André Griffin de Almeida Favinha

Bachelor Degree in Biochemistry

Protein-carbohydrate recognition by the human commensal *Bacteroides thetaiotaomicron*: expression, purification and preliminary 3D structure solution of highly specific proteins

Dissertation to obtain the Master Degree in Biochemistry

Supervisor: Doutora Maria Angelina de Sá Palma, UCIBIO, FCT/NOVA

Co-Supervisor: Doutora Ana Luísa Moreira de Carvalho, UCIBIO, FCT/NOVA

Jury

President: Professora Doutora Sofia Rocha Pauleta, Microbial Stress Research Lab UCIBIO, FCT/NOVA

Examiner: Doutor Pedro Miguel Bule Dias Gomes, Laboratório de Nutrição Animal e Biotecnologia - CIISA – FMV-UL

November of 2020

FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE NOVA DE LISBOA

Lombada

Protein-carbohydrate recognition by the human commensal *Bacteroides thetaiotaomicron*: expression, purification and preliminary 3D structure solution of highly specific proteins

André Favinha

2020

André Griffin de Almeida Favinha

Bachelor Degree in Biochemistry

Protein-carbohydrate recognition by the human commensal *Bacteroides thetaiotaomicron*: expression, purification and preliminary 3D structure solution of highly specific proteins

Dissertation to obtain the Master Degree in
Biochemistry

:

Supervisor: Doutora Maria Angelina de Sá Palma, UCIBIO, FCT/NOVA

Co-Supervisor: Doutora Ana Luísa Moreira de Carvalho, UCIBIO, FCT/NOVA

Jury

President: Professora Doutora Sofia Rocha Pauleta, Microbial Stress Research Lab UCIBIO, FCT/NOVA

Examiner: Doutor Pedro Miguel Bule Dias Gomes, Laboratório de Nutrição Animal e Biotecnologia - CIISA – FMV-UL

November of 2020

FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE NOVA DE LISBOA

André Griffin de Almeida Favinha

Graduated in Biochemistry

Protein-carbohydrate recognition by the human commensal *Bacteroides thetaiotaomicron*: expression, purification and preliminary 3D structure solution of highly specific proteins

Dissertation to obtain the Master's Degree in
Biochemistry

Supervisor: Doutora Maria Angelina de Sá Palma, UCIBIO, FCT/NOVA

Co-Supervisor: Ana Luísa Moreira de Carvalho, UCIBIO, FCT/NOVA

Jury

President: Professora Doutora Sofia Rocha Pauleta, Microbial Stress Research Lab UCIBIO, FCT/NOVA

Examiner: Doutor Pedro Miguel Bule Dias Gomes, Laboratório de Nutrição Animal e Biotecnologia - CIISA – FMV-UL

November of 2020

Protein-carbohydrate recognition by the human commensal *Bacteroides thetaiotaomicron*: expression, purification and preliminary 3D structure solution of highly specific proteins

Acknowledgments

À minha orientadora, **Doutora Angelina Palma** (GlycoLab - Functional Glycobiology), e co-orientadora, **Doutora Ana Luísa Carvalho** (XTAL – Macromolecular Crystallography) por terem acreditado em mim e nas minhas capacidades, dando-me a oportunidade para trabalhar neste grupo incrível. Deram-me o apoio que necessitava e ensinamentos que me fizeram crescer não só como estudante, mas também como pessoa. Convosco aprendi imenso e estou imensamente grato!

À **Professora Doutora Maria João Romão**, líder do grupo de investigação XTAL da FCT/NOVA e Diretora da Unidade de Investigação UCIBIO (Unidade de Ciências Biomoleculares Aplicadas) tendo sido amavelmente recebido e englobado neste ambiente de trabalho único.

À **Doutora Márcia Correia** e **Doutora Benedita Pinheiro** pelo acompanhamento do meu trabalho e por muitas vezes me terem ajudado com tarefas de laboratório quando me encontrava sozinho.

À **Raquel Costa** por ter acompanhado todo o meu trabalho, ter sido a minha ajuda incansável, quer dentro, quer fora do ambiente laboratorial. Esteve sempre disponível para ajudar, para me aturar e responder às minhas dúvidas existenciais, principalmente no início. Foste essencial para a elaboração deste trabalho e agradeço-te imenso.

À **Filipa Trovão** que acompanhou e ajudou-me incansavelmente em algumas partes do meu trabalho e me fez sentir à vontade em tarefas onde sentia algumas dificuldades.

À **Viviana Correia** por ter tido um papel fundamental nos estudos prévios a este trabalho e também pelas importantes dicas de elaboração de apresentações que me ajudaram imenso.

Aos colegas do grupo XTAL **Professora Doutora Teresa Santos Silva, Doutor Marino Santos e Doutor Cristiano Mota** por me terem recebido tão bem e por todo o apoio que me deram.

À **Liliana Ringler, Filipa Lopes, Patrícia Borges** e **Raquel Gama** por terem sido durante grande parte deste trajeto a minha companhia dentro do laboratório. Proporcionaram um ambiente de trabalho incrível e a vossa simpatia e alegria contagiante ajudaram-me imenso neste percurso.

Ao **Professor Eurico Cabrita** e á **Comissão Organizadora** da **Escola de Verão (UCIBIO Summer School**) por me terem dado a oportunidade de enriquecer a minha prática laboratorial, bem como o trabalho de investigação.

Aos colaboradores da NZYTech, o **Professor Carlos Fontes** e a **Doutora Joana Brás** e também aos colaboradores do Glycosciences Laboratory, Imperial College London, a **Professora Ten Feizi** e o **Doutor Wengang Chai,** pela contribuição essencial para o projeto no âmbito do qual foi desenvolvido o meu plano de trabalhos.

À **Fundação para a Ciência e Tecnologia - Ministério da Ciência, Tecnologia e Ensino Superior** pelo financiamento no âmbito do projeto PTDC/BIA-MIB/31730/2017 e da Unidade de Investigação UCIBIO (UID/Multi/04378/2019); e pela bolsa de investigação de 3 meses no âmbito da Escola de Verão UCIBIO, em colaboração com a Direção Geral do Ensino Superior (DGES).

À **Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa**, a casa que me acolheu durante os últimos cinco anos académicos da minha vida, onde também agradeço a todos os docentes da minha Licenciatura e Mestrado em Bioquímica que foram fundamentais para o meu crescimento académico e também pessoal. Uma palavra de agradecimento ao coordenador do Mestrado em Bioquímica, o **Professor Doutor Ricardo Franco** pelo apoio nestes 2 anos de mestrado.

Aos meus grandes amigos **Diogo Paulo** e **João Alves** por me terem feito companhia não só ao longo desta tese, mas a partir do momento em que nos conhecemos. Considero vos meus irmãos e assim quero que continuem porque como vocês há poucos.

Ao **resto dos meus amigos** que estiveram sempre presentes para me apoiar e proporcionar momentos de boa disposição que me ajudaram a levar avante este projeto. Sem vocês não teria a mesma graça!

À **Família Pinheiro** que me apoiou incansavelmente em momentos bons e maus desta caminhada e que me proporcionaram uma segunda casa onde fui sempre bem vindo e recebido. Agradeço do fundo do meu coração todo o apoio e companhia prestados.

Aos **meus pais, irmã e avós** por todo o apoio que me deram ao longo da minha vida e me darem todas as condições necessárias para chegar aqui. Sem vocês isto não seria possível. Obrigado por me acompanharem nos meus sonhos e por todos os incentivos de maneira a poder concretizá-los. Devo-vos tudo. Amo-vos.

À **Sara**, the love of my life. Nunca me deixaste desistir, estando sempre presente do meu lado em todos os momentos, uns bons, outros menos bons. Sabes muito bem o que significas para mim e quero continuar a crescer e aprender contigo, tal como me fizeste crescer e aprender neste último ano. Amo-te!

Por fim, e num tom de profunda tristeza, frustração e eterna saudade quero dedicar esta tese ao meu melhor amigo, ao companheiro de uma vida, a um guerreiro que infelizmente não poderá vivenciar a entrega nem a defesa desta tese: **o meu avô**. Partiste a uma semana de me ver entregar esta tese e não há palavras para descrever a tristeza e raiva que sinto neste momento em que te vejo partir. Ensinaste-me muita coisa, contigo cresci imenso, ajudaste-me em tudo o que precisava sem pedir nada em troca. Eras um grande senhor, o meu melhor amigo e nunca te irei esquecer. O teu sofrimento acabou e sei que a partir deste momento estarás a olhar por mim e por todos nós que sentimos a tua falta. A ti te dedico esta dissertação, meu herói e ídolo. **Amo-te avô**.

Outputs of the Thesis

3-month scholarship

Project entitled "Protein-glycan interactions in the recognition of the human gastrointestinal epithelium by the microbiome", **UCIBIO Summer School**, 1st August- 31st October

Oral Communications

A.G. Favinha "Carbohydrate-recognition in the Human Gut – Are Bacteria friends or foes?", Jornadas Intercalares dos Mestrados do DQ e do DCV, February 2020

A.G. Favinha "Protein-glycan interactions in the recognition of human gastrointestinal epithelium by the microbiome", Symposium of the UCIBIO Summer School, October 2020

Poster communications:

R.L. Costa, V.G. Correia, B.A. Pinheiro, **A.G. Favinha**, J.L.A. Brás, Y. Liu, L.M. Silva, M.J. Romão, W. Chai, T. Feizi, C.M.G.A. Fontes, A.L. Carvalho & A.S. Palma. "Deciphering the molecular details of human host O-glycan recognition by *Bacteroides thetaiotaomicron*", Annual PhD meeting, ITQB, Oeiras, Portugal, January 2020

R.L. Costa, V.G. Correia, B.A. Pinheiro, **A.G. Favinha**, J.L.A. Brás, M.J. Romão, C.M.G.A. Fontes, A.S. Palma & A.L. Carvalho "Strategies to solve the 3D structures of two proteins from a prominent human gut commensal", CCP4 Data Collection and Structure Solution Workshop, December 2020 (**Best Poster award**)

Manuscript in preparation:

The results described in Chapter 3 will be included in the following manuscript in preparation:
R.L. Costa, **A.G. Favinha**, V.G. Correia, B.A. Pinheiro, J.L.A. Brás, Y. Liu, L.M. Silva, T. Feizi, C.M.G.A. Fontes, A.L. Carvalho & A.S. Palma. "Structural basis for the divergent specificity of *Bacteroides thetaiotaomicron* family 32 carbohydrate binding modules"

Abstract

The human intestine has an important role in human digestion where diverse carbohydrates are digested to absorbable short chain fatty acids (SCFAs), crucial for human health. Commensal bacteria within this niche play an important role in this process and are crucial to maintain human nutritional balance by efficiently degrading diverse carbohydrates as sources of nutrients. These are either obtained through our diet or constituents of glycoproteins that cover the intestinal epithelium. A major predominant bacterium is *Bacteroides thetaiotaomicron*. In the genome of this bacterium, there are gene clusters termed *Polysaccharide utilization loci* (PUL), which code for proteins involved in the specific targeting of carbohydrate substrates, such as the non-catalytic carbohydrate-binding modules (CBMs), associated with catalytic enzymes, and Starch utilization system D-like proteins (SusD-like). Information about these proteins is continuously deposited in the Carbohydrate Active enZymes database (CAZy). However, the majority lack structural and functional characterization.

The main goal of this thesis was to structurally characterize *B. thetaiotaomicron* PUL proteins involved in the recognition of host-like carbohydrates to understand at molecular level their carbohydrate-binding specificity. To achieve this, different CBMs of CAZY family 32 (CBM32) and SusD-like proteins were recombinantly expressed in *Escherichia coli*. Successful expression and purification to homogeneity was achieved for pairs of CBM32/SusD: *Bt*0865-CBM32/*Bt*0866-SusD from PUL 12 and *Bt*4040-CBM32/*Bt*4038-SusD from PUL 73. Structural characterization was followed up for *Bt*4040-CBM32. The 3D structure, solved at a maximum resolution of 1.93Å, revealed a β-sandwich fold, common to known members of this family. The structural comparison with characterized CBM32 revealed an overall conservation of the core tertiary structure, but also structural features unique to *Bt*4040-CBM32, particularly the lack of conservation of amino acid residues in the putative binding site. These differences could account for the distinct binding specificity of this CBM to epitopes containing fucosylated Lewis A determinants and could reflect adaptive pressures on *B. thetaiotaomicron* from its natural habitat.

Keywords: commensal bacteria, carbohydrates, *Polysaccharide utilization loci*, CBM32, SusD-like, heterologous expression, X-ray crystallography

Resumo

O intestino humano tem um papel importante na digestão, onde hidratos de carbono são digeridos e transformados em ácidos gordos de cadeia curta (SCFAs), cruciais para a saúde humana. As bactérias comensais presentes neste nicho desempenham um papel importante neste processo e são cruciais para manter o equilíbrio nutricional sendo capazes de degradar eficazmente diversos hidratos de carbono como fonte de nutrientes. Estes podem ser obtidos através da nossa dieta ou de glicoproteínas do hospedeiro que revestem o epitélio intestinal. Uma das bactérias predominantes é a bactéria *Bacteroides thetaiotaomicron*. No genoma desta bactéria, existem conjuntos de genes denominados *Polysaccharide utilization loci* (PUL), que codificam para proteínas envolvidas na interação específica com hidratos de carbono, tais como módulos não-catalíticos de ligação a hidratos de carbono (CBMs), associados a enzimas catalíticas, e proteínas *Starch utilization System D* (SusD)-*like*. A informação sobre estas proteínas está continuamente a ser depositada na base de dados Carbohydrate Active enZymes (CAZy). No entanto, a maioria carece de caracterização estrutural e funcional.

O principal objetivo desta tese é a caracterização estrutural de proteínas constituintes de PULs de *B. thetaiotaomicron*, envolvidas no reconhecimento de hidratos de carbono do hospedeiro, de maneira a compreender a nível molecular a sua especificidade de ligação. Para o conseguir, diferentes CBMs da família 32 (CBM32) e proteínas SusD-like foram expressas de forma heteróloga em *Escherichia coli*. Os pares CBM32/SusD: *Bt*0865-CBM32/*Bt*0866-SusD pertencentes ao PUL 12 e *Bt*4040-CBM32/*Bt*4038-SusD pertencentes ao PUL 73, foram produzidos e as proteínas recombinantes purificadas em larga escala. A caracterização estrutural foi efetuada para o *Bt*4040-CBM32. A estrutura 3D foi determinada com uma resolução máxima de 1,93Å, e revelou um enrolamento em β-sandwich, comum aos membros conhecidos desta família de CBMs. A comparação estrutural com CBM32s caracterizados revelou uma conservação global da estrutura secundária, mas também características estruturais únicas do *Bt*4040-CBM32, destacando-se a falta de conservação de aminoácidos situados no local de ligação putativo. Esta divergência de sequência, pode explicar as diferenças de especificidade da ligação a hidratos de carbono deste CBM, nomeadamente a especificidade para determinantes fucosilados Lewis A, como consequência de possíveis pressões adaptativas do *B. thetaiotaomicron* ao seu habitat natural.

Palavras-chave: bactérias comensais, hidratos de carbono, *Polysaccharide utilization loci*, CBM32, SusD-like, expressão heteróloga, cristalografia de raios-X

Contents

List of Figures

List of Tables

Abbreviations and Symbols

% (v/v) –volume/volume percentage

Å- Angstrom

Asn- Asparagine

Asp- Aspartate

B. thetaiotaomicron- Bacteroides thetaiotaomicron

BSA- Bovine serum albumin

CAZy- Carbohydrate active enzyme

CBD- Cellulose binding domain

CBM- Carbohydrate Binding module

CE- Carbohydrate esterase

DHPE- 1,2-dihexadecyl-sn-glycero-3-phosphoethanolamine

DTT- Dithiothreitol

Fuc- Fucose

GAG-glycosaminoglycan

Gal- Galactose

GalNAc- N-acetylgalactosamine

GBP- Glycan binding protein

GH- Glycoside hydrolase

GI- gastrointestinal

Glc- Glucose

GlcNAc- N-acetylglucosamine

GPI- glycosylphosphatidylinositol

GT- Glycosyltransferase

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

His- Histidine

HMO- Human milk oligosaccharide

IBD- Inflammatory bowel disease

IMAC- Immobilized metal affinity chromatography

IPTG- Isopropyl β-D-1-thiogalactopyranoside

Kd- Dissociation constant

kDa- Kilodaltons

LB- Luria Bertani

Leu- Leucine

LNnT- lacto-N-neotetraose

LNT- lacto-N-tetraose

MAD- Multi-wavelength anomalous dispersion

Man- Mannose

MIR- Multiple isomorphous replacement

MR- Molecular replacement

MST- MicroScale thermophoresis

Muc1- Mucin type 1 protein

Muc2- Mucin type 1 protein

MUC2- Mucin type 2 gene

n- Stoichiometry

Neu5Ac

NGL- Neoglycolipid

NGP- Neoglycoprotein

NHS- N-hydroxysuccinimide

OD_{600}- Optic density at 600 nm

Phe- Phenylalanine

PL - Polysaccharide lyase

PO- Polysaccharide oxidase

PTS- Proline, threonine, serine rich domain

PUL- *Polysaccharide utilization loci*

RMSD - Root-mean-square deviation

Rpm- Revolutions per minute

SAD- Single-wavelength anomalous dispersion

SCFA- Short chain fatty acid

SDS-PAGE- Sodium dodecyl sulphate–polyacrylamide gel electrophoresis

Ser- Serine

SGBP - Surface glycan-binding protein

Sus- Starch utilization system

TBDTS- TonB-dependent transporter

Thr- Threonine

Tm- Melting temperature

TPR- Tetratricopeptide repeat

TSA- Thermal shift assay

Tyr- Tyrosine

Tyr- Tyrosine

UC- Ulcerative colitis

Val- Valine

σ- Sigma

Chapter 1 – Introduction and objectives

1.1. Overview of the human gut microbiome

The adult human intestine is a well-known and studied organ that plays a role of extreme importance in digestion and can be characterized as an anaerobic bioreactor for different types of organisms [1]. These organisms are representative of the *Archaea*, *Bacteria* and *Eukarya* domains, the three domains of life. The human gut is an organ that comprises different cell lineages that have the ability of communicating with each other. It can maintain and repair itself through self-replication and has the ability of storing and redistributing energy.

The anaerobic microenvironment observed in the human gut is considered one of the densest microbial communities know in nature and the size of the population is around 100 trillion microorganisms. This number, to have some perspective, far exceeds the number of organisms present in any other community associated with our body's surfaces and it is around 10 times greater than the number of somatic and germ cells in our bodies [2].

The relationship between humans and the organisms present in the gut can be described as a commensal relationship, where both humans and bacteria benefits.

Adult and healthy microbiota are dominated by two bacterial phyla: the Firmicutes and the Bacteroidetes [2], [3]. The Bacteroidetes is by far the dominating phyla in the mammalian gut and can establish long term stable relationships with the host, conferring numerous health benefits like the decrease in pathogen susceptibility and digestion of complex dietary polysaccharides [4]. From a researcher's point of view, they offer some advantages in comparison to other microorganisms, including being genetically tractable and readily cultured. In fact, the Bacteroidetes phyla has been one of the major focus of interest in the field of human gut microbiology since its discovery, living and thriving exclusively on the gastrointestinal tract (GI tract) of mammals, which suggests an adaptation to its environment [5]. This makes *Bacteroides spp.* a perfect model for modern and future society to have a deeper and better understanding about the fundamental questions that still have found no answer, such as the mechanism that allows microbiota to persist not only during the life span of the host but over an evolutionary timescale point of view [4].

There have been several studies suggesting that bacteria from the Bacteroidetes phyla can produce a cytochrome *bd* oxidase, hypothesized to lower the levels of oxygen, making it more advantageous for the growth of obligated anaerobes that otherwise would be killed by the presence of oxygen. This alteration of oxygen levels to a lower level may be one of the many reasons why Bacteroidetes is so widespread across mammals [6], [7].

This class of Gram negative and obligate anaerobe bacteria comprises the rod-shaped *Bacteroides thetaiotaomicron*, which is both a commensal and mutualist and was the first member

of Bacteroidetes to have its genome sequenced after isolation from a human faecal sample [8]. This particular organism is one of the most studied concerning the gut microbiota, and is of an extreme importance because it is a carbohydrate-degrading bacterium allowing humans and other mammals to obtain important nutrients from dietary polysaccharides that could not be obtained otherwise through their endogenous enzyme repertoire [9]. Furthermore, recent evidence has demonstrated that this bacterium has adapted to the gut ecological niche by developing the capability of adhering and using host-derived glycans as sources of nutrients (**Figure 1.1**) [10]. Therefore, it is extremely important to have a balanced diet that allows gut bacteria, including *B. thetaiotaomicron* to thrive.

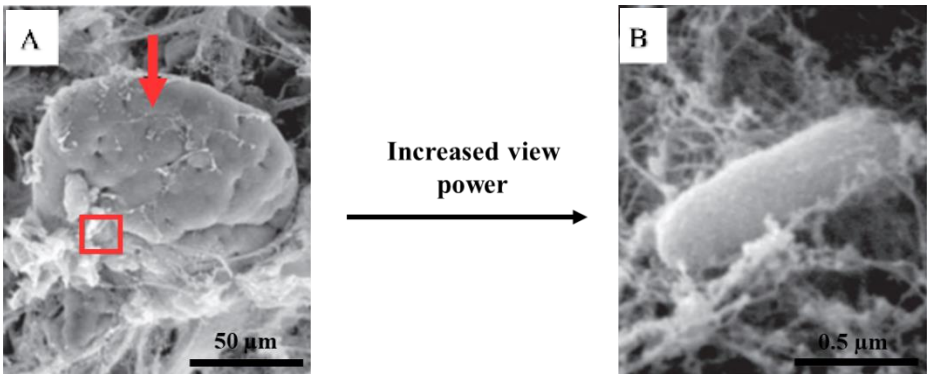


Figure 1.1- Scanning electron microscopy images that show the distribution of *B. thetaiotaomicron* within the intestinal niche. (A) Low power view of the distal small intestine showing a villus (arrow) viewed from above B) Close-up view of one *B. thetaiotaomicron* cell associated on the mucus layer of the intestinal wall. Adapted from [10].

The fermentation of dietary polysaccharides by anaerobic intestinal microbiota leads to the formation of end products called short-chain fatty acids (SCFAs) [11]. These consist of linear or branched fatty acids that contain six or less carbons in their structure, that are rapidly absorbed by colonocytes (≈95%) while the remaining 5% is secreted in the faeces. SCFAs can also be referred as volatile fatty acids and consist mainly of acetic, propionic and butyric acid [11]. These molecules have been shown to provide multiple beneficial effects on mammalian metabolism, leading to intensive research during the past years. When taken up, SCFAs provide approximately 10% of the daily caloric requirements and have several effects on host metabolism like fatty acid, glucose and cholesterol regulation [12]. SCFAs are not only beneficial for the host, but also for the microbial community, required to balance redox equivalent production in the anaerobic environment of the gut. In the last decades, researchers hypothesized that SCFAs might have an important role in preventing metabolic syndromes, cancer and bowel disorders [13].

Several studies reported that the administration of SCFAs led to the treatment of conditions like ulcerative colitis and Crohn's disease [14]. The mechanisms by which SCFAs exert these effects are not totally clear and comprise a constant subject in present research, mainly

due to the lack of human data, because most of the effects exerted in rats cannot be directly extrapolated to humans.

To counteract the side effects of a low fiber diet that could potentially harm the host microbiota and consequently lead to disease prebiotics and probiotics can be used [11]. A prebiotic is simply a specific component that is present in food products that can be used to selectively shape the abundance or physiology of a group of bacteria in the microbiome, while a probiotic consist of live microorganisms that provide health benefits. Plant polysaccharides are an example of a prebiotic that is nowadays used in prebiotic therapies to manipulate the microbiota of the human gut, promoting host health. The long-chain β-fructan inulin and also smaller fructo-oligosaccharides promote the growth of *Bifidobacterium spp.* [15]. Supplementation with these fibers has shown to control the adverse metabolic effects of a high-fat diet. In addition, certain prebiotics can also lower the levels of cholesterol by stimulating the formation of SCFAs, mainly propionate. Propionate is then absorbed into the bloodstream and when it reaches the liver it inhibits cholesterol production [16].

Taking all this in consideration a future goal is to develop new strategies not only involving pre- and probiotics, but also pharmacological strategies. For this to happen, a deeper knowledge and insight into the microbiota composition as well as their metabolism is imperative.

1.2. Carbohydrates and glycosylation

Carbohydrates, also generally known as glycans, accessible by gut bacteria can be from exogenous sources originated from the diet or can be expressed and secreted endogenously by host cells [17]. In the first case, carbohydrates can be derived from various sources such as plant dietary polysaccharides (pectins, xylan), oligosaccharides present in the milk or in meat and cartilages. However, carbohydrates are also present in the host's various mucosal surfaces, being also available as a source of energy for the bacteria that live among our bodies. All of these processes result in glycan's availability to the carbohydrate-degrading bacteria that comprise our gut microbiome, leading to the extreme importance of glycans in the normal functioning of the microbiome. The field of glycobiology is the one interested in the study of the structure, biosynthesis and role of these glycans [18]. Most glycans are found on the surface of the cells associated to proteins, lipids or proteoglycans. A representation of the different major types of glycans at the surface of a mammalian cell is presented in **Figure 1.2** below: *N*-linked glycoproteins, *O*-linked glycoproteins, glycolipids, glycosaminoglycans (GAGs) and glycosylphosphatidylinositol (GPI) anchor proteins [19]. However they can also exist in the nucleus and cytoplasm of the host cells in the form of simpler glycans attached to proteins that exhibit regulatory effects [20].

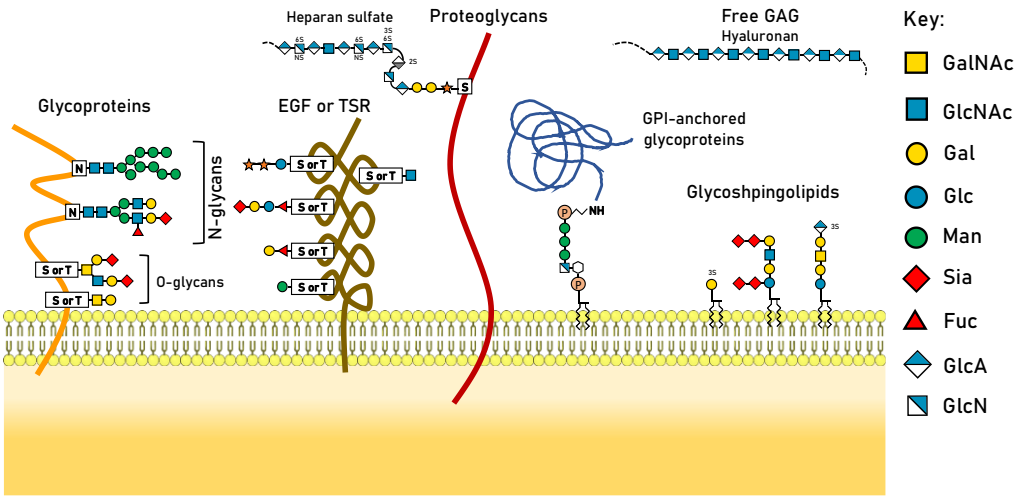


Figure 1.2- Representation of the major types of glycoconjugates of a mammalian cell. Glycans can be attached to proteins or lipids giving rise to diverse glycoproteins, proteoglycans, GPI-anchored glycoproteins, glycosphingolipids and glycosaminoglycans. Adapted from [20].

Glycans can also be found in the form of free oligosaccharides that are secreted in the milk (human milk oligosaccharides- HMOs). These can be found in high concentrations, particularly in human breast milk and have a prebiotic effect, serving as a major energy and food source for the bacteria present in the gut of newborn child, such as bifidobacteria [21]. With very few exceptions, all human milk oligosaccharides consist of a lactose or N-acetyllactosamine-based backbone; examples include the lacto-N-neotetraose (LNnT), with the sequence Gal-β-1,4-GlcNAc-β-1,3-Gal-β-1,4-Glc, and lacto-N-tetraose (LNT) with the sequence Gal-β-1,3-GlcNAc-β-1,3-Gal-β-1,4-Glc.

The addition of glycans to a macromolecule is a process denominated glycosylation and the resulting molecule is designated a glycoconjugate. The addition of monosaccharides is mediated by enzymes denominated glycosyltransferases (GT), which catalyse the transference of a monosaccharide residue form a donor activated substrate to an acceptor substrate in highly specific manner [22]. A wide range of naturally occurring monosaccharides can be added to lipids and proteins in a determined sequence to create multiple unique glycoconjugates (glycolipids or glycoproteins, respectively), all different from one another and each one exerting their own unique function. Glucose (Glc), galactose (Gal), N-acetylglucosamine (GlcNAc) and N-acetylgalactosamine (GalNAc) are common monosaccharides that make up the composition of glycan backbone structures present on mammalian-type cells. Monosaccharides such as fucose (Fuc) and N-acetyl neuraminic acid (Neu5Ac) are frequently terminal monosaccharides [20]**.** A study from 1994 revealed the possibility of the existence of 10^{12} different branched glycans [23].

The addition of glycans to proteins occurs post-translationally and it can be divided into two major glycosylation types: N- glycosylation and O-glycosylation. Many proteins can be modified through N-glycosylation, in which a GlcNAc residue is linked to the nitrogen atom of an asparagine (Asn) side chain via a glycosidic bond (β-1N linkage). These Asn-linked

glycoconjugates possess a conserved core formed by two GlcNAc residues and 3 mannose (Man) residues ($GlcNAc_2Man_3$) and to this core, distinct monosaccharides can be added through fucosylation, galactosylation and sialylation, for example. These N-glycans can be found in most living organisms and they have an extremely important role in the regulation of both intracellular and extracellular functions.

The O-glycosylation of proteins occurs on amino acids that contain functional hydroxyl groups, Serine (Ser) or Threonine (Thr), where the glycan is attached to the oxygen atom of that hydroxyl group by a covalent bond, via an O-glycosidic bond. Tyrosine (Tyr) O-glycosylation, in which the glycan attaches to the hydroxyl group of tyrosine also exists, but less frequent. One of the first and only examples of this type of glycosylation is the addition of a glucose residue to glycogenin, a glycosyltransferase [19]. The most common sugars that are linked to these amino acids are GlcNAc and GalNAc. One type of unique glycosylation that occurs mainly in the nucleus and cytoplasm is referred as O-GlcNAc glycosylation, synthesized by O-GlcNAc transferase [24]. GalNAc-linked O-glycans are major constituents of mucin glycoproteins of the gut mucosal surface that forms an extremely important interface between epithelial gut cells and the intestinal lumen [24]. This is the reason why GalNAc-linked glycans are often called mucin-type O-glycans [25], [26]. The biosynthesis of GalNAc O-glycans starts by the transference of the GalNAc monosaccharide to Ser/Thr residues of the polypeptide chain, which is catalysed by different GalNAc transferases. The glycan backbone can be further extended by the action of different glycosyltransferases (section 1.2.1 below).

Glycosylation is extremely important in a wide range of biological processes, especially in proteins, conferring structure and stability to undergo their specific biological pathways. However, glycans are also the main food source for microbes, especially its carbon content, leading to the idea that it is important to have a balanced diet in order to minimize the effects on the composition of gut microbiome, microbe gene expression, mucin composition and immune adaptive responses.

1.2.1. The intestinal mucus layer and mucin type O-glycosylation

Our intestinal tract is unique when compared to other mucosal tissues, mainly because it is heavily colonized by bacteria and other microorganisms from birth and throughout the course of our lives. In the colon, microorganism density can reach a number of $1x10^{12}$ per gram of lumen content [27], [28]. The intestinal tract, particularly the epithelium surface is covered by a layer of mucus that helps to keep maintenance over the host health and lower its susceptibility of disease, by forming a barrier between the luminal microbiota and our immune cells. This mucus layer varies in thickness as well as in composition across the full length of the GI tract, being thinner along the small intestine and growing in thickness along the large intestine [29]. In the colon, the mucus layer is complex, exhibiting two layers: an outer thicker layer that is loosely attached to a

second inner and thinner layer that is firmly attached to the epithelium. These two layers exhibit different scenarios when interactions with microbiota are taken into consideration: whereas the outer layer is in direct contact, and consequently, heavily colonized by intestinal microorganisms, the inner layer is almost devoid of bacteria. This leaves a bacteria-free zone that is adjacent to the epithelium, leading to the conclusion that intestinal mucus is an important mediator of host-bacteria interactions [30].

