Raquel Lopes Costa
Graduated in Biochemistry

Ligand discovery and structural-functional analysis of proteins involved in plant cell wall degradation

Dissertation to obtain Master degree in
Biochemistry

Supervisor: Maria Angelina de Sá Palma, UCIBIO, FCT/UNL
Co-supervisor: Ana Luísa Carvalho, UCIBIO, FCT/UNL

Jury:

Examiner: Prof. Doutora Filipa Margarida Barradas de Morais Marcelo
Vowel : Prof. Doutora Doutora Maria Angelina de Sá Palma

FACULDADE DE
CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE NOVA DE LISBOA

September 2016

Raquel Lopes Costa
Graduated in Biochemistry

Ligand discovery and structural-functional analysis of proteins involved in plant cell wall degradation

Dissertation to obtain Master degree in
Biochemistry

Supervisor: Maria Angelina de Sá Palma, UCIBIO, FCT/UNL
Co-supervisor: Ana Luísa Carvalho, UCIBIO, FCT/UNL

Jury:

Examiner: Prof. Doutora Filipa Margarida Barradas de Morais Marcelo
Vowel: Prof. Doutora Maria Angelina de Sá Palma

September 2016

"Ligand discovery and structural-functional analysis of proteins involved in plant cell wall degradation"

Acknowledgments

À minha orientadora, **Doutora Angelina Palma**, e co-orientadora, **Doutora Ana Luísa Carvalho**, por me terem dado esta oportunidade e por acreditarem em mim e no meu trabalho. Obrigada por todo o apoio e ensinamentos que me transmitiram, foram incansáveis. Tive muita sorte em ter-vos como orientadora e co-orientadora, aprendi muito convosco!

À **Professora Doutora Maria João Romão** não apenas como líder do grupo de investigação XTAL da FCT-UNL (onde amavelmente me recebeu), mas também na qualidade de Presidente do Departamento de Química da FCT-UNL e Directora da Unidade de Investigação UCIBIO@REQUIMTE (Unidade de Ciências Biomoleculares Aplicadas), instituição de acolhimento do trabalho aqui reportado.

À **Doutora Márcia Correia** e **Doutora Benedita Pinheiro** pelo acompanhamento e orientação nas técnicas de preparação de células competentes e toda a parte de biologia molecular.

À **Diana Ribeiro** que esteve comigo em toda a parte laboratorial, que me ajudou e me apoiou bastante durante todo o ano. Sempre disponível quando surgia alguma dúvida. Obrigado pelo teu acompanhamento e apoio durante toda a tese, sei que nem sempre foi fácil principalmente no inicio, não me esquecerei da tua ajuda ☺. Obrigado querida.

À **Viviana Correia** por suportar as minhas queixas no laboratório e também por tudo o que me transmitiu em termos de conhecimento, gosto muito de ti.

Ao **Doutor Filipe Freire** e ao **Francisco Leisico** por todas as dúvidas laboratoriais que esclareceram principalmente em técnicas de expressão e purificação e todo o ensinamento transmitido sobre o thermofluor.

Ao **Professor Carlos Fontes e Doutora Joana Brás**, da FMV-UL e NZYTech, pela colaboração e discussão científica e pela produção, à escala *highthroughput*, das proteínas de *C. thermocellum* e *R. flavefaciens* FD-1 que foram analisadas nos microarrays de hidratos de carbono.

À **Professora Ten Feizi**, ao **Doutor Wengang Chai** e à **Doutora Lisete Silva** do Glycosciences Laboratory, Imperial College London pela colaboração na construção dos microarrays de hidratos de carbono.

Ao Professor **Manuel Coimbra, Doutora Cláudia Nunes** e **Doutora Cláudia Passos** da Universidade de Aveiro. À **Professora Berit Smestad Paulsen** da Universidade de Oslo. Pela colaboração na obtenção de polissacáridos para a construção dos microarrays.

Aos colegas do grupo **XTAL** Doutora Catarina Coelho, Doutora Teresa Santos-Silva, Doutor Marino Santos, Rita Otrelo, Cecília Bonifácio por todo o apoio e por tão bem me terem recebido no laboratório. A todos os colegas do XTAL pelo excelente ambiente de trabalho que é proporcionado neste laboratório e pela simpatia e abertura com que recebem quem chega de fora.

À **Doutora Isabel Bento** por me ter iniciado ao mundo da cristalografia.

A todos os **docentes** do mestrado e licenciatura em Bioquímica pelos ensinamentos transmitidos e ao coordenador do mestrado em Bioquímica Professor **Ricardo Franco**.

Ao **BAG Portugal** pela oportunidade de realizar trabalho experimental no ESRF (European Synchrotron Radiation Facility, em Grenoble) e assim poder aprender todo o processo de recolher dados de difração de cristais. Um especial obrigado à minha co-orientadora Doutora Ana Luísa Carvalho que me acompanhou nesta viagem.

À **Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa**, instituição que me acolheu durante a Licenciatura e Mestrado em Bioquímica onde me foi proporcionada toda a minha formação académica.

O trabalho desenvolvido ao longo deste período e reportado nesta dissertação foi financiado pela **Fundação para a Ciência e Tecnologia (FCT-MCTES)** através dos projectos RECI/BBB-BEP/0124/2012 e PTDC-BBB-BEP-0869-2014 e um contrato no âmbito do programa Investigador FCT 2012. A unidade de investigação UCIBIO (Unidade de Ciências Biomoleculares Aplicadas) é financiada por fundos nacionais da FCT-MCTES (UID/Multi/04378/2013) e co-financiada pelo FEDER no âmbito do PT2020 (POCI-01-0145-FEDER-007728).

Obrigado ao meu namorado **Mário** por todos os dias estar comigo nos bons e maus momentos, por me apoiar e me incentivar mesmo quando queria desistir. AMO-TE. Obrigada também aos meus **sogros**, que me ajudam SEMPRE!.

Obrigado aos meus **pais e maninhas**, amo-vos.

Obrigado aos meus **avós**, que sempre acreditaram em mim, em especial ao meu avô que sempre suportou os custos associados à minha educação. Muito obrigado **Avô Carlos**.

Obrigado **família** pelo vosso amor e carinho, sem vocês isto não era possível e por isso dedico-vos esta tese. ☺

Abstract

The plant cell wall is constituted by recalcitrant polysaccharides with diverse sequences that comprise an abundant source of terrestrial biomass. To efficiently degrade plant cell wall polysaccharides, some cellulolytic bacterial organisms, such as *Clostridium thermocellum* and *Ruminococcus flavefaciens*, have an extracellular multi-enzyme complex with catalytic and non-catalytic carbohydrate-binding modules (CBMs). CBMs play a crucial role in enhancing the catalytic efficiency of the enzymes by proximity effect, cell attachment or targeting and disruptive function. The Carbohydrate Active enZymes database (CAZy) organizes the identified CBMs by sequence similarity into different families. Deposition of CBM sequences in the CAZy database is continually growing for which characterization and structure-function analysis is required.

In this study we aim to characterize the carbohydrate ligand specificities of *C. thermocellum* ATCC 27405 and *R. flavefaciens* FD-1 CBMs assigned to different families in the CAZy database. We performed carbohydrate microarray screening analysis for ligand discovery and crystallization screenings aiming to solve the 3D structures of the CBM-ligand complexes by X-ray crystallography. To complement the information provided by these methodologies we also performed ITC (Isothermal Titration Calorimetry), MST (Microscale Thermophoresis) and affinity gel electrophoresis. With the implementation of this approach it was possible to elucidate different carbohydrate binding specificities for biotechnologically relevant CBMs. The results from the initial carbohydrate microarray screening constitute a functional start point to target CBMs for structural-functional analysis of carbohydrate-recognition. *C. thermocellum* family 50 (*Ct*CBM50) reveals to be a novel chitin binding LysM domain and binding with insoluble chitin and a β-(1-4)-GlcNAc chitin oligosaccharide was identified. *R. flavefaciens* FD-1 family 62 CBM (*Rf*CBM62) reveals to be highly specific for a pectic polysaccharide for which the structure is being investigated and binding to galacturonan DP4 was observed. In the scope of this thesis, and as the structural characterization was not achieved in due time, the sequence similarity to known structures inspired the attempt to computationally produce similarity models for the two CBMs. The (hypothetical) conservation of the secondary structures revealed some structural features of the proteins under study.

An important outcome from this integrative study is the possibility to understand the versatility of plant and fungal saccharide sequences and their recognition by the different CBM families. The different binding patterns observed could reflect adaptive pressures of the microorganisms to their respective ecological niches, translating in divergent evolution of the proteome.

Keywords: Plant cell wall degradation, carbohydrate microarray, crystallization, ITC, MST, affinity gel electrophoresis.

Resumo

A parede celular vegetal é constituída por polissacáridos com diversas sequências e que constituem uma fonte abundante de biomassa terrestre. Para degradar de forma eficiente os polissacáridos da parede celular vegetal, alguns organismos celulolíticos bacterianos, tais como o *Clostridium thermocellum* e o *Ruminococcus flavefaciens*, desenvolveram um sistema extracellular multienzimático complexo com módulos catalíticos e módulos não catalíticos de ligação a hidratos de carbono (CBMs). Estes CBMs têm um papel crucial em aumentar a eficiência catalítica de algumas enzimas pela sua avidez para o substrato. A base de dados Carbohydrate Active enZymes (CAZy) organiza os CBMs pela similaridade de sequências primárias em diferentes famílias. A deposição de CBMs na CAZy está continuamente a crescer, sendo que a caracterização e análise da estrutura-função destes módulos é necessária.

Este estudo teve como objetivo caracterizar a especificidade para hidratos de carbono de CBMs atribuídos a diferentes famílias na CAZy, pertencentes às bactérias *C. thermocellum* ATCC 2745 and *R. flavefaciens* FD-1. Foram realizados ensaios com a tecnologia de *microarrays* de hidratos de carbono de forma a identificar ligandos para os diferentes CBMs e ensaios de cristalização com o objectivo de resolver a estrutura tridimensional de complexos CBM-ligando por cristalografia de raios-X. Foram aplicados também outros métodos biofísicos para o estudo da interacção CBM-ligando tais como ITC (*Isothermal Titration Calorimetry*) e MST (*Microscale thermophoresis*) e ensaios de electroforese de afinidade. Com a implementação desta abordagem foi possível elucidar a especificidade de CBMs para hidratos de carbono que são biologicamente relevantes. A análise nos microarrays constitui uma base funcional para seleccionar CBMs para estudos de estrutura-função de reconhecimento a hidratos de carbono. O *Ct*CBM50 de *C. thermocellum* revelou ser um um novo dominio LysM que liga a chitina insolúvel e a interacção com o oligossacárido de quitina (β-(1-4)-GlcNAc) foi identificada. O *Rf*CBM62-de *R. flavefaciens* FD-1 revelou ser altamente específico para um polissacárido de pectina, cuja estrutura está a ser elucidada, e a interacção com um oligossacárido derivado de pectina, galacturonan DP4, foi observada. No âmbito desta tese, e sendo que não foi conseguida a caracterização estrutural, a semelhança com estruturas conhecidas inspirou a produção de modelos de similaridade para os dois CBMs computacionalmente. A conservação (hipotética) da estrutura secundária sugeriu características estruturais e interacção com o ligando para os CBMs em estudo.

Este estudo integrativo permitiu aferir sobre a verstatilidade do reconhecimento de hidratos de carbono de plantas e fungos por diferentes famílias de CBMs, que poderão resultar da adaptação selectiva das bactérias aos seus nichos ecológicos respectivos, resultando na evolução divergente dos seus proteomas.

Palavras-chave: Degradação da parede celular vegetal, *microarrays* de hidratos de carbono, cristalização, ITC, MST, electroforese de afinidade.

Contents

List of Figures

List of Tables

Abbreviations and Symbols

% - Percentage

% (w/v) –weight/volume percentage

Å – Angstrom

Ala – Alanine

Arg - Arginine

Asp –Aspartic acid

Asn - Aspargine

BSA – Bovine serum albumin

CaCl2 – Calcium cloride

CAZymes – Carbohydrate active enzymes

C. termocellum – Clostridium thermocellum

CBM – Carbohydrate-binding module

CBMs - Carbohydrate-binding modules

D.O.600nm – Optical density at 600 nm

DP – Degree of polymerization

E. *coli* - *Escherichia coli*

GBPs – Glycan binding proteins

Glu – Glutamate

Gly - Glycine

HEPES - 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid

His - Histidine

Ile - Isoleucine

IMAC – Immobilized metal affinity chromatography

IPTG - Isopropyl β-D-1-thiogalactopyranoside

ITC – Isothermal Titration Calorimetry

KDa – Kilodalton

LB –Luria-Bertani

Leu – Leucine

Lys – Lysine

NGL - Neoglycolipid

NMR - Nuclear Magnetic Ressonance

MAD –Multiple Wavelength Anomalous Dispersion

MIR - Multiple Isomorphous Replacement

MR – Molecular Replacement

MST – Microscale Thermophoresis

PDB - Protein Data Bank

PEG - Polyethylene glycol

Phe - Phenylalanine

R. flavefaciens – Ruminococcus flavefaciens

RMSD - Root-mean-square deviation

Rpm – Rotations per minute

SAD - Single Wavelength Anomalous Dispersion

Ser – Serine

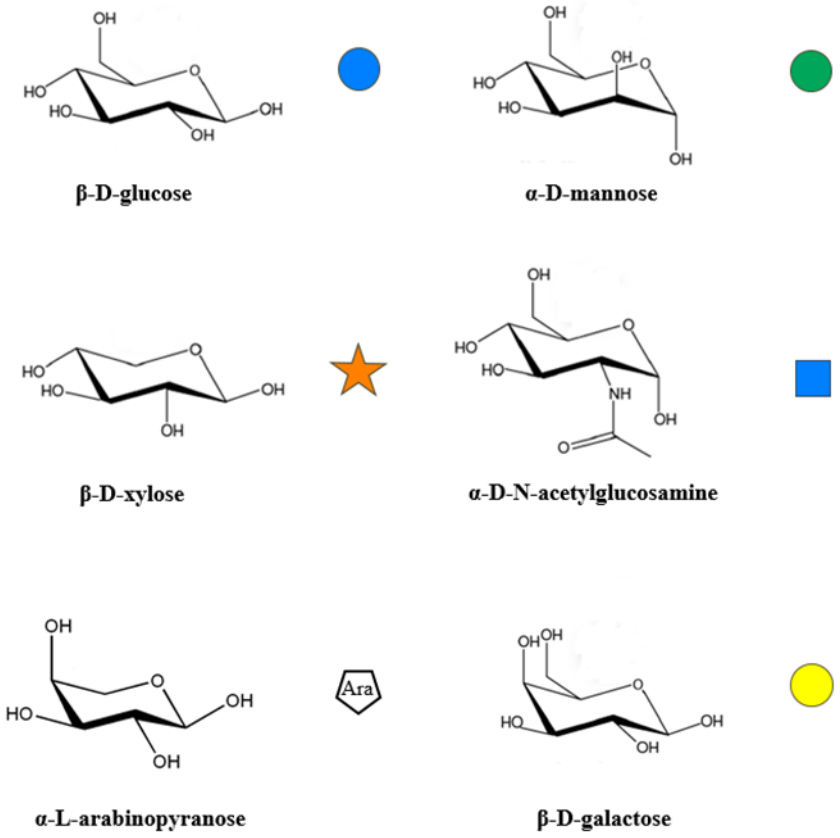
SPR – Surface Plasmon Ressonance

STD-NMR – Saturation-Tranferer Difference – Nuclear Magnetic Ressonance

Thr – Treonine

Tyr - Tyirosine

Chemical structures of monosaccharides present in derivatized NGLs probes presented in this thesis



The symbol notation for monosaccharides is in accord with the proposed in the latest edition of *Essentials of Glycobiology* (www.ncbi.nlm.nih.gov/books/NBK310273/).

Throughout this thesis, oligo- and polysaccharide structures were written in condensed or short forms according to the most recent version of carbohydrates nomenclature, established by International Union of Pure and Applied Chemistry (IUPAC).

Chapter 1 – Introduction and objectives

1.1 Overview of the plant cell wall

The plant cells are surrounded by a relatively thin but strong and rigid wall. This wall has a functional role in plant growth, development and reproduction by contributing to structural integrity, cell adhesion, and mediation of defense responses (Varner *et al.* 1989).

In the biotechnology field, plant cell walls are an important source of renewable energy and biomass, as they fix carbon that is then integrated into cell wall polymers. Plant cell wall material is also important for human economics as a natural source of fibers for the production of textile, paper-based products and wood products (Taiz & Zeiger 2002). The organic substances that create humus in the soil and improve its fertility are also derived from the plant cell walls (Taiz & Zeiger 2002).

The architecture of the plant cell wall and its composition is introduced in this section.

1.1.1 Plant cell wall morphology

The plant cell wall is a complex extracellular matrix that is present at the surface of the plasma membrane. The cell wall acts as an exoskeleton, conferring resistance to the cell, flexibility to resist to cell disruption and permeability to allow the intercellular transport. The structure and composition of plant cell walls is not uniform; factors such as plant tax, tissue, age, cell type and cell layer have influence on their composition (Taiz & Zeiger 2002). Morphologically, it is possible to differentiate three zones: middle lamella, primary wall and secondary wall (Figure 1.1).

The most external layer is the middle lamella, a thin layer of material that is deposited just after cell division and can usually be seen at the junction between the walls of neighboring cells, acting as a sort of separating panel (Harris & Stone 2008). It is the first zone to be formed and it is constituted almost totally by pectic polysaccharides (Heredia *et al.* 1995).

Over the middle lamella is the primary wall that controls the growth of the cell and forms the structural basis of the plant. It is composed of approximately 25% cellulose, 25% hemicelluloses, and 35% pectins, with 1 to 8% structural protein (Taiz & Zeiger 2002). These components form three essential networks: 1) cellulose microfibrils linked to each other by different glucans, 2) gel-forming network of pectin linked by calcium ion bridges and 3) structural proteins that are covalently bound to each other and to other cell wall components. Some cells are formed only by middle lamella and primary wall, but specialized cells have another component, the secondary wall (Heredia *et al.* 1995). When the secondary wall is formed some changes occur in the middle lamella and primary wall, for example lignification, a complex process in which lignins are deposited on the extracellular polysaccharide matrix and the main function is to strengthen the plant vascular body.

After the cell growth terminates or when the cells differentiate, secondary walls are formed between the plasma membrane and the primary cell wall. The secondary wall contains cellulose (in higher content than primary wall) (Taiz & Zeiger, 2002), hemicellulose and lignin (Sticklen 2008). Secondary walls can be divided into three layers (from the outside to the inside): S1, S2 and S3, which differ in the orientation of the cellulose microfibrils (Figure 1.1).

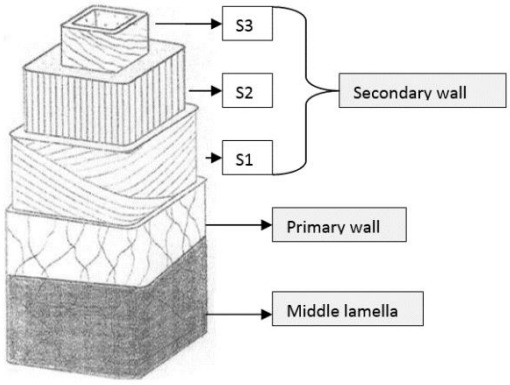


Figure 1.1 – Illustration of the different layers of the plant cell wall. Adapted from Plomion *et al.*

1.1.2 Plant cell wall composition

1.1.2.1 Cellulose

Cellulose is the major component of the plant cell wall and represents the most abundant organic polymer on Earth (Overmann 2006). Cellulose is a polysaccharide with approximately 7000 to 15000 D-glucose units that are linked by β-(1-4) glycosidic bonds. The disaccharide cellobiose (Figure 1.2) constitutes the structural repeating unit.

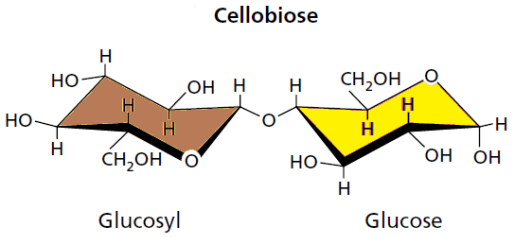


Figure 1.2 – Cellobiose structure. Adapted from Taiz & Zeiger 2002.

Cellulose molecules are aligned parallel to each other and linked by hydrogen bonds and van der Waals interactions to form elementary fibrils. These elementary fibrils make a highly ordered insoluble (crystalline) ribbon that excludes water and is relatively inaccessible to enzymatic attack.

Elementary fibrils are packed into microfibrils that contribute to the strength and structural basis of the plant cell (Figure 1.3) (Taiz & Zeiger 2002). The cellulose microfibrils include both crystalline and amorphous regions, in ratios dependent on the degree of polymerization (DP), the extent of hydrogen bonding and the source of cellulose (Bayer *et al.*, 2000). The microfibrils are then assembled into cellulose fibers, which are packed together by both intra- and inter-fiber hydrogen bonds (Heredia *et al.* 1995). Adjacent sheets overlie one another and are held together by weak van der Waals forces. This crystalline structure is tight enough to prevent chemical and biological degradation and the diffusion of small molecules such as water.

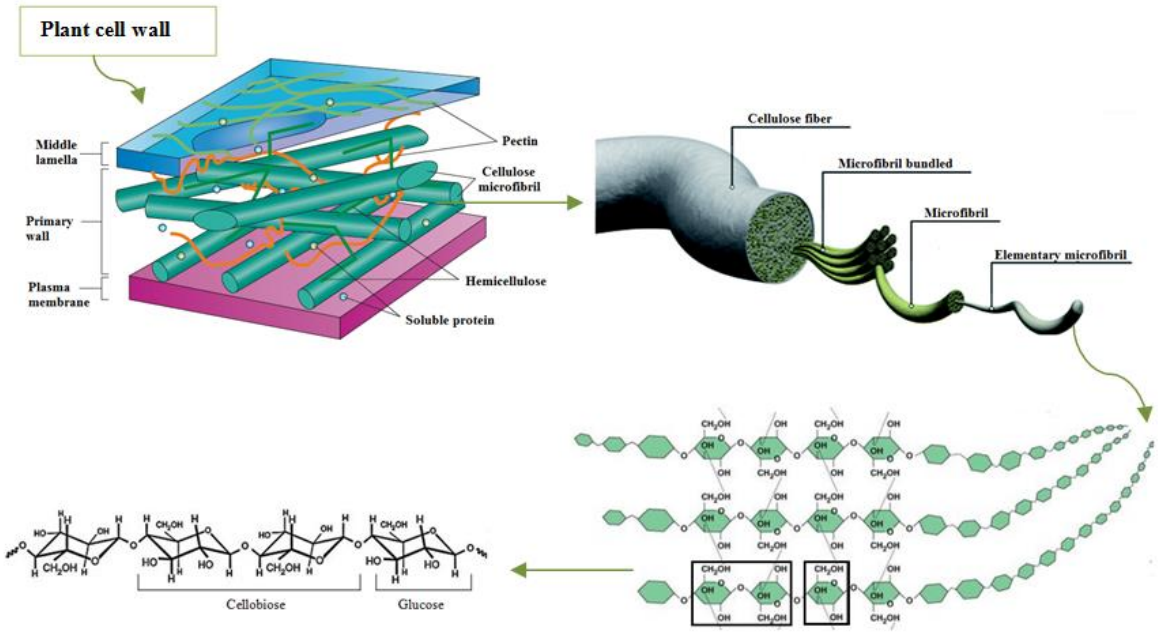


Figure 1.3 –Representation of the plant cell wall and cellulose structure. Adapted from Stiklen 2008 and www.joostdvree.nlshmlscellulose.sthm, last acess 22.09.2016

1.1.2.2 Non-cellulosic major components

Most important non-cellulosic polysaccharides are divided into two groups: Pectic polysaccharides and hemicelluloses or "cross-linking glycans".

Pectic polysaccharides contain galacturonic acid and can be divided into: 1) galacturonans, which are composed of a linear chain of α-(1-4)-linked D-galacturonic acids; 2) rhamnogalacturonan I, which consists of a backbone of alternating α-(1-4)-linked galacturonic acid and α-(1-2)-linked rhamnose units with side branches that contain other pectin domains (primarily α-(1-5)-linked arabinan and β-(1-4)-linked galactan side chains); and 3) rhamnogalacturonan II that is smaller than rhamnogalacturonan I but has a more complex structure, it comprises an α-(1-4)-homogalacturonan backbone of seven to nine residues substituted by up to five side chains comprising 12 monosaccharides bound by 20 linkages (Gorshkova *et al.* 2010). Pectins form a hydrated gel phase in which the cellulose-hemicellulose network is surrounded; acting as an hydrophilic filler they prevent aggregation and collapse of the cellulose network (Taiz & Zeiger, 2002).

Hemicelluloses are also designated "cross-linking glycans" because they can be bound to cellulose microfibrils. Hemicelluloses are polysaccharides that have β-(1-4)-linked backbones with an equatorial configuration in which are included xyloglucans, xylans, mannans, glucomannans, and mixed-linked β-(1-3;1-4)-glucans and that can be decorated with a diverse range of carbohydrate side-chains. The main characteristic of hemicelluloses is their capacity of strengthening the cell wall by interaction with cellulose and, in some cases, with lignin (Scheller & Ulvskov 2010).

Besides these two groups of non-cellulosic polysaccharides, the cell wall also contains structural proteins, enzymes and lignin. Lignin is the generic term for cross-linked phenolic polymers. Due to its structure and association with cellulose and hemicellulose, lignin is important to prevent the degradation of these plant polysaccharides (Vanholme *et al.*, 2010).

1.1.3 Plant cell wall degradation

The degradation of the plant cell wall by microorganisms is a fundamental biological process that is of considerable industrial importance, as plant cell wall polysaccharides are a major reservoir of carbon and energy. However, the plant cell wall is very complex and is composed by a variety of carbohydrates that are recalcitrant to the enzymatic attack. Only a restricted number of microorganisms have acquired the ability to deconstruct these structural carbohydrates (Fontes & Gilbert, 2010).

To achieve total or partial degradation of these carbohydrates a consortium of enzymes, free or in complex, is required. The extracellular plant cell wall degrading machinery is different for anaerobic and aerobic microorganisms as these use different strategies to target the polysaccharides (Bayer *et al.* 2004).

Aerobic microorganisms secrete free extracellular enzymes, such as endoglucanases, exoglucanases and β-glucosidases, which act individually in the degradation of plant cell wall and are considered as non-integrating systems. As example of aerobic microorganisms are bacterium from *Bacillus, Micromonospora, Cellvibrio* and *Pseudomonas* (Lynd, 2002) and fungi of genera *Aspergillus* (de Vries *et al.*, 2001).

The most anaerobic microorganisms have plant cell wall degrading enzymes associated in a supramolecular complex (molecular weight >3 MDa), termed the ‘cellulosome’ and are considered as integrating systems. As example of anaerobic microorganisms are bacterium of the genera *Clostridium, Ruminococcus*, *Thermotoga* (Bergquist *et al.*, 1999) and fungi of the genera *Neocallimastix*, *Piromyces* and *Orpinomyces* (Fontes & Gilbert, 2010). Anaerobic bacteria and fungi in the rumen have developed a wide array of multimodular cellulases and hemicellulases that can act individually or as organized cellulosomes for the hydrolysis of plant cell wall polysaccharides to soluble sugars. Many of anaerobic microorganisms are found in the digestive track of invertebrate and vertebrate, as well as other specific ecosystems, such as soils, sediments and water bodies.

1.2 The bacterial Cellulosome: identification and organization

Anaerobe microorganisms such as *Clostridia* and rumen bacterium organize cellulases and hemicellulases into a large multienzyme complex termed the cellulosome (Bayer *et al.*, 2004). It is likely that anaerobic environments impose selective pressures that have directed to the formation of cellulosomes. However, the evolutionary drivers that directed to the formation of these enzyme complexes are still currently unclear (Fontes & Gilbert, 2010).

In this Section, the two bacterium subject of this Thesis, *Clostridium thermocellum* and *Ruminococcus flavefaciens,* and their respective cellulosomes are described. The *C. thermocellum* cellulosome was the first to be identified and characterized in the 80's while *R. flavefaciens* cellulossome was only recently identified and is not as well characterized.

1.2.1 The highly efficient cellulose-degrading bacterium *Clostridium thermocellum*

C. thermocellum is an anaerobic, rod shaped and gram-positive thermophile. This bacterium has gained research interest due to its cellulolytic and ethanologenic capabilities as it can convert biomass into energy. The ecological niche of *C. thermocellum* is abundant in soil worldwide (Freier *et al.*, 1988), however, it can also be found in water bodies and digestive flora of animals (*Clostridium thermocellum,* www.microbewiki.kenyon.edu last acessed 16.09.2016)

In the 1980, Bayer & Lamed and their colleagues identified and characterized the first cellulosome, on the basis of studies of the cellulolytic system expressed by the anaerobic thermophilic bacterium *C. thermocellum* (Lamed *et al.*, 1983). Through the 80´s and 90´s there was an effort to elucidate the molecular mechanism for the assembly of *C. thermocellum* cellulosome, enabling the identification of its different structural components and how it is presented at the surface of the bacterium. In Figure 1.4 the cellulosomes are presented in the form of protuberances (Fontes & Gilbert, 2010).

C. thermocellum has one of the highest rates of cellulose utilization known, and the cellulosome of this bacterium is described as display a specific activity against crystalline cellulose (about 50 times more than *Trichoderma)* which may have been evolutionary imposed by the anaerobic environment (Fontes & Gilbert, 2010).

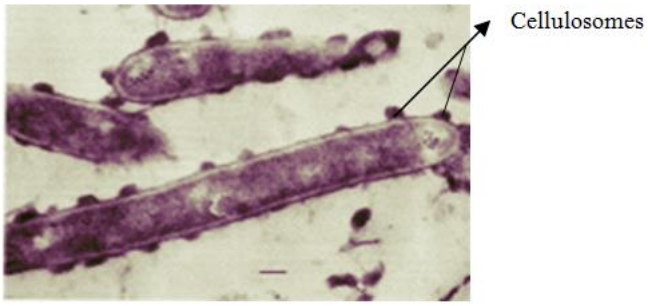


Figure 1.4 – Cellulosomes at the surface of *Clostridium thermocellum* in form of protuberances. Adapted from Fontes & Gilbert 2010.

The main structural component of the cellulosome is a scaffoldin subunit termed CipA (Figure 1.5). This component is a large non-catalytic enzyme-integrating protein that contains nine highly conserved modules, cohesin type I modules, which allow the incorporation of different Carbohydrate-Active enZymes (CAZymes) and associated carbohydrate-binding modules (CBMs). The cohesins have a complementary non-catalytic module, the dockerin type I, which binds to the cohesin type I modules of

the scaffoldin subunit. The type I cohesin – dockerin interactions allow the integration of the hydrolytic enzymes into the complex and potentiates the stability and organization of the cellulossome (Bayer *et al.* 2004). To lead the scaffoldin subunit to the surface of the bacterium cell, membrane-associated proteins are bound to a type II cohesin. The type II cohesin-dockerin interactions support the anchoring of cellulosomes into cell surfaces. The scaffoldins containing type II cohesins are termed anchoring scaffoldins, while those containing type I are termed primary scaffoldins. Primary scaffoldins usually have a CBM3a that recognizes and binds to the recalcitrant cellulosic substrate and thus play a key role in bringing the cellulosome into close proximity with the plant cell wall.

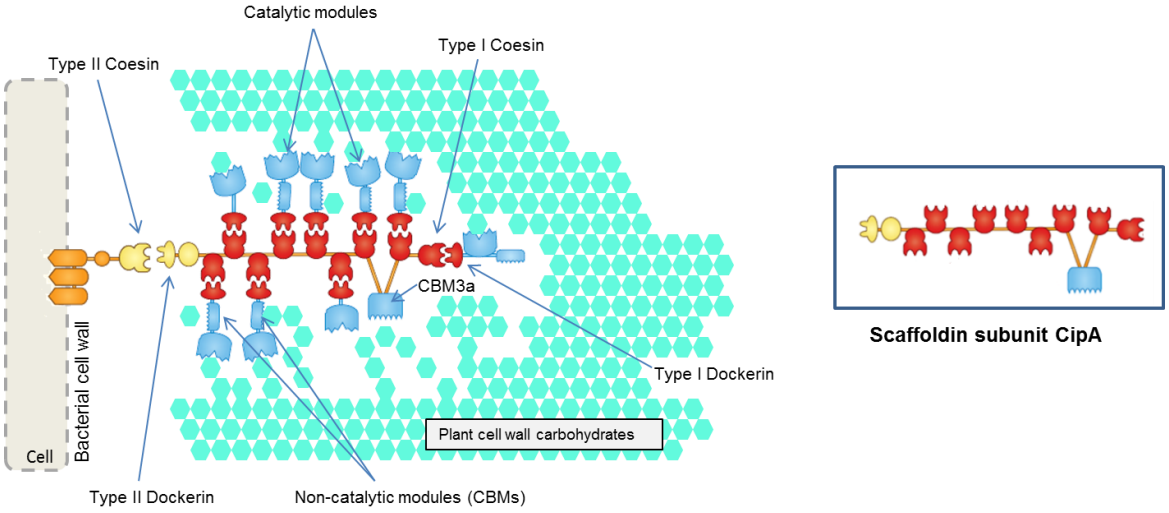


Figure 1.5 – The *Clostridium thermocellum* cellulosome. Representation of the plant cell wall degradation by the Clostridium thermocellum cellulosome. Adapted from Fontes & Gilbert 2010.

1.2.2 *Ruminococcus flavefaciens*

The cellulosome of *C. thermocellum* is one of the best characterized and one expressing the highest rates of cellulose hydrolysis. Recently, a range of anaerobic bacteria were shown to produce cellulosome systems similar to those of *C. thermocellum*, such as *Ruminococcus flavefaciens.*

R. flavefaciens are cellulosic Gram-positive cocci of the order "Clostridiales", an anaerobic bacterium that inhabits the rumen community. They are responsible for the digestion of plant cell wall polysaccharides in the large intestine of herbivorous mammals and humans (Berg Miller *et al.* 2009).

In the past years, the sequencing of the genome of *R. flavefaciens* FD1 unraveled the complexity and diversity of this rumen bacterial cellulosome, revealing information about enzymatic and structural components of the cellulosome (Berg Miller *et al.* 2009). Also, the discovery of cellulases and other proteins involved in plant cell wall degradation was important to understand how the host organisms extract energy from their diet. The large number of protein-encoding sequences identified containing dockerin modules dictates that *R. flavefaciens* FD-1 has the largest collection of cellulosome-associated proteins of any known fiber-degrading bacterium thus far described.

This cellulossome is considered as the most intricate and potentially versatile multienzyme complex described (Figure 1.6), with a genome that encodes 275 dockerins (in comparison with the 76 from *C. thermocellum*) that are likely to integrate the multienzyme complex (Berg Miller *et al.* 2009). The enzymes with homology for glycosyde hydrolases, carbohydrate esterases and polysaccharide lyases of known function constitute a diverse potential of substrate specificity (Fontes & Gilbert 2010). As the aim of several scientific groups is to elucidate the mechanism for the complete deconstruction of structural polysaccharides, *R. flavefaciens* cellulosome plays an important role in the discovery of new CAZymes and CBMs and thus to clarify their functions.

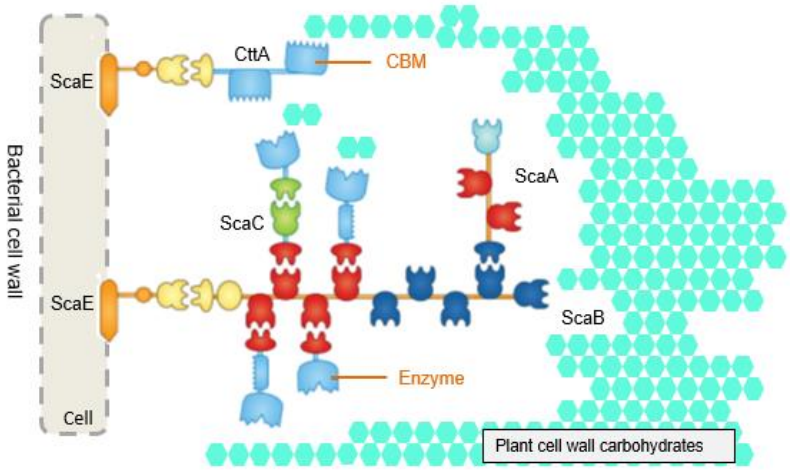


Figure 1.6 – The *Ruminococcus flavefaciens* FD-1 cellulosome. Representation of the pant cell wall degradation complexity. Coesins colored in red have different specificities than coesins colored in dark blue. Adapted from Fontes & Gilbert 2010.

Althougth less is known about the *R. flavefaciens* cellulosome, it is evident that this cellulosome has a different organization from the *C. thermocellum* cellulosome. Several studies have suggested sequence and structural differences between the cohesins and dockerins of *C. thermocellum* and *R. flavefaciens* (Fontes & Gilbert 2010).

In Figure 1.6, the *R. flavefaciens* cellulosome, the single cell-anchoring scaffoldin (ScaE) may bind a protein termed CttA, which carries two putative CBMs that mediate the primary anchorage to the plant cell wall. In addition, ScaE cohesin binds to the C-terminal dockerin of the primary scaffoldin (ScaB) which contains nine cohesins with two different specificities: four cohesins (Figure 1.6 in red) recognize the dockerin of the catalytic subunits or ScaC; five cohesins (Figure 1.6 in dark blue) bind to ScaA, which functions as an adaptor and primary scaffoldin. ScaA contains two cohesins that bind the cellulosomal catalytic subunit (presenting a similar specificity to the cohesins of ScaB, Figure 1.6 in red) and thus amplifies the number of enzymes in *R. flavefaciens* cellulosome. ScaC, an adaptor scaffoldin, can bind to ScaA and ScaB due to the specificity of the catalytic subunits (Fontes & Gilbert 2010).

Cellulosome structural organization varies between different strains of *R. flavefaciens*, which may reflect the complexity and diversity of substrates found in the rumen.

1.3 The modular organization and recognition mechanisms of Carbohydrate-Active enZymes

Microorganisms that degrade plant-cell carbohydrates produce a high variety of polysaccharide-degrading enzymes, such as glycoside hydrolases, polysaccharide lyases, carbohydrate esterases and polysaccharide oxidases (Gilbert 2010). Many of these enzymes are part of the cellulosome to increase the efficiency for degrading the substrate. These enzymes are multimodular and often contain non-catalytic carbohydrate-binding modules (CBMs), which are connected through highly flexible linker sequences.

CAZymes represent these classes of enzymes that are involved in assembly and modification/cleavage of polysaccharides. CAZymes and CBMs have been classified over 25 years into sequence-based families in the CAZy database (www.CAZy.org) (Davies & Henrissat 2013). Since September 1998 CAZy is dedicated to the display and analysis of genomic, structural and biochemical information on the carbohydrate-active enzymes and associated CBMs involved in the synthesis and degradation of complex carbohydrates (Cantarel *et al.* 2009). Currently there are numerous different CAZymes and CBM families identified: 135 for glycoside hydrolases, 100 for glycosyl transferases, 24 for polysaccharide lyases, 14 for carbohydrate estereases and 80 for CBMs. In addition, CAZy also displays available PDB codes for those proteins that have been structurally characterized.

1.3.2 Carbohydrate-binding modules

As stated in the previous section, glycoside hydrolases involved in biodegradation of the plant cell wall are normally modular and have appended non-catalytic modules, CBMs. CBMs are defined as a contiguous sequence of 30 to 200 amino acids, that are normally appended to the associated catalytic module by a flexible linker for the recognition of the plant cell wall polysaccharides and to help increase the enzyme activity (Gilbert *et al.* 2013). CBMs are not always appended to a catalytic module and they can be present in isolated or tandem forms not coupled with an enzyme (Cantarel *et al.* 2009).

Lectins share some structural similarities with CBMs and they can bind to their target ligand through similar mechanisms, however, CBMs are normally found in enzymes that are degrading complex carbohydrates primarily to provide nutrients. This functionality distinguishes CBMs from lectins into a separated protein group (Gilbert *et al.* 2013). CBMs were initially classified as cellulose-binding domains (CBDs) based on the initial discovery of several modules that bound cellulose as primary ligand (Boraston *et al.* 2004). However, the term evolved to carbohydrate-binding modules due to the diverse ligand specificity exhibited.

1.3.2.1 Classification of CBMs

There are currently 80 CBM families in CAZy database, in which the major part contains members that target plant cell walls. They are divided into 7 different-fold families based on the conservation of

protein fold (Boraston *et al.* 2004). Although there are 7 different-fold families, the majority of CBMs identified to date are included in the β-sandwich fold family (Table 1.1).

Table 2.1 – Typical 'fold families' identified for CBMs and examples of CBM families with these folds. Adapted from Boraston *et al.* 2004 and www.CAZy.org, last acess at 14.09.20016.

Fold Family	Fold	CBM families
1	β-Sandwich	2,3,4,6,9,15,17,22,27,28,29,32,34,36,,47,51,70
2	β-Trefoil	13,42
3	Cystein knot	1
4	Unique	5,12
5	OB fold	10
6	Hevein fold	18
7	Unique; contains hevein-like fold	14

To provide additional functional relevance to the CBM classification these modules have been grouped into three types: A, B, and C according to its interaction with the carbohydrate. These three types define CBMs that bind crystalline surfaces, oligosaccharide sequences, or short oligosaccharide/monosaccharides, respectively.

CBMs from **Type A** have a planar hydrophobic surface composed by aromatic residues. This planar configuration in the active site interacts with flat crystalline polysaccharides, such as chitin or cellulose (Boraston *et al.* 2004). The properties of this CBM-type differ significantly from the other types of carbohydrate-binding proteins.

Recently, the crystal structure of a CBM63-containing *Bacillus* expansin (proteins that disrupt the cellulose–hemicellulose interface) in complex with cellohexaose was determined. A typical Type A, CBM63 contains a planar surface comprising three aromatic residues that make parallel π/C-H contact with the ligand (Figure1.7) (Georgelis *et al.* 2012).

Another study showed that a cohort of type A CBMs that belong to families 2 and 3 bind not only to crystalline cellulose but also to xyloglucan (Hernandez-Gomez *et al.* 2015).

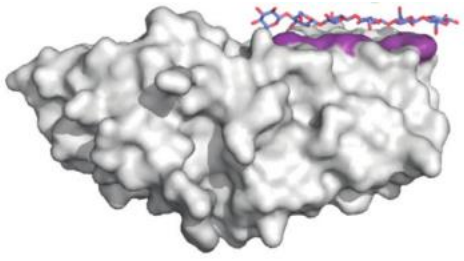


Figure 1.7 – The crystallographic dimer of CBM63 from *Bacillus subtilis* (PDB code 4FER) shown in complex with cellohexaose (blue sticks). Surface contributed by aromatic amino acids shown in purple. Adapted from Gilbert *et al.* 2013.

CBMs from **Type B** or endo-type are classified as CBMs that bind to internal oligosaccharide sequences (Gilbert *et al.* 2013). This type of CBMs display a cleft arrangement in which the binding site accommodates longer glycan chains with four or more monosaccharide units (Figure 1.8). This type of CBMs have clearly evolved a binding site for the interaction with individual glycan chains rather than crystalline surfaces (Boraston *et al.* 2004).

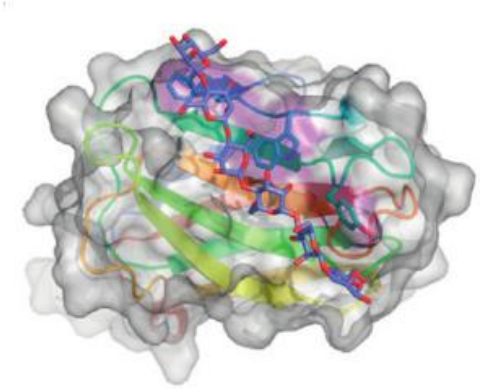


Figure 1.8 – The X-ray crystal structure of the family 29 CBM from *Piromyces equii* in complex with mannohexaose (PDB code 1GWL). The secondary structures are shown as colored ramped cartoons with relevant amino acid side chains involved in carbohydrate binding shown as sticks. Solvent accessible surfaces are shown in gray with the surfaces contributed by the aromatic residues colored purple. Adapted from Gilbert *et al.* 2013.

CBMs from **Type C** or exo-type are classified as CBMs that bind the termini of an oligosaccharide sequence (Gilbert *et al.* 2013). This type of CBMs, also described as 'lectin-like' CBMs, have the property of binding in an optimal way to mono-, di- or tri-saccharides, due to steric restriction in the binding site (Boraston *et al.* 2004). Unlike the Type B CBMs, Type C do not contain the extended grooves on binding-sites. One example of this CBM type is the family 6 CBM from *Bacillus halodurans* in complex with lamirarihexaose (Figure 1.9).

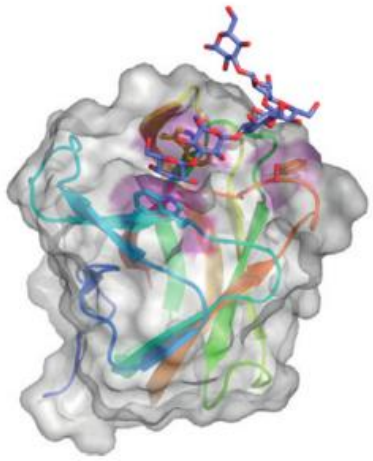


Figure 1.9 – The X-ray crystal structure of the family 6 CBM from *Bacillus halodurans* in complex with lamirarihexaose (PBD code 1W9W). The CBM specifically recognizes the non-reducing end of the sugar. The secondary structures are shown as colored ramped cartoons with relevant amino acid side chains involved in carbohydrate binding shown as sticks. Solvent accessible surfaces are shown in gray with the surfaces contributed by the aromatic residues colored purple. Adapted from Gilbert *et al.* 2013

1.3.2.2 Functional roles of CBMs

In general, CBMs contribute in the binding of the target substrates to carbohydrate degrading enzymes. According to experiments in the field reviewed by Boraston *et al.* 2004 and Gilbert *et al.* 2013, CBMs have four major roles: Proximity effect, targeting function, disruptive function and cell attachment.

Targeting function

As described in the previous section, CBMs are classified into different types according to the architecture of the binding site. Depending on this architecture, CBMs can target the enzyme to distinct regions within a larger macromolecular polysaccharide substrate. In general, there is a tight correlation between the ligands recognized by bacterial CBMs and the substrate specificity of the appended catalytic modules. As an example, *Ct*CBM11 recognizes mixed-linked β-(1-3;1-4) glucans and is associated to the enzymes GH5 and GH6, endoglucanases with catalytic activity for β-(1-3) and β-(1-3;1-4) glucans, respectively (Carvalho *et al.* 2004).

Another study (Najmudin *et al.* 2006) shows that the function of CBM44 from *C. thermocellum* Cel44A is associated with the targeting of Cel44A to its substrates. Cel44A is a typical endoglucanase that is capable of cleaving a variety of glucan-based plant cell wall polysaccharides such as cellulose, β-glucans, xyloglucans, and glucomannans. To mediate efficient targeting, ligand recognition by CBM44 needs to mirror the substrate specificity of Cel44A; this study showed that this promiscuity in carbohydrate recognition is intrinsic to CBM44.

Proximity effect

CBMs increase the concentration of enzyme in close proximity to its polysaccharide substrate, which will lead to a rapid and efficient degradation. The proximity between the enzyme and the substrate potentiates the degradation of the polysaccharide (Herve *et al.* 2010). Herve and colleagues showed that CBMs can potentiate the action of a similar catalytic module toward polysaccharides in intact cell walls through the recognition of non-substrate polysaccharides. The targeting actions of CBMs therefore have strong proximity effects within cell wall structures, explaining why cellulose-directed CBMs are appended to many non-cellulase cell wall hydrolases (Herve *et al.* 2010).

Disruption of polysaccharide structure

Some CBMs have showed the capacity to disrupt the surface of packed polysaccharide structure, such as cellulose fibers and starch granules. This disruptive effect causes the substrate to release and become more exposed to the catalytic module and thus enhance its degradation capacity.

This function was first documented for the N-terminal family 2a CBM of Cel6A from *Cellulomonas fimi* (Din *et al.* 1994) that appeared to mediate non-catalytic disruption of the crystalline structure of cellulose; furthermore, this disruptive effect enhanced the degradative capacity of the catalytic module.

Cell attachment

CBMs have been shown to adhere enzymes onto the surface of bacterial cell wall components while exhibiting catalytic activity on an external neighboring carbohydrate substrate. A family 35 CBM module has been shown to interact with the surface of glucuronic acid containing sugars in the cell wall of *Amycolatopsis orientalis*, while the catalytic module is active on external chitosan. This study shows that the biological role of CBM35s is not dictated solely by their carbohydrate specificities but also by the context of their target ligands (Montanier *et al.* 2009).

A unique family of carbohydrate-binding modules (CBM37) of *Ruminococcus albus*, located at the C-terminus of different glycoside hydrolases, appears to be responsible both for anchoring these enzymes to the bacterial cell surface and for substrate binding (Ezer *et al.* 2008).

1.3.2.3 Biotechnological applications of CBMs

The practical uses for CBMs in different fields of biotechnology are constantly on rise. CBMs have three properties that make them perfect candidates for different applications: (i) CBMs are usually independently folded units and therefore can function autonomously in chimeric proteins; (ii) the attachment matrices are abundant and inexpensive and have excellent chemical and physical properties; and (iii) the binding specificities can be controlled (Shoseyov *et al.* 2006).

CBMs are used to enhance bioprocessing enzymes in biomass degradation for industrial uses in biofuel production. A study demonstrates that the fusion of a CBM to the wild type cellulases Cel9A and Cel5A enhanced their activity as much as three fold on two insoluble lignocellulosic substrates (Reyes-Ortiz *et al.* 2013).

CBMs are also used for purification of biomolecules in immobilized affinity ligand technology. They act as affinity support for enzyme immobilization with high capacity, while retaining enzymatic activity and in some cases increased enzymatic activity is reported (Shoseyov *et al.* 2006). Studies demonstrate that CBMs can be utilized as biosensors; an example is a CBM that was used for glucose sensing in bioreactors (Verma *et al.* 2015).

Cellulose is a major component of numerous commercial products, several of which are capable of being recycled. Therefore, CBMs can also be used for the targeting of functional molecules to materials containing cellulose. The commercial potential of CBMs in this context was first realized for denim stonewashing, where cellulases were used as an alternative to the original abrasive stones (Shoseyov *et al.* 2006).

Cell immobilization is another application of CBMs. This technique was explored for whole cell immobilization by cellulosic material (Verma *et al.* 2015) and for bioremediation. Another property that

has been demonstrated for CBMs is that they can modify the characteristics of some enzymes. Replacing or adding a CBM can improve the hydrolytic activity, for example the addition of a CBM derived from cellohydrolase II of *Trichoderma reesei* to *Trichoderma harzianum* chitinase resulted in increased hydrolytic activity of insoluble substrates (Shoseyov *et al.* 2006).

