Investigation of bacterial transport for improved uptake of hemicellulose-derived sugars and oligosaccharides

Konstantinos Drousiotis

PhD

University of York

Biology

September 2017

To loving memory of my grandparents Eleni and Kyriakos

Abstract

Plant lignocellulosic biomass can be a source of fermentable sugars for the production of second generation biofuels and provide a viable alternative to fossil fuels. The hemicellulotic component of many bioenergy grasses, including *Miscanthus*, is comprised of arabinoglucuronoxylans. As xylans constitute the third most abundant biopolymer, we set out to investigate the uptake of its sugars and oligosaccharides, *ie*. L-arabinofuranose and xylobiose, by bacteria. Using a phylogenetic, phylogenomic and '*in silico*' analyses of Galacto/Arabinofuranose SBP from *E. coli* (ie. GafAEc) we detected bacterial orthologs which could show 'bias' towards arabinose binding. The chosen SBPs for further characterisation were GafASw and GafASm from *Shewanella* sp. ANA-3 and *Sinorhizobium meliloti* 1021, respectively. Biochemical characterisation showed that GafASm bound L-arabinose 5 times higher affinity than GafAEc, and D-fucose changed the thermal stability of the protein at similar levels to D-galactose. The enhanced affinity for both L-arabinose and D-fucose might suggest that GafSm system is involved in chemotaxis towards root mucilage. Strikingly, GafASw is unable to bind D-galactose which is an uncommon feature of arabinose-binding bacterial proteins. We resolved the GafASw and discovered that tilting of its Asp88 residue, prohibits the formation of a salt bridge found in GafAEc structure with D-galactofuranose bound. This is attributed to the presence of a Phe residue at position 17 instead of a Gln. Furthermore, the presence of a Phe residue instead of an Asn near the binding cavity, displaces a water molecule which would otherwise interact with the O_6 in D-galactofuranose, as exemplified in GafAEc. Lastly, we created Transport Deletion (TD) mutants of *E. coli* which are unable to grow on L-arabinose (*ie*. TDara), D-xylose (*ie*. TDxyl) and D-glucuronic acid (*ie*. TDglcA). The strains recovered their growth when their respective secondary (MFS) were ectopically expressed, thus verifying their applicability as tools for detection of useful bacterial transporters. The usability of these strains in the discovery of endogenous and exogenous transporters for biomass-derived sugars was demonstrated by the use of TDxyl to show that an endogenous MFS transporter, *ie*. YagG, can import xylobiose.

Contents

List of Tables

List of Figures

Acknowledgements

I would foremost like to thank my research supervisor, Dr. Gavin Thomas, for providing me with the opportunity to undertake this research project. His endless support, generous guidance and undisputable knowledge has proved valuable throughout the completion of this thesis.

Secondly, I want to thank former and present members of the Thomas group. Many thanks to Judith, for assisting with protein overproduction during the late stages of the project and Reyme, for leading the crystallisation and structural study. Special recognition to Andrew Leech for helping me with the experiments in the molecular interactions lab and for the guidance provided.

Next, I extend my thanks to my close family and friends who have been with me every step of the way along the journey. Special, heartfelt thanks to my mother, Anna, and my godmother, Demetra; both of you have provided me with the love, support and strength I needed throughout this.

To my close friend and colleague, Stavroula, for sharing this incredible journey together and for the much-needed coffee breaks. My close friends, Tom, Laura and Namrata, for the humour and endless laughs we share.

Author's Declaration

I declare that this thesis is a presentation of original work and I am the sole author. This work has not previously been presented for an award at this, or any other, University. All sources are acknowledged as References. Work involving the crystallisation and structural study of GafASw was performed by Reyme Herman in collaboration with Prof. Anthony Wilkinson. Additionally, the SEC-MALLS assays were conducted by Dr. Andrew Leech at the University of York in the Molecular interactions lab of the Technology facility. Contributions from any others are mentioned within the main body of the thesis. All results presented here are prepared for publications.

Chapter 1

General Introduction

The following sections of the introductory chapter will provide the reader with a review of literature, pertinent to the research detailed in subsequent chapters of the text. The General Introduction starts with a brief description of the terminology which describes the sugars in the Haworth projection, as those will be used all throughout the thesis to refer to xylan-derived sugars. Thereafter, the importance and the structure of the xylan, with particular emphasis to bioenergy grasses is explained. The introduction proceeds to give details about the main stages involved in the consolidated bioprocessing of biomass by bacteria and re-orient focus to a neglected stage *ie*. transport. The bacterial secondary and ABC primary transporters will be described in greater detail because of the direct relevance of the following chapters to these kinds of systems. Finally, the importance in engineering of transport systems for biofuel production is demonstrated by examples from yeast transporters which have been more extensively studied compared to bacteria. The chapter ends with the collective aims of the study and a brief thesis outline.

1. 1 Composition and conformation of sugars in solution

Sugars are complex species which can take multiple forms in solution, exhibiting a great constitutional, configurational and conformational isomerism. The present section will provide information about the terminology from carbohydrate chemistry used all throughout the thesis such as the pyranose and furanose forms of sugars, their D- and L- enantiomers and the α- and β- anomeric configurations.

1. 1. 1 Pyranose and furanose conformations of sugars in solution

Acyclic aldehyde sugars in solution are spontaneously cyclized to form hemiacetals which are normally presented using the Haworth projection. These cyclic forms are either found in the six-membered ring form ie. pyranose or the five-membered ring *ie*. furanose forms (Figure 1. 1, *bottom*). The pyranose forms are thermodynamically most stable for most sugars in aqueous solution due to easily reachable transition states (Lough, 1972). The difference between the free energies of the two forms is large, reaching near 8 kJ/mol when neither of them exhibits steric clashes amongst their C_1-C_5 (or C_6) constituents (Angyal, 1984). As the constituents of pyranoses are fully staggered, these are more stable than five-membered rings of the furanoses. The staggered structure of pyranoses with the tetrahydrally co-ordinated oxygen atoms interacts better with the 'structured' component of liquid water as compared to furanoses (Angyal, 1984). The percentages of the pyranose and furanose forms of the sugars relevant to this study are listed in Table 1. 1.

Table 1. 1 Percentages of pyranose and furanose forms of selected sugars

Sugar	Pyranose		Furanose		Temperature (°C)
	α	β	α	β	
Arabinose	60	35.5	2.5	2.0	31
Galactose*	31.8	60.5	3.1	4.6	31
Xylose	36.5	63.0	<1	<1	31
Allose	14.0	77.5	3.5	5.0	31
Glucose	38.0	62	0.0	0.14	31
Fucose**	28	67	~5		31
Talose	42.0	29.0	16.0	13.0	22

Distribution of furan and pyran as percentages in D2O derived from Collins and Ferrier R., 1995.
* Galactose distribution is from Barlow and Blanchard, 2000.
** Fucose distribution is from Angyal, 1972

1. 1. 2 D and L enantiomers of sugars

The linear form of sugars *ie*. Fischer projection is normally used to differentiate between the enantiomers D and L of the sugars. In a monosaccharide when the

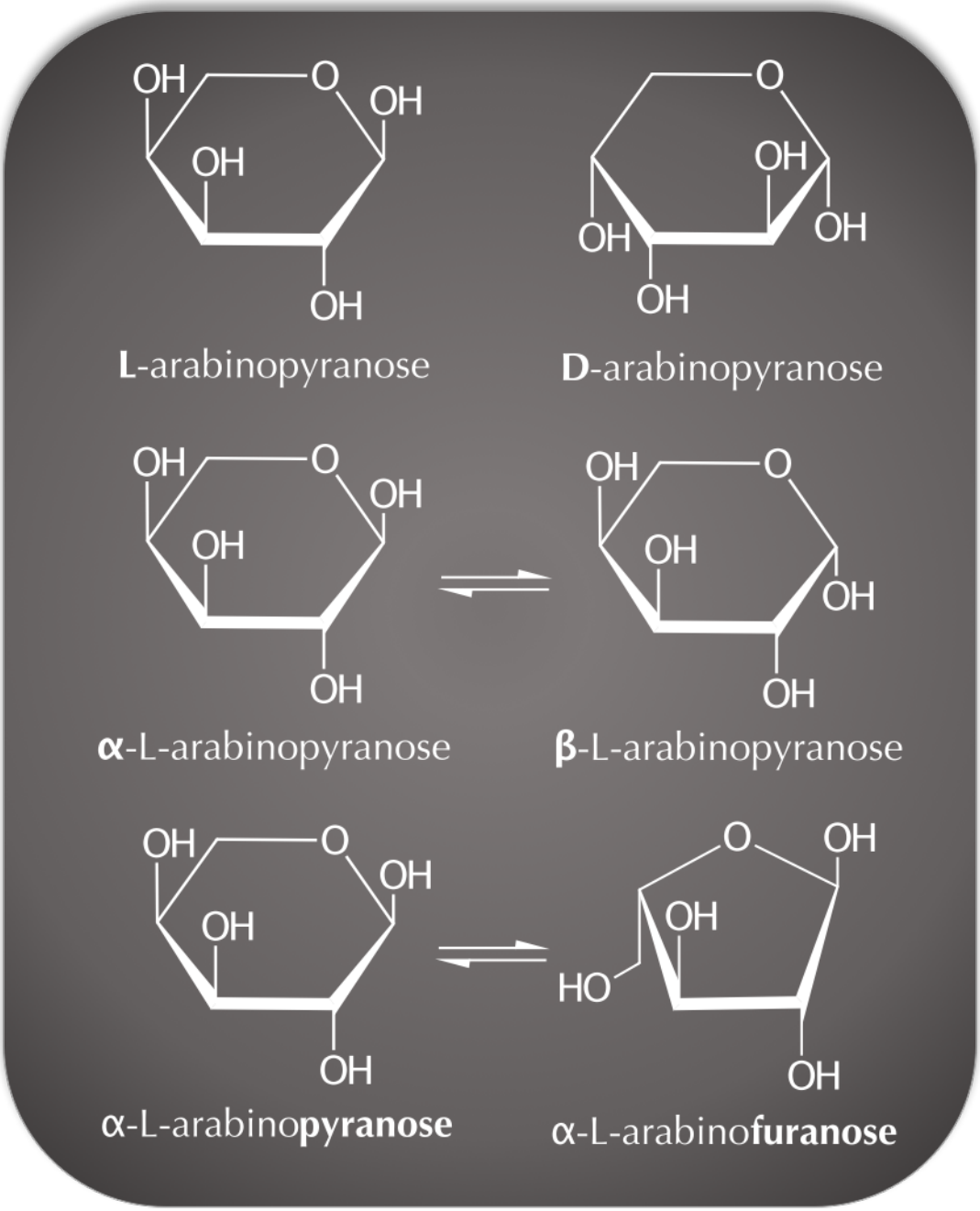


Figure 1. 1 Conformations of the sugar arabinose
All conformers and enantiomers are shown in the Haworth projection.

penultimate carbon is shown with a hydrogen on the left and a hydroxyl group on the right then it describes the D configuration of the sugar. However, when the opposite occurs *ie*. the last stereogenic carbon exhibits a hydrogen on the right and a hydroxyl group on the left, then the sugar is said to be an L enantiomer. In the Haworth projection, the cyclohexane structure of hexoses, will typically have their hemiacetal group (*ie*. C_6 constituent) pointing up. In the case of pentoses this distinction is becoming evident in the furanose form when the C5 constituent, for example, L-arabinofuranose is pointing down (Figure 1. 1, *top and bottom*) (Berg, 1980).

1. 1. 3 Sugar epimerisation

The epimerisation of a sugar relates to the anomeric center of a sugar which is created from the intramolecular formation of an acetal of a sugar hydroxyl group and an aldehyde *ie*. Fischer to Haworth projection (Osborn, 2006). The two possible stereoisomers formed are called anomers. In the Haworth projection, the a-anomer is the isomer with the anomeric substituent (ie. C1 or the carbon next to the anomeric oxygen of the ring) pointing on the opposite face to C5 hydroxymethyl group of hexoses. The β anomer is described as the one with the anomeric carbon facing the same direction as the C5 of hexoses. This applies for the D- configuration, in the case of L – the α anomer has its anomeric group pointing upwards and in the β anomer pointing down (Figure 1. 1, *middle*).

1. 2 Xylan biomass for biofuel production

Fast growth in both world energy consumption and carbon dioxide emissions correlated with the usage of fossil fuels has motivated the quest to search for alternative sources which are renewable and impose a minimal threat to the environment stability (IPCC, 2013). Biofuels are a viable alternative and have the potential to occupy a high percentage of a non petrochemical-based fuel economy. Current EU directives aim to implement an increase in the percentage of renewable sources mixed in the transport fuels and expect to reach a 10% share by 2020 (Directive

2009/28/EC). Popular candidate feedstocks for biofuel production are the first-generation including sugar crops (sugarcane), starch crops (corn, wheat) and oilseeds. Corn ethanol in US and sugarcane ethanol in Brazil are examples of the first-generation biofuels actively produced to meet the increasing global demand which has quadrupled between 2000 and 2008 (Sims *et al*., 2010). However, the use of food crops as bioenergy feedstocks has come under high scrutiny as it can be a detriment to the food supplies and potentially increase their cost (Sims *et al*., 2010). Instead, the second-generation, or advanced biofuels, are an economically viable alternative to the first-generation as these are derived from non-food crops, such as grasses and agricultural waste.

1. 2. 1 Lignocellulose component of plant biomass

The growing concerns for the use of first generation feedstocks have increased the interest in developing second-generation biofuels which are derived from non-food biomass. Such material includes the by-products of agriculture (*eg*. cereal straw, sugar cane bagasse and forest residues) and dedicated bioenegry feedstocks (*eg*. switchgrass, *Miscanthus* spp. and *Arundo donax*) which are normally low input and high yielding grasses. This type of biomass is continuously receiving attention as a successful resource of fermentable sugars for biofuel production because it doesn't raise the debate of food versus fuel.

Production of second-generation biofuels from plant biomass normally refers to the lignocellulosic component of them, as this makes up the majority of the nonfood materials derived from grasses and agricultural waste (Naik *et al*., 2010). Lignocellulosic biomass describes the secondary walls of plant cells and constitutes the most abundant raw material for the production of biofuels (Ebringerová and Heinze 2000; Naik *et al*., 2010). The lignocellulose is a complex structure formed by the association of polysaccharides including cellulose and hemicellulose (Haldar, Sen and Gayen, 2016). Cellulose is a linear polysaccharide consisted of glucose monomers linked by a β- 1, 4- glycosidic linkage and is the most abundant component of plant

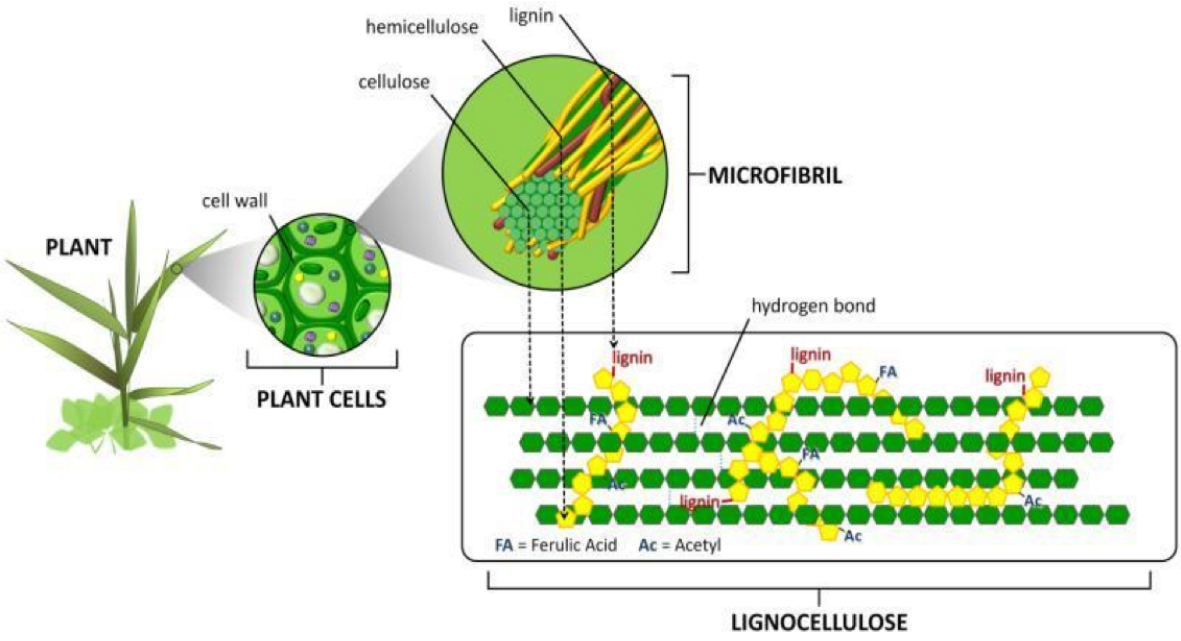


Figure 1. 2 Structure of lignocellulose

A simplified structure of plant fiber showing the crystalline cellulose in green and hemicellulose in yellow. The polymer is shaped as a mesh of highly structured polysaccharides interlinked with eachother and phenols (*ie*. lignin). The linkages between hemicellulose and lignin shown here are acetyl (Ac), and ferulic acid (FA). Figure obtained from Ribiero *et al*., 2016.

biomass on the earth (Tomme *et al*., 1995) (Figure 1. 2). The spaces between the two polysaccharides mentioned are filled by a complex organic polymer known as lignin (Figure 1. 2). The current study concentrates on the hemicellulotic component of the lignocellulose, and more specifically it focuses on the most common type of hemicellulose found in plants, *ie*. xylan.

