccharide-based methods. Microarrays from carbohydrates derived from these sources have been developed, however obtaining sequence-defined and discrete oligosaccharide sequences from natural sources poses several challenges. Unambiguous determination of plant-derived oligosaccharides is hampered by the heterogeneous nature of polysaccharides and the difficulties in their separation and purification.

Intending to surpass this constrain, a method that would enable the deconvolution of structurally similar oligosaccharide mixtures was attempted and was described in Chapter 2. This involved bi-functional conjugation of xyloglucan fractions with UV/fluorescence tag DAN, aiming at sensitive detection in HPLC to allow detailed fine separation/purification of its components. The method showed good yields of separation when starting with relatively simple mixtures that had previously been subjected to purification steps, allowing the construction of sequence-defined NGL-microarrays. However, conjugation was not successful when applied to large mixtures that had only been fractionated by size, most likely due to the high heterogeneity of each fraction. Although promising, this method needs to be further optimised, possibly by adding different steps of purification, in order to successfully generate more structurally diverse sequence-defined oligosaccharide probes. Such method would be of relevance, not only for the generation of hemicellulose-related sequence-defined microarrays, but also for the development of pectin-derived oligosaccharide microarrays that could be used for the characterization of pectin-recognising systems, such as *R. flavefaciens* family 13 CBMs, which were described in Chapters 3 and 6. Additionally, this work highlighted the important interface between carbohydrate microarrays and mass spectrometry, and the need for the further development of high-sensitivity methods for the determination of oligosaccharide linkages and sequences.

While the microarray platforms constructed allowed to successfully assign the carbohydrate specificities for over half *R. flavefaciens* CBMs tested from families 4, 6, 13, 22 and 35, several questions remain open. Ligand-binding is still unassigned for members of families 3, 48 and 63, which would be of interest to analyse in the NGL-oligosaccharide, as well as in the xyloglucan microarray platforms. Additionally, further work is needed to clarify the carbohydrate ligands for all CBMs from families 13 and 62. These would be worth testing in the pectin polysaccharides

microarrays. The structural-functional characterization of the carbohydrate-binding specificities of families 6 and 22 CBMs to xylan- and arabinoxylan-derived sequences should also be pursued, hence elucidating the role of these CBM families and their possible complementary function in *R. flavefaciens*. Furthermore, in the absence of protein crystals of the ligand-bound forms, such as in the case of *RfCBM13-1*, complementary methods should be explored for the structural characterization of the CBM ligand specificity. A possible follow-up of the work presented for *RfCBM13-1*, would be NMR titrations of the CBM with different chain-length ligands, that could inform on the protein aminoacid residues involved in the binding recognition.

The characterization of the plant cell wall carbohydrate recognition by *R. flavefaciens* can be an important step towards understanding the action of other cellulosome-producing ruminococcal bacteria from different biological systems, in particular from the human gut. For instance, the bacterium *Ruminococcus champanellensis* found in the human colon expresses a cellulosome system similar to that of *R. flavefaciens*<sup>256</sup>. The elucidation of *R. flavefaciens* cellulosome, may contribute to understanding *R. champanellensis*' cellulolytic capabilities, and hence promote the development of strategies for microbial manipulation and personalized medicine.

Moreover, the integrative approach described in this thesis, combining high-throughput techniques for protein expression and purification and for carbohydrate ligand discovery with structural biology methods, can be applied to any biological system, promoting the understanding of other complex systems, such as the human gut microbiome.

# REFERENCES

---

*REFERENCES*

## References

1. Neelamegham, S. *et al.* Updates to the Symbol Nomenclature for Glycans guidelines. *Glycobiology* **29**, 620–624 (2019).
2. *Essentials of Glycobiology*. (Cold Spring Harbor Laboratory Press, 2017).
3. Carpita, N. C. & Gibeaut, D. M. Structural models of primary cell walls in flowering plants: consistency of molecular structure with the physical properties of the walls during growth. *Plant J.* **3**, 1–30 (1993).
4. Cosgrove, D. J. Growth of the plant cell wall. *Nat. Rev. Mol. Cell Biol.* **6**, 850–861 (2005).
5. Knox, J. P. Revealing the structural and functional diversity of plant cell walls. *Curr. Opin. Plant Biol.* **11**, 308–13 (2008).
6. Burton, R. A., Gidley, M. J. & Fincher, G. B. Heterogeneity in the chemistry, structure and function of plant cell walls. *Nat. Chem. Biol.* **6**, 724–732 (2010).
7. Pérez, S. & Samain, D. Structure and Engineering of Celluloses. in *Advances in Carbohydrate Chemistry and Biochemistry* **64**, 25–116 (Elsevier Inc., 2010).
8. Scheller, H. V. & Ulvskov, P. Hemicelluloses. *Annu. Rev. Plant Biol.* **61**, 263–289 (2010).
9. Stone, B. A. Chemistry of  $\beta$ -Glucans. in *Chemistry, Biochemistry, and Biology of 1-3 Beta Glucans and Related Polysaccharides* 5–46 (Elsevier, 2009).
10. Mohnen, D. Pectin structure and biosynthesis. *Curr. Opin. Plant Biol.* **11**, 266–277 (2008).
11. Artzi, L., Bayer, E. A. & Morais, S. Cellulosomes: bacterial nanomachines for dismantling plant polysaccharides. *Nat. Rev. Microbiol.* **15**, 83–95 (2016).
12. Fontes, C. M. G. A. & Gilbert, H. J. Cellulosomes: Highly Efficient Nanomachines Designed to Deconstruct Plant Cell Wall Complex Carbohydrates. *Annu. Rev. Biochem.* **79**, 655–681 (2010).
13. Yaniv, O. *et al.* Structure of a family 3a carbohydrate-binding module from the cellulosomal scaffoldin CipA of *Clostridium thermocellum* with flanking linkers: Implications for cellulosome structure. *Acta Crystallogr. Sect. F Struct. Biol. Cryst. Commun.* **69**, 733–737 (2013).
14. Bule, P., Pires, V. M., Fontes, C. M. & Alves, V. D. Cellulosome assembly: paradigms are meant to be broken! *Curr. Opin. Struct. Biol.* **49**, 154–161 (2018).
15. Lamed, R., Setter, E., Kenig, R. & Bayer, E. A. The cellulosome: a discrete cell surface organelle of *Clostridium thermocellum* which exhibits separate antigenic, cellulose-binding and various cellulolytic activities. *Biotechnol. Prog.* **13**, 163–181 (1983).
16. Bergquist, P. L. *et al.* Molecular diversity of thermophilic cellulolytic and hemicellulolytic bacteria. *FEMS Microbiol. Ecol.* **28**, 99–110 (1999).
17. Brás, J. L. A. *et al.* Diverse specificity of cellulosome attachment to the bacterial cell surface. *Sci. Rep.* **6**, 38292 (2016).
18. Flint, H. J., Bayer, E. A., Rincon, M. T., Lamed, R. & White, B. A. Polysaccharide utilization by gut bacteria: potential for new insights from genomic analysis. *Nat. Rev. Microbiol.* **6**, 121–131 (2008).
19. Hemsworth, G. R., Déjean, G., Davies, G. J. & Brumer, H. Learning from microbial strategies for polysaccharide degradation. *Biochem. Soc. Trans.* **44**, 94–108 (2016).
20. Lynd, L. Microbial Cellulose Utilization: Fundamentals and Biotechnology. *Microbiol. Mol. Biol. Rev.* **66**, 506–577 (2002).
21. de Vries, R. P., Visser, J., Ronald, P., de Vries, R. & P. Aspergillus Enzymes Involved in Degradation of Plant Cell Wall Polysaccharides. *Microbiol. Mol. Biol. Rev.* **65**, 497–522 (2001).
22. Lombard, V., Golaconda Ramulu, H., Drula, E., Coutinho, P. M. & Henrissat, B. The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res.* **42**, D490–D495 (2014).
23. Boraston, A. B., Bolam, D. N., Gilbert, H. J. & Davies, G. J. Carbohydrate-binding modules: fine-tuning polysaccharide recognition. *Biochem. J.* **382**, 769–781 (2004).
24. Gilbert, H. J., Knox, J. P. & Boraston, A. B. Advances in understanding the molecular basis of plant cell wall polysaccharide recognition by carbohydrate-binding modules. *Curr. Opin. Struct. Biol.* **23**, 669–677 (2013).
25. Gilkes, N. R., Warren, R. A., Miller, R. C. & Kilburn, D. G. Precise excision of the cellulose binding domains from two *Cellulomonas fimi* cellulases by a homologous protease and the effect on catalysis. *J. Biol. Chem.* **263**, 10401–10407 (1988).
26. Gilbert, H. J. The Biochemistry and Structural Biology of Plant Cell Wall Deconstruction. *Plant Physiol.* **153**, 444–455 (2010).
27. Berg Miller, M. E. *et al.* Diversity and strain specificity of plant cell wall degrading enzymes revealed by the draft genome of *Ruminococcus flavefaciens* FD-1. *PLoS One* **4**, (2009).
28. Georgelis, N., Yennawar, N. H. & Cosgrove, D. J. Structural basis for entropy-driven cellulose binding by a type-A cellulose-binding module (CBM) and bacterial expansin. *Proc. Natl. Acad. Sci.* **109**, 14830–14835 (2012).
29. Hernandez-Gomez, M. C. *et al.* Recognition of xyloglucan by the crystalline cellulose-binding site of a family 3a carbohydrate-binding module. *FEBS Lett.* **589**, 2297–2303 (2015).

30. Pires, V. M. R. *et al.* Stability and Ligand Promiscuity of Type A Carbohydrate-binding Modules Are Illustrated by the Structure of Spirochaeta thermophila StCBM64C. *J. Biol. Chem.* **292**, 4847–4860 (2017).
31. Boraston, A. B. *et al.* Differential Oligosaccharide Recognition by Evolutionarily-related  $\beta$ -1,4 and  $\beta$ -1,3 Glucan-binding Modules. *J. Mol. Biol.* **319**, 1143–1156 (2002).
32. Palma, A. S. *et al.* Unravelling Glucan Recognition Systems by Glycome Microarrays Using the Designer Approach and Mass Spectrometry. *Mol. Cell. Proteomics* **14**, 974–988 (2015).
33. Carvalho, A. L. *et al.* 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. *J. Biol. Chem.* **279**, 34785–34793 (2004).
34. Ribeiro, D. O. *et al.* Molecular basis for the preferential recognition of  $\beta$ 1,3-1,4- glucans by the family 11 carbohydrate-binding module from Clostridium thermocellum. *FEBS J.* 1–21 (2019).
35. van Bueren, A. L., Morland, C., Gilbert, H. J. & Boraston, A. B. Family 6 Carbohydrate Binding Modules Recognize the Non-reducing End of  $\beta$ -1,3-Linked Glucans by Presenting a Unique Ligand Binding Surface. *J. Biol. Chem.* **280**, 530–537 (2005).
36. Ribeiro, T. *et al.* Family 42 carbohydrate-binding modules display multiple arabinoxylan-binding interfaces presenting different ligand affinities. *Biochim. Biophys. Acta - Proteins Proteomics* **1804**, 2054–2062 (2010).
37. Notenboom, V., Boraston, A. B., Williams, S. J., Kilburn, D. G. & Rose, D. R. High-resolution crystal structures of the lectin-like xylan binding domain from Streptomyces lividans xylanase 10A with bound substrates reveal a novel mode of xylan binding. *Biochemistry* **41**, 4246–4254 (2002).
38. Pires, V. M. R. R. *et al.* 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. *J. Biol. Chem.* **279**, 21560–21568 (2004).
39. Henshaw, J. L. *et al.* The Family 6 Carbohydrate Binding Module Cm CBM6-2 Contains Two Ligand-binding Sites with Distinct Specificities. *J. Biol. Chem.* **279**, 21552–21559 (2004).
40. Pettersen, E. F. *et al.* UCSF Chimera - A visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004).
41. Yaniv, O. *et al.* Fine-structural variance of family 3 carbohydrate-binding modules as extracellular biomass-sensing components of Clostridium thermocellum anti- $\alpha$ l factors. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **70**, 522–534 (2014).
42. Buist, G., Steen, A., Kok, J. & Kuipers, O. P. LysM, a widely distributed protein motif for binding to (peptido)glycans. *Mol. Microbiol.* **68**, 838–847 (2008).
43. Demain, A. L., Newcomb, M. & Wu, J. H. D. Cellulase, Clostridia, and Ethanol. *Microbiol. Mol. Biol. Rev.* **69**, 124–154 (2005).
44. Coelho, D. *et al.* Novel combination of feed enzymes to improve the degradation of Chlorella vulgaris recalcitrant cell wall. *Sci. Rep.* **9**, 5382 (2019).
45. Karmakar, M. & Ray, R. R. Current Trends in Research and Application of Microbial Cellulases. *Res. J. Microbiol.* **6**, 41–53 (2011).
46. Bayer, E. A., Morag, E. & Lamed, R. The cellulosome — A treasure-trove for biotechnology. *Trends Biotechnol.* **12**, 379–386 (1994).
47. Hyeon, J. E., Jeon, S. D. & Han, S. O. Cellulosome-based, Clostridium-derived multi-functional enzyme complexes for advanced biotechnology tool development: Advances and applications. *Biotechnol. Adv.* **31**, 936–944 (2013).
48. Stern, J., Moraš, S., Lamed, R. & Bayer, E. A. Adaptor Scaffoldins: An Original Strategy for Extended Designer Cellulosomes, Inspired from Nature. *MBio* **7**, 1–10 (2016).
49. Shoseyov, O., Shani, Z. & Levy, I. Carbohydrate Binding Modules: Biochemical Properties and Novel Applications. *Microbiol. Mol. Biol. Rev.* **70**, 283–295 (2006).
50. Oliveira, C., Carvalho, V., Domingues, L. & Gama, F. M. Recombinant CBM-fusion technology — Applications overview. *Biotechnol. Adv.* **33**, 358–369 (2015).
51. Reyes-Ortiz, V. *et al.* Addition of a carbohydrate-binding module enhances cellulase penetration into cellulose substrates. *Biotechnol. Biofuels* **6**, 93 (2013).
52. McCartney, L., Gilbert, H. J., Bolam, D. N., Boraston, A. B. & Knox, J. P. Glycoside hydrolase carbohydrate-binding modules as molecular probes for the analysis of plant cell wall polymers. *Anal. Biochem.* **326**, 49–54 (2004).
53. Gao, S., You, C., Renneckar, S., Bao, J. & Zhang, Y.-H. New insights into enzymatic hydrolysis of heterogeneous cellulose by using carbohydrate-binding module 3 containing GFP and carbohydrate-binding module 17 containing CFP. *Biotechnol. Biofuels* **7**, 24 (2014).
54. Beguin, P. & Alzarit, P. M. The cellulosome of Clostridium thermocellum. *Biochem. Soc. Trans.* **26**, 178–185 (1998).
55. Paye, J. M. D. *et al.* Biological lignocellulose solubilization: comparative evaluation of biocatalysts and enhancement via cotreatment. *Biotechnol. Biofuels* **9**, 8 (2016).
56. Xu, Q. *et al.* Dramatic performance of Clostridium thermocellum explained by its wide range of cellulase modalities. *Sci. Adv.* **2**, 1–12 (2016).



57. Hirano, K. *et al.* Enzymatic diversity of the *Clostridium thermocellum* cellulosome is crucial for the degradation of crystalline cellulose and plant biomass. *Sci. Rep.* **6**, 35709 (2016).
58. Flint, H. J., Duncan, S. H., Scott, K. P. & Louis, P. Interactions and competition within the microbial community of the human colon: links between diet and health. *Environ. Microbiol.* **9**, 1101–1111 (2007).
59. Venditto, I. *et al.* Complexity of the *Ruminococcus flavefaciens* cellulosome reflects an expansion in glycan recognition. *Proc. Natl. Acad. Sci.* **113**, 7136–7141 (2016).
60. Abbott, D. W. *Protein-Carbohydrate Interactions*. **1588**, (Springer New York, 2017).
61. Ribeiro, D. O., Pinheiro, B. A., Carvalho, A. L. & Palma, A. S. Targeting protein-carbohydrate interactions in plant cell-wall biodegradation: the power of carbohydrate microarrays. in *Carbohydrate Chemistry: Chemical and biological approaches* (eds. Rauter, A. P., Lindhorst, T. & Queneau, Y.) 159–176 (Royal Society of Chemistry, 2017).
62. Abbott, D. W. & Boraston, A. B. Quantitative approaches to the analysis of carbohydrate-binding module function. in *Methods in Enzymology* **510**, 211–31 (Elsevier Inc., 2012).
63. *ELISA Methods and Protocols. Methods in Molecular Biology* **1318**, (2012).
64. Makyio, H. & Kato, R. X-Ray Crystallography of Sugar Related Proteins. in *Glycoscience: Biology and Medicine* 175–182 (Springer Japan, 2015).
65. Carvalho, A. L., Santos-Silva, T., Romao, M. J., Cabrita, E. J. & Marcelo, F. Structural Elucidation of Macromolecules. in *Essential Techniques for Medical and Life Scientists: A Guide to Contemporary Methods and Current Applications with the Protocols* (ed. Tutar, Y.) 30–91 (Bentham Science Publishers, 2018).
66. Sarkar, A. & Pérez, S. *Structural Glycobiology. Structural Glycobiology* (CRC Press, 2012).
67. Fukui, S., Feizi, T., Galustian, C., Lawson, A. M. & Chai, W. Oligosaccharide microarrays for high-throughput detection and specificity assignments of carbohydrate-protein interactions. *Nat. Biotechnol.* **20**, 1011–1017 (2002).
68. Wang, D., Liu, S., Trummer, B. J., Deng, C. & Wang, A. Carbohydrate microarrays for the recognition of cross-reactive molecular markers of microbes and host cells. *Nat. Biotechnol.* **20**, 275–281 (2002).
69. Willats, W. G. T., Rasmussen, S. E., Kristensen, T., Mikkelsen, J. D. & Knox, J. P. Sugar-coated microarrays: A novel slide surface for the high-throughput analysis of glycans. *Proteomics* **2**, 1666–1671 (2002).
70. Blixt, O. *et al.* Printed covalent glycan array for ligand profiling of diverse glycan binding proteins. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 17033–17038 (2004).
71. Feizi, T. & Chai, W. Oligosaccharide microarrays to decipher the glyco code. *Nat. Rev. Mol. Cell Biol.* **5**, 582–588 (2004).
72. Palma, A. S. *et al.* Ligands for the  $\beta$ -Glucan Receptor, Dectin-1, Assigned Using “Designer” Microarrays of Oligosaccharide Probes (Neoglycolipids) Generated from Glucan Polysaccharides. *J. Biol. Chem.* **281**, 5771–5779 (2006).
73. Hyun, J. Y., Pai, J. & Shin, I. The Glycan Microarray Story from Construction to Applications. *Acc. Chem. Res.* **50**, 1069–1078 (2017).
74. Liu, Y., Palma, A. S. & Feizi, T. Carbohydrate microarrays: key developments in glycobiology. *Biol. Chem.* **390**, 647–56 (2009).
75. Rillahan, C. D. & Paulson, J. C. Glycan Microarrays for Decoding the Glycome. *Annu. Rev. Biochem.* **80**, 797–823 (2011).
76. Park, S., Gildersleeve, J. C., Blixt, O. & Shin, I. Carbohydrate microarrays. *Chem. Soc. Rev.* **42**, 4310–4326 (2013).
77. Donczko, B., Kerekyarto, J., Szurmai, Z. & Guttman, A. Glycan microarrays: new angles and new strategies. *Analyst* **139**, 2650 (2014).
78. Palma, A. S., Feizi, T., Childs, R. A., Chai, W. & Liu, Y. The neoglycolipid (NGL)-based oligosaccharide microarray system poised to decipher the meta-glycome. *Curr. Opin. Chem. Biol.* **18**, 87–94 (2014).
79. Song, X., Heimbürg-Molinari, J., Cummings, R. D. & Smith, D. F. Chemistry of Natural Glycan Microarray. *Curr Opin Chem Biol.* **0**, 70–77 (2014).
80. *Arrays. Current Opinion in Chemical Biology* **18**, (Current Opinion in Chemical Biology, 2014).
81. Palma, A. S. & Chai, W. Glycan Microarrays with Semi-synthetic Neoglycoconjugate Probes in Understanding Glycobiology. in *Synthetic Glycomes* 421–446 (Royal Society of Chemistry, 2019).
82. Tang, P. W., Gool, H. C., Hardy, M., Lee, Y. C. & Feizi, T. Novel approach to the study of the antigenicities and receptor functions of carbohydrate chains of glycoproteins. *Biochem. Biophys. Res. Commun.* **132**, 474–480 (1985).
83. Feizi, T. Carbohydrate recognition in the immune system: contributions of neoglycolipid-based microarrays to carbohydrate ligand discovery. *Ann. N. Y. Acad. Sci.* **1292**, 33–44 (2013).
84. Li, Z. & Feizi, T. The neoglycolipid (NGL) technology-based microarrays and future prospects. *FEBS Lett.* **592**, 3976–3991 (2018).
85. Liu, Y. *et al.* Neoglycolipid-Based Oligosaccharide Microarray System: Preparation of NGLs and Their Noncovalent Immobilization on Nitrocellulose-Coated Glass Slides for Microarray Analyses. in

- Carbohydrate Microarrays: Methods and Protocols, Methods in Molecular Biology* (ed. Chevolut, Y.) **808**, 117–136 (Springer Science+Business Media, 2012).
86. Liu, Y. *et al.* Neoglycolipid Probes Prepared via Oxime Ligation for Microarray Analysis of Oligosaccharide-Protein Interactions. *Chem. Biol.* **14**, 847–859 (2007).
  87. Smith, D. F., Song, X. & Cummings, R. D. Use of Glycan Microarrays to Explore Specificity of Glycan-Binding Proteins. in *Methods in Enzymology: Functional Glycomics* **480**, 417–444 (Elsevier Inc., 2010).
  88. Song, X., Lasanajak, Y., Xia, B., Smith, D. F. & Cummings, R. D. Fluorescent Glycosylamides Produced by Microscale Derivatization of Free Glycans for Natural Glycan Microarrays. *ACS Chem. Biol.* **4**, 741–750 (2009).
  89. Padler-Karavani, V. *et al.* Cross-comparison of protein recognition of sialic acid diversity on two novel sialoglycan microarrays. *J. Biol. Chem.* **287**, 22593–22608 (2012).
  90. Manimala, J. C., Roach, T. A., Li, Z. & Gildersleeve, J. C. High-throughput carbohydrate microarray analysis of 24 lectins. *Angew. Chemie - Int. Ed.* **45**, 3607–3610 (2006).
  91. Park, S., Lee, M. R., Pyo, S. J. & Shin, I. Carbohydrate Chips for Studying High-Throughput Carbohydrate-Protein Interactions. *J. Am. Chem. Soc.* **126**, 4812–4819 (2004).
  92. Pedersen, H. L. *et al.* Versatile High Resolution Oligosaccharide Microarrays for Plant Glycobiology and Cell Wall Research. *J. Biol. Chem.* **287**, 39429–39438 (2012).
  93. Feizi, T., Fazio, F., Chai, W. & Wong, C.-H. Carbohydrate microarrays — a new set of technologies at the frontiers of glycomics. *Curr. Opin. Struct. Biol.* **13**, 637–645 (2003).
  94. Varki, A. *et al.* Symbol Nomenclature for Graphical Representations of Glycans. *Glycobiology* **25**, 1323–1324 (2015).
  95. Geissner, A. *et al.* Microbe-focused glycan array screening platform. *Proc. Natl. Acad. Sci.* **116**, 1958–1967 (2019).
  96. Zhang, H. *et al.* Generation and characterization of  $\beta$ 1,2-gluco-oligosaccharide probes from *Brucella abortus* cyclic  $\beta$ -glucan and their recognition by C-type lectins of the immune system. *Glycobiology* **26**, 1086–1096 (2016).
  97. Shipp, M. *et al.* Glyco-array technology for efficient monitoring of plant cell wall glycosyltransferase activities. *Glycoconj. J.* **25**, 49–58 (2008).
  98. Kosík, O. *et al.* Polysaccharide microarrays for high-throughput screening of transglycosylase activities in plant extracts. *Glycoconj. J.* **27**, 79–87 (2010).
  99. Moller, I. *et al.* High-throughput mapping of cell-wall polymers within and between plants using novel microarrays. *Plant J.* **50**, 1118–1128 (2007).
  100. Wood, I. P. *et al.* Carbohydrate microarrays and their use for the identification of molecular markers for plant cell wall composition. *Proc. Natl. Acad. Sci.* **114**, 201619033 (2017).
  101. van Munster, J. M. *et al.* Application of carbohydrate arrays coupled with mass spectrometry to detect activity of plant-polysaccharide degradative enzymes from the fungus *Aspergillus niger*. *Sci. Rep.* **7**, 43117 (2017).
  102. Vidal-Melgosa, S. *et al.* A new versatile microarray-based method for high throughput screening of carbohydrate-active enzymes. *J. Biol. Chem.* **290**, 9020–9036 (2015).
  103. Nepogodiev, S. A., Fais, M., Hughes, D. L. & Field, R. A. Synthesis of apiose-containing oligosaccharide fragments of the plant cell wall: fragments of rhamnogalacturonan-II side chains A and B, and apiogalacturonan. *Org. Biomol. Chem.* **9**, 6670 (2011).
  104. Kaeothip, S. & Boons, G.-J. Chemical synthesis of [small beta]-arabinofuranosyl containing oligosaccharides derived from plant cell wall extensins. *Org. Biomol. Chem.* **11**, 5136–5146 (2013).
  105. Zakharova, A. N., Madsen, R. & Clausen, M. H. Synthesis of a Backbone Hexasaccharide Fragment of the Pectic Polysaccharide Rhamnogalacturonan I. *Org. Lett.* **15**, 1826–1829 (2013).
  106. Lopez, M., Fort, S., Bizot, H., Buléon, A. & Driguez, H. Chemo-enzymatic synthesis of xylogluco-oligosaccharides and their interactions with cellulose. *Carbohydr. Polym.* **88**, 185–193 (2012).
  107. Ndeh, D. *et al.* Complex pectin metabolism by gut bacteria reveals novel catalytic functions. *Nat. Publ. Gr.* (2017).
  108. Tanackovic, V. *et al.* High throughput screening of starch structures using carbohydrate microarrays. *Sci. Rep.* **6**, 1–9 (2016).
  109. Ruprecht, C. *et al.* A Synthetic Glycan Microarray Enables Epitope Mapping of Plant Cell Wall Glycan-Directed Antibodies. *Plant Physiol.* **175**, 1094–1104 (2017).
  110. Bartetzko, M. P. & Pfrenkle, F. Automated Glycan Assembly of Plant Oligosaccharides and Their Application in Cell-Wall Biology. *ChemBioChem* 1–10 (2019).
  111. Rupp, B. Biomolecular Crystallography: Principles, Practice, and Application to Structural Biology - Chapter 5. in *Garland Science Chapter 5* (Garland Science, 2009).
  112. Rhodes, G. *Crystallography Made Crystal Clear, A Guide for Users of Macromolecular Models.* Elsevier Inc. (Elsevier AP, 2006).
  113. Pérez, S. & Tvaroška, I. Carbohydrate-Protein Interactions. in *Advances in Carbohydrate Chemistry and Biochemistry* **71**, 9–136 (2014).

114. Brás, J. L. A. *et al.* Escherichia coli Expression, Purification, Crystallization, and Structure Determination of Bacterial Cohesin–Dockerin Complexes. in *Elsevier Inc.* (ed. Gilbert, H. J.) **510**, 395–415 (Elsevier Inc., 2012).
115. Puseya, M. L. *et al.* Life in the fast lane for protein crystallization and X-ray crystallography. *Prog. Biophys. Mol. Biol.* **88**, 359–386 (2005).
116. Karplus, P. A. & Diederichs, K. Linking crystallographic model and data quality. M&M, supporting info. *Science* **336**, 1030–3 (2012).
117. Pérez, S. & De Sanctis, D. Glycoscience@Synchrotron: Synchrotron radiation applied to structural glycoscience. *Beilstein J. Org. Chem.* **13**, 1145–1167 (2017).
118. Brünger, A. T. Free R value: a novel statistical quantity for assessing the accuracy of crystal structures. *Nature* **355**, 472–475 (1992).
119. Joosten, R. P. & Lutteke, T. Carbohydrate 3D structure validation. *Curr. Opin. Struct. Biol.* **44**, 9–17 (2017).
120. Agirre, J. *et al.* Privateer: software for the conformational validation of carbohydrate structures. *Nat. Struct. Mol. Biol.* **22**, 833–834 (2015).
121. Nishio, M., Umezawa, Y., Fantini, J., Weiss, M. S. & Chakrabarti, P. CH– $\pi$  hydrogen bonds in biological macromolecules. *Phys. Chem. Chem. Phys.* **16**, 12648–12683 (2014).
122. Hudson, K. L. *et al.* Carbohydrate-Aromatic Interactions in Proteins. *J. Am. Chem. Soc.* **137**, 15152–15160 (2015).
123. Spiwok, V. CH/ $\pi$  Interactions in Carbohydrate Recognition. *Molecules* **22**, 1038 (2017).
124. Darby, J. F. *et al.* Water Networks Can Determine the Affinity of Ligand Binding to Proteins. *J. Am. Chem. Soc.* **141**, 15818–15826 (2019).
125. Spyrakakis, F. *et al.* The Roles of Water in the Protein Matrix: A Largely Untapped Resource for Drug Discovery. *J. Med. Chem.* **60**, 6781–6827 (2017).
126. Boraston, A. B. *et al.* A structural and functional analysis of  $\alpha$ -glucan recognition by family 25 and 26 carbohydrate-binding modules reveals a conserved mode of starch recognition. *J. Biol. Chem.* **281**, 587–598 (2006).
127. Blumer-Schuette, S. E. *et al.* Thermophilic lignocellulose deconstruction. *FEMS Microbiol. Rev.* **38**, 393–448 (2014).
128. Okarter, N. & Liu, R. H. Health Benefits of Whole Grain Phytochemicals. *Crit. Rev. Food Sci. Nutr.* **50**, 193–208 (2010).
129. Haab, B. B. & Klamer, Z. Advances in Tools to Determine the Glycan-Binding Specificities of Lectins and Antibodies. *Mol. Cell. Proteomics* **19**, 224–232 (2020).
130. Pattathil, S., Avci, U., Zhang, T., Cardenas, C. L. & Hahn, M. G. Immunological Approaches to Biomass Characterization and Utilization. *Front. Bioeng. Biotechnol.* **3**, 1–14 (2015).
131. Liu, Y., Palma, A. S., Feizi, T. & Chai, W. Insights Into Glucan Polysaccharide Recognition Using Glucooligosaccharide Microarrays With Oxime-Linked Neoglycolipid Probes. in *Methods in Enzymology* **598**, 139–167 (Elsevier Inc., 2018).
132. Schallus, T. *et al.* Malectin: A Novel Carbohydrate-binding Protein of the Endoplasmic Reticulum and a Candidate Player in the Early Steps of Protein N -Glycosylation. *Mol. Biol. Cell* **19**, 3404–3414 (2008).
133. Palma, A. S. *et al.* Multifaceted approaches including neoglycolipid oligosaccharide microarrays to ligand discovery for malectin. *Methods Enzymol.* **478**, 265–86 (2010).
134. Meikle, P. J., Hoogenraad, N. J., Bonig, I., Clarke, A. E. & Stone, B. A. A (1 $\rightarrow$ 3, 1 $\rightarrow$ 4)- $\beta$ -glucan-specific monoclonal antibody and its use in the quantitation and immunocytochemical location of (1 $\rightarrow$ 3, 1 $\rightarrow$ 4)- $\beta$ -glucans. *Plant J.* **5**, 1–9 (1994).
135. McCartney, L., Marcus, S. E. & Knox, J. P. Monoclonal antibodies to plant cell wall xylans and arabinoxylans. *J. Histochem. Cytochem.* **53**, 543–546 (2005).
136. Charnock, S. J. *et al.* The X6 ‘thermostabilizing’ domains of xylanases are carbohydrate-binding modules: Structure and biochemistry of the Clostridium thermocellum X6b domain. *Biochemistry* **39**, 5013–5021 (2000).
137. Willats, W. G. T., Marcus, S. E. & Knox, J. P. Generation of a monoclonal antibody specific to (1 $\rightarrow$ 5)- $\alpha$ -l-arabinan. *Carbohydr. Res.* **308**, 149–152 (1998).
138. Pettolino, F. A. *et al.* A (1 $\rightarrow$ 4)- $\beta$ -mannan-specific monoclonal antibody and its use in the immunocytochemical location of galactomannans. *Planta* **214**, 235–242 (2001).
139. Marcus, S. E. *et al.* Restricted access of proteins to mannan polysaccharides in intact plant cell walls. *Plant J.* **64**, 191–203 (2010).
140. Pattathil, S. *et al.* A Comprehensive Toolkit of Plant Cell Wall Glycan-Directed Monoclonal Antibodies. *Plant Physiol.* **153**, 514–525 (2010).
141. Ghosh, A. *et al.* Mannan specific family 35 carbohydrate-binding module (CtCBM35) of Clostridium thermocellum: structure analysis and ligand binding. *Biologia (Bratisl)*. **69**, 1271–1282 (2014).
142. Plancot, B. *et al.* Desiccation tolerance in plants: Structural characterization of the cell wall hemicellulosic polysaccharides in three Selaginella species. *Carbohydr. Polym.* **208**, 180–190 (2019).