Patients suffering from a condition denominated human ulcerative colitis (UC) have been a subject of attention by several studies that revealed defects on the normal functionality of mucus. For instance, the alterations in functionality comprised decreased mucus layer thickness, consequently leading to a higher rate of bacteria penetration through the mucus barrier, a decrease of mucin synthesis and secretion of mucins and altered O-glycosylation of mucins [31]. A study conducted in mice models suffering from inflammatory bowel disease (IBD), revealed a deficiency in mucin synthesis prior to the disease development [32]. The study showed that mice deficient in the mucin-type 2 glycoprotein gene (MUC2) lacked the formation of a mucus layer around the epithelium and consequently failed to restrict bacterial attachment. This condition led to the appearance of complications such as colitis and more severely, colorectal cancer. The information taken from these studies leads to the conclusion that mucins have an extreme importance in maintaining a healthy gastrointestinal tract, as well as maintaining the homeostasis between the host and gut microbiota.

As mentioned above, mucins represent a group of high molecular weight O-linked glycoproteins that are widely expressed in our body, produced by the epithelial tissues in most animals. Mucins main characteristic is their ability to form gel like substances, therefore incorporating most gel-like secretions and having diverse functions, including lubrication, cell signalling and forming chemical barriers [33]. The previously referred mucin-type 2 is the main component of intestinal mucus. However, the high expression of mucins is not always beneficial, as the overexpression of mucin-type 1 protein (Muc1) is associated with several types of cancer [34]. In the case of humans there are 20 known genes that code for mucins with the majority expressed across the GI tract and they can be classified into three major classes: secretory gel-forming, secretory nongel-forming and membrane bound [35].

Muc2, the main secretory gel-forming mucin in the gut, is secreted by goblet cells and its biosynthesis is complex and highly regulated. Membrane mucins are important contributors for the glycocalyx of mucosal surfaces, playing important roles in cell-cell and cell-matrix interactions [36]. Despite having miscellaneous functions, the mucins from these families share a common characteristic which is the presence of large domains containing proline, serine and threonine called proline, threonine, serine-rich, or simply PTS domains. In the case of gel-forming mucins, these PTS domains can be found in repeated sequences that display different patterns of

extensive glycosylation (glycans can represent up to 80% of mucin total mass), conferring diversity in function and structure to mucins.

Regarding mucin glycosylation, these proteins contain a widespread array of glycan structures, depending on the glycosyltransferase profile of the host, which will determine the specific glycan structures and linkages between the monomers present on the secreted mucins [37]. As referred earlier, the mucin O-glycan synthesis starts with the addition of GalNAc to Ser/Thr residues of the mucin peptide, which is extended by other monosaccharides, comprising eight major core structures of mucins represented in **Figure 1.3**. Core 1 to 4 are the most common in intestinal mucins.

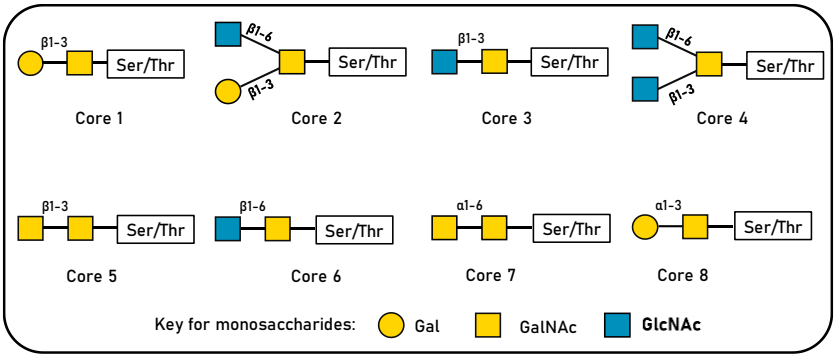


Figure 1.3 - Representation of major mucin glycan cores. Cores 1-4 are the most found in gastrointestinal mucins. Adapted from [38].

Core 3 (GlcNAc-β-1,3-GalNAc-α-Ser/Thr) structures are predominant in the small intestine while core 3 and 4 (GlcNAc-β-1,6-(GlcNAc-β1,3)-GalNAc-α-Ser/Thr) are mostly found in the colonic mucin glycans. Taking the example of Muc2 found in the colon, it was shown that it mainly contained the core 3 structure [39]. The mucin O-glycan backbone can be extended from these cores, and monosaccharides such as Gal and GlcNAc can be added by specific glycosyltransferases. In addition, Fuc, GalNAc and sialic acid can also be added, normally in terminal positions of the chain. Examples of terminal structures present in glycoproteins that are recognized by bacteria are represented in **Figure 1.4**. These include the Lewis antigens, the A, B, and O blood group antigens and polylactosamine sequences, comprising the I and small i antigens [40].

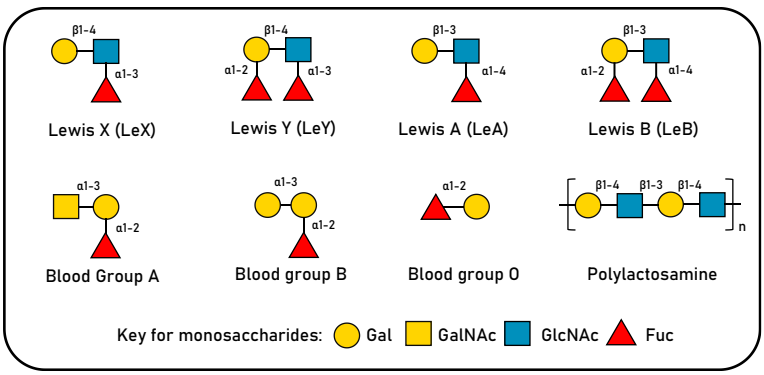


Figure 1.4- Terminal oligosaccharide structures that compose the Lewis antigens, the A, B and O blood group antigens and polylactosamine. Adapted from [40].

1.2.2. Interactions between gut microbiome and mucins

The GI tract is inhabitted by a multitude of bacteria species mainly belonging to the Firmicutes and Bacteroidetes phyla, also with fewer *Proteobacteria*, *Verrucomicrobia* and *Actinobacteria* strains [41]. As previously stated, the colon's epithelium is covered by a thick mucus bilayer. A study from Johansson and collaborators showed that the most outer mucus layer contains a higher percentage of bacteria, while the layer directly attached to the epithelium is practically absent of microorganisms [42]. It is believed that in healthy conditions, or in a high fiber diet scenario, the mucosa-associated bacteria are only present in the outer mucus layer, and not in direct contact with the gut epithelium. On the other hand, considering a low-fiber scenario, bacteria only depend on the mucins that cover our gut epithelium as a source of carbon, as no nutrient rich carbohydrates are present. This situation leads to the degradation, first of the mucus layer and then the epithelium, leading to possible inflamation and disease caused by increased susceptibility to pathogens [43]. These two scenarios are represented in **Figure 1.5** below.

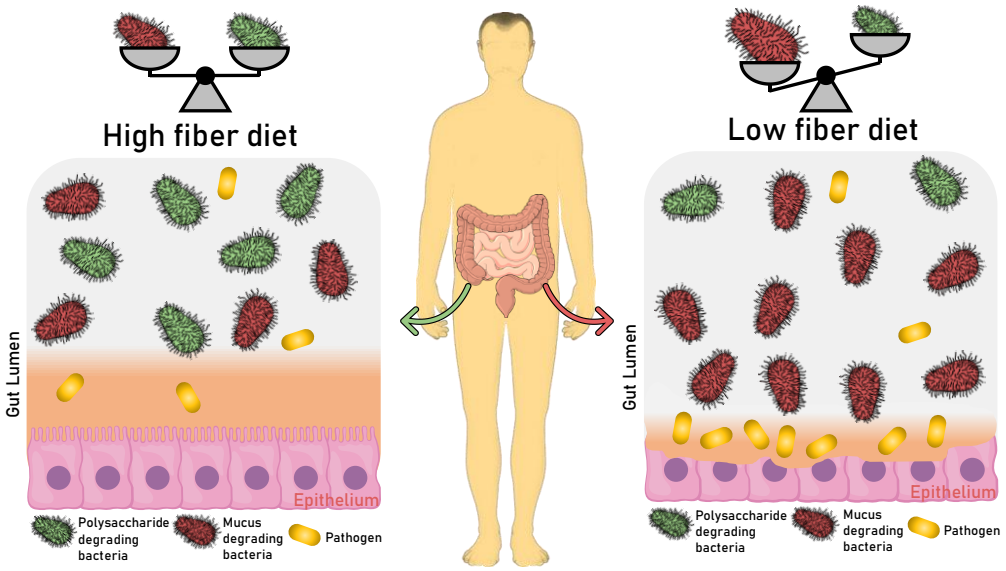


Figure 1.5- Illustration of how a fiber-deprived microbiota leads to degradation of intestinal mucus layer, increasing pathogen susceptibility. In healthy conditions (high fiber diet scenario), the mucosa-associated bacteria are only present in the outer mucus layer. In a low-fiber scenario, bacteria will mostly depend on the mucins that cover our gut epithelium as a source of carbon, leading to possible inflammation and disease caused by the action of pathogenic agents. Adapted from [43].

Mucin degradation has been associated to pathogenicity for a long period of time, mainly because the first mucin-degrading bacteria described, also denominated mucinolytic bacteria, were pathogens. However, that idea evolved and nowadays it is well understood that mucin degradation is part of a natural turn-over process that takes place months after birth [44].

Mucin degrading bacteria like the ones belonging to the Bacteroidetes phylum have been extensively studied, and a recent study showed that *Bacteroides thetaiotaomicron*, in particular the strain VPI-5482 was able to grow on fractions containing O-glycans purified from pig gastric

mucosa. The transcriptomic analyses showed the up-regulation of specific gene clusters that coded for enzymes responsible for mucin degradation [45]. Mutations on those specific gene clusters coding for mucing-degrading enzymes, led to colonizaton in germ free mice fed with a plant-derived glycan rich diet, but in a lower level when compared to the wild type strain. This was indicative that bacteria such as *B.thetaiotaomicron* relies also on host glycans, which are present on mucin glycoproteins, for colonization [45]. Another bacteria that demonstrated the abilty of mucin degradation was *Bacteroides fragilis*, which contained genes in its genome that encoded for proteins responsible for the binding, degradation and transport of host mucin O-glycans [46]. The anaerobe *Bifidobacteria*, very abundant in early life also has the ability of digesting host-mucins in addition to diet derived-carbohydrates [47]. Evidence of carbohydrate transport systems and other proteins involved in the metabolism of polysaccharides such as glycogen, pullulan, starch and mucins suport this feature. In a study by Hosking and colaborators [48], strains ATCC 35914 of *Bifidobacterium bifidum* and VIII-240 of *Bifidobacterium longum*, both from the Actinobacteria phylum, were isolated and possible mucin degrading activity was detected as both strains grew on a mucin rich fraction. More recently, the mucin utilization ability of *Bifidobacterium bifidum* was confirmed for several strains using human Muc2, among other mucins [38]. *Akkermansia* is another group of bacteria that have an important role in mucin degradation in the human gut. It is present in the gut of infants, adults and the elderly, and many strains possess the enzymes necessary for carbohydrate metabolism, and being able to degrade host mucin O-glycans [49]. All of the previosly mentioned bacteria have a cross-feeding trait, meaning that even if some strains are not able to express all enzymes necessary for polysaccharide degradation, the total intestinal flora at that location is capable of providing monosaccharides for all strains present [50] .

1.3. Bacterial enzyme toolbox and *Polysaccharide utilization loci* (PUL)

Complex polysaccharides comprise the largest known acess point to carbon for microorganisms. But these sources of carbon are not readily accessed and so bacteria need to produce a multitude of enzymes known as polysaccharide-degrading enzymes. The uptake of carbohydrate energy is directly related to the persistence of microorganisms in the most varied ecosystems, especially the gut where the competition for nutrients is at high stakes. The strategy developed by bacteria is the presence of certain enzymes that are responsible for the modification and degradation of polysaccharides. These enzymes are known as carbohydrate-active enzymes (CAZymes) and can be classified as glycoside hydrolases (GHs), polysaccharide lyases (PLs), carbohydrate esterases (CEs) and polysaccharide oxidases (POs). Enzymes like these can also contain modules designated carbohydrate-binding modules (CBMs) that are connected by a flexible linker [51].

The CAZy database organizes the enzymes in families based on their aminoacid sequence. This database to this date comprises 169 families of GHs, 114 families of GTs, 41 families of PLs, 18 families of CEs and 88 families of CBMs (www.CAZy.org) – reviewed on 05/02/2021.

1.3.1. How bacteria use glycans to thrive

The ability that bacteria possess not only to degrade and consume host glycans but also dietary fibers has been known for many years.

A unique mechanism developed by bacteria, and especially present in the genome of Bacteroidetes is the presence of clusters of co-localized and co-regulated genes that code for the machinery responsible for the detection, sequestration, digestion and transport of carbohydrates. These clusters are denominated *polysaccharide utilization loci* (PULs), a term that was first used in 2006 by Martens and colleagues [52]. Over the consecutive years, a growing number of PULs were discovered in several organims, including *B. thetaiotaomicron*, and information on these PULs is available at PUL Database (http://www.cazy.org/PULDB/).

These clusters of genes encode for certain CAZymes, such as glycoside hydrolases, polysaccharide lyases and carbohydrate esterases but also for cell surface glycan-binding proteins (SGBPs), such as Starch utilization system D protein (SusD), TonB-dependent transporters (TBDTs), such as Starch utilization system C protein (SusC) and carbohydrate-sensor/transcription regulators. Frequently, the complexity of PULs is proportional to the complexity of their designated substrate and can code for additional enzymes such as proteases, phosphatases and sulfatases. *Polysaccharide utilization loci* comprise the main strategy for bacteria such as *Bacteroides thetaiotaomicron* to acquire energy and nutrients, being directly linked to the constitution of microbial ecosystems in the gut [53]. The first hints about the existence of an organized system designed for complex carbohydrate sequestration and degradation were discovered in organisms from the Bacteroidetes phylum. These studies focused on dietary starch utilization in *Bacteroidetes thetaiotaomicron* and were performed by Abigail Salyers and colleagues in 1977[54] .

The first PUL was described in *Bacteroides thetaiotaomicron* by Abigail Salyers and colleagues in the 1980's. A total of eight genes responsible for coding proteins that are fundamental in the starch adherence to the cell surface and it's hydrolysis were discovered. This gene cluster was named starch utilization system (Sus) due to it's apparent function [55], [56]. A total of eight genes were identified (SusA,B,C,D,E,F,G and R) and to this date, the Sus of *Bacteroides thetaiotaomicron* continues to be genetically and biochemically studied and it still serves as a base model for the study of other PULs [11].

The model for this *Bacteroides thetaiotaomicron* Sus operon is represented below at **Figure 1.6**.

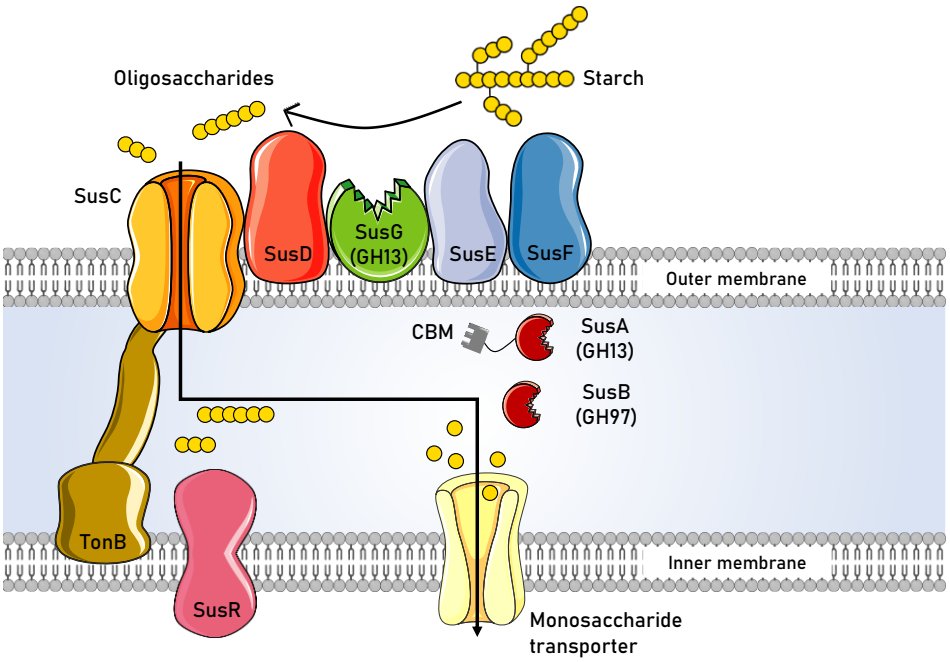


Figure 1.6- A model of the *Bacteroides thetaiotaomicron* Sus. A total of eight genes were identified (SusA,B,C,D,E,F,G and R), acting in concert for the degradation of starch polysaccharide. To this date the starch utilization system of *Bacteroides thetaiotaomicron* continues to be genetically and biochemically studied and it still serves as a base model for the study of other PULs. Adapted from [11].

1.3.1.1. Carbohydrate-binding modules (CBMs)

As previously stated, enzymes involved in the degradation of carbohydrates are often modular and contain non-catalytic modules attached by a linker, which are denominated carbohydrate-binding modules (CBMs). Initially, these modules were named cellulose-binding domains (CBDs), because when they were discovered cellulose appeared to be their primary ligand. Later, this designation was changed to CBM due to the knowledge that these modules exhibit specificity towards a variety of carbohydrates. These CBMs can be defined as sugar-binding proteins with a 30 to 200 amino acids sequence that fold into a discrete domain that is contained in a larger multi-modular enzyme, for example a glycoside hydrolase. Conventionally, the role of a CBM is to bind directly to the polysaccharide and direct the enzyme to undergo its catalytic activity, and consequently potentiating the efficiency of the multimodular carbohydrate active enzyme itself [57].

Generally, these CBMs are attached to enzymes that can degrade insoluble polysaccharides, and many CBMs have components of the plant cell wall as their main target, however there are CBM families that contain proteins that bind to starch and glycogen. CBMs can also be present in polysaccharide lyases, polysaccharide oxidases and glycosyltransferases [58]. Commensal bacteria, such as *B.thetaiotaomicron*, have also developed carbohydrate-active enzyme systems that are capable to interact with the complex mammalian glycans [59]. For example, CBMs from family 32 have a high specificity for Gal- and GalNAc-containing glycans

[60] and CBMs from family 40 have a high specificity for terminal sialic acid glycan sequences, normally associated with sialidases [61].

1.3.1.1.1. CBM classification and functional roles

In terms of classification, CBMs are classified in 88 different families according to the information available in the CAZy database. Each CBM is attributed to a specific family based on sequence comparison with already characterised CBMs, and also, each CBM can fall into one of 7 distinct fold families with the most common fold being the β-sandwich fold. For the CBMs with this type of fold the sugar binding site can be located on one face of the β-sheet [62] or within the variable loop region of the β-sheet [63]. However, there are CBMs within the β-sandwich fold family that have two binding sites: an example is the dual starch-binding site for CBM20 [64]. CBM fold families also include β-trefoil fold, OB fold, cysteine knot, the hevein/hevein-like and unique folds [57].

CBMs can be even further classified based on the mode of interaction with their specific ligand. There are three types of CBMs: type A, type B and type C, each one with their own characteristics [57], [58], [65]:

Type A CBMs can bind to crystalline surfaces of chitin and cellulose. Some example of type A CBMs are CBM from family 1, 2 and 3. The binding sites of type A CBMs are planar and rich in aromatic amino acids, which creates a relatively planar surface that can bind to the planar crystalline surface of chitin and cellulose.

Type B CBMs can bind internal sequences of the glycan and are classified as endo-type. CBMs from type B are the most abundant and their binding site takes the form of a groove or cleft that is capable to accommodate long sugar chains with at least four monosaccharides.

CBMs from **type C** bind terminal glycan sequences, at the reducing or non-reducing end, also classified as exo-type CBMs. Regarding their binding sites, they take the form of small pockets, much smaller than those of type B CBMs, only capable of recognizing short sugar ligands containing up to three monosaccharides.

Four main roles are attributed to CBMs: targeting, proximity, disruptive and adhesion roles. Regarding targeting and proximity roles, CBMs can target the enzyme to a particular region of the carbohydrate substrate and then increase the concentration of the enzyme near that same substrate, leading to a more rapid and efficient degradation. The disruptive role of CBMs is evidenced when we look at cellulose or starch molecules. Some CBMs have the ability of disrupting the structure of these polysaccharides, causing them to loosen and be more easily accessible for enzyme degradation. This effect was demonstrated in CBMs from family 20 from *Aspergillus niger* in the disruption of starch. Also, CBM41 family modules are thought to have a disruptive effect in the degradation of glycogen granules [66], [67]. Finally, the adhesion role has been demonstrated on a module from the CBM family 35. It occurs when a CBM has the ability

to adhere enzymes to the surface of a certain bacterial wall component while demonstrating a catalytic activity on an external neighbouring substrate. CBM35 modules showed to interact with sugars containing glucuronic acid in the cell wall of *Amycolatopsis orientalis*, while the catalytic part is focused on a neighbouring external chitosan [68].

1.3.1.1.2. The SusD proteins

The previously discussed starch utilization system from *B. thetaiotaomicron* is still the best studied PUL-encoded glycan uptake system and is often the canonical model for homologous PUL encoded proteins. One of the proteins that make up the starch utilization system is the SusD protein, which is a surface glycan-binding lipoprotein, that in concert with SusE, F and G aid in the task of glycan recognition and adhesion to the cell surface (**Figure 1.4**).

The SusD from the VPI-5482 strain of *B. thetaiotaomicron* has a role in starch-binding to the cell and is noticeably larger than any carbohydrate-binding protein (molecular weight around 65 kDa) [53]. A study from Martens and colleagues that performed bacterial genome sequencing revealed the presence of homologues of SusD and SusC across all PULs of Bacteroidetes present in the gut [52].

The characterised SusD crystal structures, showed primarily the presence of α-helices, with a single binding site for starch molecules [69]. SusD is anchored to the outer membrane by an N-terminal cysteine that is lipidated. This lipidated cysteine is preceded by a 16-residue flexible linker that projects the protein above the outer membrane surface. The conservation of four helix-turn-helix motifs is by far the main characteristic of SusD and all of its homologues. These motifs are known as tetratricopeptide repeats (TPR), and together they form a right-handed superhelix along the protein [70]. Despite being abundant in nature and involved in protein-protein interactions by serving as a site for those interactions, TPR motifs are practically invariable across all SusD-like proteins. The binding site for starch in SusD can be characterized as a shallow pocket that contains an arc of aromatic amino acids in the shape of an amylose helix [69].

A study from Koropatkin, Martens and colleagues focused on *Bt*1043-SusD, an outer membrane lipoprotein from PUL13 that is hypothesized to be involved in targeting mucin *O*-glycans [71]. When compared to the starch-binding protein SusD from *B. thetaiotaomicron*, *Bt*1043-SusD is considered a structural homologue due to the presence of the same four TPR repeats. A structural study showed that in addition to the presence of the TPR repeats, a conserved α-helical fold was observed, suggesting a homologous function in glycan utilization. The *Bt*1043-SusD was predicted to target host mucin O-glycans because its expression was up regulated (approximately 190 fold) in mice that undergo a poor glycan diet [10]. The glycan-binding site found on *Bt*1043-SusD was also homologous to the one seen in SusD, and interactions with a

GlcNAc residue of a LacNAc sequence identified through X-ray crystallography also pointed that *Bt*1043-SusD may recognize mucin O-glycans [71].

Despite all strong evidence taken from the above studies, it is important to state that the knowledge involving the interaction between SusD/SusD-like proteins with host glycans is relatively scarce and more work is needed in order to fully understand this intricate concept.

1.4. Overview of the methods used to study protein-carbohydrate interaction

1.4.1. Carbohydrate microarrays

Carbohydrate microarrays, also known as glycan microarrays, is a technology that proved to be a powerful high throughput method to array extensive glycan libraries and consequently probe these with diverse glycan recognition systems to identify glycan binding proteins (GBPs) and unravel their biological roles [72]. It is an important tool that also allows the identification of glycans that are directly involved in different biological contexts. This technique has gained interest over the years and since their introduction, the applications of glycan microarrays have grown at an exponential rate [73][74].

One advantage of using this technology is the use of minimal amounts of sample, enabling detection in the range of femtomoles of probe and testing of a wide variety of carbohydrate sequences that are immobilized on solid matrices, and therefore retrieving important information about the protein binding specificity. Glycan microarrays are also extremely sensitive as can detect multivalent low affinity interactions of a protein to a clustered immobilized glycan. On another point of view, this immobilization of various glycan sequences is also advantageous, as it is a strategy to mimic at some extent glycan presentation at the surface of our cells. There can be two types of carbohydrate microarrays: polysaccharide microarrays and oligosaccharide microarrays [75], however the focus on this section will be on oligosaccharide microarrays.

For construction of a sequence-defined carbohydrate microarray, the oligosaccharides need to be purified and structurally characterized [73]. The glycans can be naturally isolated or chemically or chemoenzymatically synthesized. Although glycans from natural sources are advantageous for achieving a desired glycan structural diversity and extrapolation of meaningful results, their isolation and structure assignment can be a challenge. The yield of natural isolation could be lower, as it requires a deconvolution of the heterogeneous isolated mixtures [76]. Probes that are chemically synthesized have the advantage of being much more accessible in large quantities, since carbohydrates can be synthesized in a relatively pure form in large quantities. Ideally a glycan microarray platform will comprise oligosaccharides obtained from both strategies. The structural diversity in current glycan microarray comprise mammalian glycans present on glycoproteins or glycolipids, glycosaminoglycans, and also bacterial-, fungal- or plant-derived oligosaccharides [73].

Prior to printing and immobilization on a solid surface for construction of the microarrays (e.g. glass slides or gold/nitrocellulose membranes or coated slides), the oligosaccharides need to be prepared as suitable probes due to their hydrophilic nature. This normally requires an initial derivatization step. One approach used to achieve oligosaccharide derivatization is through a process designated reductive amination of natural or synthesized oligosaccharides to a lipid [77] or to BSA [78] to originate neoglycolipid (NGL) or neoglycoprotein (NGP) probes, respectively. These probes have amphipathic properties suitable for non-covalent immobilization through adsorption of the lipid or BSA onto a solid surface such as nitrocellulose. Covalent immobilization comprises adding a functional group at the reducing end of the oligosaccharide suitable to react with a reactive group on the surface of the slide. An example of covalent immobilization involves the covalent reaction of synthetic oligosaccharides derivatized at the reducing end to an amino-terminating linker onto N-hydroxysuccinimide (NHS)-functionalized glass slides [79].

1.4.1.1. Neoglycolipid-based oligosaccharide technology

The neoglycolipid principle was first introduced in 1985 by Feizi and colleagues to address the study of antigenicities and receptor functions of carbohydrates [80]. Feizi and colleagues, adapted the NGL technology to achieve the first microarray system for complex oligosaccharides [81]. In this system, the several oligosaccharides are conjugated via reductive amination with the aminolipid 1,2-dihexadecyl-sn-glycero-3-phosphoethanolamine (DHPE) using microscale lipid conjugation, however oxime ligation can also be used in the conjugation process.

The NGL-based microarrays (**Figure 1.7**) include NGLs (prepared from natural or chemically synthesized oligosaccharides) and glycolipids, which can be from natural sources or synthetically synthesized. These probes are then robotically printed in a liposome formulation in the presence of carrier lipids, approach that contributes to a level of flexibility and movement of the oligosaccharide which can be important for some recognition systems [82]. The NGL based technology has the advantage of expanding the library of probes through the application of the designer microarray approach, which is a term that is applied to a microarray of oligosaccharide probes generated from ligand bearing glycomes [74]. These oligosaccharides can be posteriorly revealed to be isolated and individually characterized.

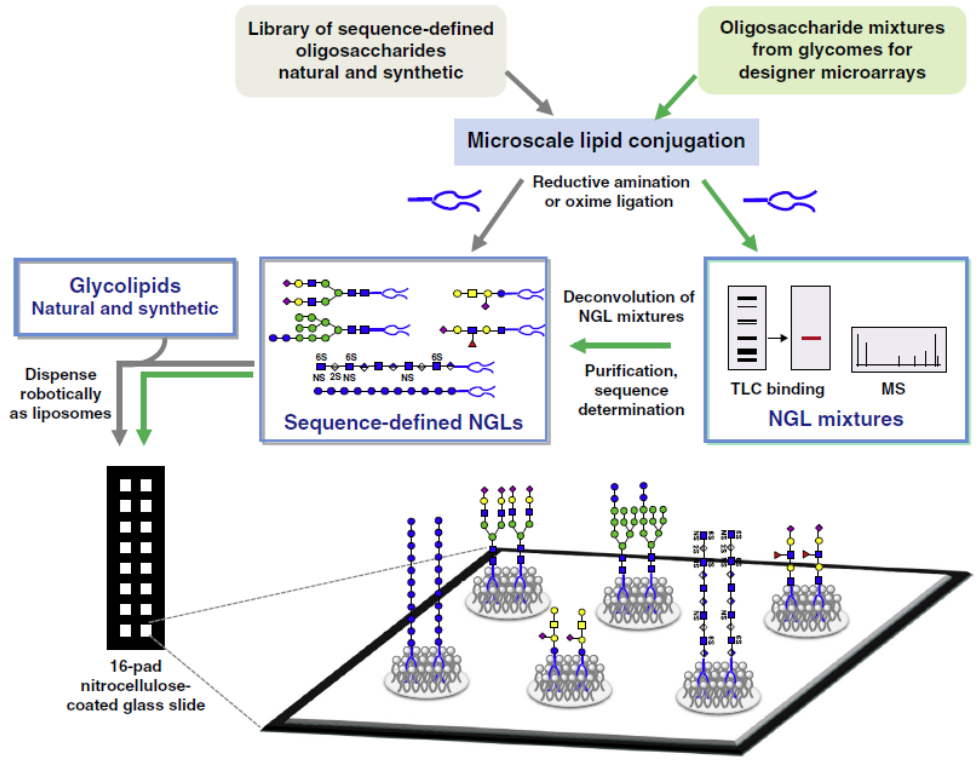


Figure 1.7- Scheme showing a representation of the NGL-based glycan microarray technology. Taken from [73].

1.4.2. MicroScale Thermophoresis

MicroScale Thermophoresis (MST) is a technique that takes advantage of the biophysical process of thermophoresis to study biomolecular interactions. Thermophoresis is a term to define the movement of molecules in a temperature gradient. This movement depends on several aspects of the targets such as the molecule size and conformation, as well as its charge and hydration shell [83]. MST is an extremely sensitive technique that can measure dissociation constants (K_D) that are down to picomolar levels (10^{-12} M). From an MST experiment, conclusions about binding modes and stoichiometries (n) can also be retrieved.

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 [83].

For an MST experiment to work, several aspects must be considered but mainly heat has to be delivered to the capillary containing the sample and a fluorescent molecule has to be present in order to analyse the thermophoretic movement of particles (**Figure 1.8**). Heat is produced by an infrared laser (IR) and the visualization of the thermophoretic movement of particles is achieved by a specific fluorescent dye (e.g. NT-647 dye), a fusion protein or by taking advantage of the protein tryptophan content [84].