1. 2. 2 Structure of the xylan polysaccharide

Xylan is the second most abundant polysaccharide in plants after cellulose, and constitutes the primary polysaccharide in the hemicellulose of plant cell walls (Siqueira and Filho 2010, He *et al*., 2014). It is structured in such manner so that its residues are supporting the formation of an overlaying layer characterised by hydrogen and covalent bonds which link the xylan to cellulose and lignin, respectively (Beg *et al*., 2001) (Figure 1. 2). Xylan polysaccharide is a heteropolymeric polymer comprised of a backbone of repeating xylopyranosyl residues (*shown in black in* Figure 1. 3) linked by a β-1,4 glycosidic bond. The backbone is decorated with branch chains of 4-O-methyl-D-glucuronic acid (Chen, 2014) (*shown in red in* Figure 1. 3).

Specifically, in the monocotyledonous Gramineae species, *eg*. bioenergy grasses, the hemicellulose also contains L-arabinofuranose (*shown in green in* Figure 1. 3) linking to the main chain as branch chains to the O2, ie. α-(1,2)-glycosidic bonds, and O3 positions, *ie*. α-(1,3)-glycosidic bonds, of the D-xylopyranose residues. The α-L-arabinofuranose residues are covalently bonded to ferulate residues (*shown in blue in* Figure 1. 3) which in turn link the AGX polysaccharide to lignin, as designated in Figure 1. 2 (Anna, Zdenka and Thomas, 2005). The xylan from hardwoods, *eg*. dicotyledonous angiosperm trees, differs from softwood xylans in that they lack acetyl groups (*shown in orange in* Figure 1. 3) and the aforementioned L-arabinofuranose units (Heinze, Barsett and Ebringerová, 2005). The studies on the type of sugars and the quantity of thereof in xylan of bioenergy grasses are important for the use of this polysaccharide in the biofuel production by consolidated bioprocessing.

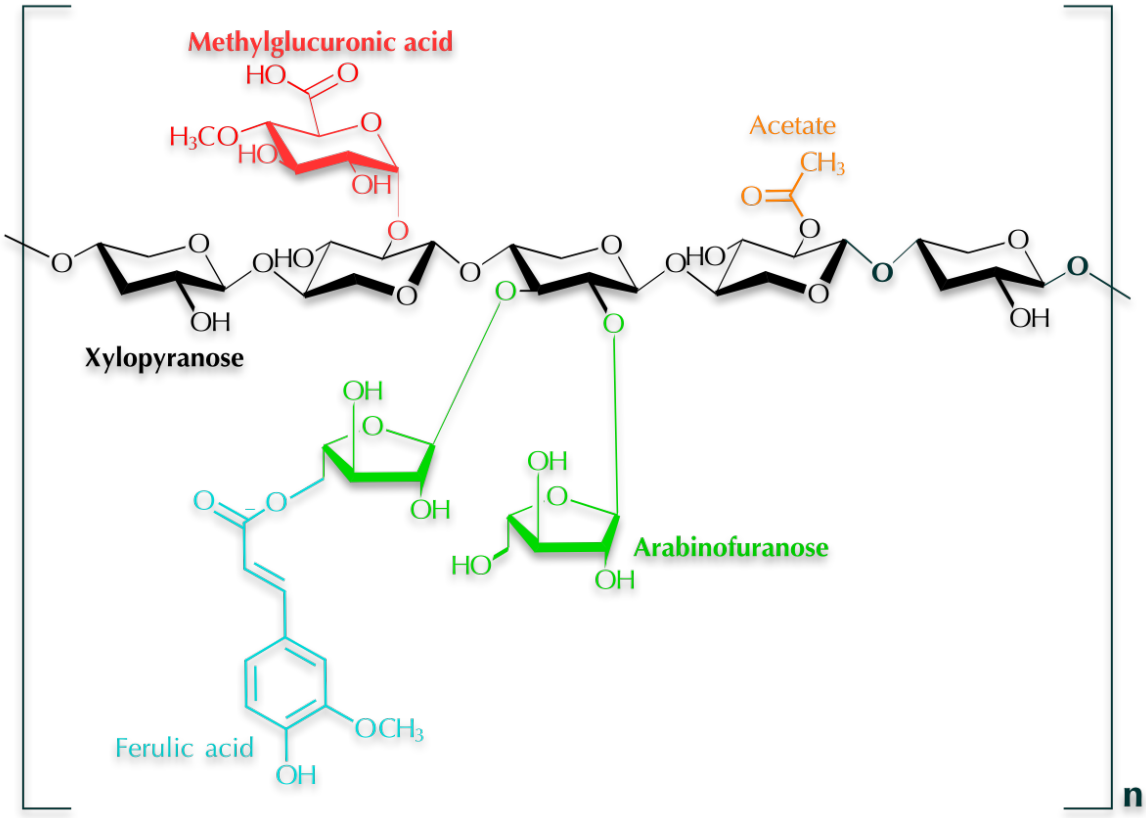


Figure 1. 3 Basic structure of xylan polysaccharide.
The xylan polysaccharide is comprised of a backbone of repeating xylopyranose residues with frequent decorations of arabinofuranose residues at positions 2- and 3-OH. Some of the arabinofuranose residues are linked ferulic acids which act as brigding units between xylan and lignin. Methyl glucuronic acids branch off O4 of xyloses. The structure was produced in ChemDoodle.

1. 1. 3 Consolidated bioprocessing of xylan by bacteria

Consolidated bioprocessing (CBP) of lignocellulose refers to the combining of the biological events involved in biomass-to-biofuel conversion. The process typically considers the optimisation of the following extracellular and intracellular reactions: 1. production of saccharolytic enzymes for the hydrolysis of polysaccharides present in pretreated biomass and 2. fermentation of hexose and pentose sugars in a bioreactor (Brodeur *et al.*, 2011). The current study is concerned with a stage involved in the bioconversion but is hugely neglected in the studies conducted so far which improved CBP processes. This stage is the transport of the hydrolysed sugars and oligosaccharides into the cytoplasm of the engineered bacterium that produces the biofuel. As discussed later, the optimisation of transport processes in the fungal organism, *Saccharomyces cerevisiae*, is essential for enabling the utilisation of xylose derived from xylan. The bacterial transport systems haven't received as much interest for this purpose, with the bulk of information gained for such systems derived from the study of their roles in bacterial pathogenesis, nutrient acquisition and response to extracellular ion content (Maloney, 2002). Limited studies exist to optimise the function of these transporters towards xylan utilisation and/ or engineer these systems in *E. coli* for biofuel production. The current study actively works in gaining more insights about the bacterial transport systems of the xylan-derived sugars arabinofuranose and xylobiose. Therefore, we suggest a review of the CBP as defined in literature to include the optimisation of the following stages: 1. Enzymatic hydrolysis of the pretreated biomass, 2. Uptake of hydrolysed sugars/ oligosaccharides via membrane transport systems, 3. Catabolism of imported sugars/ oligosaccharides and 4. Efflux systems for the produced biofuel (Figure 1. 4).

1. 3. Enzymatic hydrolysis of xylan

Enzymatic hydrolysis of biomass is a commonly employed step following its chemical and/ or mechanical pretreatment (Brodeur *et al.*, 2011). The diverse and heterogenic nature of xylan substitutions between different plant species gave rise to a plethora of

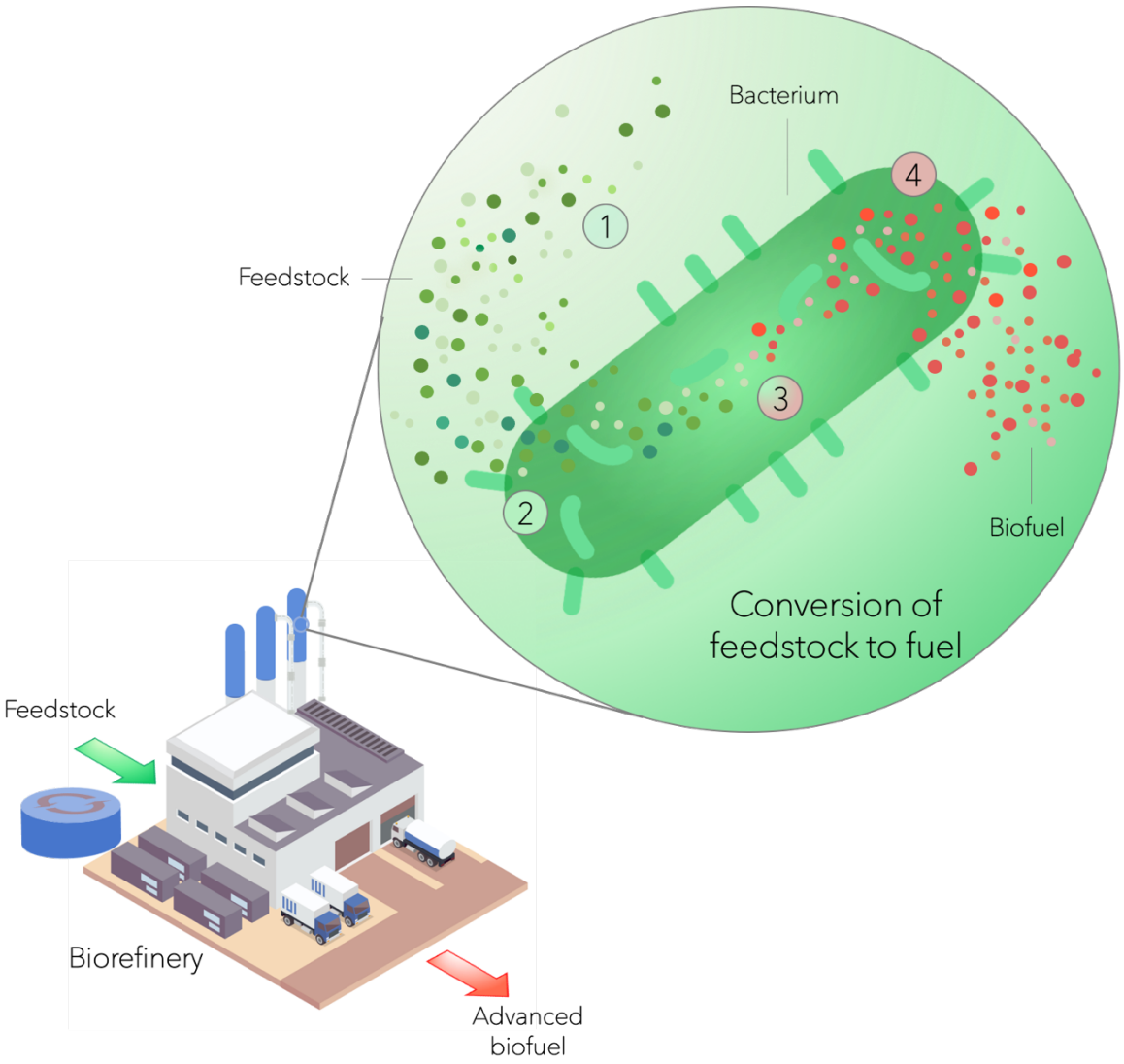


Figure 1. 4 Consolidated bioprocessing of biomass by bacteria
The four stages the CBP should be considering for efficient bioconversion of xylan to biofuel are: **1.** Enzymatic hydrolysis of the hemicellulotic polysaccharide, **2.** Uptake of oligo-, mono-saccharides and sugar acids via membrane transport systems, **3.** Intracellular catabolism of the importer sugars into biofuels and **4.** Diffusion/ Assisted efflux of biofuels.

glycosyl hydrolases which enable the bacteria to inhabit such niches. The xylan enzymatic hydrolysis requires the collaborative activity action of several main chain and side group cleaving enzymes, including endo-β-1,4-xylanases (EC 3.2.1.8), β-xylosidases (EC 3.2.1.37), α-arabinofuranosidases (EC 3.2.1.55), α-glucuronidases (EC 3.2.1.139), feruloyl esterases (EC 3.1.1.73) and acetylxylan esterases (EC 3.1.1.72) (de Vries and Visser 2001) (Figure 1. 5). The following sections provide an insight into the activity of these enzymes from studied examples, with emphasis to xylanases/ xylosidases and arabinofuranosidases which are most relevant to the current study. All of the enzymes described below are classified into glycoside hydrolase families (GH) in the Carbohydrate Active Enzyme (CAZy) (www.cazy.org) database on the basis of their amino acid sequence similarities (Cantarel *et al.*, 2009), and therefore will be referred to as such.

1. 3. 1 Xylanases and xylosidases

Xylanases are defined as hydrolases which randomly cleave β-1,4-linked D-xylopyranose residues from the xylan backbone and are widespread within bacterial and fungal phyla (Collins *et al.*, 2005; Juturu and Wu, 2012). Sequence based classification by CAZy database created two distinct groups *ie*. GH10 and 11 with also in GH5, 7 and 8 (Lafond *et al.*, 2014). GH10 family comprises primarily by endo-1, 4-β-xylanases with some of its members being endo-1, 3-β-xylanases. This family also shows versatility on substrates, with some GH10 enzymes able to hydrolyse cellulose. Further, GH10 can tolerate glucose substitutions on xylose monomers, commonly found xyloglucans (Chakdar *et al.*, 2016). Moreover, these enzymes are highly active on small xylooligosaccharides, indicating they can still processes further xylan hydrolysis products (Pollet, Delcour and Courtin, 2010) (*shown in black in* Figure 1. 5). GH10 normally have a high molecular mass, low pI and a $(\alpha\text{-}\beta)_8$-barrel fold conformation (Teplitsky *et al.*, 2000). GH11 typically cleave xylan and xylooligosaccharides at β-1,4-xylosidic bonds, and their catalytic products are normally substrates for GH10 enzymes (Chakdar *et al.*, 2016). GH11 enzymes have smaller catalytic versatility as they only cleave unsubstituted xylose monomers which

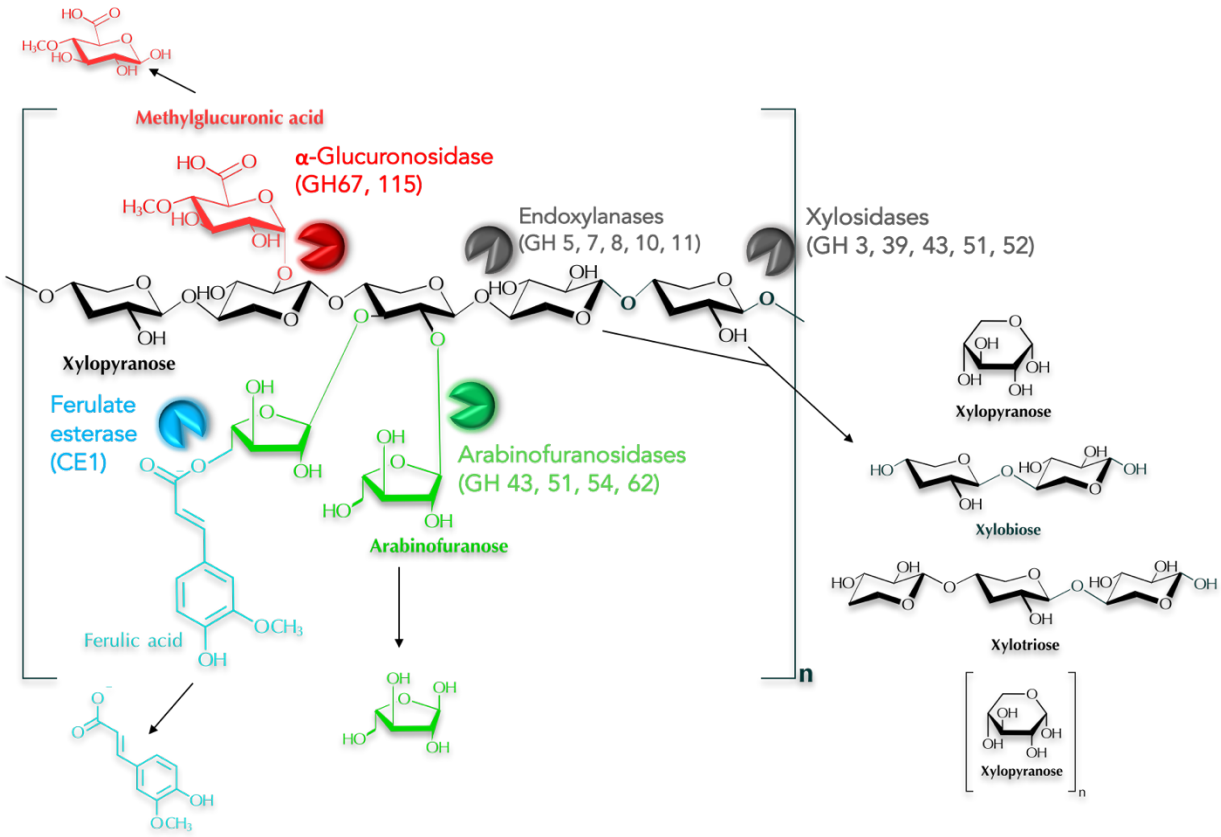


Figure 1. 5 Consortium of bacterial hydrolases which act on xylan

The enzymes are represented by circular sectors near the cleavage point. The structure of the monomers released are represented in black for xylose, green for arabinofuranose, blue for ferulate and red for methylglucuronic acid. The GH families are designated for each enzymatic function. The structure was produced in ChemDoodle.

was confirmed by their inability to attack the xylosidic linkage towards the non-reducing-end next to a branched xylose. The rarer GH8 enzymes share similarity to GH11 in terms of their substrate specificity and action pattern, with 6 members already identified to be able to act on xylan (Collins, 2006; 2008).

Xylosidases (EC 3.2.1.37) are required following xylanase and arabinofuranosidase activities. These cleave xylose residues from the non-reducing end of the xylose chain in an exo-acting manner. The xylanases act in close synergy with the xylosidases as the first create more reducing ends for the action of the latter. Further, arabinofuranosidases remove the arabinose side chains from AX and AGX polysaccharides to allow binding of β-xylosidases. Also, xylosidases remove the end products that otherwise prohibit xylanase activity (Sunna and Antranikian, 1997).