## REFERENCES

143. Kochibe, N. & C, K. Purification and properties of a novel fucose-specific hemagglutinin of *Aleuria aurantia*. *Biochemistry* **19**, 2841–2846 (1980).
144. Lahaye, M., Rondeau-Mouro, C., Deniaud, E. & Buléon, A. Solid-state <sup>13</sup>C NMR spectroscopy studies of xylans in the cell wall of *Palmaria palmata* (L. Kuntze, Rhodophyta). *Carbohydr. Res.* **338**, 1559–1569 (2003).
145. Chai, W., Stoll, M. S., Galustian, C., Lawson, A. M. & Feizi, T. Neoglycolipid Technology: Deciphering Information Content of Glycome. in 160–195 (2003).
146. Liu, Y. *et al.* The minimum information required for a glycomics experiment project: improving the standards for reporting glycan microarray-based data. *Glycobiology* **27**, 1–5 (2016).
147. Moller, I. *et al.* High-throughput screening of monoclonal antibodies against plant cell wall glycans by hierarchical clustering of their carbohydrate microarray binding profiles. *Glycoconj Journal* **37–48** (2008).
148. Stoll, M. & Feizi, T. Software Tools for Storing, Processing and Displaying Carbohydrate Microarray Data. in *Glyco-Bioinformatics – Bits ‘n’ Bytes of Sugars* 123–140 (Beilstein-Institut, 2009).
149. Taylor, M. E. & Drickamer, K. Convergent and divergent mechanisms of sugar recognition across kingdoms. *Curr. Opin. Struct. Biol.* **28**, 14–22 (2014).
150. Wang, L. *et al.* Cross-platform comparison of glycan microarray formats. *Glycobiology* **24**, 507–517 (2014).
151. Meikle, P. J., Bonig, I., Hoogenraad, N. J., Clarke, a E. & Stone, B. a. The location of (1→3)- $\beta$ -glucans in the walls of pollen tubes of *Nicotiana glauca* using a (1→3)- $\beta$ -glucan-specific monoclonal antibody. *Planta* **185**, 1–8 (1991).
152. Wang, Y. *et al.* Specificities of *Ricinus communis* agglutinin 120 interaction with sulfated galactose. *FEBS Lett.* **585**, 3927–3934 (2011).
153. Tormo, J. *et al.* Crystal structure of a bacterial family-III cellulose-binding domain: a general mechanism for attachment to cellulose. *EMBO J.* **15**, 5739–5751 (1996).
154. Petkun, S. *et al.* Reassembly and co-crystallization of a family 9 processive endoglucanase from its component parts: Structural and functional significance of the intermodular linker. *PeerJ Prepr.* **3**, e1382 (2015).
155. Alahuhta, M. *et al.* The unique binding mode of cellulosomal CBM4 from *Clostridium thermocellum* cellobiohydrolase A. *J. Mol. Biol.* **402**, 374–387 (2010).
156. Jiang, D. *et al.* Crystal structure of 1,3Gal43A, an exo- $\beta$ -1,3-galactanase from *Clostridium thermocellum*. *J. Struct. Biol.* **180**, 447–457 (2012).
157. Najmudin, S. *et al.* Xyloglucan Is Recognized by Carbohydrate-binding Modules That Interact with  $\beta$ -Glucan Chains. *J. Biol. Chem.* **281**, 8815–8828 (2006).
158. Mizutani, K. *et al.* Influence of a mannan binding family 32 carbohydrate binding module on the activity of the appended mannanase. *Appl. Environ. Microbiol.* **78**, 4781–4787 (2012).
159. Montanier, C. Y. *et al.* A novel, noncatalytic carbohydrate-binding module displays specificity for galactose-containing polysaccharides through calcium-mediated oligomerization. *J. Biol. Chem.* **286**, 22499–22509 (2011).
160. Poole, D. M. *et al.* Identification of the cellulose-binding domain of the cellulosome subunit S1 from *Clostridium thermocellum* YS. *FEMS Microbiol. Lett.* **78**, 181–6 (1992).
161. Yaniv, O., Frolow, F., Levy-Assraf, M., Lamed, R. & Bayer, E. A. *Interactions between family 3 carbohydrate binding modules (CBMs) and cellulosomal linker peptides. Methods in Enzymology* **510**, (Elsevier Inc., 2012).
162. Shimon, L. J. W. *et al.* Structure of a family IIIa scaffoldin CBD from the cellulosome of *Clostridium cellulolyticum* at 2.2 Å resolution. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **56**, 1560–1568 (2000).
163. Eklöf, J. & Hehemann, J.-H. Glycoside Hydrolase Family 16. Available at: [https://www.cazypedia.org/index.php/Glycoside\\_Hydrolase\\_Family\\_16](https://www.cazypedia.org/index.php/Glycoside_Hydrolase_Family_16). (Accessed: 24th March 2020)
164. Wilson, D. B. & Urbanowicz, B. Glycoside Hydrolase Family 9. Available at: [https://www.cazypedia.org/index.php/Glycoside\\_Hydrolase\\_Family\\_9](https://www.cazypedia.org/index.php/Glycoside_Hydrolase_Family_9). (Accessed: 24th March 2020)
165. Najmudin, S. *et al.* Xyloglucan is recognized by carbohydrate-binding modules that interact with beta-glucan chains. *J. Biol. Chem.* **281**, 8815–28 (2006).
166. Correia, M. A. S. *et al.* Signature Active Site Architectures Illuminate the Molecular Basis for Ligand Specificity in Family 35 Carbohydrate Binding Module. *Biochemistry* **49**, 6193–6205 (2010).
167. Ghosh, A., Verma, A. K., Gautam, S., Gupta, M. N. & Goyal, A. Structure and functional investigation of ligand binding by a family 35 carbohydrate binding module (CtCBM35) of  $\beta$ -mannanase of family 26 glycoside hydrolase from *Clostridium thermocellum*. *Biochem.* **79**, 672–686 (2014).
168. Mizutani, K., Sakka, M., Kimura, T. & Sakka, K. Essential role of a family-32 carbohydrate-binding module in substrate recognition by *Clostridium thermocellum* mannanase CtMan5A. *FEBS Lett.* **588**, 1726–1730 (2014).
169. Fujimoto, Z. Structure and Function of Carbohydrate-Binding Module Families 13 and 42 of Glycoside Hydrolases, Comprising a  $\beta$ -Trefoil Fold. *Biosci. Biotechnol. Biochem.* **77**, 1363–1371

- (2013).
170. Correia, M. A. S. *et al.* Structure and function of an arabinoxylan-specific xylanase. *J. Biol. Chem.* **286**, 22510–22520 (2011).
  171. Boraston, A. B. *et al.* Structure and Ligand Binding of Carbohydrate-binding Module CsCBM6-3 Reveals Similarities with Fucose-specific Lectins and “Galactose-binding” Domains. *J. Mol. Biol.* **327**, 659–669 (2003).
  172. Czjzek, M. *et al.* The Location of the Ligand-binding Site of Carbohydrate-binding Modules That Have Evolved from a Common Sequence Is Not Conserved. *J. Biol. Chem.* **276**, 48580–48587 (2001).
  173. Mewis, K., Lenfant, N., Lombard, V. & Henrissat, B. Dividing the Large Glycoside Hydrolase Family 43 into Subfamilies: a Motivation for Detailed Enzyme Characterization. *Appl. Environ. Microbiol.* **82**, 1686–1692 (2016).
  174. Bae, B. *et al.* Molecular Basis for the Selectivity and Specificity of Ligand Recognition by the Family 16 Carbohydrate-binding Modules from *Thermoanaerobacterium polysaccharolyticum* ManA. *J. Biol. Chem.* **283**, 12415–12425 (2008).
  175. Janecek, S. & Svensson, B. Carbohydrate Binding Module Family 48. Available at: [https://www.cazypedia.org/index.php/Carbohydrate\\_Binding\\_Module\\_Family\\_48](https://www.cazypedia.org/index.php/Carbohydrate_Binding_Module_Family_48). (Accessed: 24th March 2020)
  176. Georgelis, N., Tabuchi, A., Nikolaidis, N. & Cosgrove, D. J. Structure-function analysis of the bacterial expansin EXLX1. *J. Biol. Chem.* **286**, 16814–16823 (2011).
  177. Pereira, F. C. *et al.* A LysM Domain Intervenes in Sequential Protein-Protein and Protein-Peptidoglycan Interactions Important for Spore Coat Assembly in *Bacillus subtilis*. *J. Bacteriol.* **201**, 1–19 (2018).
  178. Dai, X. *et al.* Metatranscriptomic analyses of plant cell wall polysaccharide degradation by microorganisms in the cow rumen. *Appl. Environ. Microbiol.* **81**, 1375–1386 (2015).
  179. Sequeira, A. F. *et al.* A Novel Platform for High-Throughput Gene Synthesis to Maximize Recombinant Expression in *Escherichia coli*. in *PCR: Methods and Protocols, Methods in Molecular Biology* (ed. Lucilia Domingues) **1620**, 113–128 (Springer Science+Business Media LLC, 2017).
  180. Fry, S. C., Nesselrode, B. H. W. A., Miller, J. G. & Mewburn, B. R. Mixed-linkage (1→3,1→4)-β-D-glucan is a major hemicellulose of *Equisetum* (horsetail) cell walls. *New Phytol.* **179**, 104–115 (2008).
  181. Burton, R. A. & Fincher, G. B. (1,3;1,4)-β-D-Glucans in Cell Walls of the Poaceae, Lower Plants, and Fungi: A Tale of Two Linkages. *Mol. Plant* **2**, 873–882 (2009).
  182. Burton, R. A. & Fincher, G. B. Evolution and development of cell walls in cereal grains. *Front. Plant Sci.* **5**, 1–15 (2014).
  183. Stone, B. A. & Clarke, A. E. (1-3), (1-4)-β-Glucans in Higher Plants. in *Chemistry and Biology of (1-3)-β-Glucans* 431–491 (La Trobe University Press, 1992).
  184. Tamura, K. *et al.* Molecular Mechanism by which Prominent Human Gut Bacteroidetes Utilize Mixed-Linkage Beta-Glucans, Major Health-Promoting Cereal Polysaccharides. *Cell Rep.* **21**, 417–430 (2017).
  185. Kiemle, S. N. *et al.* Role of (1,3)(1,4)-β-Glucan in Cell Walls: Interaction with Cellulose. *Biomacromolecules* **15**, 1727–1736 (2014).
  186. Wade Abbott, D. & Boraston, A. B. Interactions between Proteins and (1,3)-β-Glucans and Related Polysaccharides. in *Chemistry, Biochemistry, and Biology of 1-3 Beta Glucans and Related Polysaccharides* 171–199 (Elsevier, 2009).
  187. Baumann, M. J. *et al.* Structural Evidence for the Evolution of Xyloglucanase Activity from Xyloglucan Endo -Transglycosylases: Biological Implications for Cell Wall Metabolism. *Plant Cell* **19**, 1947–1963 (2007).
  188. Himmel, M. E. *et al.* Biomass Recalcitrance: Engineering Plants and Enzymes for Biofuels Production. *Science (80-. )*. **315**, 804–807 (2007).
  189. Sommer, P., Georgieva, T. & Ahring, B. K. Potential for using thermophilic anaerobic bacteria for bioethanol production from hemicellulose. *Biochem. Soc. Trans.* **32**, 283–289 (2004).
  190. Volkov, I. I., Lunina, N. A. & Velikodvorskaia, G. A. Prospects for practical application of substrate-binding modules of glycosyl hydrolases (A review). *Prikl. Biokhim. Mikrobiol.* **40**, 499–504 (2004).
  191. Viegas, A. *et al.* Solution structure, dynamics and binding studies of a family 11 carbohydrate-binding module from *Clostridium thermocellum* (CtCBM11). *Biochem. J.* **451**, 289–300 (2013).
  192. Viegas, A. *et al.* Molecular determinants of ligand specificity in family 11 carbohydrate binding modules - an NMR, X-ray crystallography and computational chemistry approach. *FEBS J.* **275**, 2524–2535 (2008).
  193. Fonseca-Maldonado, R. *et al.* Lignocellulose binding of a Cel5A-RtCBM11 chimera with enhanced β-glucanase activity monitored by electron paramagnetic resonance. *Biotechnol. Biofuels* **10**, 269 (2017).
  194. Cattaneo, C., Cesaro, P., Spertino, S., Icardi, S. & Cavaletto, M. Enhanced features of Dictyoglomus

- turgidum Cellulase A engineered with carbohydrate binding module 11 from *Clostridium thermocellum*. *Sci. Rep.* **8**, 4402 (2018).
195. Furtado, G. P. *et al.* Engineering the affinity of a family 11 carbohydrate binding module to improve binding of branched over unbranched polysaccharides. *Int. J. Biol. Macromol.* **120**, 2509–2516 (2018).
  196. Chojnacki, S., Cowley, A., Lee, J., Foix, A. & Lopez, R. Programmatic access to bioinformatics tools from EMBL-EBI update: 2017. *Nucleic Acids Res.* **45**, W550–W553 (2017).
  197. Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M. & Barton, G. J. Jalview Version 2 - a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**, 1189–1191 (2009).
  198. Yuan, S. F. *et al.* Biochemical characterization and structural analysis of a bifunctional cellulase/xylanase from *Clostridium thermocellum*. *J. Biol. Chem.* **290**, 5739–5748 (2015).
  199. Taylor, E. J. *et al.* How family 26 glycoside hydrolases orchestrate catalysis on different polysaccharides: Structure and activity of a *Clostridium thermocellum* lichenase, CtLic26A. *J. Biol. Chem.* **280**, 32761–32767 (2005).
  200. Venditto, I. *et al.* Family 46 Carbohydrate-binding Modules Contribute to the Enzymatic Hydrolysis of Xyloglucan and  $\beta$ -1,3–1,4-Glucans through Distinct Mechanisms. *J. Biol. Chem.* **290**, 10572–10586 (2015).
  201. Luís, A. S. *et al.* Understanding How Noncatalytic Carbohydrate Binding Modules Can Display Specificity for Xyloglucan. *J. Biol. Chem.* **288**, 4799–4809 (2013).
  202. Leslie, A. G. W. Recent changes to the MOSFLM package for processing film and image plate data. *CCP4 ESF-EAMCB Newsl. Protein Crystallogr.* **26**, (1992).
  203. Kabsch, W. Automatic indexing of rotation diffraction patterns. *J. Appl. Crystallogr.* **21**, 67–72 (1988).
  204. Collaborative Computational Project, N. 4. The CCP4 suite: programs for protein crystallography. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **50**, 760–763 (1994).
  205. McCoy, A. J. *et al.* Phaser crystallographic software. *J. Appl. Crystallogr.* **40**, 658–674 (2007).
  206. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **66**, 486–501 (2010).
  207. Murshudov, G. N. *et al.* REFMAC5 for the refinement of macromolecular crystal structures. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **67**, 355–367 (2011).
  208. Adams, P. D. *et al.* PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **66**, 213–221 (2010).
  209. Chen, V. B. *et al.* MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **66**, 12–21 (2010).
  210. Joosten, R. P., Joosten, K., Cohen, S. X., Vriend, G. & Perrakis, A. Automatic rebuilding and optimization of crystallographic structures in the Protein Data Bank. *Bioinformatics* **27**, 3392–3398 (2011).
  211. Bolam, D. N. *et al.* X4 Modules Represent a New Family of Carbohydrate-binding Modules That Display Novel Properties. *J. Biol. Chem.* **279**, 22953–22963 (2004).
  212. Ohnuma, T. & Taira, T. Carbohydrate Binding Module Family 50. *CAZypedia* Available at: [https://www.cazypedia.org/index.php/Carbohydrate\\_Binding\\_Module\\_Family\\_50](https://www.cazypedia.org/index.php/Carbohydrate_Binding_Module_Family_50). (Accessed: 24th March 2020)
  213. Garvey, K. J., Saedi, M. S. & Ito, J. Nucleotide sequence of *Bacillus* phage  $\emptyset$ 29 genes 14 and 15: homology of gene 15 with other phage lysozymes. *Nucleic Acids Res.* **14**, 10001–10008 (1986).
  214. Desvaux, M., Dumas, E., Chafsey, I. & HÅ©braud, M. Protein cell surface display in Gram-positive bacteria: from single protein to macromolecular protein structure. *FEMS Microbiol. Lett.* **256**, 1–15 (2006).
  215. Mesnage, S. *et al.* Molecular basis for bacterial peptidoglycan recognition by LysM domains. *Nat. Commun.* **5**, 4269 (2014).
  216. Ohnuma, T., Onaga, S., Murata, K., Taira, T. & Katoh, E. LysM Domains from *Pteris ryukyuensis* Chitinase. *J. Biol. Chem.* **283**, 5178–5187 (2008).
  217. Meroueh, S. O. *et al.* Three-dimensional structure of the bacterial cell wall peptidoglycan. *Proc. Natl. Acad. Sci.* **103**, 4404–4409 (2006).
  218. Wong, J. E. M. M. *et al.* Cooperative binding of LysM domains determines the carbohydrate affinity of a bacterial endopeptidase protein. *FEBS J.* **281**, 1196–1208 (2014).
  219. Visweswaran, G. R. R., Leenhouts, K., Van Roosmalen, M., Kok, J. & Buist, G. Exploiting the peptidoglycan-binding motif, LysM, for medical and industrial applications. *Appl. Microbiol. Biotechnol.* **98**, 4331–4345 (2014).
  220. Bosma, T. *et al.* Novel Surface Display System for Proteins on Non-Genetically Modified Gram-Positive Bacteria. *Appl. Environ. Microbiol.* **72**, 880–889 (2006).
  221. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
  222. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* **47**, D506–D515 (2019).
  223. Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).

224. Wong, J. E. M. M. M. *et al.* An intermolecular binding mechanism involving multiple LysM domains mediates carbohydrate recognition by an endopeptidase. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **71**, 592–605 (2015).
225. Ohnuma, T. *et al.* Crystal structure and thermodynamic dissection of chitin oligosaccharide binding to the LysM module of chitinase-A from *Pteris ryukyensis*. *Biochem. Biophys. Res. Commun.* **494**, 736–741 (2017).
226. Vaz, F. *et al.* Accessibility to Peptidoglycan Is Important for the Recognition of Gram-Positive Bacteria in *Drosophila*. *Cell Rep.* **27**, 2480–2492 (2019).
227. Sánchez-Vallet, A. *et al.* Fungal effector Ecp6 outcompetes host immune receptor for chitin binding through intrachain LysM dimerization. *Elife* **2**, 1–16 (2013).
228. Kitaoku, Y., Fukamizo, T., Numata, T. & Ohnuma, T. Chitin oligosaccharide binding to the lysin motif of a novel type of chitinase from the multicellular green alga, *Volvox carteri*. *Plant Mol. Biol.* **93**, 97–108 (2017).
229. Robert, X. & Gouet, P. Deciphering key features in protein structures with the new ENDscript server. *Nucleic Acids Res.* **42**, W320–W324 (2014).
230. Mearls, E. B., Izquierdo, J. A. & Lynd, L. R. Formation and characterization of non-growth states in *Clostridium thermocellum*: spores and L-forms. *BMC Microbiol.* **12**, 180 (2012).
231. Glauner, B., Höltje, J. V & Schwarz, U. The composition of the murein of *Escherichia coli*. *J. Biol. Chem.* **263**, 10088–95 (1988).
232. Atilano, M. L., Yates, J., Glittenberg, M., Filipe, S. R. & Ligoxygakis, P. Wall Teichoic Acids of *Staphylococcus aureus* Limit Recognition by the *Drosophila* Peptidoglycan Recognition Protein-SA to Promote Pathogenicity. *PLoS Pathog.* **7**, 1–13 (2011).
233. Laskowski, R. A., MacArthur, M. W., Moss, D. S. & Thornton, J. M. PROCHECK: a program to check the stereochemical quality of protein structures. *J. Appl. Cryst.* **26**, 283–291 (1993).
234. Vaguine, A. A., Richelle, J. & Wodak, S. J. SFCHECK: a unified set of procedures for evaluating the quality of macromolecular structure-factor data and their agreement with the atomic model. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **55**, 191–205 (1999).
235. Case, D. A. *et al.* AMBER 12. (2012).
236. Hornak, V. *et al.* Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins Struct. Funct. Bioinforma.* **65**, 712–725 (2006).
237. Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A. & Case, D. A. Development and testing of a general amber force field. *J. Comput. Chem.* **25**, 1157–1174 (2004).
238. Izaguirre, J. A., Catarello, D. P., Wozniak, J. M. & Skeel, R. D. Langevin stabilization of molecular dynamics. *J. Chem. Phys.* **114**, 2090–2098 (2001).
239. Akinoshio, H., Yee, K., Close, D. & Ragauskas, A. The emergence of *Clostridium thermocellum* as a high utility candidate for consolidated bioprocessing applications. *Front. Chem.* **2**, 66 (2014).
240. Essmann, U. *et al.* A smooth particle mesh Ewald method. *J. Chem. Phys.* **103**, 8577–8593 (1995).
241. Ryckaert, J.-P., Ciccotti, G. & Berendsen, H. J. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comput. Phys.* **23**, 327–341 (1977).
242. Roe, D. R. & Cheatham, T. E. PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *J. Chem. Theory Comput.* **9**, 3084–3095 (2013).
243. Humphrey, W., Dalke, A. & Schulten, K. VMD: Visual molecular dynamics. *J. Mol. Graph.* **14**, 33–38 (1996).
244. Kollman, P. A. *et al.* Calculating Structures and Free Energies of Complex Molecules: Combining Molecular Mechanics and Continuum Models. *Acc. Chem. Res.* **33**, 889–897 (2000).
245. Wefers, D., Flörchinger, R. & Bunzel, M. Detailed Structural Characterization of Arabinans and Galactans of 14 Apple Cultivars Before and After Cold Storage. *Front. Plant Sci.* **9**, 1–12 (2018).
246. Anderson, C. T. Pectic Polysaccharides in Plants: Structure, Biosynthesis, Functions, and Applications. in *Extracellular Sugar-Based Biopolymers Matrices. Biologically-Inspired Systems* (eds. Cohen, E. & Merzendorfer, H.) **12**, 433–484 (Springer International Publishing, 2019).
247. Harholt, J., Suttangkakul, A. & Vibe Scheller, H. Biosynthesis of Pectin. *Plant Physiol.* **153**, 384–395 (2010).
248. Verhertbruggen, Y. *et al.* Developmental complexity of arabinan polysaccharides and their processing in plant cell walls. *Plant J.* **59**, 413–425 (2009).
249. Hashimoto, H. Recent structural studies of carbohydrate-binding modules. *Cell. Mol. Life Sci.* **63**, 2954–2967 (2006).
250. Fujimoto, Z. *et al.* Crystal structure of *Streptomyces olivaceoviridis* E-86  $\beta$ -xylanase containing xylan-binding domain. *J. Mol. Biol.* **300**, 575–585 (2000).
251. Ichinose, H. *et al.* A  $\beta$ -l-Arabinopyranosidase from *Streptomyces avermitilis* Is a Novel Member of Glycoside Hydrolase Family 27. *J. Biol. Chem.* **284**, 25097–25106 (2009).
252. Marchler-Bauer, A. *et al.* CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. *Nucleic Acids Res.* **45**, D200–D203 (2017).
253. Fujimoto, Z. *et al.* Crystal structures of the sugar complexes of *Streptomyces olivaceoviridis* E-86

- xylanase: sugar binding structure of the family 13 carbohydrate binding module. *J. Mol. Biol.* **316**, 65–78 (2002).
254. Labourel, A. *et al.* The Mechanism by Which Arabinoxylanases Can Recognize Highly Decorated Xylans. *J. Biol. Chem.* **291**, 22149–22159 (2016).
255. Jancarik, J. & Kim, S. H. Sparse matrix sampling: a screening method for crystallization of proteins. *J. Appl. Crystallogr.* **24**, 409–411 (1991).
256. Ben David, Y. *et al.* Ruminococcal cellulosome systems from rumen to human. *Environ. Microbiol.* **17**, (2015).
257. Fry, S. C. *et al.* An unambiguous nomenclature for xyloglucan-derived oligosaccharides. *Physiol. Plant.* **89**, 1–3 (1993).
258. Puhlmann, J. *et al.* Generation of Monoclonal Antibodies against Plant Cell-Wall Polysaccharides: I. Characterization of a Monoclonal Antibody to a Terminal  $\alpha$ -(1→2)-Linked Fucosyl-Containing Epitope. *Plant Physiol.* **104**, 699–710 (1994).
259. Jones, L., Seymour, G. B. & Knox, J. P. Localization of Pectic Galactan in Tomato Cell Walls Using a Monoclonal Antibody Specific to (1-4)- $\beta$ -D-Galactan. *Plant Physiol.* **113**, 1405–1412 (1997).
260. Nagata, Y. & Burger, M. M. Wheat germ agglutinin. Molecular characteristics and specificity for sugar binding. *J. Biol. Chem.* **249**, 3116–22 (1974).
261. Crowley, J. F., Goldstein, I. J., Arnarp, J. & Lönngrén, J. Carbohydrate binding studies on the lectin from *Datura stramonium* seeds. *Arch. Biochem. Biophys.* **231**, 524–533 (1984).
262. Baenziger, J. U. & Fiete, D. Structural determinants of Ricinus communis agglutinin and toxin specificity for oligosaccharides. *J. Biol. Chem.* **254**, 9795–9 (1979).
263. Khamlue, R., Naksupan, N., Ounaroon, A. & Saelim, N. Skin Wound Healing Promoting Effect of Polysaccharides Extracts from *Tremella fuciformis* and *Auricularia auricula* on the ex-vivo Porcine Skin Wound Healing Model. *IPCBE* **43**, 93–98 (2012).
264. Takahara, K. *et al.* Difference in Fine Specificity to Polysaccharides of *Candida albicans* Mannoprotein between Mouse SIGNR1 and Human DC-SIGN. *Infect. Immun.* **80**, 1699–1706 (2012).
265. Pinto, M., Coelho, E., Nunes, A., Brandão, T. & Coimbra, M. A. Valuation of brewers spent yeast polysaccharides: A structural characterization approach. *Carbohydr. Polym.* **116**, 215–222 (2015).
266. Haworth, W. N., Hirst, E. L. & Isherwood, F. A. Polysaccharides. Part XXIV. Yeast mannan. *J. Chem. Soc.* 784 (1937).
267. McCleary, B. V. & Matheson, N. K. Enzymic Analysis of Polysaccharide Structure. in 147–276 (1987).
268. Zhang, H.-T. *et al.* Improved curdlan fermentation process based on optimization of dissolved oxygen combined with pH control and metabolic characterization of *Agrobacterium* sp. ATCC 31749. *Appl. Microbiol. Biotechnol.* **93**, 367–379 (2012).
269. de la Cruz, J., Pintor-Toro, J. A., Benítez, T. & Llobell, A. Purification and characterization of an endo-beta-1,6-glucanase from *Trichoderma harzianum* that is related to its mycoparasitism. *J. Bacteriol.* **177**, 1864–1871 (1995).
270. Hong, F. *et al.* Beta-glucan functions as an adjuvant for monoclonal antibody immunotherapy by recruiting tumoricidal granulocytes as killer cells. *Cancer Res.* **63**, 9023–31 (2003).
271. Jamas, S., Easson, D. D., Ostroff, G. R. & Onderdonk, A. B. PGG-Glucans. A Novel Class of Macrophage-Activating Immunomodulators. in 44–51 (1991).
272. Wang, X., Xu, X. & Zhang, L. Thermally Induced Conformation Transition of Triple-Helical Lentinan in NaCl Aqueous Solution. *J. Phys. Chem. B* **112**, 10343–10351 (2008).
273. Du, Y. *et al.* Synthesis and antitumor activities of glucan derivatives. *Tetrahedron* **60**, 6345–6351 (2004).
274. Yoo, D.-H., Lee, B.-H., Chang, P.-S., Lee, H. G. & Yoo, S.-H. Improved Quantitative Analysis of Oligosaccharides from Lichenase-Hydrolyzed Water-Soluble Barley  $\beta$ -Glucans by High-Performance Anion-Exchange Chromatography. *J. Agric. Food Chem.* **55**, 1656–1662 (2007).
275. Nunes, C., Saraiva, J. A. & Coimbra, M. A. Effect of candying on cell wall polysaccharides of plums (*Prunus domestica* L.) and influence of cell wall enzymes. *Food Chem.* **111**, 538–548 (2008).
276. Coelho, E., Rocha, M. A. M., Moreira, A. S. P., Domingues, M. R. M. & Coimbra, M. A. Revisiting the structural features of arabinoxylans from brewers' spent grain. *Carbohydr. Polym.* **139**, 167–176 (2016).
277. Passos, C. P. & Coimbra, M. A. Microwave superheated water extraction of polysaccharides from spent coffee grounds. *Carbohydr. Polym.* **94**, 626–633 (2013).
278. Ho, G. T. T., Zou, Y.-F., Aslaksen, T. H., Wangensteen, H. & Barsett, H. Structural characterization of bioactive pectic polysaccharides from elderflowers (*Sambuciflos*). *Carbohydr. Polym.* **135**, 128–137 (2016).
279. Ho, G. T. T., Zou, Y.-F., Wangensteen, H. & Barsett, H. RG-I regions from elderflower pectins substituted on GalA are strong immunomodulators. *Int. J. Biol. Macromol.* **92**, 731–738 (2016).
280. Grønhaug, T. E. *et al.* Beta-D-(1→4)-galactan-containing side chains in RG-I regions of pectic polysaccharides from *Biophytum petersianum* Klotzsch. contribute to expression of immunomodulating activity against intestinal Peyer's patch cells and macrophages. *Phytochemistry*



- 72**, 2139–2147 (2011).
281. Inngjerdingen, K. T. *et al.* Bioactive pectic polysaccharides from *Glinus oppositifolius* (L.) Aug. DC., a Malian medicinal plant, isolation and partial characterization. *J. Ethnopharmacol.* **101**, 204–214 (2005).
282. Nergard, C. S. *et al.* Isolation, partial characterisation and immunomodulating activities of polysaccharides from *Vernonia kotschyana* Sch. Bip. ex Walp. *J. Ethnopharmacol.* **91**, 141–152 (2004).
283. Tvette Inngjerdingen, K. *et al.* A comparison of bioactive aqueous extracts and polysaccharide fractions from roots of wild and cultivated *Cochlospermum tinctorium* A. Rich. *Phytochemistry* **93**, 136–143 (2013).
284. Gronhaug, T. E. *et al.* Bioactive arabinogalactans from the leaves of *Opilia celtidifolia* Endl. ex Walp. (Opiliaceae). *Glycobiology* **20**, 1654–1664 (2010).
285. Austarheim, I. *et al.* Chemical and biological characterization of pectin-like polysaccharides from the bark of the Malian medicinal tree *Cola cordifolia*. *Carbohydr. Polym.* **89**, 259–268 (2012).
286. Zou, Y.-F. *et al.* Polysaccharides with immunomodulating properties from the bark of *Parkia biglobosa*. *Carbohydr. Polym.* **101**, 457–463 (2014).
287. Cheng, Y.-S. *et al.* Crystal structure and substrate-binding mode of cellulase 12A from *Thermotoga maritima*. *Proteins Struct. Funct. Bioinforma.* **79**, 1193–1204 (2011).
288. Krissinel, E. & Henrick, K. Inference of Macromolecular Assemblies from Crystalline State. *J. Mol. Biol.* **372**, 774–797 (2007).

*REFERENCES*

# APPENDIX

---

## SUPPLEMENTARY INFORMATION



## Chapter 2 - Supplementary Information

## Supplementary Figures

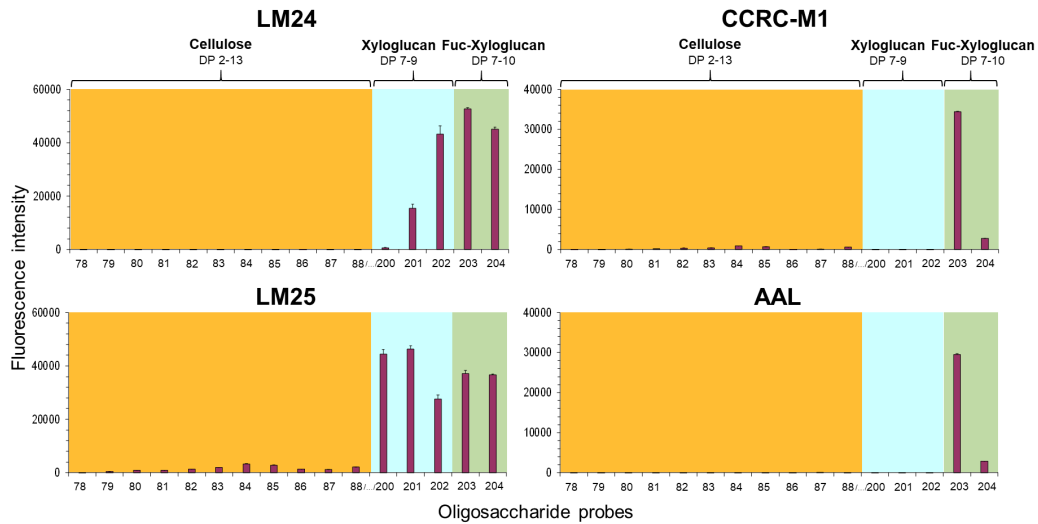


Figure S2.1. Validation of the xyloglucan series included in the glucan and hemicellulose oligosaccharide microarrays.

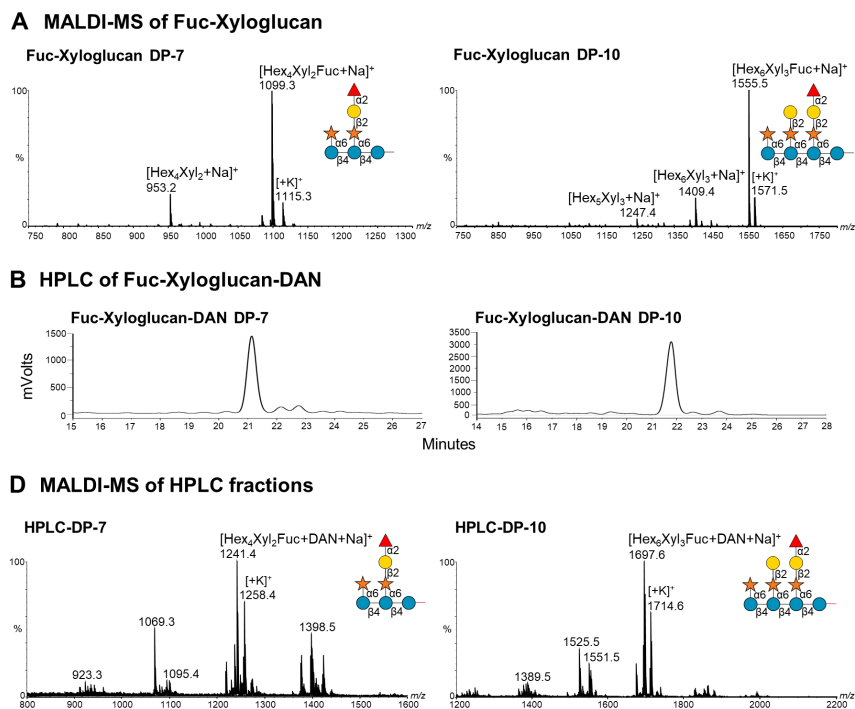
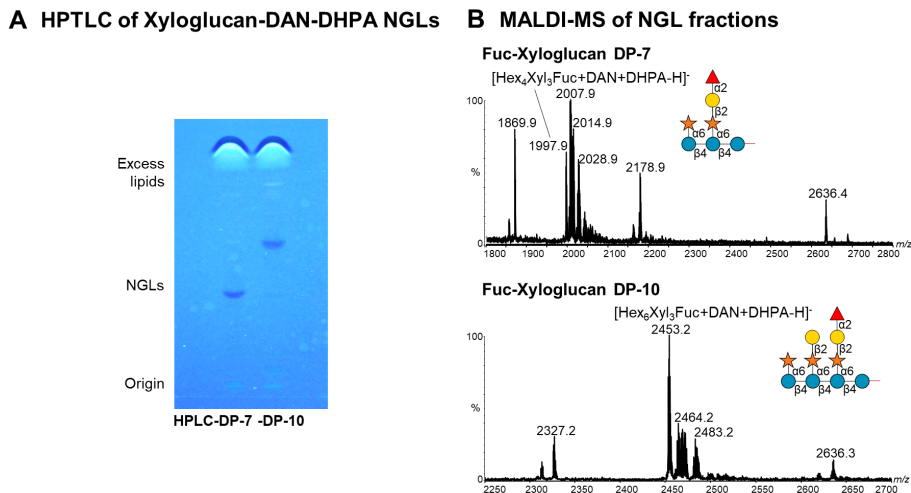
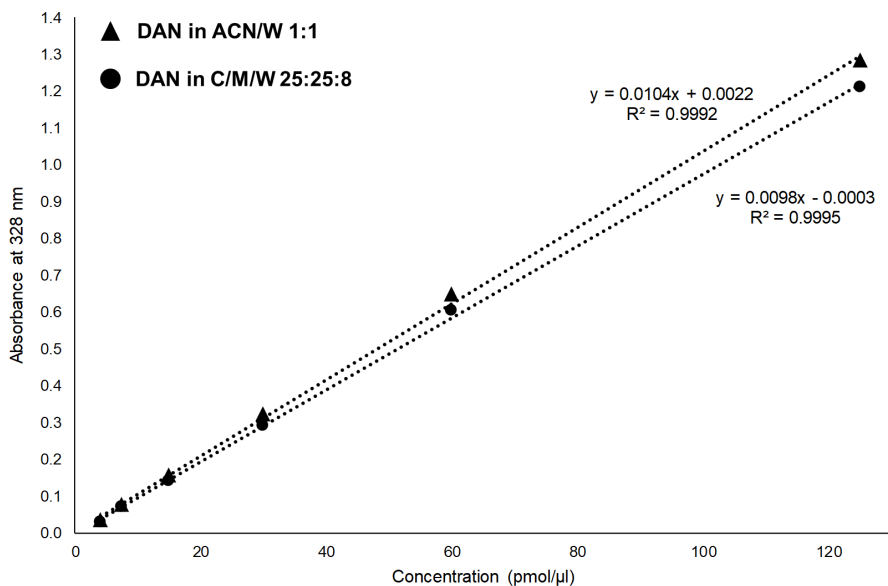


Figure S2.2. Deconvolution of the fucosylated-xyloglucan oligosaccharides from apple. (A) MALDI-MS spectra of fucosylated-xyloglucan oligosaccharides investigated from apple comprising DP-7 and DP-10. (B) HPLC separation of the DAN-conjugated fractions. (C) MALDI-MS spectra of the HPLC fractions obtained. Xyloglucan oligosaccharides are depicted. The red link represents the oligosaccharides reducing end conjugated to DAN.



**Figure S2.3. Preparation of the fucosylated-xyloglucan-DAN-NGL probes from apple included in the new xyloglucan microarrays. (A)** HPTLC analysis of conjugation mixtures DAN-DHPA-NGLs of fucosylated-xyloglucan DP-7 and DP-10 fractions (Figure S2.2B) revealed by primulin-staining. **(B)** MALDI-MS analysis of the fucosylated-xyloglucan-DAN-DHPA NGL-probes printed in the xyloglucan microarrays. Xyloglucan oligosaccharides are depicted. The red link represents the oligosaccharides reducing end conjugated to DAN.



**Figure S2.4. DAN standard curve used for the quantitation of the DAN-conjugated xyloglucan oligosaccharides and NGLs.** Serial dilutions of DAN solution were prepared at 4, 7.5, 15, 30, 60 and 125 pmol/μL in acetonitrile/water (ACN/H<sub>2</sub>O 1:1) for quantitation of xyloglucan-DAN-conjugated samples, and in chloroform/methanol/water (C/M/W 25:25:258) for quantitation of the xyloglucan-DAN-DHPA NGLs generated.