1.4 Overview of the methodologies to study CBM-carbohydrate interactions

Carbohydrates play an intriguing role inside and at the surface of the cell. Carbohydrates occur as mono- and polysaccharides but are also part of glycan structures such as glycoproteins, glycolipids, peptidoglycans. Over the last decade, evidence has accumulated indicating that interactions between carbohydrates and particular proteins that recognize them play critical roles not only in the context of polysaccharide utilization and biotechnological applications but also in many other biological processes in health and disease, such as cell adhesion, signal transduction, host-pathogen interactions and inflammation processes (Park *et al.* 2008; Liu *et al.* 2009; De Schutter & Van Damme 2015). In many cases, the protein-carbohydrate interaction is only a first step in a series of events, often leading to a complex recognition process or signaling cascade. To fully understand the protein-carbohydrate interaction, different techniques for its characterization are required.

Biophysical methods such as carbohydrate microarray technology, Microscale Thermophoresis (MST), Isothermal Titration Calorimetry (ITC) and Surface Plasmon Resonance (SPR) and also structural methods such as X-ray crystallography and Saturation Transfer Difference NMR (STD-NMR) are used to study the protein-carbohydrate interactions.

In this Section these methods are briefly described giving more emphasis to the carbohydrate microarray technology and X-ray crystallography.

1.4.1 Carbohydrate microarray technology

The development of the carbohydrate (or glycan) microarrays in the beginning of the twenty-first century addressed the need for high-throughput methods that systematically array glycan libraries and identify glycan binding proteins (GBPs) to enable the investigation of their biological roles. Since their introduction in 2002, applications of the glycan microarrays have grown exponentially (Rillahan & Paulson 2011; Palma *et al.* 2014; Palma *et al.* 2015)

Carbohydrate microarray technologies are novel tools that allow the studies of carbohydrate-proteins interactions and elucidation of carbohydrate ligands involved in different biological contexts. Carbohydrate microarray technology allows to study a wide range of carbohydrate sequences that are immobilized on solid matrices using only minute amounts of sample (minute spots). This wide range of immobilized carbohydrates on a solid surface allows to mimic at some extent the cell surface and at the same time it is ideal to detect low affinity interactions. In general, there are two categories for carbohydrate microarrays: polysaccharide (or glycoconjugate) microarrays and oligosaccharide microarrays (Liu *et al.* 2009).

1.4.1.1 Carbohydrate probe libraries and immobilization methods

The first step in the construction of a carbohydrate microarray is to obtain pure and characterised carbohydrates, either from isolation from natural sources or from chemical or chemoenzymatic synthesis, and prepare these as suitable carbohydrate probes for the microarrays.

The oligosaccharides obtained from natural sources can include those derived from glycoproteins and glycolipids of different sources (mammalian, non-mammalian), bacterial-, fungal- and plant polysaccharides obtained through acid/alkaline hydrolysis and enzymatic degradation or depolymerization, as well as free-oligosaccharides isolated from mammalian milk and urine (Feizi 2003). The isolation of pure compounds, assignment of their structures and their immobilization on the array surface are challenges of using glycans from natural sources.

The polysaccharides or glycoconjugates, such as glycoproteins or glycolipids, can be directly immobilized on solid matrices by hydrophobic physical adsorption (Pedersen *et al.* 2012; Palma *et al.* 2015b) or charge-based interaction in the case of polysaccharides (Shipp & Hsieh-Wilson 2007). Polysaccharide or glycoconjugate microarrays are valuable for comparative antigenicity analysis and for ligand-discovery microarray projects.

The oligosaccharide microarrays are required for assignment of the carbohydrate ligand specificity of a given recognition system and are thus key tools to provide detailed information on structure-activity relationships in carbohydrate recognition (Liu *et al.* 2009). The oligosaccharide immobilization normally requires derivatization processes before arraying, due to their hydrophilic nature. Many different methodologies have been developed for immobilizing oligosaccharides. One of the approaches is to conjugate natural or chemically synthesized oligosaccharides to lipid by reductive amination or oxyme ligation to generate neoglycolipid (NGL) probes with amphipathic properties for efficient immobilization on nitrocellulose membranes (Liu *et al.* 2012; Palma *et al.* 2014). The advantage is that natural and synthetic glycans can be combined to generate diverse libraries, the strategy is to derivatize glycans from both sources with the same lipid tag for the immobilization. Most other mono- or oligosaccharide probes generated for printing have been de novo synthesized chemically or chemo-enzymatically, defining structures that incorporate specific functional groups for covalent attachment to solid matrices (Rillahan & Paulson 2011). Chemical synthesis has the advantage over natural isolation in that large quantities of relatively pure carbohydrates can be produced while the yield of the natural glycans is usually low, requiring deconvolution due to the heterogeneity of isolated mixtures (Horlacher & Seeberger 2008). However, achievement of glycan diversity offers a challenge and methods have been developed to prepare ‘designer’ or ‘shot-gun’ microarrays from a natural glycome source (Palma *et al.* 2015b; Song *et al.* 2011)

Over the past years, different surfaces and methodologies for the immobilization of glycans on microarray surfaces have emerged (Park *et al.* 2008). These methodologies can be categorized into covalent or non-covalent immobilization. The solid surfaces for the immobilization that are used include glass slides, gold- or nitrocellulose membranes. The covalent immobilization involves coupling of a

functional group on carbohydrates and a reactive group on the surface through several reactions. The majority of this methods use thiol and amine chemistries (Rillahan & Paulson 2011). The non-covalent immobilization methods rely on non-covalent interactions, such as electrostatic and hydrophobic interactions. It has been shown that polysaccharides, proteoglycans and neoglycoproteins can be efficiently immobilized in membranes of nitrocellulose or oxidized polystyrene (Rillahan & Paulson 2011). Another example of the non-covalent immobilization is the work of Feizi and colleagues that have shown the noncovalent immobilization of neoglycolipids and glycolipids on nitrocellulose-coated glass slides (Liu *et al.* 2012).

1.4.1.2 Neoglycolypid-based oligosaccharide technology

The concept of NGL technology was introduced by Feizi and colleagues in 1985, as a novel approach for study the antigenicities and receptor functions of carbohydrate chains of glycoproteins (Tang *et al.* 1985). In 2002 studies Wang and colleagues (Wang *et al.* 2002) showed that polysaccharides and glycoproteins can be immobilized on nitrocellulose by noncovalent interaction. After that, Feizi and colleagues adapted their NGL technology to generate the first microarray system for complex oligosaccharides (Fukui *et al.* 2002). The technology involves conjugating oligosaccharides by microscale lipid conjugation via reductive amination to an aminolipid, 1,2-dihexadecyl-sn-glycero- 3-phosphoethanolamine (DHPE). The oxime ligation can also be used to conjugate the lipid with the oligosaccharide (Liu *et al.* 2007). The NGLs can be immobilized on solid matrices, such as nitrocellulose membranes and silica plates, and are also suitable for thin layer chromatography (TLC)-binding experiments in conjunction with mass spectrometry for deconvolution and structural characterization of the probes that are giving binding signals (Chai *et al.* 2003).

The NGL-based microarrays comprise both NGL (prepared from natural or chemically synthetized oligosaccharides) and glycolipids (natural or synthetic) (Figure 10). These probes are robotically printed onto nitrocellulose-coated glass slides at low femtomol (fmol equivalent to 10^{-15} mol) levels in a liposome formulation in the presence of carrier lipids (Liu *et al.* 2012). This approach promotes a certain level of flexibility and movement of the oligosaccharide being presented to the protein, which could be essential for particular recognition systems. The NGL-technology allows the expansion of the library of probes by the "designer" microarray approach (Figure 1.10). The term "designer" microarray, refers to a microarray of oligosaccharide probes generated from ligand-bearing glycomes to reveal the oligosaccharide ligands they harbor, so that these can be isolated and characterized (Palma *et al.* 2012; Palma *et al.* 2014)

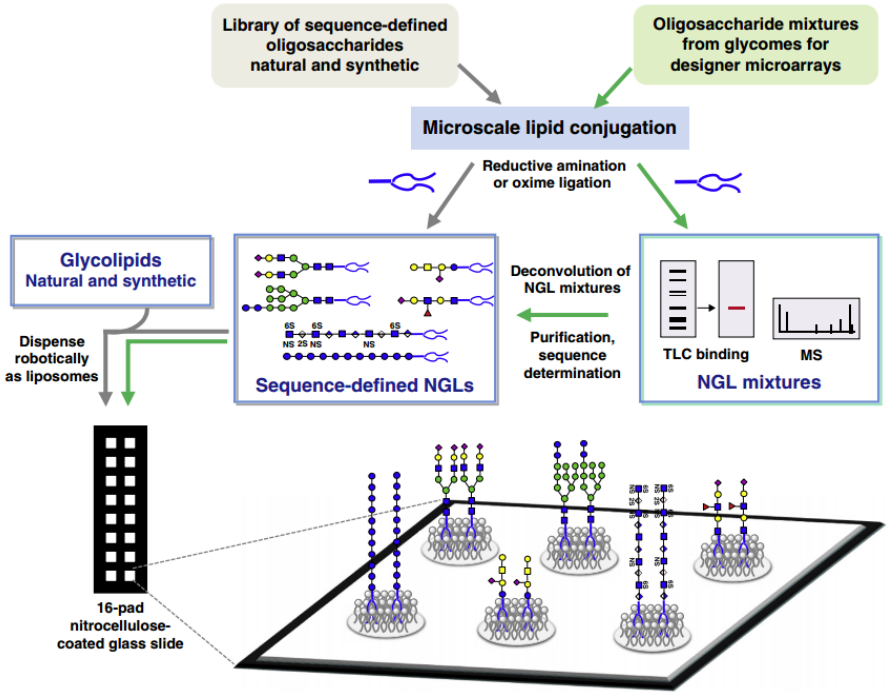


Figure 1.10 – Schematic representation of the NGL-based technology. Taken from Palma *et al.* 2014.

1.4.1.3 Biological Applications

Carbohydrate microarrays have become a powerful tool to map out interactions involving carbohydrates in a high-throughput manner and gave a massive contribution to significant advances in glycomics. Nowadays, there are several applications for the carbohydrate microarray technology in health and disease (Palma *et al.* 2014) but this section is not an exhaustive description of all the applications, instead, will highlight 2 important examples in the context of this Thesis.

The first one was the development of carbohydrate microarrays from plant polysaccharides and derived oligossacharides as a sensitive method for probing microbial CBMs and anti-plant antibodies important for plant biology (Pedersen *et al.* 2012). The second example was the implementation of the NGL-based microarrays, coupled with mass spectrometry for construction of "designer" microarrays from plant, fungal and bacterial glycomes (Palma *et al.* 2015). This microarray has been applied to identify and assign novel ligands for diverse carbohydrate recognition systems, including lectins on the immune system, therapeutic antibodies and CBMs with biological relevance.

The study of CBM-carbohydrate interactions allows the understanding of the divergent evolution of different microorganisms, as exemplified in this Thesis. It is also a means to create new strategies to improve the energy obtained by the microbial degradation of the plant cell wall carbohydrates, as well as understanding the symbiosis among these microorganisms.

1.4.2 Other biophysical methods

1.4.2.1 Microscale Thermophoresis

Microscale Thermophoresis (MST) is a recently developed technique by NanoTemper (Germany) that is based on the directed movement of molecules in a temperature gradient (Jerabek-Willemsen *et al.* 2014). Thermophoresis strongly depends on a variety of molecular properties such as size, charge, hydration shell or conformation. This technique allows a precise quantification of molecular events, independent of the size or nature of the investigated sample, due to high sensitivity (measures dissociation constants down to 1 picomolar). In a MST experiment, a microscopic temperature gradient is induced by an infrared laser, and the directed movement of molecules is detected and quantified using either covalently attached dyes, fluorescent fusion proteins or intrinsic tryptophan fluorescence (Jerabek-Willemsen *et al.* 2014).

Besides providing a precise determination of binding constants, it can also be used to derive additional information about the molecular mechanism of the investigated interaction. For instance, MST can be used to discriminate between different binding modes, and can also be used to determine interaction stoichiometries. A study from Wong and colleagues showed interactions between a protein called LysM domain with bacterial cell wall fragments and chitin oligomers using MST (Wong *et al.* 2014).

A typical MST experiment requires: one fluorescent binding partner and one nonfluorescent binding partner (the ligand). The fluorescence can be provided by an extrinsic label (e.g. NT-647 dye) covalently attached to the molecule of interest. The concentration of the fluorescent molecule is kept constant. The non-fluorescence molecule is then varied and 16 samples prepared at different concentrations. This dilution can be done in microtiter plates and a multichannel pipette to prepare multiple dilutions at once. After mixing the interaction partners, the samples are left to reach the equilibrium. The capillaries load themselves through capillary action once placed in the sample and are then loaded into the capillary tray.

The NanoTemper analysis software is used to calculate the binding parameters. After the Monolith has finished recording MST curves for each capillary the software automatically plots them on to a graph (Figure 1.11 A). The software then calculates the extent of binding by plotting the ratio between the fluorescence when the laser is on and the fluorescence before the laser is turned on. As each curve represents a different concentration of binding partner, these ratios are plotted as a function of binding partner concentration to give a binding curve (Figure 1.11 B). With the user instructions, the software automatically fits the data and provides a K_d.

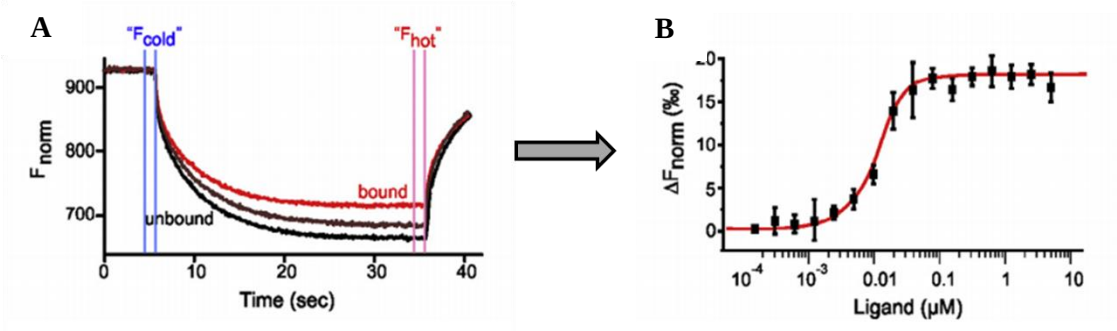


Figure 1.11– Typical MST binding experiment. A) The thermophoretic movement of a fluorescence unbound molecule (black trace) changes upon binding to a non-fluorescent ligand (red trace), resulting in different traces. B) The change in thermophoresis is expressed as the change in the normalized fluorescence (ΔFnorm) which is defined as Fhot/Fcold. Titration of the non-fluorescent ligand results in a gradual change in thermophoresis, whih is plotted as ΔFnorm to yield a binding curve that can be fitted to derive binding constants. Adapted from Jerabek-Willemsem *et al.* 2014.

1.4.2.2 Surface plasmon resonance

Surface plasmon resonance (SPR) was first introduced in early 1990s and is a powerful technique to determine specificity, affinity and kinetic parameters during the binding event of macromolecules. SPR is an optical technique that measures the refractive index changes in the vicinity of thin metal layers, such as gold, in response to biomolecular interactions. For this approach it is necessary to first immobilize the probe on the surface of the SPR chip and then a solution with the testing samples flows across this surface. During binding, changes in the SPR angle can be determined by varying the incidence angle and recording the reflected light intensity (Nguyen *et al.* 2015).

SPR technology has an important role in biomedical applications such as interaction analyses, conformational change studies, and mutation detection. SPR is widely used to study protein-carbohydrate interactions, including CBM-carbohydrate interactions measuring affinity binding constants between the two molecules (Linman *et al.* 2009).

1.4.2.3 Isothermal Titration Calorimetry

Isothermal Titration Calorimetry (ITC) is a technique used in quantitative studies of a wide variety of biomolecular interactions.

This technique works by directly measuring the heat that is either released or absorbed during a biomolecular binding event (Freire *et al.* 1990). The ITC instrument operates on the heat compensation principle where the instrumental response (measured signal) is the amount of power (micro calories per second) necessary to maintain the temperature difference between the reaction and reference cells. ITC is the only technique that can determine all binding parameters in a single experiment (Leavitt & Freire 2001). Requiring no modification of binding partners, either with fluorescent tags or through immobilization, ITC measures the affinity of binding partners in their native states.

The solution with the target biomolecule is placed in the sample cell, and a ligand solution in a corresponding buffer is placed in the syringe (Figure 1.12 A). When the ligand solution is injected into the cell, the ITC instrument detects heat that is released or absorbed as a result of the interaction (Figure 1.12 B). Injections are performed constantly, and result in peaks that become smaller as the biomolecule becomes saturated. After titration is complete, the individual peaks are integrated by the instrument software (Figure 1.12 C). An appropriate binding model is chosen and the isotherm is fitted to yield the binding enthalpy ΔH, the K_d, and the stoichiometry, n. From these data, Gibb's free energy, ΔG and entropy, ΔS are calculated (Isothermal titration calorimetry, www.malvern.com last access on 14.09.16). This provides a complete thermodynamic profile of the molecular interaction.

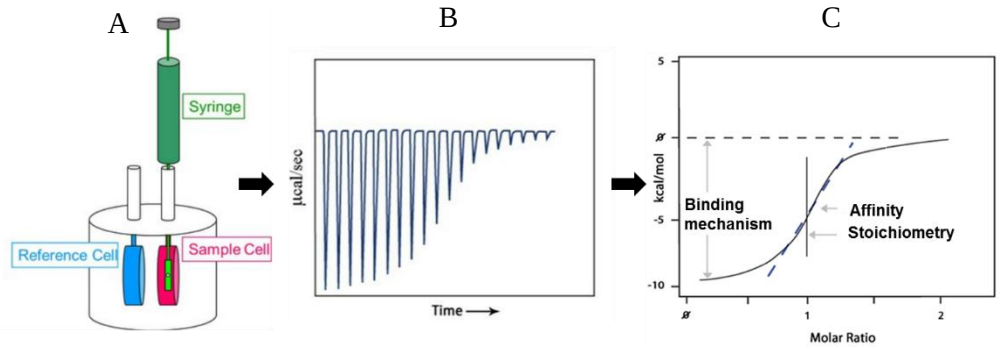


Figure 1.12 - The ITC binding experiment. A: The ligand is titrated into the sample cell. B: Raw heat data of negative peaks from an exothermic reaction. C: Data is treated with a respective model and reported. Adapted from www.malvern.com, last access on 14.09.16.

1.4.3 Structural methods

1.4.3.1 X-ray Crystallography

X-ray Crystallography is one of the most useful methods for the determination of the three-dimensional structure of a protein-carbohydrate complex. The history of crystallography has started in the year of 1895 when Wilhelm Röntgen discovered the X-ray radiation. Looking at the X-ray crystallography history one of the most important milestones was the first determination of a protein structure, myoglobin and hemoglobin, by Max Perutz and John Kendrew in the 60's decade. In the last two decades the structures solved by X-ray crystallography increased exponentially, and developments such as new software and synchrotron radiation contributed to improve the quality and precision of structures.

To obtain a three-dimensional protein structure it is necessary to go through several steps: Protein expression and purification (in case of recombinant proteins), crystallization, X-ray diffraction experiment, structure resolution and structure refinement and validation (Figure 1.13).

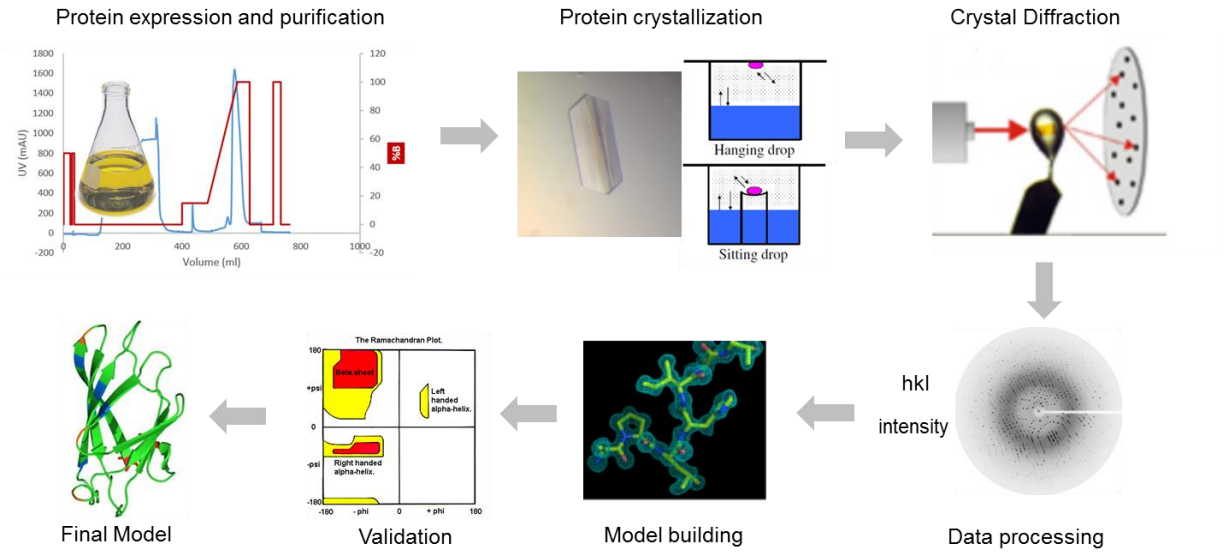


Figure 1.13 - Schematic representation of the crucial steps to obtain a threedimensional structure with X-ray crystallography. From protein expression and purification to the final model.

Obtaining high quality crystals can be the most difficult step in the resolution of a crystallographic structure. The vapor diffusion principle can be explained by the analysis of the crystallization phase diagram (Figure 1.14 A). The crystallization of a protein occurs in two stages: nucleation and growth. Initially the protein present in the drop is in the undersaturation state where, typically, the drop contains a lower reagent (precipitant) concentration than the reservoir. To achieve the equilibrium, water vapor leaves the drop and will induce the nucleation process. As water continues to leave the drop, the sample undergoes an increase in relative supersaturation. Both the sample and reagent increase in concentration as water leaves the drop to the reservoir. Equilibration is reached when the reagent concentration in the drop is approximately the same as that in the reservoir. The crystal growth process is slow and occurs in the metastable zone, by ordered and slow addition of new molecules and formation of the crystalline network (Rhodes 2006).

In vapor diffusion, crystals are prepared either using the hanging or sitting drop systems (Figure 1.14 B). Basically in the hanging drop method the crystallization drop is hanging on top of the precipitant while in the sitting drop method it is laying on top of a small bridge.

Usually, protein crystallization is a challenging process dependent on the type of protein in study. Determination of the crystallization conditions usually requires testing different screenings, which are composed of precipitant solutions with different nature and concentration of salts, pH, additives and testing different temperatures. Among all precipitants, PEG (polyethylene glycols) with a low molecular weight (400-20000) is the most used (Romão 1996).

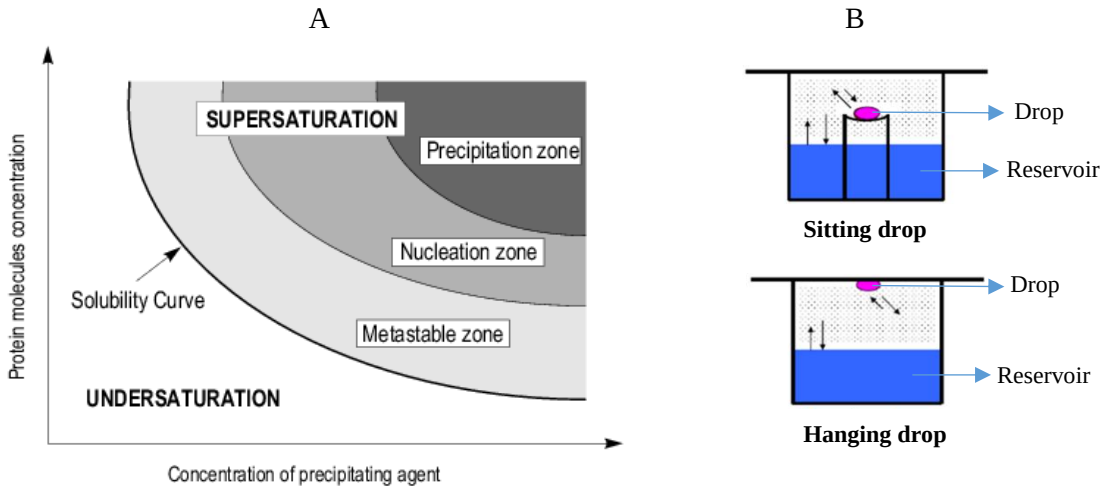


Figure 1.14 - Schematic representation of the vapor difusion method. A) Phase diagram representing the variation of protein and precipitanting agent in the protein crystallization process. B) Two techniques used for crystallization by vapor diffusion, sitting and hanging drop.

After obtaining a single crystal for an X-ray experiment, an X-ray diffractometer is used to collect crystal data. This instrument is designed to collimate a monochromatic X-ray beam through a crystal, intercept the diffracted photons and record the intensity of the reflection. Once data collection is finished, the positions of the reflections are converted into h, k, l indices and the intensities of the reflections into amplitudes (Rhodes 2006). This information, together with the estimated phase angles, allows to determine the atoms position. However, unlike the intensity of each reflection, the phase angle cannot be measured experimentally. This problem is commonly known as the Phase Problem in crystallography (Taylor 2003).

Different methodologies are used to solve the phase problem, such as molecular replacement (MR), multiple isomorphous replacement (MIR) and multiple/ single wavelength anomalous dispersion (MAD/SAD).

The MR method is used when a structure with a similar primary sequence is accessible. A sequence can be considered as similar and thus a potential model for MR method if the identity between the two sequences is superior to 25% (Taylor 2003). Patterson methods are usually used to obtain first the orientation of the model in the new unit cell and then the translation of the correctly oriented model relative to the origin of the new unit cell (Taylor 2003).

When the MR method cannot be used to solve the phase problem, other methods as multiple isomorphous replacement (which requires a derivative dataset from a protein crystal with incorporated heavy atoms) or multiple/ single wavelength anomalous dispersion (where anomalous scattering of X-rays according to the wavelength is used), have been developed.

The determination of the initial phases allows the determination of the first map of electron density, which can provide the positions of all atoms in the crystal.

1.4.3.2 Saturation-Transfer Difference

Apart from X-ray crystallography, NMR (nuclear magnetic resonance) spectroscopy is also a method of choice in the structural elucidation of macromolecules. Associated to NMR, Saturation Transfer Difference (STD)-NMR is one of the most popular ligand-based techniques for the study of protein-ligand interactions. This technique is based on the Nuclear Overhauser effect and in the observation of the ligand resonance signals. This method has the following advantages 1) No need for NMR processing information about the receptor and 2) use of small quantities of the non-labeled macromolecule. The STD-NMR method is an indispensable tool in drug discovery as it identifies binding epitope(s) at the atomic resolution of small molecule ligands, while interacting with their receptors (Bhunia *et al.* 2012; Coelho *et al.* 2015).

The STD experiment (Figure x) involves subtracting a spectrum in which the protein was selectively saturated (*on-resonance* spectrum) with signal intensities I_{SAT}, from one recorded without protein saturation (*off-resonance* spectrum), with signal intensities I_0. This subtraction gives rise to the difference spectrum ($I_{STD} = I_0 - I_{SAT}$) where only the signals of the ligand(s) that received saturation transfer from the protein (via spin diffusion) will remain (Figure 1.15).

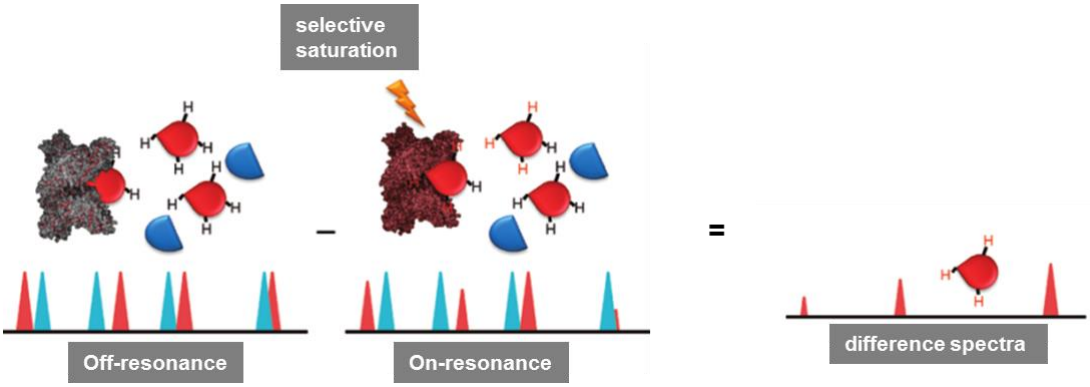


Figure 1.15 – Scheme of the STD-NMR experiment. The exchange between free and bound ligand allows transfer of magnetization from the receptor to the bound small molecule. Adapted from Viegas *et al.* 2011.

Compounds that do not bind to the receptor will not receive any saturation transfer and therefore no signals will appear in the difference spectrum. The difference in intensity as the difference between the intensities from the *off-resonance* spectrum and the intensities from the *on-resonance* spectrum ($I_{STD} = I_0 - I_{SAT}$) constitutes an indication of binding (Viegas *et al.* 2011).

This technique was applied to study CBM-carbohydrate interactions as demonstrated by Viegas and colleagues with the study of *Ct*CBM11 binding to cellohexaose (Viegas *et al.* 2008).

1.5 Objectives of this Thesis

An important step to understand efficient biodegradation of polysaccharides by cellulolytic microorganisms is to unravel the ligand specificities of the complex carbohydrate-degrading protein machinery within their genomes. The identification of the carbohydrate ligands/substrates and a full understanding of their structure-function relationships is extremely challenging. The GlycoLab group

has been addressing this complexity by combining carbohydrate microarrays with X-ray crystallography and other complementar techiques to study carbohydrate-CBM interactions. The work developed on this Thesis reflects this effort and aims to elucidate ligand specificities of identified CBMs in the genomes of two cellulolytic bacteria: *Ruminococcus flavefaciens* FD-1 and *Clostridium thermocellum* ATCC27405. The specific objectives of Thesis can be summarized as follows:

- To perform carbohydrate microarray screening analysis of 14 CAZY CBMs, including those from *Ruminococcus flavefaciens* FD-1 and *Clostridium thermocellum* ATCC27405, to characterize their carbohydrate ligand specificities. This is the subject of Chapter 2.

- To compare two recombinant protein expression protocols to investigate the best conditions for for large-scale production of CBMs; to express and purify in large-scale two of the analyzed CBMs and to evaluate the protein stability. These are the subject of Chapter 3.

- To explore the ligand specificity for chitin by a novel LysM domain (*Ct*CBM50), by performing an Isothermal Titration Calorimetry experiment and assess the interaction with insoluble chitin; to perform crystallization assays in order to structurally characterize the protein-glycan ligand complex; to perform a similarity structure model for insight representation on the predicted 3D structure for the *Ct*CBM50. These are the subject of Chapter 4.

- To explore the binding and ligand specificity for pectins of *R. flavefaciens* family 62 by Microscale thermophoresis and affinity gel electrophoresis; to perform crystallization assays in order to structurally characterize the protein; to perform a similarity structure model for insight representation on the predicted *Rf*CBM62-1 3D structure. These are the subject of Chapter 5.

Due to the high number of samples involved, parts of the work were performed in a concerted team effort together with PhD student Diana Ribeiro, under the scope of her doctoral studies.

The results obtained throughout this Thesis were presented as poster communications in 2 Meetings as indicate below. The posters presented are included in Supplementary information 3.

2016-June, D. Ribeiro, J.L.A. Brás, R. Costa, Y. Zhang, W. Chai, Y. Liu, J.A.M. Prates L.M.A. Ferreira, M.J. Romão, T. Feizi, C.M.G.A. Fontes, A.L. Carvalho& A.S. Palma. Unravelling carbohydrate binding soecifcities of CBMomes from two cellulolytic bacteria using an integrative high-throughput approach. Summer Course Glycosciences Groningen, Netherlands

2016-June, R.L. Costa, D. Ribeiro, J.L.A. Brás, M.J. Romão, T. Feizi, C.M.G.A. Fontes, A.L. Carvalho& A.S. Palma. Ligand discovery and structural-functional analysis of proteins involved in plant cell wall degradation. 5th Meeting of Portuguese Synchrotron Radiation Users, ITQB, Lisbon.

The carbohydrate microarray results described in Chapter 2 will be included in the following manuscript:

Diana Ribeiro, Joana Brás, Raquel Costa, Ana Luisa Carvalho, Angelina S. Palma *et al.* Cellulolytic bacteria express CBMomes that dictate their ecological niche polysaccharide utilization (*in preparation).*

Chapter 2 – Carbohydrate microarray screening analysis for ligand discovery

2.1 Introductory remarks

Carbohydrate microarray screening analysis is a powerful means to identify interactions between proteins and carbohydrates and thus elucidate the carbohydrate ligands of a given protein and assign its specificity.

In this chapter we applied the carbohydrate microarrays to characterize the different carbohydrate binding specificities of 14 CBMs assigned to different families in the CAZy database. The CBMs selected for this study are from two bacteria residing in different ecological niches: *Clostridium thermocellum*, an anaerobic, thermophilic soil bacterium (Cuív *et al.* 2013) and *Ruminococcus flavefaciens*, a specialist cellulolytic bacterial species characterized from the rumen, other herbivorous animals and humans (Brulc *et al.* 2011). In Table 2.1 is summarized the information about the CBMs analyzed, such as their modular organization, protein identification and assigned family. A more detailed description of each CBM is in the Supplementary Table 1 which includes in addiction the recombinant protein sequence, molecular weight and isoelectric point.

Table 2.1 - Carbohydrate binding modules (CBMs) investigated for carbohydrate binding using carbohydrate microarrays and their modular organization. CBMs under study are highlighted in blue and respective protein identification and family are depicted.

Microorganism	Modular organization	Protein ID	Family
C. thermocellum	CBM3	Cthe_0059	3
	CBM4-GH9-CBM3-DOC1	Cthe_0413	4
	SLH-SLH-SLH-CBM54-GH16-CBM4-1-CBM4-2-CBM4-3-CBM4-4	Cthe_2809	4
	GH2-CBM6-DOC1	Cthe_2197	6
	CBM6-DOC1	Cthe_2195	6
	GH43-CBM13-DOC1	Cthe_0661	13
	CBM22-GH10-DOC1	Cthe_2590	22
	CBM25	Cthe_0956	25
	GH5-CBM32-DOC1	Cthe_0821	32
	CBM42	Cthe_1273	42
	CBM50	Cthe_0300	50
R. flavefaciens	CBM6-DOC1	3747	6
	GH43-CBM13-1-CBM13-2-DOC1	2115	13
	GH30-CBM62-1-CBM62-2-DOC1	3398	62

2.2 Materials and methods

2.2.1 DNA, bacterial strains and plasmids

The protocol for cloning is property from NZYTech®. CBMs were cloned, expressed and purified by a confidential small-scale high-throughput method. For the cloning a plasmid that confers kanamycin resistance was used. This plasmid encodes for recombinant proteins containing an N-terminal tag with 6 histidines with the following sequence: MGSSHHHHHHSSGPQQGLR. For the expression, the autoinduction protocol was used (described in more detail in Chapter 3/Section 3.2.4.2).

2.2.2 CBMs quantification and SDS-PAGE analysis

The CBMs were delivered in a micro-well plate at a final volume of 180µl each in 50 mM HEPES at a pH of 7.5, 100 mM NaCl, 2 mM $CaCl_2$ and 500 mM imidazole. For the carbohydrate microarray analysis, the CBMs were analyzed directly in this buffer without the need of further buffer-exchange or desalting (described in Section 2.2.3).

The CBMs concentration was determined with a SpectraMax®190 Absorbance Plate Reader from Molecular Devices. Absorbance was measured at 280nm, which correspondes to the wavelength in which proteins absorb light due to their aromatic amino acid residues. As a control, the protein buffer (50 mM HEPES at a pH of 7.5, 100 mM NaCl, 2 mM $CaCl_2$ and 500 mM imidazole) was used. Concentration was calculated by applying the Lambert-Beer equation: $A = \varepsilon cl$, considering the reading absorbance, the molar extinction coefficient and the pathlengh of the glass.

The proteins were analysed by polyacrylamide gel electrophoresis in the presence of the anionic detergent sodium-dodecyl sulfate (SDS-PAGE) and a reducing agent β-mercaptoethanol (BME). SDS disrupts the protein structure to produce a linear polypeptide chain coated with a uniform layer of negatively charged SDS molecules. BME contributes to the protein denaturation by reducing all disulfide bonds. As a consequence, the proteins will be separated according to their molecular weight, without the influence of the intrinsic negative charge. The protocol followed was according to Roy & Kumar 2014 using resolving gel of 13 % acrylamide.

The preparation of protein samples (15 µl) was carried out by mixing the respective CBM solution (at the desired concentration) with the sample buffer (4x Tris-HCl, 10 % SDS, 0.6 M DTT, 0.012 % bromophenol blue, 30 % glycerol) and boiling in a heatblock at 100 ºC for 10 minutes. After a fast spinning, the denatured protein sample was delivered to each well. The run paramters for the electrophoresis were setup to a fixed voltage of 150 V.

2.2.3 Construction of the carbohydrate microarray

The carbohydrate microarray constructed for the analysis of CBMs was designated Fungal and Plant PS set 1 and was comprised of a total of 64 different carbohydrate probes representative of major sequences found in fungal and plant cell walls (Supplementary table 2). Of these, 48 were soluble polysaccharides from different sources and 16 were oligosaccharides prepared as lipid-linked probes, neoglycolipids (NGLs).

Among the polysaccharide samples were featuring 2 major groups of plant-related polysaccharides: 1) the hemicelluloses and 2) the pectins. The hemicelluloses included were the following: glucans containing α-(1-4) and α-(1-6)-glucose, linear β-(1-3)-glucose, linear β-(1-6)-glucose, branched β-(1-3)/(1-6) glucose or mixed linked β-(1-3;1-4)-glucose; xylans containing β-(1-4)-xylose; xyloglucans containing branched β-(1-4)-glucose;α-(1-6) xylose; arabinoxylans containing branched β-xylose;α-arabinose; arabinogalactans containing β-galactose;α-arabinose; and galactomannans containing branched β-(1-4)-mannose;α-galactose. The collection of pectin samples were received from Prof. Berit

Smestad Paulsen (University of Oslo, Oslo) and were purified from the Malian medicinal tree *Terminalia macroptera* (Zou *et al.* 2015). The sequence for these pectin samples is under analysis. Included in the microarray were also polysaccharides typically present in fungal cell walls, such as mannans (α-mannose) and glucurono-xylomannan from *Tremella fuciformi*, which is a polysaccharide that contains α-(1-3)-linked mannose backbone and is branched with xylose, glucuronic acid and fucose.

The oligosaccharides included that have defined sequences and degrees of polymerization (DP) comprised the following sequences: xylans (linear β-(1-4)-xylose DP5 and DP6), arabinans (linear α-(1-5)-arabinone DP6, DP7, DP8 and DP9), mannans (β-(1-4)-mannose DP4 and DP6), xyloglucan (branched β-(1-4)-glucose;α-(1-6) xylose, DP7, which is designated XXXG motif) and chitin, β-(1-4)-*N*-acetylglucosamine (DP5). Also included were oligosaccharide mixtures derived from chemical or enzymatic hydrolysis of xyloglucans (DP 7, DP8 and DP9) and galactomannans (DP6, DP7 and DP8). These were designated as Xylo-glucan DP8-AO (XXLG motif), Xylo-glucan DP9-AO (XLLG motif), galactomannan DP6, galactomannan DP7a, galactomannan DP7b and galactomannan DP8 (positions 57 to 62, Supplementary Table 2).

For the construction of the microarray, the probes were immobilized noncovalently onto 16-pad nitrocellulose-coated FAST™ glass slides (Z721158, Sigma), using a noncontact arrayer robot (Piezorray, Perkin Elmer, Sear Green, UK), with a spot delivery volume of approximately 330 pL.

Each probe was printed in duplicate at two levels: polysaccharides at 0.1 and 0.5 mg (dry weight) /mL (30 and 150 pg/spot) and NGLs at 5 and 15 µM (aproximatly 2 and 5 fmol/spot).

The NGLs were prepared by reductive amination with the amino lipid 1,2-dihexadecyl-*sn*-glycero-3-phosphoethanolamine (DHPE), generating DH-NGLs, or by oxime ligation with an aminooxy (AO) functionalized DHPE (AOPE), generating AO-NGLs as described. For the noncovalent immobilization NGLs were prepared as liposomes for an efficient printing (Liu *et al.* 2012). 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. The sample solutions were transferred into a 384-well plate for robotic printing according to the planned microarray layout. This part of the work was carried out by my Supervisor Dr Angelina Palma in collaboration with the group of Prof. Ten Feizi at the Glycosciences Laboratory, Imperial College London.

2.2.4 Carbohydrate microarray binding assay

In this thesis, 14 CBMs were analyzed using the Fungal and Plant PS set1. In order to perform the quality control and to validate the microarray, different proteins with known carbohydrate-binding specificities were analyzed in parallel. These included 10 carbohydrate-directed monoclonal antibodies, 4 CBMs and a lectin. The information on these and their reported specificities are presented in Table 2.2. CBMs were tested at the concentration of 5 µg/ml, and ConA lectin at 2 µg/ml. The carbohydrate-directed monoclonal antibodies from Plat Probes were probed at 1:10 ratio, as described by Moller *et al.* 2008, and from Biosuplies at 10 µg/ml.

Table 2.2 – List of the antibodies, CBMs and the lectin and their reported specificities used for quality control of the microarray set.

Monoclonal Antibodies	Specificity	Source	Methods	Reference
Anti-β-(1-3)-glucan (mouse IgG)	β-(1-3) oligosaccharide segments in β-(1-3)-glucans.	Biosupplies	ELISA	(Meikle *et al.* 1991)
Anti-β-(1-3;1-4)-glucan (mouse IgG)	Linear β-(1-3,1-4) oligosaccharide segments in β-(1-3,1-4)-glucans.	Biosupplies	ELISA	(Meikle *et al.* 1994)
Anti-Xyloglucan (rat IgM)	Xyloglucan polysaccharides; XLLG, XXLG and XXXG oligosaccharides of xyloglucan.	Plant probes	Microarray	(Pedersen *et al.* 2012).
Anti-Xyloglucan (rat IgG)	Xyloglucan polysaccharides; preferentially to the XLLG motif of xyloglucan.	Plant probes	Microarray	(Pedersen *et al.* 2012).
Anti-Xylan (rat IgM)	Unsubstituted and relatively low-substituted β-(1-4)-xylans. Can accommodate more extensive substitution of the xylan backbone with α-arabinose and binds strongly to wheat arabinoxylan.	Plant probes	Immunocytochemistry ELISA	(McCartney *et al.* 2005)
Anti-Xylan (rat IgG)	Unsubstituted and relatively low-substituted β-(1-4)-xylans.	Plant probes	Immunocytochemistry ELISA	(McCartney *et al.* 2005)
Anti-α-(1-5)-arabinan (rat IgG)	α-(1-5)- arabinan polysaccharides and α-(1-5)-arabinose oligosaccharides; α-(1-5)-arabinose residues found in the arabinan components of certain pectic polymers; arabinogalactan-proteins.	Plant probes	ELISA	(Willats *et al.* 1998)
Anti-Heteromannan (rat IgM)	β-(1-4)-linked mannan; Glucomannan and galactomannan polysaccharides; β-(1-4)-manno-oligosaccharides (DP2 to DP5).	Plant probes	Immunocytochemistry ELISA Microarray	(Marcus *et al.* 2010)
Anti-(1-4)-β-D-mannan (mouse IgG)	Linear β-(1-4)-manno-oligosaccharides in β-(1-4)-mannans and galactomannans.	Biosupplies	ELISA	(Pettolino *et al.* 2001)

Anti-β-(1-4)-Galactan (rat IgG)	β-(1-4)-galactosyl residues found in the galactan components of certain pectic polymers.; Linear tetrasaccharide in β-(1-4)-galactan.	Plant probes	Immunodot ELISA	(Jones *et al.* 1997)
Lectin	**Specificity**	**Source**	**Methods**	**Reference**
Concanavalin A (Con A)	α-mannose-linked oligosaccharides α-(1-2)- and α-(1-3)-mannose oligosaccharides	Vector (B-1005)	X-ray crystallography	(Naismith & Field 1996)
			Microarray	(Wang *et al.* 2014)
CBMs	**Specificity**	**Source**	**Methods**	**Reference**
***Ct*CBM11**	β-(1-4) and mixed-linked β-(1-3;1-4)-glucans	Prepared in house	ITC, X-ray crystallography	(Carvalho *et al.* 2004).
	Barley-derived oligosaccharides		Microarray	(Palma *et al.* 2015)
***Cm*CBM6-2**	Linear β-(1-3) and β-(1-4)-glucans Mixed-linked β-(1-3;1-4)-glucans Linear β-(1-2)-gluco-oligosaccharides with DP-2 and longer	Harry Gilbert	Microarray	(Palma *et al.* 2015)
***Ct*CBM30**	Insoluble and soluble cellulosic materials β-(1-3;1-4)-mixed glucans (lichenan and barley β-glucan)	Prepared in house	ITC	(Arai *et al.* 2003)
***Tm*CBM41**	α-glucans	Alisdair Boraston	Microarrays, ITC	(Lammerts Van Bueren *et al.* 2004)**.**
	α-(1-4)-glucans linear α-(1,4) and α-(1-6)-linked oligosaccharides derived from pullulan		Microarray	(Palma *et al.* 2015)

Before the microarray binding assays each microarray slide was scanned for Cy3 (532nm) using GenePix® 4300A microarray scanner (Molecular Devices), for visualization of the printed probe spots. After scanning, each pad of a nitrocellulose-coated 16-pad glass slide was wetted with 150 µl of water. After blotting the inverted frame to remove the water, a blocking solution (100 µL) was added to each pad and incubated for 1h at room temperature. The blocking solution was comprised of a protein solution such as 0.02% Casein with 1% BSA made in HEPES saline buffer (HBS): 5 mM HEPES pH 7.4, 150 mM NaCl; or 3% BSA (w/v) also made in HBS. The stock solutions used to prepare the different blocking solutions were 30% (w/v) bovine serum albumin (BSA) (Sigma-Aldrich, A8577) and 1% Casein (Pierce, 37528). For some proteins that require calcium for interaction 5 mM CaCl2 was added to each solution during the entire protocol.

After 1h incubation, the blocking solution was removed and the pads were washed once with HBS. After the HBS was removed, the protocol followed for CBMs, the lectin and the antibodies was different ans described below.

For **antibodies**, 100 µL of the primary antibody (prepared at the desired concentration in the blocking solution) was added to each pad and incubated for 1.30h at ambient temperature. After this period, 100 µL of biotinylated anti-species isotype secondary antibody (prepared at the desired concentration in the blocking solution) was added to the pad and incubated for 1h.

The His-tagged **CBMs** were analyzed as above, but pre-complexed with the detection antibodies as follows: protein-antibody complexes were prepared by pre-incubating the primary detection antibody (mouse anti-his) with the secondary detection antibody (biotinylated anti mouse IgG) for 15 minutes at ambient temperature, followed by addition of protein for a further 15 minutes. After this period, the proteins-antibody complexes were diluted to a final volume of 100 µl with blocking solution and the pre-complexes solution was added to the pads (Figure 2.1)

The biotinylated **ConA** was analyzed by one step, adding 100µL onto the pad at the desired concentration (prepared in the blocking solution) and incubated for 1.30h.

For the detection of binding signals, Alexa Fluor-647-labeled streptavidin (S-21374, Molecular Probes) (100 µL at 1 µg/mL in the respective blocker solution for each protein) was added to each pad and incubated for 45 minutes.

Between incubations, each pad was washed four times with HBS buffer. After the last incubation the pads were washed four times with HBS buffer and two times with Mili-Q water and dryed in the dark.

Each microarray slide was then scanned for Alexa Fluor-647 (647 nm) for detection of the binding event using the GenePix® 4300A microarray (Figure 2.1) and the Cy3 and Alexa Fluor-647 fluorescence intensity for each spot was quantified using the GenePixPro7 Software (Molecular Devices).

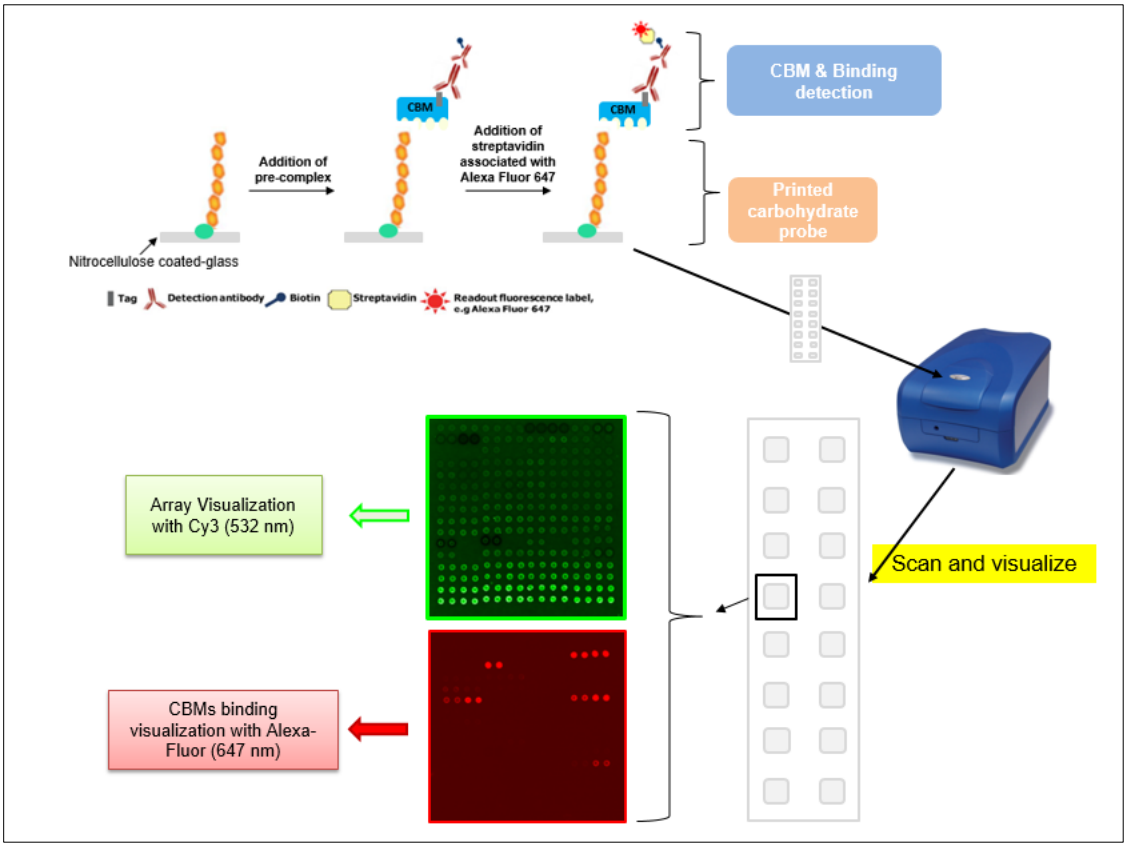


Figure 2.1 - Overview of the carbohydrate microarray binding experiment for CBMs using slides of 16 pads.