Xylosidases from family **GH3** are active in a range of substrates including AX, xylan and xylooligomers. The degree of polymerisation (DP) of the substrates for two GH3 enzymes from *Hypocrea jecorina* and *Talanomyces emersonic* ranged from $(xyl)_2$ to $(xyl)_5$. Their ability to hydrolyse showed a decrease at oligomers with DP higher than 6 (*ie*. equal or longer in size than xylohexaose). **GH39** xylosidases from *Thermoanaerobacterium saccharolyticum* and *Clostridium stercorarium* are active on xylooligomers with minimal activity on xylobiose, and much higher on xylotriose (Lee and Zeikus 1993, Wagschal *et al*., 2005). Neither show activity on AX, however *C. stercorarium* still binds the polymer (Adelsberger *et al*., 2009). The **GH43** xylosidases from *S. ruminantium*, *G. thermoleovorius* and *B. adolescentis* are all active on xylooligomers ranging from $(xyl)_2$ to $(xyl)_6$. The *S. ruminantium* was particularly active on xylotriose, distinct to the rest which didn't show any preference on the latter (Chakdar *et al*., 2016). The **GH52** enzymes are a small family of xylosidases. Two examples have been found in *Aeromonas caviae* and *Geobacillus stearothermophilus* 21 which hydrolyse xylooligomers to xylose (Nanmori *et al*., 1990). The enzyme from *Aeromonas caviae* were found to possess transglycolysation activity is observed as incubation leads to hydrolysis but production of xylotetraose and xylopentaose. The same is observed with *G. stearothermophilus* T-6 when incubated with pN-xylose, the

released xylose is made into xylooligomers. The study of the substrate specificity, activity and other biochemical properties of the abovementioned enzymes helps evaluate their usefulness in xylan hydrolysis.

1. 3. 2 Arabinofuranosidases (AFs)

The L-arabinose moieties that decorate the xylopyranose backbone can be released enzymatically by the activity of bacterial α-L-arabinofuranosidases (AFs) (Saha, 2000) (*shown in green in* Figure 1. 5), of which a number are known and are being investigated for commercial applications (Kurakake *et al.*, 2011; McKee *et al.*, 2012). These ezymes are members of five different families of glycosyl hydrolyses including GH3, GH43, GH51, GH54, and GH62, and can hydrolyze by net inversion (GH43) or retention (GH51, GH54) mechanism. Both reaction mechanisms release an arabinofuranosyl residue with the first causing inversion of the anomeric carbon whereas the latter conserves the carbon at its initial position (McCarter and Withers, 1994). In the context of AGX, the xylose monomers may be either singly or doubly substituted with arabinose, and this has significant consequences in terms of the type of enzyme needed to remove the arabinose side chains (Lagaert *et al.*, 2014). Also, some arabinose molecules may have ferulic acid molecules esterified to the 5'-OH which may subsequently be covalently tethered to ferulic acid residues attached to other arabinoxylan chains. Thus, most studies aim to characterize the exact substrates of the AFs in order to determine how appropriate they are to be used in enzyme cocktails or to decide what GHs be expressed by fermenting bacteria during deconstruction of xylan. The AFs of GH62 and GH54 families can release arabinose from AGX and branched arabinan (Lagaert *et al.*, 2014). Similarly, AFs from GH3 can cleave arabinose from arabinan, AGX and arabino-oligosaccharides (Lagaert *et al.*, 2014). AFs from the GH51 family are more interesting in that they have a broader range of substrates including the AGX, arabinoxylan- oligosaccharides, arabinan and arabino- oligosaccharides. Equally interesting is the family GH43 which has members that hydrolyse arabinose at both C(O)3 and C(O)2 positions (Lagaert *et al.*, 2010; McKee *et al.*, 2012, Sørensen *et al.*, 2006, 2007). A very intriguing candidate from this family is ACF39706, which was isolated from a compost mixture, and displays

bifunctional xylosidase/arabinofuranosidase activity. It was shown to be able to cleave both xylose and arabinose from arabino-oligosaccharides, arabinan and AGX (Wagschal *et al.*, 2009). Its activity agrees with its positioning between xylosidases and arabinofuranosidases in a phylogenetic tree constructed by Lagaert *et al.*, (2013), indicating that it could be an evolutionary indermediate. Such candidate AFs with broad substrate specificity and bifunctionality are great for biotechnological applications as they decrease the cost and complexity attached with using multiple enzymes. The extensive research for the efficient enzymatic release of L-arabinose derived from xylan underlines the importance of this sugar in the production of biofuels from lignocellulose-based feedstocks. Other enzymes specific to arabinan ie. polysaccharide of repetitively connected L-arabinofuranose sugars are the arabinosidases (EC 3. 2. 1. 99). These catalyses the endohydrolysis of 1-5-α arabinofuranosidic linkages to release arabinosides from arabinan which can be further decreased in size by the terminal-processing arabinofuranosidases.

1. 4 Transport and catabolism of xylan-derived sugars

The bacterial cell membrane is comprised of phospholipids which separate the bacterial cytoplasm from the ever-changing extracellular environment. These membranes owe their existence to the amphiphilic nature of the phospholipids (Figure 1. 6). A single phospholipid is comprised by a polar and hydrophilic head and a non-polar hydrophobic fatty acid tail. This characteristic allows them to arrange themselves in a bilayer, in such way that their heads are pointing outwards interacting with the water molecules via their hydrophilic phosphate groups, and their hydrophobic tails facing inwards towards the middle of the bilayer (Wilkins *et al.*, 1971). In Gram (*+ve*) bacteria, the membrane is attached to a thick layer of protective peptidoglycan, which collectively are known as the cell envelope. In the case of Gram (*-ve*) bacteria, these possess an extra membrane, *ie*. outer membrane, above a much thinner peptidoglycan layer (Figure 1. 6). The compartment formed in between the inner and the outer membrane is known as the periplasm. The bacterial membrane is

one of the most dynamic and fluid structure in the cell, and its predominant function is to regulate the passage of substances into and out of the cytoplasm. The membrane embedded or attached proteins that mediate the passage of the solutes through the membrane are universally referred to as transport systems or permeases (Romano, 1986; Padan, 2009). These are the gateways for the entry of raw materials which are utilised by bacteria to produce products of economic value (Romano, 1986). Bacteria own a range of transporters which have been classified based on their mode of action and how they fuel the passage of the substrate across the membrane. These systems are described in detail in the following sections.

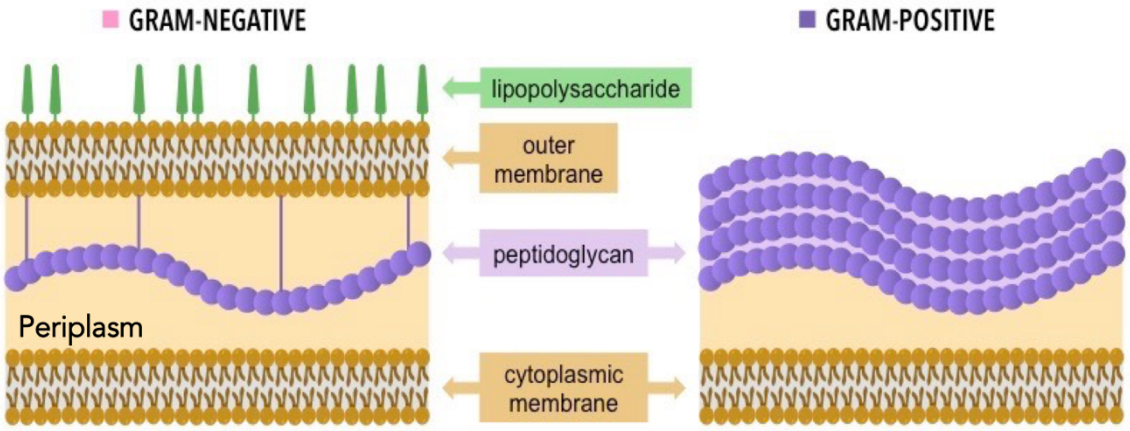


Figure 1. 6 The cell membrane of Gram (-ve) and Gram (+ve) bacteria
The distinct difference between the cell membrane of Gram negative and positive bacteria is the lack of an outer membrane in the latter. A thick peptidoglycan layer is a characteristic of the Gram (+ve) bacteria whereas the (-ve) posses only a thin layer found inside the periplasmic compartment. Figure taken from Berg, 2015.

1. 4. 1 Major types of bacterial transporters

The released sugars/ sugar acids from the chemical pretreatment and the enzymatic hydrolysis of the xylan will have to face and surpass the double barrier present on the cell membrane of Gram (*-ve*) bacteria. In the current study, we are only focused on how the sugars make their way into the cytoplasm through the inner membrane with limited reference on the outer membrane transport.

1. 4. 1. 1 Facilitated diffusion

The outcome of unaided diffusion is an equilibration of solute concentrations on both sides of the bacterial membrane. When this equilibration is established by the function of a protein channel, the process is known as facilitated diffucion and constitutes the simplest form of transport. This process is not energy consuming and is known as passive transport. These channels form the first class of the transport classification (TC) system (TC class 1, subclass 1. A. 1-36). One of the firstly discovered diffusion channels is the glycerol facilitator, GlpF (TC 1. A. 8. 1. 1) of *Escherichia coli* (Figure 1.7A) (Sweet et al., 1990, Fu et al., 2000).

1. 4. 1. 2 Secondary transporters

Unlike facilitated diffusion, the accumulation of a substance against a concentration gradient requires active transport and the expense of energy. The laws of thermodynamics define that some sort of energy is present to perform the uphill transport of substrates; this is achieved by the downhill gradient that could be dissipated (*ie*. translocation of a solute, different to the transported substrate, from its high to low concentration areas) or by the expense of chemical energy (Krämer, 1994). The type of transporters which employ the first method are known as secondary systems and the primary carriers use the latter (Krämer, 1994). The secondary active transporters (TC class 2, subclass 2. A. 1-80) energise the thermodynamic uphill transport of a substrate utilising the energy obtained from the thermodynamically favourable import or export of another solute, normally H^+ or Na^+ (Saier Jr, 2000b)

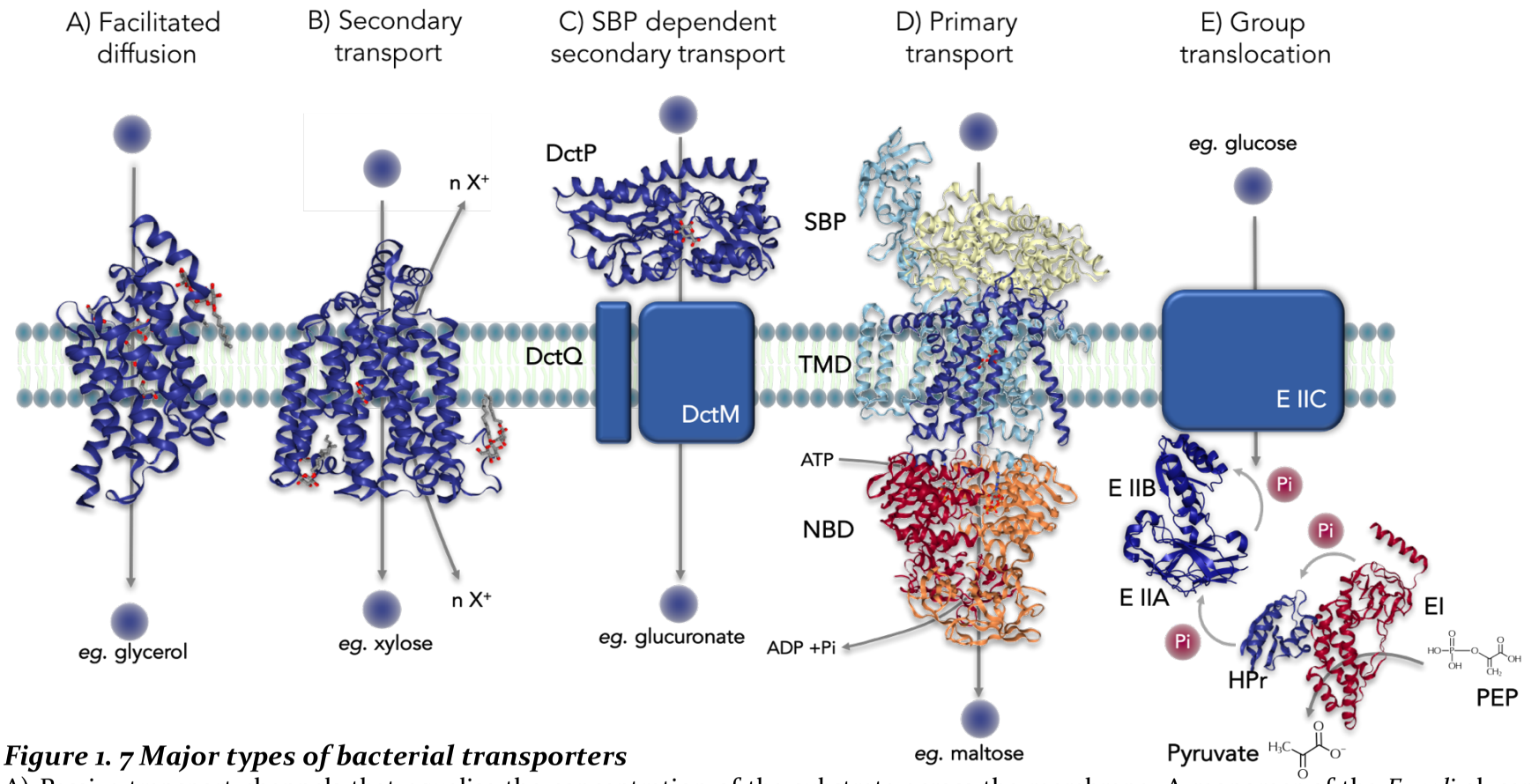


Figure 1. 7 Major types of bacterial transporters

A) Passive transport channels that equalise the concentration of the substrate across the membrane. A monomer of the *E. coli* glycerol facilitator (GlpF, 1FX8) is shown with three molecules of glycerol bound. B) Secondary active transporters couple translocation of the substrate with the electrochemical gradient across the membrane. The example given here is the *E. coli* xylose MFS symporter, XylE (4GBY). C) SBP dependent secondary transporters utilise a SBP to recognise their substrate and the energy of secondary transport to facilitate import. The example of a TRAP SBP shown here is a D-glucuronic acid DctP (4N15) from *Burkholderia ambifaria* D) Primary transporters use ATP energy to translocate the substrate. The example given here is a catalytic intermediate of the *E. coli* maltose transporter, MalEFGK2 which hydrolyses ATP hydrolysis to allow substrate translocation. E) Group translocators. This group catalyse transport of the substrate via phosphorylation. The example shown here the PTS for glucose from *E. coli* containing the following components: EIIC-(EIIB-EIIA, 1O2F) enzyme II and HPr-EI (3EZB). In each example the substrate is represented with a blue sphere.

(Figure 1. 7B). These ion-driven systems are classified into three groups; uniporters, antiporters and symporters. Uniporters translocate one substrate at a time in one direction and the latter two transport two different substrates at the same time with inwardly or outwardly directed polarity. Symporters move two solutes towards the same direction; conversely, antiporters transport two substrates in opposite directions. A huge family comprised of members from the class in question, is the Major Facilitater Superfamily (MFS) (Pao, Paulsen and Saier, 1998). An example of a well-studied symporter is the xylose MFS transporter from *E. coli* (Figure 1.7B). This is a xylose:H^+ symporter where the translocation of extracellular protons into the low proton environment of the cytoplasm drives the accumulation of intracellular lactose (Davis and Henderson *et al*., 1970). Despite the inherent difficulty in the crystallisation and structure elucidation of membrane embedded systems, there are XylE structure deposited in the Protein Data Bank (PDB) reporting the symporter in different conformational states (Quistgaard *et al*., 2013). This is mainly because of its high relatedness to the human glucose transporters (GLUT), which have been found to be implicated in brain diseases such as Alzheimer's (Shah, DeSilva and Abbruscato).

1. 4. 1. 3 Primary transporters

Another major class of active transporters are the primary carriers. Primary transport uses energy directly; the light or chemical energy is converted into electrochemical energy which is used to transport a substrate against its concentration gradient.

Primary transporters related to sugar uptake include the Phosphotrasferase systems (PTS) (TC class 4, subclass 4. A. 1-6), known also as group translocators. The bacterial PTS is a multiple component carbohydrate uptake system which facilitates the transport of monosaccharides across the inner membrane while concomitantly catalysing sugar phosphorylation. The series of events involved constitute a phosphorelay catalysed by a chain of soluble phosphotransferases (EI, HPr, EIIB and EIIA) which leads to an integral membrane protein (EIIC) which in turn imports the sugar into the cell (McCoy, Levin and Zhou, 2015) (Figure 1. 7E). The initial donor of

the phosphate group is the intermediate phosphoenolpyruvate (Kaback, 1968). Enzyme I (EI) catalyses the removal of the phosphate to produce pyruvate and subsequently donate it to the heat-stable phosphocarrier protein (HPr). Next, the phosphate group is passed on to the multimeric enzyme II (EII). EII is composed of three separate domains named EIIA, EIIB, and EIIC (Figure 1. 7E). The phosphate moves serially from HPr to EIIA, and then to EIIB. Binding of the sugar to the EIIC, causes the phosphate to be transferred from EIIB to the sugar and allow its passage through the EIIC pore (Begley, 1982). This last transfer occurs without the need of a covalent EIIC-phosphate intermediate (Begley, 1982).