## Supplementary Tables

**Table S2.1. Information on the oligosaccharide neoglycolipid probes printed and validated in the glucan and hemicellulose oligosaccharide microarrays. The probes are sorted by linkage type and degree of polymerization.**

ID <sup>a</sup>	Linkages	Sources <sup>b</sup>	Probe Designation <sup>c</sup>	Probe Sequence <sup>d,e</sup>	
1	Linear Glcα2	Gluc-fructoside ( <i>Cyanobacterium</i> ) hydrolysate (Palma <i>et al.</i> 2015 <sup>32</sup> , Liu <i>et al.</i> 2018 <sup>131</sup> )	Cyano-2	Glcα-2Glc-AO	
2			Cyano-3	Glcα-2Glcα-2Glc-AO	
3			Cyano-4	Glcα-2Glcα-2Glcα-2Glc-AO	
4			Cyano-5	Glcα-2Glcα-2Glcα-2Glcα-2Glc-AO	
5			Cyano-6	Glcα-2Glcα-2Glcα-2Glcα-2Glcα-2Glc-AO*	
6			Cyano-7	Glcα-2Glcα-2Glcα-2Glcα-2Glcα-2Glcα-2Glc-AO*	
7			Cyano-8	Glcα-2Glcα-2Glcα-2Glcα-2Glcα-2Glcα-2Glcα-2Glc-AO*	
8			Cyano-9	Glcα-2Glcα-2Glcα-2Glcα-2Glcα-2Glcα-2Glcα-2Glcα-2Glc-AO*	
9	Linear Glcα3	Wako Chemicals (Palma <i>et al.</i> 2015 <sup>32</sup> )	Nigerose	Glcα-3Glc-AO	
10		Glucan ( <i>Poria cocos</i> ) hydrolysate (Palma <i>et al.</i> 2015 <sup>32</sup> , Liu <i>et al.</i> 2018 <sup>131</sup> )	Poria-3	Glcα-3Glcα-3Glc-AO	
11			Poria-4	Glcα1-3Glcα1-3Glcα1-3Glc-AO	
12			Poria-5	Glcα-3Glcα-3Glcα-3Glcα-3Glc-AO	
13			Poria-6	Glcα-3Glcα-3Glcα-3Glcα-3Glcα-3Glc-AO	
14			Poria-7	Glcα-3Glcα-3Glcα-3Glcα-3Glcα-3Glcα-3Glc-AO	
15			Poria-8	Glcα-3Glcα-3Glcα-3Glcα-3Glcα-3Glcα-3Glcα-3Glc-AO*	
16			Poria-9	Glcα-3Glcα-3Glcα-3Glcα-3Glcα-3Glcα-3Glcα-3Glcα-3Glc-AO*	
17			Poria-10	Glcα-3Glcα-3Glcα-3Glcα-3Glcα-3Glcα-3Glcα-3Glcα-3Glcα-3Glc-AO*	
18			Poria-11	Glcα-3Glcα-3Glcα-3Glcα-3Glcα-3Glcα-3Glcα-3Glcα-3Glcα-3Glcα-3Glc-AO*	
19			Poria-12	Glcα-3Glcα-3Glcα-3Glcα-3Glcα-3Glcα-3Glcα-3Glcα-3Glcα-3Glcα-3Glcα-3Glc-AO*	
20			Poria-13	Glcα-3Glcα-3Glcα-3Glcα-3Glcα-3Glcα-3Glcα-3Glcα-3Glcα-3Glcα-3Glcα-3Glc-AO*	
21			Linear Glcα4	Sigma-Aldrich (Palma <i>et al.</i> 2015 <sup>32</sup> )	Malto-2
22	Malto-3				Glcα-4Glcα-4Glc-AO
23	Malto-4	Glcα-4Glcα-4Glcα-4Glc-AO			
24	Malto-5	Glcα-4Glcα-4Glcα-4Glcα-4Glc-AO			
25	Malto-6	Glcα-4Glcα-4Glcα-4Glcα-4Glcα-4Glc-AO			
26	Malto-7	Glcα-4Glcα-4Glcα-4Glcα-4Glcα-4Glcα-4Glc-AO			
27	Maltodextrin ( <i>Zea mays</i> ) hydrolysate; Vector Laboratories (Palma <i>et al.</i> 2015 <sup>32</sup> , Liu <i>et al.</i> 2018 <sup>131</sup> )	Malto-8		Glcα-4Glcα-4Glcα-4Glcα-4Glcα-4Glcα-4Glc-AO*	
28		Malto-9		Glcα-4Glcα-4Glcα-4Glcα-4Glcα-4Glcα-4Glcα-4Glc-AO*	
29		Malto-10		Glcα-4Glcα-4Glcα-4Glcα-4Glcα-4Glcα-4Glcα-4Glcα-4Glc-AO*	
30		Malto-11		Glcα-4Glcα-4Glcα-4Glcα-4Glcα-4Glcα-4Glcα-4Glcα-4Glcα-4Glcα-4Glc-AO*	
31		Malto-12		Glcα-4Glcα-4Glcα-4Glcα-4Glcα-4Glcα-4Glcα-4Glcα-4Glcα-4Glcα-4Glcα-4Glc-AO*	
32		Malto-13		Glcα-4Glcα-4Glcα-4Glcα-4Glcα-4Glcα-4Glcα-4Glcα-4Glcα-4Glcα-4Glcα-4Glcα-4Glc-AO*	
33	Linear Glcα6	Dextran ( <i>Leuconostoc Mesenteroides</i> ) hydrolysate	Dext-2	Glcα-6Glc-AO	
34			Dext-3	Glcα-6Glcα-6Glc-AO	

35		(Palma <i>et al.</i> 2015 <sup>32</sup> , Liu <i>et al.</i> 2018 <sup>131</sup> )	Dext-4	Glcα-6Glcα-6Glcα-6Glc-AO
36			Dext-5	Glcα-6Glcα-6Glcα-6Glcα-6Glc-AO*
37			Dext-6	Glcα-6Glcα-6Glcα-6Glcα-6Glcα-6Glc-AO*
38			Dext-7	Glcα-6Glcα-6Glcα-6Glcα-6Glcα-6Glcα-6Glc-AO*
39			Dext-8	Glcα-6Glcα-6Glcα-6Glcα-6Glcα-6Glcα-6Glcα-6Glc-AO*
40			Dext-9	Glcα-6Glcα-6Glcα-6Glcα-6Glcα-6Glcα-6Glcα-6Glcα-6Glc-AO*
41			Dext-10	Glcα-6Glcα-6Glcα-6Glcα-6Glcα-6Glcα-6Glcα-6Glcα-6Glcα-6Glc-AO*
42			Dext-11	Glcα-6Glcα-6Glcα-6Glcα-6Glcα-6Glcα-6Glcα-6Glcα-6Glcα-6Glcα-6Glc-AO*
43			Dext-12	Glcα-6Glcα-6Glcα-6Glcα-6Glcα-6Glcα-6Glcα-6Glcα-6Glcα-6Glcα-6Glcα-6Glc-AO*
44			Dext-13	Glcα-6Glcα-6Glcα-6Glcα-6Glcα-6Glcα-6Glcα-6Glcα-6Glcα-6Glcα-6Glcα-6Glcα-6Glc-AO*
45	Mixed-linked Glcα4-6	Sigma-Aldrich; Megazyme (Palma <i>et al.</i> 2015 <sup>32</sup> )	Pano-3	Glcα-6Glcα-4Glc-AO
46			i-Pano-3	Glcα-4Glcα-6Glc-AO
47			Pullu-4	Glcα-6Glcα-4Glcα-4Glc-AO
48			Pullu-6	Glcα-4Glcα-4Glcα-6Glcα-4Glcα-4Glc-AO
49			Pullu-7	Glcα-6Glcα-4Glcα-4Glcα-6Glcα-4Glcα-4Glc-AO
50	Linear Glcβ2	Cyclic β-glucan ( <i>Brucella</i> spp.) hydrolysate (Palma <i>et al.</i> 2015 <sup>32</sup> , Liu <i>et al.</i> 2018 <sup>131</sup> )	CβG-2	Glcβ-2Glc-AO
51			CβG-3	Glcβ-2Glcβ-2Glc-AO
52			CβG-4	Glcβ-2Glcβ-2Glcβ-2Glc-AO
53			CβG-5	Glcβ-2Glcβ-2Glcβ-2Glcβ-2Glc-AO*
54			CβG-6	Glcβ-2Glcβ-2Glcβ-2Glcβ-2Glcβ-2Glc-AO*
55			CβG-7	Glcβ-2Glcβ-2Glcβ-2Glcβ-2Glcβ-2Glcβ-2Glc-AO*
56			CβG-8	Glcβ-2Glcβ-2Glcβ-2Glcβ-2Glcβ-2Glcβ-2Glcβ-2Glc-AO*
57			CβG-9	Glcβ-2Glcβ-2Glcβ-2Glcβ-2Glcβ-2Glcβ-2Glcβ-2Glcβ-2Glc-AO*
58			CβG-10	Glcβ-2Glcβ-2Glcβ-2Glcβ-2Glcβ-2Glcβ-2Glcβ-2Glcβ-2Glcβ-2Glc-AO*
59			CβG-11	Glcβ-2Glcβ-2Glcβ-2Glcβ-2Glcβ-2Glcβ-2Glcβ-2Glcβ-2Glcβ-2Glcβ-2Glc-AO*
60			CβG-12	Glcβ-2Glcβ-2Glcβ-2Glcβ-2Glcβ-2Glcβ-2Glcβ-2Glcβ-2Glcβ-2Glcβ-2Glcβ-2Glc-AO*
61			CβG-13	Glcβ-2Glcβ-2Glcβ-2Glcβ-2Glcβ-2Glcβ-2Glcβ-2Glcβ-2Glcβ-2Glcβ-2Glcβ-2Glcβ-2Glc-AO*
62			Linear Glcβ3	Dextra Laboratories (Palma <i>et al.</i> 2015 <sup>32</sup> )
63	Lam-3	Glcβ-3Glcβ-3Glc-AO		
64	Lam-4	Glcβ-3Glcβ-3Glcβ-3Glc-AO		
65	Megazyme (Palma <i>et al.</i> 2015 <sup>32</sup> )	Lam-5		Glcβ-3Glcβ-3Glcβ-3Glcβ-3Glc-AO
66		Lam-6		Glcβ-3Glcβ-3Glcβ-3Glcβ-3Glcβ-3Glc-AO*
67	Seikagaku AMS Biotechnology (Palma <i>et al.</i> 2015 <sup>32</sup> )	Lam-7		Glcβ-3Glcβ-3Glcβ-3Glcβ-3Glcβ-3Glcβ-3Glc-AO
68	Curdlan ( <i>Agrobacterium</i> sp.) hydrolysate (Palma <i>et al.</i> 2015 <sup>32</sup> , Liu <i>et al.</i> 2018 <sup>131</sup> )	Curd-8		Glcβ-3Glcβ-3Glcβ-3Glcβ-3Glcβ-3Glcβ-3Glcβ-3Glc-AO*
69		Curd-9		Glcβ-3Glcβ-3Glcβ-3Glcβ-3Glcβ-3Glcβ-3Glcβ-3Glcβ-3Glc-AO*
70		Curd-10		Glcβ-3Glcβ-3Glcβ-3Glcβ-3Glcβ-3Glcβ-3Glcβ-3Glcβ-3Glcβ-3Glc-AO*
71		Curd-11		Glcβ-3Glcβ-3Glcβ-3Glcβ-3Glcβ-3Glcβ-3Glcβ-3Glcβ-3Glcβ-3Glcβ-3Glc-AO*
72		Curd-12	Glcβ-3Glcβ-3Glcβ-3Glcβ-3Glcβ-3Glcβ-3Glcβ-3Glcβ-3Glcβ-3Glcβ-3Glc-AO*	
73	Curd-13	Glcβ-3Glcβ-3Glcβ-3Glcβ-3Glcβ-3Glcβ-3Glcβ-3Glcβ-3Glcβ-3Glcβ-3Glcβ-3Glc-AO*		
74	Neutral soluble glucan ( <i>S.cerevisiae</i> ) hydrolysate (Palma <i>et al.</i> 2006 <sup>72</sup> )	NSG-11	Glcβ-3Glcβ-3Glcβ-3Glcβ-3Glcβ-3Glcβ-3Glcβ-3Glcβ-3Glcβ-3Glcβ-3Glcβ-3Glc-AO*	





117			Barley-13	(Glcβ-4/3Glcβ) <sub>6</sub> -3Glc-AO*
118			Barley-14	(Glcβ-4/3Glcβ) <sub>6</sub> -4Glcβ-3Glc-AO*
119			Barley-15	(Glcβ-4/3Glcβ) <sub>7</sub> -3Glc-AO*
120			Barley-16	(Glcβ-4/3Glcβ) <sub>7</sub> -4Glcβ-3Glc-AO*
121	Branched Glcβ3(6)	Grifolan hydrolysate (Palma <i>et al.</i> 2015 <sup>32</sup> , Liu <i>et al.</i> 2018 <sup>131</sup> )	Grifo-3*	Glcβ-3 (Glcβ-6) <sub>3-16</sub> -AO*
122			Grifo-4*	
123			Grifo-5*	
124			Grifo-6*	
125			Grifo-7*	
126			Grifo-8*	
127			Grifo-9*	
128			Grifo-10*	
129			Grifo-11*	
130			Grifo-12*	
131			Grifo-13*	
132			Grifo-14*	
133			Grifo-15*	
134			Grifo-16*	
135	Branched Glcβ3(6)	Lentinan hydrolysate (Palma <i>et al.</i> 2015 <sup>32</sup> , Liu <i>et al.</i> 2018 <sup>131</sup> )	Lentin-2	Glcβ-3 (Glcβ-6) <sub>2-13</sub> -AO*
136			Lentin-3	
137			Lentin-4*	
138			Lentin-5*	
139			Lentin-6*	
140			Lentin-7*	
141			Lentin-8*	
142			Lentin-9*	
143			Lentin-10*	
144			Lentin-11*	
145	Lentin-12*			
146	Lentin-13*			
147	Branched Glcβ3(6)	Chemical synthesis (Palma <i>et al.</i> 2015 <sup>32</sup> )	HE-9B7	Glcβ-3Glcβ-3Glcβ-3Glcβ-3Glcβ-3Glcβ-3Glcβ-3Glcβ-AO   Glcβ-6
148			HE-10B2	Glcβ-3Glcβ-3Glcβ-3Glcβ-3Glcβ-3Glcβ-3Glcβ-3Glcβ-AO   Glcβ-6
149			HE-10B3	Glcβ-3Glcβ-3Glcβ-3Glcβ-3Glcβ-3Glcβ-3Glcβ-3Glcβ-AO   Glcβ-6
150			HE-10B5	Glcβ-3Glcβ-3Glcβ-3Glcβ-3Glcβ-3Glcβ-3Glcβ-3Glcβ-AO   Glcβ-6
151			HE-10B7	Glcβ-3Glcβ-3Glcβ-3Glcβ-3Glcβ-3Glcβ-3Glcβ-3Glcβ-AO   Glcβ-6

152			Gu-6B1/3	Glc $\beta$ -3Glc $\alpha$ -3Glc $\beta$ -3Glc-AO $\begin{array}{c}   \\ \text{Glc}\beta\text{-6} \end{array}$ $\begin{array}{c}   \\ \text{Glc}\beta\text{-6} \end{array}$
153			HE-11B3/6	Glc $\beta$ -3Glc $\beta$ -3Glc $\beta$ -3Glc $\beta$ -3Glc $\beta$ -3Glc $\beta$ -3Glc $\beta$ -3Glc $\beta$ -3Glc-AO $\begin{array}{c}   \\ \text{Glc}\beta\text{-6} \end{array}$ $\begin{array}{c}   \\ \text{Glc}\beta\text{-6} \end{array}$
154	Mixed-linked Xyl $\beta$ 3-4	Xylan ( <i>Palmaria palmata</i> ) hydrolysate; Elicityl (Table S2.2)	Xyl-3	Xyl $\beta$ -4Xyl $\beta$ -4Xyl-AO or Xyl $\beta$ -4Xyl $\beta$ -3Xyl-AO
155			Xyl-4	Xyl $\beta$ -4Xyl $\beta$ -4Xyl $\beta$ -4Xyl-AO or Xyl $\beta$ -4Xyl $\beta$ -3Xyl $\beta$ -4Xyl-AO
156	Linear Xyl $\beta$ 4	Xylopentaose mixture Megazyme (Table S2.2)	Xyl-5( $\beta$ 4)	Xyl $\beta$ -4Xyl $\beta$ -4Xyl $\beta$ -4Xyl $\beta$ -4Xyl-AO
157			Xyl-6( $\beta$ 4)	Xyl $\beta$ -4Xyl $\beta$ -4Xyl $\beta$ -4Xyl $\beta$ -4Xyl $\beta$ -4Xyl-AO
158	Mixed-linked Xyl $\beta$ 3-4	Xylan ( <i>Palmaria palmata</i> ) hydrolysate; Elicityl (Table S2.2)	Xyl-7	(Xyl $\beta$ -4/3Xyl $\beta$ ) <sub>3</sub> -4/3Xyl-AO*
159			Xyl-8	(Xyl $\beta$ -4/3Xyl $\beta$ ) <sub>4</sub> -AO*
160			Xyl-9	(Xyl $\beta$ -4/3Xyl $\beta$ ) <sub>4</sub> -4/3Xyl-AO*
161			Xyl-10	(Xyl $\beta$ -4/3Xyl $\beta$ ) <sub>5</sub> -AO*
162			Xyl-11	(Xyl $\beta$ -4/3Xyl $\beta$ ) <sub>5</sub> -4/3Xyl-AO*
163			Xyl-12	(Xyl $\beta$ -4/3Xyl $\beta$ ) <sub>6</sub> -AO*
164			Xyl-13	(Xyl $\beta$ -4/3Xyl $\beta$ ) <sub>6</sub> -4/3Xyl-AO*
165	Branched Xyl $\beta$ 4(Ara $\alpha$ 3/2)	Arabinoxylan (wheat flour) hydrolysate tri- to hexasaccharides Megazyme (Table S2.2)	Ara-Xylan-3	Ara $\alpha$ -3   Xyl $\beta$ -4Xyl-AO
166			Ara-Xylan-4a	Xyl $\beta$ -4Xyl $\beta$ -4Xyl-AO   Ara $\alpha$ -2
167			Ara-Xylan-4b	Xyl $\beta$ -4Xyl $\beta$ -4Xyl-AO   Ara $\alpha$ -2   Ara $\alpha$ -3   Xyl $\beta$ -4Xyl $\beta$ -4Xyl-AO
168			Ara-Xylan-5a	Ara $\alpha$ -3   Xyl $\beta$ -4Xyl $\beta$ -4Xyl $\beta$ -4Xyl-AO
169			Ara-Xylan-5b	Xyl $\beta$ -4Xyl $\beta$ -4Xyl $\beta$ -4Xyl-AO   Ara $\alpha$ -2   Ara $\alpha$ -3   Xyl $\beta$ -4Xyl $\beta$ -4Xyl $\beta$ -4Xyl-AO
170			Ara-Xylan-5c	Ara $\alpha$ -3   Xyl $\beta$ -4Xyl $\beta$ -4Xyl-AO   Ara $\alpha$ -2
171			Ara-Xylan-6	Ara $\alpha$ -3   Xyl $\beta$ -4Xyl $\beta$ -4Xyl $\beta$ -4Xyl-AO   Ara $\alpha$ -2
172	Linear Ara $\alpha$ 5		Ara-2( $\alpha$ 5)	Ara $\alpha$ -5Ara-AO

173 174 175 176 177 178 179		Arabinan (sugar beet) hydrolysate hexa- to octasaccharides Megazyme (Table S2.2)	Ara-3( $\alpha$ 5) Ara-4( $\alpha$ 5) Ara-5( $\alpha$ 5) Ara-6( $\alpha$ 5) Ara-7( $\alpha$ 5) Ara-8( $\alpha$ 5) Ara-9( $\alpha$ 5)	Ara $\alpha$ -5Ara $\alpha$ -5Ara-AO Ara $\alpha$ -5Ara $\alpha$ -5Ara $\alpha$ -5Ara-AO Ara $\alpha$ -5Ara $\alpha$ -5Ara $\alpha$ -5Ara $\alpha$ -5Ara-AO Ara $\alpha$ -5Ara $\alpha$ -5Ara $\alpha$ -5Ara $\alpha$ -5Ara $\alpha$ -5Ara-AO Ara $\alpha$ -5Ara $\alpha$ -5Ara $\alpha$ -5Ara $\alpha$ -5Ara $\alpha$ -5Ara $\alpha$ -5Ara-AO Ara $\alpha$ -5Ara $\alpha$ -5Ara $\alpha$ -5Ara $\alpha$ -5Ara $\alpha$ -5Ara $\alpha$ -5Ara-5Ara-AO Ara $\alpha$ -5Ara $\alpha$ -5Ara $\alpha$ -5Ara $\alpha$ -5Ara $\alpha$ -5Ara $\alpha$ -5Ara-5Ara-AO*
180	Branched Ara $\alpha$ 5(Ara $\alpha$ 3/2)	Arabinan (sugar beet) hydrolysate tetra- and pentasaccharides Megazyme (Table S2.2)	Ara-4B3	Ara-3   Ara $\alpha$ -5Ara $\alpha$ -5Ara-AO
181			Ara-5B	Ara-3   Ara $\alpha$ -5Ara $\alpha$ -5Ara-AO   Ara $\alpha$ -2  Ara-3   Ara $\alpha$ -5Ara $\alpha$ -5Ara-AO
182 183 184 185	Linear Man $\beta$ 4	Mannan hydrolysate tetra- to hexasaccharides Megazyme (Table S2.2)	Man-4( $\beta$ 4) Man-5( $\beta$ 4) Man-6( $\beta$ 4)	Man $\beta$ -4Man $\beta$ -4Man $\beta$ -4Man-AO Man $\beta$ -4Man $\beta$ -4Man $\beta$ -4Man $\beta$ -4Man-AO Man $\beta$ -4Man $\beta$ -4Man $\beta$ -4Man $\beta$ -4Man $\beta$ -4Man-AO
		Mannan (ivory nut) hydrolysate Elicityl (Table S2.2)	Man-8( $\beta$ 4)	Man $\beta$ -4Man $\beta$ -4Man $\beta$ -4Man $\beta$ -4Man $\beta$ -4Man $\beta$ -4Man $\beta$ -4Man-AO
186 187 188 189 190 191 192	Branched Man $\beta$ 4(Gal $\alpha$ 6)	Galactomannan (carob) hydrolysate Elicityl (Table S2.2)	Gal-Mannan-2e Gal-Mannan-3e Gal-Mannan-4e Gal-Mannan-5e Gal-Mannan-6e* Gal-Mannan-7e Gal-Mannan-8e	Man $\beta$ 4 (Gal $\alpha$ 6) <sub>2-8</sub> -AO*
193 194 195 196		di-Galactosyl-mannopentasaccharide (carob) Megazyme (Table S2.2)	Gal-Mannan-5m Gal-Mannan-6m Gal-Mannan-7m Gal-Mannan-8m	
197 198 199		Galactomannan (carob) hydrolysate Elicityl (Table S2.2)	Gal-Mannan-9e Gal-Mannan-10e* Gal-Mannan-11e*	
200	Branched Glc $\beta$ 4(Xyl $\alpha$ 6Gal $\beta$ 4)	Xyloglucan (Tamarind) heptasaccharide Megazyme (Table S2.2)	Xyl-Glucan-7	Xyl $\alpha$ -6   Glc $\beta$ -4Glc $\beta$ -4Glc $\beta$ -4Glc-AO                          Xyl $\alpha$ -6                Xyl $\alpha$ -6

201		Xyloglucan (Tamarind) hydrolysate Megazyme (Table S2.2)	Xyl-Glucan-8*	$\begin{array}{c} \text{Xyl}\alpha\text{-6} \\   \\ \text{Glc}\beta\text{-4Glc}\beta\text{-4Glc}\beta\text{-4Glc}\text{-AO}^* \\   \quad   \\ \text{Xyl}\alpha\text{-6} \quad \text{Xyl}\alpha\text{-6} \\   \\ \text{Gal}\beta\text{-2} \\   \\ \text{Xyl}\alpha\text{-6} \\   \\ \text{Glc}\beta\text{-4Glc}\beta\text{-4Glc}\beta\text{-4Glc}\text{-AO} \\   \quad   \\ \text{Xyl}\alpha\text{-6} \quad \text{Xyl}\alpha\text{-6} \end{array}$
202			Xyl-Glucan-9	$\begin{array}{c} \text{Gal}\beta\text{-2} \\   \\ \text{Xyl}\alpha\text{-6} \\   \\ \text{Glc}\beta\text{-4Glc}\beta\text{-4Glc}\beta\text{-4Glc}\text{-AO} \\   \quad   \\ \text{Xyl}\alpha\text{-6} \quad \text{Xyl}\alpha\text{-6} \\   \\ \text{Gal}\beta\text{-2} \end{array}$
203	Branched Glc $\beta$ 4(Xyl $\alpha$ 6Gal $\beta$ 4 Fuc $\alpha$ 2)	Fucosylated-xyloglucan (apple) XFG and XLFG oligosaccharide mixtures Elicityl (Table S2.2)	Fuc-Xyl-Glucan-6*	$\begin{array}{c} \text{(Fuc}\alpha\text{-2)} \\   \\ \text{Gal}\beta\text{-2} \\   \\ \text{Xyl}\alpha\text{-6} \\   \\ \text{Glc}\beta\text{-4Glc}\beta\text{-4Glc}\text{-AO}^* \\   \\ \text{Xyl}\alpha\text{-6} \end{array}$
204				Fuc-Xyl-Glucan-9*
205	Linear GlcNAc $\beta$ 4	Chemical synthesis	GlcNAc-2	GlcNAc $\beta$ -4GlcNAc-AO
206			GlcNAc-3	GlcNAc $\beta$ -4GlcNAc $\beta$ -4GlcNAc-AO
207			GlcNAc-4	GlcNAc $\beta$ -4GlcNAc $\beta$ -4GlcNAc $\beta$ -4GlcNAc-AO
208			GlcNAc-5	GlcNAc $\beta$ -4GlcNAc $\beta$ -4GlcNAc $\beta$ -4GlcNAc $\beta$ -4GlcNAc-AO
209			GlcNAc-6	GlcNAc $\beta$ -4GlcNAc $\beta$ -4GlcNAc $\beta$ -4GlcNAc $\beta$ -4GlcNAc $\beta$ -4GlcNAc-AO
210			GlcNAc-7	GlcNAc $\beta$ -4GlcNAc $\beta$ -4GlcNAc $\beta$ -4GlcNAc $\beta$ -4GlcNAc $\beta$ -4GlcNAc $\beta$ -4GlcNAc-AO
211			GlcNAc-8	GlcNAc $\beta$ -4GlcNAc $\beta$ -4GlcNAc $\beta$ -4GlcNAc $\beta$ -4GlcNAc $\beta$ -4GlcNAc $\beta$ -4GlcNAc $\beta$ -4GlcNAc-AO
212	Linear GlcN $\beta$ 4	Chemical synthesis	GlcN-4	GlcN $\beta$ -4GlcN $\beta$ -4GlcN $\beta$ -4GlcN-AO
213			GlcN-5	GlcN $\beta$ -4GlcN $\beta$ -4GlcN $\beta$ -4GlcN $\beta$ -4GlcN-AO

214			GlcN-6	GlcN $\beta$ -4GlcN $\beta$ -4GlcN $\beta$ -4GlcN $\beta$ -4GlcN $\beta$ -4GlcN-AO
215		Miscellaneous	GlcNAc2(1,6)	GlcNAc $\beta$ -6GlcNAc-AO
216	-		GalAra(1,3)	Gal $\beta$ -3Ara-AO
217			Gal2GlcNAc(1,3)	Gal $\alpha$ -3Gal $\beta$ -3GlcNAc-AO
218			GalManNAc(1,4)	Gal $\beta$ -4ManNAc-AO

<sup>a</sup>ID, Probe position in the microarrays matching the position in the heatmaps and binding-charts;

<sup>b</sup>The sources of the oligosaccharide fragments (depolymerised polysaccharides or oligosaccharides) to prepare the NGL probes with the commercial supply, as appropriate, are indicated; detailed information on the preparation of the oligosaccharide fragments can be found in references Palma *et al.* 2015<sup>32</sup>, Liu *et al.* 2018<sup>131</sup> for probes 1-153, and in Table S2.2 for probes 154 to 204; probes 205 to 218 are control AO-NGL probes, from the collection of the Glycociences Laboratory (Imperial College, London), included in the microarrays, and are presented and validated in Chapter 3.

<sup>c</sup>Abbreviations for oligosaccharide moieties are as follows: Cyano-, from cyanobacterium gluco-fructosides; Poria-, from  $\beta$ 1,3-linked glucan polysaccharide isolated from *Poria cocos* mycelia; Malto-, from maltodextrins; Dext-, from dextran (MW 200 kDa); Pullu-, from Pullulan; C $\beta$ G-, from cyclic  $\beta$ 1,2-linked glucan isolated from *Brucella spp*; Lam-, laminarioligosaccharides; Curd-, from Curdlan; Cello-, from cellulose; Pust-, from pustulan; Barley-, from barley glucan; Grifo-, from branched glucan polysaccharide grifolan (95 kDa) isolated from the barmy mycelium of *Grifola frondosa*; Lenti- from branched glucan polysaccharide lentinan from *Lentinus edodes*; HE and Gu are synthetic oligosaccharides; Xyl-, from xylan; Ara-Xylan-, from arabinoxylan; Ara-, from arabinan; Man-, from mannan; Gal-Mannan-, from galactomannan; Xyl-Glucan-, from non-fucosylated tamarind xyloglucan; and Fuc-Xyl-Glucan-, from fucosylated apple xyloglucan. In the lentinan and poria derived fractions a minor 1,4-linked glucose contaminant was detected (Palma *et al.* 2015<sup>32</sup>, Liu *et al.* 2018<sup>131</sup>).

<sup>d</sup>The oligosaccharide probes are all lipid-linked, NGLs; AO, NGLs prepared from reducing oligosaccharides by oxime ligation with an aminoxy (AO) functionalized DHPE (1,2-dihexadecyl-sn-glycero-3-phosphoethanolamine)<sup>32,86</sup>.

<sup>e</sup>An asterisk indicates the major component when multiple components are present. The NGL-probes Fuc-Xyl-Glucan-6 (probe 203) and Fuc-Xyl-Glucan-9 (probe 204) are mixtures and both fucosylated and non-fucosylated components are present as determined by MALDI-MS analysis (see Table 2.1) and the predicted major sequences are depicted.

**Table S2.2. Sources of plant-related oligosaccharides and analysis performed for the preparation of the AO-NGL probes included in the hemicellulose microarrays.**

Sample	Source and DP	Analysis
<b>Palmaria palmata Xylan hydrolysate mixture</b>	Elicityl Xyl1111; Mixture comprising DP-2 to DP-25 (cut-off: 650Da-3kDa)	Size exclusion chromatography MALDI-MS
<b>Arabinoxylan xylopentaose mixture</b>	Megazyme O-XPE (acid hydrolysis of arabinoxylan); Mixture comprising DP-5 and DP-6	MALDI-MS HPTLC purification after lipid derivatization
<b>Wheat flour Ara-Xylan-3</b>	Megazyme O-A3X (acid hydrolysis of wheat flour arabinoxylan); DP-3	MALDI-MS
<b>Wheat flour Ara-Xylan-4a</b>	Megazyme O-A2XX (acid hydrolysis of wheat flour arabinoxylan); DP-4	MALDI-MS
<b>Wheat flour Ara-Xylan-4b</b>	Megazyme O-AX3MIX (acid hydrolysis of wheat flour arabinoxylan); DP-4	MALDI-MS
<b>Wheat flour Ara-Xylan-5a</b>	Megazyme O-XA3XX (acid hydrolysis of wheat flour arabinoxylan); DP-5	MALDI-MS
<b>Wheat flour Ara-Xylan-5b</b>	Megazyme O-XAXXMIX (acid hydrolysis of wheat flour arabinoxylan); DP-5	MALDI-MS
<b>Wheat flour Ara-Xylan-5c</b>	Megazyme O-A23XX (acid hydrolysis of wheat flour arabinoxylan); DP-5	MALDI-MS
<b>Wheat flour Ara-Xylan-6</b>	Megazyme O-XA23XX; (acid hydrolysis of wheat flour arabinoxylan); DP-6	MALDI-MS
<b>Sugar beet Arabino-hexaose</b>	Megazyme O-AHE (enzymatic hydrolysis of debranched sugar beet arabinan); DP-6	MALDI-MS
<b>Sugar beet Arabino-heptaose</b>	Megazyme O-AHP (enzymatic hydrolysis of debranched sugar beet arabinan); DP-7	MALDI-MS
<b>Sugar beet Arabino-octaose mixture</b>	Megazyme O-AOC (enzymatic hydrolysis of debranched sugar beet arabinan); Mixture comprising DP-2 to DP-9	MALDI-MS HPTLC purification after lipid derivatization
<b>Sugar beet Ara-4B3</b>	Megazyme O-A4B (enzymatic hydrolysis of debranched sugar beet arabinan); DP4	MALDI-MS
<b>Sugar beet Ara-5B</b>	Megazyme O-A5BMIX (enzymatic hydrolysis of debranched sugar beet arabinan); DP5	MALDI-MS
<b>Manno-tetraose</b>	Megazyme O-MTE (enzymatic hydrolysis of mannan); DP-4	MALDI-MS
<b>Manno-pentaose</b>	Megazyme O-MPE (enzymatic hydrolysis of mannan); DP-5	MALDI-MS
<b>Manno-hexaose</b>	Megazyme O-MHE (enzymatic hydrolysis of mannan); DP-6	MALDI-MS
<b>Ivory nut Mannan hydrolysate mixture</b>	Elicityl Man810; Mixture comprising DP-7 to DP-13	Size exclusion chromatography MALDI-MS
<b>Carob galactomannan hydrolysate mixture</b>	Elicityl Man219 (designated <i>e</i> -series); Mixture comprising DP-2 to DP-11	Size exclusion chromatography MALDI-MS HPTLC purification after lipid derivatization (DP 6-7)
<b>Carob galactosyl-mannopentaose mixture</b>	Megazyme O-GGM5 (designated <i>m</i> -series) (enzymatic hydrolysis of carob galactomannan); Mixture comprising DP-5 to DP-8	MALDI-MS HPTLC purification after lipid derivatization
<b>Tamarind Xylo-Glucan-7</b>	Megazyme O-X3G4 (enzymatic hydrolysis of tamarind xyloglucan); DP-7	MALDI-MS
<b>Tamarind Xyloglucan hydrolysate mixture</b>	Megazyme O-XGHON; Mixture comprising DP-7 to DP-9	Size exclusion chromatography (Difficult to separate) MALDI-MS HPTLC purification after lipid derivatization
<b>Apple Fuc-Xyl-Glucan XFG</b>	Elicityl GLU1110 80% XFG <sup>a</sup> oligosaccharide; DP-7	MALDI-MS
<b>Apple Fuc-Xyl-Glucan XLFG</b>	Elicityl GLU1160 30% XLFG <sup>a</sup> oligosaccharide; DP-10	MALDI-MS

<sup>a</sup>Single-letter code for xyloglucan oligosaccharides according to Fry *et al.* 1993<sup>257</sup>: G denotes an unsubstituted backbone Glc monomer; X, L, and F denote Glc residues substituted with  $\alpha$ -D-Xylp,  $\beta$ -D-Galp-(1,2)- $\alpha$ -D-Xylp, and  $\alpha$ -L-Fucp-(1,2)- $\beta$ -D-Galp-(1,2)- $\alpha$ -D-Xylp side chains, respectively.

Table S2.3. Carbohydrate-directed monoclonal antibodies, lectins and CBMs investigated in the glucan and hemicellulose microarrays.

Proteins	Reported carbohydrate binding	Method of analysis	Source	Reference
<b>Monoclonal Antibodies</b>				
<b>400-2</b>	Raised against laminarin; β1,3 glucose oligosaccharide sequences in β1,3-glucans (DP ≥ 2).	Indirect competitive ELISA	Biosupplies (400-2)	Meikle <i>et al.</i> 1991 <sup>151</sup>
<b>400-3</b>	Raised against β1,3-1,4-glucan; Mixed-linked β1,3-1,4-glucose oligosaccharide sequences in β1,3-1,4-glucans; maximum binding to the heptasaccharide with the sequence G3G4G4G3G4G4G; weak cross-reactivity with β1,4-glucans; no cross-reactivity with β1,3-glucans.	Indirect competitive ELISA	Biosupplies (400-3)	Meikle <i>et al.</i> 1994 <sup>134</sup>
<b>LM10</b>	Raised against xylopentaose; Nonreducing end of the β1,4-xylose oligosaccharide (DP ≥ 2) backbone of xylans and arabinoxylans.	ELISA Microarrays	Plant probes (LM10)	McCartney <i>et al.</i> 2005 <sup>135</sup> Ruprecht <i>et al.</i> 2017 <sup>109</sup>
<b>LM11</b>	Raised against xylopentaose; β1,4-xylans (e.g. wheat arabinoxylan) (DP ≥ 4); accommodates more extensive substitutions of the xylan backbone with α-arabinose.	ELISA Microarrays	Plant probes (LM11)	McCartney <i>et al.</i> 2005 <sup>135</sup>
<b>LM6</b>	Raised against arabinoheptaose; Linear α1,5-arabinose pentasaccharide sequence in arabinans; can recognise several pectic polysaccharides and arabinogalactan-proteins.	ELISA Microarrays	Plant probes (LM6)	Willats <i>et al.</i> 1998 <sup>137</sup>
<b>400-4</b>	Raised against galactomannan oligosaccharides; Linear β1,4-mannose oligosaccharide sequences in β1,4-mannans and galactomannans (DP-3 to DP-6).	ELISA	Biosupplies (400-4)	Pettolino <i>et al.</i> 2001 <sup>138</sup>
<b>LM21</b>	Raised against mannopentaose; β1,4-linked mannan, glucomannan and galactomannan polysaccharides; β1,4 manno-oligosaccharides (DP-2 to DP-5).	ELISA Microarrays	Plant probes (LM21)	Marcus <i>et al.</i> 2010 <sup>139</sup>
<b>CCRC-M70</b>	Raised against guar galactomannan polysaccharides; Oligosaccharide binding not reported.	ELISA	Agrisera (AS16 3116)	Pattathil <i>et al.</i> 2010 <sup>140</sup>
<b>LM24</b>	Raised against xylosylated/galactosylated xyloglucan tamarind oligosaccharides (XXLG <sup>a</sup> and XLLG <sup>a</sup> ); Galactosylated xyloglucan polysaccharides and oligosaccharides, preferentially to the XLLG <sup>a</sup> motif of xyloglucan.	Microarrays	Plant probes (LM24)	Pedersen <i>et al.</i> 2012 <sup>92</sup>
<b>LM25</b>	Raised against xylosylated/galactosylated xyloglucan tamarind oligosaccharides (XXLG <sup>a</sup> and XLLG <sup>a</sup> ); Xyloglucan polysaccharides; XLLG <sup>a</sup> , XXLG <sup>a</sup> and XXXG <sup>a</sup> oligosaccharides; requires a xyloglucan epitope with at least one α-1,6-linked xylose residue linked to a β1,4-linked glucan backbone.	Microarrays	Plant probes (LM25)	Pedersen <i>et al.</i> 2012 <sup>92</sup>
<b>CCRC-M1</b>	Raised against sycamore rhamnogalacturonan; Sycamore xyloglucan polysaccharides; sycamore pectic polysaccharides; Oligosaccharide binding not reported but require the α-Fuc-(1,2)-β-Gal epitope of fucosylated-xyloglucan.	ELISA Competitive immunoassays	Agrisera (AS16 3136)	Puhlmann <i>et al.</i> 1994 <sup>258</sup>



<b>LM5</b>	Raised against galactotetrasaccharide; Linear $\beta$ 1,4-galactose tetrasaccharide sequence in galactans; can recognise several pectic polysaccharides; no cross-reactivity with $\beta$ 1,3- or $\beta$ 1,6-galactans.	Microarrays	Plant probes (LM5)	Jones <i>et al.</i> 1997 <sup>259</sup>
<b>Lectins</b>				
<b>Human Malectin</b>	Highly specific for $\alpha$ 1,3-di-glucosylated high-mannose <i>N</i> -glycans; binds to linear $\alpha$ 1,3-, $\alpha$ 1,4-, $\alpha$ 1,6- and $\beta$ 1,3-linked glucose sequences.	Microarrays	Recombinant prepared in house	Schallus <i>et al.</i> 2008 <sup>132</sup> Palma <i>et al.</i> 2010 <sup>133</sup>
<b>Concavalin A (ConA)</b>	Binding to $\alpha$ -linked mannose oligosaccharides.	ITC Microarrays	Vector (B-1005)	Wang <i>et al.</i> 2014 <sup>150</sup>
<b>Aleuria aurantia (AAL)</b>	Binding to $\alpha$ -fucosylated oligosaccharides.	Hemagglutination/ inhibitor Microarrays	Vector Lab. (B-1395)	Kochibe and Furukawa 1980 <sup>143</sup>
<b>Wheat germ agglutinin (WGA)</b>	<i>N</i> -acetylglucosamine, preferentially to DP-2 and DP-3; can also bind to bacterial cell wall peptidoglycans, chitin, cartilage glycosaminoglycans, and glycolipids.	Equilibrium dialysis Microarrays	Vector (B-1025)	Nagata <i>et al.</i> 1974 <sup>260</sup> Wang <i>et al.</i> 2014 <sup>150</sup>
<b>Datura stramonium (DSL)</b>	Binding to $\beta$ 1,4-linked <i>N</i> -acetylglucosamine oligosaccharides, preferring DP-2 and DP-3; also binds well to <i>N</i> -acetylglucosamine and oligosaccharides containing repeating <i>N</i> -acetylglucosamine sequences.	Precipitation / inhibition	Vector Lab. (B-1185)	Crowley <i>et al.</i> 1984 <sup>261</sup>
<b>Ricinus communis agglutinin I (RCA<sub>120</sub>)</b>	Binding to galactose or <i>N</i> -acetylgalactosamine sequences.	ELISA Microarrays	Vector Lab. (B-1085)	Baenziger <i>et al.</i> 1979 <sup>262</sup> Wang <i>et al.</i> 2011 <sup>152</sup>
<b>CBMs</b>				
<b>CtCBM25<sub>Cthe_0956</sub></b>	Putative starch-binding domain in the <i>Clostridium thermocellum</i> genome; uncharacterized protein with carbohydrate binding specificity not yet assigned.	-	Recombinant prepared in house	-
<b>CtCBM11<sub>Cthe_1472</sub></b>	High specificity towards mixed-linked $\beta$ 1,3-1,4-glucose oligosaccharides with DP-4 and longer; weak binding affinity to linear $\beta$ 1,4 glucose oligosaccharides.	ITC Microarrays	Recombinant prepared in house	Palma <i>et al.</i> 2015 <sup>32</sup> Ribeiro <i>et al.</i> 2019 <sup>34</sup>
<b>CmCBM6-2</b>	Broad specificity to $\beta$ -glucans; binds glucose oligosaccharides with DP-2 and longer: linear $\beta$ 1,2, $\beta$ 1,3, $\beta$ 1,4 and $\beta$ 1,6; mixed-linked $\beta$ 1,3-1,4; also binds to $\beta$ 1,4-xylose and $\beta$ 1,4-mannose oligosaccharides (weak).	Microarrays	Harry Gilbert (University of Newcastle, UK)	Palma <i>et al.</i> 2015 <sup>32</sup>
<b>CtCBM22-2<sub>Cthe_0912</sub></b>	Binding to oat spelt xylan and wheat and rye arabinoxylan polysaccharides and to $\beta$ 1,4-xylose oligosaccharides.	AGE ITC	Recombinant prepared in house	Charnock <i>et al.</i> 2000 <sup>136</sup>
<b>CmCBM32-2</b>	Binding to linear $\beta$ 1,2- and $\beta$ 1,3-glucose oligosaccharides and $\beta$ 1,3-glucose oligosaccharides with $\beta$ 1,6-glucose branches; weaker binding to $\beta$ 1,4-, $\beta$ 1,6- and mixed-linked $\beta$ 1,3-1,4-glucose oligosaccharides.	Microarrays	Harry Gilbert (University of Newcastle, UK)	Palma <i>et al.</i> 2015 <sup>32</sup>
<b>CtCBM35<sub>Cthe_2811</sub></b>	Binding to $\beta$ 1,4-linked mannan in galactomannan and glucomannan polysaccharides; higher affinity for konjac glucomannan than to carob galactomannan; no oligosaccharide-specificity reported.	AGE Fluorescence spectroscopy	Recombinant prepared in house	Ghosh <i>et al.</i> 2014 <sup>141</sup>
<b>TmCBM41</b>	Binding to linear $\alpha$ 1,4-glucose oligosaccharides and mixed-linked $\alpha$ 1,4-1,6-glucose oligosaccharides (with $\alpha$ 1,4-linked glucose at non-reducing end).	Microarrays	Alisdair Boraston (University of Victoria, Canada)	Palma <i>et al.</i> 2015 <sup>32</sup>

<sup>a</sup>Single-letter code for xyloglucan oligosaccharides according to Fry *et al.* 1993<sup>257</sup>: G denotes an unsubstituted backbone Glc residue; X, L, and F denote Glc residues substituted with  $\alpha$ -D-Xylp,  $\beta$ -D-Galp-(1,2)- $\alpha$ -D-Xylp, and  $\alpha$ -L-Fucp-(1,2)- $\beta$ -D-Galp-(1,2)- $\alpha$ -D-Xylp side chains, respectively.