2.2.5 Carbohydrate microarray data analysis

As mentioned in the previous section a scan of the fluorescence signal was made before each assay for the Cy3 fluorophore (532 nm) and after the binding assay for the Alexa Fluor 647 fluorophore (635 nm). The first scan allows visualization and location of the probe spots and the second scan visualization of the binding event. Both scans are essential for the quantitation of the results.

The program associated with the fluorescence scanner is the GenePixPro7 Software (Molecular Devices), a program for imaging analysis and quantitation. For scanning the slides, it is necessary to choose the following parameters: laser wavelength (532 nm for Cy3 and 635 nm for Alexa Fluor 647) and corresponding emission filters; photomultiplier (PMT) gain and laser power (%) according to the intensity of the fluorescence spot signals versus the background.

After the image acquisition, the process for data analysis can be divided into three steps: 1) construction of the grid for each slide, which locates all the spots in each microarray pad referred to as block; 2) quantitation of the fluorescence intensity and 3) processing and presentation of the quantified data in form of charts, tables or matrices (heatmaps).

The first step is to construct the grid that encompasses all the spots in the respective block. This grid is constructed using the Cy3 image scan as reference and then is adjusted to the Alexa Fluor 647 slide scan (Figure 2.2) for posterior quantitation of the results.

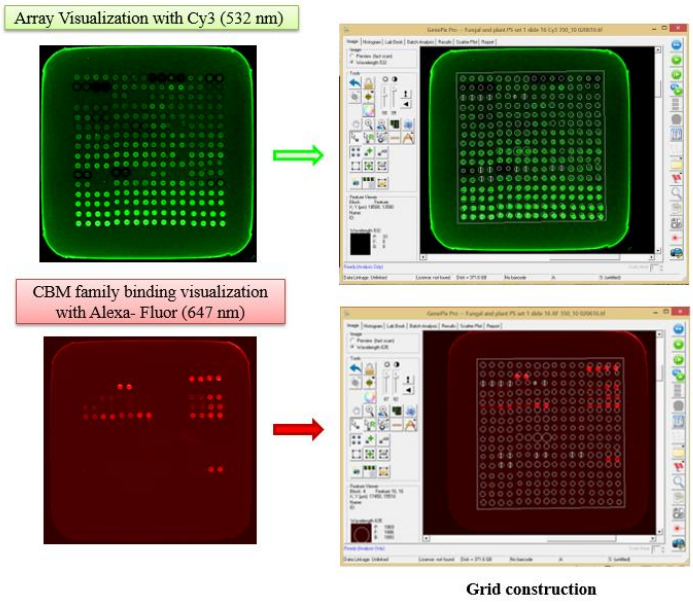


Figure 2.2 - Schematic representation of an array pad (block) spot visualization and grid construction using the GenePixPro7 Software (Molecular Devices).

The grid is saved as a GenePix settings files (.gps) and allows the location and quantitation of the fluorescence intensity of the spots in each microarray block. The program calculates the fluorescence intensity associated to each spot and outside the spot, referred to as local background. The fluorescence intensity considered for each spot is with subtracted local background. After this analysis the program saves all the scanning parameters, raw images and the numerical values in a GenePix results file (.gpr). For processing of the microarray results a dedicated microarray software developed at the Glycosciences Laboratory, Imperial College London has been developed (Stoll & Feizi 2010). This is an integrative software composed of two components (Figure 2.3): 1) the 'PiezoRay', which is an input component of all the experimental data, such as the gpr file with the raw fluorescence intensity data, overlay sample name, tested concentration, batch analyzed, date of experiment, probe set used and blocking solution; 2) the 'DisplayArray', which is an output component enabling retrieval, scrutiny and presentation of data as charts, tables and matrices (Palma *et al.* 2014). These two components communicate with a central microarray database, which holds all of the microarray metadata, the experimental conditions and information on carbohydrate probes and proteins.

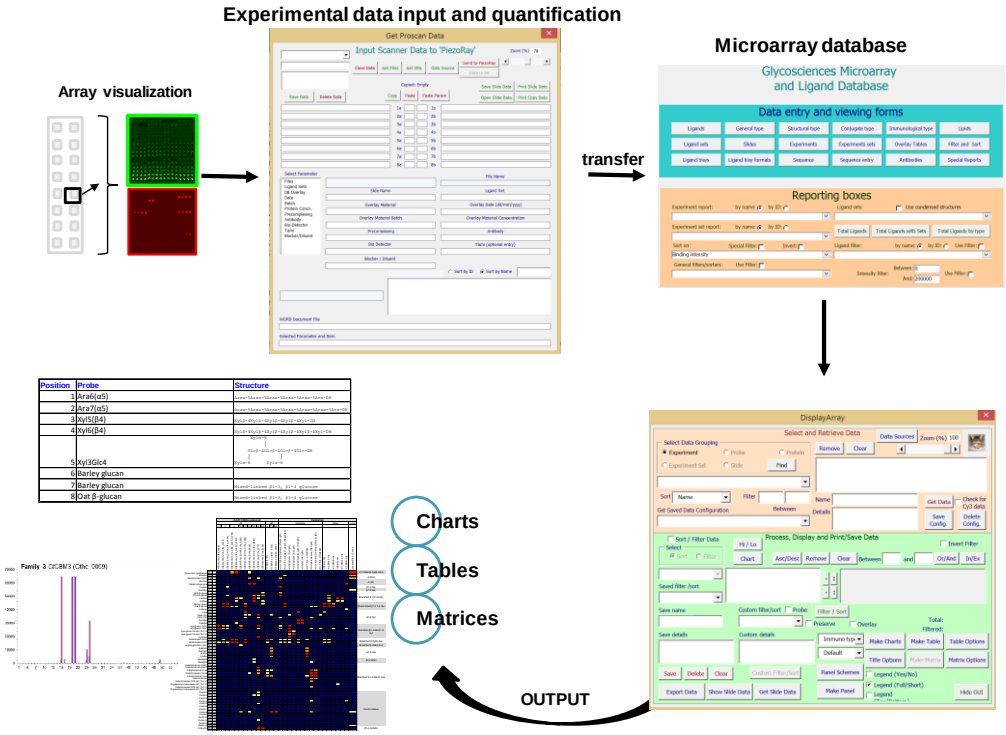


Figure 2.3 – Schematic representation of the integrated microarray data analysis: database and interactive software. Adapted from Palma *et al.* 2014.

2.3 Results and discussion

2.3.1 CBMs quantification and SDS-PAGE analysis

Before the carbohydrate microarray analysis, the 14 CBMs were quantified (Table 2.3) and analyzed by SDS-PAGE (Figure 2.4). The concentration calculated varied for each CBM, but all CBMs were obtained in good amounts to perform the microarray analysis.

Table 2.3 – Calculated concentration for each CBM

CBM	Concentration mg/ml	CBM	Concentration mg/ml	CBM	Concentration mg/ml
*Ct*CBM3 (Cthe_0059)	0.7	*Ct*CBM13 (Cthe_0661)	2.7	*Ct*CBM50 (Cthe_0300)	1
*Ct*CBM4 (Cthe_0413)	2.3	*Ct*CBM22 (Cthe_2590)	6.7	*Rf*CBM13-1 (2115)	3.7
*Ct*CBM4-3 (Cthe_2809)	5.6	*Ct*CBM25 (Cthe_0956)	1.5	*Rf*CBM6 (3747)	8.3
*Ct*CBM6 (Cthe_2197)	4.2	*Ct*CBM32 (Cthe_0821)	5.6	*Rf*CBM62-1 (3398)	4.5
*Ct*CBM6 (Cthe_2195)	4.4	*Ct*CBM42 (Cthe_1273)	1.5		

The results of the SDS-PAGE analyses confirmed the predicted molecular weight for each CBM, the purity of the sample and that no significant degradation was occurring (Figure 2.4).

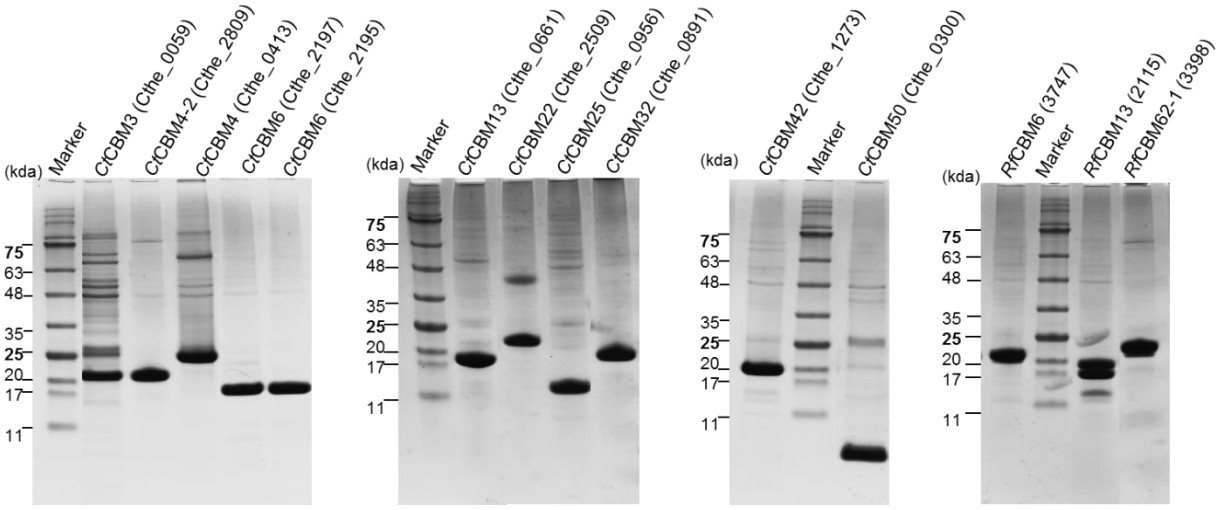


Figure 2.4 - Representative SDS-PAGE (13% acrylamide) of the 14 CBMs analyzed. Gels were performed with a constant voltage of 150V. Visualization of CBMs molecular weight was assed by coloration with Comassie blue. CBMs are identified by family and respective protein ID in each lane. M corresponds to Marker II from NZYTech®.

2.3.2 Carbohydrate microarray quality control

The quality control or validation of the carbohydrate microarray was carried out using 10 carbohydrate-directed monoclonal antibodies, 4 CBMs and 1 lectin for which the carbohydrate-binding features have been described (Table 2.2, Section 2.2.4). The rationale for the selection of these was to cover the binding to most of the carbohydrates present in the microarray. The description of the microarray set as printed is in Supplementary Table 2. For the presentation and discussion of the microarray data the following modifications to the original printed microarray set were done: repeated probes were not included; the probes were sorted according to the nature of the sample and predominant oligosaccharide sequence; and the sequence-defined oligosaccharide NGLs were highlighted. The resulting arrangement is in Table 2.4. Therefore, in this section the probe position refers to the chart/heat map position.

Table 2.4 - List of all saccharide probes analyzed in the binding charts and in the matrix (heat-map). The microarray is comprised of polysaccharide samples from different sources, representative of major sequences found in fungal and plant cell walls, and a few selected sequence-defined oligosaccharides prepared as neoglycolipids (seoquence for these is depicted). The probes are grouped according their predominant oligosaccharide sequence and glycosidic linkage and correspond to the graph/Heatmap position.

Probe	P[a]	P[b]	Predominant oligosaccharide sequence/ Monosaccharide composition
Glucurono-XyloMannan	1	42	α-(1-3)-mannose branched with xylose, glucuronic acid & fucose
Mannan	2	17	α-(1-6)- mannose
Mannoprotein	3	19	Ara (1%), Xyl (0%), Man (65%), Glc (35%)
Dextran	4	1	α-(1-6)-glucose

Pullulan	5	2	Mixed-linked α-(1-4;1-6)-glucose
Curdlan	6	3	Linear β-(1-3)-glucose
Pustulan	7	12	Linear β-(1-6)-glucose
NSG-β-glucan	8	4	Linear β-(1-3)-glucose backbone with occasional monoglucosyl β-(1-6)-glucose branches
PGG-β-glucan	9	5	Linear β-(1-3)-glucose backbone with occasional monoglucosyl β-(1-6)-glucose branches
Lentinan	10	6	β-(1-3)-glucan with β-(1-6) branching 1:1
Grifolan	11	7	β-(1-3)-glucose backbone with highly ramified oligomeric β1-6 glucose branches
Barley glucan	12	8	Mixed-linked β-(1-3;1-4)-glucose. Also contains 2% arabinose and 0.2% xylose.
Oat β-glucan	13	9	Mixed-linked β-(1-3;1-4)-glucose
Lichenan	14	48	Mixed-linked β-(1-3;1-4)-glucose
Xylan	15	40	Mixed-linked β1-3/β1-4-D-xylose 1:4
Xylan (X1)	16	13	Rha (3%), Fuc (2%), Ara (11%), Xyl (67%), Man (0%), Gal (6%), Glc (5%), Ur Ac (6%)
Xylan (X2)	17	14	Rha (3%), Fuc (3%), Ara (12%), Xyl (73%), Man (0%), Gal (0%), Glc (4%), Ur Ac (5%)
Xyl5(β4)-DH	18	53	Xylβ1-4Xylβ1-4Xylβ1-4Xylβ1-4Xyl-DH
Xyl6(β4)-DH	19	54	Xylβ1-4Xylβ1-4Xylβ1-4Xylβ1-4Xylβ1-4Xyl-DH
Xyloglucan (XG1)	20	16	Rha (2%), Fuc (5%), Ara (5%), Xyl (40%), Man (6%), Gal (13%), Glc (24%), Ur Ac (6%)
Xyloglucan (XG2)	21	22	Rha (2%), Fuc (5%), Ara (5%), Xyl (40%), Man (6%), Gal (13%), Glc (24%), Ur Ac (6%)
Xyl-glucan DP8-AO	22	57	Major component DP8 (contains small % of DP7, XXXG motif) Galβ1-2 | Xylα1-6 Xylα1-6 | | Glcβ1-4Glcβ1-4Glcβ1-4Glc-AO (designated as XXLG motif) | Xylα1-6
Xyl-glucan DP9-AO	23	58	Galβ-2 | Xylα1-6 Xylα1-6 | | Glcβ1-4Glcβ1-4Glcβ1-4Glc-AO (designated as XLLG motif) | Xylα1-6 | Galβ-2
Xyl3Glc4-DH	24	55	Xylα1-6 Xylα1-6 | | Glcβ1-4Glcβ1-4Glcβ1-4Glc-DH (designated as XXXG motif) | Xylα1-6
Arabinoxylan (AX1)	25	21	Ara (40%), Xyl (54%), Man (0%), Gal (3%), Glc (3%)
Arabinoxylan (AX2)	26	23	Ara (25%), Xyl (46%), Man (1%), Gal (3%), Glc (25%)
Arabinogalactan (AG1)	27	24	Ara (5%), Man (29%), Gal (64%), Glc (1%)
Ara6(α5)-DH	28	51	Araα1-5Araα1-5Araα1-5Araα1-5Araα1-5Ara-DH
Ara7(α5)-DH	29	52	Araα1-5Araα1-5Araα1-5Araα1-5Araα1-5Araα1-5Ara-DH
Ara8(α)-AO	30	64	Araα-5Araα-5Araα-5Araα-5Araα-5Araα-5Araα-5Ara-AO
Ara9(α)-AO	31	63	Araα-5Araα-5Araα-5Araα-5Araα-5Araα-5Araα-5Araα-5Ara-AO
Man4(β4)-DH	32	49	Manβ1-4Manβ1-4Manβ1-4Man-DH

Man6(β4)-DH	33	50	Manβ1-4Manβ1-4Manβ1-4Manβ1-4Manβ1-4Man-DH
Galactomannan (GM1)	34	25	Ara (5%), Man (62%), Gal (31%), Glc (2%)
Galactomannan (GM2)	35	26	Ara (2%), Man (87%), Gal (10%), Glc (2%)
Galactomannan Guar	36	43	Galactose: Mannose ratio = 38:62
Galactomannan Carob	37	44	Galactose: Mannose ratio = 21:79
Galactomannan Guar– Gal depleted	38	45	Galactose: Mannose ratio = 21:79
Galactomannan hydrolysed	39	46	Under analysis
Gal-Mannan DP6-AO	40	60	DP6 Mixture of β-(1-4) linked D-mannose oligosaccharides. Average Man/Gal Ratio range ~4-7
Gal-Mannan DP7a-AO	41	59	DP7 Mixture of β-(1-4) linked D-mannose oligosaccharides. Average Man/Gal Ratio range ~4-7
Gal-Mannan DP7b-AO	42	62	DP7 Prepared by controlled enzymatic hydrolysis of carob galactomannan Galβ-2 Galβ-2 | | Manβ1-4Manβ1-4Manβ1-4Manβ1-4Man-AO
Gal-Mannan DP8-AO	43	61	DP8 Prepared by controlled enzymatic hydrolysis of carob galactomannan Galβ-2 Galβ-2 Galβ-2 | | | Manβ1-4Manβ1-4Manβ1-4Manβ1-4Man-AO
Pectin-1	44	27	Under analysis
Pectin-2	45	28	Under analysis
Pectin-3	46	29	Under analysis
Pectin-4	47	30	Under analysis
Pectin-5	48	31	Under analysis
Pectin-6	49	32	Under analysis
Pectin-7	50	33	Under analysis
Pectin-8	51	34	Under analysis
Pectin-9	52	35	Under analysis
Pectin-10	53	36	Under analysis
Pectin-11	54	37	Under analysis
Pectin12	55	38	Under analysis
Pectin-13	56	39	Under analysis
GN5-AO	57	56	GlcNAcβ1-4GlcNAcβ1-4GlcNAcβ1-4GlcNAcβ1-4GlcNAc-AO

[a] Position matching the graph/Heatmap.

[b] Position matching the original microarray set.

Carbohydrate-directed monoclonal antibodies

The **anti-β-(1-3)-glucan (mouse IgG)** showed binding for the β-(1-3)-glucans curdlan, NSG-β-glucan and PGG-β-glucan (Figure 2.5 and 2.6, position 6, 8, 9 respectively), which is in accord with the reported specificity for this antibody (Meikle *et al.* 1991).

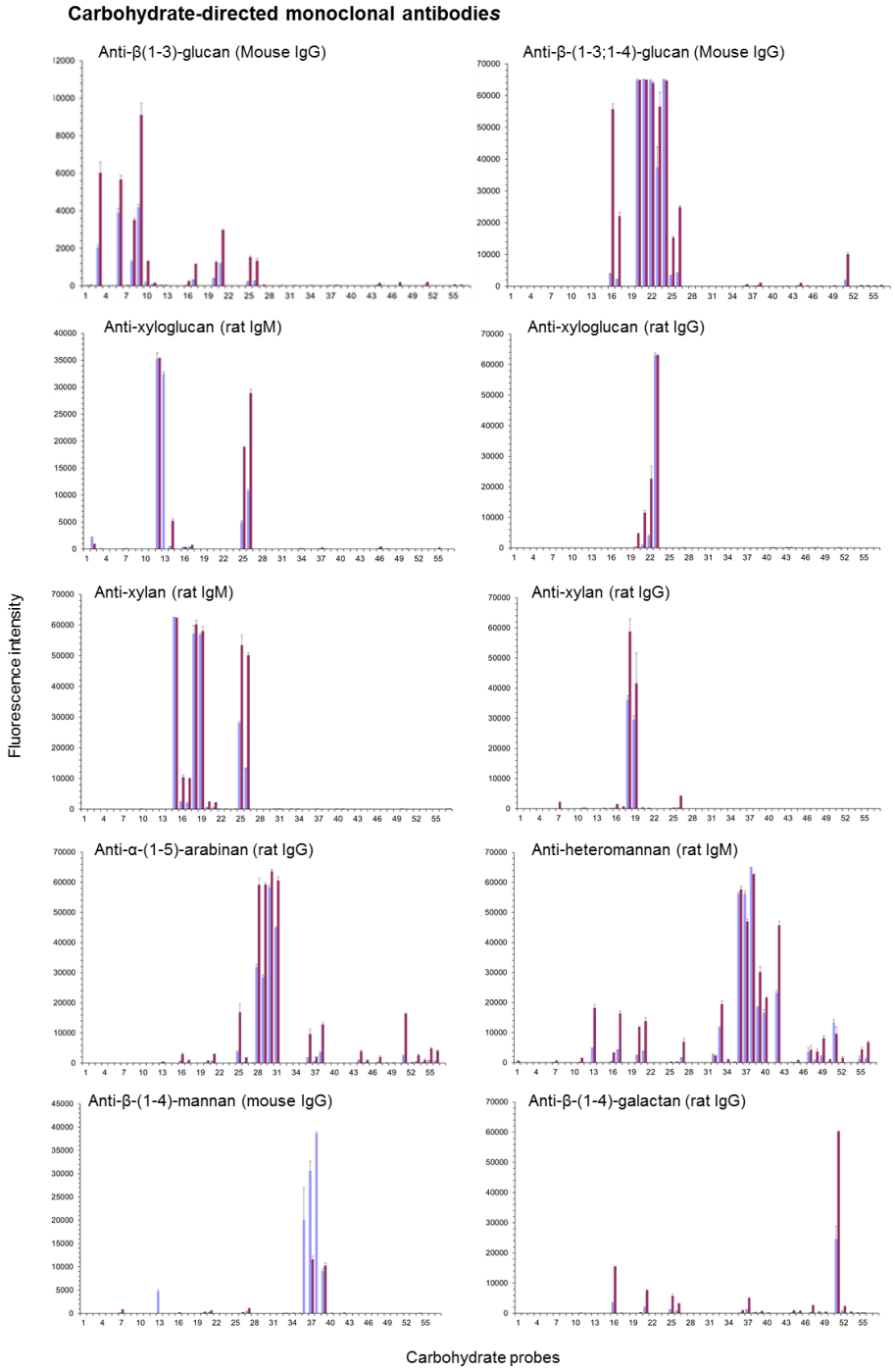


Figure 2.5 – Carbohydrate microarray data analysis of carbohydrate-directed monoclonal antibodies for quality control of the microarray set. The antibodies names are depicted on top of each graph and each recognize different epitopes and have different specificities (Table 2.2). Carbohydrate sequence information of the probes included in the microarray is in Table 2.4. The binding signals are depicted as means of fluorescence intensities of duplicate spots of probe arrayed (with error bars). Each probe was printed in duplicate at two levels: polysaccharides at 0.1 and 0.5 mg (dry weight)/Ml (30 and 150 pg/spot) and NGLs at 5 and 15 µM (aproximatly 2 and 5 fmol/spot). Quantified fluorescence intensity is plotted on the y-axis. Carbohydrate probes are plotted on the x-axis.

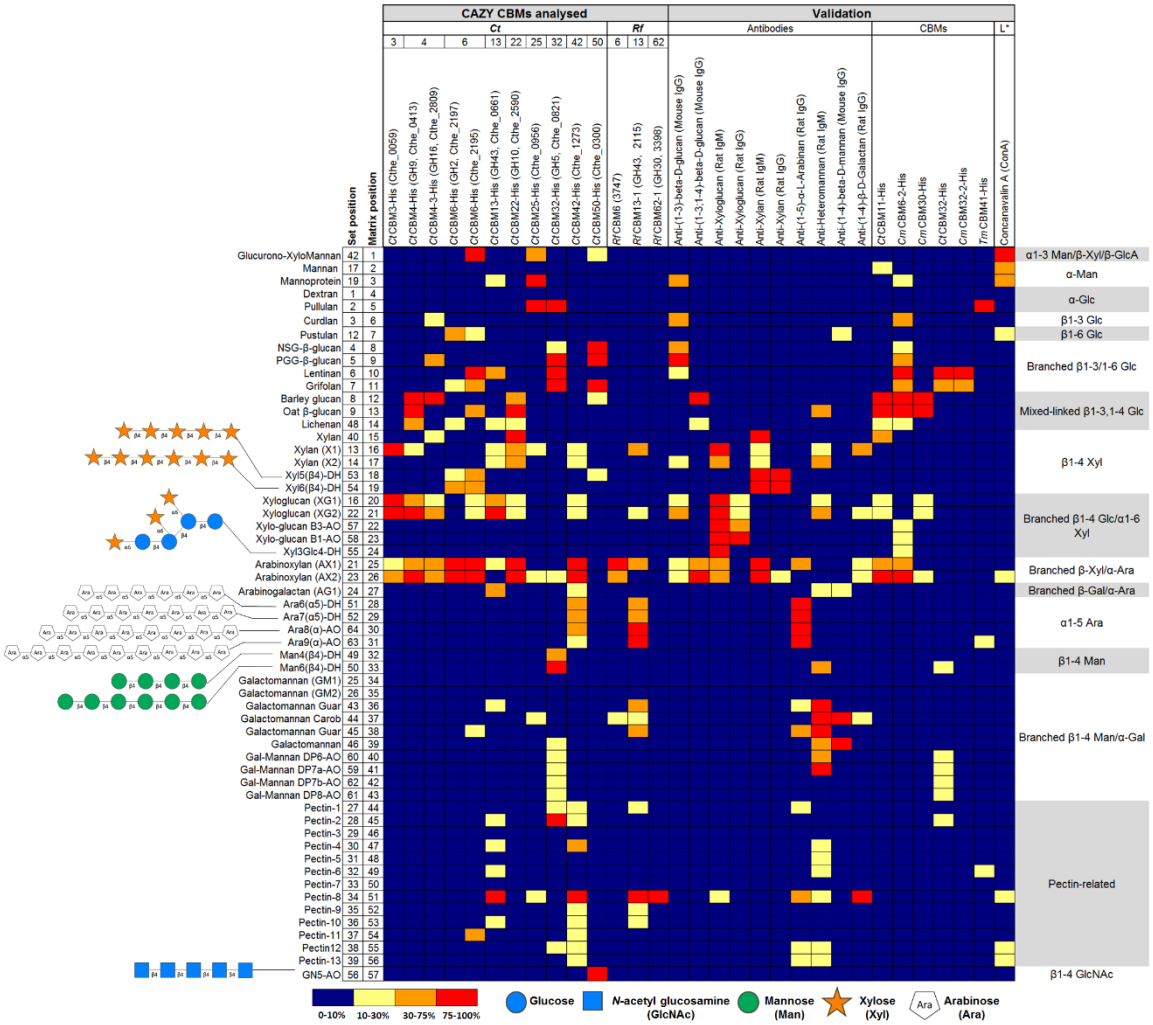


Figure 2.6 - Microarray screening analysis of CBMs from *C. thermocellum* and *R. flavefaciens*. High level heat-map of the relative binding intensities represented as a percentage of the fluorescence signal intensity relative to the probe that binds more strongly to each CBM. Sequence-defined NGLs are depicted left.

In the microarrays there was also binding detected to the mannoprotein extracted from *S. cerevisiae*, designated MP2 (Figure 2.5 and 2.6, position 3), and xyloglucan, designated as XG2 (Figure 2.5 and 2.6, position 21), extracted from plum. This could be explained by the presence of minor β1-3-linked glucose residues in these preparations.

The **anti-β-(1-3;1-4)-glucan (mouse IgG)** is described as an antibody that recognizes specifically β(1-3/1-4)-glucans and in particular the heptasaccharide sequence Glc(1-3)-Glc(1-4)-Glc(1-4)-Glc(1-3)-Glc(1-4)-Glc(1-4)-Glc (Meikle *et al.* 1994). In the microarray this antibody bound strongly to barley glucan and oat β-glucan (Figure 2.5 and 2.7 position 12, 13 respectively), both mixed linked β-(1-3/1-4) glucans, thus confirming the specificity. This antibody also bound, albeit less strongly, to arabinoxylan designated AX2 (Figure 2.5 and 2.7, position 26). This could probably be due to the presence of β-(1-4) glucose residues in this polysaccharide sample, as the content of glucose is relatively high (25%) (Table 2.4).

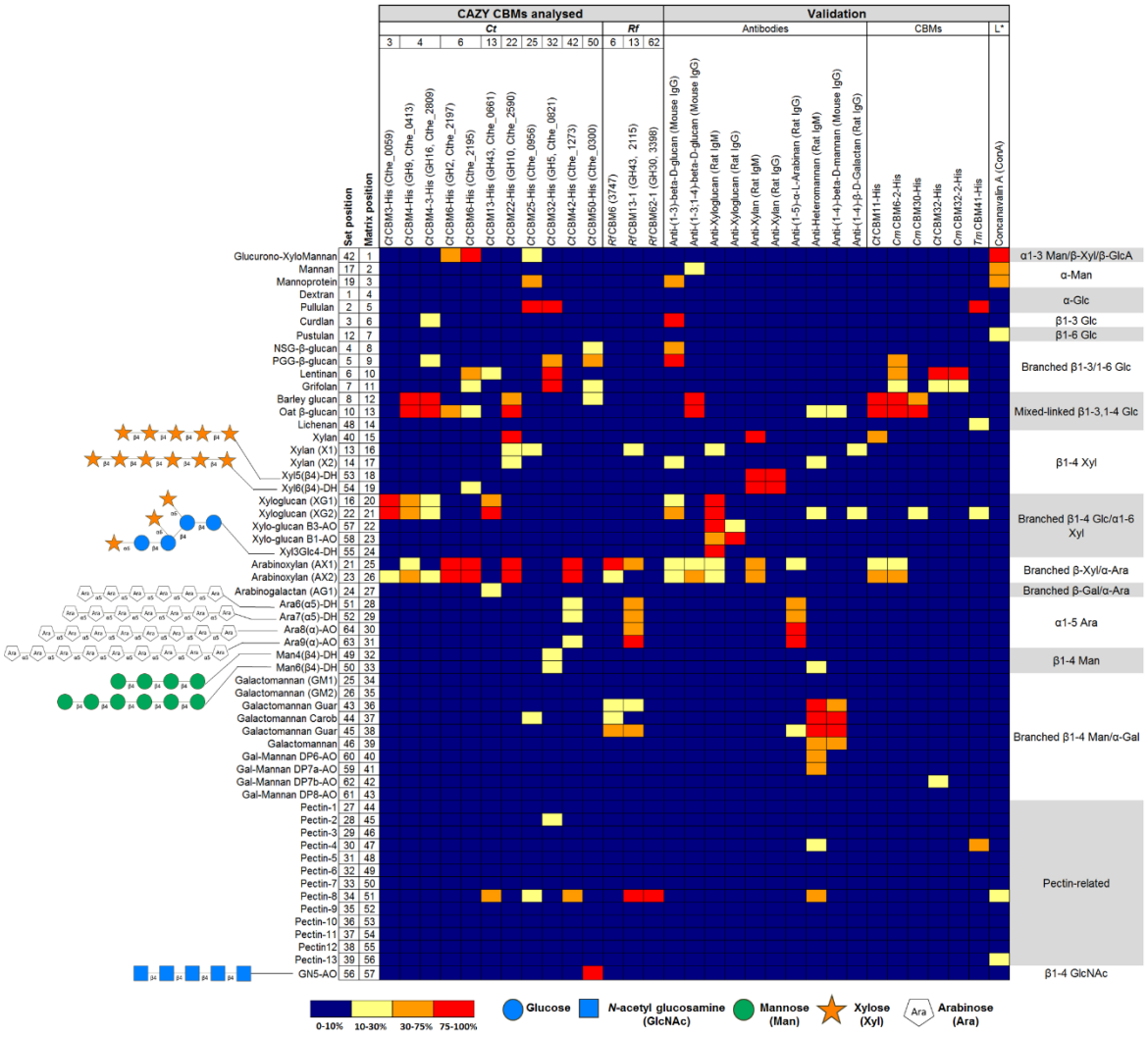


Figure 2.7 - Microarray screening analysis of CBMs from *C. thermocellum* and *R. flavefaciens*. Low level heat-map of the relative binding intensities represented as a percentage of the fluorescence signal intensity relative to the probe that binds more strongly to each CBM. Sequence-defined NGLs are depicted left.

The **anti-xyloglucan (rat IgM)** showed strong binding to the xyloglucan polysaccharides included in the microarray and also to all the xyloglucan NGL probes (Figure 2.5 and 2.7, position 16, 17, 20, 21, 25, 26 and Figure 2.5 and 2.6, position 22, 23, 24 respectively). In comparison, the **anti-xyloglucan (rat IgG)** showed a more restricted binding to two of the NGLs, strongly to xylo-glucan DP9, which contains substitution of the β-(1-4)-linked-glucose tetrasaccharide backbone by xylose and galactose (XLLG motif), and less to xylo-glucan DP8 that is one residue less substituted with galactose (XXLG) (Figure 2.5 and 2.7, position 22, 23 respectively). The heptasaccharide Xyl3Glc4 (XXXG motif) (position 24) with no galactose substitutuion was not bound by this antibody. These results are in agreement with the reported specificities for these two antibodies (Pedersen *et al.* 2012) which show anti-xyloglucan (rat IgG) is more specific for binding with xyloglucans with XLLG motif than anti-xyloglucan (rat IgM) that recognizes substituted and low substituted motifs (XLLG, XXLG and XXXG). Thus, the different binding profiles of two anti-xyloglucan antibodies gave us information about the degree of substitution of the different xyloglucan probes included in the microarray. Only the anti-xyloglucan (rat IgM)

showed binding to the xylan polysaccharides X1 and X2 (Figure 2.5 and 2.7, position 16, 17, respectively), this could be explained by the common xylose-oligosaccharide motif recognized by this antibody.

The **anti-xylan (rat IgM)** and **anti-xylan (rat IgG)** showed a binding profile according to what is reported on literature (McCartney *et al.* 2005). Both antibodies have the ability to bind unsubstituted and relatively low-substituted xylans, which was confirmed in the microarray analysis by the strong binding to the unsubstituted β-(1-4)-linked xylose penta- and hexasaccharide NGL probes (Figure 2.5 and 2.6, position 18, 19, respectively). However, the anti-xylan (rat IgM) can accommodate a more extensive and complex carbohydrate chain, which was confirmed by the strong binding to the xylan polysaccharide and albeit binding to arabinoxylan AX1 and AX2 (Figure 2.5 and 2.7, position 15, 25, 26 respectively), and also corroborated by the monosaccharide composition analysis (Table 2.4).

The **anti-α-(1-5)-arabinan (rat IgG)** showed resctricted binding to linear α-(1-5)-arabino oligosaccharides (Figure 2.5 and 2.6, position 28, 29, 30, 31). This is in conformity with this antibody being highly specific for this type of oligosaccharides (Willats *et al.* 1998).

The reported specificity of **anti-heteromannan (rat IgM)** (Marcus *et al.* 2010) was also confirmed in this microarray, the antibody interacts with galactomannan polysaccharides (Figure 2.5 and 2.7, position 36-39), galactomannan oligossacharides with DP6 and DP7 (Figure 2.5 and 2.6, position 40,41) that have only galactose and mannose and with linear hexa β-(1-4)-mannose NGL. In this experimental condition binding was also observed to pectin-8 (sequence under analysis). Binding to this pectin sample could be explained by the occurrence of β-(1-4)-galactosyl residues that have been identified in pectic galactan domanains.

In this experiment the **anti-β-(1-4)-mannan (mouse IgG)** recognizes galactomannans which is in accordance with what was described in the literature (Pettolino *et al.* 2001) . This antibody should also recognize linear (1-4)-β-manno-oligosaccharides in (1-4)-β-mannans and galactomannans. However, recognition for linear β-(1-4)-mannose NGLs was not observed for this experimental condition. In addition, strong binding for galactomannan carob and galactomannan guar was observed. The two polysaccharides have the same galactose:mannose ratio (21:79) which is consistent with the results (Figure 2.5 and 2.7, position 37, 38). Also, there was a weaker binding for the galactomannan and for the galactomannan guar with a different galactose:mannose ratio (38:62) (Figure 2.5 and 2.7, position 36, 39).

Anti-β(1-4)-galactan (rat IgG) reported specificity pointed to binding with (1-4)-β-D-galactosyl residues found in the galactan components of certain pectic polymers (Jones *et al.* 1997). The anti-(1-4)-β-D-galactan tested in the microarray is highly specific for the pectin-8 and presents a weaker interaction with xylan (X1) that contains galactose residues (Figure 2.5 and 2.6, position 51, 16 respectively). The results are consistent with the reported literature. There are some evidences that pectin-8 may have (1-4)-β-D-galactosyl residues in its composition.

Model carbohydrate-binding modules

The carbohydrate binding specificity of ***Ct*CBM11** for β-(1-3;1-4)-mixed linked glucans has been extensively analysed by different methods, including ITC, STD-NMR, carbohydrate microarrays and X-ray crystallography (Carvalho *et al.* 2004; Palma *et al.* 2015, unpublished data). In this microarray analysis, *Ct*CBM11 showed strong binding to barley- and oat-β-glucan (Figure 2.7 and 2.8, position 12, 13, respectively), which is consistent with the reported specificity. *Ct*CBM11 also bound, although less strongly, to xylan and arabinoxylan (AX2) (Figure 2.7 and 2.8, position 15, 26, respectively). Binding to arabinoxylan (AX2) could be probably due to the presence of β-(1-4) glucose residues in this polysaccharide sample, as the content of glucose is relatively high (25%).

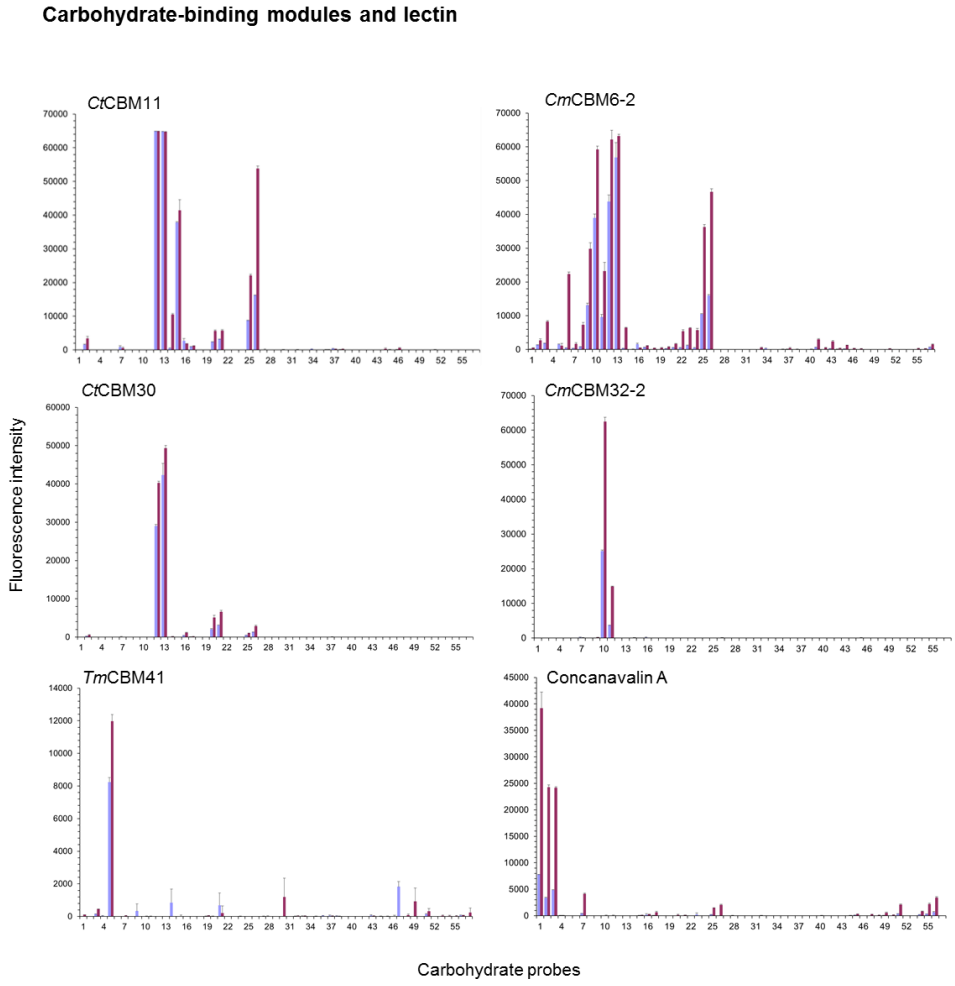


Figure 2.8 – Carbohydrate microarray data analysis of model carbohydrate-binding modules from different families and a plant lectin for quality control of the microarray set. The protein names are depicted on top of each graph and each recognize different epitopes and have different specificities. Carbohydrate sequence information of the probes included in the microarray is in Table 2.4. The binding signals are depicted as means of fluorescence intensities of duplicate spots of probe arrayed (with error bars). Each probe was printed in duplicate at two levels: polysaccharides at 0.1 and 0.5 mg (dry weight)/Ml (30 and 150 pg/spot) and NGLs at 5 and 15 µM (aproximatly 2 and 5 fmol/spot). Quantified fluorescence intensity is plotted on the y-axis. Carbohydrate probes are plotted on the x-axis.

***Cm*CBM6-2** showed a broad range of interaction, mostly in agreement with the reported binding for this CBM (Henshaw *et al.* 2004; Palma *et al.* 2015) which contains two binding clefts with different specificities for glucose and xylose oligosaccharides. In addition to binding with mixed-linked β-(1-3;1-4) glucans barley and oat-β-glucan, *Cm*CBM6-2 showed interaction with branched β-(1-3)/(1-6) glucans such as curdlan, PGG-β-glucan, lentinan and grifolan (Figure 2.6 and 2.8, position 5, 9, 10, 11 respectively), and to arabinoxylans AX1 and AX2 (Figure 2.6 and 2.8, position 25, 26, respectively).

***Ct*CBM30** is characterized by strongly binding to mixed linked β-(1,3-1,4)-glucans (Arai *et al.* 2003). Our results are concordant with the literature as *Ct*CBM30 showed binding to barley and oat-β-glucan (Figure 2.6 and 2.8, position 12, 13, respectively).

***Cm*CBM32-2** showed binding to lentinan and grifolan (Figure 2.6 and 2.8, position 10, 11, respectively), both are linear β-(1-3) glucose with branched β-(1-6) but the second probe is more ramified. *Cm*CBM32-2 reported specificity is characterized by interaction with branched β-(1-3)/(1-6) gluco-oligosaccharides (Palma *et al.* 2015) which is consistent with our results.

The results observed for ***Tm*CBM41** are consistent with the reported specificity (Palma *et al.* 2015) as the *Tm*CBM41 is highly specific for α-glucans, in this case for a mixed-linked α-1-4;1-6 glucose, pullulan (Figure 2.6 and 2.8, position 5).

Plant lectin

It is well established that **Concanavalin A** is highly specific for α-mannose-linked oligosaccharides (α-(1-2) and α-(1-3)-mannose) (Wang *et al.* 2014). The results of the microarray analysis showed a strong interaction with glucuro-xylomannan, mannan and mannoprotein (MP2) (Figure 2.6 and 2.8, position 1,2,3 respectively), in concordance with the presence of those motifs in these macromolecules (Table 2.4).

In sum, the antibodies, CBMs and lectin showed different binding patterns, some exhibiting a broad binding, others highly restricted binding, and in overall the microarray results are in agreement with the reported specificities (Table 2.2). To test the reproducibility and validate the array the analyses were also carried out using a newly prepared carbohydrate microarray of similar composition. The comparative data confirmed the validation results. Thus, the carbohydrate microarray is functional and can be used for carbohydrate binding screening analysis of CAZy CBMs, and the quality control data will be useful to interpret the binding results obtained.

2.3.3 Carbohydrate microarray analysis of CAZy CBMs

The 14 CBMs analyzed belong to 12 different CAZy families. The CBM families showed different binding patterns which reflects the diversity of recognition systems as described below.

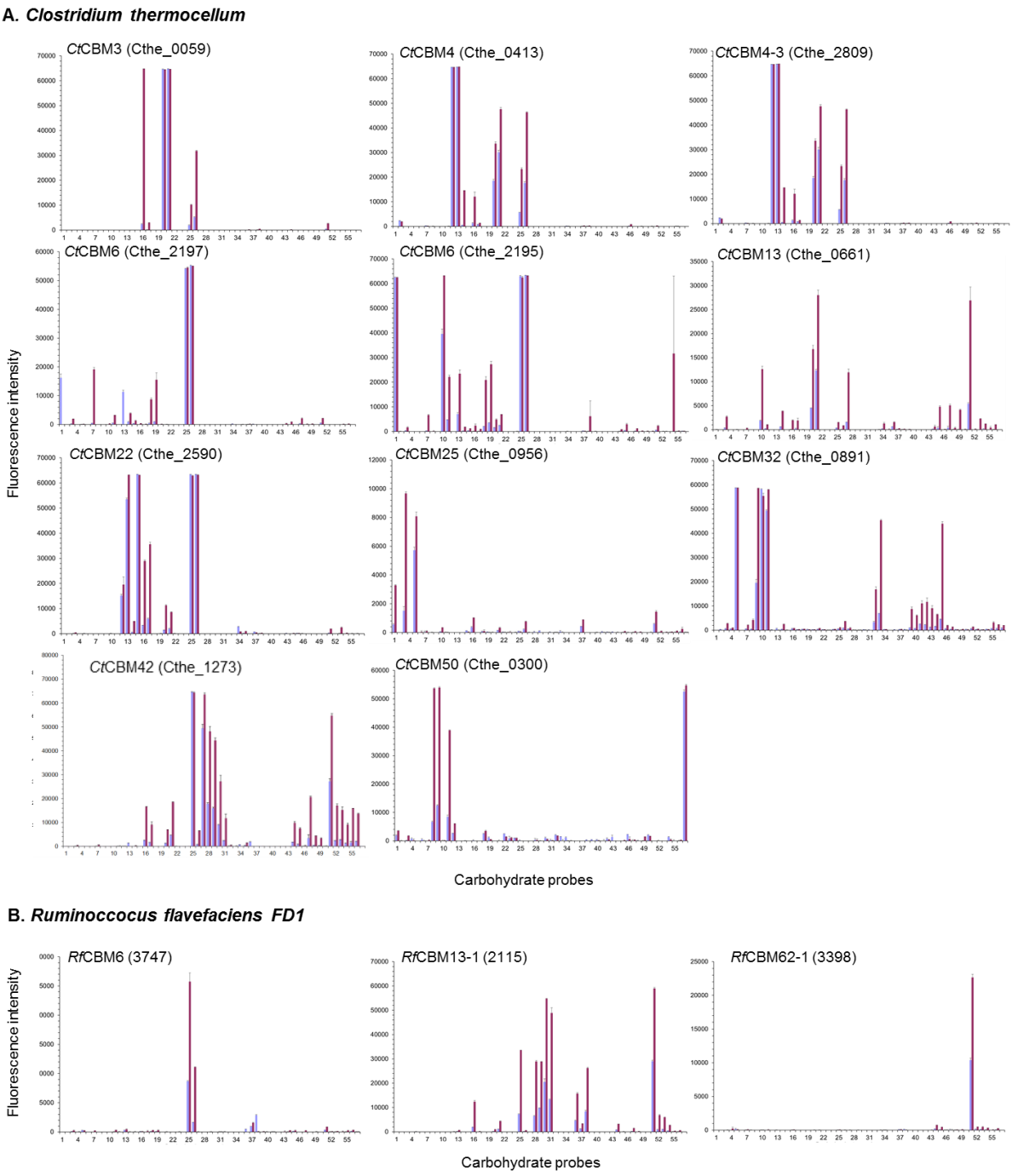


Figure 2.9 - Carbohydrate microarray data analysis of CAZy CBMs. A) CBM families analyzed from *Clostridium thermocellum*. B) CBM families analyzed from *Ruminococcus flavefaciens* FD-1. The binding signals are depicted as means of fluorescence intensities of duplicate spots of probe arrayed (with error bars). Each probe was printed in duplicate at two levels: polysaccharides at 0.1 and 0.5 mg (dry weight)/ml (30 and 150 pg/spot) and NGLs at 5 and 15 µM (approximatly 2 and 5 fmol/spot). Quantified fluorescence intensity is plotted on the y-axis. Carbohydrate probes are plotted on the x-axis.

***Ct*CBM3** displayed a highly specific binding for xyloglucan polysaccharides XG1 and XG2 (Figure 2.7 and 2.9, position 20, 21 respectively). CBMs 3 are usually associated with both cellulosomal scaffoldins and family 9 glycoside hydrolases (GH9s), which are multi-modular enzymes that act on insoluble cellulosic substrates. The solved X-ray crystal structures of these modules have established a cellulose-binding mechanism based on stacking interactions between the monosaccharide residues of cellulose and a planar array of aromatic residues located at the CBM3 surface (Yaniv *et al.* 2014). In this particular case, *Ct*CBM3 is not associated to a GH. As cellulose is an insoluble polysaccharide it was not included in the carbohydrate microarray, however the binding to xyloglucans can be explained by the β-(1-4)-glucose chain which is the target for CBM3s. The fact that this CBM could not bind to the xyloglucan NGLs, in contrast to the anti-xyloglucan (rat IgM), imply that a longer backbone chain than the tetrasaccharide might be required for binding. The *Ct*CBM3 also showed binding to xylan X1 and arabinoxylans AX1 and AX2, at the high level of the polysaccharide (Figure 2.6 and 2.9, position 16, 25, 26 respectively). These results could be explained by the presence of a β-linked glucose oligomer in these samples, as they contain glucose in their composition and the binding to AX2, which has a higher percentage of glucose, was stronger (Table 2.4).

The two ***Ct*CBMs from family 4** showed a similar binding profile in which both bound strongly to the mixed-linked β-(1-3;1-4) glucans barley-glucan and oat-β-glucan (Figure 2.5 and 2.7, position 12,13 respectively). Binding was also detected to xyloglucan XG1 and XG2, arabinoxylan AX1 and AX2 (Figure 2.7 and 2.9, position 20, 21, 25, 26 respectively) stronger for *Ct*CBM4 (Cthe_0413) then to *Ct*CBM4-3 (Cthe_2809). *Ct*CBM4 (Cthe_0413) is associated to a GH9, a glucanase, and *Ct*CBM4-3 (Cthe_2809) is associated to a GH16, glucan endo-beta-1,3-D-glucosidase. Different studies reported in the literature showed a diverse range of specificities for this *Ct*CBM family, including binding to crystalline and amorphous cellulose, xylans, β-glucans, α-glucans, chitin and chitosan (Alahuhta *et al.* 2011).