Another major class of sugar-related primary transporters are the **A**denosine **T**ri**P**hosphate - binding cassette (ABC) transporters (Figure 7D). ABC transporters form a large superfamily of proteins with common function and an ATP-binding domain. These normally exhibit high affinity for a single or multiple related substrates (Wilkens, 2015). In general, they are consisted of three overall components usually arranged in a co-transcribed operon. ABC transporters are often consisted of one or two transmembrane proteins and one or two ATPases which are attached to the cytoplasmic side of the inner membrane. The membrane associated ATPases utilise energy from the hydrolysis of ATP molecules to fuel the active transport, hence the superfamily name of ABC transporters. Most but not all of these uptake systems have a **Substrate Binding Protein** (SBP), also known as Extracytoplasmic Solute Receptor (ESR) (Wilkens, 2015). The SBPs are found in the periplasmic space of Gram (*-ve*) bacteria, or tethered on the cell wall of the Gram (*+ve*)s, and are responsible for the recognition and binding of the solute (Wilkens, 2015). The most studied example of this family is the *E. coli* maltose transporter ($MalEFGK_2$) (TC3. A. 1. 1. 1) (Oldham *et al*., 2007).

1. 4. 1. 4 SBP-dependent secondary transport

A class characterised by a fusion of primary and secondary transport are the SBP-dependent systems. These systems were identified much later than ABC transporters,

in what was an unexpected discovery in the late 1990s to find that SBPs can form part of an electrochemical gradient-driven (secondary) transporter (Forward 1997; Kelly and Thomas, 2001). The discovery was made by cloning and sequencing of genes encoding a C4-dicarboxylate SBP (DctP) from *Rhodobacter capsulatus* and a citrate SBP (TctC) from *Salmonella enterica* subsp. Typhimurium which showed that none of them was part of a usual ABC transporter-encoding operon (Forward 1997; Kelly and Thomas, 2001; Mulligan *et al*., 2007; Widenhorn *et al*., 1988). These SBPs were found to be part of a *dctPQM* of *Rhodobacter capsulatus* and *tctABC* of *Salmonella enterica* subsp. Typhimurium. These systems were showed to uptake dicarboxylates independent of ATP but with the exploiting the electrochemical ion gradient instead, making this a unique binding protein-dependent secondary transporter. The discovery of cognate systems and identification of more members of this newly found family led to the definition of the TRAP transporters (for tripartite ATP-independent periplasmic) which are encoded by *dctPQM* operons and TTT (for tripartite tricarboxylate transporters). The TRAP transporters are a large family of SBP-dependent transporters which are found in bacteria and archaea but not in eukaryotes (Mulligan *et al*., 2007). The most appropriate to the current project are the TRAP transporters which due to the specificity of their SBPs to recognise substrates with a carboxylate or sulfonate group, they are able to import D-glucuronic acid (Vetting *et al*., 2015), a xylan branching side chain. The D-glucuronate SBP (*ie*. $DctP_{glcA}$) from *Burkholderia ambifaria* is one of the recently structurally characterised DctPs (Figure 1.7C) (Vetting *et al*., 2015). This system, similarly to any other TRAP transporter, includes an SBP (or ESR) designated as DctP, and two membrane components: a small one consisted of 4 transmembrane segments (*ie*. DctQ) and a larger domain made of 12 transmembrane helices (*ie*. DctM) (Figure 6C) (Forward et al., 1997; Vetting et al., 2015).

1. 4. 2 Overview of secondary transporters

The secondary transporters (Figure 1. 6A) normally present lower affinities for their substrates compared to the primary transporters and most importantly they don't use

ATP to fuel transport. Instead, they combine the translocation of their substrates with the movement of a monovalent cation such as Na^+ or H^+, thus exploiting the available energy accumulated by the electrochemical potential gradient of the membrane. The three methods of secondary transport *ie*. symport, antiport and symport along with a description of their mode of action are provided in Figure 1. 8. There are four superfamilies which contain secondary carriers able to transport sugars (Chang *et al*., 2004). The Major Facilitator Superfamily (MFS) is one of them with six of its 29 families involved in sugar/ oligosaccharide transport (Chang *et al*., 2004). The MFS are most relevant to uptake of xylan-derived sugars with characterised members for xylose, arabinose and glucuronate uptake.

1. 4. 2. 1 Major Facilitator Superfamily (MFS)

The MFS transporters are found ubiquitously in bacteria, archaea and eukaryotes. The family includes members that act by solute uniport, solute-cation symport and solute-cation and/or solute-solute antiport (Pao, Paulsen and Saier, 1998) (Figure 1. 8). These transporters typically contain a 12 transmembrane segment (Reddy *et al*., 2012). Substrates of the MFS family include amino acids, vitamins, nucleotides and of particular relevance here; sugars and oligosaccharides (Pao, Paulsen and Saier, 1998).

The Sugar Porter family is the biggest of the MFS and includes both uniporters and solute/cation symporters. The family in question includes the phylogenetically related low affinity secondary transporters for pentose sugars: AraE (arabinose:H^+) and XylE (xylose:H^+) (Chang *et al*., 2004). The aforesaid transporters have extended sequence conservation and they both feature the 12 TM α-helical structure of the MFS (Henderson and Maiden, 1990). Their similarities extend to substrate specificity as AraE can also transport xylose (Krispin and Allmansberger, 1998).

The XylE has attracted a lot of interest from a plethora of studies due to its high homology to the human glucose transporters GLUT1-4 (Henderson and Maiden, 1990). The GLUT1-4 transporters are found in brain cells and have previously been

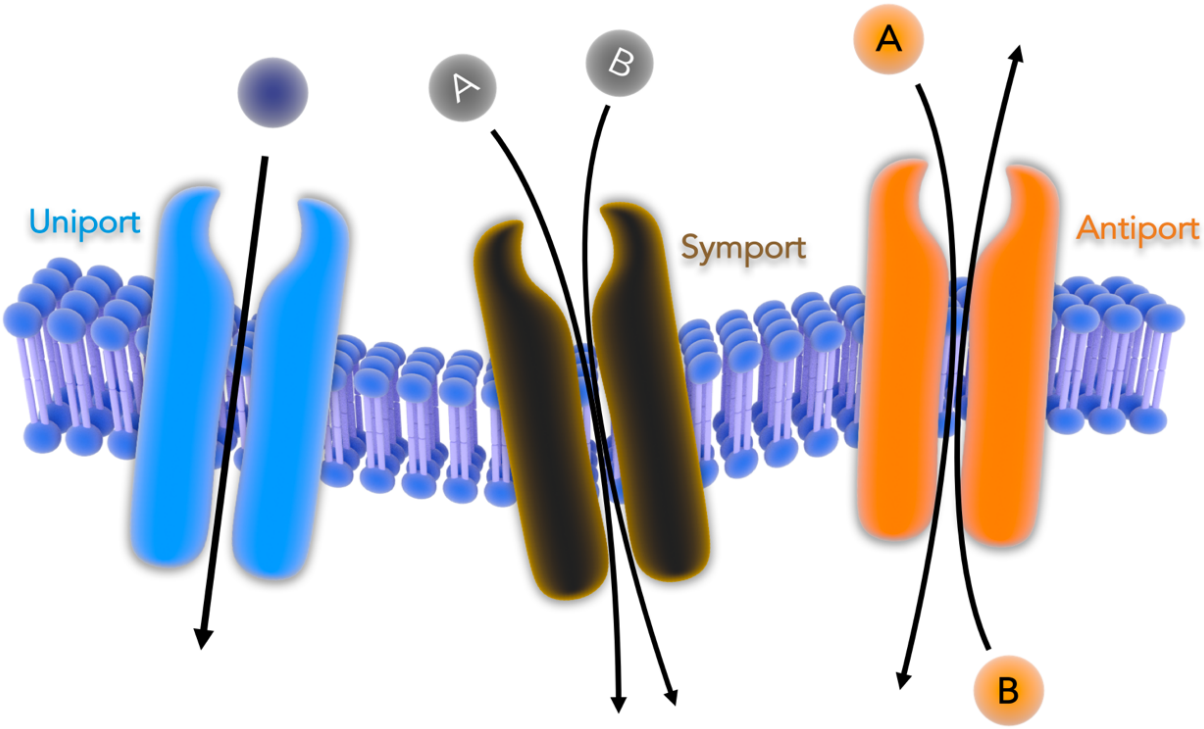


Figure 1. 8 Types of secondary transporters
"A" designates a transported solute *eg*. a sugar molecule or oligosaccharide; "B" denotes the co-transported ion. The bilayer made of phospholipids is shown in dark blue. Uniport (blue): facilitated translocation of a single molecule at a time; the solute moves down its concentration gradient and reaches equilibrium. Symporters is a type of cotransporter which couple the transport of a neutral molecule (A) against its concentration gradient utilising the energy produced by the transport of a cation in the same direction, down its concentration gradient. Antiporters catalyse the countertransport of a solute against a cation, fuelled in the same manner as symporters, accumulating the solute against its concentration gradient. In some cases antiporters can transport two solutes where one molecule is imported in exchange for another substrate (in this case B is another substrate not an ion).

implicated in Alzheimer's disease which raises their importance in research (Shah, DeSilva and Abbruscato, 2013). Even though all of the key ligand binding residues are conserved between GLUT1-4 and XylE (apart from Gln415), XylE shows smaller specificity and remains 'loyal' to xylose. XylE is inhibited by glucose and no charged residue is involved in binding of xylose (Madej *et al*., 2014).

An MFS system that has been exhaustively studied is the lactose permease from *E. coli*, *ie*. LacY. This transporter belongs to the glycoside-pentoside-hexuronide (GHP) family and it is a lactose/melibiose:H^+. The galactose containing substrates reflect the requirement for a galactopyranosyl ring of D-configuration with a free hydroxyl group to be present for the binding to occur (Sandermann, 1977). The substrate is transported with a H^+ ion in a 1:1 stochoimetry (Guan *et al*., 2006) against its concentration gradient. The 12 TM helices of the symporter are structured into two bundles of six linked by a long flexible loop. A central hydrophilic cavity is formed between the N-terminal domain (TM 1-6) and C-terminal domain (TM 7-12), with the first containing the sugar binding site and the second the residues involved in proton translocation (Kaback, 2006). LacY employs the alternating access mechanism of transport (Guan *et al*., 2006; Kaback *et al*., 2010), also known as rocker-switch mechanism (Slotboom, 2013). The mechanism involves conformational changes which facilitate the alternating access of the single sugar and proton binding sites to either side of the membrane, allowing binding when the transporter is in the outward-facing state and release of the ligand and proton when it switches to the inward-facing position (Figure 1. 9).

The current study will utilise secondary systems to complement mutant strains which are unable to grow on xylan-derived sugars. Various other studies have attempted overproduction of membrane systems for complementation or to characterisation their function. The effects of their overproduction were studied by Wagner *et al*., 2007. The group created fusions of the membrane proteins YidC, YedZ and LepI to GFP and targeted these for overproduction in the membrane. They observed less respiratory chain complexes (ie. succinate dehydrogenase and cytochrome *bd* and *b03* oxidases

were reduced) in the membrane possibly caused by the saturation of the Sec-translocon capacity. Further, the low Sec translocon capacity leads to accumulation of precursors of secretory proteins in the cytoplasm, which subsequently causes toxicity. The state of the respiratory chain in *E. coli* is controlled by various redox sensor systems including the Arc system. The data produced by Wagner *et al.*, 2007 showed that the acetate-pta pathway was induced due to Arc system activation which is normally caused when Q pool is found in a mostly reduced state (*ie*. underactive respiratory chain). The size and function of the overproduced protein did not contribute in the level of toxicity, therefore improving the fitness of *E. coli* overproducers should focus on the Sec translocon system and increasing the efficiency of respiration under such conditions.

1. 4. 3 Overview of ABC transporters

The ABC transporters are a widespread type of primary carriers with representatives in all three domains of life ranging from bacteria to eukaryotes, including humans (Holland, 2003). They constitute one of the largest superfamilies and have been implicated in many human diseases including cystic fibrosis (Gadsby *et al*., 2006). Dreyfuss was one of the first to name these transporters during the early 1960s with the discovery of the thiosulfate transporter from *Salmonella typhimurium* (Dreyfuss, 1964). However, it was not until later in 1980s that the first nucleotide sequences of these systems were reported with the first being the histidine-specific ABC system *ie*. HisJQMP (Higgins *et al.,* 1982). Later, Hyde *et al.*, (1990) formally introduced the term ABC transporters in the first comprehensive review of these systems (Hyde *et al*., 1990).

While the composition of ABC transporter components is fairly variable, these can be divided into two categories; ABC importers and exporters. ABC importers catalyze the unidirectional uptake of amino acids (Higgins *et al*., 1982), of sugars (Dean *et al*., 1989), peptides (Tame *et al*., 1994) and vitamins (Borths *et al*., 2002). ABC exporters are known to catalyse the efflux of virulence factors such as haemolysin (Dinh *et al.*, 1994).

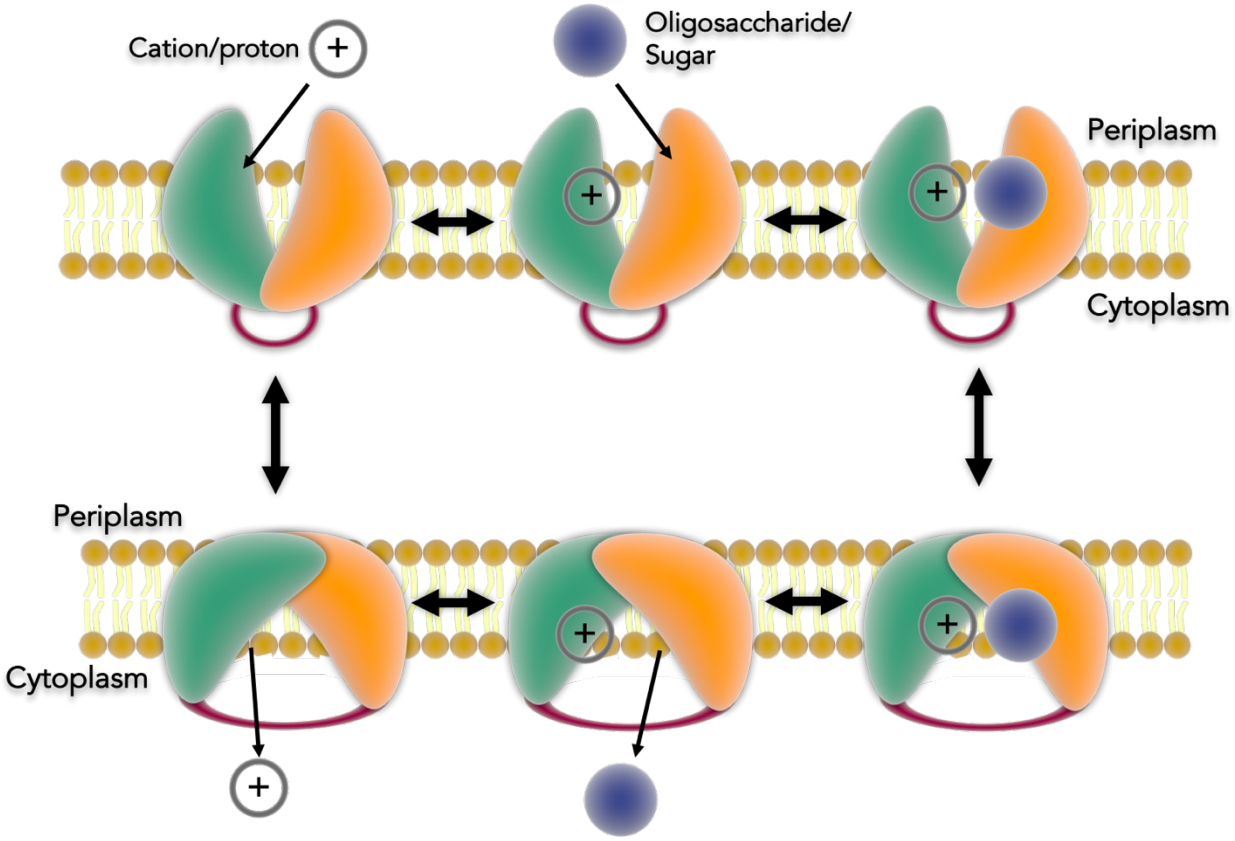


Figure 1. 9 Mechanism of transport used by cation/proton symporters.
The catalysis of transport can occur in either direction across the membrane which causes influx or efflux of molecules, indicated by the bidirectional arrows. The process of influx is described starting from the top left and moving clockwise. The cation or proton binds first at its periplasmic exposed binding site. Then the substrate (*eg.* sugar or oligosaccharide) binds from the same side of the membrane, to its specific site on the protein at the hydrophobic cavity of the transporter. Binding of both substates causes a conformational change in the symporter which switches from the outward to the inward (cytoplasmic) facing conformation. First to be released into the cytoplasm is the sugar, followed by the cation/proton. After release of the substrates the empty symporter returns to the outward (periplasmic) facing conformation to continue with another round of binding. The phospholipid bilayer is shown in brown. Green (C-terminal domain) and orange (N-terminal domain) ovals denote the two halves of the symporter protein.

1. 4. 3. 1 Architecture of ABC transporters

The part of the ABC systems which recognises and binds the substrate with high affinity are known as **Substrate Binding Proteins** (SBPs) or **Extracytoplasmic Solute Receptors** (ESRs) (Figure 1. 10). These were the firstly identified components of ABC transporters because they reside in the periplasmic space and therefore are easily extractable by cold osmotic shock (Neu and Heppel, 1965). These soluble proteins are responsible for substrate binding in a central cavity formed between two lobes, and deliver it to the membrane embedded domains, *ie*. transmembrane domains (TMDs), present in the cytoplasmic membrane. In Gram (+*ve*) and archaea, which lack an outer membrane and thus don't have an extensive periplasmic space, the SBPs are tethered on the cell surface exposed to the extracellular environment. They normally attached to the cell wall via a lipid-anchor which involves a linkage with a cysteine residue or in the case of archaea via a transmembrane peptide. Some systems, *eg*. glycine betaine ABC system OpuABCD, have two substrate-binding domains (SBDs) with each fused to one of the TMDs. In other cases, the SBDs are used to eachother with the formed SBD complex attached to the TMDs thus resulting in a tetramer with two four extracytoplasmic substrate binding sites, *eg*. in the glutamine-binding system, GlnPQ. These systems are most frequently encountered in Gram (+*ve*)s and less often in Gram (-*ve*) bacteria (van der Heide and Poolma, 2002). The presence of SBPs are not in all cases necessary as there are examples of redundant systems such as the *aglEFGAK* and *thuEFGK* from *Sinorhizobium meliloti* Rm1021 which are both active in the uptake of trehalose, sucrose and maltose (Jensen *et al*., 2002). Also, mutations can act to allow survival upon deletion of SBP- encoding genes, *eg*. deletion of *malE* led to further mutation of the respective strain and produced a maltose system which was fully functional containing the TMDs and ABC domain only (Davidson *et al*., 1992). The core of ABC systems is consisted of **TransMembrane Domains** (TMDs) and the **Nucleotide Binding Domains** (NBDs) which collectively make up the translocator (Figure 1. 10). Each one of the TMDs is normally consisted of six α-helices which transverse the cytoplasmic membrane.