**Table S2.4. Fluorescence binding intensities elicited with all the proteins investigated for validation of the glucan and hemicellulose oligosaccharide microarrays.** The numerical scores for the fluorescence binding signals are shown as means of duplicate spots at 5 fmol probe per spot (as in Figures 2.1 to 2.4 and 2.6, and Figure S2.2) and are representative of at least 2 independent experiments.

ID <sup>a</sup>	Probe <sup>b</sup>	hMalectin	CtCBM25	400-2	400-3	CtCBM11	CmCBM6-2	LM10	LM11	CtCBM22-2	LM6	400-4	LM21	CtCBM35	CCRC-M70	LM24	LM25	CCRC-M1	AAL
1	Cyano-2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
2	Cyano-3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
3	Cyano-4	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
4	Cyano-5	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
5	Cyano-6	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
6	Cyano-7	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
7	Cyano-8	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
8	Cyano-9	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
9	Nigerose	5786	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
10	Poria-3	10403	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
11	Poria-4	28415	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
12	Poria-5	29481	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
13	Poria-6	22671	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
14	Poria-7	34403	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
15	Poria-8	34760	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
16	Poria-9	13851	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
17	Poria-10	12774	810	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
18	Poria-11	14999	1817	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
19	Poria-12	28479	3319	-	-	-	692	-	-	-	-	-	-	-	-	-	-	-	-
20	Poria-13	9917	1463	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
21	Malto-2	10895	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
22	Malto-3	1503	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
23	Malto-4	7490	2008	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
24	Malto-5	7035	5253	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
25	Malto-6	6596	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
26	Malto-7	5593	5824	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
27	Malto-8	6163	7551	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
28	Malto-9	4356	4774	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
29	Malto-10	4213	7354	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
30	Malto-11	4739	9292	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
31	Malto-12	7526	9363	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
32	Malto-13	4280	3209	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
33	Dext-2	724	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
34	Dext-3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
35	Dext-4	3655	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
36	Dext-5	2431	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
37	Dext-6	3977	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
38	Dext-7	4057	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
39	Dext-8	1595	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
40	Dext-9	5475	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
41	Dext-10	2820	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

ID*	Probe <sup>b</sup>	hMalectin	CtCBM25	400-2	400-3	CtCBM11	CmCBM6-2	LM10	LM11	CtCBM22-2	LM6	400-4	LM21	CtCBM35	CCRC-M70	LM24	LM25	CCRC-M1	AAL
42	Dext-11	4324	-	643	-	-	1121	-	-	-	-	-	-	-	-	-	-	-	-
43	Dext-12	5173	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
44	Dext-13	5731	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
45	Pano-3	3029	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
46	i-Pano-3	537	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
47	Pullu-4	5817	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
48	Pullu-6	546	1322	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
49	Pullu-7	5203	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
50	CβG-2	-	-	-	-	-	8456	-	-	-	-	-	-	-	-	-	-	-	-
51	CβG-3	-	-	-	-	-	25747	-	-	-	-	-	-	-	-	-	-	-	-
52	CβG-4	-	-	-	-	-	25788	-	-	-	-	-	-	-	-	-	-	-	-
53	CβG-5	-	-	-	-	-	26575	-	-	-	-	-	-	-	-	-	-	-	-
54	CβG-6	-	-	-	-	-	29013	-	-	-	-	-	-	-	-	-	-	-	-
55	CβG-7	-	-	-	-	-	21389	-	-	-	-	-	-	-	-	-	-	-	-
56	CβG-8	-	-	-	-	-	19145	-	-	-	-	-	-	-	-	-	-	-	-
57	CβG-9	-	-	-	-	-	19735	-	-	-	-	-	-	-	-	-	-	-	-
58	CβG-10	-	-	-	-	-	19448	-	-	-	-	-	-	-	-	-	-	-	-
59	CβG-11	-	-	-	-	-	17523	-	-	-	-	-	-	-	-	-	-	-	-
60	CβG-12	-	-	-	-	-	23896	-	-	-	-	-	-	-	-	-	-	-	-
61	CβG-13	-	-	-	-	-	24428	-	-	-	-	-	-	-	-	-	-	-	-
62	Lam-2	-	-	-	-	-	16840	-	-	-	-	-	-	-	-	-	-	-	-
63	Lam-3	907	-	788	-	-	18838	-	-	-	-	-	-	-	-	-	-	-	-
64	Lam-4	3498	-	33770	-	-	30839	-	-	-	-	-	-	-	-	-	-	-	-
65	Lam-5	4137	-	39832	-	-	28217	-	-	-	-	-	-	-	-	-	-	-	-
66	Lam-6	4517	-	32174	-	-	21840	-	-	-	-	-	-	-	-	-	-	-	-
67	Lam-7	4107	-	28121	-	-	15705	-	-	-	-	-	-	-	-	-	-	-	-
68	Curd-8	3098	-	41395	-	-	25561	-	-	-	-	-	-	-	-	-	-	-	-
69	Curd-9	5654	-	44383	-	-	26357	-	-	-	-	-	-	-	-	-	-	-	-
70	Curd-10	5784	-	44162	-	-	27613	-	-	-	-	-	-	-	-	-	-	-	-
71	Curd-11	4739	-	38366	-	-	20897	-	-	-	-	-	-	-	-	-	-	-	-
72	Curd-12	4850	-	48167	876	-	28842	-	-	-	-	-	-	-	-	-	-	-	-
73	Curd-13	3537	-	39715	576	-	18338	-	-	-	-	-	-	-	-	-	-	-	-
74	NSG-11	4904	-	34940	-	-	30793	-	-	-	-	-	-	-	-	-	-	-	-
75	HE-8	3551	-	32912	-	-	20131	-	-	-	-	-	-	-	-	-	-	-	-
76	HE-9	1616	-	26274	-	-	13392	-	-	-	-	-	-	-	-	-	-	-	-
77	HE-10	1562	-	23804	-	-	8295	-	-	-	-	-	-	-	-	-	-	-	-
78	Cellobiose	748	-	-	-	-	10886	-	-	-	-	-	-	-	-	-	-	-	-
79	Cello-4	762	-	-	4451	-	21582	-	-	-	-	-	-	-	-	-	-	-	-
80	Cello-5	805	-	-	4794	-	18290	-	-	-	-	-	-	-	-	-	840	-	-
81	Cello-6	-	-	-	3638	-	18081	-	-	-	-	-	-	-	-	-	915	-	-
82	Cello-7	-	-	-	3229	-	14935	-	-	-	-	-	-	-	-	-	1353	-	-
83	Cello-8	-	-	-	4324	-	20261	-	-	-	-	-	-	-	-	-	2060	-	-
84	Cello-9	1309	-	-	7142	2306	30818	-	-	-	-	-	-	-	-	-	3174	962	-
85	Cello-10	1572	-	-	6933	3250	33677	-	-	-	-	-	-	-	-	-	2815	719	-
86	Cello-11	994	-	-	2415	4396	35561	-	-	-	-	-	-	-	-	-	1449	-	-
87	Cello-12	588	-	-	4261	2035	20385	-	-	-	-	-	-	-	-	-	1191	-	-

CHAPTER 2. SUPPLEMENTARY INFORMATION

ID*	Probe*	hMalectin	CtCBM25	400-2	400-3	CtCBM11	CmCBM6-2	LM10	LM11	CtCBM22-2	LM6	400-4	LM21	CtCBM35	CCRC-M70	LM24	LM25	CCRC-M1	AAL
88	Cello-13	1115	-	-	6721	3851	35688	-	-	-	-	-	-	-	-	-	2096	562	-
89	Gentiobiose	680	-	-	-	-	11309	-	-	-	-	-	-	-	-	-	-	-	-
90	Pust-3	-	-	-	-	-	23381	-	-	-	-	-	-	-	-	-	-	-	-
91	Pust-4	-	-	-	-	-	20850	-	-	-	-	-	-	-	-	-	-	-	1
92	Pust-5	-	-	-	-	-	32800	-	-	-	-	-	-	-	-	-	-	-	-
93	Pust-6	-	-	-	-	-	22879	-	-	-	-	-	-	-	-	-	-	-	-
94	Pust-7	-	-	-	-	-	15671	-	-	-	-	-	-	-	-	-	-	-	-
95	Pust-8	-	-	-	-	-	7613	-	-	-	-	-	-	-	-	-	-	-	-
96	Pust-9	-	-	-	-	-	9744	-	-	-	-	-	-	-	-	-	-	-	-
97	Pust-10	-	-	-	-	-	10965	-	-	-	-	-	-	-	-	-	-	-	-
98	Pust-11	-	-	-	-	-	6891	-	-	-	-	-	-	-	-	-	-	-	-
99	Pust-15	-	-	-	-	-	6641	-	-	-	-	-	-	-	-	-	-	-	-
100	Pust-15a	-	-	-	-	-	8906	-	-	-	-	-	-	-	-	-	-	-	-
101	Barley-3	-	-	-	-	-	23759	-	-	-	-	-	-	-	-	-	-	-	-
102	Barley-3a	2346	-	1620	-	-	10059	-	-	-	-	-	-	-	-	-	-	-	-
103	Barley-4	-	-	-	2993	-	31659	-	-	-	-	-	-	-	-	-	-	-	-
104	Barley-4a	2357	-	602	2362	-	25544	-	-	-	-	-	-	-	-	-	-	-	-
105	Barley-4b	-	-	-	3651	-	32173	-	-	-	-	-	-	-	-	-	-	-	-
106	Barley-4c	-	-	-	-	-	29542	-	-	-	-	-	-	-	-	-	-	-	-
107	Barley-5	-	-	-	11189	1298	28143	-	-	-	-	-	-	-	-	-	-	-	-
108	Barley-5a	4868	-	4423	16165	-	30645	-	-	-	-	-	-	-	-	-	-	-	-
109	Barley-6	-	-	-	18752	2339	23518	-	-	-	-	-	-	-	-	-	-	-	-
110	Barley-6a	8370	-	3907	14646	-	44277	-	-	-	-	-	-	-	-	-	-	-	-
111	Barley-7	-	-	-	27077	16177	34423	-	-	-	-	-	-	-	-	-	-	-	-
112	Barley-8	-	-	-	15320	12698	21560	-	-	-	-	-	-	-	-	-	-	-	-
113	Barley-9	-	-	-	24979	31593	28286	-	-	-	-	-	-	-	-	-	-	-	-
114	Barley-10	-	-	-	17785	19418	11086	-	-	-	-	-	-	-	-	-	-	-	-
115	Barley-11	-	-	-	18306	21911	24647	-	-	-	732	-	-	-	-	-	-	-	-
116	Barley-12	-	-	-	25441	30594	28306	-	-	-	-	-	-	-	-	-	-	-	-
117	Barley-13	-	-	-	20542	18931	17174	-	-	-	-	-	-	-	-	-	-	-	-
118	Barley-14	-	-	-	25852	7868	25529	-	-	-	-	-	-	-	-	-	-	-	-
119	Barley-15	-	-	-	20235	14544	24190	-	-	-	-	-	-	-	-	-	-	-	-
120	Barley-16	-	-	-	18605	28558	18166	1541	2501	6692	-	-	-	-	-	-	-	-	-
121	Grifo-3	-	-	-	-	-	25910	-	-	-	-	-	-	-	-	-	-	-	-
122	Grifo-4	-	-	966	701	-	22915	-	-	-	-	-	-	-	-	-	-	-	-
123	Grifo-5	-	-	1012	709	-	24924	-	-	-	-	-	-	-	-	-	-	-	-
124	Grifo-6	509	-	996	-	-	24261	-	-	-	-	-	-	-	-	-	-	-	-
125	Grifo-7	806	-	1194	-	-	28271	-	-	-	-	-	-	-	-	-	-	-	-
126	Grifo-8	1153	-	1234	-	-	22558	-	-	-	-	-	-	-	-	-	-	-	-
127	Grifo-9	-	-	738	-	-	21603	-	-	-	-	-	-	-	-	-	-	-	-
128	Grifo-10	-	-	-	-	-	22190	-	-	-	-	-	-	-	-	-	-	-	-
129	Grifo-11	518	-	-	-	-	21855	-	-	-	-	-	-	-	-	-	-	-	-
130	Grifo-12	-	-	715	-	-	18385	-	-	-	-	-	-	-	-	-	-	-	-
131	Grifo-13	-	-	659	-	-	22175	-	-	-	-	-	-	-	-	-	-	-	-
132	Grifo-14	-	-	249	-	-	13168	-	-	-	-	-	-	-	-	-	-	-	-
133	Grifo-15	-	-	277	-	-	17999	-	-	-	-	-	-	-	-	-	-	-	-

ID*	Probe*	hMalectin	CtCBM25	400-2	400-3	CtCBM11	CmCBM6-2	LM10	LM11	CtCBM22-2	LM6	400-4	LM21	CtCBM35	CCRC-M70	LM24	LM25	CCRC-M1	AAL
134	Grifo-16	-	-	1031	-	-	17339	-	-	-	-	-	-	-	-	-	-	-	-
135	Lentin-2	1677	-	-	-	-	16935	-	-	-	-	-	-	-	-	-	-	-	-
136	Lentin-3	1920	-	645	-	-	30439	-	-	-	-	-	-	-	-	-	-	-	-
137	Lentin-4	2307	-	2247	-	-	23914	-	-	-	-	-	-	-	-	-	-	-	-
138	Lentin-5	2135	-	22693	-	-	35202	-	-	-	-	-	-	-	-	-	-	-	-
139	Lentin-6	3266	-	15125	-	-	30787	-	-	-	-	-	-	-	-	-	-	-	-
140	Lentin-7	3117	-	11407	-	-	31686	-	-	-	-	-	-	-	-	-	-	-	-
141	Lentin-8	4761	-	15766	-	-	37287	-	-	-	-	-	-	-	-	-	-	-	-
142	Lentin-9	5620	-	12926	-	-	22775	-	-	-	-	-	-	-	-	-	-	-	-
143	Lentin-10	7036	-	9988	-	-	23159	-	-	-	-	-	-	-	-	-	-	-	-
144	Lentin-11	5961	-	8641	-	-	23897	-	-	-	-	-	-	-	-	-	-	-	-
145	Lentin-12	3446	-	8975	-	-	20509	-	-	-	-	-	-	-	-	-	-	-	-
146	Lentin-13	7773	-	24691	-	-	30322	-	-	-	-	-	-	-	-	-	-	-	-
147	HE-9B7	3145	-	12463	-	-	20693	-	-	-	-	-	-	-	-	-	-	-	-
148	HE-10B2	7089	-	38414	-	-	31028	-	-	-	-	-	-	-	-	-	-	-	-
149	HE-10B3	7254	-	39663	-	-	31276	-	-	-	-	-	-	-	-	-	-	-	-
150	HE-10B5	6478	-	28213	-	-	28830	-	-	-	-	-	-	-	-	-	-	-	-
151	HE-10B7	3505	-	20849	-	-	30752	-	-	-	-	-	-	-	-	-	-	-	-
152	Gu-6B1/3	963	-	-	-	-	29544	-	-	-	-	-	-	-	-	-	-	-	-
153	HE-11B3/6	4108	-	4556	-	-	25704	-	-	-	-	-	-	-	-	-	-	-	-
154	Xyl-3	-	-	-	-	-	4117	38464	13828	570	-	-	-	-	-	-	-	-	-
155	Xyl-4	-	-	-	-	-	2948	28021	14098	3215	-	-	-	-	-	-	-	-	-
156	Xyl-5( $\beta$ 4)	-	-	-	-	-	4092	31848	16338	11406	-	-	-	-	-	-	-	-	-
157	Xyl-6( $\beta$ 4)	-	-	-	-	-	4944	37861	17042	17999	-	-	-	-	-	-	-	-	-
158	Xyl-7	-	-	-	-	-	5613	56494	20098	34844	-	-	-	-	-	-	-	-	-
159	Xyl-8	-	-	-	-	-	5184	54652	19450	39891	-	-	-	-	-	-	-	-	-
160	Xyl-9	-	-	-	-	-	8596	49824	19699	44545	-	-	-	-	-	-	-	-	-
161	Xyl-10	-	-	-	-	-	3489	45970	17428	33815	-	-	-	-	-	-	-	-	-
162	Xyl-11	-	-	-	-	-	5456	39072	18654	45723	-	-	-	-	-	-	-	-	-
163	Xyl-12	-	-	-	-	-	2328	33903	17289	32272	-	-	-	-	-	-	-	-	-
164	Xyl-13	-	-	-	-	-	4700	38896	15291	37049	-	-	-	-	-	-	-	-	-
165	Ara-Xylan-3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
166	Ara-Xylan-4a	-	-	-	-	-	3451	2763	14452	-	-	-	-	-	-	-	-	-	-
167	Ara-Xylan-4b	-	-	-	-	-	3575	8688	15723	-	-	-	-	-	-	-	-	-	-
168	Ara-Xylan-5a	-	-	-	-	-	5789	15896	3247	7593	-	-	-	-	-	-	-	-	-
169	Ara-Xylan-5b	-	-	-	-	-	2942	35641	9488	6572	-	-	-	-	-	-	-	-	-
170	Ara-Xylan-5c	-	-	-	-	-	-	-	9344	1460	-	-	-	-	-	-	-	-	-
171	Ara-Xylan-6	-	-	-	-	-	4977	30250	1325	5944	-	-	-	-	-	-	-	-	-
172	Ara-2( $\alpha$ 5)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
173	Ara-3( $\alpha$ 5)	-	-	-	-	-	-	-	-	-	5443	-	-	-	-	-	-	-	-
174	Ara-4( $\alpha$ 5)	-	-	-	-	-	-	-	-	-	18306	-	-	-	-	-	-	-	-
175	Ara-5( $\alpha$ 5)	-	-	-	-	-	-	-	-	-	23714	-	-	-	-	-	-	-	-
176	Ara-6( $\alpha$ 5)	-	-	-	-	-	-	-	-	-	43262	-	-	-	-	-	-	-	-
177	Ara-7( $\alpha$ 5)	-	-	-	-	-	-	-	-	-	40092	-	-	-	-	-	-	-	-
178	Ara-8( $\alpha$ 5)	-	-	-	-	-	-	-	-	-	18672	-	-	-	-	-	-	-	-
179	Ara-9( $\alpha$ 5)	-	-	-	-	-	-	-	-	-	17959	-	-	-	-	-	-	-	-

CHAPTER 2. SUPPLEMENTARY INFORMATION

ID <sup>a</sup>	Probe <sup>b</sup>	hMalectin	CtCBM25	400-2	400-3	CtCBM11	CmCBM6-2	LM10	LM11	CtCBM22-2	LM6	400-4	LM21	CtCBM35	CCRC-M70	LM24	LM25	CCRC-M1	AAL
180	Ara-4B3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
181	Ara-5B	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
182	Man-4(β4)	-	-	-	-	-	5243	-	-	-	15133	10884	41819	4726	-	-	-	-	-
183	Man-5(β4)	-	-	-	-	-	7731	-	-	-	-	10114	37066	6085	-	-	-	-	-
184	Man-6(β4)	-	-	-	-	-	8721	-	-	-	-	20289	52247	14864	-	-	-	-	-
185	Man-8(β4)	-	-	-	-	-	764	-	-	-	-	21759	38401	11514	-	-	-	-	-
186	Gal-Mannan-2e	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
187	Gal-Mannan-3e	-	-	-	-	-	1322	-	-	-	-	-	7194	-	-	-	-	-	-
188	Gal-Mannan-4e	-	-	-	-	-	1970	-	-	-	-	1007	17455	-	-	-	-	-	-
189	Gal-Mannan-5e	-	-	-	-	-	1897	-	-	-	-	7435	35210	-	-	-	-	-	-
190	Gal-Mannan-6e	-	-	-	-	-	1709	-	-	-	-	5778	25781	1138	-	-	-	-	-
191	Gal-Mannan-7e	-	-	-	-	-	2172	-	-	-	-	17254	40697	6685	-	-	-	-	-
192	Gal-Mannan-8e	-	-	-	-	-	1686	-	-	-	-	6616	16554	1101	-	-	-	-	-
193	Gal-Mannan-5m	-	-	-	-	-	3698	-	-	-	-	-	3412	-	-	-	-	-	-
194	Gal-Mannan-6m	-	-	-	-	-	3525	-	-	-	-	4802	1430	-	-	-	-	-	-
195	Gal-Mannan-7m	-	-	-	-	-	6395	-	-	-	-	-	-	-	-	-	-	-	-
196	Gal-Mannan-8m	-	-	-	-	-	4417	-	-	-	-	1484	-	-	2022	-	-	-	-
197	Gal-Mannan-9e	-	-	-	-	-	4818	-	-	-	-	23010	48187	14775	1204	-	-	-	-
198	Gal-Mannan-10e	-	-	-	-	-	2128	-	-	-	-	16522	36671	16277	566	-	-	-	-
199	Gal-Mannan-11e	-	-	-	-	-	3731	-	-	-	-	21199	40755	21705	1476	-	-	-	-
200	Xyl-Glucan-7	-	-	-	-	-	16775	-	-	-	-	-	-	-	-	631	44482	-	-
201	Xyl-Glucan-8	-	-	-	-	-	10370	-	-	-	-	-	-	-	-	15406	46234	-	-
202	Xyl-Glucan-9	-	-	2120	-	-	10148	-	-	-	-	-	-	-	-	43222	27642	-	-
203	FG-Xyl-Glucan-6	-	-	-	-	-	31063	-	-	-	-	-	-	-	-	52678	37033	34433	29573
204	FG-Xyl-Glucan-9	-	-	-	-	-	16892	-	-	-	-	-	-	-	-	45090	36633	2806	2913

<sup>a</sup>ID, Probe position in the microarray matching the position in the heatmap, binding-charts and in Table S2.1;<sup>b</sup>In the β1,6-linked pustulan series, fractions containing oligomers with >DP-8 as major components (probes 100-103) there was evidence of a minor contaminant containing α-linked mannose; In the NGLs of oligosaccharide fractions derived from branched β1,3(β1,6)-lentinan (probes 142-151) there was presence of an α1,4-linked glucose contaminant (not shown)<sup>32</sup>; <sup>c</sup>The binding signals are means of fluorescence intensities of duplicate spots at 5 fmol of probe arrayed (the respective standard deviation was calculated as the associated error, overall < 5%). ‘-’ refers to a fluorescence intensity < 500.

**Table S2.5. Information on the oligosaccharide neoglycolipid probes printed and validated in the xyloglucan microarrays.** The NGL probes are sorted by source and degree of polymerization.

ID <sup>a</sup>	Linkages & Sources	Probe Designation <sup>b</sup>	Probe Sequence <sup>c,d</sup>
1	Branched Glc $\beta$ 4(Xyl $\alpha$ 6Gal $\beta$ 4) Tamarind xyloglucan	Xyl-Glucan-7-DAN-DHPA	<pre> Xyl<math>\alpha</math>-6   Glc<math>\beta</math>-4Glc<math>\beta</math>-4Glc<math>\beta</math>-4Glc-DAN- DHPA   Xyl<math>\alpha</math>-6      Xyl<math>\alpha</math>-6 </pre>
2		Xyl-Glucan-8-DAN-DHPA	<pre> Xyl<math>\alpha</math>-6   Glc<math>\beta</math>-4Glc<math>\beta</math>-4Glc<math>\beta</math>-4Glc-DAN- DHPA   Xyl<math>\alpha</math>-6      Xyl<math>\alpha</math>-6   Gal<math>\beta</math>-2 </pre>
3		Xyl-Glucan-9-DAN-DHPA	<pre> Gal<math>\beta</math>-2   Xyl<math>\alpha</math>-6   Glc<math>\beta</math>-4Glc<math>\beta</math>-4Glc<math>\beta</math>-4Glc-DAN- DHPA   Xyl<math>\alpha</math>-6      Xyl<math>\alpha</math>-6   Gal<math>\beta</math>-2 </pre>
4		Xyl-Glucan-7-AO	<pre> Xyl<math>\alpha</math>-6   Glc<math>\beta</math>-4Glc<math>\beta</math>-4Glc<math>\beta</math>-4Glc-AO   Xyl<math>\alpha</math>-6      Xyl<math>\alpha</math>-6 </pre>
5		Xyl-Glucan-8-AO*	<pre> Xyl<math>\alpha</math>-6   Glc<math>\beta</math>-4Glc<math>\beta</math>-4Glc<math>\beta</math>-4Glc-AO*   Xyl<math>\alpha</math>-6      Xyl<math>\alpha</math>-6   Gal<math>\beta</math>-2   Xyl<math>\alpha</math>-6   Glc<math>\beta</math>-4Glc<math>\beta</math>-4Glc<math>\beta</math>-4Glc-AO   Xyl<math>\alpha</math>-6      Xyl<math>\alpha</math>-6 </pre>
6		Xyl-Glucan-9-AO	<pre> Gal<math>\beta</math>-2   Xyl<math>\alpha</math>-6   Glc<math>\beta</math>-4Glc<math>\beta</math>-4Glc<math>\beta</math>-4Glc-AO   Xyl<math>\alpha</math>-6      Xyl<math>\alpha</math>-6   Gal<math>\beta</math>-2 </pre>
7	Branched Glc $\beta$ 4(Xyl $\alpha$ 6Gal $\beta$ 4Fuca $\alpha$ 2) Apple xyloglucan	FG-Xyl-Glucan-7-DAN-DHPA	<pre> Fuca<math>\alpha</math>-2   Gal<math>\beta</math>-2   Xyl<math>\alpha</math>-6   Glc<math>\beta</math>-4Glc<math>\beta</math>-4Glc-DAN-DHPA   Xyl<math>\alpha</math>-6 </pre>
8		FG-Xyl-Glucan-10-DAN-DHPA*	<pre> Gal<math>\beta</math>-2   Xyl<math>\alpha</math>-6   Glc<math>\beta</math>-4Glc<math>\beta</math>-4Glc<math>\beta</math>-4Glc-DAN- DHPA   Xyl<math>\alpha</math>-6      Xyl<math>\alpha</math>-6   Gal<math>\beta</math>-2   Fuca<math>\alpha</math>-2 </pre>
9		FG-Xyl-Glucan-6-AO*	<pre> (Fuca<math>\alpha</math>-2)   Gal<math>\beta</math>-2   Xyl<math>\alpha</math>-6   Glc<math>\beta</math>-4Glc<math>\beta</math>-4Glc-AO   Xyl<math>\alpha</math>-6 </pre>

10		FG-Xyl-Glucan-9-AO*	$  \begin{array}{c}  \text{Gal}\beta\text{-2} \\    \\  \text{Xyl}\alpha\text{-6} \\    \\  \text{Glc}\beta\text{-4Glc}\beta\text{-4Glc}\beta\text{-4Glc}\text{-AO} \\    \qquad   \\  \text{Xyl}\alpha\text{-6} \qquad \text{Xyl}\alpha\text{-6} \\    \\  \text{Gal}\beta\text{-2} \\    \\  \text{(Fuca}\alpha\text{-2)}  \end{array}  $
11	<p>Branched Glc<math>\beta</math>4(Xyl<math>\alpha</math>6Gal<math>\beta</math>4Fuca2) Apple xyloglucan</p>	Xyl-Glucan DP4-AO*	<p>Glc<math>\beta</math>4(Xyl<math>\alpha</math>6Gal<math>\beta</math>4Fuca2)<sub>4-13</sub>-AO</p>
12		Xyl-Glucan DP5-AO*	
13		Xyl-Glucan DP6a-AO*	
14		Xyl-Glucan DP6b-AO*	
15		Xyl-Glucan DP7-AO*	
16		Xyl-Glucan DP8a-AO*	
17		Xyl-Glucan DP8b-AO*	
18		Xyl-Glucan DP9-AO*	
19		Xyl-Glucan DP11/12-AO*	
20		Xyl-Glucan DP13-AO*	

<sup>a</sup>ID, Probe position in the microarray matching the position in the binding-charts.

<sup>b</sup>Abbreviations for oligosaccharide moieties: Xyl-Glucan-, from non-fucosylated xyloglucan from tamarind; and Fuc-Xyl-Glucan-, from fucosylated xyloglucan from apple;

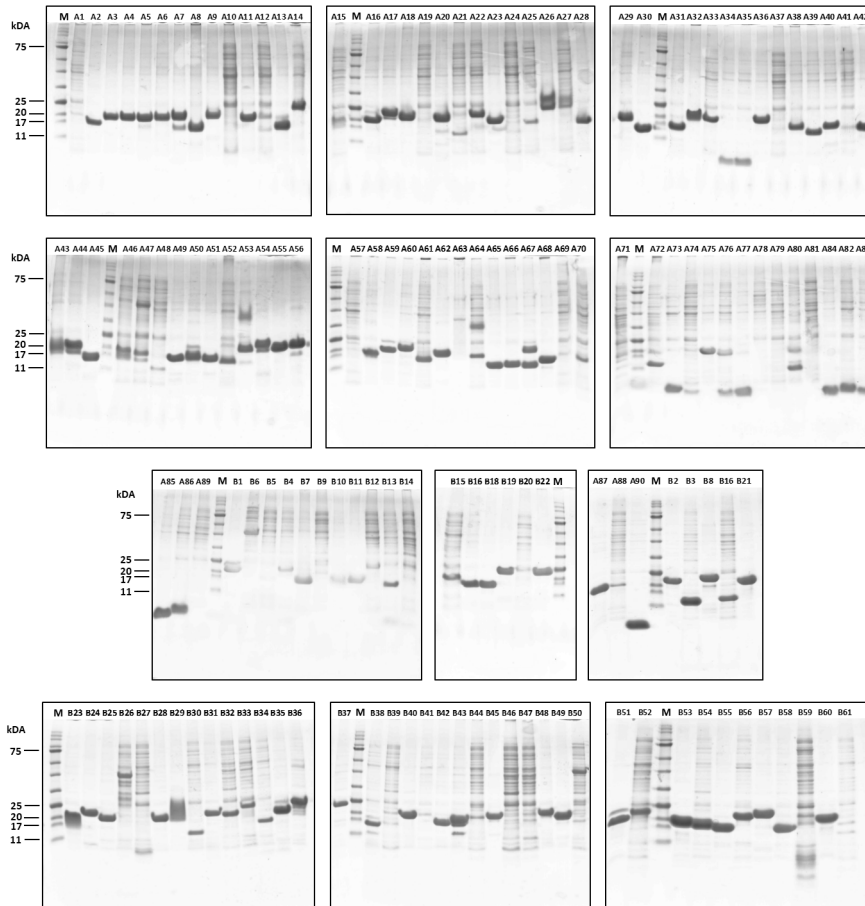
<sup>c</sup>The oligosaccharide probes are all lipid-linked, neoglycolipids (NGLs); AO-NGLs probes 4-6 and 9-10 were used as controls in this microarray, corresponding to probes 202-204 in Table S2.1; and DHPA-NGLs prepared by reductive amination with the amino lipid *N*-(4-formylbenzamide)-1,2-dihexadecyl-sn-glycero-3-phosphoethanolamine (DHPA);

<sup>d</sup>An asterisk indicates the major component when multiple components are present.

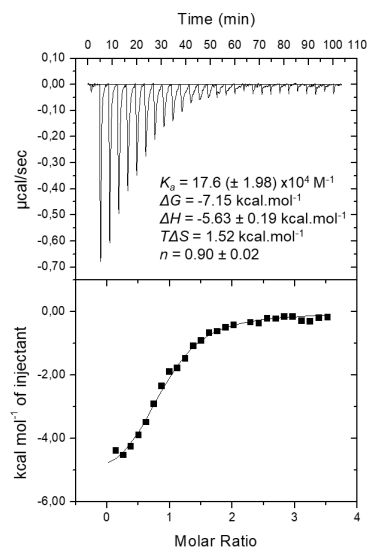


## Chapter 3 - Supplementary Information

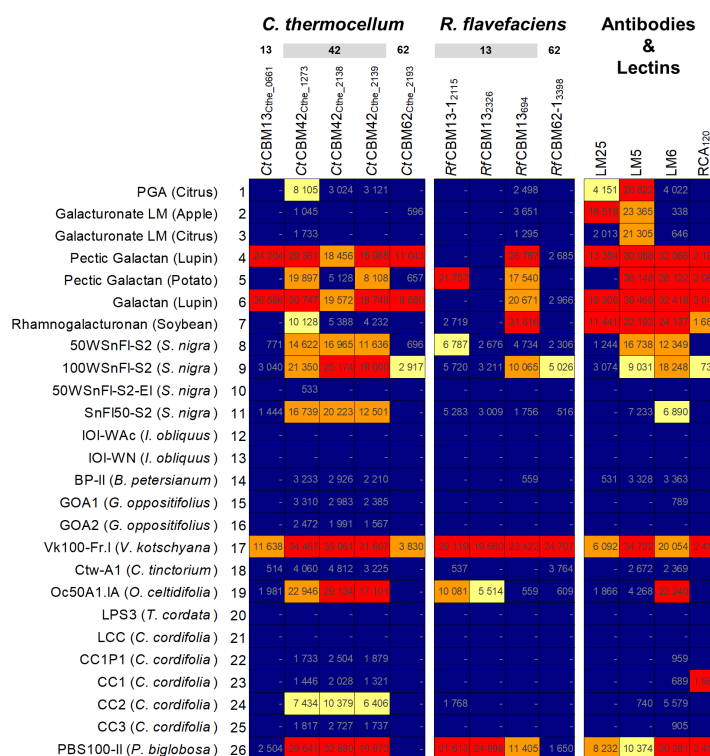
## Supplementary Figures



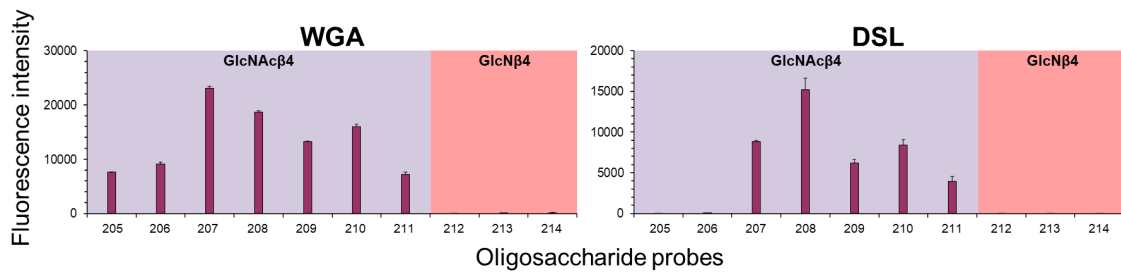
**Figure S3.1. Quality control of *C. thermocellum* and *R. flavefaciens* FD-1 recombinant CBMs produced using the high-throughput platforms.** CBMs were subjected to a denaturing gel containing 13% (w/v) acrylamide SDS-PAGE analysis using a Tris-tricine buffer system and stained with Coomassie Blue. Each lane contained 5  $\mu$ g of total protein. NZYColour Protein Marker II was used as marker (lanes M). CBMs are identified at the top using the code IDs A1-89 for *C. thermocellum* and B1-61 for *R. flavefaciens* FD-1, according to Tables S3.1.



**Figure S3.2. Isothermal calorimetry titrations of binding of CtCBM13<sub>Cthe\_0661</sub> to pustulan polysaccharide.** The top portion of each panel shows the raw power data while the bottom parts show the integrated and heat of dilution corrected data. The solid lines show the non-linear curve fits to a one site binding model with the stoichiometry fixed at 1. Thermodynamic parameters are given in the top panel.



**Figure S3.3. Analysis of *C. thermocellum* and *R. flavefaciens* CBMs on a pectin polysaccharide microarray.** The microarrays included 26 pectin-related polysaccharides. Carbohydrate sequence information on these probes is in Table S3.9. CBMs for which binding was obtained are presented at the top, organised by CAZy family for each bacterium. Monoclonal antibodies LM5 and LM6 and plant lectin RCA<sub>120</sub> used for the microarray quality control are also presented. The relative binding intensities were calculated as the percentage of the fluorescence signal intensity at 5 fmol given by the probe most strongly bound by each protein (normalized as 100%) and are representative of at least two independent experiments, with an error below 20%. Numerical scores are given in the heatmap, where ‘-’ refers to a fluorescence intensity < 500.



**Figure S3.4. Validation of the chitin and chitosan NGL probes included in the glucan, hemicellulose, chitin and chitosan NGL-microarrays.** The binding signals of lectins WGA and DSL are depicted as means of fluorescence intensities of duplicate spots at 5 fmol of oligosaccharide probe arrayed (with error bars) and are representative of at least two independent experiments. The different carbohydrate groups are indicated in the coloured panels. Carbohydrate sequence information on these probes is in Chapter 2, Table S2.1.

## Supplementary Tables

**Table S3.1. Modular architecture and primary sequences of *C. thermocellum* and *R. flavefaciens* FD-1 CBMs for which carbohydrate binding patterns were obtained.** CBMs analysed are highlighted in bold and the respective sequence of the recombinant protein expressed is shown. Each CBM is identified by organism, high-throughput ID code, and assigned protein ID.