The two ***Ct*CBMs from family 6** displayed strong binding to arabinoxylans AX1 and AX2 (Figure 2.7 and 2.9, position 25, 26, respectively). These CBMs also showed a significant interaction with linear β-(1-4)-linked xylose NGLs (Figure 2.6 and 2.9, position 18, 19) and with glucan polysaccharides glucurono-xylomannan, lentinan, and oat-β-glucan (Figure 2.7 and 2.9, position 1, 10, 13 respectively). The reported specificity for this family pinpoints the nonreducing termini of *xylo-* and *gluco-*oligosaccharides (Correia *et al.* 2011) and xylose oligosaccharides (Griffiths & Sánchez-serrano 2001) as recognition features. The recognition for xylo- and gluco-oligosaccharides is reflected in this microarray as well as the recognition for xylose residues in arabinoxylans AX1 and AX2 and in the complex fungal polysaccharide glucurono-xylomannan. **The family 6 CBMs from *R. flavefaciens* FD-1** (*Rf*CBM6 3747) showed strong binding exclusively to the arabinoxylans AX1 and AX2 (Figure 2.6 and 2.9, position 25,26, respectively). There were no previously reported CBMs characterized in this *Rf*CBMs family and these microarray analyses first identified its specificity. The specificity appears to

be more restricted for arabinoxylan polysaccharides than for *Ct*CBMs investigated, as binding to glucan polysaccharides was not observed.

***Ct*CBM13** bound strongly to polysaccharides that possess different major sequences: xyloglucans XG1 and XG2, pectin-8, arabinogalactan AG1 and lentinan (Figure 2.6 and 2.9, position 20, 21, 51, 27, 10, respectively). In the literature, a *Ct*CBM13 is reported as to bind to galactose and galactose-containing oligosaccharides (galactobiose, galactotriose and lactose) (Jiang *et al.* 2012). The microarray binding pattern for this CBM may be explained by the presence of a common galactose sequence in the xyloglucans XG1 and XG2 and arabinogalactan (Table 2.4), and possibly in pectin-8 as this pectin was strongly bound by the anti-β-(1-4)-D-galactan (Figure 2.6 and 2.9, position 51). Comparatively, ***Rf*CBM13-1** displayed the same strong binding to the pectin-8 probe but also showed binding to α-(1-5)-linked arabinose NGLs (DP 6 to DP9) (Figure 2.6 and 2.9, position 28-31, respectively), revealing a novel specificity for this microorganism CBM family. Bindind to xylan X1 and arabinoxylan AX1 (Figure 2.6 and 2.9, position 16, 25) was revealed. Also binding to galactomannan guar with different galactose:mannose ratios was detected (Figure 2.6 and 2.9, position 36, 38).

As reported in the literature **family 22 *Ct*CBMs** bind preferentially to xylans and to xylooligosaccharides (Charnock *et al.* 2000). This *Ct*CBM22 is normally associated to a GH10, which exibits a β-xylanase activity. Often CBMs that are associated with xylanases, bind to xylans and to xylooligosacharides. The *Ct*CBM22 analyzed in the microarray bound strongly to xylan, arabinoxylans AX1 and AX2 (Figure 2.7 and 2.9, position 15, 25, 26, respectively), which is consistent with the reported binding for this CBM family. However, this CBM also bound to mixed-linked β-(1-3/1-4)-glucans, barley and oat β-glucan (Figure 2.7 and 2.9, position 12, 13, respectively).

***Ct*CBM25** showed strong binding to mannoprotein MP2 and pullulan, binding less strongly to glucurono-xylomannan (Figure 2.6 and 2.9, position 3, 5, 1, respectively). This CBM is not yet characterized and there are no reports on the binding specificity for members of *Ct*CBM25. However, Boraston and colleagues reported a study on the recognition of granular starch from starch corn (branched α-(1,4)/(1-6) glucose) and α-glucooligosaccharide recognition, specially those containing α-1-6-linkages by a CBM25 from *Bacillus halodurans* (Boraston *et al.* 2006). This specificity is in accord with the *Ct*CBM25 binding in the microarrays to pullulan, which is a mixed-linked α-(1-4/1-6) glucan polysaccharide (Table 2.4).

***Ct*CBM32** exhibited strong binding to glucan polysaccharides pullulan, lentinan and grifolan (Figure 2.7 and 2.9, position 5, 10, 11, respectively). This CBM is appended to GH5, a catalytic module that displays endomannanase activity, and this CBM family from *C. thermocellum* shows binding affinity for nonreducing ends of β-1,4-mannooligosaccharides and β-1,4-gluco-oligossaccharides (Mizutani *et al.* 2012). According with this reported specificity, the *Ct*CBM32 also showed good binding to β1-4-linked mannose oligosaccharides with DP5 and DP6 (Figure 2.6 and 2.9, position 32, 33, respectively). Thus, the specificity of this CBM needs to be further explored to understand the binding to the glucan polysaccharides.

***Ct*CBM42** showed strong binding to arabinoxylan (AX1), arabinogalactan (AG1), α-(1-5)-arabinose NGLs and pectin-8 (Figure 2.6 and 2.9, position 25, 26, 28, 29, 30, 31 and 51, respectively). In the literature, three *Ct*CBMs from family 42 are described appended to GH43 (putative arabinofuranosidases) and these CBMs display similar ligand specificities binding preferentially to arabinoxylans and arabinan (Ribeiro *et al.* 2010). Although this CBM is not appended to a GH, the specificity is similar to what is reported on the literature.

***Ct*CBM50** showed a strong and very specific binding to the pentasaccharide of β-(1-4)-linked *N*-acetylglucosamine (GlcNAc), derived from chitin (Figure 2.6 and 2.9, position 57). *Ct*CBM50 sequence is a LysM (Lysin Motif) domain. Lysins, also known as endolysins or murein hydrolases, are highly evolved hydrolytic enzymes produced by bacteriophage to digest the bacterial cell wall peptidoglycan for phage progeny release. The motif typically ranges in length from 43 to 50 amino acid residues and recognizes the *N*-acetylglucosamine (GlcNAc) moieties of peptidoglycan and chitin (Buist *et al.* 2008). Thus the microarray results support that *Ct*CBM50 is a functional LysM domain.

***Rf*CBM62-1** was the CBM that showed the most restricted binding in the microarray analysis as it bound to one probe only, the pectin sample designated as pectin-8 (Figure 2.6 and 2.9, position 51). This pectin sample is among a group of pectins (designated here pectin-1 to pectin-13) that were isolated from African plants for which structure is currently being evaluated. The fact that the pectin-8 was strongly bound by the anti-β-(1-4)-galactan, leads us to hypothesize that this pectin contains a β1-4-linked galactan domain, and that this could be the recognition motif for the *Rf*CBM62-1. Of interest is that in the literature there is only one reference for a characterized CBM62 family from *C. thermocellum.* In this study the authors revealed a novel CBM family from *C. thermocellum* that recognizes xyloglucan and galactomannan-derived oligosaccharides (Montanier *et al.* 2011). Thus further studies will be required to access the specificity of the *Rf*CBM62-1 and of this CBM62 family in *R. flavefaciens*.

2.4 Conclusions

An important outcome from the carbohydrate microarray analysis was the possibility to understand the versatility of plant and fungal saccharide sequences and their recognition by different CBM families.

The Fungal and Plant PS set1 microarray was proven to be an effective, sensitive and high-throughput tool for screening newly identified CBMs or CBMs assigned to known families in the CAZy database for carbohydrate binding. These features enabled that CBMs from two different organisms could be simultaneously probed and compared. For the analyzed CAZy CBMs, it was observed that CBMs from the same family and organism could display a similar specificity. However, when the same CBM family is compared in different organisms the specificity could be considerably different. An important conclusion is that CBMs from *R. flavefaciens* FD1 revealed to be more specific with a binding pattern less broad than CBMs from *C. thermocellum*. These different binding patterns on the microarray could reflect adaptive pressures that the microorganisms suffered in their respective ecological niches, translating in divergent evolution of the proteome. The ruminants have diet rich in xylans/arabinoxylans

which is translated in a high content of GHs from family 11, endo-β-1-4 and endo-β-1-3 xylanases, and CBMs from family 22, with xylan binding function, in the *R. flavefaciens FD-1* genome (Berg Miller *et al.* 2009).

The results from this initial carbohydrate microarray screening constitute a functional starting point to: 1) target and select CBMs for assignment of carbohydrate-ligand specificities; 2) develop sequence-defined designer oligosaccharide microarrays; and 3) structurally characterize novel CBM-oligosaccharide complexes.

Chapter 3 – Expression, purification and stability analysis of CBMs

3.1 Introductory remarks

The objectives of the work described in this chapter were two-fold: 1) to investigate the best conditions for large-scale recombinant expression of the analyzed CBMs and their purification and 2) to evaluate the protein stability in different buffers.

The initial strategy consisted of small-scale expression tests for the 14 CBMs presented in Table 2.1 (Chapter 2, Section 2.1). The information about the recombinant protein sequence, modular organization, protein identification, family, molecular weight and isoelectric point of the 14 CBMs are presented in Supplementary Table 1. After this, large-scale production of two CBMs was performed using the best identified conditions. The recombinant DNAs used coded for an N-terminal tag with 6 histidines and antibiotic (kanamycin) resistance (Chapter 2, Section 2.2.1).

3.2 Materials and methods

3.2.1 Preparation of competent cells

Without competence, cells do not have the ability to take up extracellular DNA, hence cells have to be treated to make them transiently permeable to DNA. The initial stock of *Escherichia coli* BL21(DE3) and *Escherichia coli* DH5α cells were aquired from NZYTech®.

The solutions used for the preparation of competent cells are in Supplementar material 1.

Preparation of DH5α competent cells

In Day 1, 50 µL of DH5α cells were uniformly spread in Luria-Bertani (LB) agar plates (overnight). In Day 2 one colony was picked from the plate and inoculated in 5 ml LB and kept at 37 ºC, at 100 rpm, overnight (Orbital ShakerIncubator ES-20, from Grant.bio). 1ml of this pre-inoculum was used to inoculate 200 ml of super optimal broth (SOB) medium and subsequently incubated at 18 ºC and 200 rpm until optical density at 600 nm (OD_{600nm}) reached 0.4-0.5. The OD_{600nm} of bacterial cells is measured at 600 nm and 0.5 is the value where microbial growth is at the exponential phase.

The cell culture was kept on ice for 10 to 15 minutes. After that, it was centrifuged in cold centrifuge tubes (previously autoclaved) at 2500 x g for 10 minutes at 4 ºC. The supernatant was rejected and the cell pellet resuspended in 80 ml of cold transformation buffer (TB). The resuspension was kept on ice for 30 minutes and then centrifuged at 2500 x g for 10 minutes at 4ºC. The new pellet was resuspended in 18.6 ml of cold TB and 1.4 ml of dimethyl sulfoxide (DMSO) was slowly added. After 15 minutes on ice the cells were distributed into autoclaved cold 1.5 ml microcentrifuge tubes and stored immediately at -80 ºC.

Preparation of BL21 competent cells

In Day 1, 50 µL of BL21(DE3) cells were uniformly spread in LB agar plates (overnight). In Day 2 one colony was picked from the plate to inoculate 5ml LB (previously autoclaved) that was kept at 37

ºC, at 100 rpm, overnight (Orbital ShakerIncubator ES-20, from Grant.bio). 1ml of this pre-inoculum was used to inoculate 200 ml of LB (previously autoclaved) and added 20 mM $MgSO_4$ (filtrated), which was subsequently incubated at 37 ºC at 200 rpm until OD_{600nm} reached 0.4-0.5. The cell culture was kept on ice for 10 to 15 minutes. After that, it was centrifuged in cold and previously autoclaved centrifuge tubes for 4000 x g for 5 minutes at 4 ºC. The supernatant was rejected and cell pellet was resuspended in 60 ml of cold transformation buffer 1 (TBF1). The resuspension was kept on ice for 90 minutes (shaking occasionaly) and then centrifuged for 4000 x g for 5 minutes at 4 ºC. The new pellet was resuspended in 8 ml of cold transformation buffer 2 (TBF2) and distributed in autoclaved cold 1.5 ml microcentrifuge tubes, and stored immediately at -80 ºC. All the procedures were carried out in a sterile environment.

3.2.2 Transformation

The transformation of DH5α competent cells used for cloning and of BL21(DE3) cells used for expression was performed in a sterile environment. For the transformation process, 2 µl of each CBM DNA was added to 50 µl of competent cells. The cells were kept on ice for 30 minutes, then heated at 42 ºC on a heatblock for 1 minute and thenincubation on ice for 10 minutes. Afterwords, the cells were re-suspended in 1 ml of LB medium and incubated at 37 ºC at 200 rpm for about 1:30 h (Orbital ShakerIncubator ES-20, from Grant.bio). The cell suspension was centrifuged for 1 minute at 8.6 rpm and 1 ml of supernatant was discarded. The cells were re-suspended in the remaining medium and spread in LB-agar plate with kanamycin antibiotic at 50 µg/ml. The DNA plasmid contains a gene with kanamycin resistance and consequently only cells that contain the plasmid will be able to grow/divide and form colonies.

3.2.3 DNA amplification, isolation and sequencing

The plates from DNA transformation with DH5α were incubated overnight at 37 ºC and then stored at 4 ºC. One colony was picked from plates and inoculated in 10 ml LB medium with 50 µg/ml kanamycin. DNA amplification was carried out at 37 ºC at 130 rpm (Orbital Shaker-Incubator ES-20, from Grant.bio) until OD_{600nm} reached 0.5.

A miniprep protocol from NZYTech® (NZYMiniprep, MB01002) was performed for extraction and isolation of the DNA. This protocol is designed for rapid and small-scale preparation of highly pure plasmid DNA from recombinant *E. coli* strains.

The isolated DNA was sent for sequencing (STAB VIDA, FCT-NOVA) to confirm the protein sequence of the 14 CBM DNAs and to identify possible mutations, insertions or deletions. The primers used were the T7 (forward primer) and pET24a (reverse primer) (Table 3.1).

Table 3.1 – Primers sequences used to confirm the DNA sequence of the 14 CBMs.

Primer	Sequence
T7 forward	TAATACGACTCACTATAGGG
pET24a reverse	GGGTTATGCTAGTTATTGCTCAG

3.2.4 Small and large-scale expression of CBMs

In order to obtain the most favorable expression condition for each CBM, small-scale expression tests were performed using two expression protocols (Table 3.2). One protocol referred to as IPTG-induction and the other as Auto-induction. The main differences between the two procedures are: 1) different cell culture medium, 2) the use of isopropyl β-D-1-thiogalactopyranoside (IPTG) for inducing the expression and 3) the temperature of expression.

Table 3.2 – Main differences in the IPTG-induction and auto-induction protocols for protein expression

Protocol	Cell culture medium	IPTG	Temperature (ºC)	Period (Hours)
IPTG-induction	LB[a]	Yes	19ºC	15
			37ºC	5
Auto-induction	Auto-induction[b]	No	25ºC	15

[a] LB – Luria-Bertani medium culture (composition in Supplementary material 2)

[b] Auto-induction medium from NZYTech® (composition in Supplementary material 2)

3.2.4.1 Protocol for expression with IPTG-induction

The first step was to follow the protocol of transformation described in Section 3.3.2 using the *E. coli strain* BL21(DE3). After transformation, one colony was pre-inoculated in 5 ml (10 ml for large scale) LB medium with 50 μg/ml kanamycin. The pre-inoculum was incubated at 37 ºC at 200 rpm for 15 hours (overnight) (Orbital Shaker- Incubator ES-20, Grant.bio). In the following day, 1 ml from the pre-inoculum was added to 50 ml (600 ml for large scale) of LB medium with 50 μg/ml kanamycin.

Each cell culture was incubated at 37 ºC at 200 rpm (Shaker IS-971R from Lab Companion) until OD_{600nm} reached 0.6-0.8. Once cell culture reached this density, 1 mM of IPTG was added and the incubation was carried out at the desired temperature and period of induction at 160 rpm.

Large-scale expression was performed using the condition that yielded the most soluble protein fraction for each CBM.

3.2.4.2 Protocol for expression with auto-induction

The protocol for transformation is the same as described in Section 3.2.2. One colony was picked up from LB plates and incubated in 50 ml of auto-induction medium with 50 μg/ml kanamycin. Each cell

culture was incubated at 37 ºC and 200 rpm (Shaker IS-971R from Lab Companion) until OD_{600nm} reached 1.5 and furtherincubated at 25 ºC at 160 rpm for protein expression.

3.2.5 Cell harvesting and lysis

The cells were harvested by centrifugation at 5.000 x g for 20 minutes (JA-10 rotor, Avanti J-26 XPI from Beckman Coulter). After discarding the supernatant, the cells were re suspended in buffer containing 50 mM Hepes pH 7.5, 100 mM NaCl, 2 mM $CaCl_2$ and 10 mM imidazole (10 ml per gram of cell pellet).

Cells were disrupted using a sonicator (UP100H, from Hielsher) with 3-5 cycles of 1 minute. This method was used to disrupt cell membranes and release cellular content. The soluble fraction was clarified by centrifugation (Rotor JA-10, Avanti J-26 XPI from Beckman Coulter) at 5000 x g and 4 ºC for 30 minutes and continued to purification protocol.

3.2.6 Purification of CBMs by affinity chromatography

An efficient method to purify the CBMs from other bacterial proteins is the immobilized-metal affinity chromatography (IMAC), and that is the reason for including an N-terminal His-tag (Chapter 2, Section 2.2.1) in the CBM constructs.

Immobilized-metal affinity chromatography (IMAC) is based on the strong affinity and selectivity that transition bivalent metal ions such as Zn^{2+}, Cu^{2+}, Ni^{2+}, and Co^{2+} have to bind histidine and cysteine residues in aqueous solutions. These metals are immobilized in the chromatography column and bind to his-tagged and cys-tagged proteins (Block *et al.* 2009).

3.2.6.1 Purification with ÄKTA START

For large-scale purification each CBM was submitted to an IMAC using a His Trap™ column of 5 ml Ni^{2+}, which is coupled to the chromatograph ÄKTA START, both from GE-Healthcare. The protocol was performed together with the UNICORN™ start 1.0 control software, and always applying a pressure under 0.5 MPa.

In this protocol two different buffer solutions were used for the concentration gradient: Buffer A – 50 mM HEPES, 1 M NaCl, 2 mM $CaCl_2$ and 10 mM at a final pH of 7.5; Buffer B - 50mM HEPES at pH 7.5,1M NaCl, 2 mM $CaCl_2$ and 300 mM imidazole at a final pH of 7.5. The column was first washed with approximately 50 mL of Milli-Q water and subsequently equilibrated with 20 ml of buffer A. The cell extract (soluble fraction) was then loaded into the column. A first wash with 15% with buffer B was applied in order to separate the constituents with low affinity for nickel. Afterwords, aconcentration gradient of imidazole, from 15% to 100% of 100 ml of buffer B, was setup in order to collect the respective CBM. The collected fractions were subsequently analyzed by SDS-PAGE and Native PAGE

(when possible, due to the proteins isoelectric point) electrophoresis for confirmation of the purity and stability of the protein.

3.2.6.2 Affinity purification with His Gravitrap™ columns

His GraviTrap™ (GE Healthcare) is a prepared, single-use column for purification of histidine-tagged proteins by IMAC, that allows fast and simple gravity-flow purifications without any need for a purification system like ÄKTA START.

This purification can be divided into four steps: equilibration, sample application, washing and elution. For this purification 3 different buffers were used: Buffer A - 50mM HEPES, 1 M NaCl, 2 mM $CaCl_2$ and 10 mM at a final pH of 7.5; Buffer B - 50 mM HEPES at pH 7.5 ,1 M NaCl, 2 mM $CaCl_2$ and 60 mM imidazole at a final pH of 7.5; Buffer C - 50 mM HEPES at pH 7.5 ,1 M NaCl, 2 mM $CaCl_2$ and 500 mM imidazole at a final pH of 7.5.

First the columns were washed with 10 ml of Milli-Q water to remove the ethanol. After this, the columns were equilibrated in buffer A. The extracted sample was then loaded into the column and the His-tag from the protein binds to the Ni^{2+} of the column. In order to wash *E. coli* proteins, and thus prevent contamination of our sample, 10 ml of buffer A and then another 10 ml of buffer B were loaded. The CBM was eluted with 6 ml of buffer C. The fractions were eluted in separate: one fraction of 3 ml and three fractions of 1 ml each for posterior analysis in SDS-PAGE and Native PAGE electrophoresis.

3.2.7 Analysis by polyacrylamide gel electrophoresis

In this work two different types of polyacrylamide gel electrophoresis were performed: in non-denaturing conditions, Native PAGE, and in denaturing conditions in the presence of sodium-dodecyl sulfate (SDS), SDS-PAGE. The protocol for SDS-PAGE was performed as described in Section 2.2.2 from Chapter 2.

In the Native PAGE proteins are analysed in their native state. In the native state, the mobility of each protein during electrophoresis is influenced by the respective size and charge, which will depend on the primary aminoacid sequence of the protein (and consequently its isoelectric point) and the pH during electrophoresis.

Native gels were performed using 12.5% acrylamide. The protein samples were mixed with the sample buffer (4x Tris-HCl, 0.012 % bromophenol blue, 30 % glycerol) at a final volume of 20 µl with 5 µg protein. The running buffer used for the electrophoreis was 1.5 M Tris-HCl at a pH of 8.8. The run paramters for the electrophoresis were setup to a fixed voltage of 150 V.

3.2.8 Protein thermal shift assay

Thermal shift assay (TSA) or Thermofluor experiment is a rapid, temperature-based assay to screen suitable conditions that maximize protein stability (Boivin *et al.* 2013). TSA induces protein

denaturation that is monitored by a fluorophore, such as SYPRO® Orange. The fluorescence of the dye increases proportionally to the hydrophobicity of the protein in each well. This is dictated by the amount of protein in each well and the state of folding. If a protein is properly folded, in its native state, the hydrophobic residues are buried within the protein, which means that are not exposed to the dye and do not emit fluorescence. When a protein is submitted to high temperatures, denaturation occurs and the protein unfolds. With the unfolding, the buried hydrophobic residues are exposed to the dye and fluorescence occurs. The monitoring of the temperature and fluorescence is done by using a Real Time PCR instrument (StepOnePlus™ system from Applied Biosystems).

The Protein Thermal Shift™ software will input the raw fluorescence data from all 96-wells during the qPCR machine's thermal melt program. After collecting the raw data, the fluorescence melt curves are plotted showing noise-reduced fluorescence data from the protein melt curve run. The temperature representing the midpoint of protein unfolding or melting transition (*Tm*) is a quantifiable value of the thermal stability of the protein.

3.3 Results and discussion

3.3.1 DNA sequencing of CBMs

As described in Section 3.2.3, the DNA was amplified and purified by using a miniprep protocol from NZYTech® (NZYMiniprep, MB01002). The 14 CBMs were sent for sequencing (STAB VIDA, FCT-NOVA) and the analysis could confirm that, with one exception, all the CBM's DNA sequences were correct. The exception was *Ct*CBM3 (Cthe_0059) for which an insertion of an adenine was found, resulting in frame shifting and, consequently, in the wrong protein sequence. Thus, it was decided to not continue with the recombinant expression of this CBM. For the microarray analysis (Chapter 2) this protein was provided by NZYTech® and was confirmed to be functional. Further studies for this protein are required in order to perceive the origin of the DNA insertion, however this is not within the scope of this thesis work.

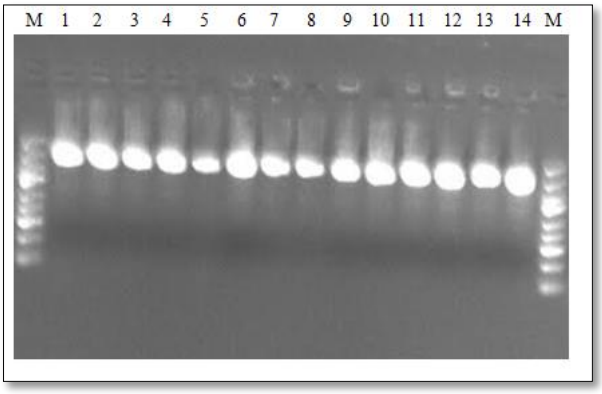


Figure 3.1 - Agarose gel electrophoresis of 14 DNA sequences from CBMs of different families. Wells from left to right are identified by organism and CBM family and protein ID brackets follows: 1-*Ct*CBM13 (Cthe_0661), 2-*Ct*CBM3 (Cthe_0059), 3-*Ct*CBM4 (Cthe_0413), 4-*Ct*CBM32 (Cthe_0821), 5-*Ct*CBM42 (Cthe_1273), 6- *Ct*CBM6 (Cthe_2197), 7– *Ct*CBM6 (Cthe_2195), 8-*Ct*CBM22 (Cthe_2590), 9-*Ct*CBM4 (Cthe_2809), 10-*Ct*CBM50 (Cthe_0300), 11- *Ct*CBM25 (Cthe_0956), 12-*Rf*CBM62 (3398), 13- *Rf*CBM13 (2115), 14 – *Rf*CBM6 (3747). The photograph was provided by STAB VIDA.

3.3.2 Small-scale expression tests

As mentioned in Section 3.2.4, several expression tests were performed for each CBM. Two protocols for expression were tested: auto-induction (Panel A) and IPTG-induction (Panel B) (Figure 3.2). The auto-induction protocol from NZYTech® was tested in order to follow the conditions in which the proteins, that were previously analyzed in the microarrays (Chapter 2), were expressed.

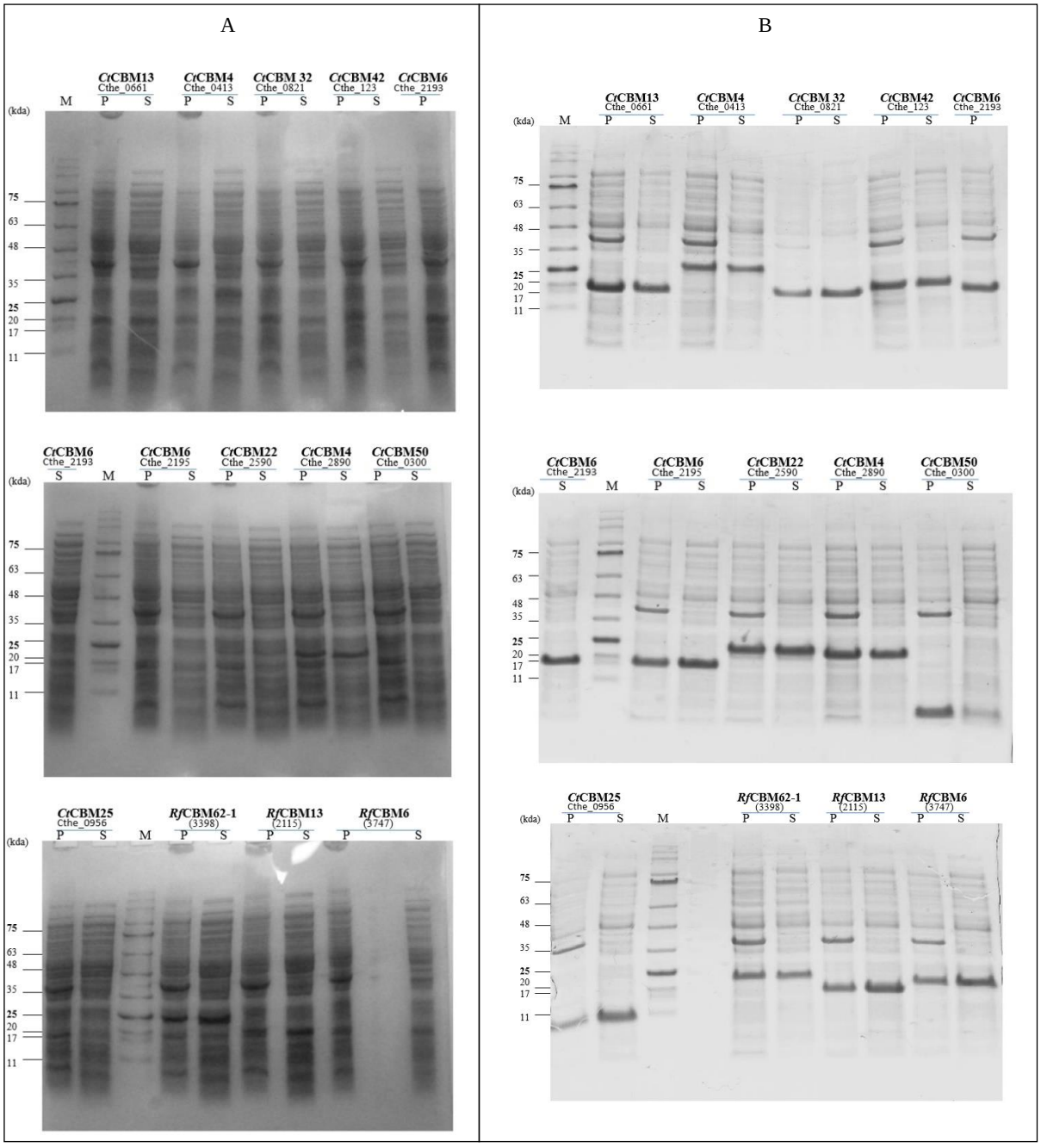


Figure 3.2 - SDS-PAGE (10% acrylamide) analysis showing the expression levels of the 13 CBMs using two expression protocols. The CBMs are identified by family and protein ID. A.) Levels of expression of 13 CBMs were tested when *E. coli* BL21 strains were grown in auto-induction medium following the NZYTech® protocol, inducing expression at 25ºC, overnight at 160 rpm. B.) Levels of expression of 13 CBMs were tested when *E. coli* BL21 strains were grown in LB medium, inducing expression with IPTG at 19ºC, overnight at 160 rpm. The respective protein lanes are identified by a blue line**.** P-Pellet (insoluble fraction); S-Supernatant (soluble fraction); M- Marker II from NZYTech®.

For each condition, the expression of the CBMs was monitored using SDS–PAGE analysis. Figure 3.1 shows soluble and insoluble fraction of each CBM after expression with the two protocols.

The SDS-PAGE analysis showed that, in general, the expression of the CBMs was much lower using the auto-induction protocol in comparison with IPTG-induction.

In order to confirm the expression of the desired proteins, affinity purification using His-GraviTrap™ columns (Supplementary Figure 1) was carried out. However, even after the purification step, the detection of CBM protein bands was very weak, confirming low levels of protein expression, not desirable to proceed to large scale production of the CBMs using this protocol.

The analysis of soluble fractions showed that, with the exception of *Ct*CBM50, there was a high level of expression of the CBMs using the IPTG-induction protocol (Figure 3.2). Despite this, there was a high level of all CBMs retained in the insoluble fractions, thus further optimization studies were performed, and described in the next section.

For *Ct*CBM50 the SDS-PAGE analysis showed that it was mainly retained in the insoluble fraction when expressed using the IPTG-induction protocol, suggesting that an improvement for expression conditions was also needed.

In face of these results, the next studies were performed using the IPTG-induction protocol.

3.3.3 CBMs expression optimization tests

With the aim to optimize the expression conditions for *Ct*CBM50 mainly, but also for the other CBMs, an optimization study was performed using the IPTG-induction protocol, at two different temperatures (19 ºC and 37 ºC) and two different induction periods (5 and 15 hours) for each CBM.

For these studies, we selected four proteins, of which two were chosen for the large-scale production required for the follow-up biophysical and structural studies. The four CBMs chosen were: *Ct*CBM50, *Ct*CBM25, *Rf*CBM62-1 and *Rf*CBM6.

The procedure, up to the stage of the induction of the expression, was the same as described before (Section 3.2.4.1). The expression of CBMs was evaluated using two different conditions: 19 ºC for 15 hours (overnight) and 37 ºC for 5 hours, both at 160 rpm. After expression, cells were harvested and lysed and the CBMs were purified using His GraviTrap™ columns, in order to obtain high quantity of pure protein. The results from the optimization tests and further purification are presented in Figures 3.3 and 3.4 respectively.

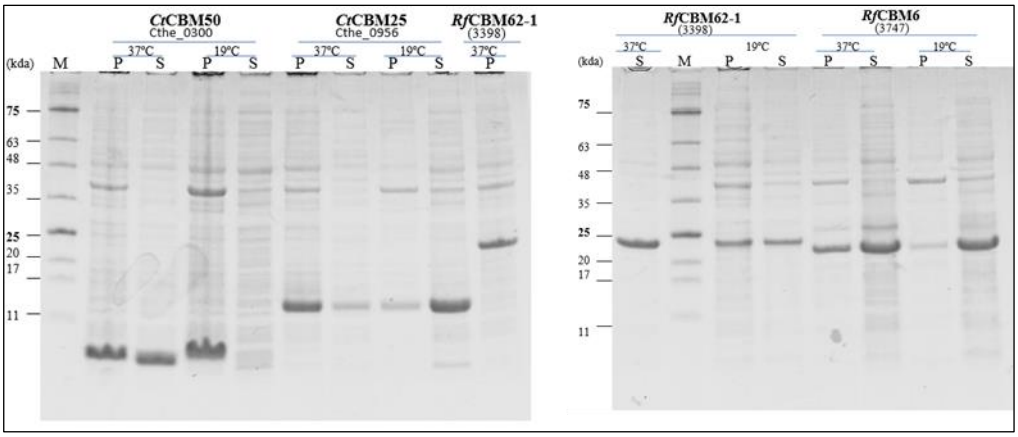


Figure 3.3 - SDS-PAGE (10% acrylamide) results showing expression levels of CBMs from different families. The CBMs are identified by family and protein ID. Levels of expression of 4 CBMs were tested with IPTG-induction protocol at two different conditions: 19ºC for 15 hours (overnight) and 37ºC for 5 hours, both at 160 rpm. Respective protein lanes are identified by a blue line. P- Pellet (insoluble fraction); S-Supernatant (soluble fraction); M- Marker II from NZYTech®.

In Section 3.3.2, it was mentioned that *Ct*CBM50 was in the insoluble fraction when expressed at 19 ºC for 15 hours (overnight). After performing the optimization studies, we could identify a new condition in which *Ct*CBM50 was expressed in good yields, in the soluble form. This corresponded to the expression condition at 37 ºC for 5 hours, and hence was selected to test the expression of *Ct*CBM50 in large scale.

For *Ct*CBM25 a different behavior was observed as the expression at 37 ºC for 5 hours resulting in most of the protein to be in the insoluble fraction. The best expression condition identified for *Ct*CBM25 was at 19 ºC for 15 hours.

For *Rf*CBM62-1 it was clear that the condition for the expression in large scale was at 37 ºC for 5 hours, as it was the condition with the highest amount of soluble protein.

For *Rf*CBM6 both conditions tested for expression resulted in similar high levels of soluble protein. In this work, we selected the condition of 19 ºC for 15 hours, as a close observation of the gel confirmed that it resulted in a higher ratio of the CBM to contaminating host proteins.

The soluble fractions of each CBM were purified by IMAC, in His Gravitrap™ columns, following the protocol from Section 3.2.6.2. After the purification, the eluted fractions were analyzed by SDS- and Native-PAGE (Figure 3.4).

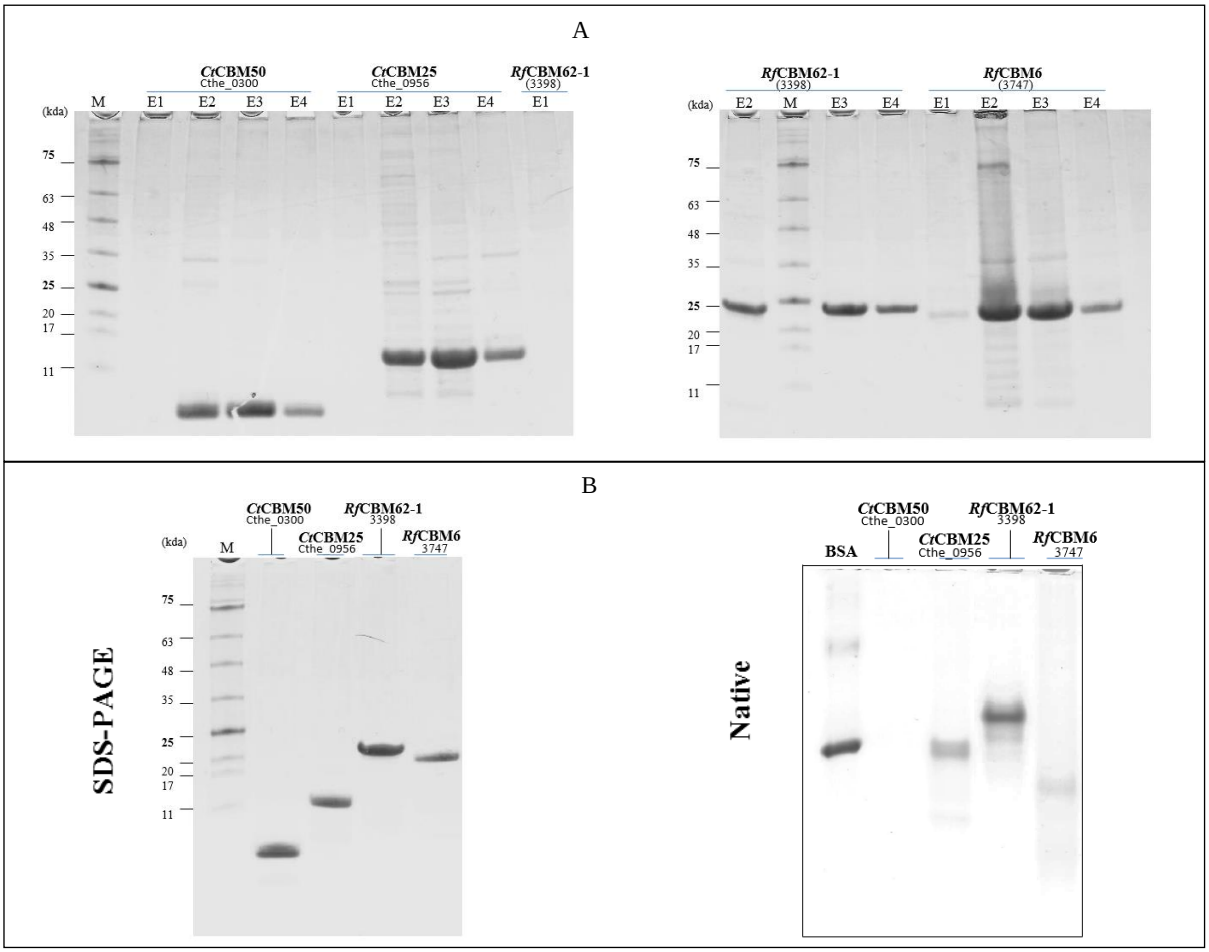


Figure 3.4 - SDS-PAGE (10% acrylamide) and native gel results showing purification of CBMs from different families. The CBMs are identified by family and protein ID. Respective protein lanes are identified by a blue line**.** Panel A.) Results of purification with His Gravitrap columns. CBMs were eluted with 6 ml of 50 mM HEPES at a pH of 7.5, 1M NaCl, 2 mM $CaCl_2$ and 500 mM imidazole. The fractions were eluted and analyzed in separated: one fraction with 3 ml and three fractions with 1 ml each. E1(Eluted fration 1) - 3ml; E2(Eluted fraction 2) - 1 ml; E3 (Eluted fraction 3) – 1ml; E3 (Eluted fraction 3) – 1ml. Panel B.) SDS-PAGE and native gel of combined eluted fractions 3 and 4 of each CBM. For both gels were used the respective concentrations: *Ct*CBM50, *Ct*CBM25 and *Rf*CBM62-1 7.5µg; *Rf*CBM6 – 6.5µg. P- Pellet (insoluble fraction); S-Supernatant (soluble fraction); For SDS-PAGE a Marker II from NZYTech® was used (M) and for native gel BSA was used as reference.

Panel A shows that the eluted fractions that contained less contamination and higher amounts of protein were fractions 3 and 4. These two fractions were combined and analyzed by SDS- and Native-PAGE (Panel B) to confirm it purity and if the protein is well folded and not degraded.

As *Ct*CBM50 has an isoelectric point of 9.78, the migration in the gel occurs to the positive pole so it could not be analyzed in a regular native gel. A protocol for Native PAGE with basic proteins is described from Bio-Rad® (Bulletin 2376) and should be used for this CBM. The other proteins were apparently well folded but *Rf*CBM6 needs to be analyzed in higher concentration to sharpen the protein band.

The combined fractions were stored at 4 ºC for posterior studies, upon the addition of 0.02 % sodium azide. The best conditions for large scale expression of the four selected CBMs are summarized in Table 3.3.

Table 3.3– Best conditions for large scale expression of the four CBMs

Protein	Medium culture	IPTG[b] (mM)	Induction temperature (ºC)	Induction time (H)
***Ct*CBM50**	LB[a]	1	37	5
***Ct*CBM25**	LB[a]	1	19	15 (overnight)
***Rf*CBM62-1**	LB[a]	1	37	5
***Rf*CBM6**	LB[a]	1	19	15 (overnight)

[a]LB – Luria-Bertani medium culture (composition in Supplementary information 2)
[b] IPTG - isopropyl β-D-1-thiogalactopyranoside

3.3.4 Large-scale expression and purification

In order to proceed to structural characterization studies, large-scale production of the CBMs was carried out. As mentioned in the previous section, four proteins were tested in order to choose two to proceed for the biophysical and structural studies.

The CBMs selected were CBM50 from *C. thermocellum* and CBM62-1 from *R. flavefaciens*. We chose these two CBMs based on their unique specificities towards carbohydrate ligands revealed by the microarray analysis. *Ct*CBM50 is a novel LysM chitin-binding domain. *Rf*CBM62-1 revealed to be highly specific for a pectic polysaccharide, and on the other hand, for the *R. flavefaciens* FD-1 the CBM specificities are largely unknown. Although challenging, both CBMs revealed to be interesting candidates for structural characterization. The modular organization where the *Rf*CBM62-1 is biologically integrated is represented in the Figure 3.5, together with the primary sequence of *Ct*CBM50.

The conditions for the large-scale expression have been presented in Table 3 from previous Section and the protocol for purification is described in Section 3.2.6.1.

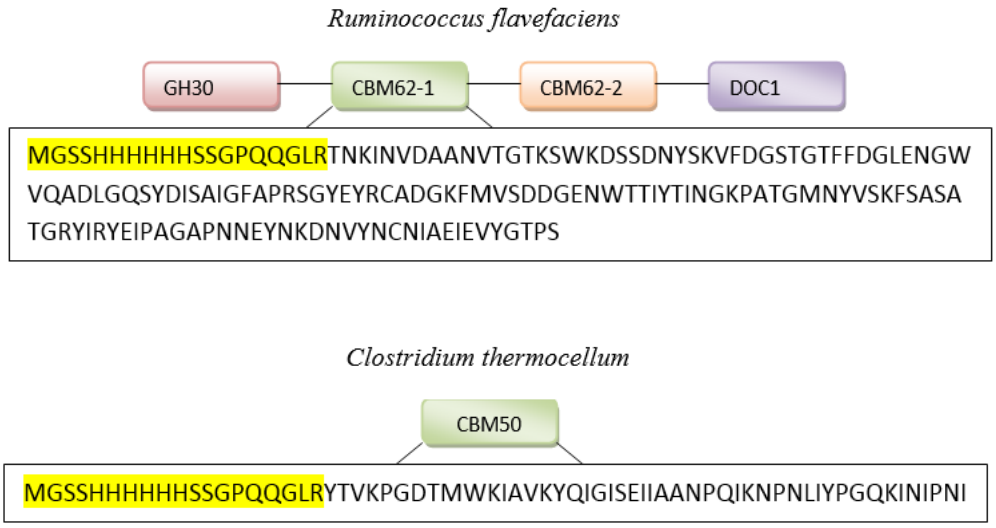


Figure 3.5 - Modular organization where the *Rf*CBM62-1 is biologically integrated and respective primary sequence is represented, together with the primary sequence of *Ct*CBM50. The N-terminal His-tag is depicted at yellow.

Purification of CBMs was performed by IMAC in a HisTrap™ (GE- Healthcare) column of 5 mL coupled to an ÄKTASTART chromatograph. The buffers used in the purification are listed in Table 3.4. The purification chromatograms obtained for *Ct*CBM50 and *Rf*CBM62-1 are shown in Figures 3.6 and 3.7, respectively. All collected fractions were analyzed by SDS-PAGE, allowing us to select those corresponding to the respective purified CBM (Figure 3.6 and 3.7 inset panels). These same fractions were then submitted to a desalting protocol and concentrated using concentrators Vivaspin® 3KD (GE Healthcare) for posterior structural studies.

Table 3.4 – Buffers used in the IMAC purification of *Ct*CBM50 and *Rf*CBM62-1

Buffer A	50 mM HEPES, 1 M NaCl, 2 mM CaCl2 and 10 mM imidazole; pH=7.5
Buffer B	50 mM HEPES, 1 M NaCl, 2 mM $CaCl_2$ and 300 mM imidazole; pH=7.5

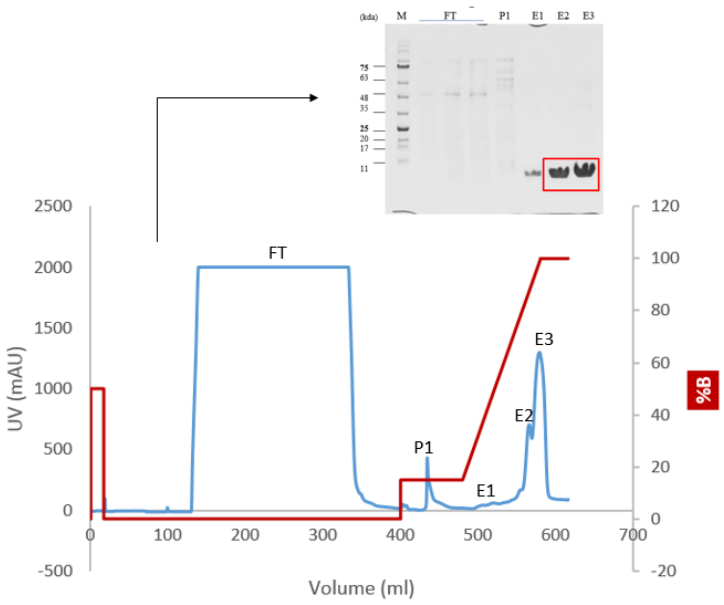


Figure 3.6 - Results from purification of *Ct*CBM50: Chromatogram from IMAC and SDS-PAGE analysis with 13% acrylamide. FT- Flow trought; P – Peak; E- Elution; M- Marker II from NZYTech®. The blue line represents the volume as function of UV. The red line represents the volume as function of a gradient of the buffer B. Red rectangles in SDS-PAGE gel show the fractions that were collected and stored for posterior desalinization protocol.

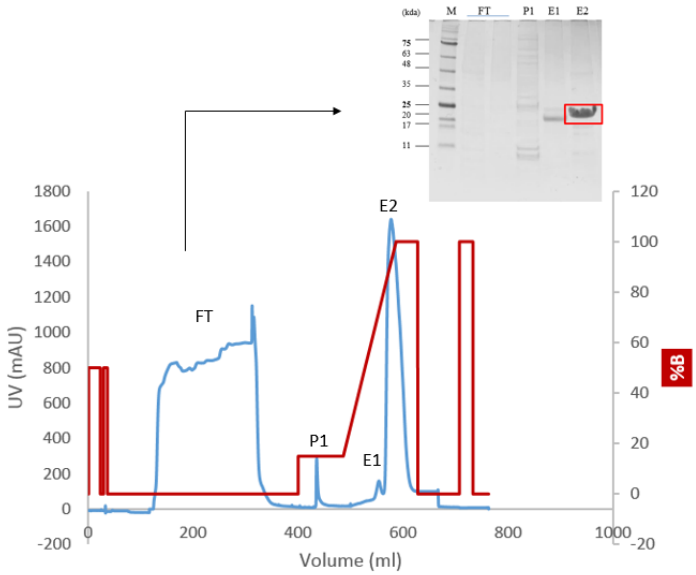


Figure 3.7- Results from purification of *Rf*CBM62-1: Chromatogram from IMAC and SDS-PAGE analysis with 13% acrylamide. FT- Flow trought; P – Peak; E- Elution; M- Marker II from NZYTech®.The blue line represents the volume as a function of UV. The red line represents the volume in function of a gradient of the buffer B. Red rectangles in SDS-PAGE gel show the fractions that were collected and stored for posterior desalinization protocol.

After the expression in 2400 ml of LB medium, soluble extracts of 130 ml and 180 ml for *Ct*CBM50 and *Rf*CBM62-1, respectively, were obtained and then submitted to purification. Ultimately, 50 ml and 60 ml of *C*tCBM50 and *Rf*CBM62-1 with a concentration of 1.3 mg/ml and 1.6 mg/ml, respectively, were obtained (Table 3.5).

Table 3.5 - Summary of the purification yields for *Ct*CBM50 and *Rf*CBM62-1

Protein	Volume Medium culture	Volume Expression extract	Volume Purified protein	Protein quantification
***Ct*CBM50**	2400 ml	130 ml	50 ml	1.3 mg/ml
***Rf*CBM62-1**	2400 ml	180 ml	65 ml	1.6 mg/ml

3.3.5 Protein stability analysis

One of the key issues for working with new proteins is to choose the appropriate buffer to use during protein purification, storage and for experimental assays. When proteins are produced in large scale it is necessary to optimize conditions that stabilize the protein. Proteins used in Chapter 2 of this thesis were provided by NZYTech® and were stable in 50 mM HEPES, 100 mM NaCl and 2 mM $CaCl_2$ at a pH of 7.5 at the range of concentrations provided. However, when produced in large-scale the protein behavior in solution could change and the protein stability in different buffers needed to be assessed.

Thermal Shift Assay (TSA) or Thermofluor experiment provides a comprehensive method to analyze buffers in order to determine those the condition that maximize protein stability. The ability to perform

Thermofluor experiments in a 96-well format provides the opportunity to test several buffers in a short time. The protocol used for the experiments is described in section 4.2.8.

An initial study was made with different concentrations of each CBM in 50mM HEPES, 100 mM NaCl and 2 mM $CaCl_2$ at a pH of 7.5 in order to search for a concentration to work in further studies. The results are shown in Table 3.6 and Figure 3.8 1) and 3.8 2).