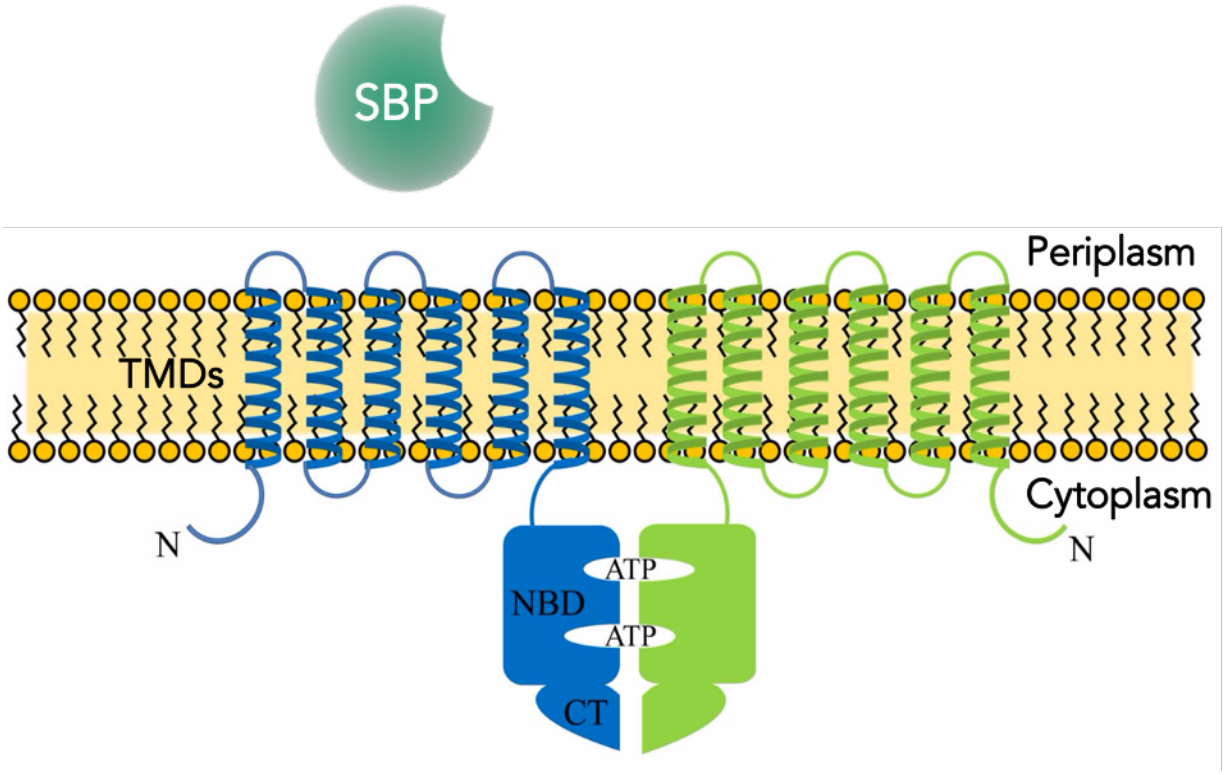


Figure 1. 10 General architecture of ABC transporters
The substrate binding proteins (SBPs) are water soluble and found in the periplasmic space. The Transmembrane Domains (TMDs, green and blue inside the membrane) are normally formed by 6 α-helices each which span through the cytoplasmic membrane of the cell. The Nucleotide Binding Domains (NBDs, green and blue inside the cytoplasm). Figure adapted from Chuang et al., 2014.

The NBDs or, ATP-binding domains, bind and hydrolyse ATP thereby fuelling transport. In the majority of these systems, there are two NBDs and two different TMDs (*eg*. maltose, histidine and glucose transporters). Less often, two copies of NBDs and one TMD comprise the translocator such as in the case of the vitamin B12 (*ie*. Btu) transporter. There are examples of systems with these two components fused to each other (*eg*. siderophore and ribose transporters). The minimum number of subunits an ABC transporter can have are two (eg. glycine betaine and glutamate ABC systems). In these instances, the transporter exhibits a unique subunit formed by the fusion of SBPs to the N- or C- terminus of the TMDs, and a second subunit created by the NBD (Heide and Poolman, 2002).

1. 4. 3. 2 E. coli maltose (MalEFGK2) ABC transporter as a model system

E. coli genome contains around 65 predicted of ABC transporters which equal to 5% of its total genome (Moussatova *et al*., 2008). One of them and the most comprehensively studied is the maltose ABC transporter. It belongs to the carbohydrate uptake transporter family (CUT) (Saier Jr, 2000a) and is able to bind both maltose (*ie*. maltobiose) and oligosaccharides of thereof (*ie*. maltodextrins). The maltose uptake system from *E. coli* (ie. MalEFGK2) is comprised of an SBP found in the periplasm (*ie*. MalE or maltose binding protein, MBP), two TMDs (MalF and G) and a homodimer of MalK units which constitute the NBD (Figure 1. 11).

The MalE (*shown in beige in* Figure 1. 11), like most SBPs, is produced in the cytoplasmic space and then is post-translationally exported in the periplasm via the Sec-translocon which recognises an N-terminal short peptide, referred to as signal peptide. During translocation of the MalE across the cytoplasmic membrane the signal peptide is cleaved off to produce the mature SBP. The MalE has been shown to bind both maltose and maltodextrin with high affinity (translates to low dissociation constant, *ie*. K_D = 1 μM). The high affinity exhibited by the SBP makes this ABC transporter a scavenging system; a property which is useful in chemotaxis. Structures of the MalE exist in both the bound (Sharf *et al*., 1992) and unbound form (Spurlino *et al*., 1991). Its structure

revealed it features the typical morphology featured of Class I carbohydrate SBPs, *ie*. two globular domains linked by a flexible hinge region consisted of three peptide stretches. The cleft formed in the middle of the two domains constitutes the binding cavity which is normally exposed to the solvent prior to ligand binding. The binding of the ligand causes drastic conformational changes in the MalE, thereby excluding the solvent in the cleft. Binding is stabilised by a number of van der Waals and hydrogen bond interactions with a contribution of residues from both domains (Sharf *et al*., 1992).

The TMD component of the system is made of two integral membrane proteins, *ie*. MalF and MalG (*shown in light blue in* Figure 1. 11). MalF is comprised of 8 transmembrane α-helices and MalG is made out of 6. Together they form the transmembrane channel which allows translocation of maltose from the SBP through to the NBD and into the cytoplasm (Jasco *et al*., 2009). MalF exhibits a peculiar feature not commonly found in the TMDs; it has an extra loop (*ie*. MalF P2) which extends in the periplasm and reaches behind the MalE as revealed by the protein crystal structure (Figure 1. 11). The interaction between P2 loop and MalE is thought to be strong and stable as is maintained when the P2 loop is expressed without the rest of MalF structure (Jasco *et al*., 2009). Comparison of the MalG sequence to other ABC transporter TMDs has led to the detection of a consensus sequence motif. The motif is referred to as EAA (*ie*. EAAX3GX9IXLP, X could represent any amino acid) and encodes for a cytoplasmic loop, thought to interact with the NBD and therefore indispensable for the function of the system. This was confirmed by site-specific chemical crosslinking, and by mutation of the most conserved residues of the EAA motif (Hunke *et al*., 2000, Mourez *et al*., 1997). The mutagenesis studies led to decreased or full abolishment of transport functionality. The EAA was proved to be crucial for the stabilisation and association of the ATP-binding monomer MalK on the cytoplasmic membrane of *E. coli*, as it was found localised in the cytoplasm instead (Mourez *et al*., 1997).

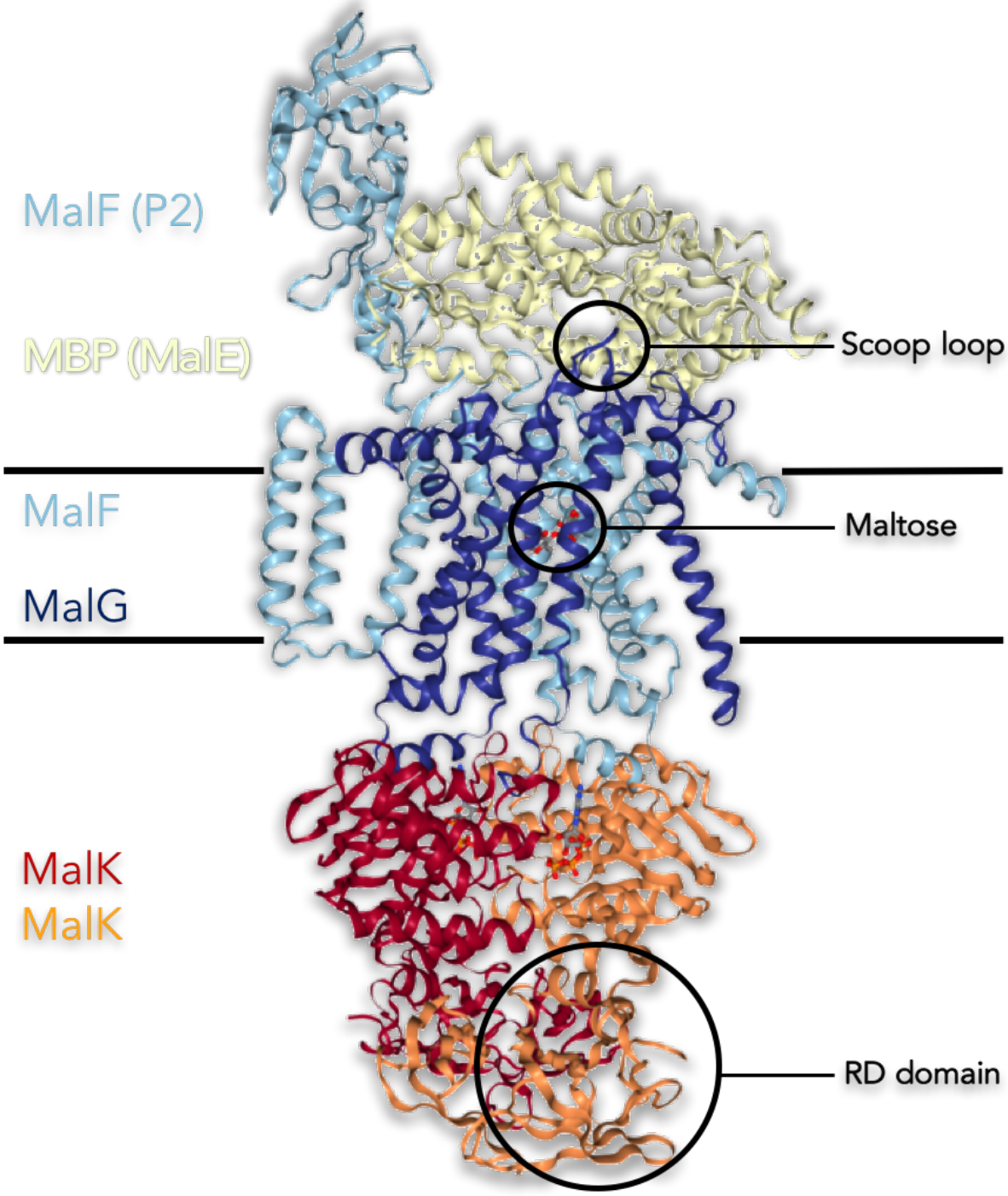


Figure 1. 11 MalEFGK2 as a model system: architecture of its crystal strucutre
Each polypeptide is coloured and labelled accordingly on the left-hand side subunits. The structure shown is the MalEFGK2 system at an intermediate state which is able to bind, but not hydrolyze ATP. In this state the NBDs (red and orange) are in the closed state with two molecules bound to them. MalE (MBP, beige) is open and maltose is bound in the translocation channel of MalF (light blue) and MalG (blue). A loop from MalG "scoop loop" is inserted into the binding pocket of MalE. PDB ID of the structure is 2R6G.

The ATPase domain is a homodimer (*ie*. MalK2, *shown in red and orange in* Figure 1. 11), and it is one of the most well-studied of all ABC transporter subunits so far. When produced independently to the rest of the system the MalK2 catalyses ATP hydrolysis spontaneously (Morbach *et al*., 1993), unlike when encoded as a full ABC unit at which case its hydrolytic activity depends on the binding of the liganded SBP MalE (Davidson *et al*., 1992). The first crystal structure of the MalK ortholog from Thermococcus litoralis was not reflecting the physiological state of the NBD due to the short-timed existence as a dimer and the unusual arrangement of the monomers (Diederichs *et al*., 2000). The structures of MalK from *E. coli* produced in isolation and as part of the complex MalEFGK2 (Chen *et al*., 2003; Lu *et al*., 2005) have been elucidated and are in agreement with its biochemical analysis (Oldham *et al*., 2007). The structure of the MalK monomer is comprised of two separate domains *ie*. an N-terminal nucleotide binding domain (NBD) and a C-terminus regulatory domain (RD) (Figure 1. 11). Its tertiary structure is highly conserved with a typical phosphate loop known as Walker A and a magnesium binding domain referred to as Walker B. They also own a switch region which harbors a histidine loop thought to coordinate the incoming molecule for ATP hydrolysis; a signature motif (LSGGQ) which is also conserved and a Q-motif (ie. conserved glutamine) positioned in between the Walker A and the signature (Biemans-Oldehinkel, Doeven and Poolman, 2005). The Walker A and B motifs are present in a RecA-like domain of the MalE, which is conserved in many ATPases (Story and Steitz, 1992; Walker *et al*., 1982). The signature motif is part of a helical subdomain which is specific to ABC transporters, thereby named as such (*ie*. 'ABC signature sequence') (Hyde *et al*., 1990). Mutational studies have attempted to assign roles to these domains. Variants of Walker A and B motifs prohibited ATP binding and hydrolysis thus abolishing transport. Q-motif has been implicated to participating in the communication between the NBD and the TMDs; although its mutation led to transport inhibition, the binding of ATP is left unaffected (Schultz, Merten and Klug, 2011). When MalK is mutated further with modifications of the conserved regions of the TMD, these reversed the inhibition of transport caused by Q-motif mutations (Hunke *et al*., 2000, Mourez *et al*., 1997). The LSGGQ motif had

similar effect to Q-motif when mutated thus also implicated in the signalling between NBD and TMD (Browne, McClendon and Bedwell, 1996).

1. 4. 3. 3 Structure of MalEFGK2

The first ever fully crystallised structure of an ABC transporter is the one of B12 transporter from *E. coli* (Locher *et al*., 2002). Since its elucidation, many more crystal structures of ABC transporter complexes have been solved. The MalEFGK2 structure was solved as an intact complex to a resolution of 2. 8 Å (Oldham, 2007) (Figure 1. 11). The transporter was immobilised to such state by mutating the MalK (E159Q). The latter allowed ATP binding to NBD but inhibited its hydrolysis.

The structure of MalK as a monomer (Chen et al., 2003) is essentially the same with it being crystallised in the homodimeric state. Physiologically, the MalK is a dimer with two ATP molecules bound. The difference being the dimeric MalK purified on its own as opposed to it being solved bound to MalEFGK2, is that the latter has the Q-motif arranged into β-strands, whereas in the first the motif is found in a disordered state. The EAA motifs of the TMDs are structured into 2 small helices which make direct contact with the MalK cleft. The MalK cleft, as already mentioned, is consisted of the helical subdomain, the Q loop and the helix after the Walker A motif.

The translocation domain (MalFG) is embedded in the membrane with the MalE SBP attached to it at its periplasmic interface (Figure 1. 11). The MalEFG exhibits a large solvent-filled cavity in its centre which spans from the periplasmic side to around in the middle of the lipid bilayer; this is large enough to accommodate the bulkiest ligand of the system, *ie*. maltoheptaose. At the bottom of the cavity there is a maltose ligand bound (*circled in* Figure 1. 11), stabilised by interacting with MalF but making no contacts with residues from MalG. Previous biochemical evidence, suggests that the MalF-maltose interaction is not a crystallisation artefact but a prerequisite for transport.

MalE is found bound to the MalFG dimer in a way that covers the opening of the translocator. The MalE is present in the unbound form thus is in the open conformation; this is a structural state thought to be stabilised by binding to MalFG which in turn allows release of the maltose from MalE into the MalFG binding pocket. One of the most structurally important features present in the MalEFGK2, is the presence of a periplasmic loop (P3) from the MalG which reaches into the binding cavity of MalE (*circled in* Figure 1. 11). This is known as the scoop loop and is postulated to assist in displacing maltose from MalE and/or prohibit its re-entry back in the cavity once dislodged thus allowing for unidirectional transport.