	Family	Code ID	Protein ID	Molecular architecture <sup>a</sup>	Primary sequence
<i>C. thermocellum</i>	3	A3	Cthe_3077	<b>CBM3</b> -DOC1	PVSGNLKVEFYNSNPSDTTNSINPQFKVTNTGSSAIDLKSLTLRYYYTVDQKDKQTFWCDHAAIIGSN GSYNGITSNVKGTFTVKMSSSTNNADTYLEISFTGGTLEPGAHVQIQGRFAKNDWSNYTQSNDSYFK SASQFVEWDQVTAYLNGVLVWGKEP
		A20	Cthe_0433	GH9- <b>CBM3</b> -DOC1	GGSYWVEAFGVDIVQSDGPKATEVTLYVRSRPSKSNISVRYFFDATGMSSVDPDKMEIRQLYDQ TAAETDYAAKLTGPHHYKDNIIYVEISWEGFAIANSNKYQFALGTYTWGNSWDPTDDWSYQELKI EESNYTGTPARNNRICVYDAGVLVGGIEP
		A6	Cthe_0040	GH9- <b>CBM3</b> - <b>CBM3</b>	QGKIGEVVLQYANGNAGATSNSINPRFKIINNGTKAINLSDVKIRYYYYTKEGGASQNFWCDWSSAGN SNVTGNFFNLSSPKEGADTCLVGFVGFSGAGTLDPPGGSVEVQIRFSKEDWSNYNQSNDSYFNPSAS DYTDWNRVTLYISNKLTVYGKEP
		A12	Cthe_0059	<b>CBM3</b>	QDGTKGLKIQYYSRKPHDSAGIDFSFRMFNTGNEAIDLKDVKIRYFFKEDVSIDEMNWAVFYSLGS EKDVQCRFYELPGKKEANKYLEITFKSGTLPNDVMYITGEFYKNDWTKFEQRDDYSYNPADSYSD WKRMTAYISNKLWVGGIEPN
	4	A64	Cthe_2809	SLH-SLH-SLH-CBM54-GH16- <b>CBM4</b> -CBM4-CBM4-CBM4	IYNGGFDVDDSAAVGVDGVPYTSYWTFLTASGGAATVNVVEEGVMHVQIENGTTDYGVQLLQAPIH LEKGAKYKASFDMAENPRQVKLKIGDGDGRGWKDYAAPPFTVSTEMTNYEFEFTMKDDTDVKAR FEFNMGLDDNDVWIDNVKLIKTEDAPVI
		A62	Cthe_2809	SLH-SLH-SLH-CBM54-GH16- CBM4-CBM4- <b>CBM4</b> -CBM4	ILNGVFNGLAGWGYGAYEPGSADFESHEEQFRAIISVGNWVQVLYQDNVPLEQGQTYEVSFD AKSTIDRKIIVQLQRNGTSDNNWDSYFYQEVELTNELKTFKYEFTMSKPTDSASRFNFALGNTENKT YAPHEIIDNVVVRKVATPSAL
		A14	Cthe_0413	<b>CBM4</b> -GH9-CBM3-DOC1	PYKNDLLYERTFDEGLCYPWHTCEDSGGKCSFDVVDVPGQPGNKAFVTVLQKGNRWSVQMRH RGLTLEQGHTYRVLKIWADASCKVYIKIQMGEPYAEYWNKWSPTYTLTAGKVLEIDETFVMDKPT DDTCEFTFHLGGELAATPPYTVYLDVSLYDPEY
	6	A30	Cthe_1271	GH43-CBM6- <b>CBM6</b> -DOC1	VTERSAFSKIEVEDFNDIKSSTIQIGTPNGGSGIGYIENGDWLAYKNIDFGNGATTFKALVASTLSPNI ELRLDSPTGTLIGTLKVAATGGFNAYEEQSCNISKVTGKHDLYLVFGAVNIDWFTFGSSGII
		A49	Cthe_2197	GH2- <b>CBM6</b> -DOC1	PVPRSAFTRIEAESYDAQSGIQTEDCSEGGKDVGYIENGDFVYKAIIDFGGAAAFKARVASATSGG NIELRIDSIDGPPVVICPVAGTGGWQEWADATCEVSDLKGVDHLYLKFTGGSGYLLNVNWFVFEV NSDED
		A52	Cthe_2194	CE1- <b>CBM6</b> -DOC1	RSFAFTRIEAEDFDNMSGIENESCSEGLNIGYIENGDYVAYSNIIDFGNGAKEFQARVASATSGGKIEI RLDSITGPLIGTCSVSGTGGWQQWVDVKCEVSGVSGTHDLYLKFTGGSGYLLNINWWKFTQAD
		A51	Cthe_2195	<b>CBM6</b> -DOC1	EPRSAFTRIEAESYNGQSGIQTECNSEGGMDVGYIENGDYVYKNIIDFGGAAAFKARVASATSGG NIELRIDSIDGPPVVICPVAGSGGWQQWVDATCEVSLKGVHDLYLKFTGGSGYLLNINWFVFEV NDE
		A39	Cthe_1963	CE1- <b>CBM6</b> -DOC1-GH10	ANTRIEAEDYDGINSSSIEIIGVPEGGRGIGYITSGDYL VYKSIDFGNGATSFKAKVANANTSNIELRL NGPNTLIGTLVSKSTGDWNTYEEQTCISIKVTGINDLYLVFKGPVNDWFTFGVSSS

	A65	Cthe_3012	GH30- <b>CBM6</b> -DOC1	VERNAFSKIECEEYNATNSSTVQVVGTGTGSGGLGYIENGNFYFAYKNINFGNGANSFKIRAATTGTPKI EIRLGSPTGTLAGTLQVAATGGFNAYEEQSCSINKITGVQDVYLVFGGAVNVVDWFTFE
	A66	Cthe_2972	GH11- <b>CBM6</b> -DOC1-CE4	TPRSAFSKIEAEEYNSLKSSTIQTIGTSDGGSGIGYIESGDYLVFNKINFGNGANSFKARVASGADTPT NIQLRLGSPTGTLIGTLTVASTGGWNNYEEKSCSITNTTGQHDLYLVFSGPVNIDYFIFDSNGVNP
11	A36	Cthe_1472	GH26-GH5- <b>CBM11</b> -DOC1	AVGEKMLDDFEGVLNWSYSYSGEGAKVSTKIVSGKTGNGMEVSYTGTTDGYWGTVYSLPDGDWSK WLKISFDIKSVDGSANEIRFMIAEKINGVGDGEHWVYSITPDSSWKTIPIPFSSFRRLDYQPPGQD MSGTLDLDNIDSIHFMYANNKSGKFVVDNIKLIGATSDP
13	A2	Cthe_0661	GH43- <b>CBM13</b> -DOC1	TRYKLVNKNNGKVLVDLDSVDNAAQIVQWTDNGSLSQQWYLVVDVGGGYKKIVNVKSGRALDVKD ESKEDGGVLIQYTSNGGYNQHWKFTDIGDGYKISSRHCGKIDVRKWSTRDGGIIQQWSDAGGTN QHWKLVLVSS
22	A33	Cthe_1838	<b>CBM22</b> -GH10-DOC1	ASAAALIYDDFETGLNGWGRPGPETVELTTEEAYSGRYSKLVSGRTSTWNGPMVDKTDVLTGSEY KLGYYVKFVGDSYSNEQRFLQLQYNDGAGDVYQNIKTATVYKGTWTLLEGQLTVP SHAKDVKIYV ETEFKNSPSPQDLMDFYIDDFTAT
	A53	Cthe_2590	<b>CBM22</b> -GH10-DOC1	AEGNLLFNPGFELGSTEGWYPYGECTIEAVGTEAHSNGYSVFTDRTQDWNGVAQDMLDKLTVGM TYQVSAWVKVAGTGS HQVKISMKKVETGKEPVYDNIASITVEGSEWYRLSGPYSYTGTVNLTLELYI EGPQPGVSYVDDVTVEVGS
	A27	Cthe_0912	<b>CBM22</b> -GH10-CBM22-DOC1-CE1	ASAAALIYDDFETGLNGWGRPGPETVELTTEEAYSGRYSKLVSGRTSTWNGPMVDKTDVLTGSEY KLGYYVKFVGDSYSNEQRFLQLQYNDGAGDVYQNIKTATVYKGTWTLLEGQLTVP SHAKDVKIYV ETEFKNSPSPQDLMDFYIDDFTAT
	A26	Cthe_0912	CBM22-GH10- <b>CBM22</b> -DOC1-CE1	PDANGYYYHDTFEGSVGQWTARGPAEVLVSGRTAYKGSSELLVRNRATAAWNGAQRALNPRTFVP GNTYCFSVVASFIEGASSTTFCMKLQYVDGSGTQRYDTIDMKTVPNQWVHLYNPQYRIPSDATDM YVYVETADDTINFYIDEAIGAVAGTVIEGPAPQPTQ
25	A80	Cthe_0956	<b>CBM25</b>	FRLVYSGILAKNPNENLYAVIGYGNLAWEDIESYSMRKIGDQKYELLFPVKRPGNINIAFKDDADNW DNNSGMNYCFENHVYQGS
30	A17	Cthe_0624	<b>CBM30</b> -GH9-GH44-DOC1-CBM44	SAETVAPEGYRKLDDVQIFKDSVPVWWSGSGMGELETIGDTPVDTTVTYNGLPTRLNVTTVQSG WWISLLTLRGWNTHDLSQYVENGYLEFDIKGKEGGEDFVIGFRDKVYERYVGLEIDVTTVISNYVT TDWQHVKIPLRDLMKINNGFDPSSVTCVFSKRYADPFTVWFSDIKITSEDNEK
32	A28	Cthe_0821	GH5- <b>CBM32</b> -DOC1	AGSIAQNKPVYASSTEPGLGNTPEKAVDGNIA TRWSSDYSDNQYIYVDDLDEYEIERVYIEWEAYA RQYKIQVSNDAVTDVYTEYNGDGDIDDIYLEARGRYVRIYCMQRATQYGNISIFELGVYPKGGIA
35	A66	Cthe_2811	<b>CBM35</b> -GH26-DOC1	INVSNAVLSDGDKYEFEDGIHKGAQIYTDYVQNEYGEVFDLTGSTCSFIAQKGTSTSVNVEVDKEG LYEIFICYVQPYDKNKKVQYLVNNGVNVQGEISFPFTLKWREISAGIVKLNAGINNIELESYWGTYFDY LIVKP
42	A29	Cthe_1273	<b>CBM42</b>	YGQFMKFESSNYRGGYIRVKSFSGRIDPVVNPVEDSMFKIVPGLADPSCISFESKTYPGYYLKHENF RVILKKEYEDTDLFREDATFRVVPGWADENMISFQSYNYPYRIRHRDFELYIENIKTDLDRKDATFIGI KVD
	A43	Cthe_2139	GH30- <b>CBM42</b> -GH43-DOC1	VPAVGLQSYNYPNRYVRHADFDARIDENVTPLEDSQWRLVPLANSSEGYVSIQSVNYPGYLRH WDYDFRLDKNDGTTIFAEDATFKLVPLADPSCVFSQSYNYPDRYIRHYGYLLKLERISTDLDRQDA TFLII
	A44	Cthe_2138	<b>CBM42</b> -GH43-DOC1	STGADGAIKLSYNSHMYIRANFDRVIDNVTPETDAQWVLPGLANSSEGYVSIQSVVDHLGY YLRHWNYDFRLEKNDGTRIFAEDATFKMVPGLADPSYTSFQSYNYPTRYIRHYNYLLRLDEIVTALD REDATFRVIDSSV
50	A73	Cthe_0300	<b>CBM50</b>	YTVKPGDTMWKIAVKYQIGISEIIAANPQIKPNLIYPGQKINIPNI
	A74	Cthe_2387	<b>CBM50</b>	YAIYVVRPGDTLYNIAGRFTSVNSIVTANPGINPNVINIGQRLVVPYGI
	A77	Cthe_3006	CBM50- <b>CBM50</b> -CBM50	YEIKSGDTLSKIAAKFNNTVGDILNANPGIPEKLYVGGKICIPQP
	A84	Cthe_3006	CBM50-CBM50- <b>CBM50</b>	YVIQKGDTPAIKIFNVTVQQLINANPGINPNALYVGVVICIPVA

R. flavefaciens FD-1		A35	Cthe_1800	<b>CBM50</b> -CBM50-GH18	MWYTVQPGDSLTYTISRFRGVTIAQIKSANQLTSDIIVGQRLYIPIGIQA
		A86	Cthe_3005	<b>CBM50</b> -CBM50-CBM50-CBM50	YKVQSGDTFWKIGQKYNISTAALLKANNANENTVLYPGQTIVLPIK
		A83	Cthe_3007	<b>CBM50</b> -CBM50	YTIKAGDTLAAIARIYGTTVQDIINANPDIDPYLVRVGGQICIPLT
	<b>62</b>	A45	Cthe_2193	GH5-CBM6-CBM13- <b>CBM62</b> -DOC1	PKLTGTVIGTQGSWNNIGNTIHKAFDGDNLNFFDGPANGCWLGLDFGEGVNRNVIQIKFCPRSGYE QRMIGGIFQGANKEDFSDAVTLFTITSLPGSGTLTSVDVDNPTGFRYVRYLSPDGSNGNIAELQFFGT PAGEEN
	<b>4</b>	B22	3995	<b>CBM4</b> -GH9	VGLPWHIVESAPGVMDFSIDGGTYNVTVNPGGASRGGEDRWDCQFRHRGLKIVSGHQYEVKYDIT ATESGMYTKIGNLDGDVLEWHNNMADNGPDFNGSWDLIHIDANKTNSVSLTFTANQNMEVAEWA FHLGGSGQYTPQDCPPEGTVISFDNMS
		B25	3259	GH16- <b>CBM4</b>	GDDFAPTPVTSMLGYSIEGAEAYVANKDGTCLVHIDSVGSLEYGVMLLRGQKVQAGDWTWQLEFD AVSTAEREMTVTAEDSSYTRYLDEKVTVSSEKHKHFSFDVNFAGDMSADIKFQLGNIGNAASVGSHE VTLNLIKWTCKNGS
		B40	776	<b>CBM4</b> -GH9-DOC1	QVSAAGNLISNSTFESGVKDWGTYKESGGKCSLKAEDGKLALTVSDVGKVNAYVQVFDILPLYQN GVYRLKYDISCTTDRFVEGMIQMNGGDYRAYTWKGLNLTSAPQTVDYEFTEDETDIMAKLVFNCG IQEKYEGVLPHEHTIYIDNVSLLELVDDS
		B45	2836	GH16- <b>CBM4</b>	EALDGNFVYNGDFAEAEDLTDDENWKFLLEGGKGAEEIRDNMIVITTEDEGTVDYSVQLVQPEM PIIKGKKYRVTFDAWADEERDIIVCVSAPNAGWIRYLEDTTLTITPEQTTYTYDFEMNDKDDPLGRLEF NMGHKGSTATVYITNVRLEEVE
	<b>6</b>	B49	3747	<b>CBM6</b> -DOC1	GSGGSTDDIIEAEKYDIQKGIQTENCSEGGSDVAYIENGDIYGFKNIDFGSGTDSISFRIGSNGAEASI EVRLGAADGKLIPTLPVKSTGGWQTWNTQTCAIENTSGRNDVYFVKGGDGYLFNINWWKPKPS EPI
	<b>13</b>	B18	2326	GH43- <b>CBM13</b> -DOC1	VTETVRIEKGKYTLKNVNSGLYIAEDSGNVIQSQSQPWEIKAVADGVYAIIDEKGNALTVDGSSPDDG ANISVSTFSDAENQKF TAVLNDDGSYSFMSLVSGKVRCLDVYNISKDDGANICQWEFVGGDGQKFI LEEA
		B43	2115	GH43- <b>CBM13</b> -CBM13-DOC1	SGTELLSGVPYFITNVNSGLSLDLPEGKLDNGTNIQQWDFNKLWAQQWRISVDKEWCRIVSLGDEG KCIYAKDTADDGTVNELQTYTGADNQLFKFVKSGSSYGVSKCSGGKALDVFESKENGNNV QFAYNEYACQLWNIAPV
		B60	694	PL11- <b>CBM13</b> -DOC1	TASDIIDGQIYTFKNLNSGLYLDVEGGTAANGTNVQQAETGKQKQKAVAAGDGYLVSQGLDGD ESYALDVNAKKTADGTNIEIYTFNKGENQKFRFQKNDGTYSILTKITDGKSALDVNEQSGNSGANIQ QYTFSGSANQKFIIEAV
	<b>22</b>	B19	1615	<b>CBM22</b> -DOC1-CE1	VPVSAADNDYMLHSTFEEGKDSWSGRGSASVKTVSGKSRSGQQSLYTSGRESWNGATLKLGS FKAGSDYSFSAAYVMTEDDDVSFCLTLQYKDGSGTAIYPKIAKVS GKKNWAHLENNSFSIPEDASD IELYVETEESKCSFYLDVVGAAVGTIEAEPKG
		B21	1272	GH30- <b>CBM22</b> -DOC1	PDSNGYYYHDTFENGTDNWEARGASELTLGRRPYKDTNGLLVQNREKAWNGVQKSLDSNTFKG GNSYSFVAATMLEDTSANVFLSLQYTDTSGETKYAHIASAQNGEYVHLANPNYKLPDGSYVLYI ETEETDNFYIDEAIVAKAGT
		B24	3180	GH11- <b>CBM22</b> -DOC1-GH11-CE1	KAVEPDANGYYFNDFESGKGSWRGRGEASAAIDNDNSAEGKSSLFVSGRTDNWNGAEMELDPA AFIPGKTYSFGAAMQNTESSTAMKMTLQYTDASGTEQYDEVASAAASNGKWTALGNPSYTIPEGA SNMYLYVEAPESLTDIFYIDNVMAAVKGEATFKN
		B48	3190	<b>CBM22</b> -GH10-DOC1	SVINTVNAAEKVVYDLGFESDDLKNWSNRGGDDTTELSITTDAKTGDGALLASGRSESWNGPAF RLDGVLEPNTQYYVTASVKGYTASMLSFQYTDIDGQTSYSNLALQNLNGSDWQTVTHVPVSYSDG MEGVYIYFEGGSDDLIDDFKIVEA
		B51	1766	GH11- <b>CBM22</b> -GH10-DOC1-GH11-CE4	PTKQADANGYYFNSSFNSGVDGWTGRGAATVAKDSSNYAEGNGSIFVSGRTDNWNGAAIELDPSA FGAGQTYSFGAAMQKSESSTSMKMTLQYTDASGTEQYDEVATATASNGKWTALGNPSYTIIPSGA SNLLYIEAPDSLDFYVDSAFGGVKG

	B8	2646	GH43- <b>CBM22</b> -DOC1-CE1	PDPNGYLMHSTFEGKTDGWSGRGAASVESTGTEHFEGSSSVYVSGRTASWNGVTHALGSKIKSGT EYSFSTNVKYTDGPDEQLFFFTLQYEDTDGEVKYDKIAKGYIRKGEWAQLANTNYMLPAGATNMQIY VETEEDGDFYIDDTIVAEAGRLIEGAAPAESG	
	B28	2002	GH30- <b>CBM22</b> -DOC1	DYLLHDTFETSADSWEGRAASVSRSGGTLFCEGRTASWNGAAKNLSTDFVPGKEYSFSVNHAMH NGTGTETFKLTLQYNDASGTANYPNIAQATASAGEWVTLKNENFLIPADATDLILYVETDDSMDFYI DEAVAAKGGTS	
	B35	2649	GH43-CBM6- <b>CBM22</b> -DOC1-CE1	ADSGNYIFNDTFESGDNDWSSRGSAAKVSSSDKKYMGSKALYCSGREASWNGALKDLGTSFKAGE SYSFSANVLSDDGGKGDVYYL TMEYKDDSEVHYVKIARSQPVKGEWVQLANSFRIPADAASDIHI YVETEKSTASFYVDDVKAAGTVIEGAKG	
	B1	3077	GH11- <b>CBM22</b> -GH10-DOC1- CBM22-CE4	QTVKADSNNGYFNFESFESGAGDWEGRGAAKVSKDTANYAEGKSSLYVSGRTDNWNGAAIQLDSS AFVAGNTYSFGAAVMQNTESSTAMKMTLQYTDADGKEQYDEVATATASNGKWTALSNPSYTIPTG ATGLLLYIEAPDSLDFYVDSAMAGVKGKEVTVSGG	
	B2	3077	GH11-CBM22-GH10-DOC1- <b>CBM22</b> -CE4	ATTQTPAASNKTYIAADFGSSSNFESRGGASVELNKSTYYSAPSSLYVTGRDNWHGASIALGSDF VPGNTYSFSAAVLQTSGSADTVKMTLQYKDADGTEQYDEVASVKADSKTWTDLTNEKYTIPAGATD LLLYVEMPDSLADFFVDDVTVAASGT	
	B37	1737	CBM22-GH10- <b>CBM22</b> -DOC1- GH43-CBM6	PDADGYWFHSTFEGSDGGWGGGRGSASVTTSGRTFYKGAEALLVQDREAAWNGASYPLSSRIFKP GEEYSFSVNVQFLDGDSDAEYKFTLQYQGSDDGEAHYDQIAVGTAPKGEWLQLANNTYKIPADATDC QIYVETTDNTGNFYIDEAIGAPAGTAIDGPGQPKV	
	B57	1878	CBM22-GH10- <b>CBM22</b> -DOC1	PDENGWYFHSTFEDGTDGWSARGSAEILVSGRKGFEQPQLLVRERTSSWHGASYALDTRAFPLG NEYSFSTNVTYFDGDDGDKFYLLKLYTDSEGKARYSTIAEGTGIKEQWVQLENTAYKIPDGASMSI YVETADTANNFYIDETIGAVAGTVINGAGQPEI	
	B32	3270	<b>CBM22</b> -CBM22	IIEYIPFNVTDSGKIDGWDMRGDKGTGFGVKGWNDPYAGATNIYISGRSQDWQGAKEYELATDKYS AGHSYSFGIFARNEGQDAKFTMTLEYFNGSKTDYTIASATLKPGEWTEIKNPNFTIPVGATKCCVA IETPGSKPNFRIDEFVSAQPNT	
	35	B17	1364	CBM35-CE3-DOC1- <b>CBM35</b> -GH26	FLAVYEAE NAVISGNIASVDDSSASGGKAVGSFSDDGDDLAFTIEVPAAGSYCFTLTSKGMGGDKYN EVLVDGENIGGFESKGNVYSETSLRRVMLTAGKHTVSIKKSXGWIMVDSLKVTTDDVISNSVYNVEN KLINSN
		B29	933	<b>CBM35</b> -GH26-DOC1	DANKYEFEDAEFTGDVTVEEDANASGGSMLKMTDSGTITLKVNVETAGSYKLTIFYALGIGGDKQQN LTVNGDSQGAIGIPKSSEYEEISVPAIMLKAGENTITIEKSWGWSQFDYMTVTSMADAKITATQTK
B58		2302	<b>CBM35</b> -GH26	YEFEEGTISNSGENEAEIISVKGASAGQAVDLKDGNTVTVKVNAAESGMHRITLRYCQPYDEDGKY QNVIVNGKNAGEIFCEYTGDEQFSTVSISAVLNQGENDIAVEASWGWMTMIDSLLEIKGDFSAYT	

<sup>a</sup>DOC, dockerin; GH, glycoside hydrolases; CBM, carbohydrate binding module; CE, carbohydrate esterase; PL, polysaccharide lyases; SLH, S-layer homology domain.

**Table S3.2. Modular architecture and primary sequences of *C. thermocellum* and *R. flavefaciens* FD-1 CBMs that were cloned but that have not expressed or for which no binding was detected in the microarray analysis.** CBMs investigated are highlighted in bold and the respective sequence of the recombinant protein expressed is shown. CBM clones that have not expressed are shaded in grey. Each CBM is identified by organism, high-throughput ID code, and assigned protein ID.

	Family	Code ID	Protein ID	Molecular architecture <sup>a</sup>	Primary sequence	
<i>C. thermocellum</i>	3	A4	Cthe_0271	<b>CBM3</b>	EKKGPIITVQYKNGDSTSSVTAIYPIFKITNNGDTSVKLSDIIIRYYYTKEGNETFWCNEFTRDGSQ VYGTFFVKMSKPKENADHYLEIGFYDKAGSLKPGESVELKVGFAKNGWTKYNQFNDYSYNRVNNRFI NWDHITVYLSGKLVYGKEP	
		A5	Cthe_0043	GH9- <b>CBM3</b> -DOC1	IVEYFCRGWIIYEGYGTLNLLLQVNNRSGWPPTMKDKLSVRYFMDLTVFESGGTVDDVQISLGQN EGAKLIGLKHYRDNIIYFTVDFTGTMIMPAEWEMCEKDAHVTIKYRDGITGSNENDWSYQNLKDP DYDATSFAGLTPYIPVYDNGVLLWGEEP	
		A9	Cthe_0267	<b>CBM3</b>	DLLTKIELQAYNHIRTSETKELQPRIKLINTGNTPTLSEVKIRYYYYTKDQVINEIYTCDSNITSSKITGT VVQMSNPKPNADSYVEIGFTNSAGVLNPGEYVEIISRIGNSYALSLATPPYSEWNYMYDQNSDYSFN NSSSDFVVDKITYYISGTLVWGIEP	
		A13	Cthe_0413	CBM4-GH9- <b>CBM3</b> -DOC1	DVKVQYLCENTQTSTQEIKGKFNIVNTGNRDYSLKDIVLRYYYFTKEHNSQLQFICYTPIGSGNLIPIPF GGSGDEHYLQLEFKDVKLPAGGQTGEIQFVIRYADNSFHDQSDNDYSFDPTIKAFQDYGKVTLYKNG ELVWGTPP	
		A7	Cthe_0040	GH9- <b>CBM3</b> -CBM3		
		A16	Cthe_0404	<b>CBM3</b>	DGEQSVKRVFYNNNTLSETGVIMRINVINTGNAPLDLSDLKRYYYTIDSESEQRFNCDWSSIGAH NVTGSFGKVNPSRNGADTYVEIGFTKEAGMLQPGESVELNARFSKTDNTQYNKADDYSFNSHYYE YVDWDRITAYISGILKWGREP	
		A18	Cthe_0578	GH9- <b>CBM3</b> -DOC1	NEEIYVEATANSNNGVELKTYLYNKSGWPARVCDKLSFRYFMDLTVYSAGYNPNDITVSIYSAAPT AKISKPILYDASKNIYYCEIDLSGTKIFPGSNSDHQKETQFRIQPPAGAPWDNTNDFSQGIKKNGEV VKEMPVYEDGILIFGVEP	
		A22	Cthe_0745	GH9- <b>CBM3</b> -DOC1	EDEFMVEAYVSSDKNYVEIKTRLNRTAWPARVSEGLSFRYFIDLTEVIEAGYGPNDLIISGGQGSS GKVS GPHLWNKEKNIIYIEVDYTGDRFLPGGQDHYRRDSSLRIAVPGNSGCWNSNDPSPFKGLSK TSEFKKAEYIPVYEGVKVAGIEP	
		A23	Cthe_0625	GH9- <b>CBM3</b> -DOC1	DDEFFVEAAINQASDHFTTEIKALLNRRSSWPARLIKDLSYNYMDLTVFEAGYSVDDIKVTIGYCES GMDVEISPITHLYDNIYIKISYIDGTNICPIGQEYAAELQFRIAAPQGTKFWDPNTDFSQGLTRELA KTKYMPVFDGATKIFGEVP	
		A25	Cthe_1257	<b>CBM3</b> -CBM4	NSANISLEFYNGDFGASVSSISMFRITNNGSSQISLSDIKLRYYYTDDGVSPITVFIDYANNNGRGIN NDVTYTIKINSSGANKYIEFGFNAQAGSLEPNTSVLMRARAYQSEYKQSFQTNDYSFCQSNDF AAWNKVTGYLNGVLFS	
		A42	Cthe_2147	<b>CBM3</b> -GH5-DOC1	GLSIHYMDGTLVDKYQSMRPYIIHHNNSGMDVDMADLRVRYYYEKEGVTEEVLTCFYTAIGADKIFAE FHPELGYAEIGFTSDAGIISGGNSGQLQLVLKKSINGYDQSDNDYSYDPSYTDYAEYDKITLYYKKG LVWGKEG	
		A54	Cthe_2506	<b>CBM3</b> -SLH-SLH-SLH	VKAEVPLKLEFFNNVKDDNVTLISPYFRVINNSSSDEIYLQHVKIRYYFTLDSSDSEETMNYEIIYAG KSNIDGTGAVEDIKPNITVIAKMDIPTDMADHYLEIGFDESCGTIGPDKKVEVMVVISKEYKFKFIQTN DYSYNDSAENYVSEWKTLYLDGELISGIEP	
		A55	Cthe_2360	GH9- <b>CBM3</b> - <b>CBM3</b> -DOC1	KPSLEVLYKYGDTTAATKDIRGSIKIKNTGTPVNLSDVKVRYWFTKDGASSQEFVCDYAHLSSEMIT AKFVDLENPVENADNYLEIGFDSNAGILGPGSDTGEIQFRIVKGDYESYDQSDNDYSCMATAKDFTAN PNITAYVNSVLVYGNPPVD	



	A58	Cthe_2761	GH9-CBM3-DOC1	PPVYYADAKIYEENESGITVDLNMYNIVTSPQQYESDLSCRYFVDLSEYAGENIDMSKFVTKVYYSPA GATISELKPYPDKENIYYVEISFPNPVYARTYVQFCIYYENKLWSSNDFSYQGIGDXYTKLENIPIY KNGVLVAGKEP
	A60	Cthe_2760	GH9-CBM3-CBM3-DOC1	TDDYFCEAKIVRETKDSTQVLLRIHNESTRPPHYETGMMARYFFNISELIENGQSIDDVIFTIEYDEQIS MQQEPVVYRGPFWDDAGTYFDFWDSGRKIYGDRELQISFRVKQDSNYMTHWDSNDYSRQGL TNEYAISKNPVYLVNGVKVYGEEP
	A56	Cthe_2360	GH9-CBM3-CBM3-DOC1	VEEYVVEGKIEQENKERTQVTIKIFNDTCHPPRFETGLMARYFFNISELLDAGQSIDDVVKIEVYYDENK ASYDGAPEVVRGPIKYDDAGTYVEVDWDSGRIIYKREIQLALISSLDSNYKSNWNPENDYSREGLGK EFVRTEKIPLYLNGVKVFGNEP
	A59	Cthe_2760	GH9-CBM3-CBM3-DOC1	DANASISVSYKCGVKDGTNTIRATINIKNTGTPVNLSDIKVRYWFTSDGNEQNNFVCDYAAFSTD KVKGIVKKIENSVPADTYCEISFTEDAGRLAPGGSTGTIPFRIEGAAEYDQTDYDYSNSEMDDDFG DNTKITAYIKDKLKYGVEPVT
	A11	Cthe_0071	GH48-CBM3	NNTVGRLLIQYANGNGSDTTNTINPRFKLINNSGSPVKLSKVIRYYTIDGEGKQQFWCDWSSAGN SNVTGKFKLSSPKNNADYYLEIGFTEGAGSIEPGMSVEVQARFSKDDWSNYSQANDYSFSASAN DYGNSNHIALYISGRLVSGNEP
	A19	Cthe_0543	GH9-CBM3-DOC1	GEEFYVEAAVNAAGPGFVNIKASIINKSGWPARGSDKLSAKYFVDISEAVAKGITLDQITVQSTTNGG AKVSQLLPWDPDNHYYNIDFTGINIFPGGNEYKRDVYFTITAPYGEWNWNTNDFSFQGLEQGF TSKKTEYIPLYDGNVRVWGKVP
	A88	Cthe_2423	CBM3	VYVTLKNIKTGVPSTIALKIGIINLNKAINLNLDIKLRYFTNDGCSPIQVNIKLFGTETESFNPELVKT SVVTGLSYPGADSYVEIGFTGSVELNCDRKPPIYIELDIKENSPDRNFDQSNDFSNNNYYTPFLPEEFF ASGRVPVFMYPDPKKR
4	A32	Cthe_1257	CBM3-CBM4	VPNGDFEGSGVFWFSFYCDLSGANATNLHSEPSGNKMSKTSITNAGSNHWAIQLKHDGIVLENLKT YRLTFDAKSTVPRNIRVSLQANATSSMIEYFGKIVEVEPKMKTYTCEFTFNSTGTNVAIVFEMGKIGT ETDKAHDIVLDNVHIEKIASPS
	A15	Cthe_0412	CBM4-GH9-DOC1	GEPGNKAFRLTVIDKGQNKWSVQMRHRGITLEQGHYTVRFTIWSDKSCRVYAKIGQMGEPYTEY WNNNWNPFNLTPGQKLTVEQNFTMNYPTDDTCEFTFHLGGELAAGTPYYVYLDVSLYDP
	A61	Cthe_2809	SLH-SLH-SLH-CBM54-GH16- CBM4-CBM4-CBM4-CBM4	ILNGTFDDGMDHLLYWGDEGEGNCDVTDGELEINIKVGTADYMPQIKQENIALQEGVYTYLSLKR ALEARSIKVDILDSSYNWYGGTIFDLTTEDAVYFTFTQSKSINNGVLTINLGTIEGKTSAAATVYLDL LLEQQ
	A63	Cthe_2809	SLH-SLH-SLH-CBM54-GH16- CBM4-CBM4-CBM4-CBM4	IYNGTFDQGNRMGFVNFVVDSTAKATYYIGSDVNERRFETRIEKGTSRGAIRLVQPGINIENGKT YKVSFEASAANTRTIEVEIASNLHNSIFATTFEISKESKIYEFFTMDKSDKNGELRFNLGSSNVV YIDNVVMKRVSTDEVE
6	A40	Cthe_1911	CBM6	TGTINALSVIQAENCENHGLEIEDCPDEGGTKNLAYIANGDYTAYYVYFPKGTGFIARVSSDTEG GYIELRLDSISAEVVGRCRVENTGGWEKYEYVYCELNKSVEGVHTLYMGFAGERDGLFNVNWFRRF TKSPYEPVM
	A47	Cthe_2193	GH5-CBM6-CBM13-CBM62-DOC1	DFDWGGNGVSYDDTDSVNVGGQYRPDEGV DIEKTSDTGGGYNVGWISEGEWLEYTIRVRNPGYY NLSLRVAGISGSRVQVSGFNQDKTGVWELPATGGFQWTWTTATRQVFLGAGLQKLRINALSGGFNL NWIELSP
	A50	Cthe_2196	GH43-CBM6-DOC1	IGTLNPYVRTEAETICWSSGIETEKCEGGMNVGFIENGDIYKVKGNVFGTGAASFARVASATNGG NIEIRLDSPTGKLVGTCTVTGTGGWQTWTTKSCPVSAGVHDLVYFVFKGGSGYLFNIDWWKFTPA NPDPTPTM
	A31	Cthe_1271	GH43-CBM6-CBM6-DOC1	LKYVDPYTKNLAVTMHKEGIEETECCSEGGRNVAFIENGDWIQVKGVDGNGVPTSFEARVASATN GGNIEIRLDSPTGLTIGTCKVEGTGDWQKWVTKCSVSKVTGVHDLFFRFTGGSGYLFNFSWWKF NSDA

13	A46	Cthe_2193	GH5-CBM6-CBM13-CBM62-DOC1	STGTIPDGTYKFLNRANGKTLQEVTTGNNISITADYKITEQHWKIQHIGGGQYRISSAGRGWNWNW WMGFGTVGWWGTGSSTCFIISPTGDGYRIVLVGDGTNLQISSGDPKIEGKAFHGGANQQWAILP VSAP	
	16	A75	Cthe_3095	CBM16-GT39	NLVKNPGFEEGNDESIVYVWQTHCWEKAEGVTEFFIDESVYHSGGKSACIVNHSENDSRYMQPIKV KGDYYRLSCWVKTENVTGKTKGANISIEGSLDTSRDIRETSDNWEYLEYKGTSPNQETFLLTIGL GGYGNTNTGKIWIDDVEVEL
		A76	Cthe_2148	CBM16	NLLKNPSFEEVDNMMPLGWSTWVWVNYQNGVVEFKVEQEGAQSGQYVVTIENREARDARYLQEV VSPNSYYKLSGWIKTENVGNDVLGANLSLEGVTTYSKDIRGTVDWEVQYTELYIKTGENVETIKVSLG LGGYGNLNTGKASFDNVMLEKV
		A87	Cthe_2805	CBM16	LPKTVFTEDFENGLSSNWEIRSSKHGNSLATVTVESGTGVNNSKCLKISSLAQDEVDGCVKTLQLAP NSYYKLSALMKYENVTPGKSDGANICLYNNEGDAWIRTATATGTNTSWELVKLLFKTPDSGSVNI GLRLGFLNCETKGTVWFDNVKVEAV
		A69	Cthe_3096	CBM16	NYIFNGSFELLIDGEPNSNWMREAYDKSPGASNFRVETEGAKFGEKYVTIINKNLNSDRYSQIVLVEEN KKYKLSYIKTENVSEEKGANLSVAEQTVTSKRKIGTTDDWEVYVELYVITESGVDRIKVTVGLGGY SGMSTGKASFDNVTMEE
	25	A70	Cthe_1080	CBM25	GENLTVMYDGLLSKSGASHVYAHVGFDRDVKHVVYDYPMKRTSIGFEATIPVMEADTLNICFKDCAN NWDNNSGANVYTFDISK
		A71	Cthe_3163	CBM25	GDEVTLYYKGLLAQSGADAIFAHIGYGENWEDKTFIPMQKENDVFKATIKINHADDLNIAPKDSGDN WDNNSWANYSFKVTKKAKPAKV
	34	A21	Cthe_0795	CBM34-GH13	MKLEAIYHKPYSEFAFPVAPDTLVIRLRTAKNDVNTCILYHEKYDTSQRGKVKMDKVASDGMFDYY EVELNVGKRIKYMFLYEDNYSIKWYSSDGFDDYMPQWGHFTYS
	35	A8	Cthe_0032	CBM35-GH26-DOC1	AYSLPVDVEAEDCTLNGAVVTTNVYGTQYPGYSGDGFVWVANSQTITLEVTIPENGMVELSTRC WMYLGKEDETRMQVISINGKSHSNYFIPNKGQWIDYSFGFFYLEAGKATIEIGSSGSWGFIYDKIYF D
		A38	Cthe_2137	GH39-CBM35-CBM35-DOC1	RYEAEYARILGTATVSHGGHSGYSGTGFEVGYAGSNNASTNFVVAETDGYNVTLRYSAGPYPG APKTRYLRMVVNGGLHKDVACIQATANWDTWESTTVKVLQAGINRLDFKAFASDESDCVNIDYIDVE PT
		A67	Cthe_2950	PL1-DOC1-CBM35	NGTATYQAEDAVFSGAIFETKNAGYTGTYVNYDNVPGGYIEWTLNANAGTYTLTYANGTSSNR TVDISVNGNIVASGVVFGGTGSWTQWQTKSITASLNSGVNKRIVTGTSSDGGPNIDKLEIRRN
		A72	Cthe_3141	CE12-DOC1-CBM35-CE12	VIIQAEDAIINYNAILETVNAGYTGSCYVNYHNEVGGYIEWNVNAPSSGSYALIFRYANGTTANRPMRI TVNGNIVKPSMDFVSTGAWTTWNEAGIVANLNQGNVIRATAIASDGGPNVDYLVFSANAFQPV S
		A10	Cthe_0246	DOC1-CBM35-PL11	QTQKTRYQAEDAMLYKAFEETIHAGYDGRSYVNYDNEPGGYIEWNVNVSSTGYKLIFRYANGSNN NRPMEIRVNSNLVAGSLDFYPTSAWTVWNDQSIWVTLNAGNNVIRATGIASDG
A37		Cthe_2137	GH39-CBM35-CBM35-DOC1	TLGGAAVRQRDAAASGGQYVWIGNGSNNYLQFNNVVYPQAGTYRMVVFANAEVFGQHSYNN NVVDRYCSISVNGGPEKGYHFFNTRGWNTYRTDIIIDVYLNAGNNTIRFYNGTSGSYAPNIDKIAIAP FEGGTEPT	
A41	Cthe_2179	PL1-DOC1-CBM35-PL9	DTHPSSTPTEGVVHEAESSNHLKYAKVESNYVVDQTKDAYIEMKKNVSPVTGEVTITIVSNGSG KSLPMEIKVNSTTIESNKEFPSTGAWNIWSTLSVKANMNSGSDNVIRIKTRSNDGGPRIDKVIIVSAG		
42	A1	Cthe_0015	CBM42-DOC1-GH43	TNPITKAKFQSYNPNMYIRHANFARIDENVTPEMDSQWELVPGLANSQDGYVSIQSVNYPGYL RHSNYDLSEKNDGTSLFAESATFKIVPGLADPSYISFQSYNFPTRYIRHYNLLRLDEIVTELDRQDA TFKIISEDQ	
44	A24	Cthe_0624	CBM30-GH9-GH44-DOC1-CBM44	KTDDWNEITLTPEDVDPTWPPQMGIQVQTIDEGEFTIYVDAIDW	
48	A48	Cthe_2191	CBM48-GH13	RMTIDGVEGTLFAVWAPCAKRVSVVGNFNQWDGRRHQMRVRGSSGVWELFIPGVGEGELYKYEI KTPHNEIYKADPYAFYSELRPNTASIVYDIEG	
50	A34	Cthe_1800	CBM50-CBM50-GH18	PVVYTVRPGDLYLIARRYNTTVDLSMALNLSSTELRIGQQLTIPLYTE	