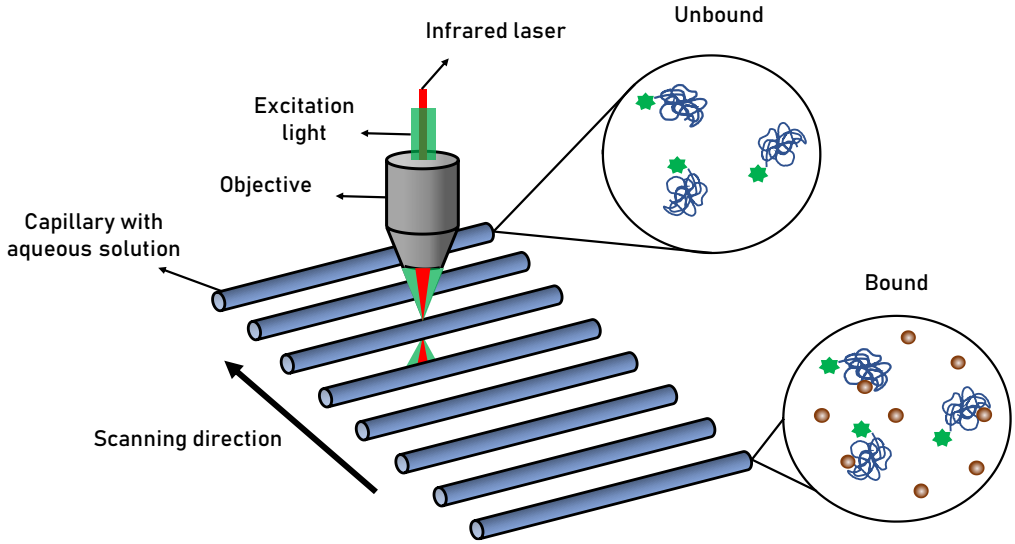


Figure 1.8- Representation of a typical MST experiment setup. An MST experiment is done using capillaries. The fluorescence within that capillary is excited and posteriorly detected through an objective and an IR laser is used to heat the sample. Adapted from [83].

1.4.3. X-ray crystallography

Another method that can be used to study the interaction between protein and carbohydrates, more importantly determine the three-dimensional structure of the protein-carbohydrate complex is X-ray crystallography. The possibility of "viewing" the inner structure of single crystals using the diffraction of X-rays was discovered more than a 100 years ago, being haemoglobin and myoglobin, the first protein crystal structures elucidated, in the 60's decade, by Max Perutz and John Kendrew. X-ray crystallography is currently the most used method for 3D structure determination of biological macromolecules at atomic resolution. This technique can also provide information about viruses, immune complexes and nucleic acid complexes, which increases the appeal of this structural method [85], [86]. The aim of crystallography is to obtain a 3D molecular structure from the analysis of a single crystal. A high concentration and purified sample containing our molecule of interest is exposed to a monochromatic X-ray beam resulting in a diffraction pattern containing a multitude of spots that can help us retrieve important information about the crystal symmetry and size of the asymmetric unit, the repeating unit that forms the crystal [86]. Obtaining a 3D structure of a protein is not a straightforward process and involves multiple steps, more specifically protein expression and purification, crystallization assays for that specific protein, X-ray diffraction experiment (data collection and processing), 3D structure solution (phasing) and electron density map calculation, and finally model building, refinement and validation (**Figure 1.9**).

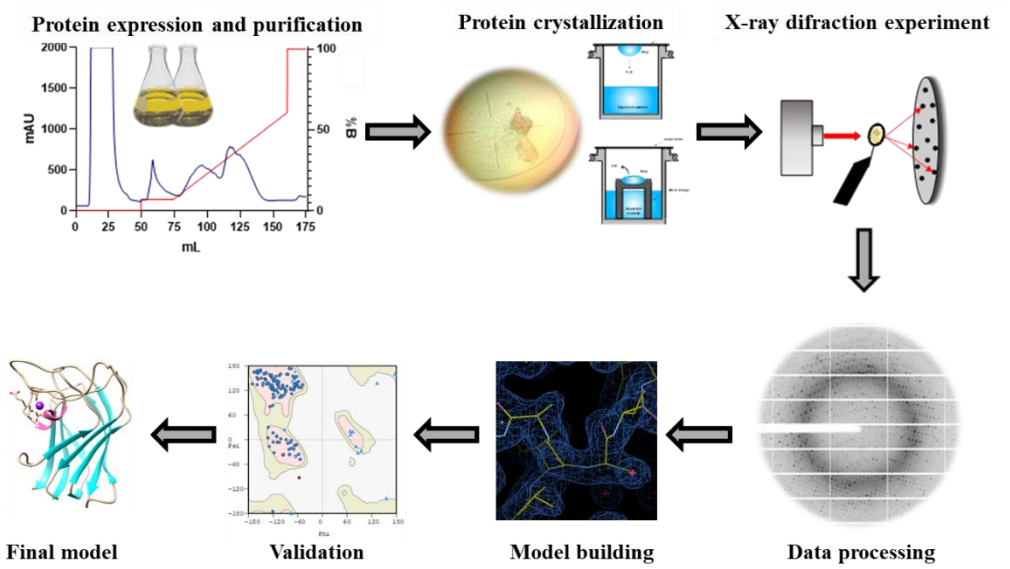


Figure 1.9- Representation of the steps to obtain the three-dimensional structure of a protein using X-ray crystallography. Obtaining the 3D structure of a protein is not a straightforward process and involves multiple steps, more specifically protein expression and purification, crystallization assays for that specific protein (isolated or in complex with ligands of interest), X-ray diffraction experiment (data collection and processing), 3D structure solution and electron density maps calculation, and finally model building, refinement and validation.

The initial step, protein expression and purification gain an increased importance because the entirety of the experiment depends on the availability of a reliable source of purified protein that will yield high quality crystals. The growth of good quality protein crystals is the rate limiting step and is the least understood.

The principle of crystallization relies on the concept of having a high concentration of a solution containing our molecule of interest and force it to come out of solution. This principle is also valid for salts, reason why sometimes the crystals obtained are salt instead of our protein. Sometimes this step occurs too fast, leading to the precipitation of the sample, but under certain circumstances crystals will grow.

In the vapor diffusion method, crystal growth occurs in a drop, containing our protein solution mixed with precipitant solution, equilibrated against a reservoir containing only the precipitant solution. This process involves two steps: nucleation and growth. In the beginning, the precipitant compound is present at a lower concentration in the drop and the protein is in an undersaturation state. In order for the concentration of precipitant agent in the drop to reach equilibrium with the reservoir solution, water vapour leaves the drop leading to an increase of saturation of the protein sample in the drop. When the concentration of precipitant agent in the drop and reservoir is the same equilibration has been achieved. Nucleation is the most difficult problem to understand because molecules pass from a totally disordered state to a highly ordered state. Nucleation is thought to occur through the formation of small crystalline intermediates from which, later the crystal will grow at a very slow rate [86].

The crystal formation conditions can be observed in the phase diagram represented below (**Figure 1.10A**), where the crystal growth arises in the metastable zone. Next to the phase diagram, two methods of vapor diffusion are represented: sitting drop and hanging drop. Both methods rely on the vapor diffusion principle discussed previously, with the only difference being that, in the hanging drop method, the drop is set upside down on a cover slide while in the sitting drop it is set on top of a horizontal support (micro bridge in **Figure 1.10B**).

Sometimes the process of obtaining protein crystals prove to be very difficult, mainly because crystallization depends on multiple variables like the pH, salt concentration, temperature, type of precipitant agent, the crystallization technique or even the need of certain additives. That is why obtaining protein crystals is a trial and error process. Techniques have been developed over the years in order to overcome this wide arrange of variables like the use of commercially available screenings that contain several solutions that test in a single assay different crystallization conditions.

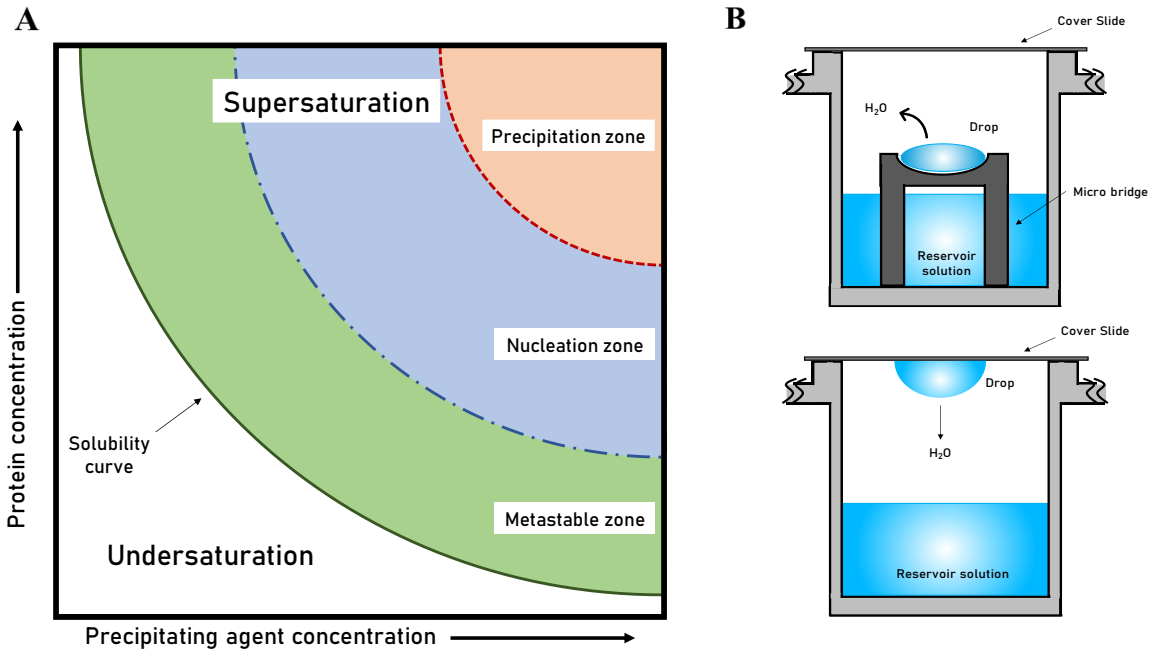


Figure 1.10- Phase diagram and vapour diffusion methods representation. A- Phase diagram for protein crystallization showing crystallization drop solution properties depending on both protein and precipitant agent concentration. B- Sitting and hanging drop vapour diffusion techniques for protein crystallization. Adapted from [85].

Once crystallization is succeeded, a single crystal is chosen for data collection. This data collection step can be done in house, in a diffractometer or in a synchrotron facility. In this stage, a monochromatic X-ray beam is collimated and forced to pass through the crystal. Multiple photon reflections are detected in the form of spots, with a photon intensity associated to each spot. From this experience we can calculate the wave amplitude of each reflection and also the h, k and l indices for the position of each reflection in the recorded diffraction patterns [87]. We

cannot directly calculate the phase angle, giving rise to a common problem in crystallography denominated the phase problem [88].

There are several methods to solve the phase problem namely MIR (Multiple Isomorphous Replacement), MAD (Multiple Wavelength Anomalous Dispersion), SAD (Single Wavelength Anomalous Dispersion) and finally MR (Molecular Replacement).

Molecular Replacement is a very popular method to solve the phase problem in crystallography requiring a previously solved 3D structure in which the primary sequence is identical to our target sequence by at least 30 %. Only if the identity is 30% or higher, the known structure is considered to be a good model for molecular replacement. From here, if the process is done correctly, we obtain a first electron density map contouring the position of every atom in the crystal structure, which, after iterative cycles of model building and refinements, could lead to the final 3D structure of our target macromolecule [88].

1.5. Objectives

Utilization of glycans that cover our intestinal wall by commensal or pathogenic microorganisms has been a topic of discussion throughout the years. To deepen the knowledge in this field it is important to assess the ligand specificities of the proteins responsible for the targeting and recognition of those glycans. With this purpose, a combination of techniques including glycan microarrays and X-ray crystallography can be used, alongside other complementary methods such as MST. The main aim of this thesis is the structural characterization of proteins from the commensal carbohydrate-degrading bacteria *Bacteroides thetaiotaomicron*, more specifically CBMs and SusD-like proteins, which are potentially involved in the recognition and degradation of host glycans. To accomplish this, the specific objectives of this thesis will be the following:

1) To assess the best conditions for recombinant heterologous expression of eight selected family 32 CBMs;
2) To perform small-scale expression and purification of selected family 32 CBMs and SusD-like proteins with the main purpose of setting-conditions for large-scale expression, purification and evaluation of protein stability;
3) To perform large-scale expression and purification of selected proteins for structural characterization using X-ray crystallography;
4) To determine and characterize the three dimensional structure of a selected protein using X-ray crystallography.

Chapter 2 - Expression, purification and stability analysis of *B. thetaiotaomicron* family 32 CBMs and SusD-like proteins

2.1. Introductory remarks

The main objectives of the work that is presented in this chapter were the following: 1) optimization of conditions for heterologous recombinant expression in *Escherichia coli* of eight different *B. thetaiotaomicron* family 32 CBMs and 2) assess the stability of family 32 CBMs and SusD-like proteins in different buffers through the thermal shift assay (TSA).

Initially, small scale expression tests were carried out, where several conditions were tested (bacterial expression cell lines, bacteria culture growth time and period, concentration of IPTG etc.). The conditions for the expression of SusD proteins were already established. After this step, small scale expression and purification of two family 32 CBMs and two SusD-like proteins was performed following a large-scale protein production protocol using the best conditions identified.

2.2.Materials and methods

The protocol for cloning is property from NZYTech®. CBMs were cloned, expressed and purified by a confidential small-scale high-throughput method using a derivative vector of pET24a, that confers kanamycin resistance and code for an N-terminal hexa-histidine tag. The information about the recombinant protein sequence, protein identification, family, molecular weight and isoelectric point for each CBM is presented in **supplementary table 1**.

2.2.1. Transformation

The transformation consists in a gene transfer process in which some bacteria take up external genetic material. Two types of *E. coli* bacterial cells were transformed using the same protocol: DH5α cells for amplification of the plasmidic DNA and BL21 cells for protein expression. The entire process of transformation was executed in sterile conditions and the first step consisted in the addition of 2µL of DNA to 50 µL of DH5α competent cells. The cells were then incubated on ice for 30 minutes and then submitted to a heat shock at 42 ^{o}C on a heatblock for 1 minute, followed by incubation on ice for 10 minutes. This sudden increase in temperature will allow the formation of pores in the plasma membrane of the bacteria, allowing plasmid DNA to enter the bacterial cell. The cells were resuspended in 1mL of LB (Luria Bertani) medium (recipe in **Supplementary information 1**) and incubated in a shaker (Orbital Shaker-Incubator ES-20 from Grant-bio), at 37^{o}C and 200 rpm, for approximately 1 hour. After incubation ceased, the cells were centrifuged for 1 minute at 8600 rpm (5804 R centrifuge from Eppendorf) and 1mL of the supernatant was discarded. The pellet was resuspended in LB medium and spread in LB-agar plates containing the antibiotic kanamycin at a concentration of 50 µg/ml.

2.2.2. DNA amplification, isolation and sequencing

The LB-agar plates that resulted from DNA transformation were incubated overnight at 37 ^{o}C (200 rpm). After incubation, the plates were stored at 4 ^{o}C. To carry out the DNA amplification, 10 mL of LB medium containing 50 µg/mL kanamycin was used to inoculate 1 pricked colony taken from LB-agar plates. To grow the bacterial cells containing the plasmid DNA, the cell resuspension was incubated overnight at 37 ^{o}C at 200 rpm (Orbital Shaker-Incubator ES-20, Grant bio).

Following this incubation step, a protocol for extraction and isolation of DNA was performed, more specifically the miniprep protocol from NZYTech® (NZYMiniprep, MB01002) due to its application in rapid and small-scale isolation of highly pure plasmid DNA from *E.coli*. The protocol involved the following steps in which spin columns were used: harvesting of the cell pellet, discarding the supernatant, followed by cell lysis and clarification of lysate in which the mixture is centrifuged for 5-10 minutes in spin columns. The supernatant from this last step, containing the DNA, is then centrifuged to bind the DNA to the silica membrane of the column. The membrane is then washed and dried before the final step which is the elution of highly pure DNA and storage at -20 ^{o}C. The complete protocol from NZYTech® is available in **supplementary information 2**.

After extraction and isolation, the plasmidic DNA was sequenced (STAB VIDA, FCT-NOVA) in order to assess if the DNA sequence of the two CBMs was correct or if there were any possible mutations present. The primers that were used were the T7 (forward primer) and the pET24a (reverse primer), and both sequences are represented in **Table 2.1.**

Table 2.1- Primer sequences used to confirm the DNA sequence of the 8 CBM32s

Primer	Sequence
T7 forward	TAATACGACTCACTATAGGG
pET24a reverse	TAATACGACTCACTATAGGG

2.2.3. Small and large scale protein expression

After applying the transformation protocol previously described in section 2.2.1, one colony was pricked and pre-inoculated in 10 mL of LB medium containing the antibiotic kanamycin at a final concentration of 50 µg/mL. This pre-inoculum was then incubated at 37 ^{o}C and 200 rpm at orbital shaker (Orbital Shaker-Incubator ES-20, Grant bio). After overnight incubation, approximately 1 ml of the pre-inoculum was added to 40 mL of previously autoclaved LB medium (or 600 mL in the case of large scale expression), containing 50 µg/mL of kanamycin.

All cell cultures were incubated at a temperature of 37 ^{o}C and 200 rpm (IS-971R refrigerated shaker from Lab. Companion) until an optical density (OD_{600}) of 0.6 was reached, or approximately 3 hours. When all cell cultures reached the specific optical density, the expression

inducer Isopropyl β-d-1-thiogalactopyranoside (IPTG) was added to the medium at a final concentration of 1mM and the expression was induced for at least 16 hours. After analysis of the SDS-PAGE results, large scale protein expression was performed using the condition that generated a larger soluble protein fraction in the small scale tests.

2.2.4. Cell harvesting

The cells were harvested by a centrifugation step at 5000x*g* for 30 minutes (JA-10 rotor, Avanti J-26 XPI from Beckman Coulter). In the case of small scale expression, the cells were centrifuged at 5000x*g* for 30 min (5804 R centrifuge from Eppendorf).

After centrifugation, the supernatant was discarded and the remaining *pellet* was re-suspended in the lysis buffer (10 mL per gram of *pellet*) containing 50mM of 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid (HEPES) pH 7.5, 1M NaCl, 10 mM imidazole, 5mM $CaCl_2$, 5mM $MgCl_2$.

2.2.5. Lysis of pre-harvested cells

To perform the lysis of the bacterial cells and recover the soluble proteins, protease inhibitor tablets, DNAse (final concentration of 10µg/mL) and 5mM $MgCl_2$ were added to the previously harvested cells resuspended in the lysis buffer. The process of disrupting the cells was performed by sonication, using a sonicator (UP100H, from Hielsher) and applying 6 cycles of 1 minute, for larger volumes (e.g. 30 ml), and 4 cycles of 10 seconds, for smaller volumes (e.g. 1 ml). This procedure is performed on ice to prevent sample overheating. After this step, the solution was clarified by centrifuging the sample for 30 minutes at 10000x*g* (4 ^{o}C). The pellet was discarded, and the soluble fraction was stored on ice for immediate purification.

2.2.6. CBM purification through affinity chromatography

One of the most used methods in order to achieve recombinant protein purification, including CBMs, is a technique denominated immobilized metal affinity chromatography, also known as IMAC. IMAC takes advantage of the use of several chelating agents that are entrapped on the surface of the chromatographic resin supports to entrap metal ions such as nickel (II), which have affinity for some amino acids that may be exposed on the surface of proteins, for example histidine or cysteine residues. These amino acids not only have affinity for Ni^{2+}, but also for other divalent metal ions such as Zn^{2+}, Cu^{2+} and Co^{2+} [89].

2.2.6.1. Nickel-affinity chromatography purification using gravity

In order to perform small scale protein purification, His *GraviTrap*^{TM} immobilized nickel (II) columns from GE-Healthcare were used. These columns facilitate the purification of proteins

that contain a Histidine tag (His-tag) using the force of gravity forcing the passage of the sample through the column resin without the use of a more complex system like the ÄKTA START. For this purification, three buffers containing three different imidazole concentrations were used that were named A, B and C: Buffer A- 50mM HEPES, 1M NaCl, 5mM $CaCl_2$, 5mM $MgCl_2$, and 10mM of imidazole with a pH of 7.5; Buffer B - 50mM HEPES, 1M NaCl, 5mM $CaCl_2$, 5mM $MgCl_2$ and 60mM of imidazole with a pH of 7.5; Buffer C - 50mM HEPES, 1M NaCl, 5mM $CaCl_2$, 5mM $MgCl_2$ and 300mM of imidazole with a pH of 7.5. This purification technique is divided into four major steps: equilibration, sample application, washing and finally, elution. First, the column was equilibrated with approximately 10 mL of Milli-Q water and then the protein lysate was loaded into the column. After this step, the column was washed with 10mL of buffer A (10mM imidazole) with the objective of removing non-specifically adsorbed *E.coli* proteins. After this step, 10 mL of Buffer B (60mM imidazole) were loaded into the column with the intention to wash loosely adsorbed proteins to the resin. To elute the His-tagged protein of interest, which should bind with high affinity to the resin through the interaction of the His-tag with the nickel, 6mL of Buffer C that contains a higher concentration of imidazole (300 mM) were used. Multiple fractions were collected: a 3 mL fraction and afterwards three 1 mL fractions for post analysis by SDS-PAGE and native PAGE electrophoresis.

2.2.6.2. Nickel-affinity chromatography using automated flow-control system

As mentioned in the previous section, in order to achieve a large scale CBM purification other methods must be used and one of these methods is using high-load His Trap™ immobilised nickel (II) columns, that are coupled with chromatography flow-control systems (e.g. ÄKTA START), provides a higher degree of purity and optimal separation in the purification process. Both the columns and chromatography systems were from GE-Healthcare and the UNICORN™ start 1.0 control software was used.

The protocol suggests the use of two different buffers to perform the buffer gradient: Buffer A- 50mM HEPES, 1M NaCl, 5mM $CaCl_2$, 5mM $MgCl_2$, and 10mM of imidazole with a pH of 7.5; Buffer B- 50mM HEPES, 1M NaCl, 5mM $CaCl_2$, 5mM $MgCl_2$, and 300mM of imidazole with a pH of 7.5.

First, the 5 mL column was washed with Milli-Q water (approximately 50 mL) and was equilibrated with 20 mL of the buffer A solution. After filtration, the cell extract was loaded into the column and a washing step with 15% of buffer B was performed to remove loosely bound proteins that may have low affinity for nickel. The next step was the elution step with the application of an imidazole gradient, which varied from 15% to 100% of Buffer B. The purification fractions were collected for post analysis by SDS-PAGE or native-PAGE electrophoresis, as appropriate. The CBM containing fractions were then concentrated on Vivaspin® 10KD (GE Healthcare) concentrators for structural studies.

2.2.7 Protein analysis using polyacrylamide gel electrophoresis

Gel electrophoresis is probably one of the most used and simple method to analyse proteins. There are two well-known types of polyacrylamide gel electrophoresis, which were also performed during this work: SDS-PAGE and native PAGE.

SDS-PAGE is a method that uses an anionic detergent, the sodium-dodecyl sulphate (SDS), to disrupt the protein structure, causing it to unfold into a linear chain. SDS also coats the protein's surface with a uniform negative charge. A reducing agent is also used, more specifically β-mercaptoethanol, which also contributes to the protein's denaturation by reducing any disulphide bonds that might be present. Due to all this, the protein movement along the polyacrylamide gel will depend only on its molecular weight. All SDS-PAGE gels were used with a resolving gel containing 10% acrylamide and a stacking gel containing 4% acrylamide. All samples were prepared to a final volume of 15 µL. These samples contained the protein prepared in the sample buffer constituted of: 4x Tris-HCl pH 6.5, 10 % (w/v) SDS, 0.6 M DTT, 0.012 % (v/v) bromophenol blue, 30 % (w/v) glycerol. After preparation, the samples were submitted to a boiling step to promote denaturation in a heatblock for 5 minutes (100 ^{o}C). After this boiling step, each sample was distributed to each well. The run parameters for all SDS-PAGE gels were fixed to a voltage of 100 V.

In the case of native PAGE, the proteins are analysed in non-denaturing conditions, allowing for the size of the protein and their charge to affect the mobility. Both size and charge will depend on many factors such as the protein amino acid sequence, the protein's isoelectric point and the pH during electrophoresis. All native gels contained an acrylamide percentage of 12.5% and the sample buffer used was constituted of 4x Tris-HCl pH 6.5, 0.012 % (v/v) bromophenol blue, 30 % (w/v) glycerol. When compared to the sample buffer used for SDS-PAGE, the only constituents that are not present are the SDS and DTT, since native conditions are present. The running buffer that was used was 1.5 M Tris-HCl at a pH of 8.8 and the voltage was fixed to 100 V for each run.

2.2.8 Protein thermal shift assay (TSA)

Protein thermal shift assay (TSA) is nowadays widely used in order to determine the most suitable conditions for a protein in order to achieve maximum stabilization. Conventional TSAs are done in a certain way in which the target proteins in the presence of an interacting ligand are subjected to a thermal gradient under specific conditions. One of the most common methods to measure protein thermal shifts is thermofluor, in which specialized fluorogenic dyes are used [90].

The protocol for TSA involves several steps: first, 10µl of buffer screen is pipetted into each well, followed by 2µL of protein at a concentration of 40µM. Next, 3µL of dye (SYPRO® Orange from ThermoFischer) diluted from 1000x to 8x is added to each well followed by 5 µL

of the protein buffer (concentration of 4x). The final step is to seal the plate, centrifuge for 1 min at 1000 rpm and incubate for 30 min away from the light.

The increase in temperature provided by the temperature gradient mentioned earlier will cause protein denaturation and unfolding. The use of the dye is very advantageous, mainly because its fluorescence increases proportionally to the protein hydrophobicity. This means that if the protein gets denatured, the hydrophobic residues that normally would face the inside of the protein, will now be exposed to the dye and fluorescence is emitted. The monitoring of this fluorescence emission is done with the use of a Real Time PCR StepOnePlus™ from Applied Biosystems. This instrument will analyse the data received from each well of the 96 well plate used for these assays. Each well contains a different candidate stabilization condition to be assayed. After data collection, the derivative-normalized melting curves for the protein in different conditions can be plotted and from each curve a melting temperature (Tm) can be estimated.

2.3 Results and Discussion

2.3.1. Expression tests of family 32 CBMs

The optimization of expression was attempted for eight different family 32 CBMs: *Bt*1775, *Bt*2194, *Bt*3014, *Bt*3592, *Bt*4132, *Bt*4245, *Bt*4270_D and *Bt*4295_D. A control comprising the Bt3015-C family 32 CBM was also used since its expression had already been achieved in our group. For these small-scale expression tests, the protocol for IPTG induction was used varying several conditions, including the E. coli expression cell strain, IPTG concentration, culture growth temperature and period and also the incubation temperature and period, which are described in Table 2.2.

Table 2.2- Optimization conditions tested for the small-scale expression of family 32 CBMs

Optimization condition	Cell strain	[IPTG]	Incubation time and temperature after induction
A	BL21 (DE3)	1mM	37 ºC for 3h/4.5h
B	*Tuner* (*E.coli*)	0.25 mM 0.5 mM 0.75 mM 1 mM	16 ºC for 16h
C	*Rosetta* (*E.coli*)	1 mM	16 ºC for 16h

The results obtained with the different expression conditions for the family 32 CBMs were assessed by SDS-PAGE electrophoresis. Both soluble and insoluble fractions for each CBM were analysed. The results obtained for optimization condition A are shown in **Figure 2.1**, where two post induction incubation periods were tested (3 hours and 4.5 hours).

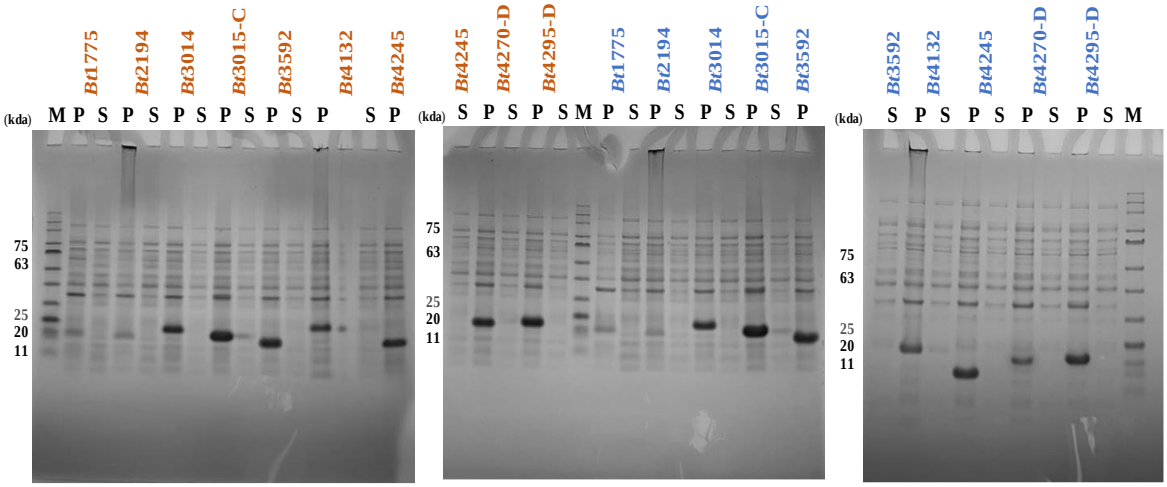


Figure 2.1 - Protein expression levels of the 8 CBMs tested using the E. coli BL21 strain and an IPTG-induction protocol. Bt3015-C was included as the control. The CBMs are identified by protein ID. The expression of CBMs was tested using the *E. coli* BL21 strain grown in LB medium, inducing the expression with 1mM IPTG at 37ºC and at two incubation periods: 3h and 4.5h. The orange and blue colour represent 3h and 4.5h respectively. SDS-PAGE (10% acrylamide) gels were run at a fixed voltage of 100 V. P-Pellet (insoluble fraction); S-supernatant (soluble fraction); M-Marker II from NZYTech®.

As we can see by analysing the results obtained, no significant differences were noticeable between the two incubation periods tested. The majority of the CBMs expression was observed in the insoluble fraction, independent from the incubation period.

Taking this into consideration, several variables were changed in attempt B, in an attempt to increase soluble expression of the CBMs. The *E.coli Tuner* strain cells were used and different concentrations of IPTG were tested (0.25 mM, 0.5 mM, 0.75 mM and 1 mM). The induction temperature was altered from 37 ºC to 16 ºC and as in the previous attempt, insoluble and soluble fractions were analysed through SDS-PAGE electrophoresis. The results are shown in **Figure 2.2**.