1. 4. 3. 4 Mechanism of action of MalEFGK2

The plethora of biochemical and genetic data have allowed the proposal of a model for the transport of maltose by MalEFGK2 (Figure 1. 12). Once maltose has made its way into the periplasm through the outer membrane channel, it binds to MalE and stabilises it. The bound form of MalE interacts with the inward-open state of the complex, in which the periplasmic interface of MalFG is shut and the MalK2 domain is open facing the cytoplasm. Suppressor mutations of the maltose-negative mutant (*ie*. malK809), which partially restored maltose transport were mapped in the periplasmic loops P2 and P1. Further data collected from cross-linking revealed that the Gly13 of unbound and bound form of MalE was in close contact to the Pro78 of MalG, proving that MalE is in permanent close cotant to the P1 loop of MalG via its N-terminus (Daus *et al*., 2007). The interaction brings about conformational changes in the TMD which in turn causes the NBDs to move closer to eachother forming an intermediate, known as the pre-translocation complex. Thereafter, the ATP binding to the MalK monomers, provokes their closure, reorientation of the TMD subunits and the opening of MalE. This shifts the state of the complex into the outward-facing conformation during which the MalE releases maltose into the translocator binding site. This triggers ATP hydrolysis; there is no structure for the complex in a post-hydrolysis state, however solved NBD isolated structures imply that ADP in their binding site is not able to maintain then into the closed MalK2 state (Lu *et al*., 2005).

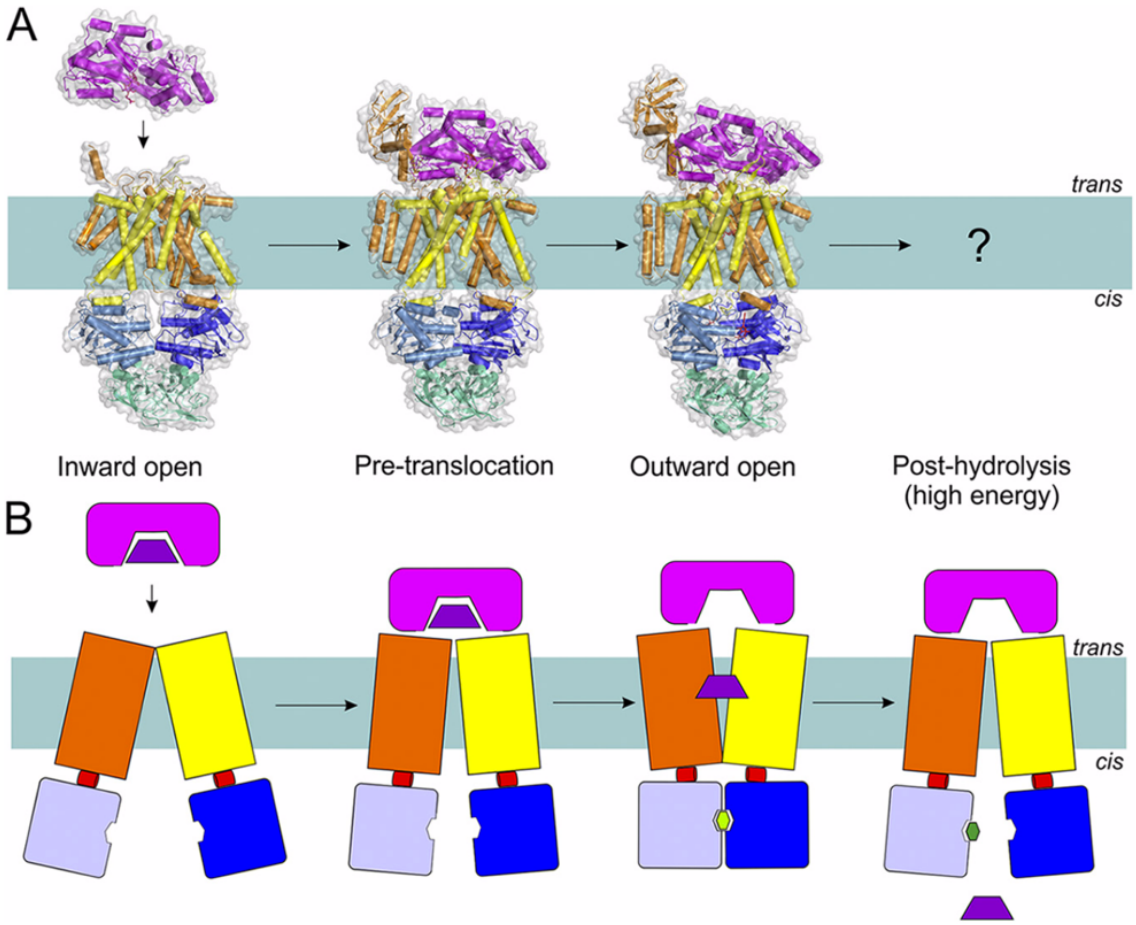


Figure 1. 12 MalEFGK2 as a model system: mode of action
The mechanism of transport of Type I importers, including sugar transport systems is exemplified here by MalEFGK2, based on the available structures (A) with the respective schematic diagrams (B) for each step. The NBDs domain (shown in blue and sky blue) are tethered to two TMDs (orange and yellow). In some transporters, additional domains are present which can act as regulators of the activity (ie. C-terminal regulatory domain - shown in green). The binding of the maltose occurs by SBP MalE (pink). The states presented here are the inward-facing (4JBW), pre-translocation (4KHZ), outward-facing (4KIo) and post-hydrolysis. The bound form of MalE docks onto the membrane embedded MalFG domains. These causes the NBDs (MalK2) to move nearer to eachother which in turn allows ATP binding. Binding of ATP causes MalK to close and reorients the TMD which adapts the outward-facing position. Subsequently the MalE opens up and releases the maltose molecule which diffuses through the membrane subunits and towards the TMD binding site. The TMD domains then hydrolyse ATP and release the maltose in the cytoplasm, resetting the structure into the inward-facing structure so it's ready to bind and translocate the next molecule. Figure obtained from Beek, Guskov and Slotboom, 2014.

Therefore, following ATP hydrolysis the TMDs are likely to reorient towards the cytoplasm causing the release of the substrate. MalE dislocates from the TMD domain and the rest of the complex enters the inward-facing resting state so the transport cycle can start again (Figure 1. 12).

1. 4. 4 Bacterial transporters and catabolic pathways involved in utilisation of xylan-derived sugars

The fermenting bacterium should preferably import the hydrolysed sugars quickly and efficiently by employing a combination of the transporters mentioned above. Following translocation of the sugars into the cytoplasm, these are shuttled into the primary catabolic pathways. The products thereof end up in the central bacterial metabolism, which allows the bacterium to obtain energy in the form of ATP and use it to reproduce (York, 2017). Very often the ecological niche of the bacterium determines its metabolic properties, similarly to its transport capabilities (Okie *et al*., 2015). Here, we identify some of the bacterial transporters reported in literature which are likely to be considered for engineering of the bacterial membrane transport during Step 2 of metabolic engineering for biofuel production (Figure 1. 3). Additionally, as the bacterial catabolic pathways (Step 3 in Figure 1. 3) have thus far been the main focus of the engineering efforts, a brief description of them for each xylan-relevant sugar is also included.

1. 4. 4. 1 Bacterial transporters and catabolic pathways for xylose

The utilization of monomeric xylose specifically in *E. coli* is mediated by the gene products of the *xyl* regulon. The transcription of this operon is controlled by the TF XylR which responds to D-xylose in the cytoplasm (Song and Park, 1997). This gene cluster is consisted of an ABC transporter encoded by the operon *xylFGH*, the catabolic enzymes *ie*. xylose isomerase and a xylokinase encoded by the operon *xylAB*. Additionally, the cluster includes the aforementioned TF encoded by *xylR*, downstream of *xylH* of the *xylFGH* operon, which is being transcribed in an

autoregulatory fashion (Song and Park, 1997). *E. coli* possesses also a low affinity proton/xylose symporter which belongs in the MFS family. This is encoded by the *xylE* (Davis and Henderson, 1987) (Figure 1. 13). Despite the scarce information available for the control of *xylE*, there is evidence that *xylE* is also regulated by XylR (Song and Park, 1997). Typical to MFS transporters, XylE binds xylose with lower affinity than its ABC counterpart, *ie*. XylF. Deletion mutants of *xylE* were used to calculate its affinity which ranges from 63 to 169 μM (Sumiya *et al*., 1995). Purified XylF was shown to import xylose with 100 to 300 times stronger affinity (Ahmed *et al*., 1982) than XylE. Import of xylose in the cytoplasm leads to activation of the XylR by direct interaction with the sugar; XylR is normally present at basal levels during growth on glucose so that sufficient number of its monomeric form are 'ready' to respond to xylose when required. The activated TF binds to promoter regions at the intergenic region between *xylFGH* and *xylAB* and activates their transcription (Song and Park, 1997). The clustering of genes encoding for transporter systems with genes for catabolic enzymes and TFs is a typical example of bacterial spatiotemporal and gene order organisation to allow for rapid induction of genes as a response to sugars or, in this case, specifically to D-xylose (Sobetzko *et al*., 2012; López-Maury, Marguerat and Bähler, 2008).

At present, only a few studies reported on the existence of dedicated ABC and MFS systems for the import of xylo-oligosaccharides released by the activity of extracellular endoxylanases (see *Section 1. 3. 1*). So far, these systems have been predominantly identified in Gram (+*ve*) bacteria. In *Streptomyces thermoviolaceus*, an ABC transporter, BxlEFG (operon doesn't include gene for an ATPase subunit), was shown to actively take up xylodextrins (Figure 1. 13). The recombinant binding protein, BxlE, presents the highest affinity for xylobiose in the lower nanomolar range (K_D = ~ 10^{-8} M) and xylotriose in the higher nanomolar range (K_D = ~ 10^{-7} M). The same study has identified BxlR as a potential repressor of the *bxlEFG* transcription is relieved in the presence of xylobiose (Tsujibo *et al*., 2004). A cognate system, known as XynEFG (Figure 1. 13), has been identified in another Gram (+*ve*) bacterium, *ie*. *Geobacillus stearothermophilus* (Shulami *et al*., 2007). The purified SBP from this system, XynE, was shown by ITC to be able to recognise and bind xylodextrins of variable length,

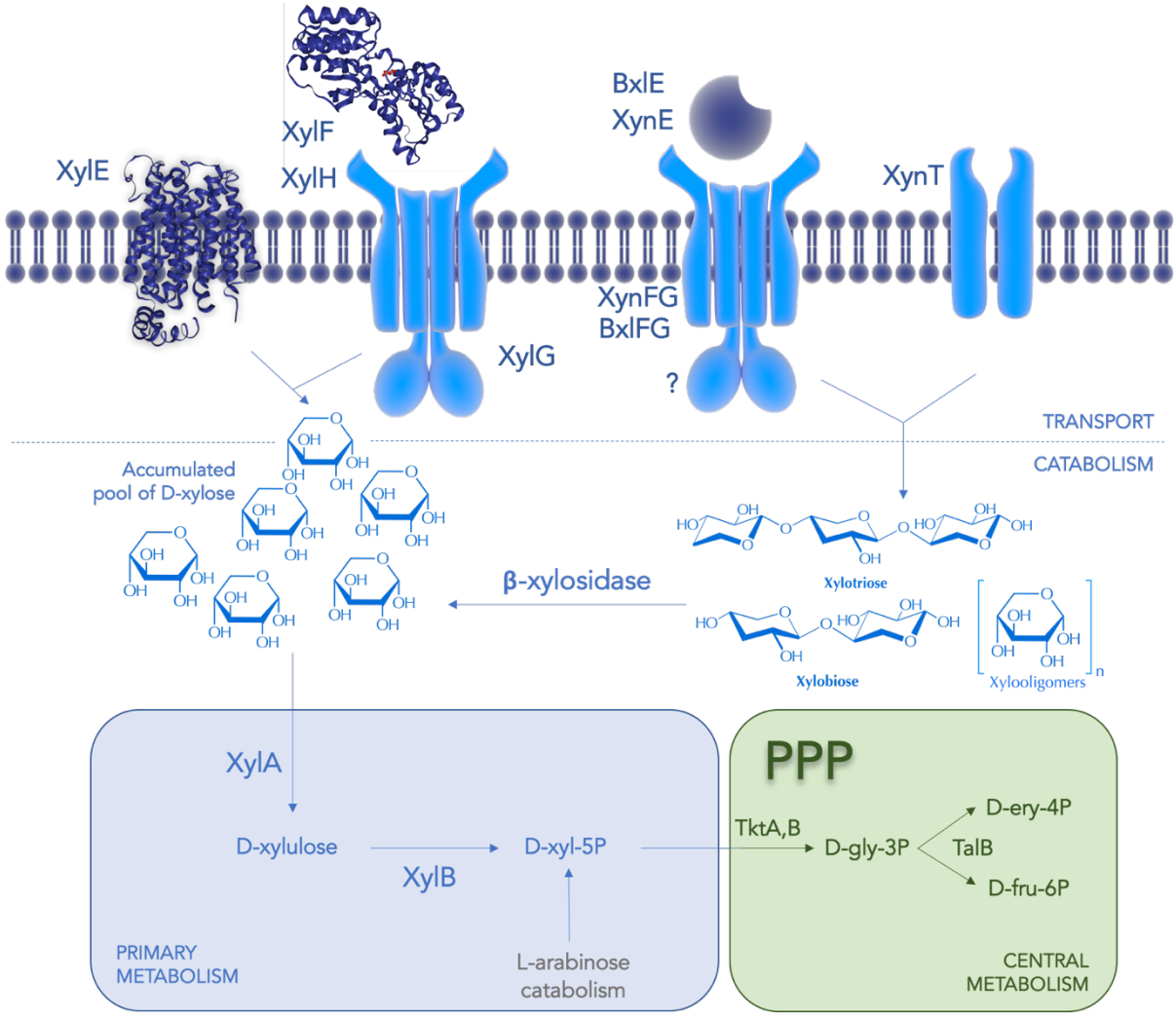


Figure 1. 13 Bacterial transport and catabolism of xylose/ xylodextrins.
Intermediate metabolites shown: D-xyl-5P = D-xylulose-5-phosphate, D-gly-3P = D-glyceraldehyde-3-phosphate, D-ery-4P = D-erythrose-4-phosphate, D-fru-6P = / Enzymes: XylA = xylose isomerase, XylB = xylulokinase, / PPP: pentose phosphate pathway, TktA,B = transketolases, TalB = transaldolase. All chemical structures were produced in the ChemDoodle. The transporter and membrane structures were obtained from the Library of Science and Medical Illustrations (www.somersault1824.com). PDB ID of protein structures used: XylE = 4JA4, XylF = 3M9X.

ranging from xylobiose to xylohexaose, and presented the highest affinity for xylotriose (K_D = 50 to 80 nM) (Shulami *et al*., 2007). Interestingly, Shulami *et al.* (1999) reported the discovery of a xylodextrin ABC transporter from *Bacillus stearothermophilus* T-6 which mediated the import of methyl-α-D-glucuronosyl-xylotriose, a degradation product of side groups of xylan (see *Section 1. 3. 3*) (Shulami *et al*., 1999). Such examples demonstrate the level of adaptability bacteria exhibit for growth on xylan.

1. 4. 4. 2 Bacterial transporters and catabolic pathways for arabinose

The uptake of monomeric arabinose has been studied extensively in the workhorse of molecular biology, *ie*. *E. coli*. The MFS transporter AraE (Griffith *et al*., 1992) can uptake arabinose with low affinity (140-320 μM) as shown by studies in membrane vesicles (Daruwalla *et al.*, 1981). Higher affinity transporter from *E. coli* is the AraFGH ABC system which binds L-arabinose with 1 μM binding affinity (Hogg, 1977) (Figure 1. 14). Both of the above systems have orthologs identified in various species including *Clostridia* (Mitchell, 2016) and *Bacilli* (Mota, Sarmento and Sa-Nogueira, 2001). A more recent discovery of an arabinose transporter in *E. coli* is the galacto/arabinofuranose ABC system denoted as YtfQRTYjfF. Horler *et al*., (2009) reported the crystal structure of SBP YtfQ bound to galactofuranose to make the first discovery of a furanose specific transporter (Figure 1. 14). This SBP binds arabinofuranose at a similar K_D with galactofuranose, and is thought to be used to supply substrates for production of UDP-galactofuranose for the cell wall in *Mycobacteria* (Horler *et al*., 2009); or for arabinofuranose uptake from arabinan and xylan by root and plant symbionts as well as pathogens. An arabinoside dedicated transport system hasn't been identified in *E. coli* yet, however such systems have been shown to exist in plant-associated bacteria. AbnEFJ, an ABC system from *Geobacillus stearothermophilus*, was shown to bind arabino-oligomers using ITC. AbnE was shown to bind xylodextrins, from xylotriose to xylooctaose, with nanomolar affinity and arabino-oligosaccharides in the micromolar range (Shulami *et al*., 2011). Further, AraNPQ and AraT are ABC transport