R. flavefaciens FD-1		A82	Cthe_3007	CBM50- <b>CBM50</b>	YYVVRPEDTLESIAAYFNITPQQLLYSNYGIDPTDLYVDQILCIPVA
		A78	Cthe_3006	<b>CBM50</b> -CBM50-CBM50	YTVRAGDTLYLIAGRFRNTTVEAILAANPGIVPERLYIGQVICVPIYA
		A85	Cthe_3005	CBM50- <b>CBM50</b> -CBM50-CBM50	YIVQSGDITYWNISQKYGINFKELLALNNANENSMLNVGDKVILPAT
		A90	Cthe_3005	CBM50-CBM50- <b>CBM50</b> -CBM50	NYTVQKGDYWTISQKFKVNFTELLKLNGANESYLDIGQVIKIPVT
		A89	Cthe_3005	CBM50-CBM50-CBM50- <b>CBM50</b>	YTVQKGDTAWSIAEKFGISMYELMEANNINSSTVLNIGQKLIKIPVH
		A79	Cthe_2489	CBM50	YHVVPQGDTLWGIACKYYGNGNQYQKIYEANKNQIKNPNIYPGQKLVIPR
		A81	Cthe_1611	<b>CBM50</b>	NTHIGFEICEPAGFSYKSGSVMVGYDAAKQEDYFFKAWQNAVELCVML
	54	A57	Cthe_2809	SLH-SLH-SLH- <b>CBM54</b> -GH16-CBM4-CBM4-CBM4-CBM4	YKNEEVAGNALINTEGVILKDTVINGDLVLAQGIQNGDVTLDGVNVKGTVFVNGGGSDSIHFINTKINRVVVNKTGVRIVTSNGTSSVESVVVKSAGAKLEEKELTGDGFKNVTVDSQLSAGNEIIFVGDFEQVDVLADDALLETKEAK
	3	B52	929	GH9- <b>CBM3</b>	FWAAGYQCESPEDTGAGVTKLTFVNTDCCLEPHTDLSIRYFFDISEFEKNTDIPGSFVLQKTYDQVE TEVTDRAATLSKPIKYKDNVYVEIAWPDYAVANSNKYQFIIGMYGDKWDSSNDWSRKGKIKELDGDYDNIVGGVELAEKCENVCVYADGKLVGGTEP
		B20	938	GH9- <b>CBM3</b> -DOC1	GPEFYVECTSKGAESSGMTISFKITNHSAPARVQDNISFRYYMDLSEVKAAGANPEDVVRCDRDQSKMYAGVTPAEISGVKHYDGDIIYVEVTLPDGRAVLPISEGMQQCEILLALVMPNYGSGWDATNDFSNKEILGAKTTTTADGVSVHGIIPTVYVYVNGKLYYGEEP
		B9	2908	GH9- <b>CBM3</b> -DOC1	PTGLYISGGKNQEQTGSVQLKVVVHNRTVNPVKFESDMKARYYFNIKELLDKGYDPKEYIFARIDYDQEKSFSNGKNEAKFTGPTKYDDNGTYVEMQWKDCDFYGSRVYQFALGYNQDKTTYEDVVWDSKNDYSYADLVSFEDDAAASAITKITLYCDDKLWGVPEP
		B12	2914	GH9- <b>CBM3</b> -DOC1	YWVEACGIDSRNDDGTGAVEVSLKVLSGETTPSKNLTIRYFIDASEVSDPSIIDTKKLYDQSEMEIEGAKCTVSPLKYYKDSDSIYVELSWEDCTIVNSGKKSQFSVGFYKGYTDPETHKYIVYKWDPENDWSYSHMKLVKEDFFAVDDPPEERCYICVYDDGVLVGGIEP
		B26	2994	GH9- <b>CBM3</b> -DOC1	WPEWEVAAVINGTEGNTYEVKAWAMNHTAWPARVAKDVEYKYFFDVSVDVLAAGLSIDDIKVEGKSQQYKEGEQYATVSGPYKYEGDATGNTYYALIKFEDGRAIQTGQSEHRDEVQFRISIPDAVDGQAVPAGAWDTSNDWSYLGGLAKATDLKADSINEHIPMYVNGELAWGEEPDKTFVAKPNTKDGKGSTDPQPTPSVTTTTSTATTSSATTATTSSATVQSTEGTTTTSGQGGNSSERVTLWGDANCDKAVDVSDAVIIMQSIANPSKYKLTDEGKANGDVNKNKGDGITGADALSIQKYKLNLTPELPEYN
4	B14	2995	<b>CBM4</b> -GH9-DOC1	TALPWHTCESQPAGQHFKEGGKYKITIDENNGPAGRWDLQFRHRGITMIQGHEYTISGDITATEDGYIYAKIGNYEGNKEYWHNLSGQEWKPYQIKAGEEFHFEDFTFLKDSVPVGPTEWAFHYSDNHGQYGNNDTGMPKGAULTFDN	
	B46	2283	<b>CBM4</b> -GH9-DOC1	IILPWRLVESQPAAGQFVYDGNALKVTVYYPEGANDRSDLQLRARGLIQAGHEYTIVSGTIKTDADGYIYSRIGNYIGNTDCWHALGGAEWMPVQMEANEPFEFSQFTTATENIEAAEWAFYYADNRGMYGNPDTGMPAGSQIWFSDL	
	B47	1485	<b>CBM4</b> -GH9-DOC1	VLPWVPVSSQPAMQDFCVQDGRLEIKLNNRGPGRWDLQLRRRGLTMIQGHEYTVKCTITADDDGYIYSKIGNYTGEKEYWHNLSGQEWMPYHITKGETYEFEDTFVLKDSVPVGPTEWFSFMYADNQGMYNNDTGMPDGGSTITFDLLE	
6	B34	2649	GH43- <b>CBM6</b> -CBM22-DOC1-CE1	LELLNPHYERVAETICWSEGIKTEECSSAGGVDIGNIEKGDYIKVSGVDFGSGAAKFTASVASDTEGGTIELHTGSKSGPIIGALQVKGTGGWQKWEEVSCDVSVSGTEDLYLVFNGGSGYLLNVDWKFSKAGAS	
	B38	1737	CBM22-GH10-CBM22-DOC1-GH43- <b>CBM6</b>	LNPYETVQAETMSNQSKNISVSGVGNNTTVKAKKGDWIKVSGVDLSNGVSSIKVKGSGNAVVKFCVGSPTGTGICIGYDLNGSENELAAAENNVSQVGDIMYVFSGDCEFDSSWFS	
13	B11	3704	PL11- <b>CBM13</b> -DOC1	PASDIVDGGIYTFKNVNSGLYLDVEGGTAANGTNIQAAAAAGKQNFKAQVAVSAGDGYFYLVSQLGDG ESYALDVNGKKTADGTNIEYTFNKGDNQKFKVKNNDGTVSILTITKITDGGQSALDVNEQSKNAGANVQQYAFNGNANQKFTIEAV	

	B53	1157	CE12-CBM13-DOC1-CBM35-CE12	AATDIIDGQIYTFKNVNSGLYLDIEGGNGANGANVQQAANGGKASQFKAVSAGNGYYYLVSQGLDGS NSYALDVNGKKTDDGANIELYTFNKGDNDQKQFVKNNDGSYAILTKITNDASALDVNEQSKNAGANV QQYKYSGGANQKFIEAV
	B55	1875	<b>CBM13</b>	VSDTILSGAEYNIVNKLSGKLMTADSDGNVMQSAQAEGASQSWLIIRNGNGYRVLVPGSDRSMALT VAEPSALNGGNICIAEYTGDDAQLFSIEWDGSAYLTTKCEGASALDVKGKRSRSDGANIHQYKYQ GSDNRQFDITPVGH
	B41	939	CE12-CBM13-DOC1-CBM35-CE12	KEVFAPEAGASFMIKVNNSGLYMEVGGANAEEGANVQQWGANEPAAHNTWFTTQATGDYFFVNS NLGDGKTWYLYVDQGSRESGGNIVIAQKNGYSQDQFFKFEDNGDDTYTIYTRSSRDACVVEVGSACT ESGANIQWESNGNCCQKW
	B44	2115	GH43-CBM13-CBM13-DOC1	YPKVNAGTYTLRNVNSGLYVGMNASGSLIQFAEPVAWDIDENSRTFLPIDDAIPVQPTYIAVDNGSN GEDVAYKGVNADSANMKLICNDKGSYSVMTGASEYKSCWDVFEKSTEAGANIGQWFEFNGGDW QKFVIEPA
22	B31	2288	GH11-CBM22-GH10-DOC1-GH11	PAVEPDSEGYFFKESFEDGIGKCVPRGEAALILDSENVYDGEKSLFVTHRDDVWHGPAIELDPSAFV PGKTYSFGAAVLQHSDDTAEVNMLQYTTDAAGYYQYSVVSSVDAEKSEWTELGNPSFTIPDDAKE MMLFIDTPDGIADFYFDSFFGGVEG
	B33	3270	<b>CBM22-CBM22</b>	NYFVNPVGGVSVLSSAGNISPWTKNDNGLTLEYVTGANAYSKQSLRISNRNKTWNGIVQQINPSA YIPGNKYSFTMYAMCEEPRKFLTLQYTSKSGGTAYKCIDDRDGEAYEWIQLSNPRYQIPDDVDTT KPMYLYVEAKHIGSNNEDDTCPFFYVDEFIEAPMGYTA
	B36	1737	<b>CBM22-GH10-CBM22-DOC1-GH43-CBM6</b>	VADVTYAAEAVKNDFEVTYEGWHGSTVDVLDIAEEGTGTAGSRGMTVTNRTSPSEGAESSKGLYL TGGINYDYSVKVYSEDETFHLSLLYIDEKTEKETVVELDSQDVKGGTWTLSSEFKAPKDVYEFRL SITTDSTNDFSFDDVLITNQVSK
	B56	1878	<b>CBM22-GH10-CBM22-DOC1</b>	VPWTSQAAEVVYNDFEESYGGWYGNADNVVLTAEDEGCGADLSRGMKVAGRTSPFDGASSAKGF YLSGDTEYDYSVKVYSTKAEFHVTLTYADEKTGKETTGLLTSDTKADTWTELRSFKAPENTCGY LLTITDDSTDDFSFDDVRITADKP
	B6	243	CE3-CBM22-DOC1-CE15	TADASDSAKVLMSCDFESGADGWTGRGSASAAVDKSKAHSGSGSLFVSKRANDWNGAVVDLGYD FSAGNTYGFEAYILQNSAASLDFKLSLEYSSGGTTQYDKIALSPVKQGEWTKVENPSYTPAGAENIK FYIEVPDDLSDFYVDDIRITGSASSAPGGPS
	B59	2539	<b>CBM22</b>	MTPGSPEEARSDSHNAESSDGSQSLFVSGRTDYWNGATIMLSTETYKPGAYHFRANVMQKSG ETATMKMTLQYDLDEGEKYDEIALAEAPSGEWITLENLAYTIPEGAENQLYIESTDSLTDYFIDVSGA ERAD
	35	B10	1368	PL11-DOC1-CBM35-CE12
B13		359	<b>CBM35-GH26-DOC1</b>	QNVFAAEATVFPYTIIEGEDMEGAELWTQNYGPAPKEWVGKGFAYLTNGTFSFTVNPAPEDGMYDVS IKAIQVLNEEGRMQTCSVNGSEKMTNMPYSADWVDFDGTFRMNKGENTIEFPKGYMAIDTVTV TK
B54		1157	CE12-CBM13-DOC1-CBM35-CE12	NVYFASDMKITNGAPEDTNKGFTGKSYVNLNNDTSAIEWTVNAPQAGNYLCTFNIANGGADNRPM KIEVNGGKDYWMQDFLTTGDWTKWEERGIVLPLKQGSNSIKMTSASAQGGPNLDYMKTELTDPIA QIYE
B7		2757	<b>CBM35-GH26</b>	FYSLYEAEDAKLSGDLKISFERDDYSGDGYVRGFTKSSIVFDIKTSAAQHYDLSFSIASDVTDCHL SLNNEGINTFRTEGGAFTYITVYGVYMEKGTSKLELTTAGGTIDIDLKVTDSVHSHKSGSKTSAET SVKKS
B16		1364	<b>CBM35-CE3-DOC1-CBM35-GH26</b>	FEFENGTVYDGTGNITVWTLGASGGKAVELKDSGDSVTVSVNAEKDGMQTLIRYSQPYDENGK YQNVIVNGENIGQIFCAYTGEQFRVTSISAGLRSGDNTVTEGSGWGWTYLDCITIGETSVSVGANPI I

		B39	2259	GH97- <b>CBM35</b> -DOC1	EAENAVLSGKASVTAGKQGKYCSNNAYVGYVGGDQSAVTFNDVTVDKAGRYTIRIYVSGERRS LKVDINGSYVFTLNLDLYANRNDWSGIRAVNLEADLKAGKNTIRLYNDKGYGPSIDRI
		B42	939	CE12-CBM13-DOC1- <b>CBM35</b> -CE12	PYIFAVDQKWDQGMTETTNAGYTDQRGYLNLNDNTVGSVDFSVNAAQDGNMTHIRFANGSAND RKMKVTVNGDQNYVWVSFTGTGSWTDWTEFGIVLPLKAGQNTIRFESLTAEGGNLDYITLTQTD EPYAET
		B50	2327	<b>CBM35</b> -CE3-GH5-DOC1	YEIENGVISAAGSGTAVVTLGASGGKALDMKDSGDSVSIIECYSENEGMMQTISIRYQCPYDEDGK YQTVIVNGQNVGDFICAYTGEKGFSTATIKAPLIKGNKNTVEIVASWGWTFDLSLTIGGQPVASASSS A
		B61	1366	GH97- <b>CBM35</b> -DOC1	AENAQLSGFASVTSKDCYSGNTYVGYVGGGRDSYITFTNVTAEKSGEYPLRIYYISGEPRLKIDVN GKYAAALDGLYANKNDWVIAAVNTNVYLNEGNTNIRLYNDEGYAPSIDRI
	<b>48</b>	B23	806	<b>CBM48</b> -GH13	GVHVRKKGRGKIFTRVWAPNAVSVSVVGDFFNNWDRTQNPMEIADGVWEAEITKLQQFDSYKYSI ETKDGRFLMKADPYGNHFETRPATASKIYESSYEWDAEWFEKKKVVQ
		B30	934	<b>CBM48</b> -GH13	MDIYGFYKGESFEVYEYLG AHLTAKGTIFRTYAPNASKVSVIGDHTKWEVPMKSVLDGNFYETVCP DAKEGMRYKYRIYDRNGNFIDHCDPYGFGMEVVRPGTCSVIRSIENY
	<b>62</b>	B4	3398	GH30- <b>CBM62</b> -CBM62-DOC1	TNKINVDAAVNTGTSWKDSSDNYSKVFDGSGTGTFDGLGLENWVQADLGQSYDISAIGFAPRSGYE YRCADGKFMVSDDGENWTTIYTINGKPATGMNYVSKFSASATGRYIRYEIPAGAPNNEYNKDNVYN CNIAEIEVYGTPS
		B5	3398	GH30-CBM62- <b>CBM62</b> -DOC1	KLADLNKIEILTSSVTGSASWRDSSNDFTKAFDGDINSFFDGLGLENWVQADLGAVYDIDTIGFSPRKA YEARCTDGKFLFSLDGENWTEAYTITNKPVFGMQYVTDLKGDTKARYIRYEIPSGAPANQYNSDNV YNCNIAEIAV
		B15	3370	GH43- <b>CBM62</b> -DOC1	ELLSQDRPASASISSKSNESPAKAFDGSYQSGFKAIDDNKKWPFYLQVDLERVCDLANIQTSWFIYK GSEAYTYTVEGSIDGQHWKLLDRTNKNDTITKTYGFTSDMLK GKARYVRLNVQNLQNNPNN NWYTPNVFEVKVFGTPISEAS
	<b>63</b>	B3	2821	EXPN- <b>CBM63</b>	EDAPISFKYKEGSTEFWCGVQVRNHRYPITKLEYLDENGDFVEIPRRPYNYFESRDMGKGPFTFRIT DIYGVVVDKDIPLSYDDTEIIPGHVQFPE

<sup>a</sup> DOC, dockerin; GH, glycoside hydrolases; CBM, carbohydrate binding module; CE, carbohydrate esterase; EXPN, expansin; PL, polysaccharide lyases; SLH, S-layer homology domain.

**Table S3.3. List of polysaccharide samples, their major sequences or monosaccharide composition and sources, included in the *Plant, Fungal and Bacterial Polysaccharide set 1*.**

ID <sup>a</sup>	Polysaccharide	Source <sup>b</sup>	Predominant oligosaccharide sequence or monosaccharide composition	Reference <sup>c</sup>
1	Glucurono-XyloMannan	<i>Tremella fuciformis</i> Elicityl (HGL200)	$\alpha$ 1,3-Man backbone with Xyl, GlcA and Fuc branches	Khamlue <i>et al.</i> 2012 <sup>263</sup>
2	Mannan	<i>Saccharomyces cerevisiae</i> Sigma-Aldrich (M7504)	$\alpha$ 1,6-Man backbone with oligomeric $\alpha$ 1,2-, $\alpha$ 1,3-Man branches	Takahara <i>et al.</i> 2012 <sup>264</sup>
3	Mannoprotein	Brewers' spent yeast	Ara (1%), Xyl (0%), Man (65%), Glc (35%)	Pinto <i>et al.</i> 2015 <sup>265</sup>
4	Dextran	<i>Leuconostoc mesenteroides</i> Sigma-Aldrich (D4876)	$\alpha$ 1,6-Glc	Haworth <i>et al.</i> 1937 <sup>266</sup>
5	Pullulan	<i>Pullularia pullulans</i> Megazyme (P-PULLN)	Mixed-linked $\alpha$ 1,6-1,4-Glc ( $\alpha$ 1,6-linked maltotriosyl repeats)	McCleary <i>et al.</i> 1987 <sup>267</sup>
6	Curdlan*	<i>Agrobacterium</i> sp. strain ATCC31749	Linear $\beta$ 1,3-Glc	Zhang <i>et al.</i> 2012 <sup>268</sup>
7	Pustulan	<i>Lasallia pustulata</i> Elicityl (GLU900)	Linear $\beta$ 1,6-Glc	de la Cruz <i>et al.</i> 1995 <sup>269</sup>
8	NSG- $\beta$ -glucan (Neutral soluble glucan)	<i>Saccharomyces cerevisiae</i> Biothera	Linear $\beta$ 1,3-Glc backbone with occasional monoglucosyl $\beta$ 1,6-Glc branches	Hong <i>et al.</i> 2003 <sup>270</sup>
9	PGG- $\beta$ -glucan (Poly-(1,6)-D-glucopyranosyl- (1,3)-D-glucopyranose)	<i>Saccharomyces cerevisiae</i> Biothera	Linear $\beta$ 1,3-Glc backbone with occasional monoglucosyl $\beta$ 1,6-Glc branches	Jamas <i>et al.</i> 1991 <sup>271</sup>
10	Lentinan	<i>Lentinus edodes</i>	Linear $\beta$ 1,3-Glc backbone with two $\beta$ 1,6-Glc branches every 5 residues	Wang <i>et al.</i> 2008 <sup>272</sup>
11	Grifolan	<i>Grifola frondosa</i>	$\beta$ 1,3-Glc backbone with highly ramified oligomeric $\beta$ 1,6-Glc branches	Du <i>et al.</i> 2004 <sup>273</sup>
12	$\beta$ -glucan (Barley)	Barley flour Megazyme (P-BGBL)	Mixed-linked $\beta$ 1,3-1,4-Glc; 1:3-4 linkage ratio; contains Ara (2%), Xyl (0.2%)	Yoo <i>et al.</i> 2007 <sup>274</sup>
13	$\beta$ -glucan (Oat)	Oat flour, Megazyme (P-BGOM)	Mixed-linked $\beta$ 1,3-1,4-Glc	-
14	Lichenan	Icelandic moss Megazyme (P-LICHN)	Mixed-linked $\beta$ 1,3-1,4-Glc; 1:2 linkage ratio	-
15	Xylan ( <i>Palmaria p.</i> )	<i>Palmaria palmata</i> Elicityl (XYL100)	Mixed-linked $\beta$ 1,3- $\beta$ 1,4-Xyl; 1:4 linkage ratio	-
16	Xylan (Plum fresh)	Fresh plum <i>Prunus domestica L.</i>	Rha (3%), Fuc (2%), Ara (11%), Xyl (67%), Man (0%), Gal (6%), Glc (5%), GalA (6%)	Nunes <i>et al.</i> 2008 <sup>275</sup>
17	Xylan (Plum boiled)	Boiled plum <i>Prunus domestica L.</i>	Rha (3%), Fuc (3%), Ara (12%), Xyl (73%), Man (0%), Gal (0%), Glc (4%), GalA (5%)	Nunes <i>et al.</i> 2008 <sup>275</sup>
18	Xyloglucan (Plum fresh)	Fresh plum <i>Prunus domestica L.</i>	Rha (2%), Fuc (5%), Ara (5%), Xyl (40%), Man (6%), Gal (13%), Glc (24%), Ur Ac (6%)	Nunes <i>et al.</i> 2008 <sup>275</sup>
19	Xyloglucan (Plum boiled)	Boiled plum <i>Prunus domestica L.</i>	Rha (2%), Fuc (6%), Ara (6%), Xyl (46%), Man (4%), Gal (14%), Glc (22%), Ur Ac (1%)	Nunes <i>et al.</i> 2008 <sup>275</sup>
20	Arabinoxylan (DP41)	Brewers' spent grain	Ara (40%), Xyl (54%), Man (0%), Gal (3%), Glc (3%)	Coelho <i>et al.</i> 2016 <sup>276</sup>

21	Arabinoxylan (DP24)	Brewers' spent grain	Ara (25%), Xyl (46%), Man (1%), Gal (3%), Glc (25%)	Coelho <i>et al.</i> 2016 <sup>276</sup>
22	Arabinogalactan	Spent coffee grounds	Ara (5%), Man (29%), Gal (64%), Glc (1%)	Passos <i>et al.</i> 2013 <sup>277</sup>
23	Galactomannan (Carob)	Carob Megazyme (P-GALML)	$\beta$ 1,4-Man backbone with $\alpha$ 1,6-Gal ramifications; Gal (24%), Man (76%)	-
24	Galactomannan (Guar)	Guar Megazyme (P-GGMM)	$\beta$ 1,4-Man backbone with $\alpha$ 1,6-Gal ramifications; Gal (38%), Man (62%)	-
25	Galactomannan Guar ( $\Delta$ Gal)	Guar (Galactose depleted) Megazyme (P-GGM21)	$\beta$ 1,4-Man backbone with $\alpha$ 1,6-Gal ramifications; Gal (21%), Man (79%)	-

<sup>a</sup> Polysaccharides are grouped according to predominant oligosaccharide sequence, glycosidic linkage or monosaccharide composition. The ID corresponds to the positions in the binding charts or heatmap;

<sup>b</sup> The sources are indicated for each carbohydrate sample; if commercial the code product number is indicated;

<sup>c</sup> References for the structural analysis or recent published work for each particular sample, if available.

**Table S3.4. Fluorescence binding intensities elicited with all the proteins investigated for the validation of the *Plant, Fungal and Bacterial Polysaccharide set 1*.** The numerical scores for the fluorescence binding signals are shown as means of duplicate spots at 150 pg/spot probe per spot (as in Figure 3.2) and are representative of at least 2 independent experiments.

ID <sup>a</sup>	Probe	ConA	Tm CBM41	Cm CBM6-2	400-3	Cm CBM32-2	400-4	LM10	LM11	LM24	LM25	CCRC-M1	AAL	LM6	LM21	400-2	CCRC-M70	RCA <sub>120</sub>
1	Glucurono-XyloMannan	39198	<sup>b</sup> -	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
2	Mannan	24284	-	2732	-	-	924	-	-	-	-	-	-	-	-	-	-	-
3	Mannoprotein	24199	-	8255	6032	-	-	-	-	-	-	-	-	-	-	-	-	-
4	Dextran	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
5	Pullulan	-	11982	1129	-	-	-	-	-	-	-	-	-	-	-	-	-	-
6	Curdlan	-	-	22350	5677	-	-	-	-	-	-	-	-	-	-	-	-	-
7	Pustulan	4153	-	1653	-	-	-	2144	-	-	-	-	-	-	-	-	-	-
8	NSG- $\beta$ -glucan	-	-	7288	3520	-	-	-	-	-	-	-	-	-	-	-	-	-
9	PGG- $\beta$ -glucan	-	-	29810	9100	-	-	-	-	-	-	-	-	-	-	-	-	-
10	Lentinan	-	-	59233	1337	62520	-	-	-	-	-	-	1029	-	-	-	-	-
11	Grifolan	-	-	23264	-	14842	-	-	-	-	-	-	584	-	-	-	-	-
12	$\beta$ -glucan (Barley)	-	-	62150	-	-	35387	-	-	-	-	-	-	-	584	617	-	-
13	$\beta$ -glucan (Oat)	-	-	63228	-	-	-	-	-	-	-	-	-	-	-	-	-	-
14	Lichenan	-	-	6440	-	-	5243	-	-	-	-	-	-	-	-	-	-	1017
15	Xylan ( <i>Palmaria p.</i> )	-	-	-	-	-	-	62371	-	-	-	-	-	-	-	-	-	-
16	Xylan (Plum fresh)	-	-	-	-	-	-	1441	10463	-	55705	15721	13481	2977	878	-	-	-
17	Xylan (Plum boiled)	670	-	1184	1168	-	724	656	10036	-	22082	794	2069	994	6361	-	-	-
18	Xyloglucan (Plum fresh)	-	-	789	1272	-	-	-	2459	4674	64926	60683	46906	790	3744	-	-	3239
19	Xyloglucan (Plum boiled)	-	-	1690	2977	-	-	-	2155	11529	64934	62257	39338	2932	4985	-	-	7075
20	Arabinoxylan (Dreche DP41)	1494	-	36243	1522	-	18914	-	53424	-	15330	578	7652	16899	-	-	-	1317
21	Arabinoxylan (Dreche DP24)	2055	-	46648	1333	-	28885	4202	50207	-	24892	-	2572	1804	-	-	-	712
22	Arabinogalactan	-	-	-	-	-	-	-	-	-	-	-	-	-	2790	1162	-	712
23	Galactomannan (Carob)	-	-	-	-	-	-	-	-	-	568	-	-	9620	44974	26612	64694	12710
24	Galactomannan (Guar)	-	-	-	-	-	-	-	-	-	-	-	-	1997	58687	26591	28824	1802
25	Galactomannan Guar ( $\Delta$ Gal)	-	-	-	-	-	-	-	-	-	1031	-	-	12780	47337	22983	16509	-

<sup>a</sup>ID, Probe position in the microarray matching the position in the heatmap and in Table S3.3; <sup>b</sup>The binding signals are means of fluorescence intensities of duplicate spots at 150 pg of probe arrayed (the respective standard deviation was calculated as the associated error, overall < 5%). '-' refers to a fluorescence intensity < 500.



**Table S3.5. Fluorescence binding intensities elicited with the CBMs from *C. thermocellum* investigated in the Plant, Fungal and Bacterial Polysaccharide set 1.** The numerical scores for the fluorescence binding signals are shown as means of duplicate spots at 150 pg/spot probe per spot (as in Figure 3.4) and are representative of at least 2 independent experiments.

Family	3				4				6						11	13	22				25	30	32	35	42			62		
ID <sup>a</sup>	Cthe_3077	Cthe_0059	Cthe_0040	Cthe_0433	Cthe_2809	Cthe_2809	Cthe_0413	Cthe_1271	Cthe_1963	Cthe_3012	Cthe_2972	Cthe_2197	Cthe_2194	Cthe_2195	Cthe_1472	Cthe_0661	Cthe_2590	Cthe_1838	Cthe_0912	Cthe_0912	Cthe_0956	Cthe_0624	Cthe_0821	Cthe_2811	Cthe_2139	Cthe_2138	Cthe_1273	Cthe_2193		
1	- <sup>b</sup>	-	-	-	-	-	-	-	-	-	-	-	16205	59536	62494	-	-	-	-	-	-	564	-	-	-	-	-	-	-	
2	-	-	-	-	-	-	1886	-	-	-	-	-	-	-	3431	-	-	-	-	-	-	-	2707	-	-	-	-	-	-	
3	-	-	-	-	-	785	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1515	-	-	-	-	-	-	-	
4	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
5	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	5725	-	11254	-	-	-	-	-	
6	-	-	-	-	31146	7202	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
7	-	-	-	-	-	-	-	-	-	-	-	-	-	-	588	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
8	-	-	-	-	-	631	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
9	-	-	-	-	9225	18573	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	575	-	
10	-	-	-	-	-	1172	-	-	-	-	-	-	37210	39548	-	1873	-	-	-	-	-	-	-	64258	-	-	-	-	-	
11	-	-	-	-	-	-	-	-	-	-	-	-	4349	4770	-	-	-	-	-	-	-	-	-	12981	-	-	-	-	-	
12	-	-	-	8694	-	989	64659	-	-	-	-	-	-	-	64904	-	15138	1660	795	47053	-	64329	-	745	-	-	-	-	-	
13	-	-	-	42281	-	2348	64787	7713	4673	5166	2136	11308	9002	7095	64737	-	53488	25230	17762	63211	-	64230	-	8431	1256	1521	1302	-	-	
14	-	-	-	-	-	-	14594	818	542	854	376	849	-	-	10528	609	519	506	829	692	-	-	-	-	-	-	-	-	1174	
15	-	-	-	-	-	3308	-	61132	63119	62592	63593	-	-	-	41447	-	63418	60161	62331	62899	-	-	-	-	-	-	-	-	-	
16	2569	64819	2969	16027	-	-	12041	14162	13121	21484	9801	-	-	-	1917	-	3302	5349	16900	17125	-	2905	-	-	1764	2143	2467	-	-	
17	-	2932	-	-	-	-	1379	16625	11811	29433	16154	-	-	-	1230	-	5977	7498	24998	26667	-	597	-	-	1681	1666	1526	-	-	
18	23114	64487	9616	63170	-	-	33596	7186	7891	12498	4653	-	-	1620	5637	4489	1435	3778	11027	9754	-	24441	-	-	1070	1322	1288	9561	-	
19	18661	64627	13940	64370	-	868	47586	8712	6587	15446	6405	-	-	2501	5736	12330	1965	4066	10562	11916	-	28051	-	-	3390	3773	4614	20418	-	
20	-	10174	-	-	-	-	23280	59736	62900	61912	63159	54149	63207	63025	22087	-	63372	59507	62397	60471	-	4856	-	-	52429	60338	64710	1270	-	
21	-	31788	-	-	-	-	46355	58149	62407	60948	62952	55250	62571	63323	53839	-	63395	58344	61410	61978	-	13547	-	-	32774	41022	49411	-	-	
22	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1546	-	-	-	-	-	-	-	-	-	-	-	-	782	-
23	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	609	-	-	-	-	-	-	-	-	40914	631	1084	686	28446	-
24	-	-	-	-	516	-	-	6919	1066	1139	547	-	-	-	-	-	684	688	899	812	-	-	-	41713	-	633	-	4279	-	
25	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	64647	1373	1928	1857	3995	-	-

<sup>a</sup>ID, Probe position in the microarray matching the position in the heatmap and in Table S3.3; <sup>b</sup>The binding signals are means of fluorescence intensities of duplicate spots at 150 pg of probe arrayed (the respective standard deviation was calculated as the associated error, overall < 5%). '-' refers to a fluorescence intensity < 500.

**Table S3.6. Fluorescence binding intensities elicited with the CBMs from *R. flavefaciens* FD-1 investigated in the *Plant, Fungal and Bacterial Polysaccharide set 1*. The numerical scores for the fluorescence binding signals are shown as means of duplicate spots at 150 pg/spot (as in Figure 3.4) and are representative of at least 2 independent experiments.**

Family	4				6	13				22											35				
ID <sup>a</sup>	2836	3259	776	3995	3747	2115	2326	694	1878	2649	1615	1272	2646	3180	2002	3077	3077	3190	1766	1737	3270	1364	933	2302	
1	- <sup>b</sup>	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
2	-	2394	6151	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
3	-	3656	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
4	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
5	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
6	-	19236	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
7	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
8	-	784	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
9	23886	5417	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
10	-	5034	-	-	-	-	-	-	-	-	-	-	-	-	-	1534	-	-	-	-	-	-	-	-	-
11	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
12	-	39898	62972	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	9960	-	-	-	-	-
13	2028	40498	64872	-	-	-	-	-	-	-	-	-	-	751	997	1389	-	-	1147	14608	-	552	995	3799	-
14	-	-	1062	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1281	-	-	-	-	-
15	-	-	14234	33072	-	-	-	-	32571	64706	56137	29657	30598	46520	37721	64685	45397	64938	64938	63865	-	-	-	-	-
16	-	655	1504	-	-	1900	-	-	-	-	-	-	-	-	-	-	-	600	-	6435	-	-	-	-	-
17	-	793	1555	-	-	-	-	-	-	-	-	-	-	-	-	540	-	-	604	5010	-	-	-	-	-
18	-	-	2905	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2259	-	-	-	-	-
19	-	639	2088	-	-	1085	-	-	-	-	-	-	-	-	-	-	-	-	-	1871	-	-	-	-	-
20	-	4228	12795	18700	25761	7231	36945	1150	29278	60957	41316	10918	23251	31108	32681	64196	31423	46105	55550	61995	1218	-	-	-	-
21	-	12162	32428	5837	11091	-	16622	722	12994	36941	12751	3225	7568	12006	13686	25968	15544	15792	32904	52529	1679	-	-	-	-
22	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
23	-	-	-	-	-	4621	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	44312	15418	64557	-
24	-	-	-	-	1561	1196	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	41936	18024	60448	-
25	-	-	-	-	-	8336	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	64685	29647	65005	-

<sup>a</sup>ID, Probe position in the microarray matching the position in the heatmap and in Table S3.3; <sup>b</sup>The binding signals are means of fluorescence intensities of duplicate spots at 150 pg of probe arrayed (the respective standard deviation was calculated as the associated error, overall < 5%). '-' refers to a fluorescence intensity < 500.

**Table S3.7. Fluorescence binding intensities elicited with the CBMs from *C. thermocellum* investigated in the Glucan, hemicellulose, chitin and chitosan NGL-microarrays.** The numerical scores for the fluorescence binding signals are shown as means of duplicate spots at 5 fmol probe per spot (as in Figure 3.5) and are representative of at least 2 independent experiments.