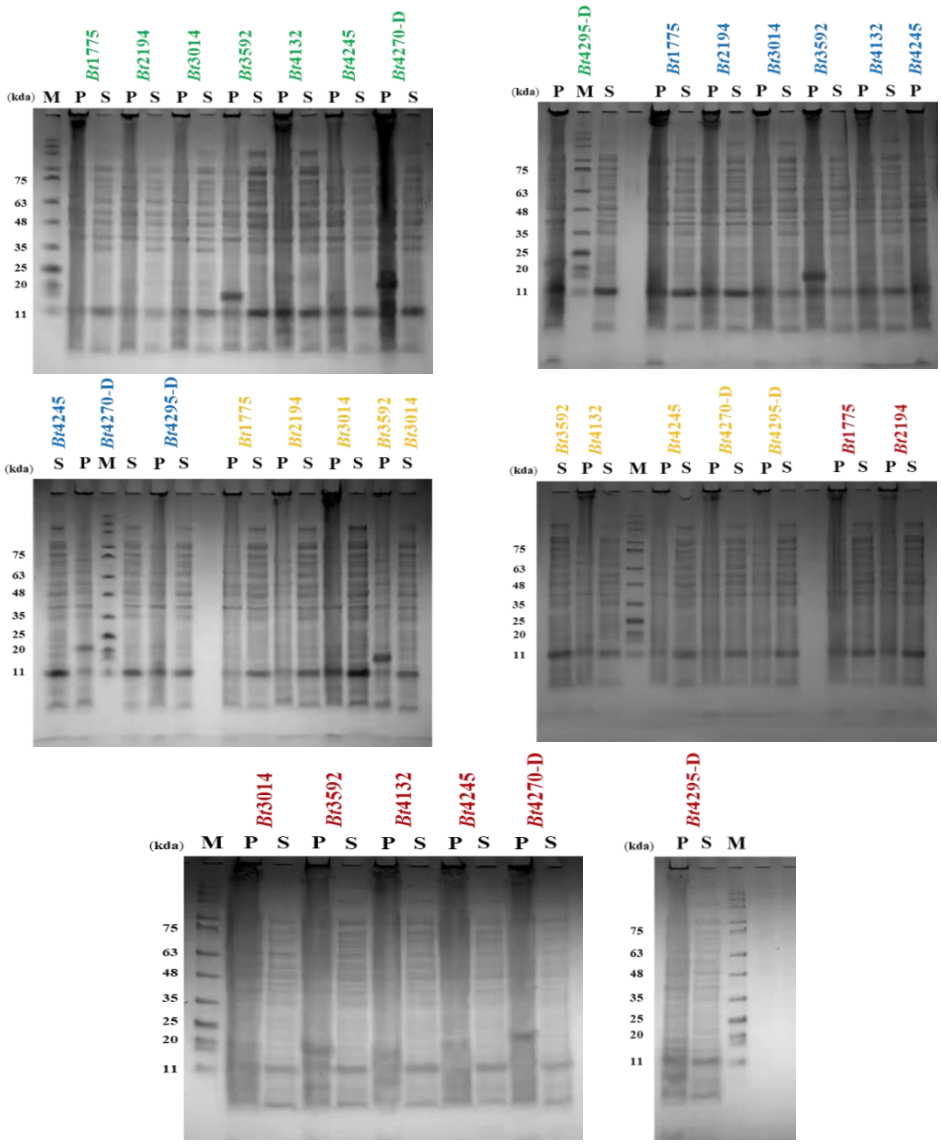


Figure 2.2 - Protein expression levels of the 8 CBMs tested using the *E. coli Tuner* strain and an IPTG-induction protocol. The CBMs are identified by protein ID. The expression of CBMs was tested using the *E. coli Tuner* strain grown in LB medium, inducing the expression with 0.25, 0.5, 0.75 and 1 mM IPTG at 16 ºC, for 16h at 160 rpm. Green, blue, yellow and red represent 0.25, 0.5, 0.75 and 1 mM of IPTG, respectively. SDS-PAGE (10% acrylamide) gels were run at a fixed voltage of 100 V. P-Pellet (insoluble fraction); S-supernatant (soluble fraction); M-Marker II from NZYTech®.

The analysis of the several SDS-PAGE results showed that there was no improvement of the expression of the CBM32 in the soluble fraction at an expected molecular weight of around 17 kDa. In fact, in Tuner cells the yields of the expression in the insoluble fraction were lower overall, taken for example the cases of *Bt*3014 or *Bt*4132. We can also conclude that the variation of IPTG concentration had no noticeable alteration in the level of expression of CBMs.

To try to achieve a higher level of CBM expression in the soluble fraction, a final attempt was carried out using *E. coli Rosetta* strain cells and the IPTG induction protocol, with an IPTG concentration of 1mM. Using these conditions, most of the CBM32s were expressed in the

insoluble fraction, while the soluble fraction appears clear for most of the CBMs tested (**Figure 2.3**). The control *Bt*3015-C CBM32 showed, as predicted, good expression levels.

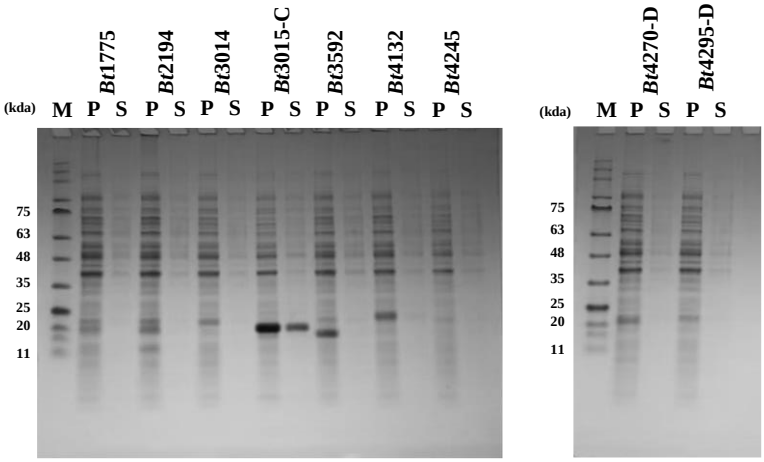


Figure 2.3 - Protein expression levels of the 8 CBMs tested using the *E. coli Rosetta* strain and an IPTG-induction protocol. The CBMs are identified by protein ID. The expression of CBMs was tested using the *E. coli Rosetta* strain grown in LB medium, inducing the expression with 1mM IPTG at 16 ºC, for 16h at 160 rpm. Fractions analysed in the gel resulted from the sonication of 1mL of sample. SDS-PAGE (10% acrylamide) gels were run at a fixed voltage of 100 V. P-Pellet (insoluble fraction); S-supernatant (soluble fraction); M-Marker II from NZYTech®.

The recurrent presence of CBMs in the insoluble fraction led to the hypothesis that the sonication was not efficient enough for the protein to be present in the soluble fraction at the expected molecular weight. The time period for the sonication of a 1 mL fraction of cell culture may not be sufficient for cell lysis and so, in the attempt of optimization, the entire sample of each CBM from this expression test was sonicated, increasing the number of cycles and longer sonication periods. Both soluble and insoluble fractions from this optimization attempt were analysed by SDS-PAGE and the results are shown below in **Figure 2.4**.

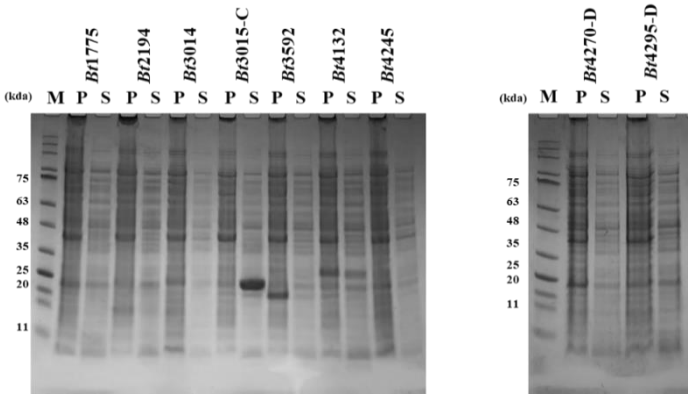


Figure 2.4 - Protein expression levels of the 8 CBMs tested using the *E. coli Rosetta* strain and an IPTG-induction protocol. The CBMs are identified by protein ID. The expression of CBMs was tested using the *E. coli Rosetta* strain grown in LB medium, inducing the expression with 1mM IPTG at 16 ºC, for 16h at 160 rpm. Fractions analysed in the gel resulted from the sonication of the entire sample. SDS-PAGE (10% acrylamide) gels were run at a fixed voltage of 100 V. P-Pellet (insoluble fraction); S-supernatant (soluble fraction); M-Marker II from NZYTech®.

Observing the results closely, it is clear that the expression of CBMs in all cases is very diminished, still with the presence of protein in the insoluble fraction. The exception still remains in the case of *Bt*3015-C, which reveals a high level of protein in the soluble fraction, in contrast with the low protein level in the insoluble portion.

In an attempt of purifying and isolating the CBMs, each soluble fraction was purified using affinity chromatography with His GraviTrap^{TM} columns (protocol described in section 2.2.6.1.) and the results are shown below in **Figure 2.5.**

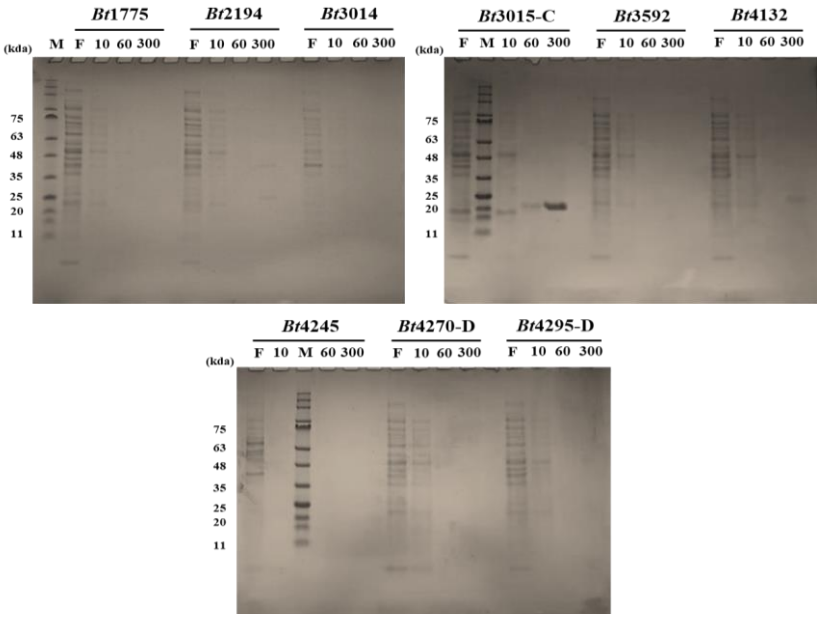


Figure 2.5 - Purification results of the 8 family 32 CBMs tested. The CBMs are identified by protein ID. Protein lanes are indicated by a black line. Results of purification with His GraviTrap^{TM} columns. CBMs were eluted with 6 ml of 50 mM HEPES at a pH of 7.5, 1M NaCl, 300 mM imidazole, 5 mM $CaCl_2$ and 5 mM $MgCl_2$ buffer. SDS-PAGE (10% acrylamide) gels were run at a fixed voltage of 100 V. F- Flowthrough; 10- Elution with 10 mM imidazole buffer; 60- Elution with 60 mM imidazole buffer; 300- elution with 300 mM imidazole buffer; M- Marker II from NZYTech®.

With the exception of the CBM *Bt*3015-C, our control, none of the other CBMs revealed positive results in this purification step. There are no bands detected after Coomassie blue staining in the lanes corresponding to elution fractions with 300 mM of imidazole, condition in which the protein should be eluted from the column. These results lead to the conclusion that the protocol for the expression of these 8 CBMs needs further optimization to express the proteins in a soluble form in good yields. One aspect to consider would be changing the protein construct.

2.3.2. Family 32 CBMs and SusD-like proteins small scale expression and purification

As mentioned in the previous section, expression could not be successfully achieved for the eight family 32 CBMs tested, and so the attention turned to two pairs of proteins, a SusD-like protein and a CBM 32 from two *polysaccharide utilization loci*: *Bt*0865-CBM32 and the SusD-like protein *Bt*0866 from PUL 12 and *Bt*4040-CBM32 and the SusD-like protein *Bt*4038 from

PUL 73. The choice of these PULs was based on the evidence that these were differentially expressed when *B. thetaiotaomicron* was grown on mucin glycans as a sole carbon source.[9] These are represented in **Figure 2.6**.

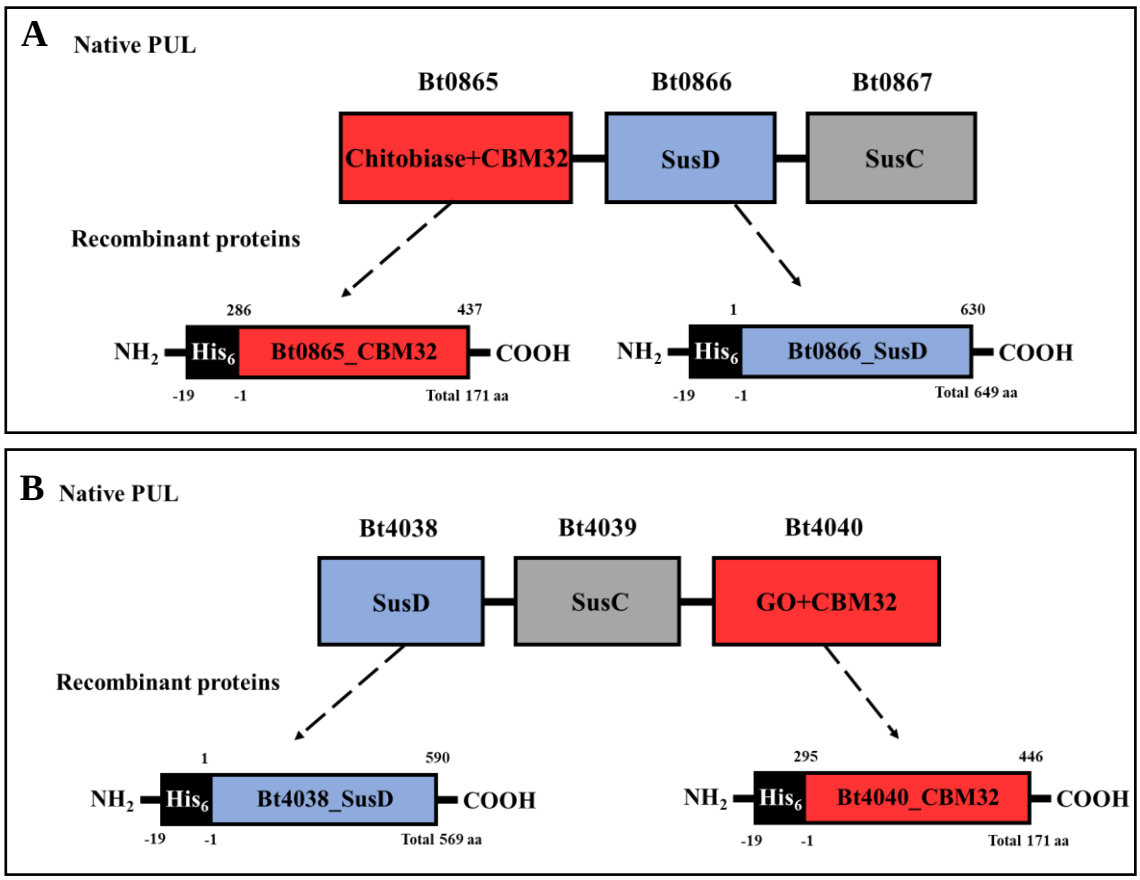


Figure 2.6 - Molecular architecture of the PULs in which the family 32 CBMs *Bt*0865, *Bt*4040 and the Sus-D like proteins *Bt*0866 and *Bt*4038 are inserted. Adapted from the information available at (http://www.cazy.org/PULDB/)

*Bt*0865-CBM32 is a protein that, according to the CAZy database (www.CAZy.org), has homology to a chitobiase, which contains an associated non-catalytic module that is homologous to a family 32 CBM; *Bt*4040-CBM32, has homology to a galactose oxidase, with a non-catalytic CBM32 module. *Bt*0866-SusD and *Bt*4038-SusD have homology with SusD-like surface glycan binding proteins. The CBM32 and SusD-like protein domains were prepared as an N-terminal hexa-histidine tag constructs in order to facilitate the purification step by affinity chromatography.

The sequences for all constructs are represented in **supplementary table 2**. The main objective of this work was to optimize the expression of these proteins and eventually undergo a large-scale expression and purification for posterior crystallization attempts and MST assays.

2.3.2.1. Small-scale expression and purification of *Bt*0866-SusD and *Bt*4038-SusD

The first step was to perform expression in a small scale to determine the optimal expression and purification conditions. The Sus-D-like proteins *Bt*0866 and *Bt*4038 were first expressed in BL21 (DE3) *E. coli* cells. The cell culture was grown at 37°C for a total of 3 hours until the OD_{600} reached a value proximate to 0.6. Once the OD_{600} reached close to this value, an IPTG-protocol was used for induction using an IPTG concentration of 1mM. Two induction temperatures and induction periods were tested: in one optimization condition the cells were induced for 16 hours at 19°C and on the second optimization condition the induction was carried over at 37°C for a total of 3 hours. After this step, the cell extract was purified through a step of affinity chromatography using His GraviTrap™ columns. The fractions denominated E1, E2 and E3 of each protein were the ones chosen for a desalting step, in order to remove the excess imidazole that still remained in solution. All samples were posteriorly analysed by SDS-PAGE electrophoresis and the results can be seen in **Figure 2.7**.

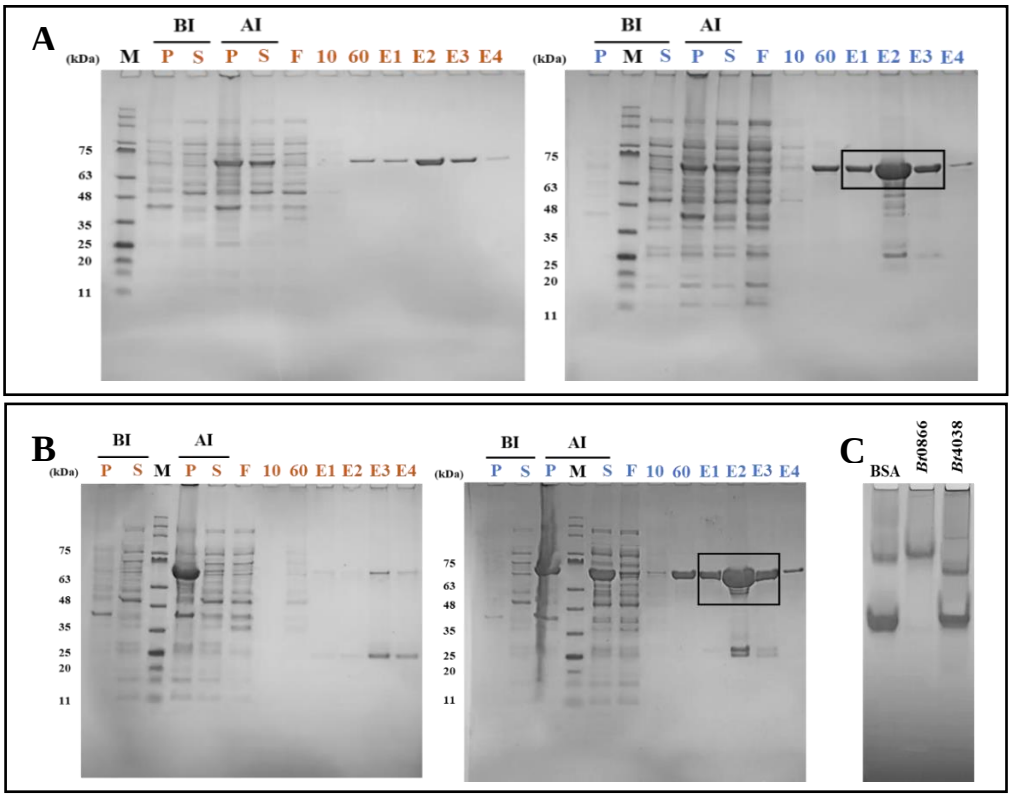


Figure 2.7 - Purification of Sus-D like proteins *Bt*0866 and *Bt*4038. Panel A) Results of purification with His GraviTrap™ columns for *Bt*0866. Panel B) Results of purification with His Gravitrap™ columns for *Bt*4038. Panel C) Native-PAGE gel of the desalted fractions of *Bt*0866 and *Bt*4038. Orange represents the results obtained for the 3 hour induction at 37 °C and the colour blue represents the results obtained for the induction at 19 °C for 16 hours. Proteins were eluted with 6 ml of 50 mM HEPES at a pH of 7.5, 1M NaCl, 300 mM imidazole, 5 mM $CaCl_2$ and 5 mM $MgCl_2$ buffer. The fractions were eluted and analysed in separate: three fractions with 1 ml each and a final 3 ml fraction. SDS-PAGE (10% acrylamide) gels were run at a fixed voltage of 100 V. P-Pellet (insoluble fraction); S-supernatant (soluble fraction); BI- Before induction; AI- After induction; F- Flowthrough; 10- Elution with 10 mM imidazole buffer; 60- Elution with 60 mM imidazole buffer; E1- Eluted fraction 1 (1ml); E2- Eluted fraction 2 (1ml); E3- Eluted fraction 3 (1ml); E4- Eluted fraction 4 (3ml); M- Marker II from NZYTech®.

The results represented in **Figure 2.7** showed that *Bt*0866-SusD and *Bt*4038-SusD were expressed in both expression conditions as a protein band was visualized at the expected molecular weight of around 70 kDa and the native-PAGE show the appearance of two bands for *Bt*4038-SusD, possibly suggesting the presence of two isoforms. The best results obtained were with the induction at 19°C for 16 hours, which showed a high level of expressed protein in the soluble fraction and after purification. This condition was selected for large-scale expression.

2.3.2.2. Small-scale expression and purification of *Bt*0865-CBM32 and *Bt*4040-CBM32

After the successful small-scale expression and purification of the SusD-like proteins *Bt*4038 and *Bt*0866, the conditions of expression and purification were tested for the family 32 CBMs *Bt*0865 and *Bt*4040. These CBMs were expressed using BL21 (DE3) *E.coli* cells that were grown for a total of 3 hours at 37°C and then induced using the same IPTG induction protocol for 16 hours at 19°C, using an IPTG concentration of 1mM. The results obtained after the analysis of both soluble and insoluble fractions are shown in **Figure 2.8.** The results showed expression of the CBMs, with both proteins present in good amounts in the soluble fraction.

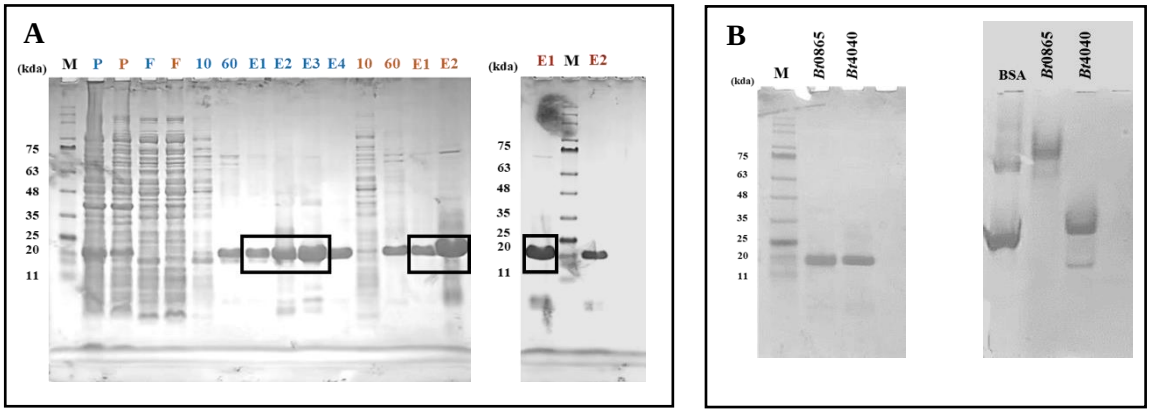
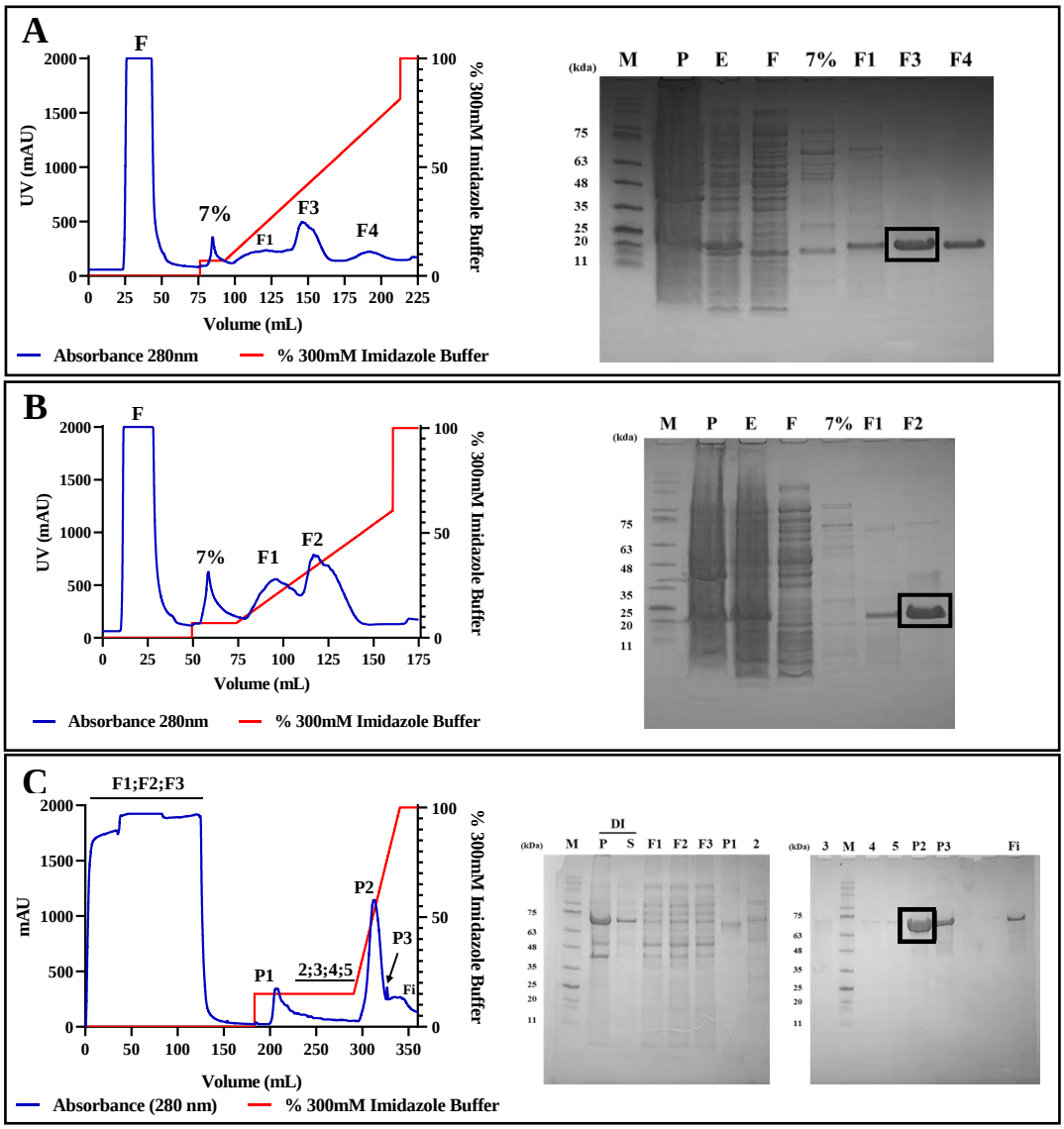


Figure 2.8 - Purification of family 32 CBMs *Bt*0865 and *Bt*4040. Panel A) Results of purification with His GraviTap™ columns for *Bt*0865-CBM32 and *Bt*4040-CBM32. Panel B) SDS-PAGE and native-PAGE results for the samples after the desalinization step. Orange represents the results obtained for Bt4040-CBM32 and the colour blue represents the results obtained for Bt0865-CBM32. Proteins were eluted with 6 ml of 50 mM HEPES at a pH of 7.5, 1M NaCl, 300 mM imidazole, 5 mM $CaCl_2$ and 5 mM $MgCl_2$ buffer. The fractions were eluted and analysed in separate: three fractions with 1 ml each and a final 3 ml fraction. SDS-PAGE (10% acrylamide) gels were run at a fixed voltage of 100 V. P-Pellet (insoluble fraction); S-supernatant (soluble fraction); BI- Before induction; AI- After induction; F- Flowthrough; 10- Elution with 10 mM imidazole buffer; 60- Elution with 60 mM imidazole buffer; E1- Eluted fraction 1 (1ml); E2- Eluted fraction 2 (1ml); E3- Eluted fraction 3 (1ml); E4- Eluted fraction 4 (3ml); M- Marker II from NZYTech®.

The elution fractions E1, E2 and E3 were chosen for desalting as these contained the highest quantity of each CBM. The samples were posteriorly analysed through SDS-PAGE and Native-PAGE electrophoresis (**Figure 2.8B**), which showed that both CBMs were purified to homogeneity with no observable protein degradation. Taking these results in consideration, this protocol was followed for large-scale production of these CBMs.

2.3.3. Large scale expression and purification

To obtain enough protein for structural and biophysical studies, we proceeded to the large scale expression and purification of *Bt*0865-CBM32, *Bt*0866-CBM32, *Bt*4040-SusD and *Bt*4038-SusD. These proteins are interesting candidates for structural and biophysical characterization, not only because of their specificity for several glycans that are present in mucins that cover our intestine wall, but also because, little information is currently known about their three-dimensional structure and interaction with specific ligands. In order to achieve a large-scale purification, the CBM32s and SusDs were purified by IMAC. The chromatograms of the purification obtained for each protein are shown in **Figure 2.9**, with the corresponding SDS-PAGE analysis.



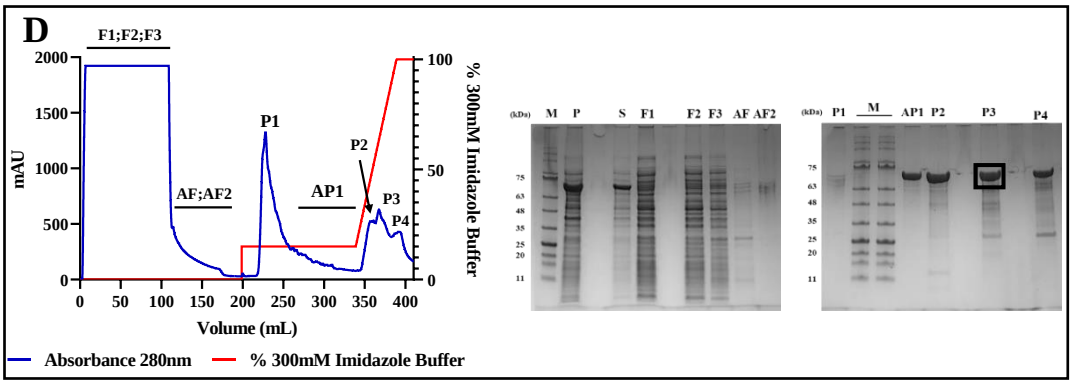


Figure 2.9 - Large-scale purification of *Bt*0865-CBM32, *Bt*4040-CBM32, *Bt*0866-SusD and *Bt*4038-SusD. A)- Chromatogram and SDS-PAGE results for *Bt*0865-CBM32; **B)-** Chromatogram and SDS-PAGE results for *Bt*4040-CBM32; **C)-** Chromatogram and SDS-PAGE results for *Bt*0866-SusD; **D)-** Chromatogram and SDS-PAGE results for *Bt*4038-SusD; Blue line is representative of the UV. The red line is representative of the gradient of 300mM imidazole buffer. Black rectangles represent the fractions selected for a desalinization step. SDS-PAGE (10% acrylamide) gels were run at a fixed voltage of 100 V P-Pellet (insoluble fraction); S-supernatant (soluble fraction); E-Extract F-Flowthrough; AF-After flowthrough; 7%- 7% of buffer B; P-Peak; AP-After peak; Fi-Final M-Marker II from NZYTech®.

The results showed that fractions F3 from *Bt*0865-CBM32, F2 from *Bt*4040-CBM32, P2 from *Bt*0866-SusD and P3 from *Bt*4038-SusD contained the protein of interest and were then submitted to a desalinization process for removal of excess imidazole and were posteriorly concentrated for structural studies.

The protein yield per litre of culture was calculated: 11.5 mg/L for *Bt*0865-CBM32, 14.6 mg/mL for *Bt*4040-CBM32, 16.25 mg/mL for *Bt*0866-SusD and 7.5 mg/mL for *Bt*4038-SusD (**Table 2.3)**. protein was stored in a buffer containing 50mM HEPES, 1M NaCl, 5mM $CaCl_2$, 5mM $MgCl_2$, and 10mM of imidazole with a pH of 7.5, at 4°C.

Table 2.3- Purification yields for *Bt*0865-CBM32, *Bt*4040-CBM32, *Bt*0866-SusD and *Bt*4038-SusD

Protein	Volume Medium culture (mL)	Volume Purified protein (mL)	Protein quantitation (mg/mL)	Total protein after IMAC (mg)
***Bt*0865-CBM32**	2400	30	0.92	27.6
***Bt*4040-CBM32**	2400	35	1	35
***Bt*0866-SusD**	2400	30	1.3	39
***Bt*4038-SusD**	2400	20	0.9	18

2.3.4. Protein stability analysis

Working with new proteins is always challenging, especially in the field of optimizing the maximum number of conditions to achieve the best outcome in protein purification. One of the conditions is the protein buffer during purification and storage. When proteins are produced in a large scale, conditions may need to be optimized in order to promote protein stability. From the gel electrophoresis analysis, the proteins produced in large scale appeared to be stable in the buffer

containing 50 mM HEPES at a pH of 7.5, 150 mM NaCl 5 mM $CaCl_2$ and 5 mM $MgCl_2$. However, a TSA experiment can give additional information regarding the stability of the protein in solution and assess which buffer maximizes that stability. The main purpose of the protein stability tests was to determine if the proteins were stable in their purification buffer (50mM HEPES pH 7.5, 150mM NaCl, 5mM $CaCl_2$, 5mM $MgCl_2$) by calculating their melting temperature. A TSA screening was performed in a 96 well format, which is very advantageous as it gives the ability to test a large variety of conditions in a short period of time. and the results are shown in **Figure 2.10** below. All proteins the proteins were analysed at the concentration of 40 µM.