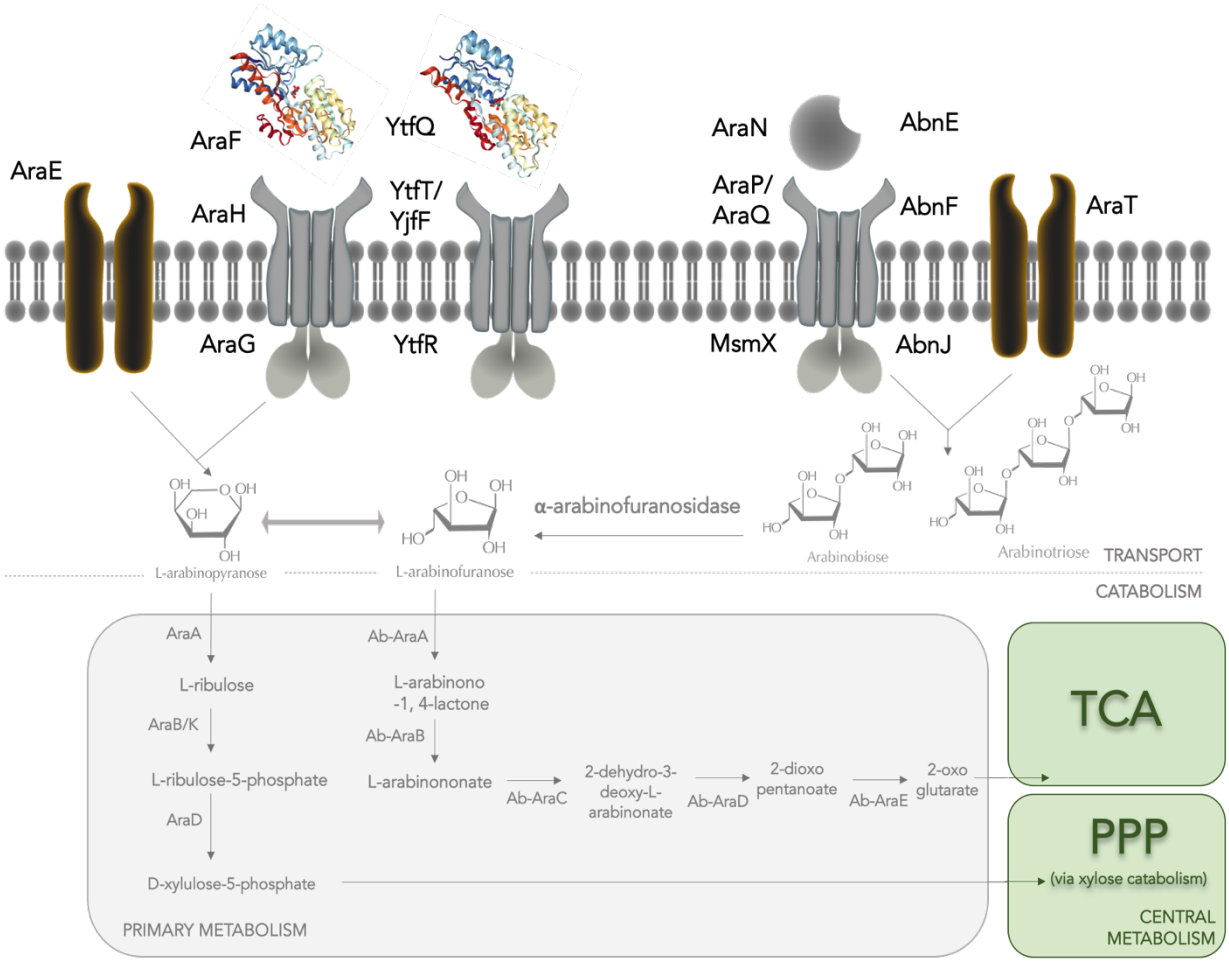


Figure 1. 14 Bacterial transport and catabolism of arabinose/ arabinosides.
Enzymes in degradation pathway I of arabinose: AraA = L-arabinose isomerase, AraB/K = Ribulokinase, AraD = L-ribulose-5-phosphate 4 epimerase/ Enzymes in degradation pathway III: Ab-AraA = L-arabinose -1-dehydrogenase, Ab-AraB = L-arabinolactonase , Ab-AraC = L-arabonate dehydratase, Ab-AraD = Deoxyarabonate dehydratase and Ab-AraE = Ketoglutaric semialdehyde dehydrogenase./ Transporter systems shown here: AraE = arabinose MFS transporter, AraFGH = arabinopyranose ABC transporter, YtfQRTYjfF = galacto-arabinofuranose ABC transporter, AraNPQ-MsmX = arabinoside ABC transporter and AraT = arabinosides MFS transporter.
TCA = tricarboxylic cycle, PPP =pentose phosphate pathway
All chemical structures were produced in the ChemDoodle. The transporter and membrane structures were obtained from the Library of Science and Medical Illustrations (www.somersault1824.com). PDB ID of protein structures used: XylE = 4JA4, XylF = 3M9X

and MFS systems from *Bacilli* and *Clostridia* predicted to bind arabino-oligomers as they are not essential for arabinose uptake, however are homologous to components of systems for transport of malto-oligosaccharides and multiple sugars (Rodionov, Mironov and Gelfand, 2001) (Figure 1. 14).

Once inside the cell, arabinose is catabolised by the AraBAD enzymes which constitute the primary catabolic pathway. An isomerase (AraA) and a kinase (AraB) convert L-arabinose to L-ribose-5-phosphate. An epimerase (AraD) converts the latter to D-xylulose-5-phosphate which is shuttled into xylose catabolism and into the pentose phosphate pathway of the central metabolism (Gross and Englesberg, 1959) (Figure 1. 14). More recently, an alternative pathway was discovered in *Azospirillum brasilense* (Watanabe *et al*., 2016) that includes the Ab-AraABCD (or AraYLCM) enzymes (Figure 1. 14). The Ab-AraA is thought to recognise the furanose form of L-arabinose which is inferred by the catalytic product of the reaction *ie*. L-arabino-1,4-lactone. Only the oxidation of L-arabinofuranose would lead to the production of the five-membered ring of L-arabino-1,4-lactone.

1. 4. 4. 3 Bacterial transporters and catabolic pathways for glucuronic acid

Glucuronic acid is imported in *E. coli* by the MFS transporter for hexuronates, namely ExuT; a member of the MFS family (MataGilsinger and Ritzenhthaler, 1983) (Figure 1. 15). Recently, TRAP systems specific to glucoronate and galacturonate have been discovered in various bacteria but not in *E. coli*. The carboxylic group of these sugar acids interact with a conserved arginine residue of the ESR binding site (Vetting *et al*., 2014). This is considered a conserved mechanism of TRAP systems in the recognition of substrates with carboxylic acid groups (Fischer *et al*., 2015).

The glucuronates, produced in the mammalian gut as a detoxification product, are imported into *E. coli* via ExuT and converted to 2-dehydro-3-deoxy-gluconate by the activity of UxaCAB, in the order mentioned (Ashwell, Wanba and Hickman, 1958). An additional two enzymatic steps catalysed by KdgK and Eda, lead into Glycolysis of the central metabolism (Figure 1. 15).

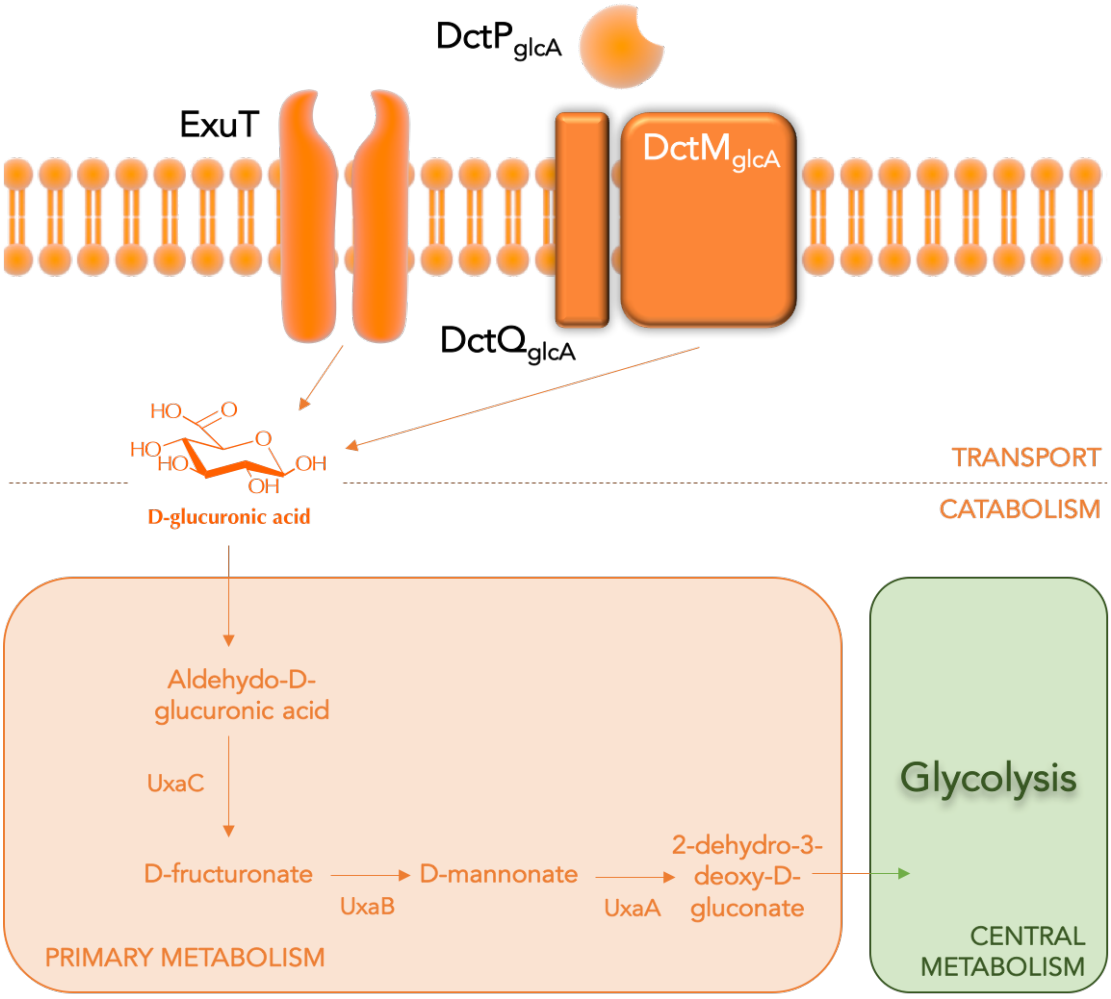


Figure 1. 15 Bacterial transport and catabolism of D-glucuronic acid.
Enzymes in degradation pathway I: UxaA = Altronate dehydrogenase, UxaB = Altronate oxidoreductase, UxaC = Uronate isomerase.
Transport: ExuT = MFS transporter for glucuronates and galacturonates, $DctPQM_{glcA}$ = TRAP transporter

1. 4. 5 Are membrane transporters important in the engineering for biofuel production?

A great amount of research for the production of biofuels and value-added products has focused on the metabolic engineering of the bacterial cell factories. Metabolic engineering can be defined as the purposeful modification of the cellular activities with the aim of strain improvement (Bailey, 1991). Progress in metabolic engineering, and synthetic biology, have allowed the engineering of bacteria to optimize the endogenous transcription regulation and activity of the enzymes that participate in limiting steps of primary and secondary cellular metabolism. The reconstruction of advanced biofuel pathways in genetically tractable heterologous hosts, such as *Escherichia coli*, has focused particularly in balancing the gene expression and the enzyme activities in order to maximize the metabolic flux. The engineering approaches applied to modify the metabolic pathways are not within the scope of this study and therefore will not be further discussed. However, in the wider context of the host engineering, metabolite transport has been a Cinderella subject mainly because the cell boundaries are normally ignored by metabolite engineers. Even though sugar transport contributes significantly to the flux towards the product, and the efflux transporters ensure the accumulation of the product in the extracellular space of the fermenters, these two cellular processes have been somewhat neglected. This is evident from published reviews which consolidate the progress made in the engineering of microbes for biofuel production. These provide current advances mainly for the stages of enzymatic hydrolysis and intracellular metabolism, and thus fail to pay attention to the intermediate stage *ie.* sugar uptake (Alper and Stephanopoulos, 2009; Wackett, 2011; Kambam and Henson, 2010; Balan, 2014; Liao *et al*., 2016).

1. 4. 6 Transport engineering efforts

Although, there have been some efforts to expand substrate utilization capability by engineering the substrate specificity of sugar transporters, no significant progress has

been made yet (Pierce, 2008). Most of the limited studies which attempted transport engineering concentrated on enabling uninhibited uptake of xylose in yeast. Xylose can enter the cells by facilitated diffusion through some yeast hexose transporters (Hamacher *et al.*, 2002; Saloheimo *et al.*, 2006). This endogenous transport is enough to allow growth on xylose, however it becomes a bottleneck when the culture is growing in low to moderate xylose concentrations as the MFS transport is inhibited by glucose (Gárdonyi *et al.*, 2003; Parachin *et al.*, 2011; Tanino *et al.*, 2012). One study has tried to shift the substrate specificity of glucose/xylose symporter, GXS1 from *Candida intermedia* and XUT3 from *Scheffersomyces stipites*, by directed evolution using random mutagenesis (Young *et al.*, 2012). The activity of the mutated transporters was tested in *Saccharomyces cerevisiae* and the study identified five enhanced transporters including one with improved growth rates on xylose by 70% and several that improved the kinetics and diauxic shift dynamics of cells when grown on glucose and xylose (Young *et al.*, 2012). Despite the accomplishments of the latter study, it didn't succeed in producing mutants without glucose competitive inhibition of xylose transport. However, in a later study from the same group the engineering of GXS1 into a xylose transporter was reported (Young *et al.*, 2013). This was achieved by targeted mutation of apparently important residues participating in the recognition of the sugar. They firstly identified a short six-residue motif (G-G/F-XXX-G) in the transmembrane span 1 (TMS1) which contained a Phe residue potentially important in the recognition of the monosaccharides. The group then used saturation mutagenesis to fully explore the effects of all possible amino acid substitutions on positions Val^{38}, Leu^{39} and Phe^{40} of the motif. One of the triple mutants they produced, showed attenuated glucose growth to the same level as the negative control with enhanced xylose uptake efficiency as compared to the wild-type GXS1 (Young *et al.*, 2013). Despite eradicating the glucose transport capacity, glucose supplied at moderate concentrations (*ie*. 5 g/L) still inhibited xylose transport. Ideally, the xylose transport shouldn't be inhibited in a lignocellulose mixture, as glucose and xylose will be hierarchically utilized and thus cause a decrease in the rate of sugar consumption by the fermenting yeast. Farwick and his coworkers addressed the issue of competitive

inhibition and achieved glucose and xylose co-fermentation (Farwick *et al.*, 2014). The group successfully engineered the galactose transporter (*ie*. Gal2) and hexose transporter (*ie*. Hxt7) from *Saccharomyces cerevisiae* so that the inhibition by glucose was abolished and the transporters were made able to bind xylose with higher affinities. They achieved this by using a growth assay to screen for mutants with glucose insensitivity (*ie*. evolutionary engineering) as well as an error-prone PCR approach followed by site-directed mutagenesis of the identified residues. One of the mutants they produced, *ie*. Gal2-N376F showed unobstructed transport of xylose by glucose, with moderate velocity (Farwick *et al.*, 2014). An additional mutant, *ie*. Hxt7-N370S, showed very high Ki value for glucose inhibition, accompanied with decreased affinity for the latter and augmented affinity for xylose as compared to the WT. In a similar study, the G-G/F-XXX-G motif of the HXT7 was identified and contradicted its position as identified by Young *et al.* (2013) (Apel *et al.*, 2016). This group evolved *S. cerevisiae* using serial dilution and plating on solid medium with 2% xylose as the sole carbon source. The evolution experiments led to the detection of a mutant which sustained significant growth on xylose. The phenotype was tracked to a single mutation in the Hxt7, *ie*. F79S, which prompt reconsideration of the position of the motif and positioned it 40 amino acids downstream the site reported by Young *et al.* (2013). The contradiction in the position of the motif as presented by Apel and her group, underlines how important it is that multiple research groups are involved in transporter engineering studies, in order to propel towards efficient biofuel production.

1. 4. 7 Importance of membrane transport engineering

The limited studies which exist, demonstrate that the molecular engineering of transport proteins represents a critical step in metabolic engineering. It is therefore, of great interest to know how important are the transporters in controlling growth rate in general. In the case of a fast-growing yeast *ie*. *Kluyveromyces marxianus*, a pHauxostat was used to select strains for even faster growth (Groeneveld *et al.*, 2008). An evolved strain was isolated which was able to grow 30% faster than the inoculated

strain. This unprecedented decrease in doubling time to 52 min, apparently produced one of the fastest eukaryotes reported. The reduced doubling time was accompanied by an approximate 40% increase in the surface area while not changing the cell volume as the shape shifted from spheroid to cylindrical (Groeneveld *et al.*, 2008). These results suggested that the processes which were most limiting to growth rate are potentially the ones occurring in the membrane processes, including substrate uptake. This assumption was tested and confirmed by quantification of the extend of control the membrane processes exert on growth; this was achieved using an equation which calculated the relative increase in maximum growth for each 1% increase in the activities of all membrane processes (Groeneveld *et al.*, 2008). Indeed, an astonishing 80% of the growth rate increase was attributed to outer membrane processes, and only 20% was ascribed to intracellular processes (Groeneveld *et al.*, 2008).

1. 5 Scope of the study

As the transport capacities of the fermenting bacteria haven't attracted much interest, this study will attempt to gain more insights on the transport of xylan-derived sugars. Most of the attention is given to the transport of the L-arabinofuranose, by trying to identify ABC transport systems specific to this sugar. The search is initiated using the YtfQ SBP from *E. coli*, a galacto/arabinofuranose protein. We will attempt to identify orthologous systems from bacteria which thrive in xylan and root mucilage- rich niches by applying an approach which combines bioinformatic, phylogenetic and synteny analysis. Any systems identified, potentially with a 'bias' towards arabinofuranose binding, will be investigated by biochemical and structural analysis to characterize any interesting features these SBPs may exhibit compared to YtfQ. Lastly, we perform gene knockouts in *E. coli* to create strains which cannot grow on xylan-derived sugars which is later used to identify a transporter that can transport xylobiose. The analysis is actively attempting to extend the engineering efforts so that they include or at least contemplate the modification of membrane systems to obtain higher biofuel yield, thus steering the CBP in different directions as is currently

exhaustively focused on the extracellular hydrolysis of the polysaccharides and their downstream intracellular catabolism.