Family	4		6						11	13	22				25	32	35	42			50						62	
ID <sup>a</sup>	Cthe_ 2809	Cthe_ 0413	Cthe_ 1271	Cthe_ 1963	Cthe_ 3012	Cthe_ 2972	Cthe_ 2197	Cthe_ 2195	Cthe_ 1472	Cthe_ 0661	Cthe_ 2590	Cthe_ 1838	Cthe_ 0912	Cthe_ 0912	Cthe_ 0956	Cthe_ 0821	Cthe_ 2811	Cthe_ 2139	Cthe_ 2138	Cthe_ 1273	Cthe_ 0300	Cthe_ 2387	Cthe_ 3006	Cthe_ 3006	Cthe_ 3005	Cthe_ 1800	Cthe_ 3007	Cthe_ 2193
1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1054	-	-	-	-	-	-	-
3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	833	-	-	-	-	-	-	-
4	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
5	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
6	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
7	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
8	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
9	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
10	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
11	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
12	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
13	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	844	-	-
14	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
15	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
16	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
17	-	-	-	-	-	-	-	-	-	-	-	-	-	-	810	-	-	-	-	-	-	-	-	-	-	-	-	-
18	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1817	-	-	-	-	-	-	-	-	-	-	-	-	-
19	-	-	-	-	-	-	-	-	-	-	-	-	-	-	3319	-	-	-	-	-	-	-	-	-	-	-	-	-
20	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1463	-	-	-	-	-	-	-	-	-	-	-	-	-
21	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
22	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
23	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2008	-	-	-	-	-	-	-	-	-	-	-	-	-
24	-	-	-	-	-	-	-	-	-	-	-	-	-	-	5253	-	-	-	-	-	-	-	-	-	-	-	-	-
25	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
26	-	-	-	-	-	-	-	-	-	-	-	-	-	-	5824	-	-	-	-	-	-	-	-	-	-	-	-	-
27	-	-	-	-	-	-	-	-	-	-	-	-	-	-	7551	-	-	-	-	-	-	-	-	-	-	-	-	-
28	-	-	-	-	-	-	-	-	-	-	-	-	-	-	4774	-	-	-	-	-	-	-	-	-	-	-	-	-
29	-	-	-	-	-	-	-	-	-	-	-	-	-	-	7354	-	-	-	-	-	-	-	-	-	-	-	-	-
30	-	-	-	-	-	-	-	-	-	-	-	-	-	-	9292	-	-	-	-	-	-	-	-	-	-	-	-	-

CHAPTER 3. SUPPLEMENTARY INFORMATION

Family	4		6						11	13	22				25	32	35	42			50							62
ID <sup>a</sup>	Cthe_2809	Cthe_0413	Cthe_1271	Cthe_1963	Cthe_3012	Cthe_2972	Cthe_2197	Cthe_2195	Cthe_1472	Cthe_0661	Cthe_2590	Cthe_1838	Cthe_0912	Cthe_0912	Cthe_0956	Cthe_0821	Cthe_2811	Cthe_2139	Cthe_2138	Cthe_1273	Cthe_0300	Cthe_2387	Cthe_3006	Cthe_3006	Cthe_3005	Cthe_1800	Cthe_3007	Cthe_2193
31	-	-	-	-	-	-	-	-	-	-	-	-	-	-	9363	-	-	-	-	-	-	-	-	-	-	-	-	-
32	-	-	-	-	-	-	-	-	-	-	-	-	-	-	3209	-	-	-	-	-	-	-	-	-	-	-	-	-
33	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
34	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
35	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
36	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
37	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
38	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
39	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
40	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
41	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
42	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
43	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
44	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
45	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
46	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
47	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
48	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1322	-	-	-	-	-	-	-	-	-	-	-	-	-
49	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
50	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
51	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
52	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
53	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
54	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
55	-	-	-	-	-	-	-	519	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
56	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
57	-	-	622	561	591	571	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
58	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
59	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
60	-	-	-	-	-	-	-	808	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
61	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
62	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
63	78.5	-	-	-	-	-	-	-	-	-	-	-	-	-	-	651	-	-	-	-	-	-	-	-	-	-	-	-

Family	4		6						11	13	22				25	32	35	42				50						62
ID <sup>a</sup>	Cthe_2809	Cthe_0413	Cthe_1271	Cthe_1963	Cthe_3012	Cthe_2972	Cthe_2197	Cthe_2195	Cthe_1472	Cthe_0661	Cthe_2590	Cthe_1838	Cthe_0912	Cthe_0912	Cthe_0956	Cthe_0821	Cthe_2811	Cthe_2139	Cthe_2138	Cthe_1273	Cthe_0300	Cthe_2387	Cthe_3006	Cthe_3006	Cthe_3005	Cthe_1800	Cthe_3007	Cthe_2193
64	809.5	-	-	-	-	-	-	-	-	-	-	-	-	-	-	3897	-	-	-	-	-	-	-	-	-	-	-	-
65	5056	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2424	-	-	-	-	-	-	-	-	-	-	-	-
66	7752	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1623	-	-	-	-	-	-	-	-	-	-	-	-
67	7625	-	-	-	-	-	-	-	-	-	-	-	-	-	-	918	-	-	-	-	-	-	-	-	-	-	-	-
68	2069	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
69	21573	-	-	-	-	-	-	-	-	-	-	-	-	-	-	873	-	-	-	-	-	-	-	-	-	-	-	-
70	28113	-	-	-	-	-	-	-	-	-	-	-	-	-	-	591	-	-	-	-	-	-	-	-	-	-	-	-
71	25935	-	-	-	-	-	-	-	-	-	-	-	-	-	-	576	-	-	-	-	-	-	-	-	-	-	-	-
72	33165	-	-	-	-	506	-	-	-	-	-	-	-	-	-	676	-	-	-	-	-	-	-	-	-	-	-	-
73	24100	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
74	17929	-	-	-	-	-	-	1136	-	-	-	-	-	-	-	1488	-	-	-	-	-	-	-	-	-	-	-	-
75	6932	-	-	-	-	-	-	-	-	-	-	-	-	-	-	944	-	-	-	-	-	-	-	-	-	-	-	-
76	6428	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
77	6943	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
78	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	3773	-	-	-	-	-	-	-	-	-	-	-	-
79	-	6366	-	-	-	-	-	-	-	-	-	-	-	-	-	2508	-	-	-	-	-	-	-	-	-	-	-	-
80	-	11752	-	-	-	-	-	-	-	-	-	-	-	-	-	1538	-	-	-	-	-	-	-	-	-	-	-	-
81	-	10400	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
82	-	9288	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
83	1290	20301	904	469	800	1053	-	-	-	-	-	-	-	-	-	924	-	-	-	-	-	-	-	-	-	-	-	-
84	2485	25790	-	-	-	594	-	-	2306	-	-	-	-	-	-	2718	-	-	-	-	-	-	-	-	-	-	-	-
85	4227	27656	-	-	-	1035	-	811	3250	-	-	-	-	-	-	3683	-	-	-	-	-	-	-	-	-	-	-	-
86	2630	26122	-	-	-	1527	-	588	4396	-	-	-	-	-	-	1948	-	-	-	-	-	-	-	-	-	-	-	-
87	3748	16041	-	-	-	-	-	-	2035	-	-	-	-	-	-	2219	-	-	-	-	-	-	-	-	-	-	-	-
88	4688	32325	-	-	-	-	-	-	3851	-	-	-	-	-	-	1933	-	-	-	-	-	-	-	-	-	-	-	-
89	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
90	-	-	-	-	-	-	-	-	-	2354	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
91	-	-	-	-	-	-	-	-	-	1794	-	-	-	-	-	218.5	-	-	-	-	-	-	-	-	-	-	-	-
92	-	-	-	-	-	-	-	-	-	6391	-	-	-	-	-	1167	-	-	-	-	-	-	-	-	-	-	-	-
93	-	-	-	-	-	-	-	-	-	7507	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
94	-	-	-	-	-	-	-	-	-	4944	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
95	-	-	-	-	-	-	-	-	-	1941	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
96	-	-	-	-	-	-	-	-	-	1893	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
97	-	-	-	-	-	-	-	-	-	2966	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
98	-	-	-	-	-	-	-	-	-	656	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

CHAPTER 3. SUPPLEMENTARY INFORMATION

Family	4		6						11	13	22				25	32	35	42			50						62	
ID <sup>a</sup>	Cthe_2809	Cthe_0413	Cthe_1271	Cthe_1963	Cthe_3012	Cthe_2972	Cthe_2197	Cthe_2195	Cthe_1472	Cthe_0661	Cthe_2590	Cthe_1838	Cthe_0912	Cthe_0912	Cthe_0956	Cthe_0821	Cthe_2811	Cthe_2139	Cthe_2138	Cthe_1273	Cthe_0300	Cthe_2387	Cthe_3006	Cthe_3006	Cthe_3005	Cthe_1800	Cthe_3007	Cthe_2193
99	-	-	-	-	-	-	-	-	-	752	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
100	-	-	-	-	-	-	-	-	-	562	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
101	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1478	-	-	-	-	-	3833	-	-	-	-	-	-
102	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
103	-	621	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
104	582	-	-	-	-	-	-	-	-	-	-	-	-	-	-	796	-	-	-	-	-	-	-	-	-	-	-	-
105	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	582	-	-	-	-	-	-	-	-	-	-	-	-
106	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	725	-	-	-	-	-	-	-	-	-	-	-	-
107	1204	1405	-	-	-	-	-	-	1298	-	-	-	-	-	-	1614	-	-	-	-	-	-	-	-	-	-	-	-
108	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1679	-	-	-	-	-	-	-	-	-	-	-	-
109	-	4907	-	-	-	-	-	-	2339	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
110	2763	-	-	-	-	-	-	-	-	-	-	-	-	-	-	3388	-	-	-	-	-	-	-	-	-	-	-	-
111	4142	12137	-	-	-	-	-	-	16177	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
112	5381	10929	-	-	-	-	-	-	12698	-	-	-	-	-	-	1028	-	-	-	-	-	-	-	-	-	-	-	-
113	12249	15778	-	-	-	-	-	-	31593	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
114	4551	11384	-	-	-	-	-	-	19418	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
115	13234	13251	-	-	-	-	-	-	21911	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
116	13060	16152	-	-	-	-	-	-	30594	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
117	8065	11941	-	-	-	-	-	-	18931	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
118	20452	19515	-	-	-	-	-	-	7868	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
119	12927	11263	-	-	-	-	1869	-	14544	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
120 <sup>b</sup>	17421	14218	15570	16308	14666	14729	-	2264	28558	6963	13436	16021	6695	6692	-	-	-	3918	-	3552	-	-	-	-	-	-	-	-
121	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
122	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1031	-	-	-	-	-	-	-	-	-	-	-	-
123	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
124	-	-	-	-	-	-	-	-	-	747	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
125	-	-	-	-	-	-	-	-	-	702	-	-	-	-	-	969	-	-	-	-	-	-	-	-	-	-	-	-
126	-	-	-	-	-	532	-	-	-	844	-	-	-	-	-	729	-	-	-	-	-	-	-	-	-	-	-	-
127	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
128	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
129	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
130	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
131	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
132	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
133	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Family	4		6						11	13	22				25	32	35	42				50						62
ID <sup>a</sup>	Cthe_2809	Cthe_0413	Cthe_1271	Cthe_1963	Cthe_3012	Cthe_2972	Cthe_2197	Cthe_2195	Cthe_1472	Cthe_0661	Cthe_2590	Cthe_1838	Cthe_0912	Cthe_0912	Cthe_0956	Cthe_0821	Cthe_2811	Cthe_2139	Cthe_2138	Cthe_1273	Cthe_0300	Cthe_2387	Cthe_3006	Cthe_3006	Cthe_3005	Cthe_1800	Cthe_3007	Cthe_2193
134	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
135	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	620	-	-	-	-	-	-	-	-	-	-	-	-
136	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
137	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
138	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2624	-	-	-	-	-	-	-	-	-	-	-	-
139	525	-	-	-	-	-	-	576	-	-	-	-	-	-	-	1669	-	-	-	-	-	-	-	-	-	-	-	-
140	1627	-	-	-	-	964	-	985	-	-	-	-	-	-	-	967	-	-	-	-	-	-	-	-	-	-	-	-
141	5943	-	-	-	-	662	-	814	-	-	-	-	-	-	-	2027	-	-	-	-	-	-	-	-	-	-	-	-
142	5840	-	-	-	-	715	-	501	-	-	-	-	-	-	-	1656	-	-	-	-	-	-	-	-	-	-	-	-
143	5416	-	-	-	-	538	-	506	-	-	-	-	-	-	-	566	-	-	-	-	-	-	-	-	-	-	-	-
144	10726	-	-	-	-	859	-	892	-	-	-	-	-	-	-	677	-	-	-	-	-	-	-	-	-	-	-	-
145	10775	-	-	-	-	614	-	-	-	-	-	-	-	-	-	593	-	-	-	-	-	-	-	-	-	-	-	-
146	22247	-	-	-	-	1418	-	1239	-	-	-	-	-	-	-	2446	-	-	-	-	-	-	-	-	-	-	-	-
147	1043	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1067	-	-	-	-	-	-	-	-	-	-	-	-
148	26737	-	-	-	-	1002	-	-	-	-	-	-	-	-	-	869	-	-	-	-	-	-	-	-	-	-	-	-
149	22370	-	-	-	-	1958	-	769	-	-	-	-	-	-	-	1594	-	-	-	-	-	-	-	-	-	-	-	-
150	12900	-	-	-	-	-	-	631	-	-	-	-	-	-	-	877	-	-	-	-	-	810	-	-	-	-	-	-
151	11866	-	-	-	-	-	-	-	-	-	-	-	-	-	-	685	-	-	-	-	-	-	-	-	-	-	-	-
152	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	515	-	-	-	-	-	-	-	-	-	-	-	-
153	11491	-	-	-	-	-	-	-	-	-	-	-	-	-	-	839	-	-	-	-	-	-	-	-	-	-	-	-
154	-	-	-	-	1834	-	3757	14636	34338	-	-	-	3156	-	570	-	-	-	-	-	-	1096	-	-	-	-	-	-
155	-	-	-	4388	6735	2974	10940	11750	31264	-	-	-	4278	-	646	3215	-	-	-	-	-	-	-	-	-	-	-	-
156	-	-	-	23273	38062	17167	43017	12238	35387	-	-	1820	11043	8104	11406	-	-	-	-	-	-	842	-	-	-	-	-	-
157	-	-	-	49668	46691	35912	59301	21884	33666	-	-	8800	38052	14534	17999	-	-	-	-	-	-	1539	-	-	-	-	-	-
158	-	-	-	33997	44878	27333	40112	26380	38666	-	-	7833	59920	28132	34844	-	-	-	-	-	-	3379	-	-	-	-	-	-
159	-	-	-	58216	58662	50687	58899	30820	54509	-	-	15451	37484	30698	39892	-	-	-	-	-	-	3432	-	-	-	-	-	-
160	-	-	-	58397	58786	58848	58793	47872	61897	-	-	21139	47615	39566	44545	-	-	-	-	-	-	7574	-	-	-	-	-	-
161	-	-	-	54512	57867	50810	59392	27318	44125	-	-	28534	45413	33640	33815	-	-	-	-	-	-	3758	-	-	-	-	-	-
162	-	-	-	48460	53196	44115	53621	25936	34355	-	-	27820	59556	32356	45723	-	-	-	-	-	-	7152	-	-	-	-	-	-
163	-	-	-	53852	57788	50353	58857	20489	42478	-	-	34670	34064	29560	32272	-	-	-	-	-	-	4873	-	-	-	-	-	-
164	-	-	-	53636	57485	58342	58773	19215	38825	-	-	30638	35954	32992	37049	-	-	-	-	-	-	8022	-	-	-	-	-	-
165	-	-	-	-	-	-	-	624.5	13884	-	-	-	-	-	-	-	-	-	-	-	-	29250	44234	43808	-	-	-	-
166	-	-	-	-	-	-	-	21509	49587	-	-	-	624	-	-	-	-	-	-	-	-	12107	19687	25533	-	-	-	-
167	-	-	-	719	-	-	-	26315	54786	-	-	-	916	-	-	-	-	-	-	-	-	22780	34479	32213	-	-	-	-
168	-	-	-	4213	16219	3748	4627	3330	12348	-	-	-	24153	5173	7593	-	-	-	-	-	-	15971	17812	27485	-	-	-	-

CHAPTER 3. SUPPLEMENTARY INFORMATION

Family	4		6						11	13	22				25	32	35	42			50							62
ID <sup>a</sup>	Cthe_2809	Cthe_0413	Cthe_1271	Cthe_1963	Cthe_3012	Cthe_2972	Cthe_2197	Cthe_2195	Cthe_1472	Cthe_0661	Cthe_2590	Cthe_1838	Cthe_0912	Cthe_0912	Cthe_0956	Cthe_0821	Cthe_2811	Cthe_2139	Cthe_2138	Cthe_1273	Cthe_0300	Cthe_2387	Cthe_3006	Cthe_3006	Cthe_3005	Cthe_1800	Cthe_3007	Cthe_2193
169	-	-	13763	25495	11829	24194	15847	46145	-	-	-	12322	4027	6572	-	-	-	17143	33879	22644	-	-	-	-	-	-	-	-
170	-	-	-	-	-	-	-	1916	-	-	-	-	-	1460	-	-	-	23401	44820	27240	-	-	-	-	-	-	-	-
171	-	-	2817	4716	3308	1238	6436	37404	-	-	-	2907	1690	5944	-	-	-	19147	45782	26445	-	-	-	-	-	-	-	-
172	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	10759	14419	18107	-	-	-	-	-	-	-	-
173	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	18378	18095	25177	-	-	-	-	-	-	-	-
174	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	14934	32191	23482	-	-	-	-	-	-	-	-
175	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	17761	39453	27425	-	-	-	-	-	-	-	-
176	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	33864	48928	35841	-	-	-	-	-	-	-	-
177	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	36622	55954	39124	-	-	-	-	-	-	-	-
178	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	17398	29816	21811	-	-	-	-	-	-	-	-
179	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	17664	31637	24895	-	-	-	-	-	-	-	-
180	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	38463	58396	48610	-	-	-	-	-	-	-	-
181	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	51265	46560	60235	-	-	-	-	-	-	-	-
182	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	25062	4726	-	-	-	-	-	-	-	-	-	-	-
183	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	18294	6085	-	-	-	-	-	-	-	-	-	-	-
184	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	31797	14864	-	-	-	-	-	-	-	-	-	-	-
185	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	8784	11514	-	-	-	-	-	-	-	-	-	-	-
186	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2433	-	-	-	-	-	-	-	-	-	-	-	-
187	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	11672	-	-	-	-	-	-	-	-	-	-	-	-
188	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	6120	-	-	-	-	-	-	-	-	-	-	-	-
189	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	10465	-	-	-	-	-	-	-	-	-	-	-	-
190	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2949	1138	-	-	-	-	6282	-	-	-	-	-	-
191	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	11625	6685	-	-	-	-	-	-	-	-	-	-	-
192	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1546	1101	-	-	-	-	-	-	-	-	-	-	-
193	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	11415	-	-	-	-	-	-	-	-	-	-	-	-
194	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	10691	-	-	-	-	-	-	-	-	-	-	-	1201
195	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	5060	-	-	-	-	-	-	-	-	-	-	-	-
196	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	11225	-	-	-	-	-	-	-	-	-	-	-	690
197	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	36495	14775	-	-	-	-	-	-	-	-	-	-	-
198	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	12171	16277	-	-	-	-	-	-	-	-	-	-	-
199	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	15189	21705	-	-	-	-	-	-	-	-	-	-	1242
200	-	-	-	-	-	-	-	14594	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
201	-	-	-	-	-	-	-	6621	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2519
202	-	-	-	727	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	7022
203	-	-	-	-	-	-	-	4024	-	-	-	-	-	-	-	1181	-	-	-	-	-	-	-	-	-	-	-	3343



Family	4		6						11	13	22				25	32	35	42			50						62	
ID <sup>a</sup>	Cthe_2809	Cthe_0413	Cthe_1271	Cthe_1963	Cthe_3012	Cthe_2972	Cthe_2197	Cthe_2195	Cthe_1472	Cthe_0661	Cthe_2590	Cthe_1838	Cthe_0912	Cthe_0912	Cthe_0956	Cthe_0821	Cthe_2811	Cthe_2139	Cthe_2138	Cthe_1273	Cthe_0300	Cthe_2387	Cthe_3006	Cthe_3006	Cthe_3005	Cthe_1800	Cthe_3007	Cthe_2193
204	-	-	-	-	-	-	-	1145	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	3443
205	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1983	-	-	-	-	-	-	-	-	-	-	-	-
206	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	7400	-	-	-	-	2223	669	-	2061	-	-	-	
207	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	9714	-	-	-	-	23614	5754	21545	23342	-	-	-	
208	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	8004	-	-	-	-	19214	6930	19759	11998	3794	1531	1612	
209	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	4381	-	-	-	-	11060	2751	4849	2457	2780	2201	-	
210	-	-	-	-	-	-	-	-	-	-	-	686	-	-	-	7107	-	-	-	-	6499	4616	9008	6753	5776	1202	-	
211	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1828	-	-	-	-	10835	3044	12530	5828	7147	-	-	
212	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
213	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
214	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
215	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
216	-	-	-	-	-	-	-	-	-	5138	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	13447	
217	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	6686	
218	-	-	-	-	-	-	-	-	-	983	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2039	

<sup>a</sup>ID, Probe position in the microarray matching the position in the heatmap, binding-charts and Table S2.1 in Chapter 2; <sup>b</sup>In the  $\beta$ 1,3-1,4-linked barley series DP-16 fraction (probe 120) there was evidence of a minor contaminant containing  $\beta$  linked xylose (data not shown); <sup>c</sup>The binding signals are means of fluorescence intensities of duplicate spots at 5 fmol of probe arrayed (the respective standard deviation was calculated as the associated error, overall < 5%). '-' refers to a fluorescence intensity < 500.

**Table S3.8. Fluorescence binding intensities elicited with the CBMs from *R. flavefaciens* FD-1 investigated in the Glucan, hemicellulose, chitin and chitosan NGL-microarrays.** The numerical scores for the fluorescence binding signals are shown as means of duplicate spots at 5 fmol probe per spot (as in Figure 3.5) and are representative of at least 2 independent experiments.

Family	4			6	13			22										35				
ID <sup>a</sup>	3259	776	3995	3747	2115	2326	694	1878	2649	1615	1272	2646	3180	2002	1737	3077	3077	3190	1766	933	2302	
1	- <sup>c</sup>	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
4	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
5	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
6	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
7	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
8	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
9	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
10	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
11	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
12	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
13	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
14	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
15	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
16	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
17	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
18	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
19	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
20	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
21	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
22	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
23	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
24	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
25	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
26	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
27	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
28	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
29	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
30	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Family	4			6	13			22										35				
ID <sup>a</sup>	3259	776	3995	3747	2115	2326	694	1878	2649	1615	1272	2646	3180	2002	1737	3077	3077	3190	1766	933	2302	
31	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
32	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
33	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
34	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
35	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
36	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
37	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
38	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
39	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
40	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
41	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
42	1156	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
43	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
44	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
45	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
46	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
47	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
48	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
49	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
50	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
51	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
52	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
53	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
54	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
55	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
56	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
57	-	-	-	-	-	-	-	-	694	-	-	-	-	-	-	-	-	-	-	-	-	-
58	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
59	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
60	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
61	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
62	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
63	2213	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

CHAPTER 3. SUPPLEMENTARY INFORMATION

Family	4			6	13			22											35			
ID <sup>a</sup>	3259	776	3995	3747	2115	2326	694	1878	2649	1615	1272	2646	3180	2002	1737	3077	3077	3190	1766	933	2302	
64	16321	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
65	20790	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
66	23469	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
67	22443	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
68	9723	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
69	37919	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	5761	-	-	-	-
70	46394	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
71	34613	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
72	55639	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
73	36939	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
74	59075	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
75	20070	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
76	14615	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
77	10471	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
78	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
79	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
80	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
81	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
82	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
83	-	1668	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
84	-	3860	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
85	-	8607	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
86	-	3491	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
87	-	7323	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
88	-	10372	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
89	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
90	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
91	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
92	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
93	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
94	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
95	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
96	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
97	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
98	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Family	4			6	13			22											35			
ID <sup>a</sup>	3259	776	3995	3747	2115	2326	694	1878	2649	1615	1272	2646	3180	2002	1737	3077	3077	3190	1766	933	2302	
99	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
100	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
101	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
102	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
103	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
104	1908	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
105	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
106	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
107	387	702	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
108	5388	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
109	428	1430	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
110	5770	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
111	3546	9413	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
112	3817	10345	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
113	7087	35257	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
114	3354	22985	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
115	6786	22067	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
116	15158	47076	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
117	8409	34402	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
118	16616	37301	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
119	10172	27038	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
120 <sup>b</sup>	13023	51244	16586	-	-	-	-	21962	33348	16937	19672	11739	11275	10484	10944	12207	12255	11387	11640	-	-	
121	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
122	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
123	670	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
124	863	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
125	5114	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
126	9436	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
127	3337	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
128	4169	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
129	4241	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
130	9633	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
131	6117	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
132	3253	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
133	4900	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

CHAPTER 3. SUPPLEMENTARY INFORMATION

Family	4			6	13			22											35			
ID <sup>a</sup>	3259	776	3995	3747	2115	2326	694	1878	2649	1615	1272	2646	3180	2002	1737	3077	3077	3190	1766	933	2302	
134	10127	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
135	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
136	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
137	4373	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
138	10073	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
139	11953	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
140	25145	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
141	43833	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
142	28156	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
143	25392	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
144	13665	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
145	29405	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
146	52988	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
147	22482	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
148	56868	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
149	56793	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
150	25453	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
151	27752	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
152	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
153	33902	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
154	-	-	-	2033	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
155	-	-	-	2118	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
156	-	-	-	1856	-	-	-	-	-	-	-	-	-	-	-	-	1657	678	1996	5294	-	-
157	-	-	-	787	-	-	-	-	-	-	-	-	-	-	-	-	6680	2919	3538	6966	-	-
158	-	-	-	3878	-	-	-	-	-	-	-	-	-	-	-	928	3098	569	6197	14912	-	-
159	-	-	-	2100	-	-	-	-	-	-	-	630	1359	1555	4143	13385	3733	17705	24536	-	-	
160	-	-	-	19126	-	-	-	-	679	-	-	2846	4310	6129	13380	30927	9878	26693	46816	-	-	
161	-	-	5249	-	-	-	-	-	4210	1860	6411	4808	6575	9601	13379	25837	17918	12210	36047	-	-	
162	-	-	11348	4586	-	1081	-	786	7030	5711	14697	11183	19977	7638	21644	29436	11783	15860	44346	-	-	
163	-	-	11673	2215	-	-	-	7647	9963	10548	15858	18405	16229	13969	27786	31246	24597	25559	36414	-	-	
164	-	-	11803	7564	-	-	-	3330	12472	10960	11935	21570	18769	17901	24389	32678	22986	30728	46753	-	-	
165	-	-	-	-	1156	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	21849	-
166	-	-	-	32899	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	5314	-
167	-	-	-	25483	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	14434	-
168	-	-	-	-	1091	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	19044	-

Family	4			6	13			22											35				
	ID <sup>a</sup>	3259	776	3995	3747	2115	2326	694	1878	2649	1615	1272	2646	3180	2002	1737	3077	3077	3190	1766	933	2302	
169	-	-	-	1811	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	21670	-
170	-	-	-	-	1921	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	26958	-
171	-	-	-	-	1214	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	32227	-
172	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1809	-
173	-	-	-	-	2199	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	6156	-
174	-	-	-	-	3292	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	8544	-
175	-	-	-	-	11188	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	9881	-
176	-	-	-	-	31115	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	18556	-
177	-	-	-	-	39510	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	18289	-
178	-	-	-	-	27414	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	8893	-
179	-	-	-	-	24948	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	8958	-
180	-	-	-	-	2168	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	20915	-
181	-	-	-	-	6917	7403	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	28216	-
182	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	9417	-
183	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	8992	-
184	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	24731	-
185	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	25680	-
186	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
187	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
188	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2774	-
189	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2117	-
190	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	4183	-
191	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	15664	-
192	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	5456	-
193	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
194	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
195	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
196	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2101	-
197	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	47842	-
198	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	32147	-
199	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1136	38079
200	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
201	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
202	912	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
203	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Family	4			6	13			22											35			
ID <sup>a</sup>	3259	776	3995	3747	2115	2326	694	1878	2649	1615	1272	2646	3180	2002	1737	3077	3077	3190	1766	933	2302	
204	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
205	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
206	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
207	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
208	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
209	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
210	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
211	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
212	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
213	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
214	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
215	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
216	-	-	-	-	-	-	1893	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
217	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
218	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

<sup>a</sup>ID, Probe position in the microarray matching the position in the heatmap, binding-charts and in Table S2.1 in Chapter 2; <sup>b</sup>In the  $\beta$ 1,3-1,4-linked barley series, DP-16 fraction (probe 120) there was evidence of a minor contaminant containing  $\beta$  linked xylose (data not shown); <sup>c</sup>The binding signals are means of fluorescence intensities of duplicate spots at 5 fmol of probe arrayed (the respective standard deviation was calculated as the associated error, overall < 5%). '-' refers to a fluorescence intensity < 500.



Table S3.9. List of polysaccharide samples, their major sequences or monosaccharide composition and sources included in the *Pectin polysaccharide set 1*.

ID <sup>a</sup>	Probe	Source <sup>b</sup>	Predominant monosaccharide composition	Reference <sup>c</sup>
1	PGA (Citrus)	Citrus pectin (sodium salt) Megazyme (P-PGACT)	Linear $\alpha$ 1,4 GalA; Ara (0.81%), Rha (0.47%), Fuc (0.01%) Xyl (0.09%), Gal (4.51%), Glc (3.03%), GalA (88.6%)	-
2	Galacturonate LM (Apple)	<i>Pyrus malus</i> (low methylated, sodium salt) OligoTech (GAT100)	Main linear chain $\alpha$ 1,4-linked GalA and $\alpha$ 1,2-linked Rha units w/side chains of neutral sugars	-
3	Galacturonate LM (Citrus)	<i>Citrus aurantifolia</i> peel (low methylated, sodium salt) OligoTech (GAT102)	Main linear chain $\alpha$ 1,4-linked GalA and $\alpha$ 1,2-linked Rha units w/side chains of neutral sugars: Ara (0.21%), Rha (0.42%), Fuc (0.01%) Xyl (1.5%), Gal (3.67%), Glc (0.94%), UA (92.6%)	-
4	Pectic Galactan (Lupin)	Lupin seed fiber Megazyme (P-PGALU)	Linear $\beta$ 1,4 Gal; Gal: Ara: Rha: Xyl: GalA = 77: 14: 3: 0.6: 5.4	-
5	Pectic Galactan (Potato)	Potato fiber Megazyme (P-PGAPT)	Linear $\beta$ 1,4 Gal; Gal: Ara: Rha: GalA = 78: 9: 4: 9	-
6	Galactan (Lupin)	Lupin seed Megazyme (P-GALLU)	Linear $\beta$ 1,4 Gal; Gal:Ara:Rha:Xyl:other sugars = 82 : 5.8 : 5.1 : 1.4 : 5.7, GalA 14.6%.	-
7	Rhamnogalacturonan (Soybean)	Soybean Megazyme (P-RHAGN)	Main linear chain $\alpha$ 1,4-linked GalA and $\alpha$ 1,2-linked Rha units	-
8	50WSnFI-S2 ( <i>S. nigra</i> )	<i>Sambucus nigra</i> Berit S. Paulsen	Ara (28%), Rha (4.4%), Xyl (1.3%), Man (1%), Gal (19.2%), Glc (2%), GlcA (0.4%), GalA (42.3%), 4-O-Me-GlcA (1.4%)	Ho <i>et al.</i> 2016a <sup>278</sup>
9	100WSnFI-S2 ( <i>S. nigra</i> )	<i>Sambucus nigra</i> Berit S. Paulsen	Ara (18.2%), Rha (16.8%), Xyl (3.4%), Man (0.5%), Gal (17.8%), Glc (2.8%), GlcA (0.3%), GalA (40.5%), 4-O-Me-GlcA (0.9%)	Ho <i>et al.</i> 2016a <sup>278</sup>
10	50WSnFI-S2-EI ( <i>S. nigra</i> )	<i>Sambucus nigra</i> Berit S. Paulsen	Ara (29.5%), Rha (14.3%), Fuc (0.4%), Xyl (1.8%), Man (2.0%), Gal (25.5%), Glc (3.6%), GlcA (2.3%), GalA (17.9%), 4-O-Me-GlcA (2.7%)	Ho <i>et al.</i> 2016b <sup>279</sup>
11	SnFI50-S2 ( <i>S. nigra</i> )	<i>Sambucus nigra</i> Berit S. Paulsen	Ara (19.4%), Rha (5.3%), Xyl (0.7%), Man (1.1%), Gal (22.9%), Glc (2.8%), GlcA (2.1%), GalA (44.7%), 4-O-Me-GlcA (1%)	Ho <i>et al.</i> 2016a <sup>278</sup>
12	IOI-WAc ( <i>I. obliquus</i> )	<i>Inonotus obliquus</i> Berit S. Paulsen	-	-
13	IOI-WN ( <i>I. obliquus</i> )	<i>Inonotus obliquus</i> Berit S. Paulsen	-	-
14	BP-II ( <i>B. petersianum</i> )	<i>Biophytum petersianum</i> Berit S. Paulsen	Ara (5.1%), Rha (8.2%), Fuc (0.5%), 2-Me-Fuc (trace), Xyl (6.3%), 2-Me-Xyl (trace), Man (0.7%), Gal (8.3%), Glc (4.4%), GlcA (1.3%), GalA (65.1%)	Grønhaug <i>et al.</i> 2011 <sup>280</sup>
15	GOA1 ( <i>G. oppositifolius</i> )	<i>Glinus oppositifolius</i> Berit S. Paulsen	Ara (26.4%), Rha (4.2%), Xyl (3.9%), Man (4.3%), Gal (42.9%), Glc (3.5%), GalA (12.1%), 4-O-Me-GlcA (2.9%)	Inngjerdigen <i>et al.</i> 2005 <sup>281</sup>
16	GOA2 ( <i>G. oppositifolius</i> )	<i>Glinus oppositifolius</i> Berit S. Paulsen	Ara (5.5%), Rha (10.3%), Fuc (1.3%), Xyl (0.5%), Man (0.6%), Gal (9.7%), Glc (3.3%), GalA (68.3%), 4-O-Me-GlcA (0.4%)	Inngjerdigen <i>et al.</i> 2005 <sup>281</sup>
17	Vk100-Fr.I ( <i>V. kotschyana</i> )	<i>Vernonia kotschyana</i> Berit S. Paulsen	Ara (2%), Rha (1%), Fru (83%), Gal (2%), Glc (3%), GalA (1%)	Nergard <i>et al.</i> 2004 <sup>282</sup>
18	Ctw-A1 ( <i>C. tinctorium</i> )	<i>Cochlospermum tinctorium</i> Berit S. Paulsen	Ara (16.3%), Rha (17.9%), Man (1.8%), Gal (45.8%), Glc (4%), GlcA (8.8%), GalA (5.8%), Fru (4.9%)	Inngjerdigen <i>et al.</i> 2013 <sup>283</sup>

19	Oc50A1.IA ( <i>O. celtidifolia</i> )	<i>Opilia celtidifolia</i> Berit S. Paulsen	Ara (38.9%), Rha (4.2%), Man (5.8%), Gal (30.9%), Glc (5.4%), GlcA (trace), GalA (11.5%), 4-O-Me-GlcA (3.3%)	Grønhaug <i>et al.</i> 2010 <sup>284</sup>
20	LPS3 ( <i>T. cordata</i> )	<i>Tilia cordata</i> Berit S. Paulsen	-	-
21	LCC ( <i>C. cordifolia</i> )	<i>Cola cordifolia</i> bark Berit S. Paulsen	-	-
22	CC1P1 ( <i>C. cordifolia</i> )	<i>Cola cordifolia</i> bark Berit S. Paulsen	Ara (trace), Rha (32%), Gal (31%), Glc (2%), GalA (35%)	Austarheim <i>et al.</i> 2012 <sup>285</sup>
23	CC1 ( <i>C. cordifolia</i> )	<i>Cola cordifolia</i> bark Berit S. Paulsen	Ara (3.7%), Rha (22.1%), Gal (20.2%), Glc (0.5%), GalA (29.6%), 2-O-Me-Gal (6.5%), 4-O-Me-GlcA (17.4%)	Austarheim <i>et al.</i> 2012 <sup>285</sup>
24	CC2 ( <i>C. cordifolia</i> )	<i>Cola cordifolia</i> bark Berit S. Paulsen	Ara (37.2%), Rha (8.5%), Gal (31.3%), Glc (1.1%), GalA (11.5%), GlcA (3.4%), 2-O-Me-Gal (0.4%), 4-O-Me-GlcA (6.6%)	Austarheim <i>et al.</i> 2012 <sup>285</sup>
25	CC3 ( <i>C. cordifolia</i> )	<i>Cola cordifolia</i> bark Berit S. Paulsen	Ara (3.0%), Rha (22.8%), Gal (17.3%), Glc (1%), GalA (32.8%), 4-O-Me-GlcA (17.8%), 2-O-Me-Gal (5.3%)	Austarheim <i>et al.</i> 2012 <sup>285</sup>
26	PBS100-II ( <i>P. biglobosa</i> )	<i>Parkia biglobosa</i> Berit S. Paulsen	Ara (21.2%), Rha (7.3%), Xyl (0.2%), Gal (18%), Glc (6.1%), GalA (30.1%), GlcA (10.5%), 4-O-Me-GlcA (1.3%)	Zou <i>et al.</i> 2014 <sup>286</sup>

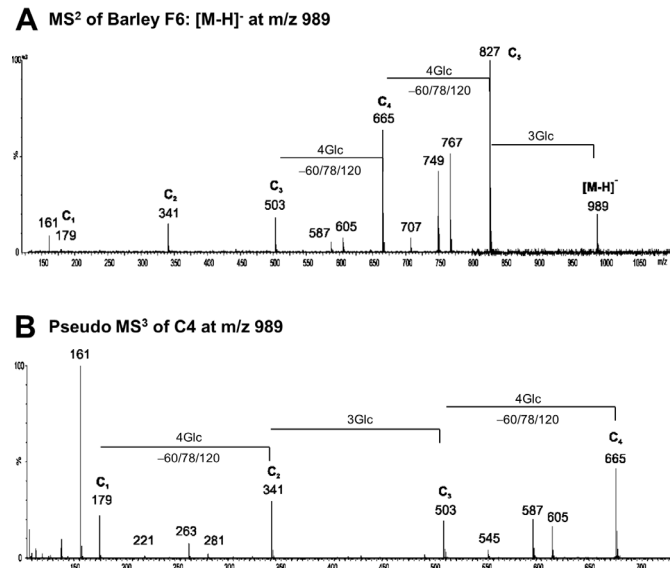
<sup>a</sup> Probes are grouped according to predominant oligosaccharide sequence and glycosidic linkage. The ID corresponds to the positions in the binding heatmap;

<sup>b</sup> The sources are indicated for each carbohydrate sample; if commercial the code product number is indicated;

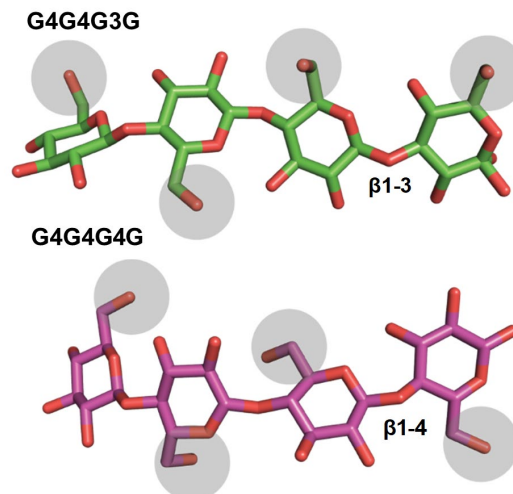
<sup>c</sup> References for the structural analysis or recent published work for each particular sample, if available.

## Chapter 4 - Supplementary Information

## Supplementary Figures



**Figure S4.1.** MS/MS sequence analysis of barley hexasaccharide. Negative-ion ESI-CID-MS/MS product-ion spectrum (A) MS2 and (B) quasi MS3. The characteristic 3- and 4-linkage diagnostic ions assigned in the spectra are as described in Palma *et al.* 2015<sup>32</sup>.



**Figure S4.2.** Crystal structures of bound G4G4G3G (top - PDB ID 6R3M) and G4G4G4G (bottom - PDB ID 3AMM<sup>287</sup>). The presence of the  $\beta$ 1,3 bond causes a rotation of the consecutive glucose unit of about 180°, positioning the hydroxymethylene group (transparent grey circles) in the opposite direction.

## Supplementary Tables

Table S4.1. Privateer validation results for G4G4G3G and G4G3G4G4G3G ligands bound to CfCBM11.

Residue Name	Conformation	Average B-factor	RSCC <sup>1</sup>	Diagnostic
<b>CfCBM11-G4G4G3G</b>				
BGC	<sup>4</sup> C <sub>1</sub>	21.85	0.75	Ok
BGC	<sup>4</sup> C <sub>1</sub>	15.87	0.92	Ok
BGC	<sup>4</sup> C <sub>1</sub>	15.62	0.89	Ok
BGC	<sup>4</sup> C <sub>1</sub>	20.79	0.84	Ok
<b>CfCBM11-G4G3G4G4G3G</b>				
BGC	<sup>4</sup> C <sub>1</sub>	22.02	0.76	Ok
BGC	<sup>4</sup> C <sub>1</sub>	21.14	0.82	Ok
BGC	<sup>4</sup> C <sub>1</sub>	21.14	0.84	Ok
BGC	<sup>4</sup> C <sub>1</sub>	22.22	0.70	Ok
BGC	<sup>4</sup> C <sub>1</sub>	23.31	0.71	Ok
BGC	<sup>4</sup> C <sub>1</sub>	23.58	0.73	Ok

<sup>1</sup>RSCC, Real Space Correlation Coefficient, measures the agreement between model and positive omit density.

Table S4.2. List of protein-ligand contacts for CfCBM11-G4G4G3G structure.