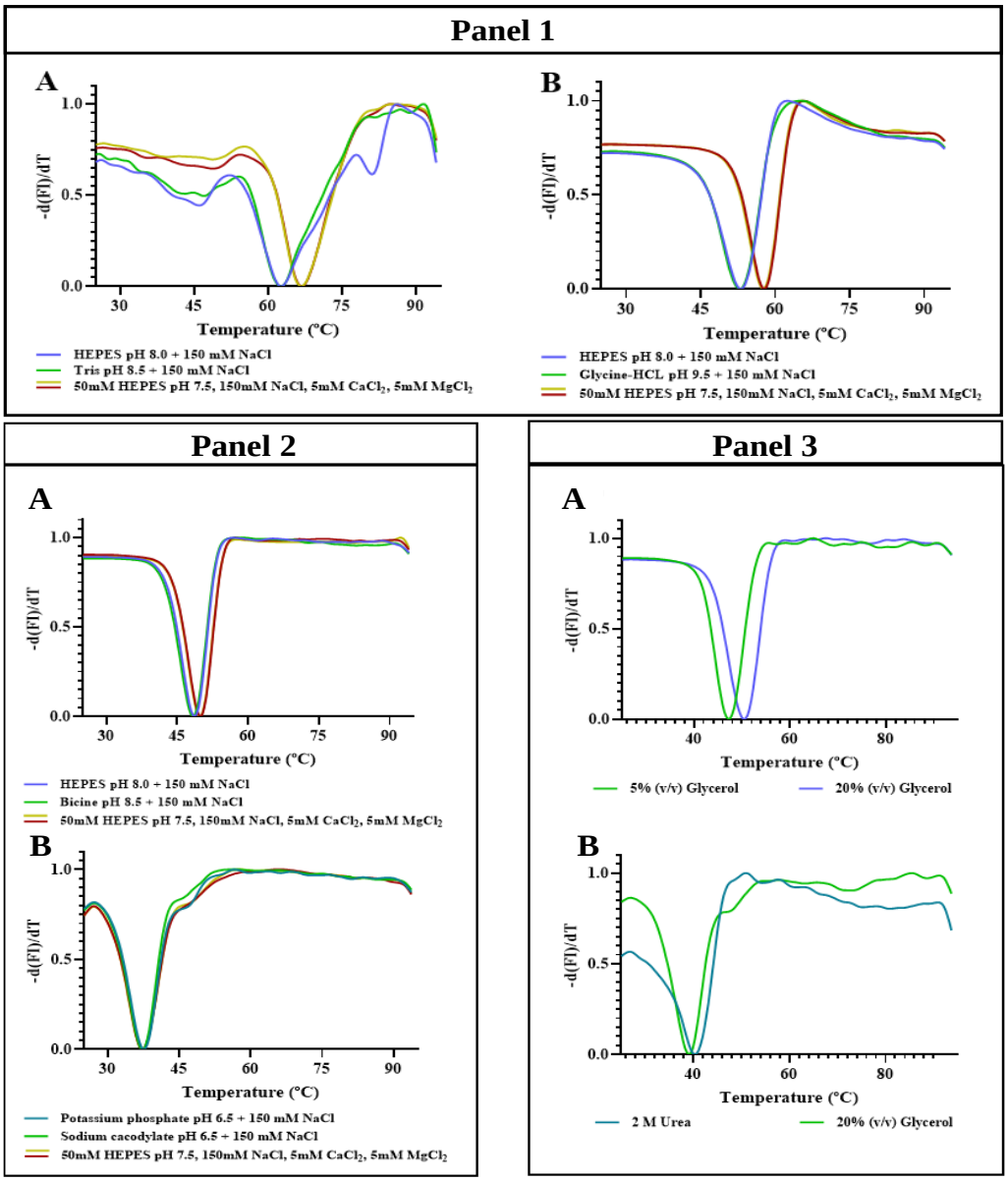


Figure 2.10 - Thermal stability assay results for *Bt*4040-CBM32, *Bt*0865-CBM32, *Bt*0866-SusD and *Bt*4038-SusD. Panel 1) Results showing the derivative normalized melting curve for the Thermofluor buffer screen: **1A**- Results for *Bt*4040-CBM32; **1B**- Results for *Bt*0865-CBM32; Panel 2) Results showing the derivative normalized melting curve for the Thermofluor buffer screen: **2A**- Results for *Bt*0866-SusD ; **2B**- Results for *Bt*4038-SusD; Panel 3- Results showing the derivative normalized melting curve for the Thermofluor additive screen: **3A**- Results for *Bt*0866-SusD ; **3B**- Results for *Bt*4038-SusD; – d(FI)/dT represents the first derivative of the raw data in which the minimum value corresponds to the protein melting temperature.

The three panels represented above show the derivative normalized melting curve for the Thermofluor buffer screen of *Bt*4040 and *Bt*0865 (1A and 1B, respectively) and *Bt*0866 and *Bt*4038 (2A and 2B). Panel 3 is representative of the derivative melting curve for the Thermofluor additive screen of *Bt*0866 and *Bt*4038 (3A and 3D). A screening of 96 buffer conditions (**Supplementary figure 1**) was used to test the stability of all four proteins.

The solution containing each protein also contains a dye (SYPRO Orange) that binds to hydrophobic amino acid residues. When the protein gets progressively more denatured, the dye gains access to the internal hydrophobic residues, consequently increasing the fluorescence intensity. It is clear when the protein reaches a complete unfolded state, the fluorescence reaches its maximum level, and then the SYPRO Orange signal decreases due to protein-dye dissociation. The inflection point is representative of the melting temperature (Tm), which is the temperature at which 50% of the protein in solution is unfolded. In the derivative graph, this inflection point corresponds to a value of zero, and the temperature at which the derivative is equal to zero is our melting temperature. Raw data from both buffer screen assays and additive assays were analysed and the most stable conditions with the melting temperature obtained are shown in **Table 2.4**.

Table 2.4- Thermal stability assay data analysis for *Bt*4040-CBM32, *Bt*0865-CBM32, *Bt*0866-SusD, *Bt*4038-SusD

Protein	Buffer	Tm (ºC)	Additive screen Tm (ºC)
***Bt*4040-CBM32**	HEPES pH 8.0 + 150 mM NaCl	62.7	Not tested
	Tris pH 8.5 + 150 mM NaCl		
	50mM HEPES pH 7.5, 150mM NaCl, 5mM $CaCl_2$, 5mM $MgCl_2$ (Control)	66.71	
***Bt*0865-CBM32**	HEPES pH 8.0 + 150 mM NaCl	53.12	Not tested
	Glycine-HCL pH 9.5 + 150 mM NaCl	52.72	
	50mM HEPES pH 7.5, 150mM NaCl, 5mM $CaCl_2$, 5mM $MgCl_2$ (Control)	58	
***Bt*0866-SusD**	HEPES pH 8.0 + 150 mM NaCl	48.75	50.34 (20% (v/v) glycerol)
	Bicine pH 8.5 + 150 mM NaCl	48.33	
	50mM HEPES pH 7.5, 150mM NaCl, 5mM $CaCl_2$, 5mM $MgCl_2$ (Control)	50	
***Bt*4038-SusD**	Potassium phosphate pH 6.5 + 150 mM NaCl	37.48	40.3 (1M Urea)
	Sodium cacodylate pH 6.5 + 150 mM NaCl	37.08	
	50mM HEPES pH 7.5, 150mM NaCl, 5mM $CaCl_2$, 5mM $MgCl_2$ (Control)	37.08	

The *Bt*4040-CBM32 was stable in the protein the purification buffer, as its Tm in this buffer (66.71 ºC) was superior to the Tm determined for the second and third most stable buffers (62.7 ºC). Regarding the *Bt*0865-CBM32, the melting temperatures obtained with purification buffer (58 ºC) and the HEPES pH 7.5 and Glycine-HCl pH 9.5 buffers (53.12 ºC and 53.72 ºC, respectively) indicated that the protein was stable in the original buffer. For *Bt*0866-SusD, the Tm obtained were also similar for the most stable buffers, indicating that this protein is also stable in the purification buffer. Finally, *Bt*4038-SusD was the protein that revealed to be less stable in the purification buffer comparing with *Bt*4040-CBM32, *Bt*0865-CBM32 and *Bt*0866-CBM32 with a melting temperature of 37.08 ºC and the Thermofluor buffer screen revealed no buffers that provided a significant increase in stability of the protein in solution. Taking this in consideration, an additive screen (**Supplementary figure 2**) was performed in order to assess the existence of compounds that could increase the protein stability. The results showed that for *Bt*0866-SusD the temperature variation from the addition of 20% (v/v) glycerol is negligeable (ΔTm=0.34 ºC) when compared to the melting temperature of the protein in the initial buffer. For *Bt*4038-SusD the addition of 1M Urea retrieved the best results increasing the melting temperature from 37.08 ºC to 40.3 ºC, but nevertheless it was not considered a significant improvement.

2.4. Conclusions

The first objective when studying proteins is obtaining our protein purified to homogeneity and stable following an expression and purification protocol that needs to be optimized so that an acceptable amount of protein is obtained for future studies. When that optimization step is not conducted the efficiency of a large scale protocol will be conditioned, possibly leading to low protein stability that consequently will result in a low purification yield due to protein precipitation or folding instability.

In the first part of the work, several expression tests were performed to assess the optimal expression conditions for eight different family 32 CBMs and obtain the higher amount of purified protein. Good levels of expression of protein in the soluble fraction were not achieved and the expression of these specific family 32 CBMs need optimization in the future.

In the second part, large scale expression and purification was tested using two pairs, each from different PULs of family 32 CBM/Sus-D like protein, more specifically *Bt*0865/*Bt*0866 (PUL 12) and *Bt*4040/*Bt*4038 (PUL73). A high level of expression was achieved for all proteins using IPTG induction and an incubation for at least 16 hours at 19 ºC.

After affinity purification with IMAC, 11.5 mg/L of *Bt*0865-CBM32, 14.6 mg/mL of *Bt*4040-CBM32, 16.25 mg/mL of *Bt*0866-Sus-D and 7.5 mg/mL of *Bt*4038-SusD were obtained per litre of culture.

The third part of the work was relative to the protein thermal stability. The main objective was to ensure that the proteins were stable in the selected protein buffer, or if there were other conditions that would improve the protein stability in solution. All proteins were stable in the purification buffer (50mM HEPES pH 7.5, 150mM NaCl, 5mM CaCl2, 5mM MgCl2) except for *Bt*4038-SusD, which presented a melting temperature of around 37 ^{o}C. The additive screen tested did not show any results leading to an increase of protein stability in solution and requires further studies.

Following the protein large scale expression and purification and thermal stability, the proteins were followed up for structural characterization by X-ray crystallography.

2.5. Selection of the protein target for structural characterization

The structural characterization by X-ray crystallography is going to be explored in the next chapter. In the time frame of the thesis, the most promising and complete results were obtained for *Bt*4040-CBM32, which enabled to solve its 3D structure at high resolution. For this reason, and as this is a novel structure of a CBM family 32 member with a distinctive glycan binding specificity, this CBM will be the focus of the next chapter. For *Bt*0865-CBM32, protein microcrystals were obtained, and optimization of the crystallization conditions needs to be performed. Regarding the SusD-like proteins, no crystals were obtained in the case of the *Bt*4038-SusD. This could be explained by the low protein stability, considering the low melting temperature observed for this protein (TSA assay present in section 2.3.4). Regarding the *Bt*0866-SusD, good-diffracting crystals were obtained and 3D structure solution is ongoing. The screenings tested for all four proteins are represented in **supplementary figures 3, 4, 5** and **6**.

Taking this in consideration, and like it was mentioned before, the following chapter was focused on the *Bt*4040-CBM32 since the results obtained showed to be more promising and complete.

Chapter 3- Structural characterization of the family 32 CBM *Bt*4040

3.1. Introductory remarks

Family 32 CBMs can be found in a multitude of microorganisms, including archaea, fungi and eubacteria. Many members of this family are identified in the genome of known human pathogens, such as *Clostridium perfringens* [91]. They can also be found in several other organisms such as, in the context of this thesis, the commensal *Bacteroides thetaiotaomicron*. CBM32s are characterized by a typical β-sandwich fold, normally containing a bound metal ion (most often calcium) and five amino acid residues that appear to be conserved among the family 32 CBMs. These residues comprise an arginine, asparagine, phenylalanine, histidine and tyrosine, making up what is known as the canonical galactose-binding site [92]. This binding site is located in the terminal loop region, and in some cases it is quite small and intended to bind monosaccharides or small oligosaccharides, thus CBM32s are considered type C CBMs [91]. CBM32s can be associated with a variety of catalytic modules and in the case of *Bt*4040-CBM32, which is the focus of this chapter, is associated with a galactose oxidase. A substantial amount of information is known for pathogen-associated CBM32s but it is also important to deepen the knowledge in commensal bacteria, such as *B. thetaiotaomicron*, to understand the interactions that take place in the GI tract.

CBM32 have been shown to bind a variety of carbohydrate ligands including galactose, polylactosamine (LacNAc) and polygalacturonic acid [91], [93]. Previous studies in the lab with *Bt*4040-CBM32 using the glycan microarrays technique demonstrated a binding specificity of *Bt*4040-CBM32 for epitopes with the Lewis A determinants (Gal-β-1,3(Fuc-α-1,4)GlcNAc). In this experiment, around 475 probes were tested, mainly probes containing sequences found in mammals and the results are represented in **Figure 3.1**.

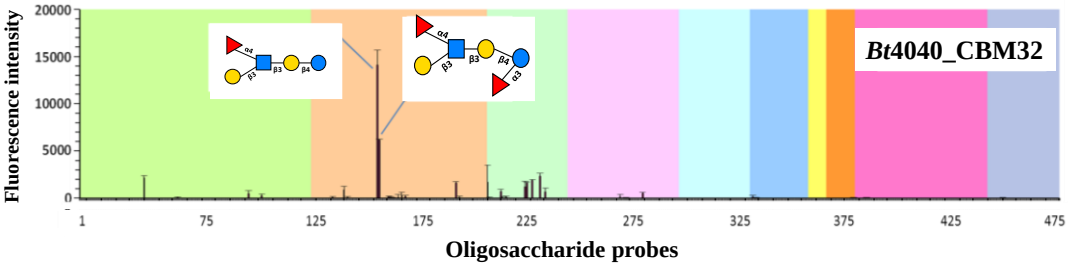


Figure 3.1- Glycan microarray analysis of *Bt*4040-CBM32. Probes are arranged in colours depending on the class to which they belong, for a total of 475 probes (mainly mammalian type) (Correia, Costa et al, unpublished).

The results obtained in the microarray analyses show the binding specificity of *Bt*4040-CBM32 for epitopes containing Lewis A determinants, being these results important in the context of future work, mainly the 3D structure solution of the protein-ligand complexes, as well as the determination of a dissociation constant for the interaction using MST.

This chapter reports: 1) the attempts to obtain good X-ray diffracting crystals for *Bt*4040-CBM32 using different crystallization screenings; 2) the determination of the three-dimensional structure of *Bt*4040-CBM32; 3) the comparative structural analysis with a family 32 CBM from *B. thetaiotaomicron* and a family 32 CBM from *Clostridium perfringens*.

3.2. Materials and methods

3.2.1. Pre-Crystallization test

To assess the optimal protein concentration for crystallization screenings, a pre-crystallization test was performed for each protein. The sample concentration is an extremely important crystallization variable, mainly because too concentrated samples can originate an amorphous precipitate and the complete opposite, too dilute samples, can lead to clear drops. These results are very often found in crystallization screening conditions that do not promote crystallization and so, the pre crystallization test by optimizing the protein concentration is very important in the reduction of the number of clear and precipitate results, leading to a more efficient sample utilization, enhancing at the same time the chances of crystallization.

To perform the pre-crystallization tests, 4 different solutions (A1, A2, B1 and B2) were used and are listed in **Table 3.1**.

Table 3.1- List of compounds and respective concentrations for pre-crystallization test (as described in https://hamptonresearch.com/uploads/support_materials/HR2-140_142_Binder.pdf**)**

Reagent	Composition
A1	0.1M Tris-HCl pH 8.5, 2 M Ammonium sulfate
A2	0.1M Tris-HCl pH 8.5, 1 M Ammonium sulfate
B1	0.1M Tris-HCl pH 8.5 0.2 M Magnesium chloride hexahydrate 30% (w/v) Polyethylene glycol 4000
B2	0.1M Tris-HCl pH 8.5 0.2 M Magnesium chloride hexahydrate 15% (w/v) Polyethylene glycol 4000

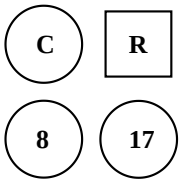
*Bt*4040-CBM32 was submitted to a pre-crystallization test to search for an optimal protein concentration for crystallization screenings. The pre-crystallization test was carried over in 24-well crystallization plates where 700 µL of solutions A1 and A2 were pipetted into reservoir A1 and A2, respectively. Next, 1 µL of protein was mixed with 1 µL of solution A1 in a single glass

cover slide. The same procedure was repeated with solution A2. The glass slides containing both drops were inverted, sealed and incubated for 30 minutes. For an ideal protein concentration, a light granular precipitate should form. The results obtained were compared with **supplementary table 3**.

3.2.2 Protein crystallization assays

The crystallization screening attempts were conducted using the vapor diffusion method (sitting drop) in 96-well plates (3 drop format), using the protein at two different concentrations. The setups were prepared using a nano drop robot (Oryx8, Douglas Instruments). Different crystallization screenings were used to quickly analyse a wide variety of conditions with a low protein quantity. Two different screens were tested the JBScreen Classic 1-4 (Jena Bioscience) (**Supplementary figure 3**) and the Morpheus screening (Molecular Dimensions) (**Supplementary figure 4**), at two different temperatures (4 ºC and 20 ºC) for the two protein concentrations, comprising a total of 768 different formulations. The plates contain three spots for the drops and all experiments were designed in order to have a drop of buffer only, our control, a drop containing the lower concentrated protein and a drop with the higher concentrated protein. Crystallization conditions and the plate layout are represented in **Table 3.2**. Prior to robot set up, 30 µL of each crystallization solution from the respective screening was pipetted into the reservoir of the 96 well plate. After this step, the WASPRun program, implemented in the crystallization robot software, allows the user to select the desired protein:precipitant ratio which in this case was 1:1. The plates for each protein were stored at 4 ºC and 20 ºC.

Table 3.2- Crystallization conditions for *Bt*4040-CBM32 and the respective plate layout. R- reservoir; C-control drop; 8-drop containing the protein at a concentration of 8 mg/mL; 17-drop containing the protein at a concentration of 17 mg/mL;

	Screenings tested	**Incubation temperatures tested (ºC)**	**Protein:precipitant ratio**	**Plate layout**
***Bt*4040-CBM32**	JBScreen 1-4 Morpheus	4 ºC/20 ºC 4 ºC/20 ºC	1:1 1:1	

3.2.3. X-ray diffraction data collection

Crystals of *Bt*4040-CBM32 were produced using two crystallization screenings (JBScreen Classic 1-4 and Morpheus) at 20 ºC and 4 ºC. The crystals obtained were then harvested and stored in liquid nitrogen, using paratone as a cryo protector.

X-ray diffraction data from a single *Bt*4040-CBM32 crystal were collected under a nitrogen stream at 100 K in BL13 beamline at ALBA synchrotron. Using radiation with a wavelength of 0.9763 Å, a maximum resolution of 1.93 Å was achieved. The *Bt*4040-CBM32 crystal indexed in space group $P4_12_12$ with cell constants: $a=b=65.83$ Å, $c=69.08$ Å, and $\alpha=\beta=\gamma=90°$. Data were automatically processed with the XDS program and a file in an mtz format was retrieved from ISPyB.

3.2.4. Bt4040-CBM32 3D structure determination and refinement

The next step after acquiring the X-ray diffraction data was the determination of the 3D structure of *Bt*4040. The method of choice was molecular replacement, a phasing method that takes advantage of previous information in the form of known structures that are sequence-related or homologous to the protein of interest, in this case the CBM32 *Bt*4040 [94]. MR is usually the method of choice for structure determination when a suitable search model is available. In this structure determination step, a family 32 CBM (*Bt*0865) from *Bacteroides thetaiotaomicron* was used as search model. This CBM32 contains a percentage of sequence identity of around 30% with our target protein and is available from the Protein Data Bank (PDB) with the code 3GGL. The program PhaserMR was used for the molecular replacement and model improvement was performed by Autobuild, both programs from Phenix platform [95]. Several cycles of refinement were carried out in REFMAC5 [96] from CCP4i2 [97] suite alternating with iterative model building using COOT [98].

3.3. Results and Discussion

3.3.1. *Bt*4040-CBM32 crystallization assays

Like previously stated in section 3.2.1, a pre-crystallization test was performed with the objective of determining the optimal concentration for the crystallization assays. After comparing the results with **supplementary table 3**, a concentration of 17 mg/mL and a lower concentration of 8 mg/mL were elected for the crystallization assays. The JBScreen Classic 1-4 and Morpheus screens were tested at 20 ^{o}C and 4 ^{o}C. No crystals were obtained for both screens at 4 ^{o}C, however crystals were observed after one week of the JBScreen 1-4 crystallization set up at 20^{o}C (**Figure 3.2**). The condition that led to the formation of this crystal was: 25% Polyethylene glycol 4000 (PEG 4K), 0.1M Sodium Citrate pH 5.5 and 0.2M Ammonium Acetate, being the crystal present in the drop that corresponds to the higher protein concentration (17 mg/mL). The crystal obtained had an approximate length of 0.6mm, however it appeared to be multiple, reason why only a smaller part of the crystal was chosen for X-ray diffraction.

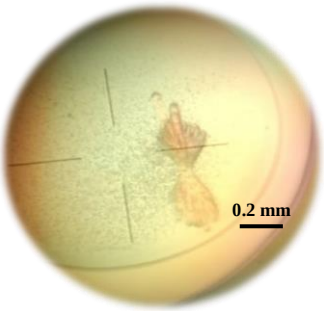


Figure 3.2 – Illustration of the crystal obtained for *Bt*4040-CBM32 using the JBScreen Classic 1-4 at 20 ºC for a protein concentration of 17 mg/mL. Crystal of *Bt*4040 obtained from a solution of 25% polyethylene glycol 4000, 0.2M sodium citrate pH 5.5 and 0.2 M ammonium acetate**.**

3.3.2. *Bt*4040-CBM32 3D structure determination and refinement

The crystal obtained in the previous section was used for diffraction data collection at ALBA synchrotron (BL13 beamline), under cryo conditions (100K) in order to protect the crystal from damage caused by radiation. The monochromatic radiation used had a wavelength of 0.9763 Å. Parameters and statistics from data collection and processing are reported in **Table 3.3**. After data collection, Phenix platform was used to solve the three-dimensional structure of *Bt*4040 based on the Molecular Replacement method, using as search model the *B. thetaiotaomicron* CBM32 *Bt*0865 (PDB code 3GGL). Several cycles of refinement were performed using REFMAC5 from CCP4i2 suite and iterative model building in the $2F_{obs}$-F_{calc} and F_{obs}-F_{calc} electron density maps took place in COOT (see **Table 3.3** for final refinement and validation statistics). The structure obtained for *Bt*4040 is shown below in **Figure 3.3**.

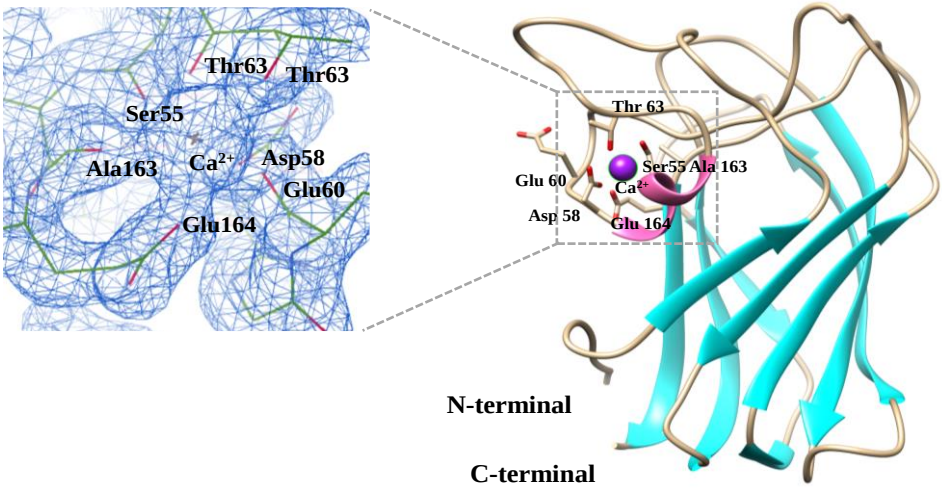


Figure 3.3- Ribbon representation of the 3D structure of *Bt*4040-CBM32 at 1.93Å resolution. Secondary structure elements are coloured: β-strands are represented in cyan for a total of nine and one α-helix is coloured in pink. The calcium ion (coloured in purple) has a heptahedral coordination from the main chain oxygens of Ala163, Thr63, Glu60 and Ser55 and from the side chain oxygen of Asp58, Glu164 and Thr63. Image rendered using Chimera [96]. On the right, the $2F_{obs}$-F_{calc} density map is depicted (σ=1), showing the calcium binding site and the coordinating amino acids.

Table 3.3- X-ray diffraction and structure refinement parameters and statistics for the 3D structure solution of Bt4040-CBM32

Data collection	
Wavelength (Å)	0.9763
Space group	*P* 4_1 2_1 2
Cell dimensions	
a, *b*, *c* (Å)	65.8, 65.8, 69.1
α, β, γ (°)	90, 90, 90
Resolution (outer shell) (Å)	47.65-1.90 (1.97-1.90)
Total number of reflections (outer shell)	183714 (16896)
Nº of unique reflections (outer shell)	12529 (1191)
R_{pim} (outer shell) (%)$^{+}$	4.1 (38.4)
R_{merge} (outer shell) (%)†	14.8 (140.8)
Mean I/σ(I) (outer shell) (%)	11.9 (1.3)
CC(1/2) (outer shell)	0.992 (0.528)
Completeness (outer shell) (%)	100 (99.9)
Redundancy (outer shell)	14.7 (14.2)
Structure refinement	
Resolution (Å)	46.55-1.93
No. of reflections	11752
R_{work}/ R_{free}‡	0.1883/0.2239
No. of atoms	
Protein	1160
Water molecules	73
Calcium ion	1
Average B factor	
Protein	29.33
Main-chain	28.33
Side-chain	30.40
Calcium ion	22.26
Water molecules	35.58
RMS deviations	
Bond length (Å)	0.007
Bond angle (°)	0.871
Ramachandran statistics (%)	
Favored	97.28
Allowed	2.04
Generously allowed	0.68
Forbidden	0

† $R_{merge} = \frac{\sum_{hkl}\sum_{i=1}^{n}|I_i(hkl)-\bar{I}(hkl)|}{\sum_{hkl}\sum_{i=1}^{n}I_i(hkl)}$, where I is the observed intensity, and $\bar{I}$ is the statistically-weighted average intensity of multiple observations.

$^{+}$ $R_{p.i.m.} = \frac{\sum_{hkl}\sqrt{1/(n-1)}\sum_{i=1}^{n}|I_i(hkl)-\bar{I}(hkl)|}{\sum_{hkl}\sum_{i=1}^{n}I_i(hkl)}$, a redundancy-independent version of R_{merge}.

‡ $R_{work} = \frac{\sum_{hkl}||F_{obs}(hkl)|-|F_{calc}(hkl)||}{\sum_{hkl}|F_{obs}(hkl)|}$, where $|F_{calc}|$ and $|F_{obs}|$ are the calculated and observed structure factor amplitudes, respectively. R_{free} is calculated for a randomly chosen 10% of the reflections.

The structure of *Bt*4040 was refined to final R_{work}=0.1883 and R_{free}=0.2239. The R-value, also known as R-factor or R_{work}, is a measure used in crystallography that reveals how well the atomic model reflects the X-ray diffraction data. In the case of *Bt*4040-CBM32, the R_{work} fits in the typical values which are around 0.2. The quality of a model normally requires a second value, in this case the R_{free} value. Analysing only R_{work} values normally introduces model bias to the process because during the refinement process the atomic model is also improved in order to fit better to the experimental data, consequently improving the R_{work}. The R_{free} is calculated using only 10% of the experimental observations and seeing how well the model predicts the lacking 90% of observations. This value is typically a little higher, with normal values around 0.26 and since the value obtained for this CBM is 0.2239 we can conclude that, not only the $R_{free,}$ but also the R_{work} are within normal values.

Overall, this CBM contains a β-sandwich structure, similar to the structure found in other CBMs from family 32. Regarding secondary structure elements, *Bt*4040-CBM32 contains 9 β-strands (indicated in cyan in **figure 3.3**) and an α-helix (indicated in pink in **figure 3.3**). A calcium ion is also present and is represented in the figure by a purple sphere. *Bt*4040-CBM32's calcium ion has a heptahedral coordination from the main chain oxygens of Ala163, Thr63, Glu60 and Ser55 and from the side chain oxygen of Asp58, Glu164 and Thr63.

3.3.3. CBM32s multiple sequence alignment and structure superpose

As a way to study the differences between several CBMs from family 32 a multiple sequence alignment was performed using the Clustal Omega online program [99]. The results obtained from this multiple sequence alignment were then analysed using another online program designated ESPript 3 with the objective of rendering sequence similarities and information about the secondary structure of aligned sequences.

In the multiple sequence alignment, and like previously mentioned, the CBMs used were the *Bt*3015C-CBM32 and CBM32 of *Cp*GH84C from *Clostridium perfringens* with the results being represented in **Figure 3.4**. A matrix of identity was also calculated (**Table 3.4**) showing the identity between all sequences, with *Bt*0865-CBM32 having a higher percentage of identity with *Bt*4040-CBM32 (35.14%), reason why it was chosen for the molecular replacement model.

Table 3.4 – Percent identity matrix calculated for CBM32s used in the sequence alignment (created by Clustal 2.1).

CBM32s	*Bt*4040	*Bt*0865	*Bt*3015C	*Cp*GH84C
***Bt*4040**	100	35.14	21.37	22.14
***Bt*0865**		100	22.46	24.82
***Bt*3015C**			100	29.71
***Cp*GH84C**				100

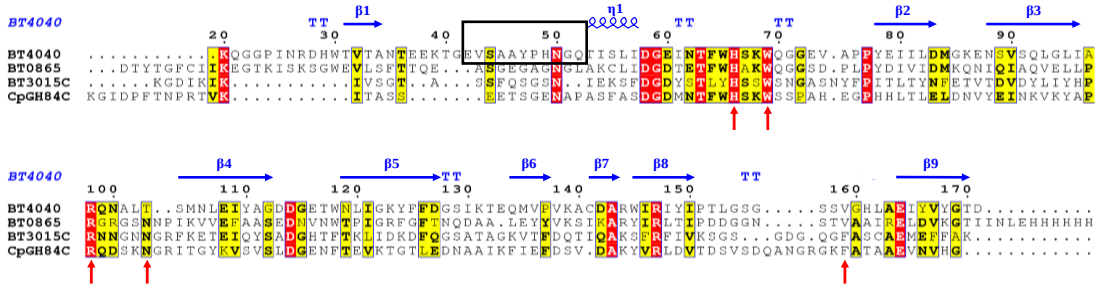


Figure 3.4 - Comparison of *Bt*4040-CBM32 with other CBM32s through multiple sequence alignment. Secondary structure motifs are indicated in blue. Red arrows indicate the residues involved in galactose binding. Numeration refers to *Bt*4040-CBM32. Highly conserved residues are shown in red and partially conserved residues are shown in yellow. Black rectangle is indicative of a less conserved loop in *Bt*4040-CBM32 in comparison with *Bt*3015C-CBM32 and CBM32 of *Cp*GH84C. Image rendered using the online program ESPript 3.