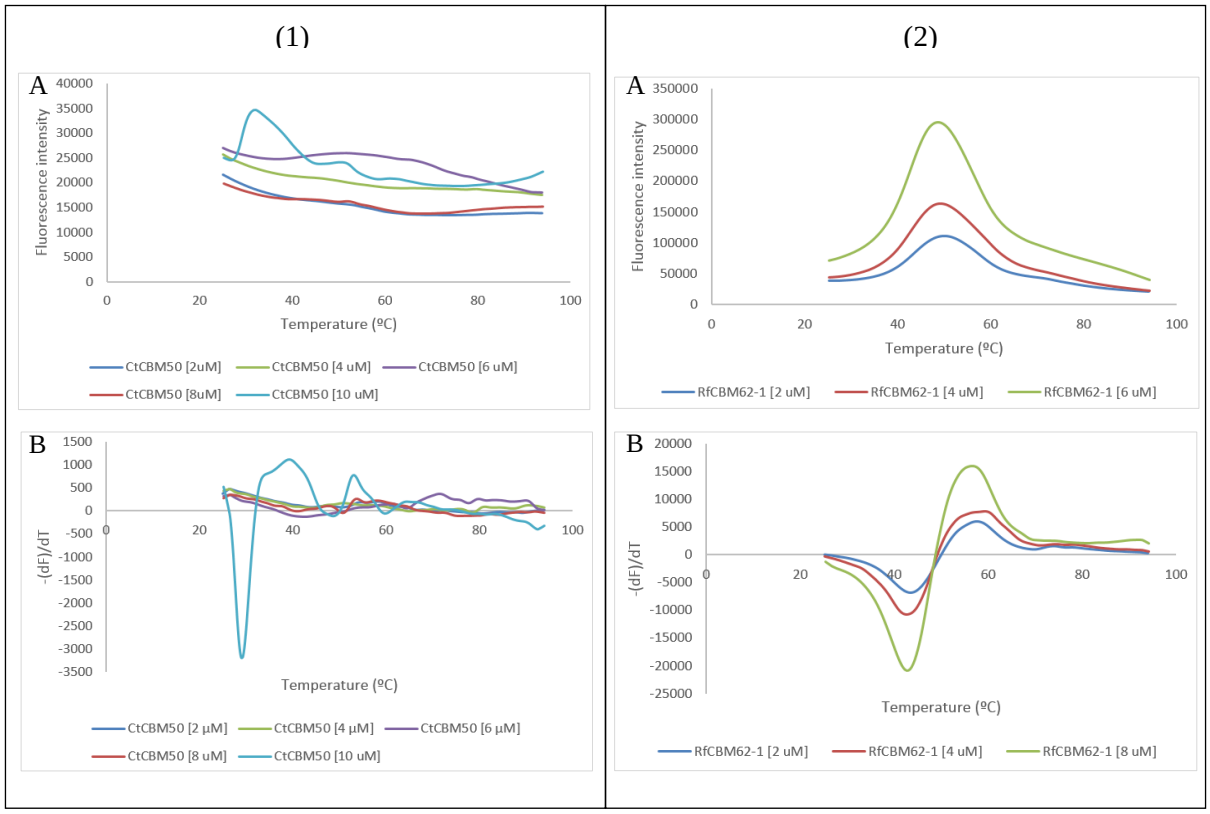


Figure 3.8 – Thermal denaturation assay using Thermofluor for *Ct*CBM50 and *Rf*CBM62-1. (1) Thermal denaturation assay using Thermofluor for *Ct*CBM50. A) Melting curves of *Rf*CBM62-1 at a concentration of 2,4,6,8 and 10 µM in 50 mM HEPES, 100 mM NaCl and 2 mM $CaCl_2$ at pH=7.5. B) Alternative representation of the melting curves using the first derivative – (dRFU)/dT of the raw data in which the minimum value corresponds to the Tm. (2) Thermal denaturation assay using Thermofluor for *Rf*CBM62-1. A) Melting curves of *Rf*CBM62-1 at a concentration of 2, 4 and 8 µM in 50 mM HEPES, 1 M NaCl and 2 mM $CaCl_2$ at pH=7.5. B) Alternative representation of the melting curves using the first derivative – (dRFU)/dT of the raw data in which the minimum value corresponds to the Tm.

Table 3.6 – Results from Thermofluor data analysis.

Protein	Concentration [µM]	Derivative *Tm* (ºc)
CtCBM50	2, 4, 6, 8, 10	-
RfCBM62	2	43.54
	4	42.53
	6	42.57

The Figure 3.8 shows the fluorescence (1A and 2A) and derivative (1B and 2B) normalized melting curves for CBMs at different concentrations. The protein solutions were heated in the presence of a

hydrophobic dye (SYPRO Orange). Upon denaturation, the dye binds to the internal hydrophobic protein core increasing significantly the fluorescence. Maximal fluorescence intensity is obtained when the protein unfolds completely, then SYPRO Orange signal decreases corresponding to dye-protein dissociation. Residual signal from SYPRO Orange is explained by the interaction from the dye with aggregated protein. The inflection point corresponds to the melting temperature (Tm), at which 50 % of the protein is unfolded in the fluorescence graph, while in the derivative graph corresponds to the minimum value.

For *Ct*CBM50 there was no tendency associated with a melting curve, actually for concentrations of 2, 4, 6 and 8 µM the line is flat, which could mean that the protein does not denature between 25 and 95ºC. This can be related with the fact that this protein belongs to a thermophile organism, *C. thermocellum*. However, it can also be an indicator of protein misfolding or quickly changing between different folds. For the concentration of 10 µM the protein has a high initial fluorescence and does not present a typical sigmoidal shape. Further studies are necessary to conclude about the stability of *Ct*CBM50, ideally with an equipment capable to go to higher temperatures. In the meantime, we decided to continue with this protein for crystallographic studies in the same buffer.

*Rf*CBM62-1 showed a typical fluorescence melting curve, with low initial fluorescence that increases with the temperature increment, as a result of the protein unfolding. Also, there was an increment of fluorescence intensity with the increase of the concentration of the protein, which is coherent with the Thermofluor principles. For *Rf*CBM62-1 it was possible to obtain a range of melting transitions between 42 and 44 ºC.

After this first initial screening with different concentrations for *Ct*CBM50 and *Rf*CBM62-1 in 50mM HEPES, 100 mM NaCl and 2 mM $CaCl_2$ at a pH of 7.5 buffer, we continued further studies with *Rf*CBM62-1 at a concentration of 6 µM. With the intent to evaluate the protein stability in different buffers, a buffer screen was performed (Supplementary Figure 2).

A screening of 96 conditions was made and analyzed. The results for the two most promising conditions are showed in Table 3.7 and Figure 3.9.

Table 3.7 – Thermofluor data analysis for *Rf*CBM62-1

Well	Buffer	Derivative *Tm* (ºC)
C1	50mM HEPES, 100 mM NaCl, 2 mM $CaCl_2$ pH=7.5	43.31
C5	2 mM $CaCl_2$	44.77

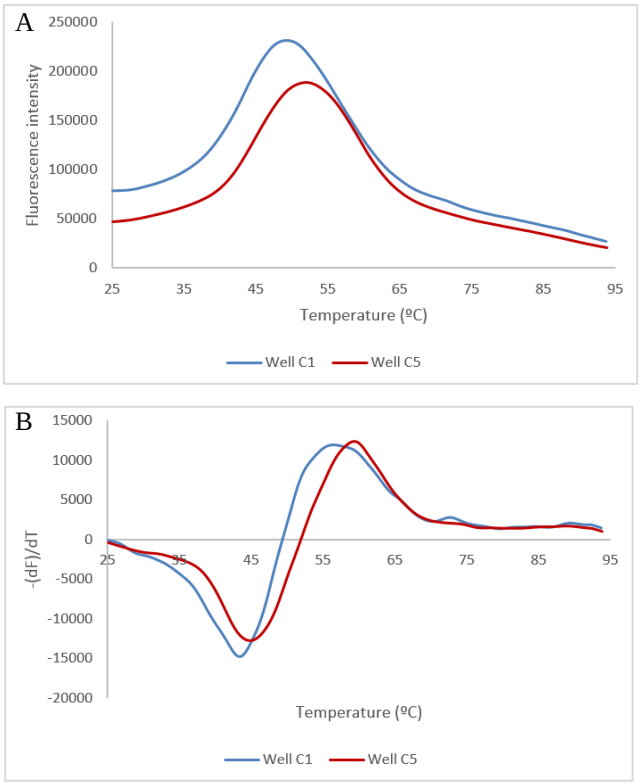


Figure 3.9 - Thermal denaturation assay using Thermofluor for *Rf*CBM62-1 in 50mM HEPES, 100 mM NaCl, 2 mM $CaCl_2$ pH=7.5 (well C1) and in H_2O with 2 mM $CaCl_2$ (well C5). (A) Melting curve of *Rf*CBM62-1 in 50mM HEPES, 100 mM NaCl, 2 mM $CaCl_2$ pH=7.5 (well C1) and in H_2O with 2 mM $CaCl_2$ (well C5). (B) Alternative representation of the melting curve using the first derivative – (dRFU)/dT of the raw data in which the minimum value corresponds to the Tm.

The melting transitions for this conditions confirmed the previous assay, which indicated melting transitions around 43 ºC. Comparing both melting curves, the two conditions have a slightly difference in the melting transition of 1.4 ºC and also differ in the initial fluorescence value. These differences between the two conditions showed that the protein is more stable in the condition that contains only 2 mM $CaCl_2$. The protein showed melting curves with weak denaturation transition, which indicates that the protein is not highly stable. However, this protein was purified in buffer 50mM HEPES, 100 mM NaCl, 2 mM $CaCl_2$ pH=7.5, concentrated and the buffer exchanged to 2 mM $CaCl_2$ for crystallographic studies.

3.4 Conclusions

The first step in any study of a particular protein begins at its expression and purification. This step is crucial to get an acceptable amount of pure protein to use in further studies. The efficient large-scale production of recombinant proteins depends on the careful conditioning of the protein as it is isolated and purified to homogeneity. Low protein stability leads to low purification yields as a result of protein degradation, precipitation and folding instability (Boivin *et al.* 2013).

In the first part of this work we performed expression tests in small-scale in order to achieve the expression condition that would provide the higher amount of pure protein.

Driven by the previous carbohydrate microarray results, two CBMs were chosen to express and purify in large-scale: *Ct*CBM50 and *Rf*CBM62-1. These CBMs presented higher expression levels when expressed for 5 hours at 37 ºC following the IPTG-induction protocol. After the affinity purification with IMAC, 50 ml at 1.4 mg/ml of *Ct*CBM50 and 65 ml at 1.6 mg/ml of *RfCBM62*-1 were obtained, which were acceptable values of purified protein for posterior concentration.

The second part of this work consisted in investigating the thermal stability of the two CBMs. It was not possible to obtain any conclusions about the *Tm* of *Ct*CBM50, which could mean that the protein does not unfold with the respective temperature variation. Also, being a relatively small protein, the unfolding process (and consequent exposure of core residues) does not occur in a clear transition, which makes it difficult to follow by thermofluorescence assays. On the other hand, *Rf*CBM62-1 exhibited a melting curve with an approximated *Tm* of 44.77ºC in buffer containing 2 mM $CaCl_2$.

In conclusion, this part of the work was the starting point for the biophysical characterization of *Ct*CBM50 and *Rf*CBM62-1.

Chapter 4 - *Ct*CBM50: A novel chitin binding LysM domain

4.1 Introductiory remarks

Following the observations of the microarray analysis in Chapter 2, the ligand specificity of *Ct*CBM50 LysM domain for chitin was explored in this chapter. The LysM domain is a highly conserved carbohydrate binding module found in proteins from virus, bacteria, fungi, plants and mammals. LysM modules consist of 43-50 amino acids with a βααβ-fold. These modules recognize polysaccharides with *N*-acetylglucosamine (GlcNAc) residues, such as peptidoglycan or chitin (components of bacterial and fungal cell wall, respectively) (Mesnage *et al.* 2014).

In this chapter the following tasks were performed: 1) assessement of binding to insoluble chitin by SDS-PAGE electrophoresis; 2) ITC measurements to evaluate the *Ct*CBM50 binding affinity to the β-(1-4)-linked GlcNAc chitin pentasaccharide; 3) several crystallization screens aiming at solving the X-ray structure; and 4) a similarity model produced computationally to evaluate the predicted 3D structure and the residues for interaction with the ligand.

4.2 Materials and methods

4.2.1 Analysis of interaction with insoluble polysaccharides

The interaction with insoluble carbohydrates can be monitored by SDS-PAGE in a procedure where the CBM is incubated with the polysaccharide before being applyed to the gel.

In the protocol followed here, 50 µg of CBM was incubated with 2 % (w/v) of insoluble polysaccharide in 200 µl of buffer (50 mM HEPES pH 7.5, 100 mM NaCl and 2 mM $CaCl_2$ at a pH 7.5) with gentle shaking each 10 minutes for 2 hours at room temperature. The mixture was centrifuged at 13000 x g for 5 minutes and 160 µl of incubated supernatant (soluble fraction) was collected. The remaining mixture was resuspended with 160 µl of buffer (50 mM HEPES pH 7.5, 100 mM NaCl and 2 mM $CaCl_2$) and centrifuged again at 13000 x g for 5 minutes. This process was repeated twice (washes). After the third wash, the mixture was re-suspended in the remaining buffer (insoluble fraction). All fractions were then analyzed by SDS–PAGE with 13 % acrylamide using a marker (Protein Marker II from NZYTech®) as reference and Coomassie blue to detect protein.

Binding was assessed by visually comparing the amount of protein in the soluble fractions, washes and insoluble fraction.

4.2.2 Isothermal titration calorimetry

As stated previously in the Introduction (Section 1.4.2.3), isothermal titration calorimetry (ITC) measures the affinity of binding partners in their native states. The heat transfer in a binding event can be used to determine thermodynamic parameters (ΔG, ΔH and ΔS), binding constants (Kd) and reaction stoichiometry (n). This is a method of high value for the quantification of protein- carbohydrate interactions, as it allows to follow the titration of a carbohydrate to a protein (Freire *et al.* 1990).

ITC measurements were performed at 25ºC using a Nano ITC instrument from TA Instruments. During experiments, the stirrer-syringe was kept rotating at 300 rpm. The binding reaction was monitored by recording the heat released upon small additions of oligosaccharide solution to the protein solution.

For ITC sample preparation, the CBM (40 µM) was equilibrated in a buffer containing 50 mM HEPES, 100 mM NaCl and 2 mM $CaCl_2$ at a pH of 7.5 in the cell of the calorimeter, and subsequently, 25 injections of 10 µl of 500 µM β-(1-4)-linked GlcNAc pentasaccharide solution were performed and the heat response recorded. After subtraction of the baseline, the integrated heat responses were fitted to an independent model using the NanoAnalyze software.

4.2.3 Crystallization assays

Crystallization screening assays were conducted using a crystallization robot, the Nano drop robot Oryx8 (from Douglas Instruments) using 96-well crystallization plates. In order to screen for crystallization conditions quickly and with the smallest amount of protein possible, as this is a newly identified CBM and no crystallization conditions are known, commercial and *in house* screens were tested (Suplementary Figures 3, 4, 5, 6, 7). Before robot set ups, 50 µl of different crystallization solutions were pipetted to 96 reservoirs of the plate in a pre-established order. The robot is implemented with an informatic tool, Wasprun, which allows choosing the exact protein: precipitant ratio added into each well, which depends on the protein concentration. The crystallization sceenings used were 80!, PEG Ion 1 & 8k, JBS 1, 2, 3, 4 and EWI & EWII.

For crystallization improvement, 24-well crystallizing plates were used. In this case, 700 µl of the respective crystallization solution was added into each of the 24 reservoirs. The protein and the precipitant were added into the lamella, which was then inverted over the reservoir with a grease on top. This procedure will seal up the system and the equilibrium between the drop and the reservoir will be established. Unlike the screenings in the robot Oryx8, all the crystallization improvement process was made by hand.

Crystallization assays were made using the vapor diffusion methods of the sitting drop for the robot Oryx8 and the hanging drop for crystallization in 24-well plates. All the crystallization assays were performed at 20 ºC and an Olympus SZ60 microscope was used for visualization of drops.

4.2.4 Phyre2: A bioinformatic tool for 3D structure prediction

Phyre2 is a suite of tools available on the web to predict and analyze protein structure, function and mutations (Kelly *et al.* 2015). Phyre2 provides a simple and intuitive interface to state-of-the-art protein bioinformatics tools. The server is available at www.sbg.bio.ic.ac.uk/phyre2. A typical structure prediction will be returned between 30 minutes and 2 h after submission. In the first stage (gathering homologous sequences), a query sequence is scanned against the protein sequence database with

HHblits. The profile calculated in stage 1, together with the predicted secondary structure with PSIPRED, is converted to a hidden Markov model (HMM). In stage 2 (fold library scanning) the calculated HMM is scanned against a database of HMMs of proteins of known structure. The top-scoring alignments from this search are used to construct crude backbone-only models with an alighnment algorithm called HHsearch. Stage 3 (loop modeling) consists in the correction of the insertions and deletions by loop modeling. The last step, stage 4 (side-chain placement), is where the amino acid side chains are added to generate the final Phyre2 model.

4.3 Results and discussion

4.3.1 *Ct*CBM50 binding to insoluble chitin

Qualitative binding experiments were carried out using 50 µg *Ct*CBM50 and 2% (w/v) of respective insoluble polysaccharide. For the *Ct*CBM50 binding analysis insoluble chitin from shrimp shells (C-8908, Sigma-Aldrich) was used. The protocol followed for this experiment is presented in Section 5.2.1.

The SDS-PAGE results showed that *Ct*CBM50 binds tightly to insoluble chitin, as the protein band was visible in the insoluble fraction (polysaccharide fraction) and only a residual band was detected in the first wash fraction (Figure 4.1). This result suggests that not only this CBM have the ability to bind chitin-oligosaccharides, but also to chitin polysaccharides. This was supported by the microarray results and also opens the way to search for different chito-oligosaccharides as ligands.

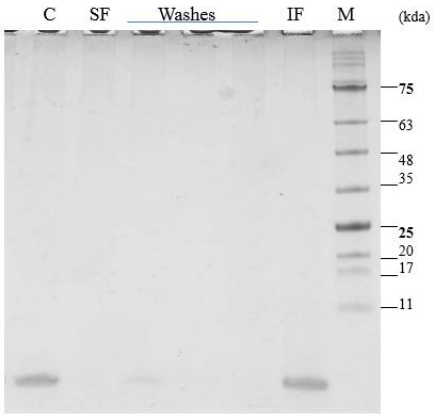


Figure 4.1 – Qualitative binding analysis of *Ct*CBM50 with insoluble chitin by SDS-PAGE with 13% acrylamide. Binding analysis of *Ct*CBM50 with 2% chitin. C – Control (50µg *Ct*CBM50, without polysaccharide); SF – Soluble fraction; IF – Insoluble fraction; M- Marker II from NZYTech®.

4.3.2 Binding affinity of *Ct*CBM50 for the β-(1-4)-linked GlcNAc chitin pentasaccharide

In order to obtain more detailed information on the binding abilities of *Ct*CBM50 for oligosaccharides, an isothermal titration calorimetry experiment was performed (Figure 4.2) with the β-(1-4)-linked GlcNAc pentasaccharide derived from chitin (O-CHI5, from Megazyme). The detailed

protocol used for the ITC experiment is presented in Section 4.2.2. The binding curve is shown in Figure 4.2 as well as the data from ITC analysis in Table 4.1.

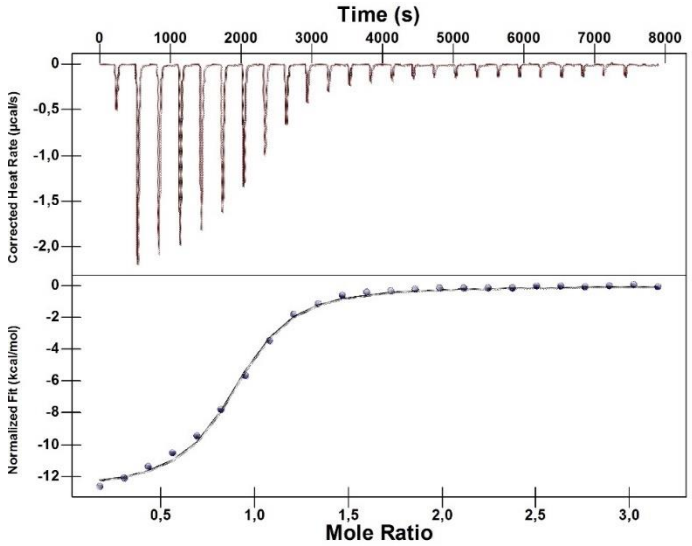


Figure 4.2 – Isothermal titration calorimetry of β-(1-4)-linked GlcNAc pentasaccharide binding to *Ct*CBM50. 40 µM of *Ct*CBM50 was titrated with N-acetyl-chitopentaose injections of 500 µM at 25ºC. Data was fitted with NanoAnalyze software. The heat response was recorded and is shown in the top graph while fitted curve is shown in bottom graph.

Table 4.1 - Data for the ITC analysis of *Ct*CBM50 binding to β(1-4)-linked GlcNAc pentasaccharide at 25 °C

K_d (M)	n	ΔH (kcal/mol)	ΔS (cal/mol.K)	K_a (M^{-1})
1.1×10^{-6}	0.863	-12.73	-15.44	9.083×10^5

n = Stoichiometry of binding

Titration of the oligosaccharide against *Ct*CBM50 produced an exothermic binding with a K_d value of 1.1×10^{-6} M and a ΔH of -12.73 kcal/mol. Also the data pointed to 1:1 stoichiometry as n is approximately 1. The K_d for binding of *Ct*CBM50 to this type of ligand is in the micromolar range as described in the literature for bacterial (Mesnage *et al.* 2014) and plant (Ohnuma *et al.* 2009) LysM modules.

A study (Mesnage *et al.* 2014) with bacterial LysM domains from *Enterococcus faecalis* AtlA, a peptidoglycan hydrolase, showed that K_d increases with DP of chito-oligosaccharide ($GlcNAc_6$>$GlcNAc_5$>$GlcNAc_4$>$GlcNAc_3$ >$GlcNAc_2$). This study showed a K_d of approximately 1.23×10^{-5} M for the bacterial LysM module binding with the pentasaccharide $GlcNAc_5$.

Another study (Ohnuma *et al.* 2009) showed that the ITC parameters for the binding affinity of a plant LysM domain from *Pteris ryukyuensis* Chitinase-A with $GlcNAc_5$ were: n=1.1, ΔH=-8.7 kcal/mol, and a K_d of 1.88×10^{-4} M. Although both K_d from bacterial and plant LysM modules described before are in the micromolar range and ITC parameters from plant LysM module presented similarity to those obtained by our experiment, our LysM domain showed a higher affinity for this chitin oligosaccharide

with a K_d of 1.1 x 10^{-6} M. No reported binding affinities for CBMs from family 50 from *Clostridium thermocellum* are described in the literature.

Further studies with chitin-oligosaccharides of different DPs are required to conclude about the chain length dependency on the affinity of *Ct*CBM50.

4.3.3 Crystallization assays

To crystallize *Ct*CBM50 several commercial screens (Table 4.2) were tested in 96-well plates using the vapor diffusion method of the sitting drop. In the first attempt for crystallization two screens were used (80! and PEG Ion 1 & 8k) with *Ct*CBM50 at a concentration of 2 mg/ml. Also other screenings (PEG Ion 1 & 8K; JBS 1, 2, 3, 4 and EWI & EWII) with *Ct*CBM50 incubated with the β-(1-4)-linked GlcNAc pentasaccharide (as suggested from the microarray and ITC studies) at a concentration of 3.2 mg/ml were made.

As no crystals were obtained, other approach was tried. A new protein batch was incubated with 1 % of maltose before concentration in order to help stabilize the protein. A concentration of 4.7 mg/ml was reached and four screens (80!, PEG ion 1 & 8k, JCSG and JBS 1, 2, 3, 4) were tested. After approximately one month, a drop with four crystals was observed in the 80! screen (Well D6: 0.1M HEPES pH:7.5 and 1M Li_2SO_4) and crystals were tested in house with a D8 Venture X-ray diffractometer.

Table 4.2 – Experimental conditions used in crystallization assays for *Ct*CBM50 with an automated nanodrop dispensing equipment (known as crystallization robot).

Sample	Buffer	Concentration (mg/ml)	Drop ratio[b]	Drop[c] (µl)	Screens
***Ct*CBM50**	A	2	2:1	1.5	80!; PEG Ion 1 & 8 k
***Ct*CBM50 +1%maltose**		4.7	2:1	1.5	80!; PEG Ion 1 & 8k; JCSG and JBS 1,2,3,4
***Ct*CBM50 + ligand[a]**		3.2	2:1	1.5	PEG Ion 1 and 8K; JBS 1,2,3,4 and EWI & EWII

Buffer A: 50 mM HEPES pH=7.5, 100 mM NaCl and 2 mM $CaCl_2$
[a] β-(1-4)-linked GlcNAc pentasaccharide
[b] Protein:precipitant ration
[c] Final drop volume with protein+precipitant

Two different solutions for handling the crystals are required, harvesting buffer and cryo-protector solution. The harvesting buffer is a stabilizing solution with the same composition of the precipitant solution except for a higher concentration of precipitating agent (Li_2SO_4) to avoid the crystal dissolution during handling. The cryo-protector solution is composed by harvesting buffer with a glycerol

percentage (10-30%) and prevents ice formation when the crystal is exposed to a cooled nitrogen gas stream (~100K).

A multiple crystal of *Ct*CBM50 presented a diffraction pattern (Supplementary Figure 8), however this crystal did not diffract well enough to provide good data. We proceeded for a scale-up optimization (Figure 4.3) by changing the protein:precipitant ratio in the crystallization drop. In order to confirm the reproducibility of the results and obtain good diffracting crystals, a *Ct*CBM50 optimization study was made using the vapor diffusion method and the hanging drop technique in a 24-well plate.

Each lamella was composed by four drops: three drops with different protein-precipitant ratios and one control drop (buffer and precipitant). After new purification and concentration with 1 % maltose, we were able to reach a concentration of 5.7 mg/ml of *Ct*CBM50. The concentration of precipitant was varied between 0.5 M and 1.5 M and the protein-precipitant ratios were 2:1, 1:2 and 1:1, to a final volume of 2 µl. The control was made with a proportion of buffer-precipitant of 1:1.

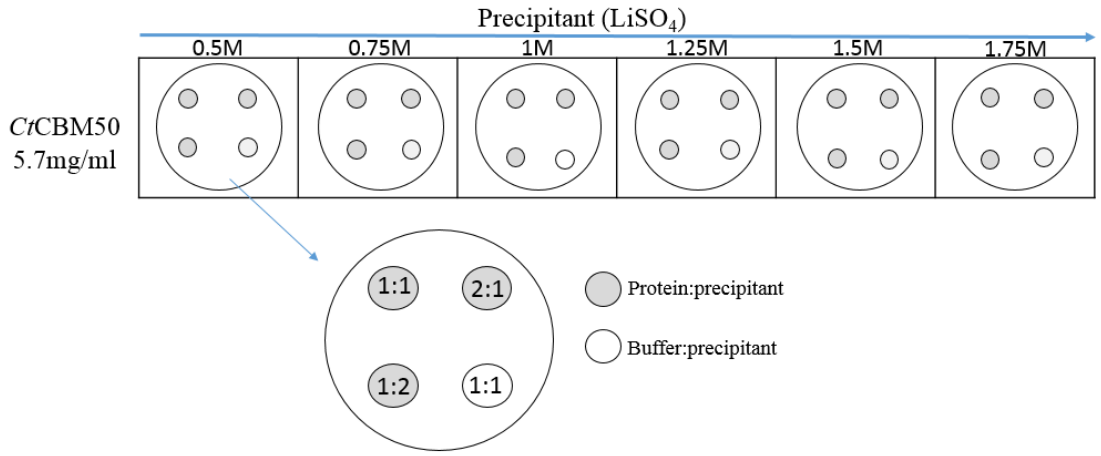


Figure 4.3 - *Ct*CBM50 optimization study from 80! Screen (Well D6: 0.1M HEPES pH:7.5 and 1M Li_2SO_4). Concentrations of the precipitant agent were varied from 0.5 M to 1.75 M. The ration 1:1, 2:1 and 1:2 are represented for grey circles. The control (buffer: precipitant) is represented as white circles. All drops were made with a *Ct*CBM50 concentration of 5.7 mg/ml.

No crystals were observed from this optimization, and in general, low precipitation was observed in the drops, which indicates that a higher concentration of protein is required. Overall, we observed that the protein was very unstable and precipitated when submitted to a concentration protocol. *Ct*CBM50 is a protein with 66 amino acid residues, in which 19 are part of the N-terminal His-tag. Normally, in small proteins, the terminal His-tag contributes for protein instability and may interfere with crystallization process. In this case a different protein-construct approach may be needed for future studies.

4.3.4 *Ct*CBM50 3D structure prediction

The rice blast fungus *Magnaporte oryzae*'s genome encodes a hypothetical protein (MGG_03307) containing a type III CVNH lectin, in which a LysM domain is inserted between individual repeats of a

single CVNH domain, linked by two highly flexible seven-residue Gly-rich linkers. The structure of type III CVNH module (designated as MoCVNH-LysM) was solved by NMR spectroscopy (Koharudin *et al.* 2011). *Ct*CBM50 3D structure (Figure 4.4) was predicted by calculating a similarity model with MoCVNH-LysM (PDB code 2L9Y).

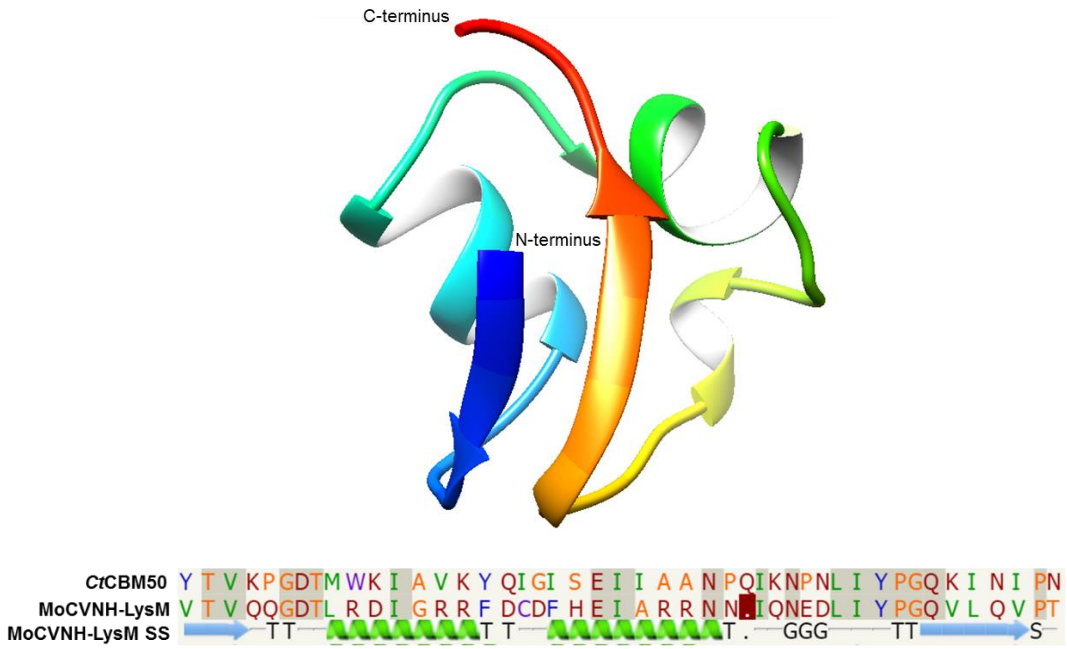


Figure 4.4 – Ribbon representation of *Ct*CBM50 similarity model and respective alignment with the template sequence. MoCVNH-LysM SS – Template secondary sctructure. In the secondary structure T corresponds to hydrogen bonded turn, G to 3-turn helix and S to bend. A similarity model was calculated by Phyre2 using a type III CVNH lectin from *Magnaporte oryzae* (PDB code 2L9Y) as a template as well as the alignment. The program Chimera was used for visualization. Rainbow ribbon spectrum from N-terminus (blue) to C-terminus (red).

For the similarity model, 46 of 47 residues (98% of *Ct*CBM50 sequence) were modelled with 99.6% confidence and 40% identity with the template sequence. With this bioinformatic tool it was possible to predict what the three-dimensional structure of *Ct*CBM50 could be. According to the criteria of Chimera program our model has a βαααβ fold. Different programs have different criteria to conclude if a certain region of the protein is considered an alpha helix or a beta foil. In this case, the program considered a very small portion of the loop as an alpha helix, however it is very small and could be consider an artefact. Nevertheless, this is just a prediction and structural differences are expected in lower similarity stretches, in solvent exposed loops and possibly in the recognition site. In the literature, a βααβ-fold for the three-dimensional structure of LysM domains is described (Buist *et al.* 2008; Gronenborn *et al.* 2015; Mesnage *et al.* 2014).

After exhaustive efforts to crystallize this protein, the authors designed a new construct in which they removed both linkers and the last two (TK) residues at the C-terminal end of the LysM domain and changed the proline to a glycin, resulted in crystals. This protein was named Mo0v and the structure for complex with $GlcNAc_3$ and $GlcNAc_4$ was solved (Gronenborn *et al.* 2015).

In order to compare our model with Mo0v in complex with $GlcNAc_4$ (PDB code 5C8Q) and look at the residues involved in ligand binding, a superposition between our model and Mo0v structure was made (Figure 4.5). The 43 residue pairs were aligned with an RMSD of 0.967Å for 38 pruned atom pairs.

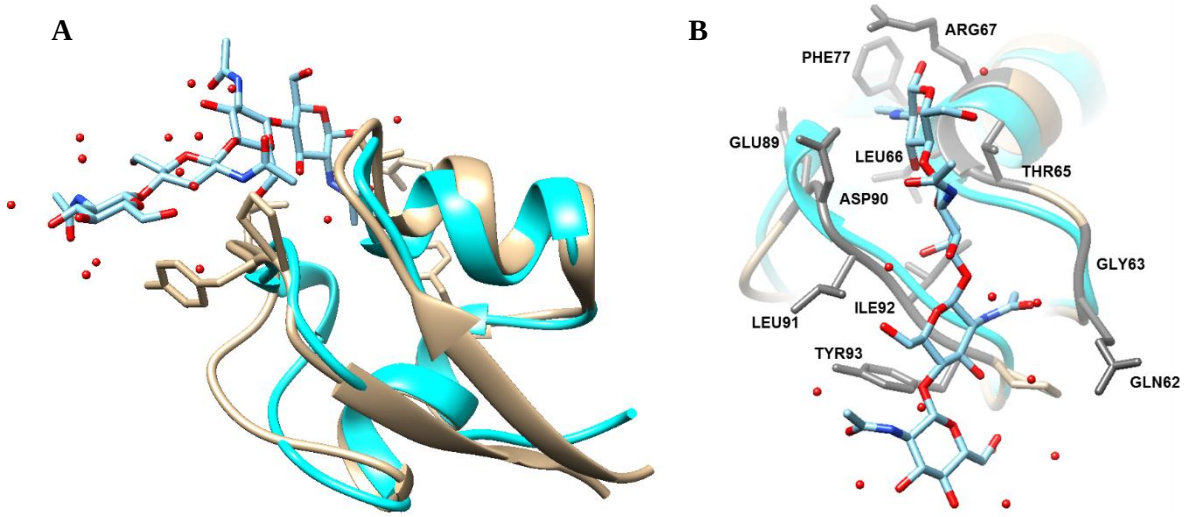


Figure 4.5 - Representation of the superposition between *Ct*CBM50 similarity model with Mo0v in complex with $GlcNAc_4$ and respective residues involved in ligand binding. A) Representation of the superposition between *Ct*CBM50 similarity model and Mo0v in complex with $GlcNAc_4$. Ribbon representation of *Ct*CBM50 in cyan and Mo0v in tan. $GlcNAc_4$ is represented by sticks. B) Mo0v residues involved in hydrogen bonding with ligand are represented by grey sticks and labeled by three-letter code and number. Ribbon representation of *Ct*CBM50 in cyan and Mo0v in tan. The $GlcNAc_4$ is represented by sticks. The program Chimera was used for visualization.

In the following discussion, the amino acid residue nomenclature adopted was the three-letter code and numbering according to the authors (Gronenborn *et al.* 2015).

The amino acids that are involved in hydrogen bonding to the $GlcNAc_4$ in the Mo0v structure are Gln62, Gly63, Thr65, Leu66, Arg67, Phe77, Glu89, Asp90, Leu91, Ile92, Tyr93. With the sequence alignment option, the program allowed to check for the common residues between the binding site of Mo0v and our *Ct*CBM50 homology model (Figure 4.6).

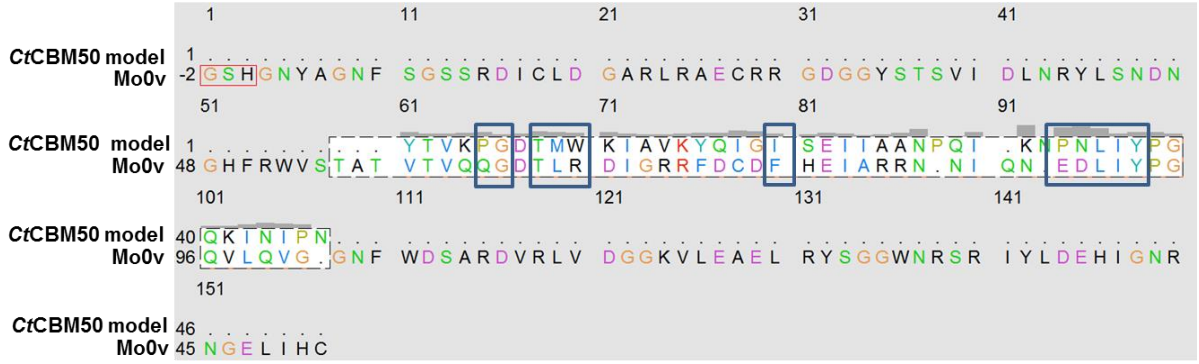


Figure 4.6 - Structure-based sequence alignment between *Ct*CBM50 and MoOv. Residues involved in Mo0v binding to $GlcNAc_4$ and respective aligned residues from *Ct*CBM50 are showed in blue rectangles.

The comparison between the binding site residues of Mo0v with the aligned residues of *Ct*CBM50 showed that 5 residues are conserved: Gly63, Thr65, Leu91, Ile92 and Tyr93. This could be an indication that *Ct*CBM50 may have a similar mode of binding to GlcNAc oligosaccharides.

4.4 Conclusions

In this Chapter we continued to study the specificity of *Ct*CBM50 based on the previous carbohydrate microarray results (Chapter 2).

*Ct*CBM50 revealed to be a novel chitin binding LysM domain, which was confirmed by binding to insoluble chitin. This study also allowed to obtain the binding affinity of *Ct*CBM50 to the pentasaccharide derived from chitin with a K_d of $1.1x10^{-6}$ M. However, further studies are required to investigate the binding to β-(1-4)-linked GlcNAc oligosaccharides with different DP and other oligosaccharides to acertain specificity.

Several crystallization screens were performed and a protein crystal was obtained in one condition, 0.1 M HEPES pH 7.5 and 1 M $LiSO_4$. However, optimization of this crystallizaytion condition was not successful and thus the screen should be repeated, as well as other screens, in order to achieve the conditions for protein crystallization. As *Ct*CBM50 is a small protein, in further work, it is recommended the N-terminal His-tag removal after purification since it can cause protein instability and can interfere with the crystallization process. Also other approaches could be implemented, such as NMR solution studies. NMR is more suitable for structural studies of difficult to crystallize, highly dynamic, small proteins as it is not necessary to obtain crystals.

Neverthless, a similarity model was made with the intention to predict the 3D structure of *Ct*CBM50. Also using this similarity model, a superposition with a mutated LysM module was made in order to look at the residues involved in binding to chitin oligosaccharide, β-(1-4)-linked GlcNAc tetrasaccharide.. The superposition showed that 5 residues are conserved in the two structures (Glycine, Threonine, Leucine, Isoleucine and Tyrosine), which can be an indication that the mode of binding to GlcNAc oligosaccharides can be similar in the two LysM domains.

Chapter 5 - *Rf*CBM62-1: A highly specific CBM for pectins

5.1 Introductiory remarks

The first CBM from family 62 was identified in 2011 (Montanier *et al.* 2011) from *Clostridium thermocellum* and is the only known crystallographic structure of a carbohydrate-binding module of family 62. Moreover, its structure has been solved in complex with xyloglucan and galactomannan-derived oligosaccharides (Montanier *et al.* 2011).

In *R. flavefaciens* this family of CBMs still lacks characterization and 3D structures of the apo and the liganded protein, that would provide information to understand the preference of this CBM for pectic polysaccharides.

Following the observations of the microarray analysis in Chapter 2, the binding of *Rf*CBM62-1 to pectin was explored in this chapter.

In this chapter the following tasks were performed: 1) assessement of binding to insoluble pectin by SDS-PAGE electrophoresis; 2) affinity gel electrophoresis with soluble pectic polysaccharides; 3) Microscale thermophoresis (MST) to evaluate the interaction of *Rf*CBM62-1 with galacturonan DP4 oligosaccharide; 3) preliminary crystallization assays aiming at solving the X-ray structure; and 4) a similarity model produced computationally to evaluate the predicted structure and the residues involved in the binding.

5.2 Materials and methods

5.2.1 Analysis of interaction with insoluble polysaccharides

The protocol for analysis of interaction with insoluble polysaccharides was performed as described in Section 4.2.1 from Chapter 4.

5.2.2 Affinity gel electrophoresis with soluble polysaccharides

Affinity gel electrophoresis consists in the preparation of a polyacrylamide matrix embedded with a soluble carbohydrate. The electrophoretic mobility is compared with a control native gel (without the polysaccharide) and the infused-carbohydrate gel. The interaction between the CBM and the carbohydrate results in a reduction of the CBM's electrophoretic mobility, being indicative of its affinity towards the ligand.

The protocol for the affinity gel electrophoresis encompasses the substitution of water for the solution with the carbohydrate in the native gel. In the described assays, we used 12.5% acrylamide gels polymerized with 0.1% (w/v) of the carbohydrate and a control gel in the absence of the carbohydrate. Bovine Serum Albumine (BSA) was included in the gels as a reference standard and a negative control.

Electrophoresis was performed in a 20 ºC room and a native buffer (1.5M Tris-HCl pH 8.8) was used for the running, for approximately 1 hour and 30 minutes (Mini-PROTEAN® tetra system from Biorad) at a constant voltage of 150V. Coomassie blue was used as staining solution to detect protein.

Binding can be visualized or calculated by measuring the migration distances in the gel of the CBMs and the reference protein. Migration distances can be measured between the top of the gel and the solvent front for both CBMs and BSA. A relative mobility for each CBM can be calculated as the migration distance of the CBM divided by the migration distance of the reference which allows to infer about the affinities (Tomme *et al.* 2000).

5.2.3 Microscale thermophoresis

The MST experiment was performed during a demonstration of the Monolith NT.115 equipment (from NanoTemper) assisted by Dr Aileen Justes. It is a microscale thermophoresis equipment that measures thermophoresis of fluorescently labeled molecules (Chapter 1, Section 1.4.2.1).

In this MST experiment a dilution series of up to 16 dilutions of non-labeled molecule (galacturonan DP4) was prepared. The concentration of the fluorescently labeled molecule was kept constant at estimated 20 µM and the concentration of the ligand was varied. Therefore, 20 µM of *Rf*CBM62-1 was labeled with a fluorescent dye (NT-647) using Monolith NT™ Protein Labeling Kits. 10 µl of the serial dilution of the galacturonan DP4 were mixed with 10 µl of the diluted fluorescent *Rf*CBM62-1. After mixing the interaction partners, the samples are left to reach the equilibrium for a couple of minutes. Mixed samples were loaded into glass capillaries and the MST analysis was performed using the Monolith NT.115.

5.2.4 Crystallization assays

Crystallization assays were performed as described in Section 5.2.3 from Chapter 5.

6.2.5 Phyre2: A bioinformatic tool for 3D structure prediction

Phyre2 procedure information is described in Section 5.2.4 from Chapter 5.

5.3 Results and discussion

5.3.1 *Rf*CBM62-1 binding with pectin

The carbohydrate microarrays results described in Section 2.3.3 in Chapter 2 showed that *Rf*CBM62-1 binds specifically to a pectic polysaccharide for which characterization is under analysis (pectin-8). With the intent to further understand the specificity of *Rf*CBM62-1, we performed affinity gel electrophoresis with an insoluble pectin derived from apple (76282, from Sigma-Aldrich).

Qualitative binding experiments were carried out with 50 µg of *Rf*CBM62-1 and 2 % (w/v) of the respective insoluble polysaccharide. The detailed protocol used for this experiment is described in Section 5.2.1.

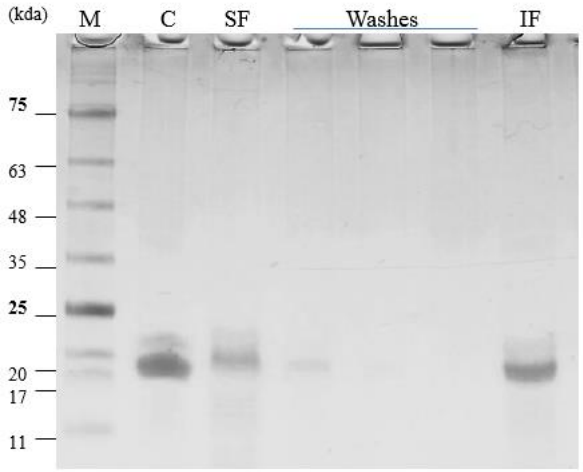


Figure 5.1– Qualitative binding analysis of *Rf*CBM62-1 with pectin from apple by SDS-PAGE with 13% acrylamide. Binding analysis of *Rf*CBM62-1 with 2% (w/v) insoluble pectin. C - Control (50µg of *Rf*CBM62-1, without polysaccharide); SF – Soluble fraction; IF – Insoluble fraction; M – Marker II from NZYTech®.

The SDS-PAGE results showed that *Rf*CBM62-1 binds to this insoluble pectin from apple, as comparing the protein band and the insoluble fraction band (polysaccharide fraction), there was a singnificant percentage of protein that precipitated along with the pectin.

The pectin-8, for which *Rf*CBM62-1 revealed to be highly specific in the carbohydrate microarray experiment (Chapter 2), could have a similar structure or contain a similar epitope present in this pectin from apple.

5.3.2 Binding affinity of *Rf*CBM62-1 with galacturonan DP4

Following the results from previous Section, which showed *Rf*CBM62-1 binding to pectin from apple, galacturonan DP4, α-(1-4)-galacturonic acid (GAT111, Elicityl), was tested in a MST equipment demonstration in our laboratory as described in Section 5.2.3. To perform this experiment with *Rf*CBM62-1, a fluorescent label (NT-647) was covalently attached to the protein. In the MST experiment we have kept the concentration of NT-647 labeled *Rf*CBM62-1 constant (20 µM), while the concentration of the non-labeled galacturonan DP4 was varied between 500 µM – 15 nM. The assay was performed in 50 mM HEPES, 100 mM NaCl and 2 mM $CaCl_2$ at a pH of 7.5.

After a short incubation the samples were loaded into MST NT.115 premium glass capillaries and the MST analysis was performed using the Monolith NT.115. A weak interaction of *Rf*CBM62-1 with galacturonan DP4 was observed (Figure 5.2), with a predicted K_d in the µM range. However, the experiment should be repeated using higher ligand concentrations, to reach the saturation plateau and thus accurately calculate K_d.

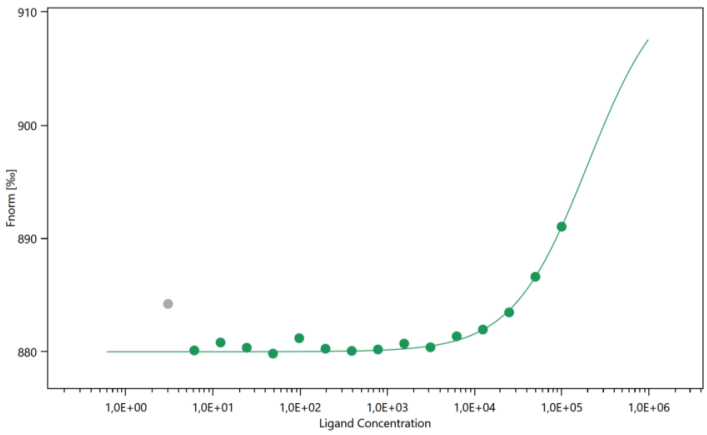


Figure 5.2 - MST binding experiment curve of *Rf*CBM62-1 with galacturonan DP4. The change in thermophoresis is expressed as the change in the normalized fluorescence. Ligand concentrations on the x-axis are plotted in nM. This graph was provided by Dr Ailen Justies (NanoTemper).

5.3.3 *Rf*CBM62-1 binding with pectin-related polysaccharides

The results from the previous section showed that an interaction occurs in the micromolar range for *Rf*CBM62-1 with galacturonan DP4.

With the intent to further understand the specificity of *Rf*CBM62-1, we performed affinity gel electrophoresis with 6 pectin-related polysaccharides. The polysaccharides tested were the following: 1) Galacturonate low methylated from apple (GAT100); 2) galacturonate high methylated from apple (GAT101); 3) galacturonate low methylated from citrus (GAT102); 4) galacturonate high methylated from citrus (GAT103) 5) rhamnogalacturonan from soy bean (P-RHAGN) and 6) polygalacturonic acid (P-PGACT). The first four polysaccharides were purchased from Elicityl and the last two from Megazyme. The results are shown in Figure 5.3.