Residue	Direct hydrogen bonds	d(Å)	Water-mediated hydrogen bonds	d(Å)	CH- $\pi$ stacking/ Hydrophobic interactions	d(Å)
Asp51			COO <sup>-</sup> ↔OH <sub>2</sub> (W356)↔OH (C4) Glc 4	3.1; 2.7		
Glu25			COO <sup>-</sup> ↔OH <sub>2</sub> (W310)↔OH (C2) Glc 4	2.5; 2.9		
Tyr152	COO <sup>-</sup> ↔OH (C6) Glc 4	2.9				
	COO <sup>-</sup> ↔OH (C3) Glc 3	3.1				
Arg126	NH <sub>2</sub> ↔OH (C3) Glc 3	2.9				
	NH <sub>2</sub> ↔OH (C2) Glc 3	3.0				
Tyr129			COO <sup>-</sup> ↔OH <sub>2</sub> (W336)↔OH (C6) Glc 3	2.7; 2.7	Arom. ring↔Glc 2	4.3
Tyr22			COO <sup>-</sup> ↔OH <sub>2</sub> (W318)↔OH (C2) Glc 3	2.8; 2.6	Arom. ring↔Glc 3	4.1
Tyr53			COO <sup>-</sup> ↔OH <sub>2</sub> (W425)↔OH (C2) Glc 2	2.9; 3.1	Arom. ring↔Glc 2	4.2
			COO <sup>-</sup> ↔OH <sub>2</sub> (W425)↔OH (C3) Glc 2	2.9; 2.9		
Asp99	COO <sup>-</sup> ↔OH (C6) Glc 2	2.7	COO <sup>-</sup> ↔OH <sub>2</sub> (W314)↔OH (C4) Glc 1	2.6; 2.5		
His102					Arom. ring↔Glc 2	4.2
His149			NH↔OH <sub>2</sub> (W314)↔OH (C4) Glc 1	3.1; 2.5		
Ser147			OH↔OH <sub>2</sub> (W314)↔OH (C4) Glc 1	2.9; 2.5		
Asp146	COO <sup>-</sup> ↔OH (C6) Glc 1	2.7	COO <sup>-</sup> ↔OH <sub>2</sub> (W448)↔OH (C6) Glc 1	3.0; 3.0		
Ser59			OH↔OH <sub>2</sub> (W368)↔OH (C1) Glc 1	2.8; 2.8		

Table S4.3. List of protein-ligand contacts for CfCBM11-G4G3G4G4G3G structure.

Residue	Direct hydrogen bonds	d(Å)	CH- $\pi$ stacking	d(Å)
Glu25	COO <sup>-</sup> ↔OH (C6) Glc 5	3.0		
Arg126	NH <sub>2</sub> ↔OH (C2) Glc 3	3.1		
	NH <sub>2</sub> ↔OH (C3) Glc 3	3.1		
Tyr129			Arom. ring↔Glc 3	4.2
Tyr53			Arom. ring↔Glc 3	4.2
Tyr22			Arom. ring↔Glc 2	4.1
Asp99	COO <sup>-</sup> ↔OH (C6) Glc 2	3.0		
His102			Arom. ring↔Glc 2	4.6

**Table S4.4. Primers used to generate the CtCBM11 mutant derivatives.** Mutation points are depicted in bold.

<b>Mutants</b>	<b>Sequence (5' - 3')</b>	<b>Direction</b>
<b>Asp99Ala</b>	gcataaacggtgtgggagccgggagaacactgg	Forward
	ccagtgttctccggctcccacaccgttatgc	Reverse
<b>Arg126Ala</b>	ctccagctttagaagagcacttgattatcagccgc	Forward
	gccgctgataatcaagtgctctctaaagctggag	Reverse
<b>Asp146Ala</b>	ggatcttgacaatatagcttcaattcacttcatgtatgcc	Forward
	ggcatacatgaagtgaattgaagctatattgtcaagatcc	Reverse
<b>Val57Ala</b>	ctggggaacagtatacgtttaccggacggcgat	Forward
	atcggcgctccggtaaagcgatactgttcccag	Reverse
<b>Ser59Ala</b>	ggctactggggaacagcatacagttaccggac	Forward
	gtccggtaaactgtatgctgttcccagtagcc	Reverse
<b>Glu25Ala</b>	ctccggtgcgggtgcaaaagttcaacaaaattg	Forward
	ctttgcaccgcaccggagtatgaacccaatttaa	Reverse
<b>Asp51</b>	gacaacggctggctactggggaacagtatac	Forward
	Cagtagccagccgtgtccgggtgtagctgac	Reverse



## Chapter 5 - Supplementary Information

## Supplementary Figures

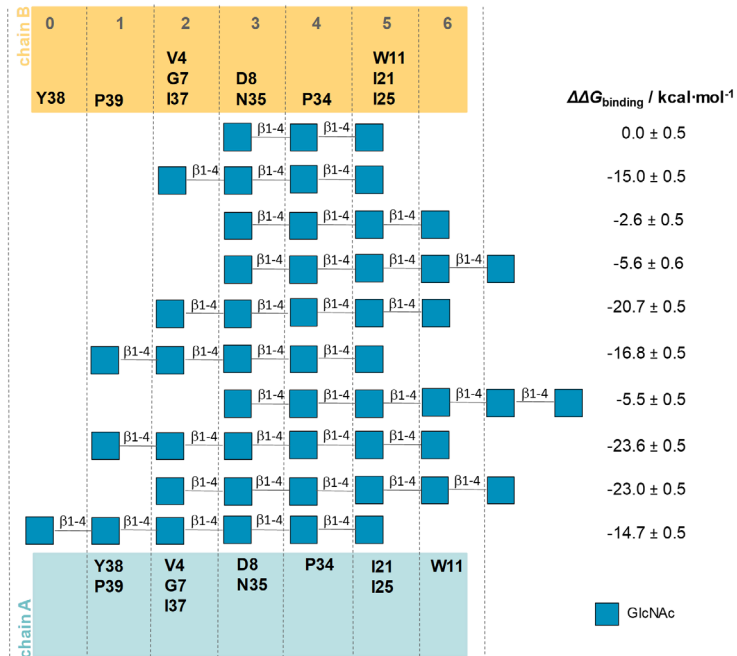


Figure S5.1. Schematic representation of the position of the various GlcNAc ligands into  $CfCBM50_{AB}$ . The respective binding affinities are shown as well.

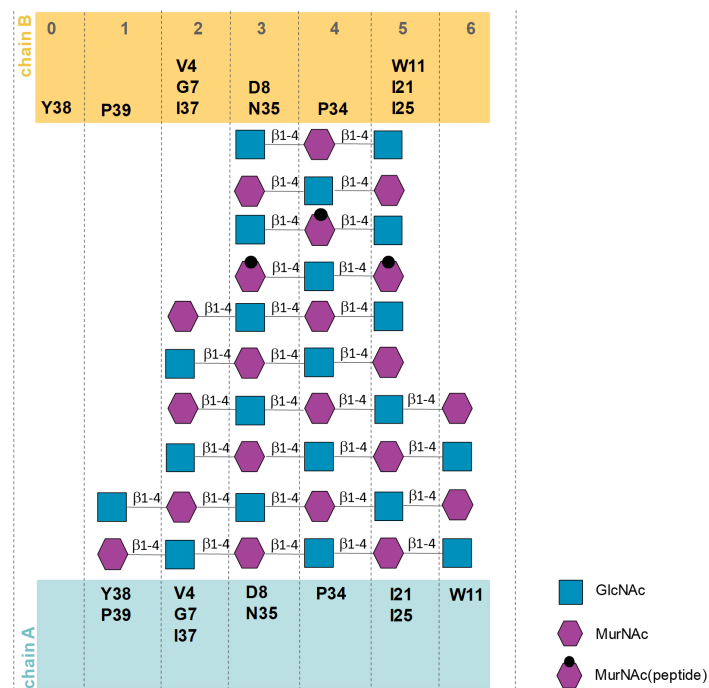
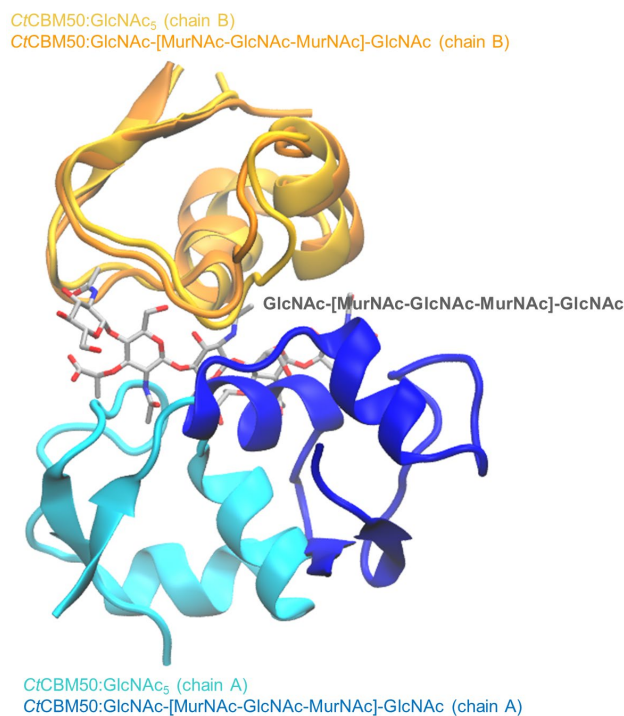
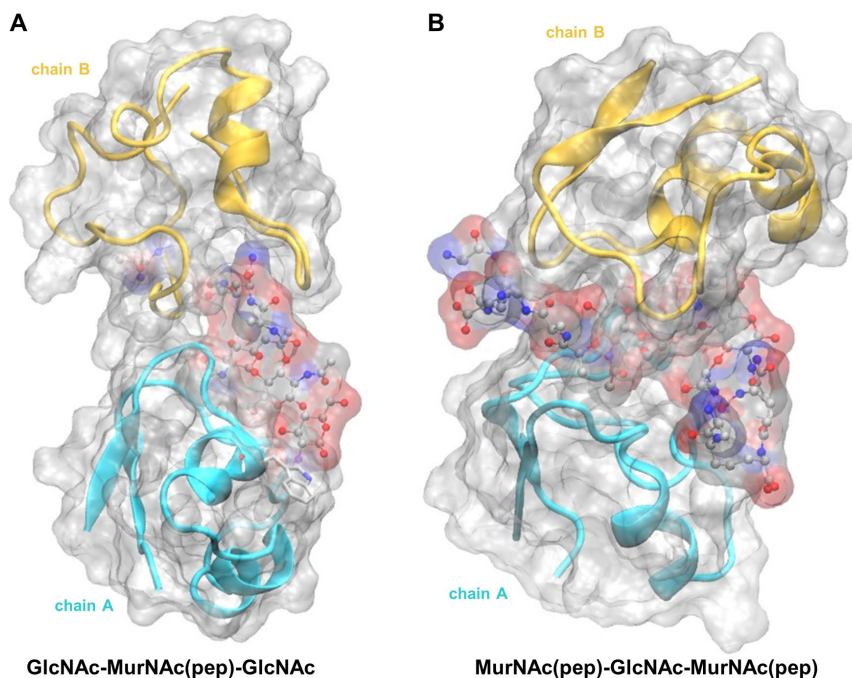


Figure S5.2. Schematic representation of the position of the various peptidoglycan ligands into the binding cleft of  $CfCBM50_{AB}$ .



**Figure S5.3. Representation of CtCBM50<sub>AB</sub> complexed with the two pentasaccharides GlcNAc<sub>5</sub> and GlcNAc-[MurNAc-GlcNAc-MurNAc]-GlcNAc.** Complex geometries are aligned by chain B backbone atoms. To simplify the visualization only the peptidoglycan fragment is represented as sticks and coloured by atom type.



**Figure S5.4. Representation of the CtCBM50:GlcNAc-MurNAc(peptide)-GlcNAc and CtCBM50:MurNAc(peptide)-GlcNAc-MurNAc(peptide) complexes.** The last frame of each simulation is represented. Assembly interface between the chains A (depicted as blue cartoon and silver surface) and B (depicted as yellow cartoon and silver surface). MurNAc(peptide)-GlcNAc fragments are represented as balls-and-sticks and surface and coloured by atom type.



## Supplementary Tables

Table S5.1. Privateer validation results for the  $\beta$ 1,4-linked GlcNAc trisaccharide bound to CtCBM50.

Residue Name	Conformation	Average B-factor	RSCC <sup>1</sup>	Diagnostic
NAG	<sup>4</sup> C <sub>1</sub>	14.90	0.91	Ok
NAG	<sup>4</sup> C <sub>1</sub>	13.56	0.92	Ok
NAG	<sup>4</sup> C <sub>1</sub>	14.45	0.90	Ok

<sup>1</sup>RSCC, Real Space Correlation Coefficient, measures the agreement between model and positive omit density.

Table S5.2. Similarity analysis between protein chains of CtCBM50-GlcNAc<sub>3</sub> structure. Rmsd's between chain pairs is presented. Sequence alignment was performed using PDBePISA<sup>288</sup>.

	A	B	C
A	-	0.70	1.05
B	-	-	0.97
C	-	-	-

Table S5.3. List of protein-ligand contacts for CtCBM50-GlcNAc<sub>3</sub> structure.

Res.	Chain	Direct hydrogen bonds	d(Å)	Water-mediated hydrogen bonds	d(Å)	CH- $\pi$ stacking	d(Å)
Trp11	A	NH <sub>2</sub> ↔ OH (NAc) GlcNAc1	3.0				
	B	NH <sub>2</sub> ↔ OH (NAc) GlcNAc2	2.8			Arom. ring ↔ GlcNAc1	4.5
Asp8	C			COO ↔ OH <sub>2</sub> (W89) ↔ OH (C1) GlcNAc1	2.9; 2.8		
Pro34	A			COO ↔ OH <sub>2</sub> (W80) ↔ OH (C3) GlcNAc1	2.8; 2.8		
	B			COO ↔ OH <sub>2</sub> (W81) ↔ OH (C3) GlcNAc2	2.8; 2.8		
Met10	A			NH ↔ OH <sub>2</sub> (W80) ↔ OH (C3) GlcNAc1	2.9; 2.8		
	B			NH ↔ OH <sub>2</sub> (W81) ↔ OH (C3) GlcNAc2	2.9; 2.8		
Thr9	B			OH ↔ OH <sub>2</sub> (W83) ↔ OH (C4) GlcNAc1	2.8; 2.9		
Asn35	A	COO ↔ NH (NAc) GlcNAc1	2.8				
	B	CO ↔ NH (NAc) GlcNAc2	2.9				
		COO ↔ OH (C3) GlcNAc2	2.8	CO ↔ OH <sub>2</sub> (W28) ↔ OH (C6) GlcNAc1	2.9; 2.9		
		COO ↔ OH (C6) GlcNAc3	2.9	COO ↔ OH <sub>2</sub> (W48) ↔ OH (NAc) GlcNAc3	2.9; 3.0		
Ile37	A	NH <sub>2</sub> ↔ OH (C6) GlcNAc2	3.0				
	B	COO ↔ NH (NAc) GlcNAc3	2.9				
Gly7	A	NH <sub>2</sub> ↔ OH (C6) GlcNAc2	2.9	COO ↔ OH <sub>2</sub> (W16) ↔ OH (C4) GlcNAc3	2.7; 2.7		
	C	NH <sub>2</sub> ↔ OH (NAc) GlcNAc3	2.8				
		NH <sub>2</sub> ↔ OH (C1) GlcNAc1	3.0				

Table S5.4. Hydrogen bonds involving the GlcNAc<sub>3</sub> to GlcNAc<sub>6</sub>, and that are present during more than 15% of the MD simulation. CO – carbonyl group; NH – amine group; SC – side chain.

Carbohydrate interactions in the CtCBM50:GlcNAc <sub>n</sub> system		Distance (Å)	%
<b>CtCBM50:GlcNAc<sub>3</sub></b>			
GlcNAc 1 (NAc)	I37 - A (CO)	2.9	61.0
GlcNAc 1 (NAc)	G7 - A (NH)	2.9	38.0
GlcNAc 1 (HO-C6)	N35 - B (CO)	2.8	24.0
GlcNAc 1 (HO-C6)	D8 - B (CO)	2.8	20.0
GlcNAc 1 (HO-C6)	I37 - B (NH)	2.9	20.0
GlcNAc 2 (NAc)	W11 - B (NH)	2.9	59.0
GlcNAc 2 (HO-C6)	N35 - A (CO)	2.8	52.0
GlcNAc 2 (HO-C6)	I37 - A (NH)	2.9	33.0
GlcNAc 2 (HO-C3)	N35 - B (CO)	2.8	17.0
GlcNAc 3 (NAc)	N35 - A (SC)	2.9	47.0
GlcNAc 3 (NAc)	W11 - A (NH)	2.9	39.0
<b>CtCBM50:GlcNAc<sub>4</sub></b>			
GlcNAc 1 (NAc)	I37 - B (CO)	2.9	73.0
GlcNAc 1 (NAc)	G7 - B (NH)	2.9	27.0

GlcNAc 2 (NAc)	I37 - A (CO)	2.9	45.0
GlcNAc 2 (NAc)	G7 - A (NH)	2.9	43.0
GlcNAc 2 (HO-C6)	I37 - B (NH)	2.9	41.0
GlcNAc 2 (HO-C6)	N35 - B (CO)	2.8	32.0
GlcNAc 2 (HO-C6)	D8 - B (CO)	2.7	19.0
GlcNAc 3 (HO-C6)	N35 - A (CO)	2.8	67.0
GlcNAc 3 (NAc)	W11 - B (NH)	2.9	63.0
GlcNAc 3 (HO-C6)	I37 - A (NH)	2.9	32.0
GlcNAc 3 (HO-C3)	N35 - B (CO)	2.8	17.0
GlcNAc 4 (NAc)	N35 - A (SC)	2.9	62.0
GlcNAc 4 (NAc)	W11 - A (NH)	2.9	33.0
<b>CtCBM50:GlcNAc<sub>5</sub></b>			
GlcNAc 1 (NAc)	I37 - B (CO)	2.9	29.0
GlcNAc 1 (NAc)	G7 - B (NH)	2.9	29.0
GlcNAc 2 (NAc)	I37 - A (CO)	2.9	59.0
GlcNAc 2 (HO-C6)	I37 - B (NH)	2.9	44.0
GlcNAc 2 (NAc)	G7 - A (NH)	2.9	36.0
GlcNAc 2 (HO-C6)	N35 - B (CO)	2.8	35.0
GlcNAc 2 (HO-C6)	D8 - B (CO)	2.7	15.0
GlcNAc 3 (NAc)	W11 - B (NH)	2.9	59.0
GlcNAc 3 (HO-C6)	N35 - A (CO)	2.8	53.0
GlcNAc 3 (NAc)	N35 - B (SC)	2.9	28.0
GlcNAc 3 (HO-C3)	N35 - B (CO)	2.8	16.0
GlcNAc 4 (NAc)	W11 - A (NH)	2.9	51.0
GlcNAc 4 (NAc)	N35 - A (SC)	2.9	21.0
<b>CtCBM50:GlcNAc<sub>6</sub></b>			
GlcNAc 1 (NAc)	Y38 - B (SC)	2.8	27.0
GlcNAc 2 (NAc)	I37 - B (CO)	2.9	74.0
GlcNAc 3 (HO-C6)	I37 - B (NH)	2.9	52.0
GlcNAc 3 (HO-C6)	N35 - B (CO)	2.8	41.0
GlcNAc 3 (NAc)	I37 - A (CO)	2.9	40.0
GlcNAc 3 (NAc)	G7 - A (NH)	2.9	38.0
GlcNAc 4 (HO-C6)	N35 - A (CO)	2.8	68.0
GlcNAc 4 (NAc)	W11 - B (NH)	2.9	58.0
GlcNAc 4 (HO-C6)	I37 - A (NH)	2.9	30.0
GlcNAc 4 (HO-C3)	N35 - B (CO)	2.8	17.0
GlcNAc 4 (NAc)	N35 - B (SC)	2.9	16.0
GlcNAc 5 (NAc)	N35 - A (SC)	2.9	55.0
GlcNAc 5 (NAc)	W11 - A (NH)	2.9	49.0

Table S5.5. Relative enthalpy energies of the individual CtCBM50 chains A and B complexed with all GlcNAc ligands.

	MD simulation	$\Delta\Delta H_{\text{binding}}$ (kcal·mol <sup>-1</sup> )
CtCBM50:GlcNAc <sub>3</sub>	chain A	0.0 ± 0.4
	chain B	12.2 ± 0.4
CtCBM50:GlcNAc <sub>4</sub>	chain A	0.0 ± 0.4
	chain B	0.4 ± 0.4
CtCBM50:GlcNAc <sub>5</sub>	chain A	0.0 ± 0.5
	chain B	1.6 ± 0.5
CtCBM50:GlcNAc <sub>6</sub>	chain A	0.0 ± 0.5
	chain B	2.0 ± 0.5

Table S5.6. Relative enthalpy and binding free energies of the individual C<sub>t</sub>CBM50 chains complexed with all GlcNAc ligands.

MD simulation	$\Delta\Delta H_{\text{binding}}$ (kcal·mol <sup>-1</sup> )	$\Delta\Delta G_{\text{binding}}$ (kcal·mol <sup>-1</sup> )
C <sub>t</sub> CBM50-chain A:GlcNAc <sub>3</sub>	0.0 ± 0.4	0.0 ± 0.1
C <sub>t</sub> CBM50-chain B:GlcNAc <sub>3</sub>	11.8 ± 0.5	9.4 ± 0.1
C <sub>t</sub> CBM50-chain A:GlcNAc <sub>4</sub>	0.0 ± 0.5	0.0 ± 0.1
C <sub>t</sub> CBM50-chain B:GlcNAc <sub>4</sub>	-4.0 ± 0.6	-3.3 ± 0.1
C <sub>t</sub> CBM50-chain A:GlcNAc <sub>5</sub>	0.0 ± 0.5	0.0 ± 0.1
C <sub>t</sub> CBM50-chain B:GlcNAc <sub>5</sub>	0.7 ± 0.5	2.1 ± 0.1
C <sub>t</sub> CBM50-chain A:GlcNAc <sub>6</sub>	0.0 ± 0.6	0.0 ± 0.1
C <sub>t</sub> CBM50-chain B:GlcNAc <sub>6</sub>	-1.8 ± 0.6	0.9 ± 0.1

Table S5.7. Relative enthalpy and binding free energies of the complexes with all MurNAc-GlcNAc ligands in relation to the C<sub>t</sub>CBM50:MurNAcGlcNAc<sub>3</sub> complex.

Carbohydrate	$\Delta\Delta H_{\text{binding}}$ (kcal·mol <sup>-1</sup> )	$\Delta\Delta G_{\text{binding}}$ (kcal·mol <sup>-1</sup> )
[MurNAc-GlcNAc-MurNAc]	0.0 ± 1.0	0.0 ± 0.1
GlcNAc-[MurNAc-GlcNAc-MurNAc]	-33.9 ± 1.0	-28.3 ± 0.1
GlcNAc-[MurNAc-GlcNAc-MurNAc]-GlcNAc	-37.7 ± 1.1	-28.7 ± 0.1
MurNAc-GlcNAc-[MurNAc-GlcNAc-MurNAc]-GlcNAc	-39.2 ± 1.0	-29.1 ± 0.1
[GlcNAc-MurNAc-GlcNAc]	0.0 ± 0.7	0.0 ± 0.1
MurNAc-[GlcNAc-MurNAc-GlcNAc]	-19.6 ± 0.8	-16.1 ± 0.1
MurNAc-[GlcNAc-MurNAc-GlcNAc]-MurNAc	-31.3 ± 0.9	-22.1 ± 0.1
GlcNAc-MurNAc-[GlcNAc-MurNAc-GlcNAc]-MurNAc	-21.2 ± 0.8	-16.7 ± 0.1

Table S5.8. Hydrogen bonds involving the [GlcNAc-MurNAc]<sub>n</sub> ligands tested, and that are present during more than 15% of the MD simulation. CO – carbonyl group; NH – amine group; SC – side chain.

Carbohydrate interactions in the C <sub>t</sub> CBM50:[GlcNAc-MurNAc] <sub>n</sub> system		Distance (Å)	%
<b>C<sub>t</sub>CBM50:[GlcNAc-MurNAc-GlcNAc]</b>			
GlcNAc 1 (NAc)	I37 - A (CO)	2.9	66.0
GlcNAc 1 (NAc)	G7 - A (NH)	2.9	41.0
MurNAc 2 (HO-C6)	I37 - A (NH)	2.9	25.0
GlcNAc 3 (HO-C3)	N35 - A (CO)	2.8	75.0
GlcNAc 3 (NAc)	N35 - A (SC)	2.9	52.0
GlcNAc 3 (NAc)	W11 - A (NH)	2.9	16.0
<b>C<sub>t</sub>CBM50:MurNAc-[GlcNAc-MurNAc-GlcNAc]</b>			
MurNAc 1 (NAc)	Y38 - A (SC)	2.8	20.0
MurNAc 1 (NAc)	N35 - B (CO)	2.9	17.0
GlcNAc 2 (NAc)	I37 - A (CO)	2.9	60.0
GlcNAc 2 (NAc)	G7 - A (NH)	2.9	40.0
MurNAc 3 (NAc)	W11 - B (NH)	2.9	60.0
MurNAc 3 (HO-C6)	N35 - A (CO)	2.8	40.0
MurNAc 3 (HO-C6)	I37 - A (NH)	2.9	26.0
MurNAc 3 (Mur)	N35 - B (SC)	2.8	35.0
GlcNAc 4 (NAc)	N35 - A (SC)	2.9	44.0
GlcNAc 4 (NAc)	W11 - A (NH)	2.9	40.0
<b>C<sub>t</sub>CBM50:MurNAc-[GlcNAc-MurNAc-GlcNAc]-MurNAc</b>			
MurNAc 1 (NSc)	G7 - B (NH)	2.9	36.0
MurNAc 1 (NAc)	I37 - B (CO)	2.9	35.0
GlcNAc 2 (HO-C6)	MurNAc 3 (Mur)	2.7	80.0
GlcNAc 2 (NAc)	G7 - A (NH)	2.9	42.0
GlcNAc 2 (NAc)	I37 - A (CO)	2.9	39.0
MurNAc 3 (NAc)	W11 - B (NH)	2.9	60.0
MurNAc 3 (HO-C6)	N35 - A (CO)	2.8	60.0
MurNAc 3 (HO-C6)	I37 - A (NH)	2.9	29.0

GlcNAc 4 (NAc)	N35 - A (SC)	2.9	66.0
GlcNAc 4 (NAc)	W11 - A (NH)	2.9	60.0
GlcNAc 4 (HO-C6)	MurNAc 5 (Mur)	2.7	33.0
<b>CtCBM50:GlcNAc-MurNAc-[GlcNAc-MurNAc-GlcNAc]-MurNAc</b>			
GlcNAc 1 (HO-C6)	MurNAc 2 (Mur)	2.7	31.0
GlcNAc 3 (NAc)	I37 - A (CO)	2.9	64.0
GlcNAc 3 (HO-C6)	MurNAc 4 (Mur)	2.7	63.0
MurNAc 4 (HO-C6)	N35 - A (CO)	2.8	37.0
MurNAc 4 (HO-C6)	I37 - A (NH)	2.9	30.0
MurNAc 4 (Mur)	N35 - B (SC)	2.8	19.0
GlcNAc 5 (NAc)	N35 - A (SC)	2.9	57.0
GlcNAc 5 (NAc)	W11 - A (NH)	2.9	51.0
GlcNAc 5 (HO-C6)	MurNAc 6 (Mur)	2.7	34.0
<b>CtCBM50:[MurNAc-GlcNAc-MurNAc]</b>			
MurNAc 1 (HO-C6)	N35 - B (CO)	2.8	47.0
MurNAc 1 (HO-C6)	I37 - B (NH)	2.9	29.0
GlcNAc 2 (NAc)	W11 - B (NH)	2.9	51.0
GlcNAc 2 (NAc)	AN35 - B (SC)	2.9	45.0
GlcNAc 2 (HO-C6)	MurNAc 3 (Mur)	2.7	43.0
MurNAc 3 (NAc)	N35 - A (SC)	2.9	34.0
<b>CtCBM50:GlcNAc-[MurNAc-GlcNAc-MurNAc]</b>			
GlcNAc 1 (NAc)	I37 - B (CO)	2.9	64.0
GlcNAc 1 (HO-C6)	MurNAc 2 (Mur)	2.7	44.0
GlcNAc 1 (NAc)	G7 - B (NH)	2.9	36.0
MurNAc 2 (NAc)	G7 - A (NH)	2.9	54.0
MurNAc 2 (HO-C6)	N35 - B (CO)	2.8	47.0
MurNAc 2 (HO-C6)	I37 - B (NH)	2.9	37.0
GlcNAc 3 (NAc)	W11 - B (NH)	2.9	59.0
GlcNAc 3 (HO-C6)	MurNAc 4 (Mur)	2.7	55.0
GlcNAc 3 (NAc)	N35 - B (SC)	2.9	44.0
MurNAc 4 (O-C3)	T9 - A(SC)	2.8	17.0
MurNAc 4 (Mur)	T9 -A (SC)	2.7	26.0
<b>CtCBM50:GlcNAc-[MurNAc-GlcNAc-MurNAc]-GlcNAc</b>			
GlcNAc 1 (NAc)	I37 - B (CO)	2.9	64.0
GlcNAc 1 (NAc)	G7 - B (NH)	2.9	42.0
GlcNAc 1 (HO-C6)	MurNAc 2 (Mur)	2.7	26.0
MurNAc 2 (HO-C6)	N35 - B (CO)	2.8	49.0
MurNAc 2 (HO-C6)	I37 - B (NH)	2.9	39.0
GlcNAc 3 (NAc)	N35 - B (SC)	2.9	53.0
GlcNAc 3 (NAc)	W11 - B (NH)	2.9	46.0
GlcNAc 3 (HO-C6)	MurNAc 4 (Mur)	2.7	33.0
GlcNAc 3 (HO-C3)	N35 - B (CO)	2.8	15.0
MurNAc 4 (Mur)	T9 - A (SC)	2.8	26.0
MurNAc 4 (NAc)	W11 - A (NH)	2.9	22.0
MurNAc 4 (Mur)	N35 - A (SC)	2.8	23.0
<b>CtCBM50:MurNAc-GlcNAc-[MurNAc-GlcNAc-MurNAc]-GlcNAc system</b>			
MurNAc 1 (NAc)	Y38 - B (SC)	2.8	20.0
GlcNAc 2 (NAc)	I37 - B (CO)	2.9	74.0
GlcNAc 2 (NAc)	G7 - B (NH)	2.9	24.0
GlcNAc 2 (HO-C6)	MurNAc 3 (Mur)	2.7	43.0
MurNAc 3 (HO-C6)	N35 - B (CO)	2.8	46.0
MurNAc 3 (HO-C6)	I37 - B (NH)	2.9	34.0
GlcNAc 4 (NAc)	W11 - B (NH)	2.9	61.0
GlcNAc 4 (HO-C6)	MurNAc 5 (Mur)	2.7	50.0
GlcNAc 4 (NAc)	N35 - B (SC)	2.9	34.0
MurNAc 5 (Mur)	T9 - A (SC)	0.8	61.0
MurNAc 5 (Mur)	W11 - A (NH)	2.9	15.0

Table S5.9. Relative enthalpy and binding free energies of the individual C<sub>t</sub>CBM50 chains complexed with all MurNAc-GlcNAc ligands.

MD simulation	$\Delta\Delta H_{\text{binding}}$ (kcal·mol <sup>-1</sup> )	$\Delta\Delta G_{\text{binding}}$ (kcal·mol <sup>-1</sup> )
C <sub>t</sub> CBM50-chain A:[MurNAc-GlcNAc-MurNAc]	0.0 ± 1.9	0.0 ± 0.2
C <sub>t</sub> CBM50-chain B:[MurNAc-GlcNAc-MurNAc]	-13.8 ± 1.4	-6.0 ± 0.2
C <sub>t</sub> CBM50-chain A:GlcNAc-[MurNAc-GlcNAc-MurNAc]	0.0 ± 0.4	0.0 ± 0.1
C <sub>t</sub> CBM50-chain B:GlcNAc-[MurNAc-GlcNAc-MurNAc]	-7.0 ± 0.5	-4.7 ± 0.1
C <sub>t</sub> CBM50-chain A:GlcNAc-[MurNAc-GlcNAc-MurNAc]-GlcNAc	0.0 ± 1.5	0.0 ± 0.1
C <sub>t</sub> CBM50-chain B:GlcNAc-[MurNAc-GlcNAc-MurNAc]-GlcNAc	-19.5 ± 1.1	-13.8 ± 0.1
C <sub>t</sub> CBM50-chain A:MurNAc-GlcNAc-[MurNAc-GlcNAc-MurNAc]-GlcNAc	0.0 ± 0.6	0.0 ± 0.1
C <sub>t</sub> CBM50-chain B:MurNAc-GlcNAc-[MurNAc-GlcNAc-MurNAc]-GlcNAc	-26.1 ± 0.6	-18.7 ± 0.1
C <sub>t</sub> CBM50-chain A:[GlcNAc-MurNAc-GlcNAc]	0.0 ± 0.6	0.0 ± 0.1
C <sub>t</sub> CBM50-chain B:[GlcNAc-MurNAc-GlcNAc]	37.9 ± 0.9	21.2 ± 0.1
C <sub>t</sub> CBM50-chain A:MurNAc-[GlcNAc-MurNAc-GlcNAc]	0.0 ± 1.4	0.0 ± 0.1
C <sub>t</sub> CBM50-chain B:MurNAc-[GlcNAc-MurNAc-GlcNAc]	-24.4 ± 1.1	-15.0 ± 0.1
C <sub>t</sub> CBM50-chain A:MurNAc-[GlcNAc-MurNAc-GlcNAc]-MurNAc	0.0 ± 0.6	0.0 ± 0.1
C <sub>t</sub> CBM50-chain B:MurNAc-[GlcNAc-MurNAc-GlcNAc]-MurNAc	14.3 ± 0.6	4.6 ± 0.1
C <sub>t</sub> CBM50-chain A:GlcNAc-MurNAc-[GlcNAc-MurNAc-GlcNAc]-MurNAc	0.0 ± 0.7	0.0 ± 0.1
C <sub>t</sub> CBM50-chain B:GlcNAc-MurNAc-[GlcNAc-MurNAc-GlcNAc]-MurNAc	41.0 ± 1.1	26.5 ± 0.1

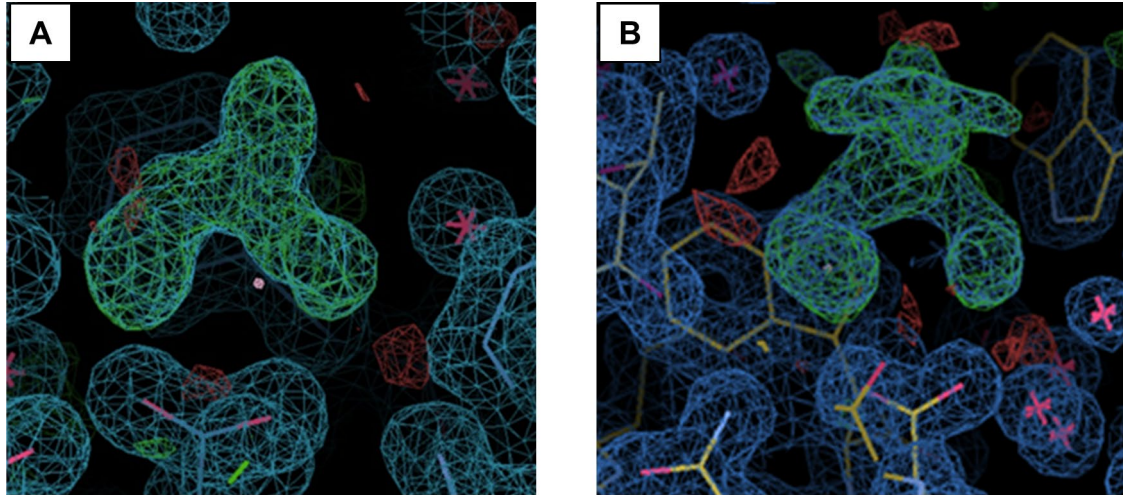
Table S5.10. Primers used to generate the C<sub>t</sub>CBM50 mutant derivatives. Mutation points are depicted in bold.

Mutants	Sequence (5' - 3')	Direction
Trp11Ala	gtcaagccgggagacactatgg <b>cg</b> gaaaattgctgtaaaatca	Forward
	ttgataatttacagcaattt <b>cg</b> ccatagtgctccccgcttgac	Reverse
Asn35Ala	agcaaatccgcaaattaaaaaccct <b>g</b> ccctcattatcccgaca	Forward
	tgccgggataaatgaggg <b>g</b> cagggttttaatt <b>gc</b> ggattgct	Reverse
Tyr38Ala	gcaaatccgcaaattaaaaaccctacctcatt <b>g</b> ctcccgacagaaaatta	Forward
	taatttctgtccgggagcaatgaggttagggtttaatt <b>gc</b> ggattgct	Reverse



## Chapter 6 - Supplementary Information

## Supplementary Figure



**Figure S6.1. *RfCBM13-12115* putative binding sites exhibiting unassigned electron density.** Two of *RfCBM13-12115* putative binding sites are shown for crystal structures obtained from co-crystallization assays with (A) arabinobiose (Ara<sub>2</sub>) and (B) arabinotriose (Ara<sub>3</sub>). The 2mF<sub>o</sub>-DF<sub>c</sub> electron density map is shown in blue (contour at 1  $\sigma$ ) and the mF<sub>o</sub>-DF<sub>c</sub> electron density map is shown in green (contour at 3  $\sigma$ ), evidencing the unexplained density in one of the putative binding sites for each data set.

## Supplementary Table

**Table S6.1. Primers used to generate the *RfCBM13-1* mutant derivatives. Mutation points are depicted in bold.**

Mutants	Sequence (5' - 3')	Direction
Trp38Ala	accaacatccagcagg <b>g</b> cggaactcaacaag	Forward
	ctgttgaagtc <b>g</b> cctgctggatgttgg	Reverse
Gln86Ala	aatgtagagctc <b>g</b> cgacctacacaggcgca	Forward
	tgccctatgtagg <b>g</b> cgagctctacatt	Reverse
Phe134Ala	gaaacgtcaaccag <b>g</b> ccgcctacaacgag	Forward
	ctcgttgtaggc <b>g</b> cctggtgacgttc	Reverse
Asp78Ala	gctaaggatactccg <b>g</b> ccgacggtacaatgta	Forward
	tacattgtaccg <b>g</b> cgccagatccttagc	Reverse
Asp79Ala	ctaaggatactccg <b>g</b> ccggtacaatgtagag	Forward
	ctctacattgtacc <b>g</b> cggtggcagatccttag	Reverse
Asp119Ala	caaggcgctctg <b>g</b> ctgtattcgagtggtcc	Forward
	ggaccactcgaatacagccagagcgccttg	Reverse
Phe121Ala	ggcgcctggatgtag <b>g</b> ccgagtggtccaagg	Forward
	ccttgaccactc <b>g</b> gctacatccagagcgcc	Reverse
Glu122Ala	gctctggatgtattc <b>g</b> cggtgtccaagaaaaacgg	Forward
	ccgtttccttgaccac <b>g</b> cgaatacatccagagc	Reverse
Trp123Ala	ctggatgtattgag <b>g</b> cgccaagaaaaacggc	Forward
	gccgtttccttgacc <b>g</b> cctcgaatacatccag	Reverse
Asn132Ala	acggcggaacgtc <b>g</b> cccagttcgctacaacg	Forward
	cgttgtaggcgaact <b>g</b> ggcgagcttccgccgt	Reverse
Glu138Ala	cagttgcctacaac <b>g</b> cgtagtgcctccagctg	Forward
	cagctggcaggcatac <b>g</b> cgttgtaggcgaactg	Reverse
Tyr139Ala	gttcgcctacaacgag <b>g</b> ctgcctgccagctgtgg	Forward
	ccacagctggcaggcag <b>g</b> cctcgtgtgtaggcgaac	Reverse
Gln142Ala	caacgaglatgcct <b>g</b> cgctgtggaatcgcg	Forward
	gcgatattccacagc <b>g</b> cgcaggcactactcgttg	Reverse