The multiple sequence alignment results **(Figure 3.4)** showed the presence of multiple conserved residues across all four CBMs. The residues highlighted with a red arrow represent the residues that form the canonical galactose binding-site present in the family 32 CBMs, more specifically histidine, tryptophan, arginine, asparagine and phenylalanine residues, as previously discussed in the introductory remarks of this chapter. These binding sites have a high specificity for galactose residues and have been previously studied in many reports such as in 2006 by Boraston and colleagues, using the CBM32 from *Clostridium perfringens* used in the above multiple sequence alignment (*Cp*GH84C), and more recently, in 2012, also by Boraston and colleagues [100]. The interaction made by some of these residues that comprise the canonical galactose-binding motif are highly conserved in all CBMs studied, namely the histidine, tryptophan and arginine residues. Only in the case of residues 103 and 159 of *Bt*4040-CBM32, the asparagine and phenylalanine residues are replaced by a threonine and a valine, respectively. This may indicate a different binding mode to galactose and Lewis A containing ligands in the case of *Bt*4040-CBM32 when compared to *Cp*GH84C. Regarding this *Clostridium perfringens* CBM32 from the 2006 study from Boraston [101] an asparagine residue is replaced by a glutamate residue in the binding motif for galactose, but a more recent study from 2012 [100] discovered that generally across the CBM32 family the residues that interact with galactose are the ones indicated by the red arrow in the alignment (**Figure 3.4**).

Taking a view at the secondary structure, we can observe that the more conserved regions are related to the secondary structure motif regions, mostly β-strands. To compare the secondary structure conservation across the CBMs tested, the MatchMaker tool from Chimera was used to superpose the structures. Two superpositions were performed. First, the isolated *Bt*4040-CBM32 structure, previously determined and refined, was superposed with the *Bt*3015C-CBM32 structure complexed with core 2 O-glycan (PDB code: 7BLJ, to be published) (**Figure 3.5 Panel A**). Then, *Bt*4040-CBM32 was superposed with the structure of *Cp*CBM32 from *Cp*GH84C complexed

with LacNAc (PDB code: 2J1E) being represented in (**Figure 3.5 Panel B**). The RMSD for the superposition of *Bt*4040 with each model was calculated and is presented in **Table 3.5**.

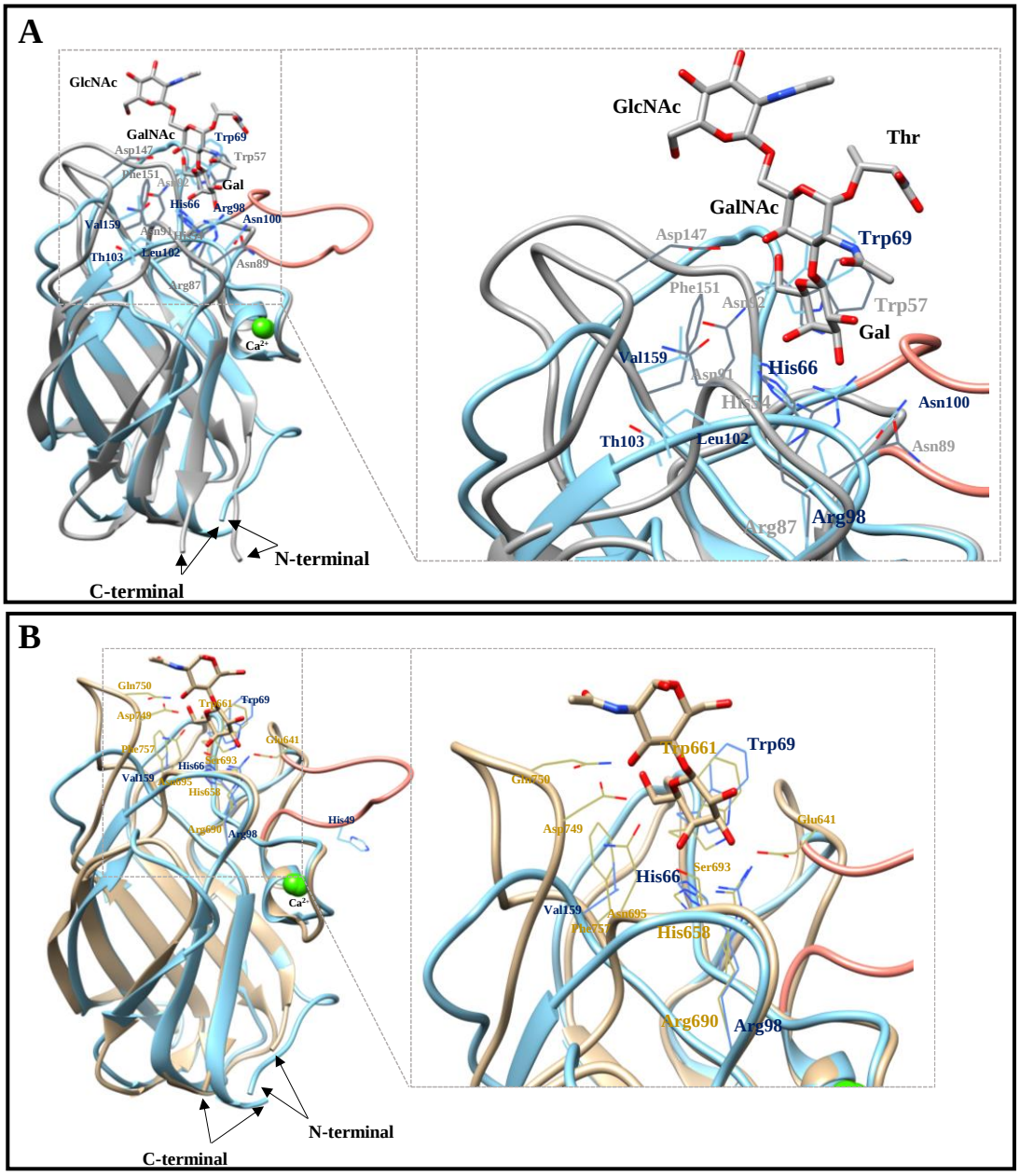


Figure 3.5- Superposition of *Bt*4040-CBM32 3D structure with *Bt*3015C-CBM32 and the CBM32 from *Cp*GH84C structures in complex with core 2 O-glycan and LacNAc, respectively. A) On the left**,** a ribbon representation of the overlay of *Bt*4040-CBM32 (sky blue) and *Bt*3015C-CBM32-core 2 O-glycan (grey, PDB accession code: 7BLJ, to be published). *Bt*4040-CBM32 amino acids are depicted as stick models and labelled in a dark blue colour and *Bt*3015C-CBM32 amino acids are coloured in light grey. *Bt*3015C-CBM32 amino acid residues responsible for the interaction with galactose and respective aligned residues of *Bt*4040-CBM32 retrieved from the multiple sequence alignment are also depicted as stick models. N- and C-termini are also labelled. On the right, a close-up view of the overlay of *Bt*4040-CBM32 (sky blue) and *Bt*3015C-CBM32-core 2 O-glycan (grey). **B.** On the left**,** a ribbon representation of the overlay of *Bt*4040-CBM32 (sky blue) and *Cp*CBM32-LacNAc (tan, PDB accession code: 2JIE). *Bt*4040-CBM32 amino acid are depicted as stick models and labelled in a dark blue colour and *Cp*CBM32 amino acids are coloured in light brown. *Cp*CBM32-LacNAc amino acid residues responsible for the interaction with galactose and respective aligned residues of *Bt*4040-CBM32 retrieved from the multiple sequence alignment are also depicted as stick models. N- and C-termini are also labelled. On the right, a close-up view of the overlay of *Bt*4040-CBM32 (sky blue) and *Cp*GH84C-LacNAc (tan). Less conserved loop from Bt4040-CBM32 is coloured in salmon.

Table 3.5- RMSDs calculated for the superpositions of *Bt*4040-CBM32 with both models

	***Bt*3015C**	***Cp*GH84C**
RMSD (Å)	0.952(88 atom pairs) 4.076 (across all 138 pairs)	0.947 (98 atom pairs) 2.955 (across all 132 pairs)

Before retrieving any conclusions regarding the residues from the binding site we need to assess which residues from both Bt3015C-CBM32 and CpGH84 interact with the ligand. Firstly, in the case of Bt3015C-CBM32 the residues that interact with the galactose moiety are the ones identified previously as the canonical binding site, more specifically His54, Trp57, Arg87, Asn92 and Phe151, denoted in Panel A from Figure 3.5. More specifically, the galactose residue makes CH-π interaction with Trp57 and makes several direct and water-mediated contacts with His54, Trp57, Arg87, Asn92 and Phe151. From the structural superposition and, in accordance with the sequence alignment, we can see that between the two structures there are three residues that are conserved: histidine, tryptophan and arginine residues. The asparagine and the phenylalanine residues are not conserved and are replaced by a tryptophan and valine respectively in Bt4040-CBM32, which could explain the different binding specificity of Bt4040-CBM32 towards Lewis A determinants. It is important to note the existence of similarity in secondary structure motifs with the conservation of the beta strands and the alpha helix that precedes the conserved structural calcium binding site. However, the loop region is less conserved, more noticeable in the loop coloured in salmon in **Figure 3.5** and also highlighted in **Figure 3.4** with the black rectangle. In fact, in the *Bt*3015C-CBM32 a residue present in one of these loops that are less conserved interacts with the core 2 O-glycan, being this residue Asp147. This residue in the sequence alignment did not align to any residue in the Bt4040-CBM32 sequence, which could be explained by the differences in this loop area, possibly leading to alteration in the binding interactions between the protein and galactose. It is also clear to see that there are residues in the loop region that interact with the galactose moiety that differ from *Bt*4040-CBM32, like Asn91 and Phe151 from Bt3015 that are replaced for Leu102 and Val159, possibly affecting ligand specificity. Regarding the family 32 CBM *Cp*GH84C in complex with LacNAc, the differences in residues from the canonical galactose binding site remain in the residues 103 and 159 of Bt4040-CBM32, that instead of being an asparagine and phenylalanine, are replaced by a tryptophan and valine which could, like mentioned earlier, explain the fact that Bt4040-CBM32 interacts with Lewis A determinants. For instance, Ser693, Asn695, Arg749, Gln750 and Phe757 are residues from *Cp*GH84C that have specificity for LacNAc and are not conserved in the case of *Bt*4040-CBM32, indicating a different ligand specificity. In one of the loops that is less conserved (coloured in salmon in Panel B) there is a glutamate residue (Glu641) in *Cp*GH84C, important for the interaction with galactose that is not conserved in *Bt*4040-CBM32, corresponding to a histidine residue (His49), possibly explaining the different ligand specificity. This structural study

elucidated the 3D structure of Bt4040, a CBM32 that has been previously studied by or group. In fact, these results regarding the differences in residue conservation corroborates with the known information that this CBM specifically bind to epitopes containing Lewis A-determinants.

Chapter 4- Integrative conclusions and future work

The work developed during this thesis focused on the main goal of studying proteins (CBMs and SusD-like proteins) from *Bacteroides thetaiotaomicron*, which are putatively involved in the recognition of carbohydrates present in the human gut.

To achieve the structural and biophysical characterization of a specific protein, recombinant protein expression and purification to yield stable proteins in suitable amounts for these studies need to take place. Before this work, recombinant protein expression in *Escherichia coli* as N-terminal His-tagged proteins has not been successful for 8 family 32 CBMs of interest and so the optimization of the expression conditions was the starting point. The main conclusion from the expression tests was that the conditions were not optimal for the CBMs production scale-up, even if different cell lines and IPTG concentrations were attempted. In future work, one of the approaches could be to alter the constructs for these CBMs. One possibility could be the insertion of the His-tag at the C-terminal.

After unsuccessfully obtaining expression for these family 32 CBMs, the focus of the work was the optimization of the recombinant expression conditions of *Bt*0865-CBM32 and *Bt*0866-SusD from PUL12, and *Bt*4040-CBM32 and *Bt*4038-SusD from PUL73. Induction with 1mM of IPTG at a temperature of 19 ^{o}C for 16 hours revealed to be successful in the expression of high amounts of soluble protein and the purification using IMAC chromatography revealed sufficient to purify the proteins to homogeneity as visualised by Coomassie blue staining.

However, guaranteeing a good protein yield may not be enough if the objective is to obtain crystals for posterior structural characterization, and the stability of the proteins in the protein purification buffer should also be tested. With this purpose, a TSA was performed. *Bt*4040-CBM32 was the most stable protein with a melting temperature of 66.71 ^{o}C. In fact, *Bt*4040-CBM32 was the protein that gave rise to good quality diffracting crystals (further discussed below). In future work, crystallization conditions will need to be optimized for *Bt*0865-CBM32, as only microcrystals were obtained for this CBM. Co- crystallizations with the glycan ligands will also be performed. This CBM has already been analysed using glycan microarrays, which determined its binding specificity for N-acetyllactosamine oligosaccharide sequences. The purpose will be to determine the 3D structure of the complex using X-ray crystallography to study at a molecular level the interaction and determine the affinity of the interaction for different glycan ligands using MST. For *Bt*4038-SusD, no crystals were obtained in the screening conditions tested. This might be related to the relatively low melting temperature of *Bt*4038-SusD, showing to be unstable and optimization of protein stability or the protein construct need to be considered for future studies. Regarding *Bt*0866-SusD, diffracting crystals were obtained, and its 3D structure solution is currently ongoing.

For *Bt*4040-CBM32, the 3D structure was successfully solved to a maximum resolution of 1.93 Å using the known structure of *B. thetaiotaomicron* CBM32 *Bt*0865 (PDB code 3GGL) as search model for molecular replacement procedures. The structure, in comparison to other characterized CBM32 structures, *Bt*3015C-CBM32 (Correia, Costa et al, unpublished) and *Cp*GH84C-CBM32 [101] revealed a conservation in secondary structure (β-sandwich fold) and a conservation of the ion metal binding site, in this case calcium. Regarding the putative binding site identified, residues that comprise the "galactose canonical" binding site were conserved, except for Asn103 and Phe159, which in the case of *Bt*4040-CBM32 are replaced by a threonine and a valine residue. This would explain a different mode of binding to the glycan ligand and a change in specificity for *Bt*4040-CBM32. Supporting this is the evidence that amino acid residues important for the interaction of the characterized CBMs with their specific glycan ligands, Asp147 in the case of *Bt*3015C-CBM32 and Glu641 in the case of *Cp*GH84C-CBM32, are positioned in less conserved loops, explaining the different mode of binding and the glycan binding specificities of the CBMs. In fact, *Bt*4040-CBM32 has been studied in the group using glycan microarrays, identifying a restricted binding specificity for Lewis A determinants, contrasting with the specificity of *Bt*3015C-CBM32 for core 1- and core 2- O-glycans and the specificity of *Cp*GH84C for LacNAc sequences. Future work will rely on the co-crystallization attempts of *Bt*4040-CBM32 with the identified glycan ligands: e.g. Lewis A trisaccharide (Gal-β-1,3(Fuc-α-1,4)GlcNAc). Similar to what is proposed for *Bt*0865-CBM32, MST will also be performed in order to calculate the affinity of the interaction for different glycan ligands.

References

[1] F. Bäckhed, R. E. Ley, J. L. Sonnenburg, D. A. Peterson, and J. I. Gordon, "Host-bacterial mutualism in the human intestine," *Science*, vol. 307, no. 5717. 2005, doi: 10.1126/science.1104816.

[2] H. Iversen, T. Lindbäck, T. M. L'Abée-Lund, N. Roos, M. Aspholm, and L. S. Arnesen, "The gut bacterium bacteroides thetaiotaomicron influences the virulence potential of the enterohemorrhagic escherichia coli O103:H25," *PLoS One*, vol. 10, no. 2, pp. 1–23, 2015, doi: 10.1371/journal.pone.0118140.

[3] F. H. Karlsson, D. W. Ussery, J. Nielsen, and I. Nookaew, "A Closer Look at Bacteroides: Phylogenetic Relationship and Genomic Implications of a Life in the Human Gut," *Microb. Ecol.*, vol. 61, no. 3, pp. 473–485, 2011, doi: 10.1007/s00248-010-9796-1.

[4] A. G. Wexler and A. L. Goodman, "An insider's perspective: Bacteroides as a window into the microbiome," *Nat. Microbiol.*, vol. 2, no. April, pp. 1–11, 2017, doi: 10.1038/nmicrobiol.2017.26.

[5] R. E. Ley *et al.*, "and Their Gut Microbes," *Wild*, vol. 1647, no. November, pp. 1647–1652, 2008, doi: 10.1126/science.1155725.

[6] E. R. Rocha and C. J. Smith, "Ferritin-like family proteins in the anaerobe Bacteroides fragilis: When an oxygen storm is coming, take your iron to the shelter," *BioMetals*, vol. 26, no. 4, pp. 577–591, 2013, doi: 10.1007/s10534-013-9650-2.

[7] A. D. Baughn and M. H. Malamy, "The strict anaerobe Bacteroides fragilis grows in and benefits from nanomolar concentrations of oxygen," *Nature*, vol. 427, no. 6973, pp. 441–444, 2004, doi: 10.1038/nature02285.

[8] J. Xu *et al.*, "A genomic view of the human-Bacteroides thetaiotaomicron symbiosis," *Science (80-.).*, vol. 299, no. 5615, pp. 2074–2076, 2003, doi: 10.1126/science.1080029.

[9] E. C. Martens, H. C. Chiang, and J. I. Gordon, "Mucosal Glycan Foraging Enhances Fitness and Transmission of a Saccharolytic Human Gut Bacterial Symbiont," *Cell Host Microbe*, vol. 4, no. 5, pp. 447–457, 2008, doi: 10.1016/j.chom.2008.09.007.

[10] J. L. Sonnenburg *et al.*, "Glycan foraging in vivo by an intestine-adapted bacterial symbiont," *Science (80-.).*, 2005, doi: 10.1126/science.1109051.

[11] N. M. Koropatkin, E. A. Cameron, and E. C. Martens, "How glycan metabolism shapes the human gut microbiota," *Nat. Rev. Microbiol.*, vol. 10, no. 5, pp. 323–335, 2012, doi: 10.1038/nrmicro2746.

[12] B. Dalile, L. Van Oudenhove, B. Vervliet, and K. Verbeke, "The role of short-chain fatty acids in microbiota–gut–brain communication," *Nat. Rev. Gastroenterol. Hepatol.*, vol. 16, no. 8, pp. 461–478, 2019, doi: 10.1038/s41575-019-0157-3.

[13] H. M. Hamer, D. Jonkers, K. Venema, S. Vanhoutvin, F. J. Troost, and R. J. Brummer, "Review article: The role of butyrate on colonic function," *Aliment. Pharmacol. Ther.*, vol. 27, no. 2, pp. 104–119, 2008, doi: 10.1111/j.1365-2036.2007.03562.x.

[14] J. P. Segain *et al.*, "Butyrate inhibits inflammatory responses through NFκB inhibition: Implications for Crohn's disease," *Gut*, vol. 47, no. 3, pp. 397–403, 2000, doi: 10.1136/gut.47.3.397.

[15] D. Meyer and M. Stasse-Wolthuis, "The bifidogenic effect of inulin and oligofructose and its consequences for gut health," *Eur. J. Clin. Nutr.*, vol. 63, no. 11, pp. 1277–1289, 2009, doi: 10.1038/ejcn.2009.64.

[16] A. M. Neyrinck, S. Possemiers, W. Verstraete, F. De Backer, P. D. Cani, and N. M. Delzenne, "Dietary modulation of clostridial cluster XIVa gut bacteria (Roseburia spp.) by chitin-glucan fiber improves host metabolic alterations induced by high-fat diet in mice," *J. Nutr. Biochem.*, vol. 23, no. 1, pp. 51–59, 2012, doi: 10.1016/j.jnutbio.2010.10.008.

[17] S. McKeen, W. Young, K. Fraser, N. C. Roy, and W. C. McNabb, "Glycan utilisation and function in the microbiome of weaning infants," *Microorganisms*, vol. 7, no. 7, pp. 1–18, 2019, doi: 10.3390/microorganisms7070190.

[18] H. Ghazarian, B. Idoni, and S. B. Oppenheimer, "A glycobiology review: Carbohydrates, lectins and implications in cancer therapeutics," *Acta Histochemica*. 2011, doi: 10.1016/j.acthis.2010.02.004.

[19] R. G. Spiro, "Protein glycosylation: Nature, distribution, enzymatic formation, and disease implications of glycopeptide bonds," *Glycobiology*, vol. 12, no. 4, 2002, doi: 10.1093/glycob/12.4.43R.

[20] C. Reily, T. J. Stewart, M. B. Renfrow, and J. Novak, "Glycosylation in health and disease," *Nat. Rev. Nephrol.*, vol. 15, no. 6, pp. 346–366, 2019, doi: 10.1038/s41581-019-0129-4.

[21] M. Wiciński, E. Sawicka, J. Gębalski, K. Kubiak, and B. Malinowski, "Human milk oligosaccharides: Health benefits, potential applications in infant formulas, and pharmacology," *Nutrients*, 2020, doi: 10.3390/nu12010266.

[22] C. Breton, L. Šnajdrová, C. Jeanneau, J. Koča, and A. Imberty, "Structures and mechanisms of glycosyltransferases," *Glycobiology*. 2006, doi: 10.1093/glycob/cwj016.

[23] R. A. Laine, "Invited commentary: A calculation of all possible oligosaccharide isomers both branched and linear yields 1.05 × 10 structures for a reducing hexasaccharide: The Isomer Barrier to development of single-method saccharide sequencing or synthesis systems," *Glycobiology*, vol. 4, no. 6, pp. 759–767, 1994, doi: 10.1093/glycob/4.6.759.

[24] D. J. Gill, H. Clausen, and F. Bard, “Location, location, location: New insights into O-GalNAc protein glycosylation,” *Trends in Cell Biology*. 2011, doi: 10.1016/j.tcb.2010.11.004.

[25] E. P. Bennett, U. Mandel, H. Clausen, T. A. Gerken, T. A. Fritz, and L. A. Tabak, “Control of mucin-type O-glycosylation: A classification of the polypeptide GalNAc-transferase gene family,” *Glycobiology*, vol. 22, no. 6, pp. 736–756, 2012, doi: 10.1093/glycob/cwr182.

[26] D. Vasudevan and R. S. Haltiwanger, “Novel roles for O-linked glycans in protein folding,” *Glycoconj. J.*, vol. 31, no. 6, pp. 417–426, 2014, doi: 10.1007/s10719-014-9556-4.

[27] K. S. B. Bergstrom and L. Xia, “Mucin-type O-glycans and their roles in intestinal homeostasis,” *Glycobiology*, vol. 23, no. 9, pp. 1026–1037, 2013, doi: 10.1093/glycob/cwt045.

[28] R. B. Sartor, “Microbial Influences in Inflammatory Bowel Diseases,” *Gastroenterology*, 2008, doi: 10.1053/j.gastro.2007.11.059.

[29] M. E. V. Johansson *et al.*, “Composition and functional role of the mucus layers in the intestine,” *Cellular and Molecular Life Sciences*. 2011, doi: 10.1007/s00018-011-0822-3.

[30] C. Atuma, V. Strugala, A. Allen, and L. Holm, “The adherent gastrointestinal mucus gel layer: Thickness and physical state in vivo,” *Am. J. Physiol. - Gastrointest. Liver Physiol.*, 2001, doi: 10.1152/ajpgi.2001.280.5.g922.

[31] A. Swidsinski *et al.*, “Comparative study of the intestinal mucus barrier in normal and inflamed colon,” *Gut*, 2007, doi: 10.1136/gut.2006.098160.

[32] N. M. J. Schwerbrock *et al.*, “Interleukin 10-deficient mice exhibit defective colonic Muc2 synthesis before and after induction of colitis by commensal bacteria,” *Inflamm. Bowel Dis.*, 2004, doi: 10.1097/00054725-200411000-00016.

[33] F. Marin, G. Luquet, B. Marie, and D. Medakovic, “Molluscan Shell Proteins: Primary Structure, Origin, and Evolution,” *Current Topics in Developmental Biology*. 2007, doi: 10.1016/S0070-2153(07)80006-8.

[34] Y. Niv, “MUC1 and colorectal cancer pathophysiology considerations,” *World Journal of Gastroenterology*. 2008, doi: 10.3748/wjg.14.2139.

[35] I. Brockhausen, “Biosynthesis of Mucin-Type O-Glycans,” in *Comprehensive Glycoscience: From Chemistry to Systems Biology*, 2007.

[36] M. Bäckström, D. Ambort, E. Thomsson, M. E. V. Johansson, and G. C. Hansson, “Increased understanding of the biochemistry and biosynthesis of MUC2 and other gel-forming mucins through the recombinant expression of their protein domains,” *Mol. Biotechnol.*, 2013, doi: 10.1007/s12033-012-9562-3.

[37] E. P. Bennett, U. Mandel, H. Clausen, T. A. Gerken, T. A. Fritz, and L. A. Tabak, "Control of mucin-type O-glycosylation: A classification of the polypeptide GalNAc-transferase gene family," *Glycobiology*. 2012, doi: 10.1093/glycob/cwr182.

[38] L. E. Tailford, E. H. Crost, D. Kavanaugh, and N. Juge, "Mucin glycan foraging in the human gut microbiome," *Front. Genet.*, vol. 5, no. FEB, 2015, doi: 10.3389/fgene.2015.00081.

[39] C. Robbe, C. Capon, B. Coddeville, and J. C. Michalski, "Structural diversity and specific distribution of O-glycans in normal human mucins along the intestinal tract," *Biochem. J.*, 2004, doi: 10.1042/BJ20040605.

[40] A. Magalhes, M. N. Ismail, and C. A. Reis, "Sweet receptors mediate the adhesion of the gastric pathogen Helicobacter pylori: Glycoproteomic strategies," *Expert Review of Proteomics*. 2010, doi: 10.1586/epr.10.18.

[41] E. G. Zoetendal, A. Von Wright, T. Vilpponen-Salmela, K. Ben-Amor, A. D. L. Akkermans, and W. M. De Vos, "Mucosa-associated bacteria in the human gastrointestinal tract are uniformly distributed along the colon and differ from the community recovered from feces," *Appl. Environ. Microbiol.*, 2002, doi: 10.1128/AEM.68.7.3401-3407.2002.

[42] M. E. V. Johansson, M. Phillipson, J. Petersson, A. Velcich, L. Holm, and G. C. Hansson, "The inner of the two Muc2 mucin-dependent mucus layers in colon is devoid of bacteria," *Proc. Natl. Acad. Sci. U. S. A.*, 2008, doi: 10.1073/pnas.0803124105.

[43] M. S. Desai *et al.*, "A Dietary Fiber-Deprived Gut Microbiota Degrades the Colonic Mucus Barrier and Enhances Pathogen Susceptibility," *Cell*, 2016, doi: 10.1016/j.cell.2016.10.043.

[44] K. E. NORIN, B. E. GUSTAFSSON, B. S. LINDBLAD, and T. MIDTVEDT, "The Establishment of Some Microflora Associated Biochemical Characteristics in Feces from Children during the First Years of Life," *Acta Pædiatrica*, 1985, doi: 10.1111/j.1651-2227.1985.tb10951.x.

[45] E. C. Martens, N. M. Koropatkin, T. J. Smith, and J. I. Gordon, "Complex glycan catabolism by the human gut microbiota: The bacteroidetes sus-like paradigm," *Journal of Biological Chemistry*. 2009, doi: 10.1074/jbc.R109.022848.

[46] J. Y. Huang, S. M. Lee, and S. K. Mazmanian, "The human commensal Bacteroides fragilis binds intestinal mucin," *Anaerobe*, 2011, doi: 10.1016/j.anaerobe.2011.05.017.

[47] A. O'Callaghan and D. van Sinderen, "Bifidobacteria and their role as members of the human gut microbiota," *Frontiers in Microbiology*. 2016, doi: 10.3389/fmicb.2016.00925.

[48] L. C. Hoskins, M. Agustines, W. B. McKee, E. T. Boulding, M. Kriaris, and G. Niedermeyer, "Mucin degradation in human colon ecosystems. Isolation and properties of fecal strains that degrade ABH- blood group antigens and oligosaccharides from mucin glycoproteins," *J. Clin. Invest.*, 1985, doi: 10.1172/JCI111795.

[49] M. C. Collado, M. Derrien, E. Isolauri, W. M. De Vos, and S. Salminen, "Intestinal integrity and Akkermansia muciniphila, a mucin-degrading member of the intestinal microbiota present in infants, adults, and the elderly," *Appl. Environ. Microbiol.*, 2007, doi: 10.1128/AEM.01477-07.

[50] F. Turroni, C. Milani, S. Duranti, J. Mahony, D. van Sinderen, and M. Ventura, "Glycan Utilization and Cross-Feeding Activities by Bifidobacteria," *Trends in Microbiology*. 2018, doi: 10.1016/j.tim.2017.10.001.

[51] A. El Kaoutari, F. Armougom, J. I. Gordon, D. Raoult, and B. Henrissat, "The abundance and variety of carbohydrate-active enzymes in the human gut microbiota," *Nat. Rev. Microbiol.*, 2013, doi: 10.1038/nrmicro3050.

[52] M. K. Bjursell, E. C. Martens, and J. I. Gordon, "Functional genomic and metabolic studies of the adaptations of a prominent adult human gut symbiont, Bacteroides thetaiotaomicron, to the suckling period," *J. Biol. Chem.*, 2006, doi: 10.1074/jbc.M606509200.

[53] M. H. Foley, D. W. Cockburn, and N. M. Koropatkin, "The Sus operon: a model system for starch uptake by the human gut Bacteroidetes," *Cellular and Molecular Life Sciences*. 2016, doi: 10.1007/s00018-016-2242-x.

[54] A. A. Salyers, J. R. Vercellotti, S. E. H. West, and T. D. Wilkins, "Fermentation of mucin and plant polysaccharides by strains of Bacteroides from the human colon," *Appl. Environ. Microbiol.*, 1977, doi: 10.1128/aem.33.2.319-322.1977.

[55] J. A. Shipman, J. E. Berleman, and A. A. Salyers, "Characterization of four outer membrane proteins involved in binding starch to the cell surface of Bacteroides thetaiotaomicron," *J. Bacteriol.*, 2000, doi: 10.1128/JB.182.19.5365-5372.2000.

[56] E. Tancula, M. J. Feldhaus, L. A. Bedzyk, and A. A. Salyers, "Location and characterization of genes involved in binding of starch to the surface of Bacteroides thetaiotaomicron," *J. Bacteriol.*, 1992, doi: 10.1128/jb.174.17.5609-5616.1992.

[57] A. B. Boraston, D. N. Bolam, H. J. Gilbert, and G. J. Davies, "Carbohydrate-binding modules: Fine-tuning polysaccharide recognition," *Biochemical Journal*. 2004, doi: 10.1042/BJ20040892.

[58] H. J. Gilbert, J. P. Knox, and A. B. Boraston, "Advances in understanding the molecular basis of plant cell wall polysaccharide recognition by carbohydrate-binding modules," *Current Opinion in Structural Biology*. 2013, doi: 10.1016/j.sbi.2013.05.005.

[59] E. Ficko-Blean and A. B. Boraston, "Insights into the recognition of the human glycome by microbial carbohydrate-binding modules," *Current Opinion in Structural Biology*. 2012, doi: 10.1016/j.sbi.2012.07.009.

[60] A. B. Boraston, E. Ficko-Blean, and M. Healey, "Carbohydrate recognition by a large sialidase toxin from Clostridium perfringens," *Biochemistry*, vol. 46, no. 40, pp. 11352–11360, 2007, doi: 10.1021/bi701317g.

[61] S. Manco, F. Hernon, H. Yesilkaya, J. C. Paton, P. W. Andrew, and A. Kadioglu, "Pneumococcal neuraminidases A and B both have essential roles during infection of the respiratory tract and sepsis," *Infect. Immun.*, 2006, doi: 10.1128/IAI.01237-05.