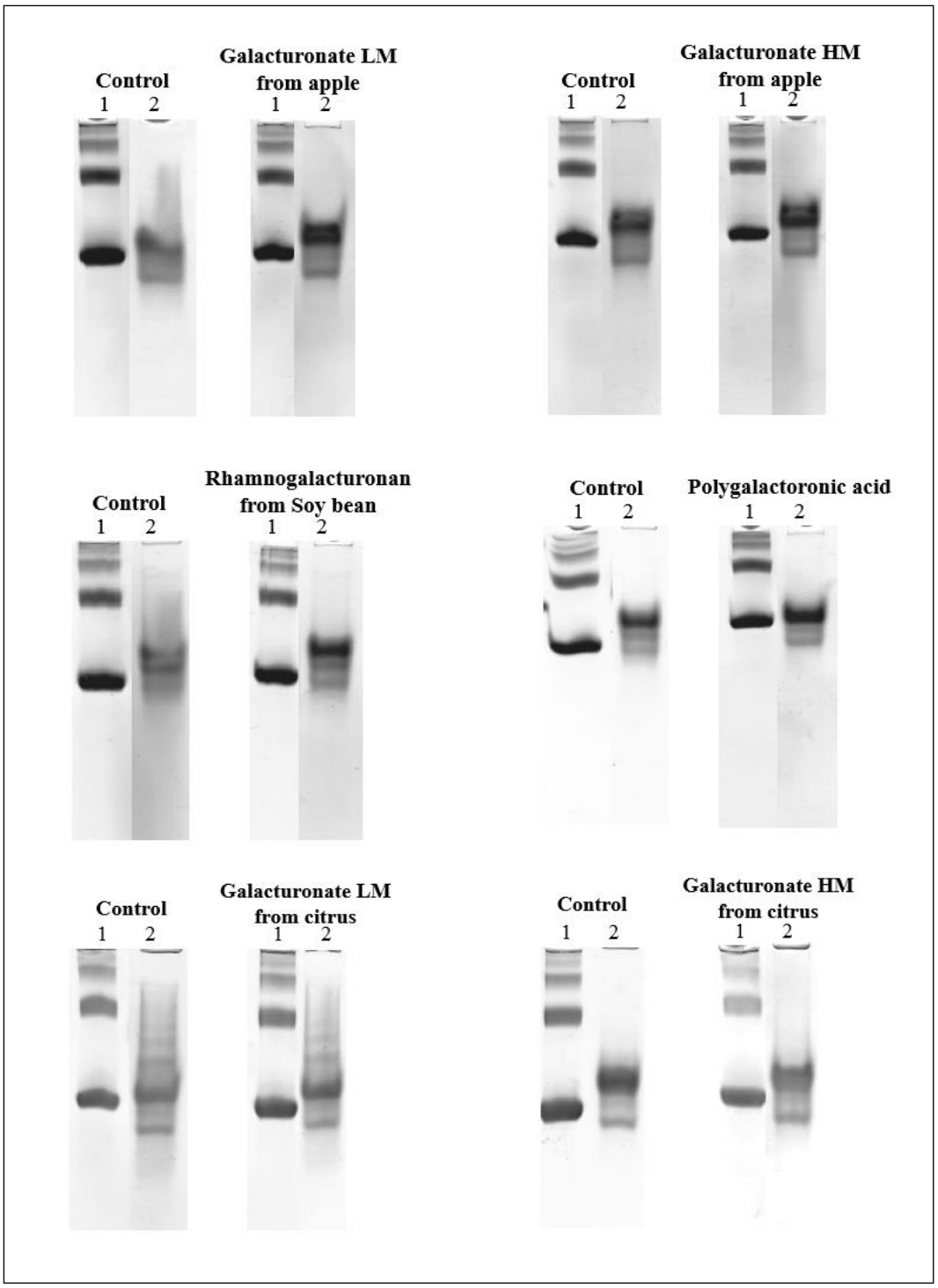


Figure 5.3 - Affinity gel electrophoresis of *Rf*CBM62-1 with pectin-related polysaccharides. Lane 1 – BSA; Lane 2 – *Rf*CBM62-1 (5µg). The respective control for each polysaccharide gel is located in the left. Polyacrylamide gels were prepared with 12.5% acrylamide and 0.1% of polysaccharide. LM – Low methylated; HM – High methylated. Control gels do not have any percentage of polysaccharide.

In the experimental conditions, no binding was observed to the different pectin-related polysaccharides.

The binding for *Rf*CBM62-1 observed in the carbohydrate microarray results (Chapter 3) showed high specificity for pectin-8, since 11 pectins more were included in the carbohydrate microarray set and no signal for those was observed. Also binding was detected for this pectin-8 by anti-β (1-4)-galactan, carbohydrate-directed monoclonal antibody used for microarray validation. The reported binding of this antibody was described to the β-(1-4)-/3 pectic galactan from lupin, of which galactose constitutes 91%, and from tomato, which are composed of > 60% galactose. The galactose percentage present in the rhamnogalacturonan from soy bean and polygalacturonic acid were 25 and 1% respectively. For the galacturonate polysaccharides from apple and citrus the percentage of galactose is not known. One hypothesis is that this pectin-8 has a higher percentage of galactose than these polysaccharides, however, further studies are needed before any conclusions. Studies with pectins from lupin and tomato are under way.

The galacturonan DP4 used in MST experiment (Section 5.3.2) was derived from galacturonate LM from apple used in this affinity electrophoresis. Although they have a similar sequence, no binding was detected in this experiment, in contrast with the MST experiment. The fact that MST detects interactions with higher sensitivity than an affinity electrophoresis experiment can be an explanation for these results. The galacturonan DP4 may only present part of the motif that is being recognized by *Rf*CBM62-1. However, further analysis should be carried to understand the specificity of *Rf*CBM62-1.

5.3.4 Crystallization assays

For crystallization assays of *Rf*CBM62-2, several screens (Table 5.1) were tested in 96-well plates using the vapor diffusion method and sitting drop technique. In the first attempt for crystallization four screens were tested (80!; PEG ion 1 & 8k; JCSG and JBS 1,2,3,4) with *Rf*CBM62-1 at a concentration of 15.5 mg/ml. As no crystals were observed a different screen was tested, PEG Ion 1 and 2, known to have generated crystals for other CBM families. No crystals were obtained for this screen. With the amount of protein available, it was only possible to test these screens. However, other screens should be tested, considering the limiting amount of protein available. As it was not abundant, the screens should be tested by stages, following the results progressively.

Table 5.1 – Experimental conditions used in crystallization assays for *Rf*CBM62-1 with an automated nanodrop dispensing equipment (known as crystallization robot).

Sample	Buffer	Concentration (mg/ml)	Drop ratio[a]	Drop[b] (µl)	Screens
***Rf*CBM62-1**	B	15.5	1:1	2	80!; PEG Ion 1 & 8k; JCSG and JBS 1,2,3,4
***Rf*CBM62-1**	A	11.4	1:1	2	PEG Ion 1 & 2

Buffer A: 50 mM HEPES pH=7.5, 100 mM NaCl and 2 mM $CaCl_2$
Buffer B: H_2O with 2 mM $CaCl_2$
[a] Protein:precipitant ration
[b] Final drop volume with protein+precipitant

5.3.5 *Rf*CBM62-1 3D structure prediction

In the absence of a 3D structure, ideally obtained using structural characterization methodologies such as X-ray crystallography or NMR, different bioinformatic tools are available that allow to infer on the expected polypeptide fold of a protein. The success of these calculations is strongly dependent on the number of possible similar 3D models of previously determined structures. This similarity is evaluated by the alignment of primary amino-acid sequences.

CBM62 (PDB code 2YFZ) is within the large *C. thermocellum* cellulosomal protein Cthe_2193 (defined as CtXyl5A) and is the only known structure of a CBM from family 62 (Montanier *et al.* 2011).

Using this structural information, associated to a sequence similarity of 35% between the 2 proteins, it was possible to have a glance on a probable 3D structure for *Rf*CBM62-1 (Figure 5.4).

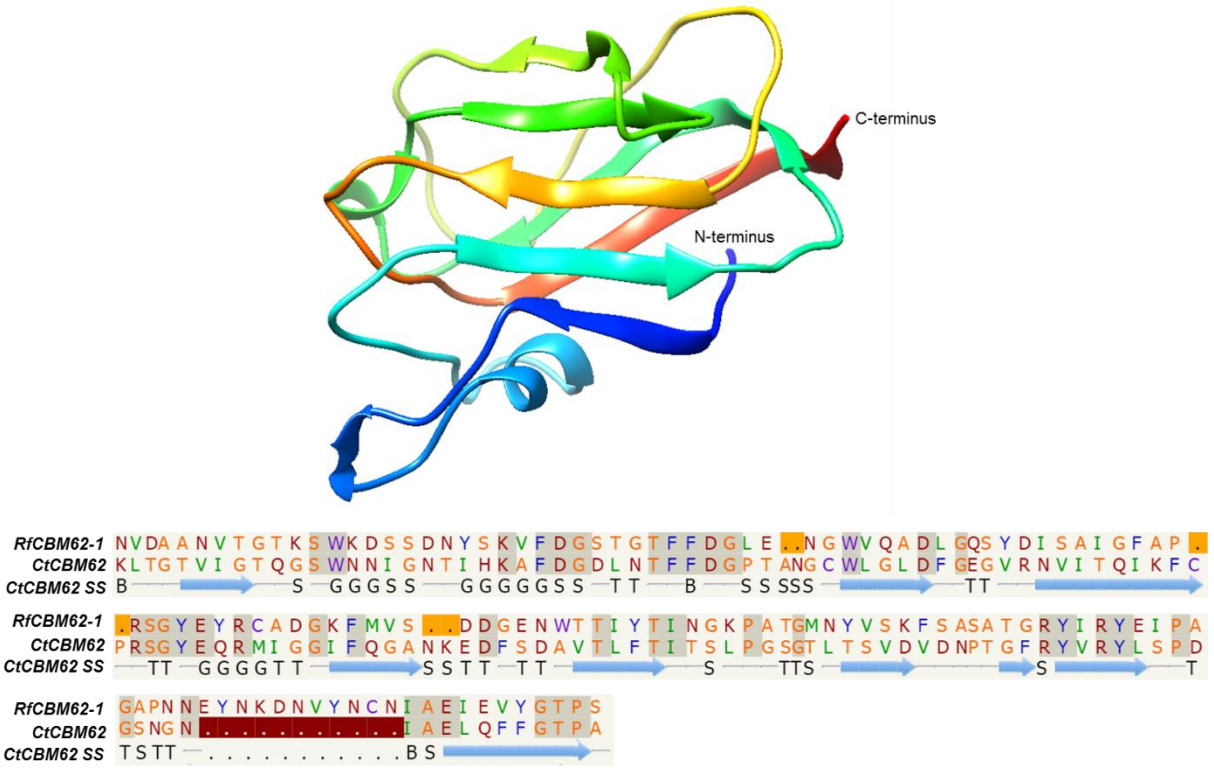


Figure 5.4 - Ribbon representation of *Rf*CBM62-1 similarity model calculated with Phyre2 and *Ct*CBM62 Chain A (PDB code 2YFZ) as a template. *Ct*CBM62 SS – Secondary structure of the template sequence. In the template secondary structure T corresponds to hydrogen bonded turn, G to 3-turn helix, S to bend and B to residue in isolated β-bridge. The program Chimera was used for visualization. Rainbow ribbon representation from N-terminus (blue) to C-terminus (red).

For the similarity model, residues from 5 to 145 of the query sequence were aligned with the template with 96% coverage. The *Rf*CBM62-1 homology model was modeled with 99.9% confidence and 35% identity with *Ct*CBM62. With this bioinformatic tool it was possible to predict the *Rf*CBM62-1 three-dimensional structure. According to the criteria of program Chimera, *Rf*CBM62-1 similarity model comprises 7 β-sheets and 3 α-helixes. Nevertheless, this is just a prediction and structural differences are expected in lower similarity stretches, in solvent exposed loops and possibly in the recognition site.

In order to compare our model with *Ct*CBM62 in complex with 6^1-α-D-Galactosyl-mannotriose (PDB code: 2YB7) and look at the residues involved in ligand binding, a structural superposition was calculated (Figure 5.5). 98 pruned atom pairs were aligned with a RMSD of 0.293Å.

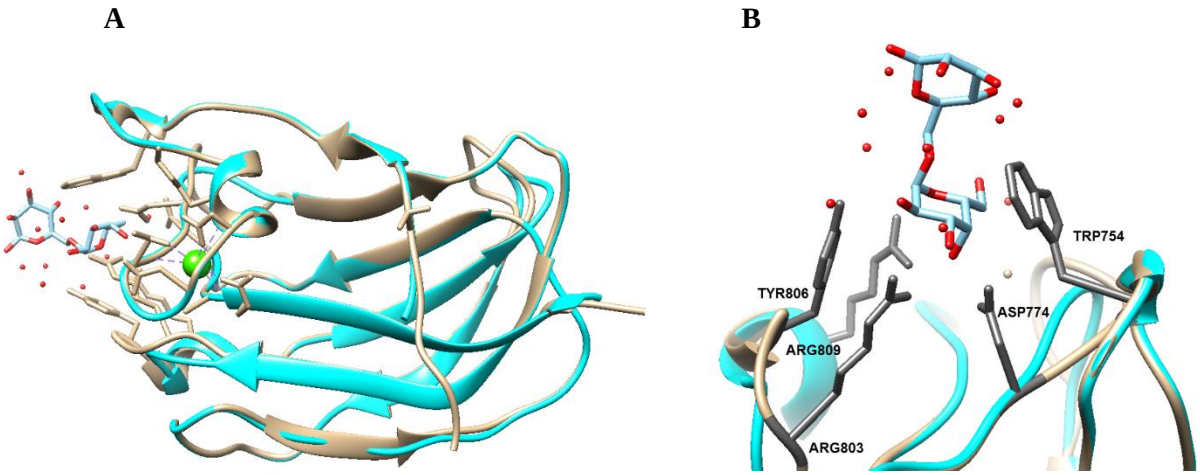


Figure 5.5 - Representation of the superposition between *Rf*CBM62-1 similarity model and CtCBM62 with 6^1-α -D-GalMan3 and respectively amino acid residues involved in ligand binding. A) Representation of the superposition between *Rf*CBM62-1 similarity model and *C*tCBM62 in complex with 6^1-α -D-GalMan3 and calcium (PDB code 2YB7). B) *Ct*CBM62 residues involved in hydrogen bonding with ligand are represented by grey sticks and labeled by three-letter code and number. Ribbon representation of *Rf*CBM62-1 in cyan and *Ct*CBM62 in tan. 6^1-α -D-GalMan3 is represented by sticks and calcium by a green sphere. The program Chimera was used for visualization.

In the following discussion, the amino acid residue nomenclature adopted were three-letter code and numbering is according to the authors (Montanier *et al.* 2011).

The amino acids involved in binding to 6^1-α-D-GalMan3 in the *Ct*CBM62 structure are Trp754, Asp774, Arg803, Tyr806 and Arg809. The superposition between the *Rf*CBM62-1 similarity model and *Ct*CBM62 structure (Figure 5.6) allowed to identify common residues between the two. The comparison between the binding site residues of *Ct*CBM62 with the aligned residues of *Rf*CBM62-1 showed that 4 of 5 residues are conserved: Trp754, Asp774, Tyr806 and Arg809.

Figure 5.6 - Structure-based sequence alignment between *Rf*CBM62-1 and *Ct*CBM62. Residues involved in *Ct*CBM62 binding to 6^1-α -D-GalMan$_3$ and respective aligned residues from *Rf*CBM62-1 are showed in blue rectangles. Residues involved in *Ct*CBM62 binding to calcium and respective aligned residues from *Rf*CBM62-1 are showed in yellow rectangles.

The authors refer that *Ct*CBM62 specificity is conferred through numerous interactions with the axial O4 of the target sugars, a feature that distinguishes galactose and arabinose from the other major sugars located in plant cell walls such as mannose, xylose, and glucose.

Also an important feature is calcium-mediated oligomerization, a feature that is conserved in several other CBM families, resulting in avidity effects that confer selectivity for polysaccharides rather than monovalent oligosaccharides (Montanier *et al.* 2011). The residues involved in calcium binding are Lys763, Asp766, Thr771, Ala888 and Glu869. These 5 residues are conserved in our model, which means that *Rf*CBM62-1 may also bind to calcium to oligomerize.

These results could be an indication that *Rf*CBM62-1 may have a similar mode of binding to the axial O4 of D-galactose and L-arabinose residues of pectic polysaccharides and to calcium.

5.4 Conclusions

In this Chapter we continued to study the specificity of *Rf*CBM62-1 based on the previous carbohydrate microarray results (Chapter 2). As the previous results showed that *Rf*CBM62-1 binds to a yet an uncharacterized pectin, we performed an analysis of interaction with a pectin from apple. Effectively, binding was observed for pectin from apple, however, this experiment should be repeated in future work.

A MST experiment was carried out with galacturonate DP4, a pectic oligosaccharide, and binding was observed with a K_d in the one digit µM range for the interaction. These results contributed to elucidate the structure of the pectin that binds to *Rf*CBM62-1 in the microarrays. This result led us to test 6 pectin-related soluble polysaccharides, using affinity gel electrophoresis, in order to understand the possible sequence features for binding to pectin-8. However, in these experimental conditions no binding was observed for these polysaccharides. These experiments should be repeated in future work.

Several crystallization screens were perfomed in order to structuraly characterize *Rf*CBM62-1, however, no crystals were obtained. *Rf*CBM62-1 is also a small protein, such as *Ct*CBM50, and the

suggestions for future work as in Chapter 4/Section 4.4 related to the removal of the N-terminal His-tag and NMR studies should also be considered for this protein.

The calculated similarity model predicted a 3D structure for *Rf*CBM62-1 comprised of 7 β-sheets and 3 α-helices. The superposition with a known structure of a *Ct*CBM62 showed that 4 of the 5 residues involved in binding with an oligosaccharide, 6^1-α -D-GalMan3, were conserved. Also the 5 residues involved in calcium binding, an important feature that contributes for CBM oligomerization and stabilization, were conserved.

Chapter 6 – Integrative conclusions and future work

6. Integrative conclusions and future work

The work developed in this thesis was focused on the study of proteins involved in plant cell-wall recognition and degradation (CBMs) from two microorganisms, *C. thermocellum* and *R. flavefaciens* FD-1.

The carbohydrate microarray designated Fungal and Plant PS set1 used in this study has proven to be a valid tool for initial high-throughput screening of newly identified CBMs or CBMs assigned to known families in the CAZy database for carbohydrate binding. The results from this initial screening will be important to select CBMs to follow-up for ligand specificity assignments and structural characterization. The emerging conclusion is that CBMs from *R. flavefaciens* FD1 showed to be more specific for a particular carbohydrate with a binding pattern that is less broad than CBMs from *C. thermocellum*. This is in agreement with results from the analysis of a wide range of CBM sequences from both organisms (Ribeiro *et al*, manuscript in preparation). This could be a reflection of the adaptive pressures of the microorganisms to their respective ecological niches. The microarray used is comprised mainly of polysaccharides with sequences that are found in plant and fungal cell walls. However, polysaccharide samples extracted from biological sources are heterogeneous macromolecules and possess heterogeneous sequences. Furthermore, these macromolecules may also contain contaminants resulting from the extraction/purification protocol. For this reason, the validation of the array and the structural analysis of the probes printed in the array is highly important to 'map' the macromolecule and be able to make assertive conclusions of the results. To understand which are the glycan epitopes being recognized by the CBM and assign their specificity, the macromolecules need to be depolymerized as oligosaccharides and thus sequence-defined 'designer' microarrays could be prepared for probing with the CBMs. To increase sequence-diversity some sequence-defined oligosaccharides or oligosaccharide mixtures were already included in the microarray analysed and enabled the assignment of a specific sequence specificity for certain CBMs. The designer microarrays are now under development to include an increased number and diversity of sequences and different degrees of polymerization of the carbohydrate chain. Future work will be implemented in order to identify the epitopes that are being recognized by CBMs.

High-throughput methods for expression and purification of CBMs are implemented so that the smallest amounts of protein are used. These methods work well for the carbohydrate microarray analysis, as proteins are stable for the period of analysis and can be analysed directly in their elution buffer. However, to produce CBMs at a larger scale required for biophysical and structural analysis, an optimization of expression and purification should be considered. The work described here showed that the conditions for high-throughput expression (Autoinduction protocol) might be not appropriate to scale-up the CBMs production. For the CBMs targeted in this Thesis the IPTG-induction protocol reveals to be a better choice to produce protein in higher amounts, as levels of expression are high and low contamination is observed when compared with the Auto-induction protocol. After purification, attempts were made to increase the proteins concentration. However, both *Ct*CBM50 and *Rf*CBM62-1

show a tendency to precipitate at higher concentration levels after storage at 4ºC, which suggests a possible instability of the proteins. Many reasons could account for this, but the influence of the relatively long N-terminal His-tags should be investigated. In future work, one of the strategies could be the comparison of the protein behavior with and without the N-terminal His-tag in thermofluorescence assays. Oligomerization should also be assessed and this can be done by dynamic light scattering (DLS), a technique used to determine the size distribution profile of small particles in suspension, or NMR.

The family 50 CBMs, or LysM domain, from *Clostridium thermocellum* still lacks structural characterization and the carbohydrate ligand specificities were not proven experimentally at the start of this thesis. We identified the *Ct*CBM50 as a novel chitin binding domains, displaying strong binding to insoluble chitin and to a chitin oligosaccharide, the pentasaccharide with the sequence GlcNAcβ4GlcNAcβ4GlcNAcβ4GlcNAcβ4GlcNAc. Efforts to crystallize *Ct*CBM50 for determination of its structure produced a preliminary crystallization condition with diffracting crystals. Unfortunately, the optimization assays were not successful.

The degradative machinery *R. flavefaciens* is not well characterized and new insights are important to understand how this microorganism degrades plant cell wall carbohydrates. In the microarrays *Rf*CBM62-1 showed to be highly specific for a pectin sample designated Pectin-8 (characterization is underway) that is part of a large collection isolated from a medicinal plant extract. In this thesis, weak binding with pectin from apple, analyzed by affinity gel electrophoresis, and a derived oligosaccharide (galacturonan DP4), analyzed by MST, was also observed. However, these results require further investigation. Crystallization assays for determination of *Rf*CBM62-1 structure were unsuccessful to this date.

It is known that the His-tag (or other tags) can interfere with the crystallization process, especially for small proteins. In future work, the removal of the N-terminal His-tag from *Ct*CBM50 and *Rf*CBM62-1 should be considered. In addition, these proteins are excellent candidates for NMR studies, due to their small size, which should be considered as a method of choice for the structural characterization. In anticipation of this, a preliminary 1D 1H NMR spectrum was performed and confirmed that *Ct*CBM50 is folded.

Trying to overcome the unsuccessful crystallization and in an attempt to derive relevant structural information, similarity models for both CBMs were made using Phyre2 and compared with known structures with >30% identity to analyze the residues at the binding site. For *Ct*CBM50, 5 residues of the model from 11 of the known template structure residues were conserved. In contrast, *Rf*CBM62-1 model contains 4 of 5 residues of the known template structure. This could mean that, when compared to the template structures, these CBMs may have a similar mode of binding to oligosaccharides.

The results described in this thesis contributed to increase the knowledge about ligand specificities of newly identified CAZY CBMs and open the way for the structural characterization of CBMs and CBM-ligand complexes. Understanding the versatility of plant and fungal saccharide sequences and

their recognition by different CBM families is of great importance. We hope to understand how these two cellulotic microorganisms evolved in different ecological niches. The adaptation of each microorganism is a reflection from their adaptive pressures to respective ecological niches, translating in divergent evolution of the proteome. These results are also important for biotechnological applications, such as production of biofuels and animal feed through the use of the multimodular complex cellulosome or modular enzymes.

References

Alahuhta, M. *et al.*, 2011. Structure of CBM4 from Clostridium thermocellum cellulase K. *Acta Crystallographica Section F: Structural Biology and Crystallization Communications*, 67(5), pp.527–530.

Arai, T. *et al.*, 2003. Characterization of a cellulase containing a family 30 carbohydrate-binding module (CBM) derived from Clostridium thermocellum CelJ: Importance of the CBM to cellulose hydrolysis. *Journal of Bacteriology*, 185(2), pp.504–512.

Bayer, E. a *et al.*, 2004. The cellulosomes: multienzyme machines for degradation of plant cell wall polysaccharides. *Annual review of microbiology*, 58, pp.521–54.

Berg Miller, M.E. *et al.*, 2009. Diversity and Strain Specificity of Plant Cell Wall Degrading Enzymes Revealed by the Draft Genome of Ruminococcus flavefaciens FD-1. *PLoS ONE*, 4(8), p.e6650.

Bergquist, P.L. *et al.*, 1999. Molecular diversity of thermophilic cellulolytic and hemicellulolytic bacteria. *FEMS Microbiology Ecology*, 28(2).

Bhunia, A., Bhattacharjya, S. & Chatterjee, S., 2012. Applications of saturation transfer difference NMR in biological systems. *Drug Discovery Today*, 17(9-10), pp.505–513.

Block, H. *et al.*, 2009. Immobilized-Metal Affinity Chromatography (IMAC). A Review. *Methods in Enzymology*, 463(C), pp.439–473.

Boivin, S., Kozak, S. & Meijers, R., 2013. Optimization of protein purification and characterization using Thermofluor screens. *Protein Expression and Purification*, 91(2), pp.192–206.

Boraston, A.B. *et al.*, 2006. A structural and functional analysis of α-glucan recognition by family 25 and 26 carbohydrate-binding modules reveals a conserved mode of starch recognition. *Journal of Biological Chemistry*, 281(1), pp.587–598.

Boraston, A.B. *et al.*, 2004. Carbohydrate-binding modules: fine-tuning polysaccharide recognition. *The Biochemical journal*, 382(Pt 3), pp.769–781.

Brulc, J.M. *et al.*, 2011. Cellulosomics, a gene-centric approach to investigating the intraspecific diversity and adaptation of ruminococcus flavefaciens within the rumen. *PLoS ONE*, 6(10).

Buist, G. *et al.*, 2008. LysM, a widely distributed protein motif for binding to (peptido)glycans. *Molecular Microbiology*, 68(4), pp.838–847.

Cantarel, B.I. *et al.*, 2009. The Carbohydrate-Active EnZymes database (CAZy): An expert resource for glycogenomics. *Nucleic Acids Research*, 37(1), pp.233–238.

Carvalho, A.L. *et al.*, 2004. The family 11 carbohydrate-binding module of Clostridium thermocellum Lic26A-Cel5E accommodates β-1,4- and β-1,3-1,4-mixed linked glucans at a single binding site. *Journal of Biological Chemistry*, 279(33), pp.34785–34793.

Chai, W. *et al.*, 2003. Neoglycolipid technology: deciphering information content of glycome. *Methods in enzymology*, 362, pp.160–95.

Charnock, S.J. *et al.*, 2000. The X6 "thermostabilizing" domains of xylanases are carbohydrate-binding modules: Structure and biochemistry of the Clostridium thermocellum X6b domain. *Biochemistry*, 39(17), pp.5013–5021.

Coelho, H. *et al.*, 2015. The Quest for Anticancer Vaccines: Deciphering the Fine-Epitope Specificity of Cancer-Related Monoclonal Antibodies by Combining Microarray Screening and Saturation Transfer Difference NMR. *Journal of the American Chemical Society*, 137(39), pp.12438–12441.

Correia, M. *et al.*, 2011. Structure and function of an arabinoxylan-specific xylanase. *Journal of Biological Chemistry*, 286(25), pp.22510–22520.

Cuív, P. *et al.*, 2013. Extending the cellulosome paradigm: The modular clostridium thermocellum cellulosomal serpin pinA is a broad-spectrum inhibitor of subtilisin-like proteases. *Applied and Environmental Microbiology*, 79(19), pp.6173–6175.

Davies, G.J. & Henrissat, B., 2013. Cracking the code, slowly: the state of carbohydrate-active enzymes in 2013. *Current Opinion in Structural Biology*, 23(5), pp.649–651.

Din, N. *et al.*, 1994. C1-Cx revisited: intramolecular synergism in a cellulase. *Proceedings of the National Academy of Sciences of the United States of America*, 91(24), pp.11383–11387.

Ezer, A. *et al.*, 2008. Cell surface enzyme attachment is mediated by family 37 carbohydrate-binding modules, unique to Ruminococcus albus. *Journal of Bacteriology*, 190(24), pp.8220–8222.

Feizi, T., 2003. Carbohydrate microarrays — a new set of technologies at the frontiers of glycomics. *Current Opinion in Structural Biology*, 13(5), pp.637–645.

Fontes, C.M.G. a. & Gilbert, H.J., 2010. Cellulosomes: Highly Efficient Nanomachines Designed to Deconstruct Plant Cell Wall Complex Carbohydrates. *Annual Review of Biochemistry*, 79(1), pp.655–681.

Freier, D., Mothershed, C.P. & Wiegel, J., 1988. Characterization of Clostridium thermocellum JW20. *Applied and environmental microbiology*, 54(1), pp.204–211.

Freire, E., Mayorgal, O.L. & Straume, M., 1990. Isothermal Titration. *Analytical Chemistry*, 62(18), pp.950–959.

Fukui, S. *et al.*, 2002. Oligosaccharide microarrays for high-throughput detection and specificity assignments of carbohydrate-protein interactions. *Nature Biotechnology*, 20(10), pp.1011–1017.

Georgelis, N., Yennawar, N.H. & Cosgrove, D.J., 2012. Structural basis for entropy-driven cellulose binding by a type-A cellulose-binding module (CBM) and bacterial expansin. *Proceedings of the National Academy of Sciences*, 109(37), pp.14830–14835.

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

Gilbert, H.J., Knox, J.P. & Boraston, A.B., 2013. Advances in understanding the molecular basis of plant cell wall polysaccharide recognition by carbohydrate-binding modules. *Current Opinion in Structural Biology*, 23(5), pp.669–677.

Gorshkova, T. a *et al.*, 2010. Formation of plant cell wall supramolecular structure. *Biochemistry.* 75(2), pp.159–72.

Griffiths, G. & Sánchez-serrano, J.J., 2001. The location of the ligand binding site of carbohydrate binding modules that have evolved from a common sequence is not conserved. *Journal of Biological Chemistry*

Gronenborn, A.M. *et al.*, 2015. Structural Insight into Fungal Cell Wall Recognition by a CVNH Protein with a Single LysM Domain Article Structural Insight into Fungal Cell Wall Recognition by a CVNH Protein with a Single LysM Domain. *Structure/Folding and Design*, 23(11), pp.2143–2154.

Harris, P.J. & Stone, B. a, 2008. Chemistry and Molecular Organization of Plant Cell Walls. *Biomass Recalcitrance: Deconstructing the Plant Cell Wall for Bioenergy.* pp 61-93

Henshaw, J.L. *et al.*, 2004. The family 6 carbohydrate binding module CmCBM6-2 contains two ligand-binding sites with distinct specificities. *Journal of Biological Chemistry*, 279(20), pp.21552–21559.

Heredia, A., Jiménez, A. & Guillénn, R., 1995. Composition of plant cell walls. *Zeitschrift für Lebensmittel-Untersuchung und -Forschung*, 200(1), pp.24–31.

Hernandez-Gomez, M.C. *et al.*, 2015. Recognition of xyloglucan by the crystalline cellulose-binding site of a family 3a carbohydrate-binding module. *FEBS Letters*, 589(18), pp.2297–2303.

Herve, C. *et al.*, 2010. Carbohydrate-binding modules promote the enzymatic deconstruction of intact plant cell walls by targeting and proximity effects. *Proceedings of the National Academy of Sciences*, 107(34), pp.15293–15298.

Horlacher, T. & Seeberger, P.H., 2008. Carbohydrate arrays as tools for research and diagnostics. *Chemical Society reviews*, 37(7), pp.1414–1422.

Jerabek-Willemsen, M. *et al.*, 2014. Microscale Thermophoresis: Interaction analysis and beyond. *Journal of Molecular Structure*, 1077, pp.101–113.

Jiang, D. *et al.*, 2012. Crystal structure of 1,3Gal43A, an exo-β-1,3-galactanase from Clostridium thermocellum. *Journal of Structural Biology*, 180(3), pp.447–457.

Jones, L., Seymour, G.B. & Knox, J.P., 1997. Localization of Pectic Galactan in Tomato Cell Walls Using a Monoclonal Antibody Specific to (1-4)-β-D-Galactan. *Plant physiology*, 113(4), pp.1405–1412.

Kelly, L. a. *et al.*, 2015. The Phyre2 web portal for protein modelling, prediction, and analysis. *Nature Protocols*, 10(6), pp.845–858.

Koharudin, L.M.I. *et al.*, 2011. Structure-function analysis of a CVNH-LysM lectin expressed during plant infection by the rice blast fungus *Magnaporthe oryzae*. *Structure*, 19(5), pp.662–674.

Lammerts Van Bueren, A. *et al.*, 2004. α-Glucan recognition by a new family of carbohydrate-binding modules found primarily in bacterial pathogens. *Biochemistry*, 43(49), pp.15633–15642.

Leavitt, S. & Freire, E., 2001. Direct measurement of protein binding energetics by isothermal titration calorimetry. *Current Opinion in Structural Biology*, 11(5), pp.560–566.

Linman, M.J. *et al.*, 2009. Surface Plasmon Resonance Study of Protein-Carbohydrate Interactions using Biotinylated Sialosides *Anal Chem* , 80(11), pp.4007–4013.

Liu, Y. *et al.*, 2007. Neoglycolipid Probes Prepared via Oxime Ligation for Microarray Analysis of Oligosaccharide-Protein Interactions. *Chemistry & Biology*, 14(7), pp.847–859.

Liu, Y. *et al.*, 2012. Neoglycolipid-Based Oligosaccharide Microarray System: Preparation of NGLs and Their Noncovalent Immobilization on Nitrocellulose-Coated Glass Slides for Microarray Analyses, 808.

Liu, Y., Palma, A.S. & Feizi, T., 2009. Carbohydrate microarrays: key developments in glycobiology. *Biological Chemistry*, 390(7), pp.647–656.

Lynd, L., 2002. Microbial Cellulose Utilization: Fundamentals and Biotechnology. *Microbiology and Molecular Biology Reviews*, 66(3), pp.506–577.

Marcus, S.E. *et al.*, 2010. Restricted access of proteins to mannan polysaccharides in intact plant cell walls. *Plant Journal*, 64(2), pp.191–203.

McCartney, L., Marcus, S.E. & Knox, J.P., 2005. Monoclonal Antibodies to Plant Cell Wall Xylans and Arabinoxylans.. *J. Histochem. Cytochem.*, 53(4), pp.543–546.

Meikle, P.J. *et al.*, 1994. A (1-3,1-4)-beta-glucan-specific monoclonal antibody and its use in the quantitation and immunocytochemical location of (1-3,1-4)-beta-glucans. *The Plant journal : for cell and molecular biology*, 5(1), pp.1–9.

Meikle, P.J. *et al.*, 1991. The location of (1-3)-β-glucans in the walls of pollen tubes of *Nicotiana alata* using a (1-3)-β-glucan-specific monoclonal antibody. *Planta*, 185(1), pp.1–8.

Mesnage, S. *et al.*, 2014. Molecular basis for bacterial peptidoglycan recognition by LysM domains. *Nature Communications*, 5.

Mizutani, K. *et al.*, 2012. Influence of a mannan binding family 32 carbohydrate binding module on the activity of the appended mannanase. *Applied and Environmental Microbiology*, 78(14), pp.4781–4787.

Moller, I. *et al.*, 2008. High-throughput screening of monoclonal antibodies against plant cell wall glycans by hierarchical clustering of their carbohydrate microarray binding profiles. *Glycoconjugate Journal*, 25(1), pp.37–48.

Montanier, C. *et al.*, 2009. Evidence that family 35 carbohydrate binding modules display conserved specificity but divergent function. *Proceedings of the National Academy of Sciences of the United States of America*, 106(9), pp.3065–3070.

Montanier, C.Y. *et al.*, 2011. A Novel , Noncatalytic Carbohydrate-binding Module Displays Specificity for Galactose-containing Polysaccharides through. , 286(25), pp.22499–22509.

Naismith, J.H. & Field, R. a., 1996. Structural basis of trimannoside recognition by concanavalin A. *Journal of Biological Chemistry*, 271(2), pp.972–976.

Najmudin, S. *et al.*, 2006. Xyloglucan is recognized by carbohydrate-binding modules that interact with beta-glucan chains. *Journal of Biological Chemistry*, 281(13), pp.8815–8828.

Nguyen, H.H. *et al.*, 2015. Surface plasmon resonance: A versatile technique for biosensor applications. *Sensors*, 15(5), pp.10481–10510.

Ohnuma, T. *et al.*, 2009. Structure and Function of Family 50 Carbohydrate Binding Modules (LysM Domains) from *Pteris ryukyuensis* Chitinase-A. *Journal of Applied Glycoscience*, 56(2), pp.97–104..

Overmann, R.G., 2006. Lignocellulose - Deomposing bacteria and their enzyme systems. *The Prokaryotes*, 2, pp.578–617

Palma, A.S. *et al.*, 2012. Neoglycolipid-Based "Designer" Oligosaccharide Microarrays to Define beta-Glucan Ligands for Dectin-1. *Methods in Molecular Biology*. 808(1), pp.337–359.

Palma, A.S. *et al.*, 2014. The neoglycolipid (NGL)-based oligosaccharide microarray system poised to decipher the meta-glycome. *Current Opinion in Chemical Biology*, 18, pp.87–94.

Palma, A.S. *et al.*, 2015. Unravelling glucan recognition systems by glucome microarrays using the designer approach and mass spectrometry. *Molecular and Cellular Proteomics*, (5), pp.3105–3117.

Park, S., Lee, M.-R. & Shin, I., 2008. Carbohydrate microarrays as powerful tools in studies of carbohydrate-mediated biological processes. *Chemical Communications*, (37), p.4389.

Pedersen, H.L. *et al.*, 2012. Versatile high resolution oligosaccharide microarrays for plant glycobiology and cell wall research. *Journal of Biological Chemistry*, 287(47), pp.39429–39438.

Pettolino, F. a *et al.*, 2001. A (1-4)-beta-mannan-specific monoclonal antibody and its use in the immunocytochemical location of galactomannans. *Planta*, 214(2), pp.235–242.

Reyes-Ortiz, V. *et al.*, 2013. Addition of a carbohydrate-binding module enhances cellulase penetration into cellulose substrates. *Biotechnology for biofuels*, 6(1), p.93.

Ribeiro, T. *et al.*, 2010. Family 42 carbohydrate-binding modules display multiple arabinoxylan-binding interfaces presenting different ligand affinities. *Biochimica et Biophysica Acta - Proteins and Proteomics*, 1804(10), pp.2054–2062

Rillahan, C.D. & Paulson, J.C., 2011. Glycan Microarrays for Decoding the Glycome. *Annual Review of Biochemistry*, 80, pp797-823

Rhodes G, 2006 Crystallography Made Crystal Clear - A Guide for Users of Macromolecular Models.

Romão, M.J., 1996. Cristalografia de Proteinas: metodologias e aplicações me Bioquímica. *Boletim de Biotecnologia*, 53, pp.18–36.

Roy, S. & Kumar, V., 2014. A Practical Approach on SDS PAGE for Separation of Protein. *International Journal of Science and Research* , 3(8), pp.955–960.

Scheller, H.V. & Ulvskov, P., 2010. Hemicelluloses. *Annual Review of Plant Biology*, 61(1), pp.263–289.

Schutter, K. & Van Damme, E.J.M., 2015. Protein-carbohydrate interactions, and Beyond. *Molecules*, 20(8), pp.15202–15205.

Shipp, E.L. & Hsieh-Wilson, L.C., 2007. Profiling the Sulfation Specificities of Glycosaminoglycan Interactions with Growth Factors and Chemotactic Proteins Using Microarrays. *Chemistry and Biology*, 14(2), pp.195–208.

Shoseyov, O., Shani, Z. & Levy, I., 2006. Carbohydrate Binding Modules: Biochemical Properties and Novel Applications. *Microbiology and Molecular Biology Reviews*, 70(2), pp.283–295.

Song, X. *et al.*, 2011. Shotgun Glycomics: A Microarray Strategy for Functional Glycomics. , 8(1), pp.85–90.

Sticklen, M.B., 2008. Plant genetic engineering for biofuel production: towards affordable cellulosic ethanol. *Nature Reviews Genetics*, 9(6), pp.433–443.

Stoll, M. & Feizi, T., 2010. Software Tools for Storing , Processing and Displaying Carbohydrate Microarray Data. *Glyco-Bioinformatics*, pp.1–22.

Taiz, L. & Zeiger, E., 2002. *Plant Physiology*.

Tang, P.W. *et al.*, 1985. Novel approach to the study of the antigenicities and receptor functions of carbohydrate chains of glycoproteins. *Biochemical and biophysical research communications*, 132(2), pp.474–80.

Taylor, G., 2003. The phase problem. *Acta Crystallographica - Section D Biological Crystallography*, 59(11), pp.1881–1890.

Tomme, P. *et al.*, 2000. Affinity electrophoresis for the identification and characterization of soluble sugar binding by carbohydrate-binding modules. *Enzyme and Microbial Technology*, 27(7), pp.453–458.

Vanholme, R. *et al.*, 2010. Lignin Biosynthesis and Structure. *Plant Physiology*, 153(3), pp.895–905.

Varner, J.E. *et al.*, 1989. Plant cell wall architecture. *Cell*, 56(2), pp.231–239..

Verma, A.K. *et al.*, 2015. The family 6 Carbohydrate Binding Module (CtCBM6) of glucuronoxylanase (CtXynGH30) of Clostridium thermocellum binds decorated and undecorated xylans through cleft A. *Archives of Biochemistry and Biophysics*, 575, pp.8–21.

Viegas, A. *et al.*, 2008. Molecular determinants of ligand specificity in family 11 carbohydrate binding modules - An NMR, X-ray crystallography and computational chemistry approach. *FEBS Journal*, 275(10), pp.2524–2535.

Viegas, A. *et al.*, 2011. Saturation-transfer difference (STD) NMR: A simple and fast method for ligand screening and characterization of protein binding. *Journal of Chemical Education*, 88(7), pp.990–994.

De Vries, R.P. *et al.*, 2001. Aspergillus Enzymes Involved in Degradation of Plant Cell Wall Polysaccharides. *Microbiology and Molecular Biology Reviews*, 65(4), pp.497–522.

Wang, D. *et al.*, 2002. Carbohydrate microarrays for the recognition of cross-reactive molecular markers of microbes and host cells. *Nat Biotech*, 20(3), pp.275–281.

Wang, L. *et al.*, 2014. Cross-platform comparison of glycan microarray formats. *Glycobiology*, 24(6), pp.507–517.

Willats, W.G.T., Marcus, S.E. & Knox, J.P., 1998. Generation of a monoclonal antibody specific to (1→5)-α-l-arabinan. *Carbohydrate Research*, 308(1-2), pp.149–152.

Wong, J.E.M.M. *et al.*, 2014. Cooperative binding of LysM domains determines the carbohydrate affinity of a bacterial endopeptidase protein. *FEBS Journal*, 281(4), pp.1196–1208.

Yaniv, O. *et al.*, 2014. Fine-structural variance of family 3 carbohydrate-binding modules as extracellular biomass-sensing components of Clostridium thermocellum anti-σI factors. *Acta Crystallographica Section D: Biological Crystallography*, 70(2), pp.522–534.

Zou, Y.F. *et al.*, 2015. Immunomodulating pectins from root bark, stem bark, and leaves of the Malian medicinal tree Terminalia macroptera, structure activity relations. *Carbohydrate Research*, 403, pp.167–173.

www.microbewiki.kenyon.edu, last acess 16.09.2016

www.cazy.org, last acess 14.09.2016

www.malvern.com last acess 14.09.2016

www.moleculardimensions.com, last acess 16.09.2016

Supplementary material

Supplementary Table 1 - Carbohydrate binding modules (CBMs) investigated for carbohydrate binding using carbohydrate microarrays and their modular organization. CBMs under study are highlighted in bold and respective recombinant protein sequence, protein identification, family, molecular weight, base pairs, extinction coefficient and isoelectric point are depicted. N-terminal His-tag is highlighted at yellow.

Microorganism	Molecular Architecture	Recombinant protein sequence	Protein ID	Family	Molecular weight (KDa)	Base pairs	Extinction Coefficient (M^{-1})	Isoeletric point
C. thermocellum	**CBM3**	MGSSHHHHHHSSGPQQGLRQDGTKGLKIQYYSRKPHDSAGIDFSFRMFNTGNEAIDLKDVKVRYYFKEDVSIDEMNWAVYFYSLGSEKDVQCRFYELPGKKEANKYLEITFKSGTLSPNDVMYITGEFYKNDWTKFEQRDDYSYNPADSYSDWKRMTAYISNKLVWGIEPN	Cthe_0059	3	20,1	516	42860	6.37
	CBM4-GH9-CBM3-DOC1	MGSSHHHHHHSSGPQQGLRPYKNDLLYERTFDEGLCYPWHTCEDSGGKCSFDVVDVPGQPGNKAFAVTVLDKGQNRWSVQMRHRGLTLEQGHTYRVRLKIWADASCKVYIKIGQMGEPYAEYWNNKWSPYTLTAGKVLEIDETFVMDKPTDDTCEFTFHLGGELAATPPYTVYLDDVSLYDPEY	Cthe_0413	4	21,0	1031	45630	5.62
	SLH-SLH-SLH-CBM54-GH16-CBM4-1-**CBM4-2**-CBM4-CBM4	MGSSHHHHHHSSGPQQGLRIYNGTFDQGPNRMGFWNFVVDSTAKATYYIGSDVNERRFETRIEKGGTSRGAIRLVQPGINIENGKTYKVSFEASAANTRTIEVEIASNLHNSSIFATTFEISKESKIYEFEFTMDKDSDKNGELRFNLGGSNVNVYIDNVVMKRVSTDEVE	Cthe_2809	4	19,2	570	14440	6.16
	GH2-**CBM6**-DOC1	MGSSHHHHHHSSGPQQGLRPVPRSAFTRIEAESYDAQSGIQTEDCSEGGKDVGYIENGDFVVYKAIDFGRGAASFKARVASATSGGNIELRIDSIDGPVVGICPVAGTGGWQEWADATCEVSDLKGVHDLYLKFTGGSGYLLNVNWFTFVEGNSDED	Cthe_2197	6	16,8	471	24075	4.82
	CBM6-DOC1	MGSSHHHHHHSSGPQQGLREPRSAFTRIEAESYNGQSGIQTENCSEGGMDVGYIENGDYVVYKNIDFGKGAASFKARVASATSGGNIELRIDSIDGPVVGICPVAGSGGWQQWVDATCEVSGLKGVHDLYLKFTGGSGYLLNINWFTFVEGNNDE	Cthe_2195	6	16,6	465	25565	5.21
	GH43-**CBM13**-DOC1	MGSSHHHHHHSSGPQQGLRTRYKLVNKNSGKVLDVLDGSVDNAAQIVQWTDNGSLSQQWYLVDVGGGYKKIVNVKSGRALDVKDESKEDGGVLIQYTSNGGYNQHWKFTDIGDGYYKISSRHCGKLIDVRKWSTEDGGIIQQWSDAGGTNQHWKLVLVSS	Cthe_0661	13	17,8	480	43430	8.60

	CBM22-GH10-DOC1	MGSSHHHHHHSSGPQQGLRAEGNLLFNPGFELGSTEGWYPYGECTIEAVGTEAHSGNYSVFVTDRTQDWNGVAQDMLDKLTVGMTYQVSAWVKVAGTGSHQVKISMKKVETGKEPVYDNIASITVEGSEWYRLSGPYSYTGTNVTNLELYIEGPQPGVSYYVDDVTVTEVGSA	Cthe_2590	22	18,9	519	38390	4.98
	CBM25	MGSSHHHHHHSSGPQQGLRFRLVYSGILAKNNPENLYAVIGYGNNLAWEDIESYSMRKIGDQKYELLFPVKRPGNINIAFKDDADNWDNNSGMNYCFENHVYQGSH	Cthe_0956	25	12,2	318	21430	6.42
	GH5-**CBM32**-DOC1	MGSSHHHHHHSSGPQQGLRAGSIAQNKPVYASSTEPGLGNTPEKAVDGNIATRWSSDYSDNQYIYVDLLDEYEIERVYIEWEAAYARQYKIQVSNDAVTWTDVYTEYNGDGDIDDIYLEARGRYVRIYCMQRATQYGNSIFELGVYPKGGIA	Cthe_0821	32	17,2	456	38850	4.97
	CBM42	MGSSHHHHHHSSGPQQGLRYGQFMKFESSNYRGYYIRVKSFSGRIDPYVNPVEDSMFKIVPGLADPSCISFESKTYPGYYLKHENFRVILKKYEDTDLFREDATFRVVPGWADENMISFQSYNYPYRYIRHRDFELYIENIKTDLDRKDATFIGIKVD	Cthe_1273	42	18,8	474	26360	7.10
	CBM50	MGSSHHHHHHSSGPQQGLRYTVKPGDTMWKIAVKYQIGISEIIAANPQIKNPNLIYPGQKINIPNI	Cthe_0300	50	7,4	198	9970	9.78
R. flavefaciens	**CBM6**-DOC1	MGSSHHHHHHSSGPQQGLRGSGGSTDDIIEAEKYDIQKGIQTENCSEGGSDVAYIENGDYIGFKNIDFGSGTDSISFRIGSNGAEASIEVRLGAADGKLIGTLPVKSTGGWQTWNTQTCAIENTSGRNDVYFVFKGGDGYLFNINWWKPDKPSEPI	3747	6	16,8	468	29575	5.08
	GH43-**CBM13**-CBM13-DOC1	MGSSHHHHHHSSGPQQGLRSGTELLSGVPYFITNVNSGLSLDLPEGKLDNGTNIQQWDFNKLWAQQWRIISVDKEWCRIVSLGDEGKCIAVAKDTADDGTNVELQTYTGADNQLFKFVKSGSSYGIVSKCSGGKGALDVFEWSKENGGNVNQFAYNEYACQLWNIAPV	2115	13	18,5	504	40700	5.60
	GH30-**CBM62-1**-CBM62-2-DOC1	MGSSHHHHHHSSGPQQGLRTNKINVDAANVTGTKSWKDSSDNYSKVFDGSTGTFFDGLENGWVQADLGQSYDISAIGFAPRSGYEYRCADGKFMVSDDGENWTTIYTINGKPATGMNYVSKFSASATGRYIRYEIPAGAPNNEYNKDNVYNCNIAEIEVYGTPS	3398	62	18,0	492	33015	5.59

Supplementary Table 2 - List of all saccharide probes included in the Fungal and Plant polysaccharide set 1. The microarray is comprised of polysaccharide samples from different sources, representative of major sequences found in fungal and plant cell walls, and a few selected sequence-defined oligosaccharides prepared as neoglycolipids (depicted in green). The probes are grouped according to the printing layout (set position).