[62] A. B. Boraston, T. J. Revett, C. M. Boraston, D. Nurizzo, and G. J. Davies, "Structural and thermodynamic dissection of specific mannan recognition by a carbohydrate binding module, TmCBM27," *Structure*, 2003, doi: 10.1016/S0969-2126(03)00100-X.

[63] A. B. Boraston, V. Notenboom, R. A. J. Warren, D. G. Kilburn, D. R. Rose, and G. Davies, "Structure and ligand binding of carbohydrate-binding module CsCBM6-3 reveals similarities with fucose-specific lectins and 'galactose-binding' domains," *J. Mol. Biol.*, 2003, doi: 10.1016/S0022-2836(03)00152-9.

[64] C. L. Lawson *et al.*, "Nucleotide sequence and X-ray structure of cyclodextrin glycosyltransferase from Bacillus circulans strain 251 in a maltose-dependent crystal form," *J. Mol. Biol.*, 1994, doi: 10.1006/jmbi.1994.1168.

[65] S. Armenta, S. Moreno-Mendieta, Z. Sánchez-Cuapio, S. Sánchez, and R. Rodríguez-Sanoja, "Advances in molecular engineering of carbohydrate-binding modules," *Proteins: Structure, Function and Bioinformatics*. 2017, doi: 10.1002/prot.25327.

[66] S. M. Southall, P. J. Simpson, H. J. Gilbert, G. Williamson, and M. P. Williamson, "The starch-binding domain from glucoamylase disrupts the structure of starch," *FEBS Lett.*, 1999, doi: 10.1016/S0014-5793(99)00263-X.

[67] A. L. Van Bueren, M. Higgins, D. Wang, R. D. Burke, and A. B. Boraston, "Identification and structural basis of binding to host lung glycogen by streptococcal virulence factors," *Nat. Struct. Mol. Biol.*, 2007, doi: 10.1038/nsmb1187.

[68] C. Montanier *et al.*, "Evidence that family 35 carbohydrate binding modules display conserved specificity but divergent function," *Proc. Natl. Acad. Sci. U. S. A.*, 2009, doi: 10.1073/pnas.0808972106.

[69] N. M. Koropatkin, E. C. Martens, J. I. Gordon, and T. J. Smith, "Starch Catabolism by a Prominent Human Gut Symbiont Is Directed by the Recognition of Amylose Helices," *Structure*, 2008, doi: 10.1016/j.str.2008.03.017.

[70] L. D. D'Andrea and L. Regan, "TPR proteins: The versatile helix," *Trends in Biochemical Sciences*. 2003, doi: 10.1016/j.tibs.2003.10.007.

[71] N. Koropatkin, E. C. Martens, J. I. Gordon, and T. J. Smith, "Structure of a SusD homologue, BT1043, involved in mucin O-glycan utilization in a prominent human gut symbiont," *Biochemistry*, 2009, doi: 10.1021/bi801942a.

[72] C. D. Rillahan and J. C. Paulson, "Glycan microarrays for decoding the glycome," *Annu. Rev. Biochem.*, 2011, doi: 10.1146/annurev-biochem-061809-152236.

[73] A. S. Palma, T. Feizi, R. A. Childs, W. Chai, and Y. Liu, "The neoglycolipid (NGL)-based oligosaccharide microarray system poised to decipher the meta-glycome," *Curr. Opin. Chem. Biol.*, vol. 18, no. 1, pp. 87–94, 2014, doi: 10.1016/j.cbpa.2014.01.007.

[74] A. S. Palma *et al.*, "Unravelling glucan recognition systems by glycome microarrays using the designer approach and mass spectrometry," *Mol. Cell. Proteomics*, 2015, doi: 10.1074/mcp.M115.048272.

[75] Y. Liu, A. S. Palma, and T. Feizi, "Carbohydrate microarrays: Key developments in glycobiology," *Biol. Chem.*, vol. 390, no. 7, pp. 647–656, 2009, doi: 10.1515/BC.2009.071.

[76] T. Horlacher and P. H. Seeberger, "Carbohydrate arrays as tools for research and diagnostics," *Chem. Soc. Rev.*, 2008, doi: 10.1039/b708016f.

[77] X. Song, J. Heimburg-Molinaro, R. D. Cummings, and D. F. Smith, "Chemistry of natural glycan microarrays," *Current Opinion in Chemical Biology*. 2014, doi: 10.1016/j.cbpa.2014.01.001.

[78] J. C. Gildersleeve, O. Oyelaran, J. T. Simpson, and B. Allred, "Improved procedure for direct coupling of carbohydrates to proteins via reductive amination," *Bioconjug. Chem.*, 2008, doi: 10.1021/bc800153t.

[79] D. O. Ribeiro, B. A. Pinheiro, A. L. Carvalho, and A. S. Palma, "Targeting protein-carbohydrate interactions in plant cell-wall biodegradation: The power of carbohydrate microarrays," *Carbohydr. Chem.*, 2018, doi: 10.1039/9781788010641-00159.

[80] P. W. Tang, H. C. Gool, M. Hardy, Y. C. Lee, and T. Felzi, "Novel approach to the study of the antigenicities and receptor functions of carbohydrate chains of glycoproteins," *Biochem. Biophys. Res. Commun.*, 1985, doi: 10.1016/0006-291X(85)91158-1.

[81] S. Fukui, T. Feizi, C. Galustian, A. M. Lawson, and W. Chai, "Oligosaccharide microarrays for high-throughput detection and specificity assignments of carbohydrate-protein interactions," *Nat. Biotechnol.*, 2002, doi: 10.1038/nbt735.

[82] Y. Liu *et al.*, "Neoglycolipid-based oligosaccharide microarray system: Preparation of ngls and their noncovalent immobilization on nitrocellulose-coated glass slides for microarray analyses," *Methods Mol. Biol.*, 2012, doi: 10.1007/978-1-61779-373-8_8.

[83] M. Jerabek-Willemsen *et al.*, "MicroScale Thermophoresis: Interaction analysis and beyond," *J. Mol. Struct.*, 2014, doi: 10.1016/j.molstruc.2014.03.009.

[84] S. A. I. Seidel *et al.*, "Label-free microscale thermophoresis discriminates sites and affinity of protein-ligand binding," *Angew. Chemie - Int. Ed.*, 2012, doi: 10.1002/anie.201204268.

[85] A. McPherson and J. A. Gavira, "Introduction to protein crystallization," *Acta Crystallographica Section F:Structural Biology Communications*. 2014, doi: 10.1107/S2053230X13033141.

[86] M. S. Smyth and J. H. J. Martin, "x Ray crystallography," *J. Clin. Pathol. - Mol. Pathol.*, vol. 53, no. 1, pp. 8–14, 2000, doi: 10.1136/mp.53.1.8.

[87] G. Rhodes, "An Overview of Protein Crystallography," in *Crystallography Made Crystal Clear*, 2006.

[88] G. Taylor, "The phase problem," in *Acta Crystallographica - Section D Biological Crystallography*, 2003, doi: 10.1107/S0907444903017815.

[89] V. Gaberc-Porekar and V. Menart, "Perspectives of immobilized-metal affinity chromatography," *Journal of Biochemical and Biophysical Methods*. 2001, doi: 10.1016/S0165-022X(01)00207-X.

[90] M. L. Dart *et al.*, "Homogeneous Assay for Target Engagement Utilizing Bioluminescent Thermal Shift," *ACS Med. Chem. Lett.*, 2018, doi: 10.1021/acsmedchemlett.8b00081.

[91] D. W. Abbott, J. M. Eirín-López, and A. B. Boraston, "Insight into ligand diversity and novel biological roles for family 32 carbohydrate-binding modules," *Mol. Biol. Evol.*, 2008, doi: 10.1093/molbev/msm243.

[92] S. Etzold and N. Juge, "Structural insights into bacterial recognition of intestinal mucins," *Current Opinion in Structural Biology*. 2014, doi: 10.1016/j.sbi.2014.07.002.

[93] S. L. Newstead, J. N. Watson, A. J. Bennet, and G. Taylor, "Galactose recognition by the carbohydrate-binding module of a bacterial sialidase," *Acta Crystallogr. Sect. D Biol. Crystallogr.*, 2005, doi: 10.1107/S0907444905026132.

[94] P. Evans and A. McCoy, "An introduction to molecular replacement," *Acta Crystallogr. Sect. D Biol. Crystallogr.*, vol. 64, no. 1, pp. 1–10, 2007, doi: 10.1107/S0907444907051554.

[95] P. D. Adams *et al.*, "PHENIX: A comprehensive Python-based system for macromolecular structure solution," *Acta Crystallogr. Sect. D Biol. Crystallogr.*, 2010, doi: 10.1107/S0907444909052925.

[96] G. N. Murshudov *et al.*, "REFMAC5 for the refinement of macromolecular crystal structures," *Acta Crystallogr. Sect. D Biol. Crystallogr.*, 2011, doi: 10.1107/S0907444911001314.

[97] "The CCP4 suite: Programs for protein crystallography," *Acta Crystallogr. Sect. D Biol. Crystallogr.*, 1994, doi: 10.1107/S0907444994003112.

[98] P. Emsley, B. Lohkamp, W. G. Scott, and K. Cowtan, “Features and development of Coot,” *Acta Crystallogr. Sect. D Biol. Crystallogr.*, 2010, doi: 10.1107/S0907444910007493.

[99] F. Sievers *et al.*, “Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega,” *Mol. Syst. Biol.*, 2011, doi: 10.1038/msb.2011.75.

[100] E. Ficko-Blean *et al.*, “Carbohydrate recognition by an architecturally complex α-n-acetylglucosaminidase from Clostridium perfringens,” *PLoS One*, 2012, doi: 10.1371/journal.pone.0033524.

[101] E. Ficko-Blean and A. B. Boraston, “The interaction of a carbohydrate-binding module from a Clostridium perfringens N-acetyl-β-hexosaminidase with its carbohydrate receptor,” *J. Biol. Chem.*, vol. 281, no. 49, pp. 37748–37757, 2006, doi: 10.1074/jbc.M606126200.

Supplementary Material

Supplementary information 1:

Luria Bertani medium culture used for the IPTG-induction protocol 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.

Supplementary information 2:

Protocol for plasmid DNA purification from *Escherichia coli* cells

All centrifugations should be carried out at room temperature in a table-top microcentrifuge at >12000 x*g* (10000-15000 rpm depending on the rotor type).

1. Cultivate and harvest bacterial cells

Pellet 1-5 mL of an *E. coli* LB culture for 30 s. Discard supernatant. Remove as much media as possible. For low copy number plasmids double the volume of cells and of lysis Buffers A1, A2 and A3.

2. Cell lysis

Re-suspend cell pellet in 250 µL Buffer A1 by vigorous vortexing.

Add 250 µL of Buffer A2 and mix gently by inverting the tube for 6-8 times. Incubate at room temperature for a maximum of 4 min. Do not vortex.

Add 300 µL Buffer A3. Mix gently by inverting the tube for 6-8 times. Do not vortex.

3. Clarification of lysate

Centrifuge for 5-10 min at room temperature, depending on initial culture volume.

4. Bind DNA

Place NZYTech spin column in a 2 mL collecting tube and load the supernatant from step 3 onto the column. Centrifuge for 1 min at 11,000 x*g*. Discard flow-through.

5. Wash silica membrane

Add 500 µL of Buffer AY onto the column. Centrifuge for 1 min. Discard flow-through. This step is crucial to increase the reading length of DNA sequencing reactions and to improve the performance of critical enzymatic reactions. When using endA+ strains, such as JM series, HB101 and its derivatives, or any wild-type strain, use pre-warmed Buffer AY (50 °C).

Add 600 µL of Buffer A4 (make sure ethanol was previously added). Centrifuge for 1 min. Discard flow-through.

6. Dry silica membrane

Re-insert the NZYTech spin column into the empty 2 mL collecting tube and centrifuge for 2 min.

7. Elute highly pure DNA

Place the dried NZYTech spin column into a clean 1.5 mL microcentrifuge tube and add 50 µL of Buffer AE. Incubate 1 min at room temperature. Centrifuge for 1 min. By repeating this step the overall yield will increase by 15-20%. To obtain a highly concentrated miniprep (1.3 times higher) reduce the volume of elution buffer to 30 µL. Store the purified DNA at -20 °C.

Supplementary Table 1 - Family 32 CBMs from *B. thetaiotaomicron* used in the small scale expression tests. Each CBM's recombinant protein sequence, protein identification, molecular weight, extinction coefficient and isoelectric point are depicted. N-terminal His-tag is highlighted in yellow.

Microorganism	Protein ID	Recombinant protein sequence	Molecular weight (kDa)	Extinction coefficient (M^{-1})	Isoelectric point
B. thetaiotaomicron	*Bt*1775	MGSSHHHHHHSSGPQQGLRAESFRVEMEACDYFFDVSAKNSGNQYRTGSLDV AKIAGVAPDKNTEWYVTSTEAGEWMEWKELPYAAGDITVKVCYAAKEDAKI RFDFGEGTKRRQGPVVELHATGGEWVTVDAFTMKSDVNGWRRTVLNIVAGK PDLNYFTIT	18.20	36565	5.96
	*Bt*2194	MGSSHHHHHHSSGPQQGLRWQVVEVSSEETSGEGSDNGHAIHAFDGLKGTFW HTQWKGNEPQPPHHIVVDMGQEVKMLGFQYVSRDHGEAWPQEMTMETSLD GSKWESAGTYSDLPAGAKEEFRSYFPGFKQARYFRLTITAVYSGKWATVVAEI NAISI	17.91	40450	5.9
	*Bt*3014	MGSSHHHHHHSSGPQQGLRAPFPTRPIKDLVDGNIATFFHSWWSSLVEMPHYL VVDLGEEVSAIKFRSTNTNRANDSSWKTINLYTSDSYNPAEWFDGVEKIDGNT VDISQAGTHKETTLTGLPDGVSEVYNSEVIPLSKPSRYLWFEVTETTKGTPYFA LGELEIYQC	19.00	37930	5.36
	*Bt*3015_C	MGSSHHHHHHSSGPQQGLRKGDIKIKIVSGTASSFQSGSNIEKSFDGDYSTLYH SSWSNGASNYFPITLTYNFETVTDVDYLIYHPRNNGNNGRFKETEIQYSADGHT FTKLIDKDFQGSATAGKVTFDQTIQAKSFRFIVKSGSGDGQGFASCAEMEFFAK	17.89	15930	6.74
	*Bt*3592	MGSSHHHHHHSSGPQQGLREEVKEIWLDELGESSYYIQDWGLPRINKAVTMTP LTVKGIVYERGIGTHAISRMLFDIGKKAKTLSGLAGADDNTPFACNLQFKILGD RKELWRSGIMRKGDPAKPFNIDLSGIDKVLLLVEECGDGMMYDRADWLNVKF TT	18.07	28085	6.65
	*Bt*4132	MGSSHHHHHHSSGPQQGLRSKEEWDEIHNQYIDVDRTGWGVEANTEELTGEG AVNGHKEALIDGNLNTFWHSQWDGEGKNPPLPHIIIFDMQQTQNILSIELARRQ NNLDLKAVMFSISDDKENWTELGKLDFPNDKVPNAQIILLPKAISGRYFRTTVT DSNNGVNASIAEIMF	19.76	30480	5.24
	*Bt*4245	MGSSHHHHHHSSGPQQGLRLSSNAIEPSEGSLANLLDGDIGTYFHSAWSVSIAD KHYVQVKLPVSTKTFRFTYTNRSNNGNAALAWFNLYTGTNENNLQLYKRFA WDEDGLPSGAAGVYVSPDVSIDNAANTLRFKCESNWTGGSFFVWSEFSLF	17.18	36440	6.2
	*Bt*4270_D	MGSSHHHHHHSSGPQQGLRFSANATEPSEGSLAALVDDNINTYYHTIWSGTSP NKQPHYLQINMSELPLQSLRFEYDGRNNGNGAGDVKRVGIWGSDNNNTWTL MGKETYTLPGSRGQHVEPNENIKAGKPYKYIRFIPEARRDADPIDPSGGNGWW NMAGIYLYK	18.57	40910	6.75
	*Bt*4295_D	MGSSHHHHHHSSGPQQGLRSDNANHVGDGGGLPALIDGKVNTYYHTKWNAP VTTEAHYVQIKLNKPLKDLCFEYDARQSGVNNGGDVKAATIYGSMNGEFFES MGNEEFNLPTTNGGHATAKNNVSGKQAYNYIRFTPTARRDKDPLDYTVAGSA WWNMSEIYLY	18.25	31400	6.54

Supplementary Table 2 – Family 32 CBMs and SusD-like proteins from *B. thetaiotaomicron* used for the large scale expression and purification. Each CBM's recombinant protein sequence, protein identification, molecular weight, extinction coefficient and isoelectric point are depicted. N-terminal His-tag is highlighted in yellow.

Microorganism	Protein ID	Recombinant protein sequence	Molecular weight (kDa)	Extinction coefficient	Isoelectric point
B. thetaiotaomicron	*Bt*0865	MGSSHHHHHHSSGPQQGLRCIIKEGTKISKSGWEVLSFTTQEASGEGAGNGLA KCLIDGDTETFWHAKWQGGSDPLPYDIVIDMKQNIQIAQVELLPRGRGSNNPIK VVEFAASEDNVNWTPIGRFGFTNQDAALEYYVKSIKARYIRLTIPDDGGNSTV AAIRELDVKGT	18.70	28085	6.14
	*Bt*4040	MGSSHHHHHHSSGPQQGLRKQGGPINRDHWTVTANTEEKTGEVSAAYPHNG QTISLIDGEINTFWHSKWQGGEVAPPYEIILDMGKENSVSQLGLIARQNALTSM NLEIYAGDDGETWNLIGKYFFDGSIKTEQMVPVKACDARWIRIYIPTLGSGSSV GHLAEIYVYGTD	18.86	37930	5.74
	*Bt*0866	MGSSHHHHHHSSGPQQGLRDYLAVSDQMSGGLQNTDQIFENVAYTKRWYAN VFAGIPDYSGINSLNVGAFKNPWAAICDELVVGYGNAAKANNSDKNAATAGF HRYGDCYKYIRQANIFLEKAHVITTSGTQGDRLEEDELNEMRANVRFMRAFY NYLLLEQYGPIILVKDKVYEATETQDVPRNTVDEVITYIDQELREVANELPQEP MHENESYRAWPTKGVALAVRAKLWLYAASPLLNGGYREALSLTNPDGTRLFP DRDDNKWNTALNACKDFIDYAETGNRYELYKEYTTSSTGEQILDVDASVYNL FQKYNKEIIWGTANNDWGGLDGDAFDRRIVPRCEKNGLGSTGVTQELVDAFY MNDGLPIKETDYLPKSTLYKEDGYGTYKDKNDGKYSKNYTNVTVSNRYLNRE ARFYNTVFFNGRQWPVTCKQVQFYNGGNAGVQEGQATTTGYMLFKRFNRSIS KTSPGVASQNRPSIIFRLADFYLIYAEVANEVNPSDSRVLTYLNLVRERAGLPK VEILNPGIVGNKELQRAAIQRERQIELATEGQRYFDVRRWMIADKDGEGRQNG YAHGMNVRGTINDTEEFNRVVETEKIVFNRKMYLQPIPDHEMRKTQNLVQNP GW	71.88	113360	5.89
	*Bt*4038	MGSSHHHHHHSSGPQQGLRSDYLDVSDDLAAELSMEEVFNNTGYARRFHRYI YSGIPDVSNIIITSSYAALTGLDNPWPAVSDELKSAQNNVKTIPTVGYHAGSAD LSRWSLYKQIRQANEFLAYSHTIPQQGDVTDYIDEDELARLKNEARFLRAYYH YLLFELYGPIPIMTEISEPSASDLDYYRNSVDEVVQFIDSELNECYNELPEKEVN DDGTPNENRSAAPTKGAALAILAKLHVYAASPLLNGGYSEAIALRDNQDKQLF PAKDDKKWQTALNALQRFIDYAKTHYHLYKEKDKDGELNAEESLYQLFQVSL NNPEAIWQTSKNSWGDVGGEGRERRCTPRAIYSGFGCVGVLQEAIDDFYMND GKSIHESGLYSEEGIGEDGIPNMYKNREPRFYQAITYSGKTWQKTTKQIYFYKG SGDDNSKADMSYSGYLLYKGMNRDLLNQGSNAKSKYRAGMLFRLADFYLLY AEALNHVNPSDERIIAHIDSVRYRAGIPLLKDIKPEIKGNQALQEEAIRKERRIEL FAEGQRYFDVRRWMCADEEGYKQGGPVHGMNMNADNLEDFMERTAFETRIF ERRMYLYPIPLNEIQKSSKLVQNPGW	69.64	105340	5.35

	1	2	3	4	5	6	7	8	9	10	11	12
A	Citric acid pH 4.0	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
B	MES pH 6.5	PIPES pH 6.7	MOPS 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 8.5	Tris pH 8.5
C	Bicine pH 9.0	CAPS pH 9.0	Glycine-HCl pH 9.5	Glycine-HCl pH 10	CAPS pH 10.5	CAPS pH 11	Citric acid pH 4.0 + 300 mM NaCl	Sodium acetate pH 4.5 + 300 mM NaCl	Sodium citrate pH 4.7 + 300 mM NaCl	Sodium acetate pH 5.0 + 300 mM NaCl	Potassium phosphate pH 5.0 + 300 mM NaCl	Sodium phosphate pH 5.5 + 300 mM NaCl
D	Sodium citrate pH 5.5 + 300 mM NaCl	MES pH 5.8 + 300 mM NaCl	Potassium phosphate pH 6.0 + 300 mM NaCl	MES pH 6.2 + 300 mM NaCl	Sodium phosphate pH 6.5 + 300 mM NaCl	Sodium cacodylate pH 6.5 + 300 mM NaCl	MES pH 6.5 + 300 mM NaCl	PIPES pH 6.7 + 300 mM NaCl	MOPS pH 7.0 + 300 mM NaCl	HEPES pH 7.0 + 300 mM NaCl	Ammonium acetate pH 7.3 + 300 mM NaCl	Sodium phosphate pH 7.5 + 300 mM NaCl
E	Tris pH 7.5 + 300 mM NaCl	Imidazole pH 8.0 + 300 mM NaCl	Hepes pH 8.0 + 300 mM NaCl	Tris pH 8.0 + 300 mM NaCl	Bicine pH 8.5 + 300 mM NaCl	Tris pH 8.5 + 300 mM NaCl	Bicine pH 9.0 + 300 mM NaCl	CAPS pH 9.0 + 300 mM NaCl	Glycine.HCl pH 9.5 + 300 mM NaCl	Glycine.HCl pH 10 + 300 mM NaCl	CAPS pH 10.5 + 300 mM NaCl	CAPS pH 11 + 300 mM NaCl
F	Citric acid pH 4.0 + 1 M NaCl	Sodium acetate pH 4.5 + 1 M NaCl	Sodium citrate pH 4.7 + 1 M NaCl	Sodium acetate pH 5.0 + 1 M NaCl	Potassium phosphate pH 5.0 + 1 M NaCl	Sodium phosphate pH 5.5 + 1 M NaCl	Sodium citrate pH 5.5 + 1 M NaCl	MES pH 5.8 + 1 M NaCl	Potassium phosphate pH 6.0 + 1 M NaCl	MES pH 6.2 + 1 M NaCl	Sodium phosphate pH 6.5 + 1 M NaCl	Sodium cacodylate pH 6.5 + 1 M NaCl
G	MES pH 6.5 + 1 M NaCl	PIPES pH 6.7 + 1 M NaCl	MOPS pH 7.0 + 1 M NaCl	HEPES pH 7.0 + 1 M NaCl	Ammonium acetate pH 7.3 + 1 M NaCl	Sodium phosphate pH 7.5 + 1 M NaCl	Tris pH 7.5 + 1 M NaCl	Imidazole pH 8.0 + 1 M NaCl	Hepes pH 8.0 + 1 M NaCl	Tris pH 8.0 + 1 M NaCl	Bicine pH 8.5 + 1 M NaCl	Tris pH 8.5 + 1 M NaCl
H	Bicine pH 9.0 + 1 M NaCl	CAPS pH 9.0 + 1 M NaCl	Glycine-HCl pH 9.5 + 1 M NaCl	Glycine-HCl pH 10 + 1 M NaCl	CAPS pH 10.5 + 1 M NaCl	CAPS pH 11 + 1 M NaCl	CONTROLO 1	CONTROLO 2	CONTROLO 3			

Supplementary Figure 1 - Thermofluor buffer screening (buffers prepared in house)

	1	2	3	4	5	6	7	8	9	10	11	12
A	water	100 mM Na Acetate	100 mM Ca Acetate	100 mM K Acetate	100 mM Na Sulfate	100 mM Na Sulfate	100 mM Mg Sulfate	100 mM K Sulfate	100 mM Ammonium Sulfate	100 mM Na Phosphate (monobasic)	100 mM Na Phosphate (dibasic)	100 mM Na Phosphate (monobasic)
B	100 mM K Phosphate (dibasic)	100 mM Na tartrate	100 mM Na Citrate (tribasic)	100 mM Na Malonate	100 mM Na nitrate	100 mM Na Formate	100 mM K Formate	100 mM NaF	100 mM KF	100 mM NH_4F	100 mM LiCl	100 mM NaCl
C	100 mM KCl	100 mM NH_4Cl	100 mM NaI	100 mM KI	100 mM NaBr	1 mM $MgCl_2$	1 mM $CaCl_2$	1 mM $MnCl_2$	1 mM $NiCl_2$	1 mM $FeCl_2$	1 mM $ZnCl_2$	1 mM $CoCl_2$
D	5mM EDTA	5 mM EGTA	0.1 M Urea	0.5 M Urea	1 M Urea	2 M Urea	4 M Urea	150 mM Guanidine-HCl	500 mM Guanidine-HCl	1 mM NDSB-195	1mM NDSB-201	1 mM Fos Choline 12
E	1 mM CHAPS	1 mM CHAPSO	1 mM OG	1 mM DM	1 mM DDM	25 mM Monosaccharides mix MD2-100-75	25 mM Glucose	25 mM Sucrose	25 mM Maltose	50 mM Carboxylic acids mix MD2-100-76	50 mM Proline	50 mM Glycine
F	25 mM Glutamic Acid	500 mM Glutamic Acid	50 mM Arginine	50 mM Arginine	50 mM Arginine 50 mM Glutamic Acid	500 mM Arginine 500 mM Glutamic Acid	5 mM Gly-Gly-Gly	5% (v/v) Oxaloacetic Acid	5% (v/v) DMSO	5% (v/v) Ethylene glycol	5% (v/v) Glycerol	20% (v/v) Glycerol
G	5% (v/v) PEG 400	5% (w/v) PEG 1000	5% (w/v) PEG 3350	5 mM DTT	5 mM TCEP	5 mM Biotin	5 mM Betaine	5 mM Coenzyme A	5 mM Nicotinic Acid	1 mM Spermidine	1 mM Spermine	1 mM Sarcosine
H	≈20 uM Desoxyribonucleic acid library <50bp	1 mM ATP 1 mM $MgCl_2$	1 mM ATPγS 1 mM $MgCl_2$	1 mM cAMP 1 mM $MgCl_2$	1 mM GTP 1 mM $MgCl_2$	1 mM GTPγS 1 mM $MgCl_2$	1 mM cGMP 1 mM $MgCl_2$	1 mM NADH 1 mM $MgCl_2$	1 mM NADPH 1 mM $MgCl_2$	5 mM Polyethylenimine	200 mM Imidazole	400 mM Imidazole

Supplementary Figure 2 - Additive thermofluor screen (Molecular Dimensions)

	1	2	3	4	5	6	7	8	9	10	11	12
A	JBS 1 1/A1-15% P 400 0.1M NaAct 4.6 0.1M $CaCl_2$	JBS 1 1/A2-15% P 400 0.1M MES 6.5	JBS 1 1/A3-15% P 400 0.1M HEPES 7.5 0.2M $MgCl_2$	JBS 1 1/A4-15% P 400 0.1M TRIS 8.5 0.2M $CitraNa_3$	JBS 1 1/A5-25% P 400 0.1M NaAct4.5 0.1M $MgCl_2$	JBS 1 1/A6-25% P 400 0.1M TRIS 8.5 0.2M Li_2SO_4	JBS 1 1/B1-28% P 400 0.1M HEPES 7.5 0.2M $CaCl_2$	JBS 1 1/B2-30% P 400 0.1M NaAct 4.6 0.1M $CaCl_2$	JBS 1 1/B3-30% P 400 0.1M MES 6.5 0.1M ActNa	JBS 1 1/B4-30% P 400 0.1M MES 6.5 0.1M $MgCl_2$	JBS 1 1/B5-30% P 400 0.1M HEPES 7.5 0.2M $MgCl_2$	JBS 1 1/B6-30% P 400 0.1M TRIS 8.5 0.2M $CitraNa_3$
B	JBS 1 1/C1-30% P 550 0.1M Bicina 9 0.1M NaCl	JBS 1 1/C2-25% P 550 0.1M MES 6.5 0.01M $ZnSO_4$	JBS 1 1/C3- 25% PEG 1K 0.1M HEPES 7.5	JBS 1 1/C4-30% PEG 1K 0.1M TRIS 8.5	JBS 1 1/C5-15% P 1.5K	JBS 1 1/C6-20% P 1.5K 0.1M HEPES 7.5	JBS 1 1/D1-30% P 1.5K	JBS 1 1/D2-20% P 2K 0.1M TRIS 8.5 0.01M $NiCl_2$	JBS 1 1/D3-25% P 2K	JBS 1 1/D4-30% P 2K 0.1M MES 6.5 0.1M ActNa	JBS 1 1/D5-20% P 3K 0.1M HEPES 7.5 0.2M ActNa	JBS 1 1/D6 -30% P 3K 0.1M TRIS 8.5 0.2M Li_2SO_4
C	JBS 2 2/A1-4% P 4K 0.1M NaAct 4.6	JBS 2 2/A2-8% P 4K	JBS 2 2/A3-8% PEG 4K 0.1M NaAct 4.6	JBS 2 2/A4-10% P 4K 0.1M MES 6.5 0.2M $MgCl_2$	JBS 2 2/A5-12% P 4K 0.1M HEPES 7.5 0.1M ActNa	JBS 2 2/A6-12% P 4K 0.1M TRIS 8.5	JBS 2 2/B1-16% P 4K 0.1M TRIS 8.5 0.2M Li_2SO_4	JBS 2 2/B2-16% P 4K 0.1M TRIS 8.5 0.2M ActNa	JBS 2 2/B3-16% P 4K 0.1M TRIS 8.5 0.1M $MgCl_2$	JBS 2 2/B4-18% P 4K 0.1M Act 4.6	JBS 2 2/B5-20% P 4K 0.1M TRIS 8.5 0.2M Li_2SO_4	JBS 2 2/B6-20% P 4K 0.1M TRIS 8.5 0.2M $CaCl_2$
D	JBS 2 2/C1-22% P 4K 0.1M HEPES 7.5 0.1M ActNa	JBS 2 2/C2-25% P 4K 0.1M NaAct 4.6	JBS 2 2/C3-25% P 4K 0.1M MES 6.5 0.2M $MgCl_2$	JBS 2 2/C4-25% P 4K 0.1M TRIS 8.5 0.2M $CaCl_2$	JBS 2 2/C5-30% P 4K	JBS 2 2/C6-30% P 4K 0.1M NaAct 4.5 0.1M $MgCl_2$	JBS 2 2/D1-30% P 4K 0.1M MES 6.5	JBS 2 2/D2-30% P 4K 0.1M HEPES 7.5 0.2M $CaCl_2$	JBS 2 2/D3-30% P 4K 0.1M TRIS 8.5 0.2M Li_2SO_4	JBS 2 2/D4-30% P 4K 0.1M TRIS 8.5 0.2M ActNa	JBS 2 2/D5-30% P 4K 0.1M TRIS 8.5 0.2M $MgCl_2$	JBS 2 2/D6 - 35% P 4K
E	JBS 3 3/A1-8% P 4K 0.1M TRIS 8.5 0.8M LiCl	JBS 3 3/A2-10% PEG 4K 20% Isopp	JBS 3 3/A3-10% P 4K 0.1M Cit 5.5 10% Isopp	JBS 3 3/A4-10% P 4K 0.1M HEPES 7.5 5% Isopp	JBS 3 3/A5-10% P 4K 0.1M HEPES 7.5 20% Isopp	JBS 3 3/A6-12% P 4K 0.1M Act 4.6 0.2M $ActNH_4$	JBS 3 3/B1-15% P 4K 0.2M $(NH_4)_2SO_4$	JBS 3 3/B2 - 15% P 4K 0.1M Cit 5.6 0.2M $ActNH_4$	JBS 3 3/B3-16% P 4K 0.1M HEPES 7.5 10% Isopp 0.2M $(NH_4)_2SO_4$	JBS 3 3/B4-20% P 4K 0.2M $(NH_4)_2SO_4$	JBS 3 3/B5-20% P 4K 10% Glicerol 0.2M $MgSO_4$	JBS 3 3/B6-20% P 4K 5% Isopp 0.1M $CitraNa_3$
F	JBS 3 3/C1-20% P 4K 20% Isopp 0.1M $CitraNa_3$	JBS 3 3/C2-20% P 4K 0.1M Mes 6.5 600mM NaCl	JBS 3 3/C3-20% P 4K 0.1M HEPES 7.5 10% Isopp	JBS 3 3/C4-22% P 4K 0.1M ActNa 0.2M $(NH_4)_2SO_4$	JBS 3 3/C5-25% P 4K 0.1M ActNa 4.6 0.2M $(NH_4)_2SO_4$	JBS 3 3/C6-25% P 4K 0.1M Cit 5.6 0.2M $ActNH_4$	JBS 3 3/D1-25% P 4K 0.1M HEPES 7.5 0.2M Li_2SO_4 0.1M ActNa	JBS 3 3/D2-25% P 4K 8% Isopp 0.1M ActNa	JBS 3 3/D3-30% P 4K 0.2M $(NH_4)_2SO_4$	JBS 3 3/D4-30% P 4K 0.1M Act 4.6 0.2M $(NH_4)_2SO_4$	JBS 3 3/D5 - 30% P 4K 0.1M $CitNa_3$ 5.6 0.1M $ActNH_4$	JBS 3 3/D6-32% P 4K 0.1M TRIS 8.5 0.8M LiCl
G	JBS 4 4/A1-25% P 5K 0.1M TRIS 8.5 0.2M Li_2SO_4	JBS 4 4/A2 -30% P 5K 0.1M MES 6.5 0.2M $(NH_4)_2SO_4$	JBS 4 4/A3-3% P 6K 0.1M TRIS 8.5 0.1M KCl	JBS 4 4/A4 -10% P 6K 0.01M $MgCl_2$	JBS 4 4/A5 - 12% P 6K 2M NaCl	JBS 4 4/A6 - 15% P 6K 5% Glicerol	JBS 4 4/B1-15% P 6K 0.05M KCl 0.01M $MgCl_2$	JBS 4 4/B2-16% P 6K 0.01M $CitraNa_3$	JBS 4 4/B3 -20% P 6K 0.05M Imid 8	JBS 4 4/B4-25% P 6K 0.1M HEPES 7.5 0.1M LiCl	JBS 4 4/B5 -28% P 6K 0.1M TRIS 8.5 0.5M LiCl	Jbs 4 4/B6 -30% P 6K 1M LiCl 0.1M ActNa
H	JBS 4 4/C1 - 33% P 6K 0.01M $CitraNa_3$	JBS 4 4/C2 -2% P 8K 0.5M Li_2SO_4	JBS 4 4/C3 - 2% P 8K 1M Li_2SO_4	JBS 4 4/C4 -4% P 8K	JBS 4 4/C5 -8% P 8K 0.2M LiCl 0.05M $MgSO_4$	JBS 4 4/C6 -8% P 8K 0.1M TRIS 8.5	JBS 4 4/D1-10% P 8K 0.1M MES 6.5 0.2M ActZn	JBS 4 4/D2-10% P 8K 0.1M HEPES 7.5 0.2M ActNa	JBS 4 4/D3-10% P 8K 0.05M ActMg 0.1M ActNa	JBS 4 4/D4-10% PEG 8K 0.2M ActMg	JBS 4 4/D5-10% P 8K 0.1M HEPES 7.5 10% Etlglicol	JBS 4 4/D6-10% P 8K 10% P 1K

Supplementary Figure 3 - JBScreen classic 1,2,3,4 crystallization screening (Jena Bioscience)

A1 Morpheus 1-0.06M Div 0.1M Buf Sist 1 50%Prec. Mix 1	A2 Morpheus 2-0.06M Div 0.1M Buf Sist 1 50%Prec. Mix 2	A3 Morpheus 3-0.06M Div 0.1M Buf Sist 1 50%Prec. Mix 3	A4 Morpheus 4-0.06M Div 0.1M Buf Sist 1 50%Prec. Mix 4	A5 Morpheus 5-0.06M Div 0.1M Buf Sist 2 50%Prec. Mix 1	A6 Morpheus 6-0.06M Div 0.1M Buf Sist 2 50%Prec. Mix 2	A7 Morpheus 7-0.06M Div 0.1M Buf Sist 2 50%Prec. Mix 3	A8 Morpheus 8-0.06M Div 0.1M Buf Sist 2 50%Prec. Mix 4	A9 Morpheus 9-0.06M Div 0.1M Buf Sist 3 50%Prec. Mix 1	A10 Morpheus 10-0.06M Div 0.1M Buf Sist 3 50%Prec. Mix 2	A11 Morpheus 11-0.06M Div 0.1M Buf Sist 3 50%Prec. Mix 3	A12 Morpheus 12-0.06M Div 0.1M Buf Sist 3 50%Prec. Mix 4
B1 Morpheus 13-0.09 M Hal 0.1M Buf Sist 1 50%Prec. Mix 1	B2 Morpheus 14-0.09 M Hal 0.1M Buf Sist 1 50%Prec. Mix 2	B3 Morpheus 15-0.09 M Hal 0.1M Buf Sist 1 50%Prec. Mix 3	B4 Morpheus 16-0.09 M Hal 0.1M Buf Sist 1 50%Prec. Mix 4	B5 Morpheus 17-0.09 M Hal 0.1M Buf Sist 2 50%Prec. Mix 1	B6 Morpheus 18-0.09 M Hal 0.1M Buf Sist 2 50%Prec. Mix 2	B7 Morpheus 19-0.09 M Hal 0.1M Buf Sist 2 50%Prec. Mix 3	B8 Morpheus 20-0.09 M Hal 0.1M Buf Sist 2 50%Prec. Mix 4	B9 Morpheus 21-0.09 M Hal 0.1M Buf Sist 3 50%Prec. Mix 1	B10 Morpheus 22-0.09 M Hal 0.1M Buf Sist 3 50%Prec. Mix 2	B11 Morpheus 23-0.09 M Hal 0.1M Buf Sist 3 50%Prec. Mix 3	B12 Morpheus 24-0.09 M Hal 0.1M Buf Sist 3 50%Prec. Mix 4
C1 Morpheus 25-0.09 NPS 0.1M Buf Sist 1 50%Prec. Mix 1	C2 Morpheus 26-0.09 M NPSl 0.1M Buf Sist 1 50%Prec. Mix 2	C3 Morpheus 27-0.09 M NPS 0.1M Buf Sist 1 50%Prec. Mix 3	C4 Morpheus 28-0.09 M NPS 0.1M Buf Sist 1 50%Prec. Mix 4	C5 Morpheus 29-0.09 M NPS 0.1M Buf Sist 2 50%Prec. Mix 1	C6 Morpheus 30-0.09 M NPS 0.1M Buf Sist 2 50%Prec. Mix 2	C7 Morpheus 31-0.09 M NPS 0.1M Buf Sist 2 50%Prec. Mix 3	C8 Morpheus 32-0.09 M NPS 0.1M Buf Sist 2 50%Prec. Mix 4	C9 Morpheus 33-0.09 M NPS 0.1M Buf Sist 3 50%Prec. Mix 1	C10 Morpheus 34-0.09 M NPS 0.1M Buf Sist 3 50%Prec. Mix 2	C11 Morpheus 35-0.09 M NPS 0.1M Buf Sist 3 50%Prec. Mix 3	C12 Morpheus 36-0.09 M NPS 0.1M Buf Sist 3 50%Prec. Mix 4
D1 Morpheus 37-0.12M Alco 0.1M Buf Sist 1 50%Prec. Mix 1	D2 Morpheus 38-0.12M Alco 0.1M Buf Sist 1 50%Prec. Mix 2	D3 Morpheus 39-0.12M Alco 0.1M Buf Sist 1 50%Prec. Mix 3	D4 Morpheus 40-0.12M Alco 0.1M Buf Sist 1 50%Prec. Mix 4	D5 Morpheus 41-0.12M Alco 0.1M Buf Sist 2 50%Prec. Mix 1	D6 Morpheus 42-0.12M Alco 0.1M Buf Sist 2 50%Prec. Mix 2	D7 Morpheus 43-0.12M Alco 0.1M Buf Sist 2 50%Prec. Mix 3	D8 Morpheus 44-0.12M Alco 0.1M Buf Sist 2 50%Prec. Mix 4	D9 Morpheus 45-0.12M Alco 0.1M Buf Sist 3 50%Prec. Mix 1	D10 Morpheus 46-0.12M Alco 0.1M Buf Sist 3 50%Prec. Mix 2	D11 Morpheus 47-0.12M Alco 0.1M Buf Sist 3 50%Prec. Mix 3	D12 Morpheus 48-0.12M Alco 0.1M Buf Sist 3 50%Prec. Mix 4
E1 Morpheus 49-0.12M EtGly 0.1M Buf Sist 1 50%Prec. Mix 1	E2 Morpheus 50-0.12M EtGly 0.1M Buf Sist 1 50%Prec. Mix 2	E3 Morpheus 51-0.12M EtGly 0.1M Buf Sist 1 50%Prec. Mix 3	E4 Morpheus 52-0.12M EtGly 0.1M Buf Sist 1 50%Prec. Mix 4	E5 Morpheus 53-0.12M EtGly 0.1M Buf Sist 2 50%Prec. Mix 1	E6 Morpheus 54-0.12M EtGly 0.1M Buf Sist 2 50%Prec. Mix 2	E7 Morpheus 55-0.12M EtGly 0.1M Buf Sist 2 50%Prec. Mix 3	E8 Morpheus 56-0.12M EtGly 0.1M Buf Sist 2 50%Prec. Mix 4	E9 Morpheus 57-0.12M EtGly 0.1M Buf Sist 3 50%Prec. Mix 1	E10 Morpheus 58-0.12M EtGly 0.1M Buf Sist 3 50%Prec. Mix 2	E11 Morpheus 59-0.12M EtGly 0.1M Buf Sist 3 50%Prec. Mix 3	E12 Morpheus 60-0.12M EtGly 0.1M Buf Sist 3 50%Prec. Mix 4
F1 Morpheus 61-0.12M MSac 0.1M Buf Sist 1 50%Prec. Mix 1	F2 Morpheus 62-0.12M MSac 0.1M Buf Sist 1 50%Prec. Mix 2	F3 Morpheus 63-0.12M MSac 0.1M Buf Sist 1 50%Prec. Mix 3	F4 Morpheus 64-0.12M MSac 0.1M Buf Sist 1 50%Prec. Mix 4	F5 Morpheus 65-0.12M MSac 0.1M Buf Sist 2 50%Prec. Mix 1	F6 Morpheus 66-0.12M MSac 0.1M Buf Sist 2 50%Prec. Mix 2	F7 Morpheus 67-0.12M MSac 0.1M Buf Sist 2 50%Prec. Mix 3	F8 Morpheus 68-0.12M MSac 0.1M Buf Sist 2 50%Prec. Mix 4	F9 Morpheus 69-0.12M MSac 0.1M Buf Sist 3 50%Prec. Mix 1	F10 Morpheus 70-0.12M MSac 0.1M Buf Sist 3 50%Prec. Mix 2	F11 Morpheus 71-0.12M MSac 0.1M Buf Sist 3 50%Prec. Mix 3	F12 Morpheus 72-0.12M MSac 0.1M Buf Sist 3 50%Prec. Mix 4
G1 Morpheus 73-0.1M CarbAc 0.1M Buf Sist 1 50%Prec. Mix 1	G2 Morpheus 74-0.1M CarbAc 0.1M Buf Sist 1 50%Prec. Mix 2	G3 Morpheus 75-0.1MCarbAc 0.1M Buf Sist 1 50%Prec. Mix 3	G4 Morpheus 76-0.1MCarbAc 0.1M Buf Sist 1 50%Prec. Mix 4	G5 Morpheus 77-0.1MCarbAc 0.1M Buf Sist 2 50%Prec. Mix 1	G6 Morpheus 78-0.1MCarbAc 0.1M Buf Sist 2 50%Prec. Mix 2	G7 Morpheus 79-0.1MCarbAc 0.1M Buf Sist 2 50%Prec. Mix 3	G8 Morpheus 80-0.1MCarbAc 0.1M Buf Sist 2 50%Prec. Mix 4	G9 Morpheus 81-0.1MCarbAc 0.1M Buf Sist 3 50%Prec. Mix 1	G10 Morpheus 82-0.1MCarbAc 0.1M Buf Sist 3 50%Prec. Mix 2	G11 Morpheus 83-0.1MCarbAc 0.1M Buf Sist 3 50%Prec. Mix 3	G12 Morpheus 84-0.1MCarbAc 0.1M Buf Sist 3 50%Prec. Mix 4
H1 Morpheus 85-0.1M AA 0.1M Buf Sist 1 50%Prec. Mix 1	H2 Morpheus 86-0.1M AA 0.1M Buf Sist 1 50%Prec. Mix 2	H3 Morpheus 87-0.1M AA 0.1M Buf Sist 1 50%Prec. Mix 3	H4 Morpheus 88-0.1M AA 0.1M Buf Sist 1 50%Prec. Mix 4	H5 Morpheus 89-0.1M AA 0.1M Buf Sist 2 50%Prec. Mix 1	H6 Morpheus 90-0.1M AA 0.1M Buf Sist 2 50%Prec. Mix 2	H7 Morpheus 91-0.1M AA 0.1M Buf Sist 2 50%Prec. Mix 3	H8 Morpheus 92-0.1M AA 0.1M Buf Sist 2 50%Prec. Mix 4	H9 Morpheus 93-0.1M AA 0.1M Buf Sist 3 50%Prec. Mix 1	H10 Morpheus 94-0.1M AA 0.1M Buf Sist 3 50%Prec. Mix 2	H11 Morpheus 95-0.1M AA 0.1M Buf Sist 3 50%Prec. Mix 3	H12 Morpheus 96-0.1M AA 0.1M Buf Sist 3 50%Prec. Mix 4

Supplementary Figure 4 - Morpheus crystallization screening (Molecular Dimensions)

	1	2	3	4	5	6	7	8	9	10	11	12
A	JBS 5 96-12% P 8K 5% Glicerol 0.1M KCl	JBS 5 97-12% P 8K 10% Glicerol 0.1M KCl	JBS 5 98-15% P 8K 0.2M $(NH_4)_2SO_4$	JBS 5 99-15% P 8K 0.5M Li_2SO_4	JBS 5 100-15% P 8K 0.1M MES 6.5 0.2M ActNa	JBS 5 101-15% P 8K 0.05M $(NH_4)_2SO_4$ 0.1M $CitraNa_3$	JBS 5 102-18% P 8K 0.1M HEPES 7.5 0.2M ActCa	JBS 5 103-18% P 8K 0.1M HEPES 7.5 2% Isopp 0.1M ActNa	JBS 5 104-18% P 8K 0.1M TRIS 8.5 0.2M Li_2SO_4	JBS 5 105-20% P 8K 0.1M HEPES 7.5	JBS 5 106-20% P 8K 0.1M MES 6.5 0.2M ActMg	JBS 5 107-20% P 8K 0.1M TRIS 9.5
B	JBS 5 108-22% P 8K 0.1M MES 6.5 0.2M $(NH_4)_2SO_4$	JBS 5 109-25% P 8K 0.2M LiCl	JBS 5 110-30% P 8K 0.2M $(NH_4)_2SO_4$	JBS 5 111-8% P 10K 0.1M Act 4.5	JBS 5 112-14% P 10K 0.1M Imid 8	JBS 5 113-16% P 10K 0.1M TRIS 8.5	JBS 5 114-18% P 10K 0.1M TRIS 8.5 20% Glicerol 0.1M NaCl	JBS 5 115-20% P 10K 0.1M HEPES 7.5	JBS 5 116-30% P 10K 0.1M TRIS 8.5	JBS 5 117-10% P 20K 0.1M MES 6.5	JBS 5 118-17% P 20K 0.1M TRIS 8.5 0.1M $MgCl_2$	JBS 5 119-20% P 20K
C	JBS 6 120- 0.5M $(NH_4)_2SO_4$ 1M Li_2SO_4 0.1M $CitraNa_3$	JBS 6 121- 1M $(NH_4)_2SO_4$	JBS 6 122- 1M $(NH_4)_2SO_4$ 0.1M Act 4.5	JBS 6 123- 1M $(NH_4)_2SO_4$ 0.1M HEPES 7.5 2% P 400	JBS 6 124- 1M $(NH_4)_2SO_4$ 0.1M TRIS 8.5	JBS 6 125- 1.2M $(NH_4)_2SO_4$ 3% Isopp 0.05M $CitraNa_3$	JBS 6 126- 1.5M $(NH_4)_2SO_4$ 0.1M TRIS 8.5 15% Glicerol	JBS 6 127- 1.6M $(NH_4)_2SO_4$ 0.1M LiCl	JBS 6 128- 1.6M $(NH_4)_2SO_4$ 1M Li_2SO_4	JBS 6 129- 1.6M $(NH_4)_2SO_4$ 0.1M HEPES 7.5 0.2M NaCl	JBS 6 130- 1.6M $(NH_4)_2SO_4$ 0.1M HEPES 7.5 2% P 1K	JBS 6 131- 1.8M $(NH_4)_2SO_4$ 0.1M MES 6.5
D	JBS 6 132- 2M $(NH_4)_2SO_4$ 2M NaCl	JBS 6 133- 2M $(NH_4)_2SO_4$ 0.1M Act 4.5	JBS 6 134- 2M $(NH_4)_2SO_4$ 0.1M MES 6.5 5% P 400	JBS 6 135- 2M $(NH_4)_2SO_4$	JBS 6 136- 2.2M $(NH_4)_2SO_4$	JBS 6 137- 2.2M $(NH_4)_2SO_4$ 20% Glicerol	JBS 6 138- 2.4M $(NH_4)_2SO_4$ 0.1M $CitraNa_3$	JBS 6 139- 3M $(NH_4)_2SO_4$ 1% MPD	JBS 6 140- 3M $(NH_4)_2SO_4$ 10% Glicerol	JBS 6 141- 3.5M $(NH_4)_2SO_4$ 0.1M HEPES 7.5	JBS 6 142- 3.5M $(NH_4)_2SO_4$ 0.1M MES 6.5 1% MPD	JBS 6 143- 3.5M $(NH_4)_2SO_4$
E	JBS 7 144-10% MPD 0.1M HEPES 7.5 0.1M $CitraNa_3$	JBS 7 145-12% MPD 0.1M TRIS 8.5 0.05M $MgCl_2$	JBS 7 146-15% MPD 0.1M Act 4.5 0.02M $CaCl_2$	JBS 7 147-15% MPD 0.1M Imid 8 5% P 4K	JBS 7 148-15% MPD 0.1M Cit 5.5 0.2M $ActNH_4$	JBS 7 149-15% MPD 0.1M MES 6.5 0.2M $MgCl_2$	JBS 7 150-15% MPD 0.1M HEPES 7.5 0.2M $CitraNa_3$	JBS 7 151-20% MPD 0.1M HEPES 7.5 0.1M $CitraNa_3$	JBS 7 152-20% MPD 0.1M Imid 8	JBS 7 153-20% MPD 4% Glicerol 0.2M NaCl	JBS 7 154-30% MPD 0.1M Act4.5 0.02M $CaCl_2$	JBS 7 155-30% MPD 0.1M Cit 5.5 0.2M $ActNH_4$
F	JBS 7 156-30% MPD 0.1M MES 6.5 0.2M ActMg	JBS 7 157-30% MPD 0.1M HEPES 7.5 0.5M $(NH_4)_2SO_4$	JBS 7 158-30% MPD 0.1M HEPES 7.5 0.2M $CitraNa_3$	JBS 7 159-30% MPD 0.1M HEPES 7.5 5% P 4K	JBS 7 160-30% MPD 0.1M Imid 8 10% P 4K	JBS 7 161-30% MPD 20% Etanol	JBS 7 162-35% MPD	JBS 7 163-35% MPD 0.1M Imid 8	JBS 7 164-40% MPD 0.1M TRIS 8.5	JBS 7 165-47% MPD 0.1M HEPES 7.5	JBS 7 166-47%MPD 3% P400	JBS 7 167-50% MPD
G	JBS 8 168-50% MPD 15% Etanol 0.01M ActNa	JBS 8 169-50% MPD 2% Isopp 0.05M ActNa 0.05M NaCl	JBS 8 170-50% MPD 0.1M TRIS 8.5 0.2M $NH_4H_2PO_4$	JBS 8 171-55% MPD	JBS 8 172-60% MPD 0.1M Act 4.5 0.01M $CaCl_2$	JBS 8 173-60% MPD 0.02M ActNa	JBS 8 174-70% MPD 0.1M MES 6.5	JBS 8 175-70% MPD 0.1M TRIS 8.5	JBS 8 176-70% MPD 0.1M TRIS 8.5 0.01M $CaCl_2$	JBS 8 177-2% Etanol 0.1M TRIS 8.5	JBS 8 178-5% Etanol 0.1M HEPES 7.5 5% MPD	JBS 8 179-5% Etanol 0.1M TRIS 8.5 5% MPD 0.2M NaCl
H	JBS 8 180-10% Etanol 0.1M TRIS 8.5	JBS 8 181-12% Etanol 0.1M Act 4.5 4% P 400	JBS 8 182-14% Etanol 0.1M TRIS 8.5	JBS 8 183-18% Etanol 0.1M TRIS 8.5	JBS 8 184-20% Etanol	JBS 8 185-20% Etanol 10% Glicerol	JBS 8 186-30% Etanol 10% PEG 6K 0.1M ActNa	JBS 8 187-45% Etanol	JBS 8 188-50% Etanol 0.01M ActNa	JBS 8 189-60% Etanol 1.5% PEG 6K 0.05M ActNa	JBS 8 190-60% Etanol 0.1M NaCl	JBS 8 191-2% Isopp 0.1M TRIS 8.5 0.01M $MgSO_4$

Supplementary Figure 5 - JBScreen classic 5,6,7,8 crystallization screening (Jena Bioscience)

A1 1-1 50% PEG 400 0.1M NaAct 4.5 0.2M Li_2SO_4	**A2 1-2** 20% PEG 3000 0.1M NaCit 5.5	**A3 1-3** 20% PEG 3350 0.2M Cit(NH_4)$_2$	**A4 1-4** 30% MPD 0.1M NaAct 4.6 0.02M $CaCl_2.2H_2O$	**A5 1-5** 20% P 3350 0.2M FormMg	**A6 1-6** 20% PEG 1000 0.1M FosCit 4.2 0.2M Li_2SO_4	**A7 1-7** 20% PEG 8K 0.1M CHES 9.5	**A8 1-8** 20% P 3350 0.2 FormNH_4	**A9 1-9** 20% P 3350 0.2M NH_4Cl	**A10 1-10** 20% P 3350 0.2M FormK	**A11 1-11** 50 % MPD 0.1M TRIS 8.5 0.2M AmmFosf monobasic	**A12 1-12** 20% PEG 3350 0.2M KNO_3
B1 1-13 0.1M Cit 4 0.8M Amm.Sulf	**B2 1-14** 20% P 3350 0.2M NaSCN	**B3 1-15** 20% PEG 6K 0.1M Bicine 9	**B4 1-16** 10% PEG 8K 8% Etilenoglicol 0.1M HEPES 7.5	**B5 1-17** 40% MPD 0.1M NaCaco 6.5 5% PEG 8K	**B6 1-18** 40% Etanol 0.1M FosCit 4.2 5% PEG 1K	**B7 1-19** 8% P 4K 0.1M NaAct 4.6	**B8 1-20** 10% P 8K 0.1M TRIS 7 0.2M $MgCl_2.6H_2O$	**B9 1-21** 20% P 6K 0.1M Cit 5	**B10 1-22** 50% P 200 0.1M NaCaco 6.5 0.2M $MgCl_2.6H_2O$	**B11 1-23** 1.6M $CitNa_3.2H_2O$	**B12 1-24** 20% P 3350 0.2M $CitK_3.H_2O$
C1 1-25 20% PEG 8K 0.1M FosCit 4.2 0.2M NaCl	**C2 1-26** 20% PEG 6K 0.1M Cit 4 1M LiCl	**C3 1-27** 20% PEG 3350 0.2M NH_4NO_3	**C4 1-28** 10% PEG 6K 0.1M HEPES 7.0	**C5 1-29** 0.8M $NaH_2PO_4.H_2O$ 0.8M KH_2PO_4 0.1M NaHEPES 7.5	**C6 1-30** 40% PEG 300 0.1M Fosf/Cit 4.2	**C7 1-31** 10% PEG 3K 0.1M NaAct 4.5 0.2M $ActZn.2H_2O$	**C8 1-32** 20% Etanol 0.1M TRIS 8.5	**C9 1-33** 10% Glicerol 25%Propanediol 0.1M Na/K Fosf 6.2	**C10 1-34** 10% PEG 20K 2%Dioxano 0.1M Bicina 9	**C11 1-35** 2M S.A. 0.1M NaAct 4.6	**C12 1-36** 10% PEG 1K 10% PEG 8K
D1 1-37 24% PEG 1.5K 20% Glicerol	**D2 1-38** 30% PEG 400 0.1M NaHEPES 7.5 0.2M $MgCl_2.6H_2O$	**D3 1-39** 50% PEG 200 0.1M Na/K Fosf 6.2 0.2M NaCl	**D4 1-40** 30% PEG 8K 0.1M NaAct 4.5 0.2M Li_2SO_4	**D5 1-41** 70% MPD 0.1M HEPES 7.5	**D6 1-42** 20% P 8K 0.1M TRIS 8.5 0.2M $MgCl_2.6H_2O$	**D7 1-43** 40% PEG 400 0.1M TRIS 8.5 0.2M Li_2SO_4	**D8 1-44** 40% MPD 0.1M TRIS 8.5	**D9 1-45** 25.5% PEG 4K 15% Glicerol 0.17M S.A.	**D10 1-46** 40% PEG 300 0.1M NaCaco 6.5 0.2M CaAct	**D11 1-47** 14% Isopp 0.07M Act 4.5 30% Glicerol 0.14M $CaCl_2.2H_2O$	**D12 1-48** 16% PEG 8K 0.04M KH_2PO_4 20% Glicerol
E1 2-1 1M $CitNa_3.2H_2O$ 0.1M NaCaco 6.5	**E2 2-2** 2M AmmSulf 0.1M NaCaco 6.5 0.2M NaCl	**E3 2-3** 10% Isopp 0.1M HEPES 7.5 0.2M NaCl	**E4 2-4** 1.26M AmmSulf 0.1M TRIS 8.5 0.2M Li_2SO_4	**E5 2-5** 40% MPD 0.1M CAPS 10.5	**E6 2-6** 20% PEG 3K 0.1M Imd 8 0.2M $ActZn.2H_2O$	**E7 2-7** 10% Isopp 0.1M NaCaco 6.5 0.2M $ActZn.2H_2O$	**E8 2-8** 1M $(NH_4)_2HPO_4$ 0.1M Act 4.5	**E9 2-9** 1.6M $MgSO_4.7H_2O$ 0.1M MES 6.5	**E10 2-10** 10% PEG 6K 0.1M Bicina 9	**E11 2-11** 14.4% PEG 8K 20% Glicerol 0.08M NaCaco 6.5 0.16 $CaAct.H_2O$	**E12 2-12** 10% PEG 8K 0.1M Imd 8
F1 2-13 30% Jeff M600 0.05M CsCl 0.1M MES 6.5	**F2 2-14** 3.2M AmmSulf 0.1M Cit 5	**F3 2-15** 20% MPD 0.1M TRIS 8	**F4 2-16** 20% Jeff M-600 0.1M HEPES 7.5	**F5 2-17** 50% Etilenoglicol 0.2M $MgCl.6H_2O$ 0.1M TRIS 8.5	**F6 2-18** 10% MPD 0.1M Bicina 9.0	**F7 2-19** 0.8M Succinic 7.0	**F8 2-20** 2.1M DL-Malic Acd 7.0	**F9 2-21** 2.4M NaMalonato dibasic.H_2O 7.0	**F10 2-22** 0.5% Jeff ED-2003 0.1M HEPES 7.0 1.1M NaMalonato dibasic.H_2O 7.0	**F11 2-23** 1% PEG-2000 MME 0.1M HEPES 7.0 1M Succinic Ac	**F12 2-24** 30% Jeff M-600 0.1M HEPES 7.0
G1 2-25 30% Jeff ED-2003 0.1M HEPES 7.0	**G2 2-26** 22% NaPolyacl 5100 0.1M HEPES 7.5 0.2M $MgCl_2.6H_2O$	**G3 2-27** 20% Polyvinilpirro 0.1M TRIS 85 0.1M $Co(II)Cl.6H_2O$	**G4 2-28** 20% P2K MME 0.1M TRIS 8.5 0.2M TMAO (Trimethylamine N-oxide)	**G5 2-29** 12% P 3350 0.1M HEPES 7.5 5mM $CoCl.6H_2O$ 5mM CdCl; 5mM $MgCl_2.6H_2O$ 5mM $NiCl.6H_2O$	**G6 2-30** 20% PEG 3350 0.2M NaMalonato dibasic.H_2O	**G7 2-31** 15% PEG 3350 0.1M Succinic ac.	**G8 2-32** 20% PEG 3350 0.15 DL- Malic Acid	**G9 2-33** 30% PEG 2K MME 0.1M KSCN	**G10 2-34** 30% PEG 2000 MME 0.15M KBr	**G11 2-35** 2M Amm.Sulf 0.1M Bis-TRIS 5.5	**G12 2-36** 3M NaCl 0.1M Bis-TRIS 5.5
H1 2-37 0.3M $FormMg.2H_2O$ 0.1M Bis-TRIS 5.5	**H2 2-38** 1% PEG 3350 0.1M Bis-TRIS 5.5 1M Amm.Suf	**H3 2-39** 25% PEG 3350 0.1M Bis-TRIS 5.5	**H4 2-40** 45% MPD 0.1M Bis-TRIS 5.5 0.2M $CaCl_2.2H_2O$	**H5 2-41** 45% MPD 0.1M Bis-TRIS 5.5 0.2M $ActNH_4$	**H6 2-42** 17% PEG 10000 0.1M BisTRIS 5.5 0.1M $ActNH_4$	**H7 2-43** 25% PEG 3350 0.1M BisTRIS 5.5 0.2M Amm.Sulf	**H8 2-44** 25% PEG 3350 0.1M BisTRIS 5.5 0.2M NaCl	**H9 2-45** 25% PEG 3350 0.1M BisTRIS 5.5 0.2M Li_2SO_4	**H10 2-46** 25% PEG 3350 0.1M BisTRIS 5.5 0.2M $ActNH_4$	**H11 2-47** 25% PEG 3350 0.1M BisTRIS 5.5 0.2M $MgCl_2.6H_2O$	**H12 2-48** 45% MPD 0.1M HEPES 7.5 0.2M $ActNH_4$

Supplementary Figure 6 –JCSG-Plus crystallization screening (Molecular Dimensions)

Supplementary Table 3 – Pre crystallization test results and recommended action

Pct Reagent A1/B1 Results	PCT Reagent A2/B2 Results	Recommended Action
Heavy Amorphous Precipitate	Heavy Amorphous Precipitate	Dilute sample 1:1, repeat test
Clear	Clear	Concentrate sample to half the original volume, repeat test
Light granular precipitate	Clear	Perform Screen
Clear	Light granular precipitate	Perform Screen
Heavy Amorphous Precipitate	Light granular precipitate	Perform Screen
Heavy Amorphous Precipitate	Clear	Perform PCT with B1 & B2
Clear	Heavy Amorphous Precipitate	Perform PCT with B1 & B2