Pos[a]	Probe	Source[b]	Predominant oligosaccharide sequence/ Monosaccharide composition
1	Dextran	*Leuconostoc mesenteroides* (Sigma, D4876)	α-(1-6) glucose
2	Pullulan	*Pullularia pullulans* (Megazyme, P-PULLN)	Mixed-linked α-(1-4;1-6)-glucose
3	Curdlan (50 mM NaOH)	*Agrobacterium* sp.ATCC31749 (IC)	Linear β-(1-3)-glucose
4	NSG-β-glucan	*Saccharomyces cerevisiae* (Biothera)	Linear β-(1-3)-glucose backbone with occasional monoglucosyl β1-6 glucose branches
5	PGG-β-glucan	*Saccharomyces cerevisiae* (Biothera)	Linear β-(1-3)-glucose backbone with occasional monoglucosyl β1-6 glucose branches
6	Lentinan	*Lentinus edodes* (IC)	β-(1-3)-glucan with β-(1-6)-branching 1:1
7	Grifolan	*Grifola frondosa* (IC)	β-(1-3)-glucose backbone with highly ramified oligomeric β-(1-6)-glucose branches
8	Barley glucan	Barley flour (Megazyme, B-BGBL)	Mixed-linked β-(1-3;1-4)-glucose. Also contains 2% arabinose and 0.2% xylose.
9*	Barley glucan (Medium viscosity)	Barley flour (Megazyme, B-BGBM)	Mixed-linked β-(1-3;1-4)-glucose Also contains 2% arabinose and 0.2% xylose.
10	Oat β-glucan (Medium viscosity)	Oat flour (Megazyme, P-BGOM)	Mixed-linked β-(1-3;1-4)-glucose
11*	Fucoidan	*Fucus vesiculosus (*Sigma F-5631)	Fucose-rich, sulfated polysaccharide
12	Pustulan	*Umbilicaria papullosa* (Calbiochem)	Linear β-(1-6)-glucose
13	Xylan (X1)	Plum (UA)	Sample rich in xylans: Rha (3%), Fuc (2%), Ara (11%), Xyl (67%), Man (0%), Gal (6%), Glc (5%), Ur Ac (6%)
14	Xylan (X2)	Plum (UA)	Sample rich in xylans: Rha (3%), Fuc (3%), Ara (12%), Xyl (73%), Man (0%), Gal (0%), Glc (4%), Ur Ac (5%)
15*	Mannoprotein (MP2)	*Saccharomyces cerevisiae* (UA)	Yeast cell wall; sequential extraction 8M KOH Ara (1%), Xyl (0%), Man (65%), Glc (35%)
16	Xyloglucan (XG1)	Plum (UA)	Sample rich in xyloglucans: Rha (2%), Fuc (5%), Ara (5%), Xyl (40%), Man (6%), Gal (13%), Glc (24%), Ur Ac (6%)

17	Mannan	*Saccharomyces cerevisiae* (Sigma, M7504)	α-(1-6)-mannose
18*	Mannoprotein (MP1)	*Saccharomyces cerevisiae* (UA)	Yeast cell wall; sequential extraction 4M KOH Ara (1%), Xyl (1%), Man (85%), Glc (13%)
19	Mannoprotein (MP2)	*Saccharomyces cerevisiae* (UA)	Yeast cell wall; sequential extraction 8M KOH Ara (1%), Xyl (0%), Man (65%), Glc (35%)
20*	Xyloglucan (XG1)	Plum (UA)	Sample rich in xyloglucans Rha (2%), Fuc (5%), Ara (5%), Xyl (40%), Man (6%), Gal (13%), Glc (24%), Ur Ac (6%)
21	Arabinoxylan (AX1)	Dreche (UA)	Sample rich in arabinoxylans; KOH 0.1 M (DP 41); Branching degree 0.57 Ara (40%), Xyl (54%), Man (0%), Gal (3%), Glc (3%)
22	Xyloglucan (XG2)	Plum (UA)	Sample rich in xyloglucans Rha (2%), Fuc (6%), Ara (6%), Xyl (46%), Man (4%), Gal (14%), Glc (22%), Ur Ac (1%)
23	Arabinoxylan (AX2)	Dreche (UA)	Sample rich in arabinoxylans (DP24); Branching degree 0.48 Ara (25%), Xyl (46%), Man (1%), Gal (3%), Glc (25%)
24	Arabinogalactan (AG1)	Coffee (UA)	Sample rich in arabinogalactans (High MW, M23pp) Ara (5%), Man (29%), Gal (64%), Glc (1%)
25	Galactomannan (GM1)	Coffee (UA)	Sample rich in galactomannans (High MW); Ara (5%), Man (62%), Gal (31%), Glc (2%)
26	Galactomannan (GM2)	Coffee (UA)	Sample rich in galactomannans (High MW); Ara (2%), Man (87%), Gal (10%), Glc (2%)
27	Pectin-1	Malian medicinal tree *Terminalia macroptera* (IC)	Under analysis
28	Pectin-2	Malian medicinal tree *Terminalia macroptera* (IC)	Under analysis
29	Pectin-3	Malian medicinal tree *Terminalia macroptera* (IC)	Under analysis
30	Pectin-4	Malian medicinal tree *Terminalia macroptera* (IC)	Under analysis
31	Pectin-5	Malian medicinal tree *Terminalia macroptera* (IC)	Under analysis
32	Pectin-6	Malian medicinal tree *Terminalia macroptera* (IC)	Under analysis
33	Pectin-7	Malian medicinal tree *Terminalia macroptera*	Under analysis

		(IC)	
34	Pectin-8	Malian medicinal tree *Terminalia macroptera* (IC)	Under analysis
35	Pectin-9	Malian medicinal tree *Terminalia macroptera* (IC)	Under analysis
36	Pectin-10	Malian medicinal tree *Terminalia macroptera* (IC)	Under analysis
37	Pectin-11	Malian medicinal tree *Terminalia macroptera* (IC)	Under analysis
38	Pectin12	Malian medicinal tree *Terminalia macroptera* (IC)	Under analysis
39	Pectin-13	Malian medicinal tree *Terminalia macroptera* (IC)	Under analysis
40	Xylan	*Palmaria palmata* (Elicityl, XYL100)	Mixed-linked β-(1-3;1-4)-D-xylose 1:4
41*	Glucurono-XyloMannan (125 mM NaCl)	*Tremella fuciformi* (Elicityl, HGL200)	α-(1-3)-mannose branched with xylose, glucuronic acid & fucose
42	Glucurono-XyloMannan (250 mM NaCl)	*Tremella fuciformi* (Elicityl, HGL200)	α-(1-3)-mannose branched with xylose, glucuronic acid & fucose
43	Galactomannan Guar	Guar (Megazyme, P-GGMM)	Galactose: Mannose ratio = 38:62
44	Galactomannan Carob (Low viscosity)	Carob (Megazyme, P-GALML)	Galactose: Mannose ratio = 21:79
45	Galactomannan Guar (Galactose depleted, high viscosity)	Guar (Megazyme, P-GGM21)	Galactose: Mannose ratio = 21:79
46	Galactomannan (Hydrolysed)	Coffee (UA)	Under analysis
47*	Poly-ManU	Algae (IC)	Under analysis
48	Lichenan	Icelandic moss (Megazyme, P-LICHN)	Mixed-linked β-(1-3;1-4)-glucose
49	Man4(β4)-DH	Megazyme (OMTE)	Manβ1-4Manβ1-4Manβ1-4Man-DH
50	Man6(β4)-DH	Megazyme (O-MHE)	Manβ1-4Manβ1-4Manβ1-4Manβ1-4Manβ1-4Man-DH
51	Ara6(α5)-DH	Megazyme, O-AHE	Araα1-5Araα1-5Araα1-5Araα1-5Araα1-5Ara-DH

52	Ara7(α5)-DH	Megazyme, O-AHP	Araα1-5Araα1-5Araα1-5Araα1-5Araα1-5Araα1-5Ara-DH
53	Xyl5(β4)-DH	Megazyme, O-XPE	Xylβ1-4Xylβ1-4Xylβ1-4Xylβ1-4Xyl-DH
54	Xyl6(β4)-DH	Megazyme, O-XHE	Xylβ1-4Xylβ1-4Xylβ1-4Xylβ1-4Xylβ1-4Xyl-DH
55	Xyl3Glc4-DH	Megazyme, O-XGHON	Xylα1-6 Xylα1-6 | | Glcβ1-4Glcβ1-4Glcβ1-4Glc-DH (designated as XXXG motif) | Xylα1-6
56	GN5-AO	Megazyme, O-CHI5	GlcNAcβ1-4GlcNAcβ1-4GlcNAcβ1-4GlcNAcβ1-4GlcNAc-AO
57	Xyl-glucan DP8-AO	Megazyme, O-XGHON	Major component DP8 (contains small % of DP7, XXXG motif) Galβ1-2 | Xylα1-6 Xylα1-6 | | Glcβ1-4Glcβ1-4Glcβ1-4Glc-AO (designated as XXLG motif) | Xylα1-6
58	Xyl-glucan DP9-AO	Megazyme, O-XGHON	Galβ-2 | Xylα1-6 Xylα1-6 | | Glcβ1-4Glcβ1-4Glcβ1-4Glc-AO (designated as XLLG motif) | Xylα1-6 | Galβ-2
59	Galactomannan DP7a-AO	Elicityl, Man219	DP7 (TLC-1) Mixture of β(1-4) linked D-mannose oligosaccharides. Average Man/Gal Ratio range ~4-7
60	Galactomannan DP6-AO	Elicityl, Man219	DP6 (TLC-2) Mixture of β(1-4) linked D-mannose oligosaccharides. Average Man/Gal Ratio range ~4-7
61	Galactomannan DP8-AO	Megazyme, O-GGM5	DP8 Prepared by controlled enzymatic hydrolysis of carob galactomannan Galβ-2 Galβ-2 Galβ-2 | | | Manβ1-4Manβ1-4Manβ1-4Manβ1-4Man-AO
62	Galactomannan DP7b -AO	Megazyme, O-GGM5	DP7 Prepared by controlled enzymatic hydrolysis of carob galactomannan

			Galβ-2 Galβ-2 | | Manβ1-4Manβ1-4Manβ1-4Manβ1-4Man-AO
63	Ara9(α5)-AO	Megazyme, O-AOC	Araα-5Araα-5Araα-5Araα-5Araα-5Araα-5Araα-5Araα-5Ara-AO
64	Ara8(α5)-AO	Megazyme, O-AOC	Araα-5Araα-5Araα-5Araα-5Araα-5Araα-5Araα-5Ara-AO

[a] Position matching the original microarray set.

[b] The sources are indicated for each carbohydrate. IC- Collection available at the carbohydrate microarray Facility at Imperial College London; UA – Collaboration with Prof. Manuel Coimbra, University of Aveiro. The collection of pectin samples were received from Prof. Berit and were purified from the Malian medicinal tree *Terminalia macroptera* (Zou *et al.* 2015).The sequence for these pectin samples is under analysis. The oligosaccharides were prepared as lipid-linked probes, neoglycolipids (NGLs); DH, NGLs prepared from reducing oligosaccharides by reductive amination with the amino lipid, 1,2-dihexadecyl-sn-glycero-3-phosphoethanolamine (DHPE) (Liu *et al.* 2012); AO, NGLs prepared from reducing oligosaccharides by oxime ligation with an aminooxy (AO) functionalized DHPE (Liu *et al.* 2007).

*Probes excluded after a quality control analysis.

Supplementary information 1:

Solutions for DH5α competent cells preparation

SOB (Super optimal broth) medium

2% peptone

0.5% Yeast extract

10 mM NaCl

2.5 mM KCL

Adjust the pH to 7.5 and autoclave

Then add 10 mM $MgCl_2$ and 10 mM $MgSO_4$.

TB (Transformation buffer)

250 M KCL

15 mM $CaCl_2.2H_2O$

10 mM PIPES

Dissolve the components and adjust the pH to 6.7. Then add:

55 mM $MnCl_2$

Filter with a 0.45 µm sterile filter and store at 4ºC.

Solutions for BL21(DE3) competent cells preparation

LB (Luria-Bertani) medium 200ml:

2 g Tryptone (Difco)

1 g Yeast Extract (Difco)

1 g NaCl

Adjust with com 1N NaOH.

TFBI (Transformation buffer I)

30 mM KOAc

100 mM RbCl

10 mM CaCl2 + 50 mM MnCl215% glycerol

Adjust to pH 5.8 with acetic acid and filter (0.45 µm) to sterilize.

TFBII (Transformation buffer II)

10 mM MOPS

75 mM $CaCl_2$

10 mM RbCl

15% glycerol

Adjust pH to 6.5 with KOH and filter to sterilize.

Supplementary information 2:

Luria Bertani and AutoInduction medium cultures used for the two protocols of expression

Luria Bertani medium culture (1L)

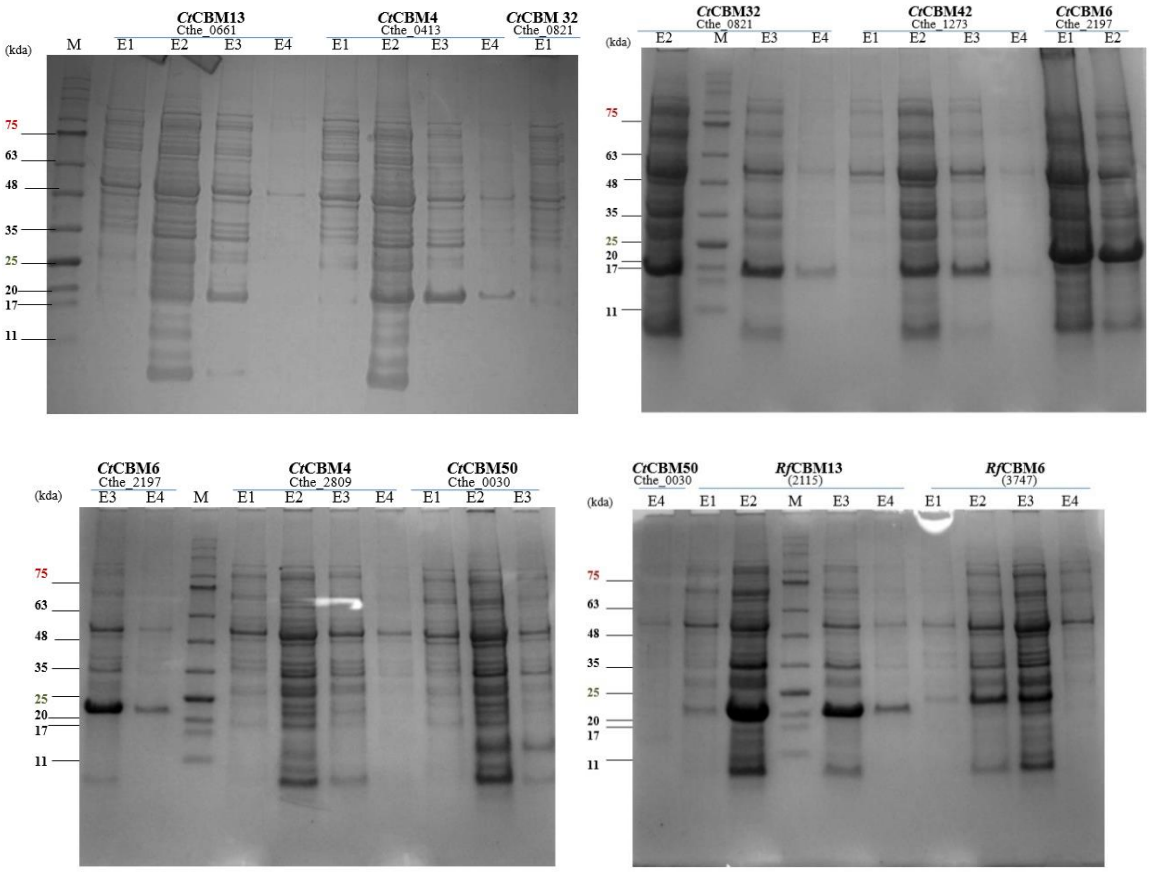
- 10g Tryptone (Sigma-Aldrich®);
- 10g NaCl (Panreac®);
- 5g Yeast (NZYeTch®);
- Required volume of distilled water.

Autoclave for 20 minutes at 120ºC.

NZYTech® Autoinduction medium (1L)

- 50g of NZY AutoInduction LB medium (powder);
- 10 ml glycerol;
- Required volume of distilled water.

Autoclave for 15 minutes at 121ºC.



Supplementary Figure 1 - SDS-PAGE (10% acrylamide) showing IMAC purification of CBMs (expressed with Autoinduction protocol) from different families with His Gravitrap® columns. The CBMs are identified by family and protein ID. Respective protein lanes are identified by a blue line**.** CBMs were eluted with 6 ml of 50 mM HEPES at a pH of 7.5, 1M NaCl, 2 mM $CaCl_2$ and 500 mM imidazole. The fractions were eluted and analyzed in separated: one fraction with 3 ml and three fractions with 1 ml each. E1(Eluted fration 1) - 3ml; E2(Eluted fraction 2) - 1 ml; E3 (Eluted fraction 3) – 1ml; E3 (Eluted fraction 3) – 1 ml.

	1	2	3	4	5	6	7	8	9	10	11	12
A	Sodium acetate pH 4.5	Sodium citrate pH 4.7	Sodium acetate pH 5.0	Potassium phosphate pH 5.0	Sodium phosphate pH 5.5	Sodium citrate pH 5.5	MES pH 5.8	Potassium phosphate pH 6.0	MES pH 6.2	Sodium phosphate pH 6.5	Sodium cacodylate pH 6.5	MES pH 6.5
B	Potassium phosphate pH 7.0	HEPES pH 7.0	Ammonium acetate pH 7.3	Sodium phosphate pH 7.5	Tris pH 7.5	Imidazole pH 8.0	Hepes pH 8.0	Tris pH 8.0	Bicine pH 8.0	Tris pH 8.5	Bicine pH 9.0	Control (without protein)
C	50mM HEPES, 1M NaCl, 2 mM $CaCl_2$ pH=7.5	Sodium citrate pH 4.7 + 100 mM NaCl	Sodium acetate pH 5.0 + 100 mM NaCl	Potassium phosphate pH 5.0 + 100 mM NaCl	H_2O + 2mM $CaCl_2$	Sodium citrate pH 5.5 + 100 mM NaCl	MES pH 5.8 + 100 mM NaCl	Potassium phosphate pH 6.0 + 100 mM NaCl	MES pH 6.2 + 100 mM NaCl	Sodium phosphate pH 6.5 + 100 mM NaCl	Sodium cacodylate pH 6.5 + 100 mM NaCl	MES pH 6.5 + 100 mM NaCl
D	Potassium phosphate pH 7.0 + 100 mM NaCl	HEPES pH 7.0 + 100 mM NaCl	Ammonium acetate pH 7.3 + 100 mM NaCl	Sodium phosphate pH 7.5 + 100 mM NaCl	Tris pH 7.5 + 100 mM NaCl	Imidazole pH 8.0 + 100 mM NaCl	Hepes pH 8.0 + 100 mM NaCl	Tris pH 8.0 + 100 mM NaCl	Bicine pH 8.0 + 100 mM NaCl	Tris pH 8.5 + 100 mM NaCl	Bicine pH 9.0 + 100 mM NaCl	Control (without dye)
E	Sodium acetate pH 4.5 **+ 250 mM NaCl**	Sodium citrate pH 4.7 + 250 mM NaCl	Sodium acetate pH 5.0 + 250 mM NaCl	Potassium phosphate pH 5.0 + 250 mM NaCl	Sodium phosphate pH 5.5 + 250 mM NaCl	Sodium citrate pH 5.5 + 250 mM NaCl	MES pH 5.8 + 250 mM NaCl	Potassium phosphate pH 6.0 + 250 mM NaCl	MES pH 6.2 + 250 mM NaCl	Sodium phosphate pH 6.5 + 250 mM NaCl	Sodium cacodylate pH 6.5 + 250 mM NaCl	MES pH 6.5 + 250 mM NaCl
F	Potassium phosphate pH 7.0 + 250 mM NaCl	HEPES pH 7.0 + 250 mM NaCl	Ammonium acetate pH 7.3 + 250 mM NaCl	Sodium phosphate pH 7.5 + 250 mM NaCl	Tris pH 7.5 + 250 mM NaCl	Imidazole pH 8.0 + 250 mM NaCl	Hepes pH 8.0 + 250 mM NaCl	Tris pH 8.0 + 250 mM NaCl	Bicine pH 8.0 + 250 mM NaCl	Tris pH 8.5 + 250 mM NaCl	Bicine pH 9.0 + 250 mM NaCl	50mM HEPES, 1M NaCl, 2 mM $CaCl_2$ pH=7.5
G	Sodium acetate pH 4.5 **+ 500 mM NaCl**	Sodium citrate pH 4.7 + 500 mM NaCl	Sodium acetate pH 5.0 + 500 mM NaCl	Potassium phosphate pH 5.0 + 500 mM NaCl	Sodium phosphate pH 5.5 + 500 mM NaCl	Sodium citrate pH 5.5 + 500 mM NaCl	MES pH 5.8 + 500 mM NaCl	Potassium phosphate pH 6.0 + 500 mM NaCl	MES pH 6.2 + 500 mM NaCl	Sodium phosphate pH 6.5 + 500 mM NaCl	Sodium cacodylate pH 6.5 + 500 mM NaCl	MES pH 6.5 + 500 mM NaCl
H	Potassium phosphate pH 7.0 + 500 mM NaCl	HEPES pH 7.0 + 500 mM NaCl	Ammonium acetate pH 7.3 + 500 mM NaCl	Sodium phosphate pH 7.5 + 500 mM NaCl	Tris pH 7.5 + 500 mM NaCl	Imidazole pH 8.0 + 500 mM NaCl	Hepes pH 8.0 + 500 mM NaCl	Tris pH 8.0 + 500 mM NaCl	Bicine pH 8.0 + 500 mM NaCl	Tris pH 8.5 + 500 mM NaCl	Bicine pH 9.0 + 500 mM NaCl	CONTROLO ÁGUA miliQ

Supplementary Figure 2 – Thermofluor screen (in house). All the buffers were made at 100 mM.

Tube #	Salt
1.	0.2 M Sodium fluoride
2.	0.2 M Potassium fluoride
3.	0.2 M Ammonium fluoride
4.	0.2 M Lithium chloride
5.	0.2 M Magnesium chloride hexahydrate
6.	0.2 M Sodium chloride
7.	0.2 M Calcium chloride dihydrate
8.	0.2 M Potassium chloride
9.	0.2 M Ammonium chloride
10.	0.2 M Sodium iodide
11.	0.2 M Potassium iodide
12.	0.2 M Ammonium iodide
13.	0.2 M Sodium thiocyanate
14.	0.2 M Potassium thiocyanate
15.	0.2 M Lithium nitrate
16.	0.2 M Magnesium nitrate hexahydrate
17.	0.2 M Sodium nitrate
18.	0.2 M Potassium nitrate
19.	0.2 M Ammonium nitrate
20.	0.2 M Magnesium formate dihydrate
21.	0.2 M Sodium formate
22.	0.2 M Potassium formate
23.	0.2 M Ammonium formate
24.	0.2 M Lithium acetate dihydrate
25.	0.2 M Magnesium acetate tetrahydrate
26.	0.2 M Zinc acetate dihydrate
27.	0.2 M Sodium acetate trihydrate
28.	0.2 M Calcium acetate hydrate
29.	0.2 M Potassium acetate
30.	0.2 M Ammonium acetate
31.	0.2 M Lithium sulfate monohydrate
32.	0.2 M Magnesium sulfate heptahydrate
33.	0.2 M Sodium sulfate decahydrate
34.	0.2 M Potassium sulfate
35.	0.2 M Ammonium sulfate
36.	0.2 M Sodium tartrate dibasic dihydrate
37.	0.2 M Potassium sodium tartrate tetrahydrate
38.	0.2 M Ammonium tartrate dibasic
39.	0.2 M Sodium phosphate monobasic monohydrate
40.	0.2 M Sodium phosphate dibasic dihydrate
41.	0.2 M Potassium phosphate monobasic
42.	0.2 M Potassium phosphate dibasic
43.	0.2 M Ammonium phosphate monobasic
44.	0.2 M Ammonium phosphate dibasic
45.	0.2 M Lithium citrate tribasic tetrahydrate
46.	0.2 M Sodium citrate tribasic dihydrate
47.	0.2 M Potassium citrate tribasic monohydrate
48.	0.2 M Ammonium citrate dibasic

Tube #	Polymer	Tube #	pH ◊
1.	20% w/v Polyethylene glycol 3,350	1.	7.3
2.	20% w/v Polyethylene glycol 3,350	2.	7.3
3.	20% w/v Polyethylene glycol 3,350	3.	6.2
4.	20% w/v Polyethylene glycol 3,350	4.	6.8
5.	20% w/v Polyethylene glycol 3,350	5.	5.9
6.	20% w/v Polyethylene glycol 3,350	6.	6.9
7.	20% w/v Polyethylene glycol 3,350	7.	5.1
8.	20% w/v Polyethylene glycol 3,350	8.	7.0
9.	20% w/v Polyethylene glycol 3,350	9.	6.3
10.	20% w/v Polyethylene glycol 3,350	10.	7.0
11.	20% w/v Polyethylene glycol 3,350	11.	7.0
12.	20% w/v Polyethylene glycol 3,350	12.	6.2
13.	20% w/v Polyethylene glycol 3,350	13.	6.9
14.	20% w/v Polyethylene glycol 3,350	14.	7.0
15.	20% w/v Polyethylene glycol 3,350	15.	7.1
16.	20% w/v Polyethylene glycol 3,350	16.	5.9
17.	20% w/v Polyethylene glycol 3,350	17.	6.8
18.	20% w/v Polyethylene glycol 3,350	18.	6.8
19.	20% w/v Polyethylene glycol 3,350	19.	6.2
20.	20% w/v Polyethylene glycol 3,350	20.	7.0
21.	20% w/v Polyethylene glycol 3,350	21.	7.2
22.	20% w/v Polyethylene glycol 3,350	22.	7.3
23.	20% w/v Polyethylene glycol 3,350	23.	6.6
24.	20% w/v Polyethylene glycol 3,350	24.	7.9
25.	20% w/v Polyethylene glycol 3,350	25.	7.9
26.	20% w/v Polyethylene glycol 3,350	26.	6.4
27.	20% w/v Polyethylene glycol 3,350	27.	8.0
28.	20% w/v Polyethylene glycol 3,350	28.	7.5
29.	20% w/v Polyethylene glycol 3,350	29.	8.1
30.	20% w/v Polyethylene glycol 3,350	30.	7.1
31.	20% w/v Polyethylene glycol 3,350	31.	6.0
32.	20% w/v Polyethylene glycol 3,350	32.	6.0
33.	20% w/v Polyethylene glycol 3,350	33.	6.7
34.	20% w/v Polyethylene glycol 3,350	34.	6.8
35.	20% w/v Polyethylene glycol 3,350	35.	6.0
36.	20% w/v Polyethylene glycol 3,350	36.	7.3
37.	20% w/v Polyethylene glycol 3,350	37.	7.4
38.	20% w/v Polyethylene glycol 3,350	38.	6.6
39.	20% w/v Polyethylene glycol 3,350	39.	4.7
40.	20% w/v Polyethylene glycol 3,350	40.	9.1
41.	20% w/v Polyethylene glycol 3,350	41.	4.8
42.	20% w/v Polyethylene glycol 3,350	42.	9.2
43.	20% w/v Polyethylene glycol 3,350	43.	4.6
44.	20% w/v Polyethylene glycol 3,350	44.	8.0
45.	20% w/v Polyethylene glycol 3,350	45.	8.4
46.	20% w/v Polyethylene glycol 3,350	46.	8.3
47.	20% w/v Polyethylene glycol 3,350	47.	8.3
48.	20% w/v Polyethylene glycol 3,350	48.	5.1

Tube #	Salt
1.	0.1 M Sodium malonate pH 4.0
2.	0.2 M Sodium malonate pH 4.0
3.	0.1 M Sodium malonate pH 5.0
4.	0.2 M Sodium malonate pH 5.0
5.	0.1 M Sodium malonate pH 6.0
6.	0.2 M Sodium malonate pH 6.0
7.	0.1 M Sodium malonate pH 7.0
8.	0.2 M Sodium malonate pH 7.0
9.	4% v/v Tacsimate™ pH 4.0
10.	8% v/v Tacsimate™ pH 4.0
11.	4% v/v Tacsimate™ pH 5.0
12.	8% v/v Tacsimate™ pH 5.0
13.	4% v/v Tacsimate™ pH 6.0
14.	8% v/v Tacsimate™ pH 6.0
15.	4% v/v Tacsimate™ pH 7.0
16.	8% v/v Tacsimate™ pH 7.0
17.	4% v/v Tacsimate™ pH 8.0
18.	8% v/v Tacsimate™ pH 8.0
19.	0.1 M Succinic acid pH 7.0
20.	0.2 M Succinic acid pH 7.0
21.	0.1 M Ammonium citrate tribasic pH 7.0
22.	0.2 M Ammonium citrate tribasic pH 7.0
23.	0.1 M DL-Malic acid pH 7.0
24.	0.2 M DL-Malic acid pH 7.0
25.	0.1 M Sodium acetate trihydrate pH 7.0
26.	0.2 M Sodium acetate trihydrate pH 7.0
27.	0.1 M Sodium formate pH 7.0
28.	0.2 M Sodium formate pH 7.0
29.	0.1 M Ammonium tartrate dibasic pH 7.0
30.	0.2 M Ammonium tartrate dibasic pH 7.0
31.	2% v/v Tacsimate™ pH 4.0
32.	2% v/v Tacsimate™ pH 5.0
33.	2% v/v Tacsimate™ pH 6.0
34.	2% v/v Tacsimate™ pH 7.0
35.	2% v/v Tacsimate™ pH 8.0
36.	None
37.	None
38.	None
39.	None
40.	None
41.	None
42.	0.02 M Calcium chloride dihydrate, 0.02 M Cadmium chloride hydrate, 0.02 M Cobalt(II) chloride hexahydrate
43.	0.01 M Magnesium chloride hexahydrate 0.005 M Nickel(II) chloride hexahydrate
44.	0.02 M Zinc chloride
45.	0.15 M Cesium chloride
46.	0.2 M Sodium bromide
47.	1% w/v Tryptone, 0.001 M Sodium azide
48.	1% w/v Tryptone, 0.001 M Sodium azide

Tube #	Buffer ◊
1.	None
2.	None
3.	None
4.	None
5.	None
6.	None
7.	None
8.	None
9.	None
10.	None
11.	None
12.	None
13.	None
14.	None
15.	None
16.	None
17.	None
18.	None
19.	None
20.	None
21.	None
22.	None
23.	None
24.	None
25.	None
26.	None
27.	None
28.	None
29.	None
30.	None
31.	0.1 M Sodium acetate trihydrate pH 4.6
32.	0.1 M Sodium citrate tribasic dihydrate pH 5.6
33.	0.1 M BIS-TRIS pH 6.5
34.	0.1 M HEPES pH 7.5
35.	0.1 M Tris pH 8.5
36.	0.07 M Citric acid, 0.03 M BIS-TRIS propane / pH 3.4
37.	0.06 M Citric acid, 0.04 M BIS-TRIS propane / pH 4.1
38.	0.05 M Citric acid, 0.05 M BIS-TRIS propane / pH 5.0
39.	0.04 M Citric acid, 0.06 M BIS-TRIS propane / pH 6.4
40.	0.03 M Citric acid, 0.07 M BIS-TRIS propane / pH 7.6
41.	0.02 M Citric acid, 0.08 M BIS-TRIS propane / pH 8.8
42.	None
43.	0.1 M HEPES sodium pH 7.0
44.	None
45.	None
46.	None
47.	0.05 M HEPES sodium pH 7.0
48.	0.05 M HEPES sodium pH 7.0

Tube #	Polymer
1.	12% w/v Polyethylene glycol 3,350
2.	20% w/v Polyethylene glycol 3,350
3.	12% w/v Polyethylene glycol 3,350
4.	20% w/v Polyethylene glycol 3,350
5.	12% w/v Polyethylene glycol 3,350
6.	20% w/v Polyethylene glycol 3,350
7.	12% w/v Polyethylene glycol 3,350
8.	20% w/v Polyethylene glycol 3,350
9.	12% w/v Polyethylene glycol 3,350
10.	20% w/v Polyethylene glycol 3,350
11.	12% w/v Polyethylene glycol 3,350
12.	20% w/v Polyethylene glycol 3,350
13.	12% w/v Polyethylene glycol 3,350
14.	20% w/v Polyethylene glycol 3,350
15.	12% w/v Polyethylene glycol 3,350
16.	20% w/v Polyethylene glycol 3,350
17.	12% w/v Polyethylene glycol 3,350
18.	20% w/v Polyethylene glycol 3,350
19.	12% w/v Polyethylene glycol 3,350
20.	20% w/v Polyethylene glycol 3,350
21.	12% w/v Polyethylene glycol 3,350
22.	20% w/v Polyethylene glycol 3,350
23.	12% w/v Polyethylene glycol 3,350
24.	20% w/v Polyethylene glycol 3,350
25.	12% w/v Polyethylene glycol 3,350
26.	20% w/v Polyethylene glycol 3,350
27.	12% w/v Polyethylene glycol 3,350
28.	20% w/v Polyethylene glycol 3,350
29.	12% w/v Polyethylene glycol 3,350
30.	20% w/v Polyethylene glycol 3,350
31.	16% w/v Polyethylene glycol 3,350
32.	16% w/v Polyethylene glycol 3,350
33.	20% w/v Polyethylene glycol 3,350
34.	20% w/v Polyethylene glycol 3,350
35.	16% w/v Polyethylene glycol 3,350
36.	16% w/v Polyethylene glycol 3,350
37.	16% w/v Polyethylene glycol 3,350
38.	16% w/v Polyethylene glycol 3,350
39.	20% w/v Polyethylene glycol 3,350
40.	20% w/v Polyethylene glycol 3,350
41.	16% w/v Polyethylene glycol 3,350
42.	20% w/v Polyethylene glycol 3,350
43.	15% w/v Polyethylene glycol 3,350
44.	20% w/v Polyethylene glycol 3,350
45.	15% w/v Polyethylene glycol 3,350
46.	20% w/v Polyethylene glycol 3,350
47.	12% w/v Polyethylene glycol 3,350
48.	20% w/v Polyethylene glycol 3,350

Supplementary Figure 3 – PEG Ion and PEG Ion2 commercial screens from Hampton Researh used in preliminary crystallization assays. A) PEG Ion screen. B) PEG Ion2 screen. Figure taken from www.hamptonresearch.com.

JCSG-*plus* HT-96 Wells A1- D12 MD1–40

Well #	Conc.	Salt	Conc.	Buffer	pH	Conc.	Precipitant
A1	0.2 M	Lithium sulfate	0.1 M	Sodium acetate	4.5	50 % w/v	PEG 400
A2		None	0.1 M	Sodium citrate	5.5	20 % w/v	PEG 3000
A3	0.2 M	Ammonium citrate dibasic		None		20 % w/v	PEG 3350
A4	0.02 M	Calcium chloride dihydrate	0.1 M	Sodium acetate	4.6	30 % v/v	MPD
A5	0.2 M	Magnesium formate dihydrate		None		20 % w/v	PEG 3350
A6	0.2 M	Lithium sulfate	0.1 M	Phosphate/citrate	4.2	20 % w/v	PEG 1000
A7		None	0.1 M	CHES	9.5	20 % w/v	PEG 8000
A8	0.2 M	Ammonium formate		None		20 % w/v	PEG 3350
A9	0.2 M	Ammonium chloride		None		20 % w/v	PEG 3350
A10	0.2 M	Potassium formate		None		20 % w/v	PEG 3350
A11	0.2 M	Ammonium phosphate monobasic	0.1 M	Tris	8.5	50 % v/v	MPD
A12	0.2 M	Potassium nitrate		None		20 % w/v	PEG 3350
B1	0.8 M	Ammonium sulfate	0.1 M	Citrate	4.0		None
B2	0.2 M	Sodium thiocyanate		None		20 % w/v	PEG 3350
B3		None	0.1 M	BICINE	9.0	20 % w/v	PEG 6000
B4		None	0.1 M	HEPES	7.5	10 % w/v	PEG 8000
						8 % v/v	Ethylene glycol
B5		None	0.1 M	Sodium cacodylate	6.5	40 % v/v	MPD
						5 % w/v	PEG 8000
B6		None	0.1 M	Phosphate/citrate	4.2	40 % v/v	Ethanol
						5 % w/v	PEG 1000
B7		None	0.1 M	Sodium acetate	4.6	8 % w/v	PEG 4000
B8	0.2 M	Magnesium chloride hexahydrate	0.1 M	Tris	7.0	10 % w/v	PEG 8000
B9		None	0.1 M	Citrate	5.0	20 % w/v	PEG 6000
B10	0.2 M	Magnesium chloride hexahydrate	0.1 M	Sodium cacodylate	6.5	50 % v/v	PEG 200
B11	1.6 M	Sodium citrate tribasic dihydrate pH 6.5		None			None
B12	0.2 M	Potassium citrate tribasic monohydrate		None		20 % w/v	PEG 3350
C1	0.2 M	Sodium chloride	0.1 M	Phosphate/citrate	4.2	20 % w/v	PEG 8000
C2	1.0 M	Lithium chloride	0.1 M	Citrate	4.0	20 % w/v	PEG 6000
C3	0.2 M	Ammonium nitrate		None		20 % w/v	PEG 3350
C4		None	0.1 M	HEPES	7.0	10 % w/v	PEG 6000
C5	0.8 M	Sodium phosphate monobasic monohydrate	0.1 M	Sodium HEPES	7.5		None
	0.8 M	Potassium phosphate monobasic		None			
C6		None	0.1 M	Phosphate/citrate	4.2	40 % v/v	PEG 300
C7	0.2 M	Zinc acetate dihydrate	0.1 M	Sodium acetate	4.5	10 % w/v	PEG 3000
C8		None	0.1 M	Tris	8.5	20 % v/v	Ethanol
C9		None	0.1 M	Sodium/potassium phosphate	6.2	25 % v/v	1,2-Propandiol
						10 %v/v	Glycerol
C10		None	0.1 M	BICINE	9.0	10 % w/v	PEG 20,000
						2 % v/v	1,4-Dioxane
C11	2.0 M	Ammonium sulfate	0.1 M	Sodium acetate	4.6		None
C12		None		None		10 % w/v	PEG 1000
						10 % w/v	PEG 8000
D1		None		None		24 % w/v	PEG 1500
						20 % v/v	Glycerol
D2	0.2 M	Magnesium chloride hexahydrate	0.1 M	Sodium HEPES	7.5	30 % v/v	PEG 400
D3	0.2 M	Sodium chloride	0.1 M	Sodium/potassium phosphate	6.2	50 % v/v	PEG 200
D4	0.2 M	Lithium sulfate	0.1 M	Sodium acetate	4.5	30 % w/v	PEG 8000
D5		None	0.1 M	HEPES	7.5	70 % v/v	MPD
D6	0.2 M	Magnesium chloride hexahydrate	0.1 M	Tris	8.5	20 % w/v	PEG 8000
D7	0.2 M	Lithium sulfate	0.1 M	Tris	8.5	40 % v/v	PEG 400
D8		None	0.1 M	Tris	8.0	40 % v/v	MPD
D9	0.17 M	Ammonium sulfate		None		25.5 % w/v	PEG 4000
						15 % v/v	Glycerol
D10	0.2 M	Calcium acetate hydrate	0.1 M	Sodium cacodylate	6.5	40 % v/v	PEG 300
D11	0.14 M	Calcium chloride dihydrate	0.07 M	Sodium acetate	4.6	14 % v/v	2-Propanol
						30 % v/v	Glycerol
D12	0.04 M	Potassium phosphate monobasic		None		16 % w/v	PEG 8000
						20 % v/v	Glycerol

JCSG-*plus* HT-96 Wells E1 – H12 MD1–40

Well #	Conc.	Salt	Conc.	Buffer	pH	Conc.	Precipitant
E1	1.0 M	Sodium citrate tribasic dihydrate	0.1 M	Sodium cacodylate	6.5		None
E2	2.0 M	Ammonium sulfate	0.1 M	Sodium cacodylate	6.5		None
	0.2 M	Sodium chloride					
E3	0.2 M	Sodium chloride	0.1 M	HEPES	7.5	10 % v/v	2-Propanol
E4	1.26 M	Ammonium sulfate	0.1 M	Tris	8.5		None
	0.2 M	Lithium sulfate					
E5		None	0.1 M	CAPS	10.5	40 % v/v	MPD
E6	0.2 M	Zinc acetate dihydrate	0.1 M	Imidazole	8.0	20 % w/v	PEG 3000
E7	0.2 M	Zinc acetate dihydrate	0.1 M	Sodium cacodylate	6.5	10 % v/v	2-Propanol
E8	1.0 M	Ammonium phosphate dibasic	0.1 M	Sodium acetate	4.5		None
E9	1.6 M	Mageisum sulfate heptahydrate	0.1 M	MES	6.5		None
E10		None	0.1 M	BICINE	9.0	10 % w/v	PEG 6000
E11	0.16 M	Calcium acetate hydrate	0.08 M	Sodium cacodylate	6.5	14.4 % w/v	PEG 8000
						20 % v/v	Glycerol
E12		None	0.1 M	Imidazole	8.0	10 % w/v	PEG 8000
F1	0.05 M	Cesium chloride	0.1 M	MES	6.5	30 % v/v	Jeffamine® M-600
F2	3.2 M	Ammonium sulfate	0.1 M	Citrate	5.0		None
F3		None	0.1 M	Tris	8.0	20 % v/v	MPD
F4		None	0.1 M	HEPES	7.5	20 % v/v	Jeffamine® M-600
F5	0.2 M	Magnesium chloride hexahydrate	0.1 M	Tris	8.5	50 % v/v	Ethylene glycol
F6		None	0.1 M	BICINE	9.0	10 % v/v	MPD
F7	0.8 M	Succinic acid pH 7.0		None			None
F8	2.1 M	DL-Malic acid pH 7.0		None			None
F9	2.4 M	Sodium malonate dibasic monohydrate pH 7.0		None			None
F10	1.1 M	Sodium malonate dibasic monohydrate	0.1 M	HEPES	7.0	0.5 % v/v	Jeffamine® ED-2003
F11	1.0 M	Succinic acid	0.1 M	HEPES	7.0	1 % w/v	PEG 2000 MME
F12		None	0.1 M	HEPES	7.0	30 % v/v	Jeffamine® M-600
G1		None	0.1 M	HEPES	7.0	30 % v/v	Jeffamine® ED-2003
G2	0.02 M	Magnesium chloride hexahydrate	0.1 M	HEPES	7.5	22 % w/v	Poly(acrylic acid sodium salt) 5100
G3	0.01 M	Cobalt(II) chloride hexahydrate	0.1 M	Tris	8.5	20 % w/v	Polyvinylpyrrolidone
G4	0.2 M	TMAO	0.1 M	Tris	8.5	20 % w/v	PEG 2000 MME
G5	0.005 M	Cobalt(II) chloride hexahydrate	0.1 M	HEPES	7.5	12 % w/v	PEG 3350
	0.005 M	Cadmium chloride hemi(pentahydrate)					
	0.005 M	Magnesium chloride hexahydrate					
	0.005 M	Nickel(II) chloride hexahydrate					
G6	0.2 M	Sodium malonate dibasic monohydrate		None		20 % w/v	PEG 3350
G7	0.1 M	Succinic acid		None		15 % w/v	PEG 3350
G8	0.15 M	DL-Malic acid		None		20 % w/v	PEG 3350
G9	0.1 M	Potassium thiocyanate		None		30 % w/v	PEG 2000 MME
G10	0.15 M	Potassium bromide		None		30 % w/v	PEG 2000 MME
G11	2.0 M	Ammonium sulfate	0.1 M	BIS-Tris	5.5		None
G12	3.0 M	Sodium chloride	0.1 M	BIS-Tris	5.5		None
H1	0.3 M	Magnesium formate dihydrate	0.1 M	BIS-Tris	5.5		None
H2	1.0 M	Ammonium sulfate	0.1 M	BIS-Tris	5.5	1 % w/v	PEG 3350
H3		None	0.1 M	BIS-Tris	5.5	25 % w/v	PEG 3350
H4	0.2 M	Calcium chloride dihydrate	0.1 M	BIS-Tris	5.5	45 % v/v	MPD
H5	0.2 M	Ammonium acetate	0.1 M	BIS-Tris	5.5	45 % v/v	MPD
H6	0.1 M	Ammonium acetate	0.1 M	BIS-Tris	5.5	17 % w/v	PEG 10,000
H7	0.2 M	Ammonium sulfate	0.1 M	BIS-Tris	5.5	25 % w/v	PEG 3350
H8	0.2 M	Sodium chloride	0.1 M	BIS-Tris	5.5	25 % w/v	PEG 3350
H9	0.2 M	Lithium sulfate	0.1 M	BIS-Tris	5.5	25 % w/v	PEG 3350
H10	0.2 M	Ammonium acetate	0.1 M	BIS-Tris	5.5	25 % w/v	PEG 3350
H11	0.2 M	Magnesium chloride hexahydrate	0.1 M	BIS-Tris	5.5	25 % w/v	PEG 3350
H12	0.2 M	Ammonium acetate	0.1 M	HEPES	7.5	45 % v/v	MPD

Supplementary Figure 4– JSCG-*plus* HT96 commercial screen used for preliminary crystallization assays. Taken from www.moleculardimensions.com.

JBS 1

No.	Precipitant	Buffer	Additive
A1	15 % w/v Polyethylene glycol 400	100 mM Sodium acetate; pH 4.6	100 mM Calcium chloride
A2	15 % w/v Polyethylene glycol 400	100 mM MES; pH 6.5	none
A3	15 % w/v Polyethylene glycol 400	100 mM HEPES; pH 7.5	200 mM Magnesium chloride
A4	15 % w/v Polyethylene glycol 400	100 mM TRIS; pH 8.5	200 mM tri-Sodium citrate
A5	25 % w/v Polyethylene glycol 400	100 mM Sodium acetate; pH 4.6	100 mM Magnesium chloride
A6	25 % w/v Polyethylene glycol 400	100 mM TRIS; pH 8.5	200 mM Lithium sulfate
B1	28 % w/v Polyethylene glycol 400	100 mM HEPES; pH 7.5	200 mM Calcium chloride
B2	30 % w/v Polyethylene glycol 400	100 mM Sodium acetate; pH 4.6	100 mM Calcium chloride
B3	30 % w/v Polyethylene glycol 400	100 mM MES; pH 6.5	100 mM Sodium acetate
B4	30 % w/v Polyethylene glycol 400	100 mM MES; pH 6.5	100 mM Magnesium chloride
B5	30 % w/v Polyethylene glycol 400	100 mM HEPES; pH 7.5	200 mM Magnesium chloride
B6	30 % w/v Polyethylene glycol 400	100 mM TRIS; pH 8.5	200 mM tri-Sodium citrate
C1	30 % w/v Polyethylene glycol monomethyl ether 550	100 mM BICINE; pH 9.0	100 mM Sodium chloride
C2	25 % w/v Polyethylene glycol monomethyl ether 550	100 mM MES; pH 6.5	10 mM Zinc sulfate
C3	25 % w/v Polyethylene glycol 1,000	100 mM HEPES; pH 7.5	none
C4	30 % w/v Polyethylene glycol 1,000	100 mM TRIS; pH 8.5	none
C5	15 % w/v Polyethylene glycol 1,500	none	none
C6	20 % w/v Polyethylene glycol 1,500	100 mM HEPES; pH 7.5	none
D1	30 % w/v Polyethylene glycol 1,500	none	none
D2	20 % w/v Polyethylene glycol monomethyl ether 2,000	100 mM TRIS; pH 8.5	10 mM Nickel (II) chloride
D3	25 % w/v Polyethylene glycol monomethyl ether 2,000	none	none
D4	30 % w/v Polyethylene glycol monomethyl ether 2,000	100 mM MES; pH 6.5	100 mM Sodium acetate
D5	20 % w/v Polyethylene glycol 3,000	100 mM HEPES; pH 7.5	200 mM Sodium acetate
D6	30 % w/v Polyethylene glycol 3,000	100 mM TRIS; pH 8.5	200 mM Lithium sulfate

JBS 2

No.	Precipitant	Buffer	Additive
A1	4 % w/v Polyethylene glycol 4,000	100 mM Sodium acetate; pH 4.6	none
A2	8 % w/v Polyethylene glycol 4,000	none	none
A3	8 % w/v Polyethylene glycol 4,000	100 mM Sodium acetate; pH 4.6	none
A4	10 % w/v Polyethylene glycol 4,000	100 mM MES; pH 6.5	200 mM Magnesium chloride
A5	12 % w/v Polyethylene glycol 4,000	100 mM HEPES; pH 7.5	100 mM Sodium acetate
A6	12 % w/v Polyethylene glycol 4,000	100 mM TRIS; pH 8.5	none
B1	16 % w/v Polyethylene glycol 4,000	100 mM TRIS; pH 8.5	200 mM Lithium sulfate
B2	16 % w/v Polyethylene glycol 4,000	100 mM TRIS; pH 8.5	200 mM Sodium acetate
B3	16 % w/v Polyethylene glycol 4,000	100 mM TRIS; pH 8.5	200 mM Magnesium chloride
B4	18 % w/v Polyethylene glycol 4,000	100 mM Sodium acetate; pH 4.6	none
B5	20 % w/v Polyethylene glycol 4,000	100 mM TRIS; pH 8.5	200 mM Lithium sulfate
B6	20 % w/v Polyethylene glycol 4,000	100 mM TRIS; pH 8.5	200 mM Calcium chloride
C1	22 % w/v Polyethylene glycol 4,000	100 mM HEPES; pH 7.5	100 mM Sodium acetate
C2	25 % w/v Polyethylene glycol 4,000	100 mM Sodium acetate; pH 4.6	none
C3	25 % w/v Polyethylene glycol 4,000	100 mM MES; pH 6.5	200 mM Magnesium chloride
C4	25 % w/v Polyethylene glycol 4,000	100 mM TRIS; pH 8.5	200 mM Calcium chloride
C5	30 % w/v Polyethylene glycol 4,000	none	none
C6	30 % w/v Polyethylene glycol 4,000	100 mM Sodium acetate; pH 4.6	100 mM Magnesium chloride
D1	30 % w/v Polyethylene glycol 4,000	100 mM MES; pH 6.5	none
D2	30 % w/v Polyethylene glycol 4,000	100 mM HEPES; pH 7.5	200 mM Calcium chloride
D3	30 % w/v Polyethylene glycol 4,000	100 mM TRIS; pH 8.5	200 mM Lithium sulfate
D4	30 % w/v Polyethylene glycol 4,000	100 mM TRIS; pH 8.5	200 mM Sodium acetate
D5	30 % w/v Polyethylene glycol 4,000	100 mM TRIS; pH 8.5	200 mM Magnesium chloride
D6	35 % w/v Polyethylene glycol 4,000	none	none

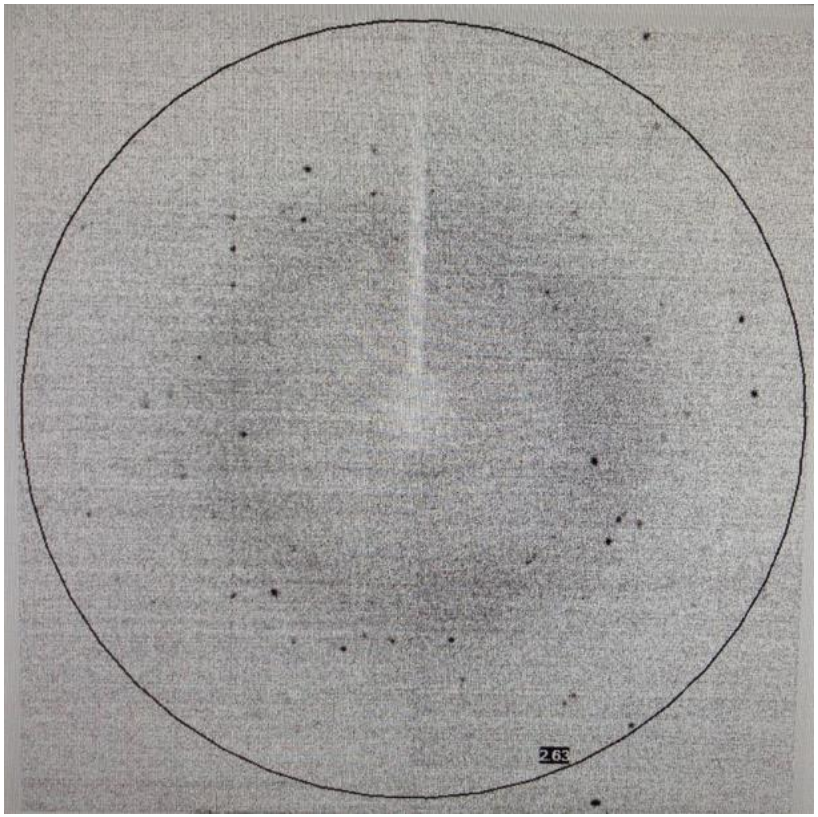
Supplementary Figure 5 –JBS 1 and JBS 2 commercial screens used in preliminary cristalization assays. Taken from www.jenabioscience.com

No.	Precipitant 1	Precipitant 2	Buffer	Additive
A1	8 % w/v Polyethylene glycol 4,000	800 mM Lithium chloride	100 mM TRIS; pH 8.5	none
A2	10 % w/v Polyethylene glycol 4,000	20 % w/v 2-Propanol	none	none
A3	10 % w/v Polyethylene glycol 4,000	10 % w/v 2-Propanol	100 mM tri-Sodium citrate; pH 5.6	none
A4	10 % w/v Polyethylene glycol 4,000	5 % w/v 2-Propanol	100 mM HEPES; pH 7.5	none
A5	10 % w/v Polyethylene glycol 4,000	20 % w/v 2-Propanol	100 mM HEPES; pH 7.5	none
A6	12 % w/v Polyethylene glycol 4,000	none	100 mM Sodium acetate; pH 4.6	200 mM Ammonium sulfate
B1	15 % w/v Polyethylene glycol 4,000	none	none	200 mM Ammonium sulfate
B2	15 % w/v Polyethylene glycol 4,000	none	100 mM tri-Sodium citrate; pH 5.6	200 mM Ammonium sulfate
B3	16 % w/v Polyethylene glycol 4,000	10 % w/v 2-Propanol	100 mM HEPES; pH 7.5	200 mM Ammonium sulfate
B4	20 % w/v Polyethylene glycol 4,000	none	none	200 mM Ammonium sulfate
B5	20 % w/v Polyethylene glycol 4,000	10 % w/v Glycerol	none	200 mM Magnesium sulfate
B6	20 % w/v Polyethylene glycol 4,000	5 % w/v 2-Propanol	none	100 mM tri-Sodium citrate
C1	20 % w/v Polyethylene glycol 4,000	20 % w/v 2-Propanol	none	100 mM tri-Sodium citrate
C2	20 % w/v Polyethylene glycol 4,000	none	100 mM MES; pH 6.5	600 mM Sodium chloride
C3	20 % w/v Polyethylene glycol 4,000	10 % w/v 2-Propanol	100 mM HEPES; pH 7.5	none
C4	22 % w/v Polyethylene glycol 4,000	none	none	100 mM Sodium acetate, 200 mM Ammonium sulfate
C5	25 % w/v Polyethylene glycol 4,000	none	100 mM Sodium acetate; pH 4.6	200 mM Ammonium sulfate
C6	25 % w/v Polyethylene glycol 4,000	none	100 mM tri-Sodium citrate; pH 5.6	200 mM Ammonium sulfate
D1	25 % w/v Polyethylene glycol 4,000	200 mM Lithium sulfate	100 mM HEPES; pH 7.5	100 mM Sodium acetate
D2	25 % w/v Polyethylene glycol 4,000	8 % w/v 2-Propanol	none	100 mM Sodium acetate
D3	30 % w/v Polyethylene glycol 4,000	none	none	200 mM Ammonium sulfate
D4	30 % w/v Polyethylene glycol 4,000	none	100 mM Sodium acetate; pH 4.6	200 mM Ammonium sulfate
D5	30 % w/v Polyethylene glycol 4,000	none	100 mM tri-Sodium citrate; pH 5.6	100 mM Ammonium sulfate
D6	32 % w/v Polyethylene glycol 4,000	none	100 mM TRIS; pH 8.5	800 mM Lithium chloride

Supplementary Figure 6 –JBS 3 commercial screen used in preliminary cristalization assays. Taken from www.jenabioscience.com

No.	Precipitant 1	Precipitant 2	Buffer	Additive
A1	25 % w/v Polyethylene glycol monomethyl ether 5,000	none	100 mM TRIS; pH 8.5	200 mM Lithium sulfate
A2	30 % w/v Polyethylene glycol monomethyl ether 5,000	none	100 mM MES; pH 6.5	200 mM Ammonium sulfate
A3	3 % w/v Polyethylene glycol 6,000	none	100 mM TRIS; pH 8.5	100 mM Potassium chloride
A4	10 % w/v Polyethylene glycol 6,000	none	none	10 mM Magnesium chloride
A5	12 % w/v Polyethylene glycol 6,000	2 M Sodium chloride	none	none
A6	15 % w/v Polyethylene glycol 6,000	5 % w/v Glycerol	none	none
B1	15 % w/v Polyethylene glycol 6,000	50 mM Potassium chloride	none	10 mM Magnesium chloride
B2	16 % w/v Polyethylene glycol 6,000	none	none	10 mM tri-Sodium citrate
B3	20 % w/v Polyethylene glycol 6,000	none	50 mM Imidazole; pH 8.0	none
B4	25 % w/v Polyethylene glycol 6,000	none	100 mM HEPES; pH 7.5	100 mM Lithium chloride
B5	28 % w/v Polyethylene glycol 6,000	500 mM Lithium chloride	100 mM TRIS; pH 8.5	none
B6	30 % w/v Polyethylene glycol 6,000	1 M Lithium chloride	none	100 mM Sodium acetate
C1	33 % w/v Polyethylene glycol 6,000	none	none	10 mM tri-Sodium citrate
C2	2 % w/v Polyethylene glycol 8,000	500 mM Lithium sulfate	none	none
C3	2 % w/v Polyethylene glycol 8,000	1 M Lithium sulfate	none	none
C4	4 % w/v Polyethylene glycol 8,000	none	none	none
C5	8 % w/v Polyethylene glycol 8,000	200 mM Lithium chloride	none	50 mM Magnesium sulfate
C6	8 % w/v Polyethylene glycol 8,000	none	100 mM TRIS; pH 8.5	none
D1	10 % w/v Polyethylene glycol 8,000	none	100 mM MES; pH 6.5	200 mM Zinc acetate
D2	10 % w/v Polyethylene glycol 8,000	none	100 mM HEPES; pH 7.5	200 mM Calcium acetate
D3	10 % w/v Polyethylene glycol 8,000	none	none	100 mM Sodium acetate, 50 mM Magnesium acetate
D4	10 % w/v Polyethylene glycol 8,000	none	none	200 mM Magnesium acetate
D5	10 % w/v Polyethylene glycol 8,000	10 % w/v Ethylene glycol	100 mM HEPES; pH 7.5	none
D6	10 % w/v Polyethylene glycol 8,000	10 % w/v Polyethylene glycol 1,000	none	none

Supplementary Figure 7 –JBS 4 commercial screen used in preliminary cristalization assays. Taken from www.jenabioscience.com



Supplementary Figure 8 – X-ray diffraction pattern from a CtCBM50 crystal. Crystal was tested in house in D8 Venture diffractometer.

Supplementary Table 3 – EWI and EWII screen used for preliminary crystallization assays. This screen was made in house and can be consulted in Macromolecular Cristallography laboratory with Cecilia Bonifácio.

A1-EWI (1) 0.1M CAPS 9.7 20 % PEG 8K	A2-EW (2) 0.2 M NaCl 0.1M HEPES7.5 10 % Isopp	A3-EWI (3) 0.1 M CAPS 9.7 15 % Etanol	A4-EWI (4) 0.2 M $MgCl_2$ 0.1 M Imidaz 8 35 % MPD	A5-EWI (5) 0.1M CAPS 10.5 30 % PEG 400	A6-EWI (6) 0.1M Citrato 5.5 20 % PEG 3K	A7-EWI (7) 0.2 M ActZn 0.1 M MES 6 10 % PEG 8K	A8-EWI (8) 0.1 M Citrato 5.5 2 M $(NH_4)_2SO_4$	A9-EWI (9) 0.1M Acetato 4.5 1 M $NH_4H_2PO_4$	A10-EWI (10) 0.1 M TRIS 7 20 % P2K MME	A11-EWI (11) 0.2 M Li_2SO_4 0.1 M MES 6 20 % Dioxan	A12-EWI (12) 0.2 M ActCa 0.1 M Imidaz 8 20 % PEG 1K
B1-EWI (13) 0.1 M Cac 6.5 1.26M$(NH_4)_2SO_4$	B2-EWI (14) 0.1 M Cac 6.5 1 M CitratoNa_3	B3-EWI (15) 0.2 M Li_2SO_4 0.1 M Imidaz 8 10 % PEG 3K	B4-EWI (16) 0.1 M Fosf 6.2 2.5 M NaCl	B5-EWI (17) 0.2 M Li_2SO_4 0.1 M Acetato 4.5 30 % PEG 8K	B6-EWI (18) 0.2 M NaCl 0.1 M Imidaz 8 1 M Tartar K/Na	B7-EWI (19) 0.1 M TRIS 7 20 % PEG 1K	B8-EW7 (20) 0.2 M NaCl 0.1 M Imidaz 8 0.4 M FosfNa 1.6 M FosfK	B9-EWI (21) 0.1M HEPES 7.5 20 % PEG 8K	B10-EWI (22) 0.1 M TRIS 8.5 10 % Isopropanol	B11-EWI (23) 0.2 M $MgCl_2$ 0.1 M Imidaz 8 15 % Etanol	B12-EWI (24) 0.2 M NaCl 0.1 M TRIS 7 35 % MPD
C1-EWI (25) 0.2 M $MgCl_2$ 0.1 M TRIS 8.5 30 % PEG 400	C2-EWI (26) 0.1 M CAPS 9.7 10 % PEG 3K	C3-EWI (27) 0.2 M Li_2SO_4 0.1M CAPS 10.5 1.2 M FosfNa 0.8 M FosfK	C4-EWI (28) 0.2 M NaCl 0.1M HEPES7.5 20 % PEG 3K	C5-EWI (29) 0.2 M NaCl 0.1 M CAPS 9.7 10 % PEG 8K	C6-EWI (30) 0.2 M NaCl 0.1M Acetato 4.5 1.26 M $(NH_4)_2SO_4$	C7-EWI (31) 0.2 M NaCl 0.1M Fosf-Cit 4.2 20 % PEG 8K	C8-EWI (32) 0.1 M Fosf 6.2 10 % PEG 3K	C9-EWI (33) 0.2 M Li_2SO_4 0.1M CAPS10.5 2 M $(NH_4)_2SO_4$	C10-EWI (34) 0.1 M Imidaz 8 1 M $NH_4H_2PO_4$	C11-EWI (35) 0.1M Acetato 4.5 20 % Dioxan	C12-EWI (36) 0.1 M Imidaz 8 1 M CitratoNa_3
D1-EWI (37) 0.1 M Imidaz 8 2.5 M NaCl	D2-EWI (38) 0.2 M Li_2SO_4 0.1 M CAPS 9.7 1 M Tartar K/Na	D3-EWI (39) 0.2 M Li_2SO_4 0.1M Fos-Cit4.2 20 % PEG 1K	D4-EWI (40) 0.2 M ActCa 0.1 M MES 6 10 % Isopropanol	D5-EWI (41) 0.1 M CAPS 9.7 30 % MPD	D6-EWI (42) 0.1 M TRIS 7 15 % Etanol	D7-EWI (43) 0.1 M Fosf 6.2 35 % MPD	D8-EWI (44) 0.2 M ActCa 0.1M Acetato 4.5 30 % PEG 400	D9-EWI (45) 0.1M Acetato 4.5 20 % PEG 3K	D10-EWI (46) 0.2 M ActCa 0.1 M Imidaz 8 10 % PEG 8K	D11-EWI (47) 0.2 M Li_2SO_4 0.1 M TRIS 7 1.26 M $(NH_4)_2SO_4$	D12-EWI (48) 0.2 M ActZn 0.1M Acetato 4.5 20 % PEG 1K
E1-EWII (1) 0.2M ActZn 0.1M Acetato 4.5 10 % PEG 3K	E2-EWII (2) 0.2 M Li_2SO_4 0.1 M MES 6 35 % MPD	E3-EWII (3) 0.2 M $MgCl_2$ 0.1 M TRIS 8 20 % PEG 8K	E4-EWII (4) 0.2 M NaCl 0.1 M Cac 6.5 2 M $(NH_4)_2SO_4$	E5-EWII (5) 0.2 M NaCl 0.1M HEPES7.5 20 % Dioxan	E6-EWII (6) 0.2 M Li_2SO_4 0.1M Fos-Cit4.2 10 % Isopropanol	E7-EWII (7) 0.2 M NaCl 0.1 M TRIS 7 30 % PEG 3K	E8-EWII (8) 0.2 M NaCl 0.1 M Fosf 6.2 10 % PEG 8K	E9-EWII (9) 0.1M Fos-Cit4.2 2 M $(NH_4)_2SO_4$	E10-EWII (10) 0.1 M TRIS 8.5 1 M $NH_4H_2PO_4$	E11-EWII (11) 0.2 M ActZn 0.1 M Cac 6.5 10 % Isopropanol	E12-EWII (12) 0.2 M Li_2SO_4 0.1 M Cac 6.5 30 % PEG 400
F1-EWII (13) 0.2 M Li_2SO_4 0.1M Citrato5.5 15 % Etanol	F2-EWII (14) 0.2 M NaCl 0.1 M Fosf 6.2 20 % PEG 1K	F3-EWII (15) 0.1M HEPES7.5 1.26M$(NH_4)_2SO_4$	F4-EWII (16) 0.1 M CAPS 9.7 1 M CitratoNa_3	F5-EWII (17) 0.2 M $MgCl_2$ 0.1 M TRIS 7 2.5 M NaCl	F6-EWII (18) 0.2 M ActCa 0.1 M TRIS 7 20 % PEG 3K	F7-EWII (19) 0.1M Fos-Cit4.2 1.26 M Fosf Na 0.4 M Fosf K	F8-EWII (20) 0.2 M ActZn 0.1 M MES 6 15 % Etanol	F9-EWII (21) 0.1M Acetato 4.5 35 % MPD	F10-EWII (22) 0.1 M Imidaz 8 10 % Isopropanol	F11-EWII (23) 0.2 M $MgCl_2$ 0.1M HEPES7.5 15 % Etanol	F12-EWII (24) 0.2 M NaCl 0.1M Imidaz 8 30 % PEG 8K
G1-EWII (25) 0.2 M NaCl 0.1M HEPES7.5 35 % MPD	G2-EWII (26) 0.1 M CAPS 9.7 30 % PEG 400	G3-EWII (27) 0.2 M $MgCl_2$ 0.1 M Cac 6.5 10 % PEG 3K	G4-EWII (28) 0.2 M ActCa 0.1 M MES 6 20 % PEG 8K	G5-EWII (29) 0.2 M NaCl 0.1 M CAPS 9.7 1.26 M $(NH_4)_2SO_4$	G6-EWII (30) 0.2 M ActZn 0.1 M Imidaz 8 20 % Dioxan	G7-EWII (31) 0.2 M NaCl 0.1 M TRIS 7 1 M CitratoNa_3	G8-EWII (32) 0.1 M TRIS 7 20 % PEG 1K	G9-EWII (33) 0.2 M NaCl 0.1M Citrato5.5 1 M $NH_4H_2PO_4$	G10-EWII (34) 0.1 M Imidaz 8 10 % PEG 8K	G11-EWII (35) 0.1 M Acetato 4.5 0.8 M Fosf Na 1.2 M Fosf K	G12-EWII (36) 0.2 M NaCl 0.1M Fos-Cit4.2 10 % PEG 3K
H1-EWII (37) 0.2 M Li_2SO_4 0.1 M TRIS 7 1 M Tart K/Na	H2-EWII (38) 0.2 M Li_2SO_4 0.1 M Acetato 4.5 2.5 M NaCl	H3-EWII (39) 0.2 M NaCl 0.1M CAPS 10.5 20 % PEG 8K	H4-EWII (40) 0.2 M ActZn 0.1 M Imidaz 8 20 % PEG 3K	H5-EWII (41) 0.2 M Li_2SO_4 0.1 M TRIS 7 2 M $(NH_4)_2SO_4$	H6-EWII (42) 0.2 M NaCl 0.1M HEPES 7.5 30 % PEG 400	H7-EWII (43) 0.2 M $MgCl_2$ 0.1 M TRIS 7 10 % PEG 8K	H8-EWII (44) 0.2 M $MgCl_2$ 0.1 M Cac 6.5 20 % PEG 1K	H9-EWII (45) 0.1 M MES 6 1.26 M $(NH_4)_2SO_4$	H10-EWII (46) 0.2 M NaCl 0.1 M Imidaz 8 1 M $NH_4H_2PO_2$	H11-EWII (47) 0.2 M ActZn 0.1 M Imidaz 8 2.5 M NaCl	H12-EWII (48) 0.1M MES 6 1 M Tart K/Na

Supplementary Table 4 – PEG Ion 1 and 8k screen used for preliminary crystallization assays. This screen was made in house and can be consulted in Macromolecular Cristallography laboratory with Cecilia Bonifácio.

A1 PEG ION 1- 0.2M NaF 20% P 4K	A2 PEG ION 2- 0.2M KF 20% P 4K	A3 PEG ION 3- 0.2M NH_4F 20% P 4K	A4 PEG ION 4- 0.2M LiCl 20% P 4K	A5 PEG ION 5- 0.2M $MgCl_2$ 20% P 4K	A6 PEG ION 6- 0.2M NaCl 20% P 4K	A7 PEG ION 7- 0.2M $CaCl_2$ 20% P 4K	A8 PEG ION 8- 0.2M KCl 20% P 4K	A9 PEG ION 9- 0.2M NH_4Cl 20% P 4K	A10 PEG ION 10- 0.2M NaI 20% P 4K	A11 PEG ION 11- 0.2M KI 20% P 4K	A12 PEG ION 12- 0.2M NH_4I 20% P 4K
B1 PEG ION 13- 0.2M NaSCN 20% P 4K	B2 PEG ION 14- 0.2M KSCN 20% P 4K	B3 PEG ION 15- 0.2M $LiNO_3$ 20% P 4K	B4 PEG ION 16- 0.2M $Mg(NO_3)_2$ 20% P 4K	B5 PEG ION 17- 0.2M $NaNO_3$ 20% P 4K	B6 PEG ION 18 -0.2M KNO_3 20% P 4K	B7 PEG ION 19- 0.2M NH_4NO_3 20% P 4K	B8 PEG ION 20- 0.2M FormMg 20% P 4K	B9 PEG ION 21- 0.2M FormNa 20% P 4K	B10 PEG ION 22- 0.2M FormK 20% P 4K	B11 PEG ION 22- 0.2M FormK 20% P 4K	B12 PEG ION 24- 0.2M ActLi 20% P 4K
C1 PEG ION 25- 0.2M ActMg 20% P 4K	C2 PEG ION 26- 0.2M ActZn 20% P 4K	C3 PEG ION 27- 0.2M ActNa 20% P 4K	C4 PEG ION 28- 0.2M ActCa 20% P 4K	C5 PEG ION 29- 0.2M ActK 20% P 4K	C6 PEG ION 30- 0.2M $ActNH_4$ 20% P 4K	C7 PEG ION 31- 0.2M Li_2SO_4 20% P 4K	C8 PEG ION 32- 0.2M $MgSO_4$ 20% P 4K	C9 PEG ION 33- 0.2M Na_2SO_4 20% P 4K	C10 PEG ION 34- 0.2M K_2SO_4 20% P 4K	C11 PEG ION 35- 0.2M S.A 20% P 4K	C12 PEG ION 36- 0.2M Tart Na_2 20% P 4K
D1 PEG ION 37- 0.2M Tart Na/K 20% P 4K	D2 PEG ION 38- 0.2M Tart$(NH_4)_2$ 20% P 4K	D3 PEG ION 39- 0.2M NaH_2PO_4 20% P 4K	D4 PEG ION 40- 0.2M Na_2HPO_4 20% P 4K	D5 PEG ION 41- 0.2M KH_2PO_4 20% P 4K	D6 PEG ION 42- 0.2M K_2HPO_4 20% P 4K	D7 PEG ION 43- 0.2M F.A. 20% P 4K	D8 PEG ION 44- 0.2M$(NH_4)_2HPO_4$ 20% P 4K	D9 PEG ION 45- 0.2M Cit Li_3 20% P 4K	D10 PEG ION 46- 0.2M Cit Na_3 20% P 4K	D11 PEG ION 47- 0.2M Cit K_3 20% P 4K	D12 PEG ION 48- 0.2M Cit$(NH_4)_2$ 20% P 4K
E1 PEG ION 8K 1- 0.2M NaF 20% P 8K	E2 PEG ION 8K 2- 0.2M KF 20% P 8K	E3 PEG ION 8K 3- 0.2M NH_4F 20% P 8K	E4 PEG ION 8K 4- 0.2M LiCl 20% P 8K	E5 PEG ION 8K 5- 0.2M $MgCl_2$ 20% P 8K	E6 PEG ION 8K 6- 0.2M NaCl 20% P 8K	E7 PEG ION 8K 7- 0.2M $CaCl_2$ 20% P 8K	E8 PEG ION 8K 8- 0.2M KCl 20% P 8K	E9 PEG ION 8K 9- 0.2M NH_4Cl 20% P 8K	E10 PEG ION 8K 10- 0.2M NaI 20% P 4K	E11 PEG ION 8K 11- 0.2M KI 20% P 8K	E12 PEG ION 8K 12- 0.2M NH_4I 20% P 8K
F1 PEG ION 8K 13- 0.2M NaSCN 20% P 8K	F2 PEG ION 8K 14- 0.2M KSCN 20% P 8K	F3 PEG ION 8K 15- 0.2M $LiNO_3$ 20% P 8K	F4 PEG ION 8K 16- 0.2M $Mg(NO_3)_2$ 20% P 8K	F5 PEG ION 8K 17- 0.2M $NaNO_3$ 20% P 8K	F6 PEG ION 8K 18- 0.2M KNO_3 20% P 8K	F7 PEG ION 8K 19- 0.2M NH_4NO_3 20% P 8K	F8 PEG ION 8K 20- 0.2M FormMg 20% P 8K	F9 PEG ION 8K 21- 0.2M FormNa 20% P 8K	F10 PEG ION 8K 22- 0.2M FormK 20% P 8K	F11 PEG ION 8K 22- 0.2M FormK 20% P 8K	F12 PEG ION8K 24- 0.2M ActLi 20% P 8K
G1 PEG ION 8k 25- 0.2M ActMg 20% P 8K	G2 PEG ION 8k 26- 0.2M ActZn 20% P 8K	G3 PEG ION 8k 27- 0.2M ActNa 20% P 8K	G4 PEG ION 8k 28- 0.2M ActCa 20% P 8K	G5 PEG ION 8k 29- 0.2M ActK 20% P 8K	G6 PEG ION 8k 30- 0.2M $ActNH_4$ 20% P 8K	G7 PEG ION 8k 31- 0.2M Li_2SO_4 20% P 8K	G8 PEG ION 8k 32- 0.2M $MgSO_4$ 20% P 8K	G9 PEG ION 8k 33- 0.2M Na_2SO_4 20% P 8K	G10 PEG ION 8k 34- 0.2M K_2SO_3 20% P 8K	G11 PEG ION 8k 35- 0.2M S.A 20% P 8K	G12 PEG ION 8k 36- 0.2M Tart Na_2 20% P 8K
H1 PEG ION 8k 37- 0.2M Tart Na/K 20% P 8K	H2 PEG ION 8k 38- 0.2M Tart$(NH_4)_2$ 20% P 8K	H3 PEG ION 8k 39- 0.2M NaH_2PO_4 20% P 8K	H4 PEG ION 8k 40- 0.2M Na_2HPO_4 20% P 8K	H5 PEG ION 8K 41- 0.2M KH_2PO_4 20% P 8K	H6 PEG ION 8K 42- 0.2M K_2HPO_4 20% P 8K	H7 PEG ION 8K 43- 0.2M F.A. 20% P 8K	H8 PEG ION 8K 44- 0.2M$(NH_4)_2$ HPO_4 20% P 8K	H9 PEG ION 8K 45- 0.2M Cit Li_3 20% P 8K	H10 PEG ION 8K 46- 0.2M Cit Na_3 20% P 8K	H11 PEG ION 8K 47- 0.2M Cit K_3 20% P 8K	H12 PEG ION 8K 48- 0.2M Cit$(NH_4)_2$ 20% P 8K

Supplementary Table 5 – 80! screen used for preliminary crystallization assays. This screen was made in house and can be consulted in Macromolecular Cristallography laboratory with Cecilia Bonifácio.

A1 80! 1- 0.2 M $CaCl_2$ 0.1 M Act 4.5 30 % MPD	A2 80! 2- 1 M Tart K/Na 0.1M MES 6.5	A3 80! 3- 0.4M F.A.	A4 80! 4- 0.1M TRIS 8.5 3M S.A.	A5 80! 5- 0.2M $CitrNa_3$ 0.1MHEPES 7.5 30 % MPD	A6 80! 6- 0.2 M $MgCl_2$ 0.1 M Act 4.5 30 % PEG 4K	A7 80! 7- 1.2M $CitrNa_3$ 0.1MHEPES 7.5	A8 80! 8-0.2M $CitrNa_3$ 2 M $(NH_4)_2SO_4$	A9 80! 9- 0.2 M $ActNH_4$ 0.1 M Cit 5.5 30 % PEG 400	A10 80! 10- 0.1M Act4.5 1.5 M F.A.	A11 80! 11- 0.2M S.A. 0.1MHEPES 7.5 1.5 M K_2HPO_4 1.5 M NaH_2PO_4	A12 80! 12- 0.2 M $CitNa_3$ 0.1 M TRIS 8.5 20 % PEG 400
B1 80! 13- 0.2 M $CaCl_2$ 0.1MHEPES 7.5 25 % PEG 4K	B2 80! 14-0.1 M $MgCl_2$ 0.1 M MES 6.5 30 % PEG 8K	B3 80! 15- 0.2M Li_2SO_4 0.1 M Cit 5.5 30 % PEG 4K	B4 80! 16- 1 M Li_2SO_4 0.1 M Act 4.5	B5 80! 17- 0.2 M F.A. 0.1 M TRIS 7.5 30 % MPD	B6 80! 18- 0.2M $ActNH_4$ 0.1 M TRIS 7.5 1.5 M K_2HPO_4 1.5 M NaH_2PO_4	B7 80! 19- 0.1 M S.A. 0.1 M Cit 5.5 30 % PEG 8K	B8 80! 20- 0.1M MES 6.5 30 % MPD	B9 80! 21-0.2 M $MgCl_2$ 0.1MHEPES 7.5 30 % PEG 4K	B10 80! 22- 0.2 M ActNa 0.1 M TRIS 8.5 30 % PEG 4K	B11 80! 23- 0.1M TRIS 7.5 1 M Tart K/Na	B12 80! 24- 0.2 M $CaCl_2$ 0.1 M TRIS 8.5
C1 80! 25- 0.5M $ActNH_4$ 0.1 M Cit 5.5 30 % MPD	C2 80! 26- 2 M ActNa 0.1 M MES 6.5	C3 80! 27- 0.2M Tart K/Na 0.1 M MES 6.5 30 % P 8K	C4 80! 28- 1 M Tart K/Na 0.1MHEPES 7.5	C5 80! 29- 0.2 M S.A. 0.1 M Act 4.5 30 % P 400	C6 80! 30- 0.1 M S.A 0.1MHEPES 7.5 20 % P 4K	C7 80! 31- 2 M S.A 0.1 M MES 6.5	C8 80! 32- 0.2 M NaCl 0.1 M MES 6.5 30 % Etanol	C9 80! 33-0.2 M $MgCl_2$ 0.1MHEPES 7.5 30 % Etanol	10 80! 34- 0.2 M $ActNH_4$ 0.1 M TRIS 8.5 30 % Etanol	C11 80! 35- 0.2 M $CaCl_2$ 0.1 M Act 4.5 30 % Etanol	C12 80! 36- 0.2 M ActNa 0.1MHEPES 7.5 30 % Etanol
D1 80! 37-0.2 M $MgCl_2$ 0.1MHEPES 7.5 30 % Isopp	D2 80! 38- 0.1 M Cac 6.5 30 % MPD	D3 80! 39- 0.1 M Act 4.5 2 M FormNa	D4 80! 40- 0.2M $CitrNa_3$ 0.1 M Cac 6.5 40 % Isopp	D5 80! 41- 0.1MHEPES 7.5 20 % PEG 400 10 % Isopp	D6 80! 42- 0.1MHEPES 7.5 1 M Li_2SO_4	D7 80! 43-0.2 M Li_2SO_4 0.1 M TRIS 8.5 30 % PEG 4K	D8 80! 44- 0.2 M S.A. 0.1 M Cac 6.5 30 % PEG 6K	D9 80! 45- 0.1M Act 4.5 1.5 M ActNa	D10 80! 46- 0.1M $CitrNa_3$ 1 M $NH_4H_2PO_4$	D11 80! 47- 4 M FormNa	D12 80! 48- 0.1MHEPES 7.5 1.2 M $CitNa_3$
E1 80! 49- 0.4M Tart K/Na	E2 80! 50-0.2 M $MgCl_2$ 0.1 M TRIS 8.5 30 % PEG 4K	E3 80! 51- 0.1 M Cac 6.5 1.4 M ActNa	E4 80! 52- 0.2M $ActNH_4$ 0.1 M Cit 5.5 30 % PEG 4K	E5 80! 53- 0.2M $ActNH_4$ 0.1 M Act 4.5 30 % PEG 4K	E6 80! 54- 0.2 M $CaCl_2$ 0.1MHEPES 7.5 28 % PEG 400	E780! 55-0.2M S.A. 0.1M Cac 6.5 30 % PEG 8K	E8 80! 56- 0.2M ActMg 0.1 M Cac 6.5 20 % PEG 8K	E9 80! 57- 0.2M $ActNH_4$ 0.1 M TRIS 8.5 30% Isopp	E10 80! 58- 0.2 M S.A. 0.1 M Act 4.5 25 % PEG 4K	E11 80! 59-0.2M ActMg 0.1 M Cac 6.5 30 % MPD	E12 80! 60- 0.2 M $CaCl_2$ 0.1 M Act 4.5 20 % Isopp
F1 80! 61- 0.1M Imid 7 1 M ActNa	F2 80! 62- 0.2 M $CitraNa_3$ 0.1 M Cac 6.5 20 % Isopp	F3 80! 63- 0.2 M ActNa 0.1 M Cac 6.5 30 % PEG 8K	F4 80! 64- 0.2 M S.A 30 % PEG 8K	F5 80! 65- 0.2 M S.A 30 % PEG 4K	F6 80! 66- 0.1MHEPES 7.5 1.6 M K_2HPO_4 1.6 M NaH_2PO_4	F7 80! 67- 0.1 M TRIS 8.5 8 % PEG 8K	F8 80! 68- 0.1M Act4.5 8 % PEG 4K	F9 80! 69- 0.1MHEPES 7.5 2 % PEG 400 1.8 M F.A	F10 80! 70- 0.1M Cit 5.5 20 % Isopp 20 % PEG 4K	F11 80! 71- 0.05M K_2HPO_4 20 % PEG 8K	F12 80! 72- 30 % PEG 1K
G1 80! 73- 0.2 M FormMg	G2 80! 74- 0.2M ActZn 0.1 M Cac 6.5 18 % PEG 8K	G3 80! 75- 0.2M ActCa 0.1 M Cac 6.5 18 % PEG 8K	G4 80! 76- 0.1 M Act 4.5 2 M $(NH_4)_2SO_4$	G5 80! 77- 0.1 M TRIS 8.5 2 M $(NH_4)_2SO_4$	G6 80! 78- 1 M Li_2SO_4 2 % PEG 8K	G7 80! 79- 0.5 M Li_2SO_4 15 % PEG 8K	G8 80! 80-0.2M $ctNH_4$ 0.1 M Cit 5.5 20 % Isopp 20 % PEG 4K	G9 SaltRx1 1- 0.1 M BTP 7 1.8 M AcetNa	G10 SaltRx1 4- 0.1 M BTP 7 1.5 M NH_4Cl	G11 SaltRx1 6- 0.1 M Act 4.5 3.5 M NH_4Cl	G12 SaltRx1 9- 0.1 M Act 4.5 2.2 M NaCl

H1 SaltRx1 14- 0.1 M TRIS 8.5 3.2 M NaCl	H2 SaltRx1 17- 0.1 M BTP 7 1 M $Cit(NH_4)_2$	H3 SaltRx1 19- 0.1 M BTP 7 0.7 M $CitNa_3$	H4 SaltRx1 26- 0.1 M BTP 7 0.7M FormMg	H5 SaltRx1 31- 0.1 M BTP 7 3.5 M FormNa	H6 SaltRx1 33- 0.1 M BTP 7 1.2M Malic Ac.	H7 SaltRx1 34- 0.1 M BTP 7 2.2M Malic Ac.	H8 SaltRx1 35- 0.1 M BTP 7 1.4 M MaltNa	H9 SaltRx1 43- 0.1 M Act 4.5 1.5M $NaNO_3$	H10 SaltRx1 44- 0.1 M BTP 7 1.5 M $NaNO_3$	H11 SaltRx1 46- 0.1 M Act 4.5 4 M $NaNaNO_3$	H12 SaltRx1 47- 0.1 M BTP 7 4 M $NaNO_3$

Supplementary information 3: Poster communications

UNRAVELLING CARBOHYDRATE BINDING SPECIFICITIES OF CBMOMES FROM TWO CELLULOLYTIC BACTERIA USING AN INTEGRATIVE HIGH-THROUGHPUT APPROACH

D. Ribeiro[1], J.L.A. Brás[2], R. Costa[1], Y. Zhang[3], W. Chai[3], Y. Liu[3], J.A.M. Prates[2,4], L.M.A. Ferreira[2,4], M.J. Romão[1], T. Feizi[3], C.M.G.A. Fontes[2,4], A.L. Carvalho[1] & A.S. Palma[1,3]

[1]UCIBIO-NOVA, Departamento de Química, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, 2829-516 Caparica, Portugal
[2]NZYTech - genes & enzymes - Estrada do Paço do Lumiar, Campus do Lumiar, Edif. E, R/C, 1649-038 Lisboa, Portugal
[3]Glycosciences Laboratory, Department of Medicine, Imperial College London, London W12 0NN, UK
[4]CIISA/FMV-UL, Faculdade de Medicina Veterinária, Universidade de Lisboa, Portugal

BACKGROUND

Plant cell-wall polysaccharides present enormous potential as resources for biotechnological applications. Several anaerobic microbial organisms have evolved a multi-protein complex, the Cellulosome (Fig.1), which efficiently contributes for plant cell-wall polysaccharide biodegradation. Cellulosomal enzymes are appended to non-catalytic protein modules, Carbohydrate Binding Modules (**CBMs**), which participate in pivotal protein-carbohydrate interactions, highly-potentiating the enzymes' catalytic efficiency and the bacteria's cellulolytic capability[1]. Bacterial genome sequencing has revealed numerous putative-CBM sequences, deposited in the CAZy database (http://www.cazy.org), for which specificities await elucidation.

AIMS

- High-throughput cloning, expression and purification of all assigned CAZy CBMs ('**CBMome**') of two cellulolytic bacteria, *Clostridium thermocellum* and *Ruminococcus flavefaciens* FD-1, that colonize the soil and the rumen of mammals, respectively.
- Investigate novel **carbohydrate interactions** and **ligand specificities of the bacterial CBMomes**, combining carbohydrate microarray technology[2] with structural characterization by X-ray crystallography, affinity gel electrophoresis and Isothermal Titration Calorimetry (ITC)[3].

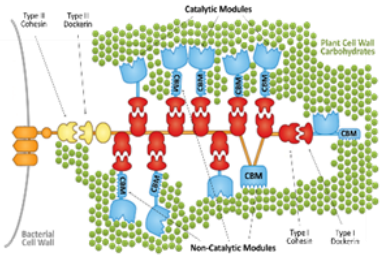


Figure 1. The Cellulosome. Representation of plant cell wall degradation by the *Clostridium thermocellum* cellulosome (adapted from Fontes & Gilbert, 2010).

CARBOHYDRATE MICROARRAY SCREENING ANALYSIS

- The carbohydrate microarray is comprised of plant and fungal cell wall representative sequences: polysaccharides and oligosaccharide fractions, prepared as NGLs[2].

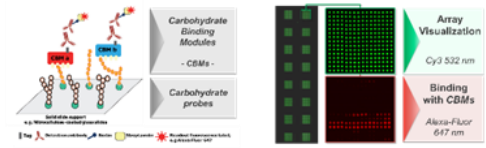


Figure 2. Representation and fluorescence imaging of carbohydrate microarrays. Polysaccharide and lipid-linked oligosaccharide probes (neoglycolipids, NGLs) are robotically printed onto 16-pad nitrocellulose-coated glass slides; binding of up to 16 CBMs to 64 different probes can be analyzed on the same slide.

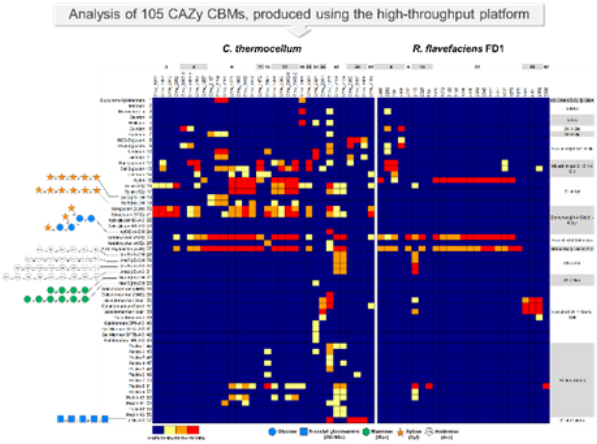


Figure 3. Microarray screening analyses of *the bacterial* CBMs. Heat-map of the relative binding intensities represented as a percentage of the fluorescence signal intensity relative to the probe that binds more strongly to each CBM. Sequence-defined NGLs are depicted 'left'.

- Different patterns of polysaccharide binding were revealed by the CBMs of the two bacteria (Fig. 3).
- Novel specificities were already assigned for some CBMs, which are being followed-up using sequence-defined oligosaccharide microarrays[2] and X-ray crystallography.

CONCLUDING REMARKS

- Comparative carbohydrate microarray analysis reveals CBMomes with different carbohydrate-binding specificities expressed by the bacteria, showing its adaptation to specific ecological niches.
- Carbohydrate microarrays combined with X-ray crystallography has proved valuable in elucidating CBMs carbohydrate-binding specificity.

NOVEL CHITIN BINDING LysM DOMAIN - *Ct*CBM50

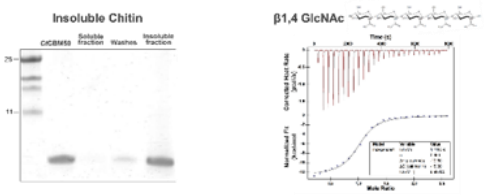


Figure 4. Binding of *Ct*CBM50 to chitin assessed by SDS-PAGE with insoluble chitin polysaccharide: using 2% chitin and **ITC with pentaacetyl-chitopentose oligosaccharide:** raw heats of binding are shown in the upper panel and the integrated heats fit to one-se binding model in the lower panel.

- *Ct*CBM50 bound strongly to chitin insoluble polysaccharide.
- The affinity exhibited for β1,4 GlcNAc was within the expect Ka values for CBMs, corroborating the microarray results (Fig. 3, Cthe_0300).

OLIGOSACCHARIDE-LIGAND SPECIFICITY OF *Ct*CBM11

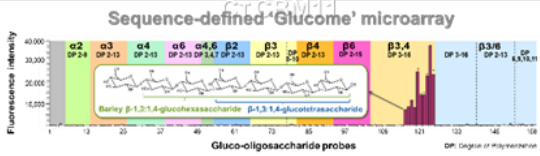


Figure 5. Assignment of the carbohydrate binding specificity of *Ct*CBM11[5]. The fluorescence intensity signals are quantified and shown as absolute values to reveal the gluco-oligosaccharide probes bound by *Ct*CBM11.

- *Ct*CBM11 is highly specific for the barley-derived β1,3:1,4-gluco-oligosaccharides (Fig. 5), in agreement with the binding preference to barley-glucans[5] (Fig. 3, Cthe_1472).
- Binding was observed only to the heptasaccharide and longer oligosaccharides.

*Ct*CBM11 mode of binding by X-ray crystallography

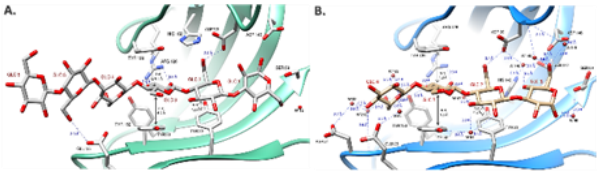


Figure 6. Structural characterization of *Ct*CBM11[5] in complex with barley-derived natural ligands. *Ct*CBM11 binding cleft occupied by (**A**) barley β1,3:1,4 glucose hexasaccharide (Glcβ1,4Glcβ1,3Glcβ1,4Glcβ1,4Glcβ1,3Glc) and (**B**) barley β1,3:1,4 glucose tetrasaccharide (Glcβ1,4Glcβ1,4Glcβ1,3Glc) (PDB 5afc). The glucose residues are numbered from the reducing end.

- The π-π stacking of the β1,4-linked Glc2 and Glc3 with Tyr22, 53 and 129 plays a key role in carbohydrate.
- Hydrogen bonding with adjacent β1,3-linked Glc1 is important for the conformation and positioning of the ligand in the binding cleft.

Check out: **Viviana Correia's work (PS, 54)** *for more on Carbohydrate Microarrays*

REFERENCES

1. C.M. Fontes, H.J. Gilbert, Annu. Rev. Biochem., **2010**, 79, 655-681.
2. A.S. Palma, *et al.*, Curr. Opin. Chem. Biol., **2014**, 18C, 87-94.
3. D. Ribeiro, *et al.* (unpublished)
4. A.S. Palma et al., Mol. Cell. Proteomics, **2015**, 14(4), 974-88.
5. A.L. Carvalho et al., J. Biol. Chem., **2004**, 279, 34785-34793.

COLLABORATIONS & ACKNOWLEDGMENTS

We would like to thank Dr. Robert Childs (Imperial College) for assistance in microarray printing and Prof. Manuel Coimbra (University of Aveiro) for kindly providing polysaccharide samples included in the microarray.

FUNDING

Supported by Portuguese Science and Technology Foundation (FCT-MEC) through grants PTDC/QUI-QUI/112537/2009, REC/IBBB-BEP/0124/2012; PTDC/BBB-BEP/0869/2014; SFRH/BD/100569/2014; Unidade de Ciências Biomoleculares Aplicadas-UCIBIO UID/Multi/04378/2013; POCI-01-0145-FEDER-007728.

www.ucibio.pt

LIGAND DISCOVERY AND STRUCTURAL-FUNCTIONAL ANALYSIS OF PROTEINS INVOLVED IN PLANT CELL WALL BIODEGRADATION

R.L.Costa[a], D.Ribeiro[a], B.A.Pinheiro[a], J.L.Brás[b], M.J.Romão[a], C.M.G.A. Fontes[bc], A.L.Carvalho[a] & A.S.Palma[a]

[a]*UCIBIO-REQUIMTE*, Dep. of Chemistry, Faculty of Science and Technology, NOVA University of Lisbon, Portugal;
[b]*NZYTech* - genes & enzymes, Lisbon, Portugal;
[c]*CIISA/FMV-UL*, Faculty of Veterinary Medicine, University of Lisbon, Portugal

THE BACTERIAL CELLULOSOME

The plant cell wall is constituted by recalcitrant polysaccharides with diverse sequences that comprise an abundant source of terrestrial biomass[1]. To efficiently degrade plant cell wall polysaccharides some cellulolytic bacterial organisms, such as *Clostridium thermocellum* and *Ruminococcus flavefaciens* FD-1, have an extracellular multi-enzyme complex called cellulosome (Fig.1) with catalytic and non-catalytic carbohydrate-binding modules (CBMs)[2]. CBMs play a crucial role in enhancing the catalytic efficiency of the enzymes by eagerly targeting the substrate. The Carbohydrate Active enZYmes database (CAZY.org) organizes the identified CBMs by sequence similarity into different families[3]. Deposition of CBM sequences in the CAZY database is continuously growing for which characterization and structure-function analysis is required.

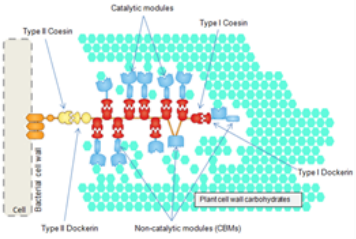
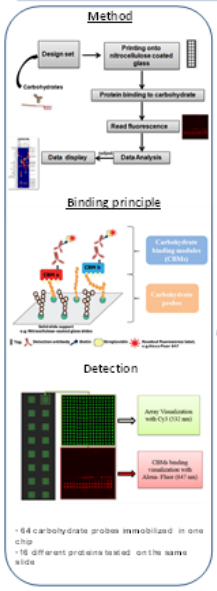


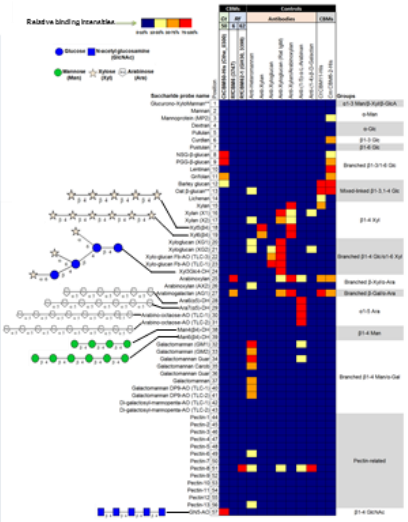
Figure 1. The Cellulosome. Representation of plant cell wall degradation by the *Clostridium thermocellum* cellulosome (adapted from Pantos & Gilbert, 2010).

AIMS

- Characterize the carbohydrate ligand specificities of *C. thermocellum* and *R. flavefaciens* FD-1 CBMs assigned to different families in the CAZY database;
- Complement the information provided by carbohydrate microarray technology using ITC (isothermal titration calorimetry) and affinity gel electrophoresis.

CARBOHYDRATE MICROARRAY SCREENING ANALYSIS

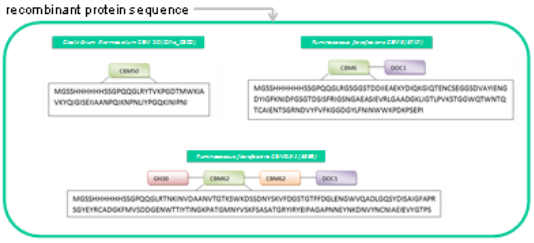




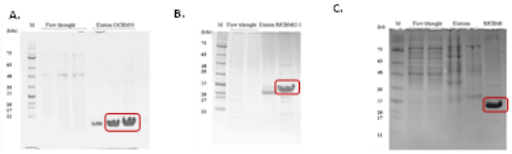
- Microarray set was validated with antibodies and CBMs of known specificity;
- Different binding specificities were obtained for the CBMs under study.

LARGE SCALE EXPRESSION AND PURIFICATION

- The modular organization of selected CBM families understudy and their respective recombinant protein sequence

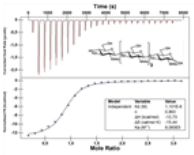


- The initial DNA was cloned into pHTP1 *Escherichia coli* prokaryotic expression vector (kanamycin resistance) with an N-terminal His-tag;
- DNA transformation was performed with *Escherichia coli* DH5α and protein expression with BL21 (DE3) as host cells;
- Purification was performed with IMAC in His-Trap™ of 5ml Ni➢ columns and their purity was assessed by SDS-page electrophoresis (panel A, B and C).

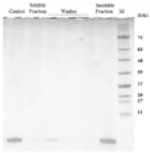


NOVEL CHITIN BINDING LYSM DOMAIN – *Ct*CBM50

ITC affinity results with pentaacetyl-chitopentaose



SDS-PAGE electrophoresis after incubation with insoluble chitin



- In agreement with microarray results *Ct*CBM50 binds pentaacetyl-chitopentaose oligosaccharide (β1,4-linked GlcNAc) with a Kd of 1.01 $\times 10^{-6}$ M;
- *Ct*CBM50 also binds strongly to chitin polysaccharide.

*Ct*CBM50 TERTIARY STRUCTURE PREDICTION

- *Ct*CBM50 homology model prediction for tertiary structure using Phyre2[5];
- LysM module from the rice blast fungus *Magnaporthe oryzae*[4] used as the template for modeling;
- 54 residues have been modeled with a coverage of 82% and confidence of 99.6%.

Crystallization attempts

- Preliminary crystallization conditions were found;
- Crystallization improvement is underway, which will be followed by structural characterization using X-ray crystallography.

Concluding Remarks

- Carbohydrate microarray technology enabled screening of a highly diverse set of probes on the same chip and ligand discovery for the CBMs analyzed.
- ITC studies and affinity gel electrophoresis complement the microarray results and allowed validation of the CBM-carbohydrate interaction.
- Further crystallization studies are needed for structural characterization of CtCBM50-carbohydrate at the atomic level.

REFERENCES

1. Varner, J.E., Lewis S., Lin, L.-S., Lewis, S., & Lin, L.-S. 1989. Plant cell wall architecture. *Cell*, 56(2) 231–239.
2. C.M. Fontes, H.J. Gilbert, Annu. Rev. Biochem. 2010, 79, 655-681.
3. A. B. Boraston, D. N. Bolam, H. J. Gilbert, & G. J.Davies. 2004. *Biochem J.*, 382(Pt 3), 769–781.
4. L.M. Kohandin et al. 2011. *Structure*, 19, 882-874.
5. http://www.sbg.bio.ic.ac.uk/~phyre2

COLLABORATIONS & ACKNOWLEDGMENTS

We would like to thank Professor Ten Feizi and colleagues from the carbohydrate microarray facility at Imperial College London for access to some of the NGLs probes and for the robotic microarray printing and Prof. Manuel Coimbra (University of Aveiro) for kindly providing polysaccharide samples included in this microarray.

FUNDING

This work was supported by the Unidade de Ciencias Biomoleculares Aplicadas-UCIBIO which is financed by national funds from FCT/MEC (UID/Multi/04378/2013) and co-financed by the ERDF under the PT2020 Partnership Agreement (POCI-01-0145-FEDER-007728). Also financed by PTDC/BBB-BEP/0892/014, RECI/BBB-BEP/0124/2012, SFRH/BDP/63083/2009 and SFRH/BPD/63963/2010.

www.ucibio.pt