1. 6 Thesis outline

The dissertation contains a general introduction followed by three main results chapters, and ends with a general summary and conclusions. First and second results Chapter are directly related whereas the third Chapter focuses autonomously in the investigation of a specific aim which forms part of the generic scope of the study. Therefore, these chapters will include an introduction to the concepts behind the research conducted and its aims are explained. The first Chapter describes the phylogenetic analysis and the literature review performed to identify SBPs from ABC transport systems which show high preference in binding the xylan-derived L-arabinofuranose. Also included in this chapter, are the optimisation efforts for the overproduction of the selected SBPs. The second results Chapter (*ie*. Chapter 3) includes the biochemical and structural analysis of these SBPs to define the ligand specificity and determine the molecular determinants for the binding of L-arabinofuranose. Further, the production and verification of engineered *E. coli* strains which cannot grow on xylan-derived sugars and sugar acids is presented in Chapter 4. The application of these strains in transport engineering of bacteria for biofuel production is explained and exemplified by their use in the identification of an *E. coli* secondary transporter for xylobiose. The final Chapter of the thesis is a recap of the main results of the study, with an outline of the conclusions and how these are implicated to the improvement of CBP for biofuel production.

Chapter 2

Identification and overproduction of bacterial Galacto/Arabino - furanose (Gaf) substrate - binding proteins

This chapter reports the definition of a family of substrate – binding proteins (SBPs) for Galacto/Arabinofuranose (Gaf) using phylogenetic and genomic neighbourhood analysis and attempts to discover Gaf proteins specific for L-arabinofuranose. The study uses a galacto/ arabinofuranose binding protein from *E. coli* (YtfQ) as the starting query to identify bacterial orthologs (*ie*. GafAs) for the construction of phylogenetic trees to infer relevance to arabinose uptake and attempts to discover Gaf proteins specific for L-arabinofuranose. It also considers the function of the proteins encoded by genes in the genetic proximity of the *gafABCD* operons and identifies systems linked specifically to arabinose uptake and catabolism for experimental study. Predictions of the transcription regulation of these systems are taken into account to aid selection of three candidate GafAs for experimental characterization. Different approaches to overproduce the cloned exogenous *gafA* genes using various *E. coli* overproducing strains and plasmids are described. The purification of two GafA proteins and their verification using biophysical techniques are then reported. Succeeding efforts to optimize the purity of the proteins by screening different periplasmic extraction methods are explained. The chapter concludes with the optimization of the protein yields by using different rich overproducing media to yield sufficient protein for further biochemical and biophysical analysis

2. 1 Introduction

Biofuels have the potential to occupy a high percentage of a non petrochemical-based fuel economy. The production of so-called first-generation biofuels involves utilizing sugar crops (sugarcane) or starch crops (corn, wheat). However, these are under increased scrutiny as they are considered to be responsible for a rise in food prices and the land conversion associated with their production may actually increase carbon dioxide emissions (Fargione et al., 2008). In comparison, second generation biofuels are made from lignocellulosic feedstocks derived from non-food crops (*eg*. Miscanthus, switchgrass), agricultural waste and municipal waste. The latter have a plethora of benefits over the first-generation feedstocks including the following: 1. no energy for replanting as the plants are perennial, 2. they are fast growing and they can be harvested twice a year, 3. they require low amounts of fertilizer, therefore minimising the adverse effects on the environment, 4. they can grow on marginal land, and 5. they are not used for food products. These advantages increase the likelihood that second generation biofuels will be used more extensively to meet the increasing demand in biofuels. In that case, the structure of the cell walls of both the first and second-generation feedstocks should be taken into account, since the agricultural waste are the remnants of crop harvesting. The cellulose, an important structural feature of the plant cell walls, is a polysaccharide comprised of β (1→4) linked D - glucose monomers, therefore also commonly known as glucan. Another common structural feature of these plant feedstocks is the hemicellulotic composition of their cell walls, which is principally comprised of xylans which can constitute about 25% of the total polysaccharide in the biomass (Table 2. 1). More specifically, the hemicellulotic component of grasses (*ie*. switchgrass, *Miscathus*), sugarcane and rice bran is the arabinoglucuronoxylan (AGX) polysaccharide (Vogel, 2008). The two major components of the plant feedstocks, cellulose and hemicellulose, are anticipated to become significant substrates in the future for bioconversion of the polysaccharides to ethanol or other higher molecular weight alcohols (Somerville, 2007). Therefore, the abundancy and the type of these polysaccharides in the plant cell wall plays a determining factor in choosing an appropriate feedstock. Another

defining factor to be taken into account in the choice is the extent of the lignification in the feedstock's cell wall. Lignin is the third major component and it's a phenolic polymer which associates itself with the rest of the cell wall polysaccharides through hydroxycinnamic acids (Somerville, 2007). The strong cross linking between the lignin and cellulose/ hemicellulose is well known to negatively affect the hydrolysis of monomeric sugars and therefore decreasing the rate of biodegradation. Ultimately,

Table 2. 1. The percentage composition(%) of polysaccharides in potential feedstocks

Cell wall component	Cellulose	Hemicellulose	Hemicellulose	Pectin	Pectin	Lignin
Type of the polysaccharide	Glucan	Xylan	Mannan	Galactan	Arabinan	-
Feedstock						
Corn stover	36.4	18	0.6	1	3	16.6
Wheat straw	38.2	21.2	0.3	0.5	2.5	23.4
Rice straw	34.2	24.5	-	-	-	11.9
Sugarcane	40.2	21.1	-	0.5	1.9	25.2
Switchgrass	31	20.4	0.3	0.9	2.8	17.9
Miscanthus x giganteus	47.3	25.5	-	0.6	1.8	12
MSW paper	56	8.3	5.6	-	-	30.1
Office paper	68.6	12.4	7.8	-	-	11.3
Pine	46.4	8.8	11.7	-	2.4	29.4
Cotton wood	43.2	13.3	2.2	0.6	0.6	26.8
Poplar	48.6	14.6	0.5	0.3	0.3	21.8
Hickory	45.6	19.9	0.6	-	1.2	24.2

All the data were obtained by samples analyzed at the National Renewable Energy Laboratory (Wiselogel and Johnson, 1996) apart for the *Miscanthus x giganteus* (Hodgson *et al.*, 2011).

the feedstock should therefore be delignified prior to hydrolysis. Examination of Table 2. 1, clearly shows that *Miscanthus x giganteus* is the most attractive candidate to be used as a second-generation bioenergy feedstock as it possesses a very high percentage of cellulose and xylan (47.3% and 25.5%, respectively) and a distinctly lower amount of lignin (12%) compared to its counterparts. Following careful selection of the bioenergy feedstock, one should ideally consider the type of the xylan present, as this is the most heterogeneous polymer and thus appropriate engineering of the fermenting bacteria and yeast is required for the efficient transport and catabolism of the polymer.

Despite this disadvantage, the first-generation biofuels are still not the preferred choice due to their limitations *ie*. raise the food vs. fuel debate, require high amounts of fertiliser and they make use of arable land. Nonetheless, the non-food parts of current food crops left behind following harvesting and extraction can still be utilised for biofuel production and are essentially falling within the second-generation feedstocks.

2. 1. 1 Chemical pretreatment of xylan from feedstocks for L- arabinose release

A large body of research has focused on the pretreatment of bioenergy feedstocks with the general aim of achieving high percentages of released monosaccharides and oligosaccharides to be fermented by biofuel-producing microbes. Pretreatment is a critical step to remove the lignin layer, increase porosity and to decrystallise the cellulose component so that the hydrolytic enzymes can access efficiently the polysaccharides (Dantas, Legey and Mazzone, 2013). Pretreatment with dilute acid is one of the most efficient pretreatment methods. Combined with high temperatures (*ie*. 100 – 160 °C), the xylan is hydrolysed to release sugars from the cell wall into the hydrolysate (Chen *et al*., 2007). Previous studies have reported that such pretreatment led to complete or near complete conversion of hemicellulose into soluble mono- and oligosaccharides in corn stover (Schell *et al.*, 2003), switchgrass (Li *et al.*, 2010) and *Miscanthus x giganteus* (Ji *et al.*, 2015; Li *et al.*, 2013). The higher xylan solubilisation compared with cellulose breakdown could be attributed to the fact that the xylan is more labile and amorphous than glucan, and it also exhibits a lower degree of polymerization (Himmel, Baker and Overend, 1996). Also, the acid hydrolysis under mild conditions is a suitable method to selectively release arabinose as the α-1 → 2/3 bonds, connecting arabinose moieties to the xylan backbone, are more sensitive to the effects of pH and temperature than the β-1 → 4 bonds between the xylose residues of the xylan, as the first are more accessible than the latter (González *et al.*, 1986). One study has detected release of up to 99% of the total arabinose from bagasse xylan following 30 mins in 0.2 M HCl prior to enzymatic saccharification. In similar studies,

high yields of arabinose were achieved following acidic pretreatment of wheat bran (94.9%), *Miscanthus x giganteus* (44.8%), corn hull (89%), corn stover (24.4%) (Table

Table 2. 2. The yield of arabinose released following acidic pretreatments of grasses.

	Yield* (%)	Duration (mins)	Type of acid	[Acid] (M)	Temperature (°C)	A/X in arabinoxylan**	Reference
Bagasse	99	30	HCl	0.2	130	0.073	(Kurakake et al., 2005)
Corn hull	89	30	HCl	0.2	125	0.59	(Kurakake et al., 2005)
Corn stover	24.4	15	H_2SO_4	0.076	90	0.6	(Fehér et al., 2014)
Wheat bran	94.9	120	HCl	0.5	130	-	(Aguedo et al., 2013)
Miscanthus x giganteus	44.8	15	H_2SO_4	0.1	150	0.12	(Ji *et al.*, 2015; Li *et al.* 2013)

* arabinose yield released based on total arabinose present in the cell wall of each grass.
** ratio of arabinose to xylose in arabinoxylan prior to any pretreatment.

2. 2) (Kurakake *et al.*, 2005; Aguedo *et al.*, 2013; Ji *et al.*, 2015). The percentages of arabinose released are not comparable as each of the grass and crop mentioned exhibit varying levels of arabinose in the AX/ AGX backbone. These levels are reflected by the arabinose to xylose ratio in the xylan component of each feedstock as reported in the Table 2. 2. As an example, the ratio of arabinose to xylose in the AX of bagasse is 8 times lower than the one in corn hull which is very likely to imply a lower arabinose content. This is evident by comparing the amount of arabinose released following acid hydrolysis; the same pre-treatment conditions achieved an 89% release of arabinose from the total amount present in corn hull as compared to 99% release from bagasse. The varying degrees of arabinose release shown in Table 2. 2 are attributable to the concentration and the type of the acid used, the temperature, the duration of the reaction, the preheating speed and whether a combination of chemical with physical techniques was employed or not. These are all parameters which are controlled to minimize the production of growth inhibitors (5-hydroxylmethylfurfural, furfural), which would decrease the total sugar yield. No matter what the exact conditions of the acidic pre-treatment, it is evident that a substantial amount of L-arabinose is released in the mixture which is then available for uptake and catabolism by fermenting bacteria and yeast.

2. 1. 2 Enzymatic hydrolysis of xylan from feedstocks for L- arabinose release

Additionally, to the chemical pretreatments described above, the *Section 1. 3. 2* in the General Introduction explains in detail the role of arabinofuranosidases and provides more information about the research performed to identify these enzymes. Their hydrolytic action to release L-arabinofuranose branches from the xylan polysaccharide can be exploited following the chemical pretreatment and increase the titer of these sugar in the fermenting media. Pretreatment of biomass described above, is necessary before enzymatic hydrolysis to allow for easy and quick accessibility to convert xylan to xylooligosaccharides and arabinose (Timung *et al*., 2016). The pretreatment leads to partial delignification, lowers crystallinity extend of cellulose and increases the volume of the pores in the surface of the feedstock therefore improving the enzymatic digestibility of the biomass. The yields are much improved in a process where acidic or alkaline pretreatment precedes enzymatic hydrolysis (Maurya *et al*., 2015). The extensive study of enzymatic hydrolysis after pretretment underlines the importance of this sugar in the production of biofuels from lignocellulose-based feedstocks and pinpoints the requirement of appropriate transport systems for the furanose form of the sugars.

2. 1. 3 Bacterial transport of the furanose form of sugars

The bacterial transport of the released arabinose has attracted much less interest than the strategies employed for its release, such as the dilute acid pre-treatment and the enzymatic hydrolysis mentioned above. While engineering of bacteria and yeast to utilise the xylan-derived xylose more efficiently has had some success (Xiao et al., 2011), L- arabinose has a unique property that has not been considered in relation to rational engineering. This is the fact that arabinose, uniquely for a cell wall polysaccharide component, is found exclusively in the furanose form in arabinoxylans (α-L-Araf) and not the pyranose form seen for other sugars (Figure 2. 1). This

important chemical difference has been given almost no consideration in strategies for improving pentose use by bacteria used in the manufacture of biofuels. Hexose and pentose sugars in solution form ring structures that are usually formed by 6-membered pyranose rings. For glucose the pyranose form accounts for over 99.9% of the molecule, which perhaps partially explains why this phenomenon had not been widely considered, but for other related hexoses (*eg*. D-galactose) a small proportion, around 8%, is found in the 5- membered furanose form. The biological implications of this significant furanose component amongst an excess of pyranose forms attracted, until very recently, little apparent interest. It is now clear that, as might have been predicted, bacteria have evolved to utilize even small furanose components of sugars through furanose-specific transporters (Horler *et al.*, 2009; Bagaria *et al.*, 2011). For arabinose an even greater fraction (over 12%) is in the furanose form (at 25 °C in water

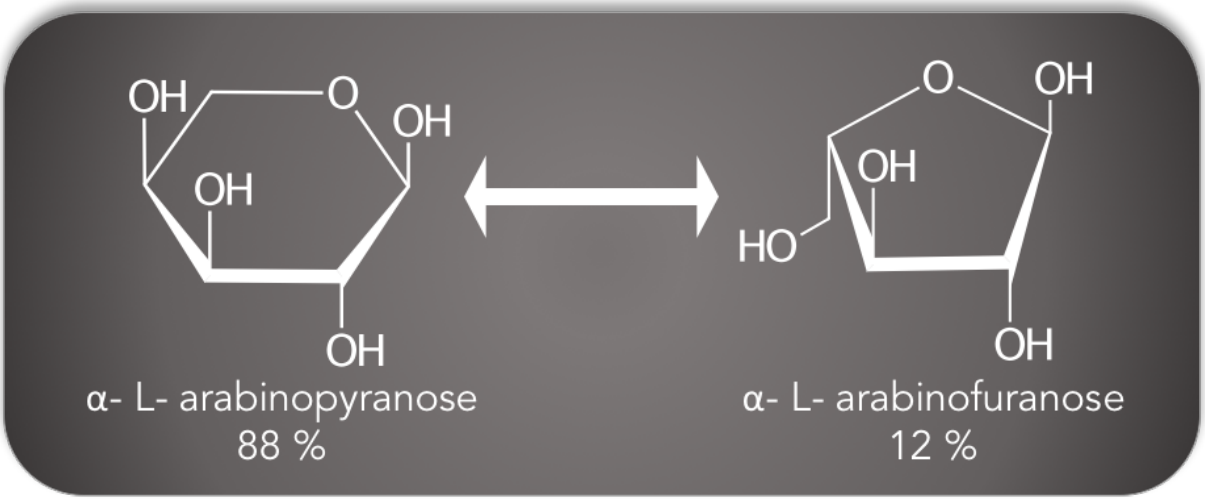


Figure 2. 1. **Equilibrium composition of the L - arabinose isoforms in aqueous solution**
The percentage of the six membered ring of L-arabinose *ie*. pyranose form and its five membered ring isoform *ie*. furanose, at 25 ºC in water.
The chemical structures were produced in ChemDoodle.

57% is α-pyranose, 30.5% β- pyranose, 8% α-furanose and 4.5% β-furanose) (H. Conner and Anderson, 1972). The existence of a specific arabinofuranose transporter in biology has not previously been considered but given that all arabinose found in

plant biomass is in the furanose form it would seem an attractive idea for bacteria that live exclusively on this substrate to have developed a specific transporter.

The study of Horler *et al.* (2009) discovered the first example of furanose – specific transport systems and showed that these exist in bacteria. The study set out to elucidate the function and structure of uncharacterized ATP-binding cassette (ABC) transporters in the model bacterium *Escherichia coli*. One of them is encoded by the *ytfQRTytfF* operon and was predicted to be a sugar transporter and the operon was known to be regulated by the galactose-responsive transcription factors GalR and GalS, suggesting that it might be a galactose-specific transporter. Using intrinsic tryptophan fluorescence spectroscopy the authors screened a range of ligands for binding to the purified recombinant YtfQ protein, the substrate binding protein (SBP) component, and demonstrated that it binds D-galactose in the low μM concentration range. Surprisingly, this binding was an order of magnitude weaker than that of a previously characterized ABC transporter for galactose, the Mgl system (Horler *et al.*, 2009). This mystery was resolved when the crystal structure of YtfQ was determined to 1.2 Å and found that it resembled other sugar binding SBPs in structure, except that the bound ligand was the furanose form of galactose and not the pyranose form. Additionally, the study used HSQC NMR methods to investigate the species bound to the protein and acquire a spectra that distinguished the pyranose and the furanose forms. Only immediately after ligand release is an excess of the furanose form seen; when remeasured over a day later the released D-galactose had reached a resting equilibrium that is almost exclusively the pyranose form.

Finally, the crystal structure of YtfQ also suggests a key binding site adaptation in YtfQ where the movement of an Asp residue out of the binding pocket permits binding of the extended furanose form (Horler *et al.*, 2009). Recalculation of the affinity based on the 7.7% of galactose that is in furanose form in solution, resulted in a K_D 0.13 μM, which is very similar to the K_D of MglB for D- galactopyranose (0.14 – 0.48 μM) (Miller *et al.*, 1980). Hence, *E. coli* expresses concomitantly two high-affinity transporters for galactose that recognise the two forms of the sugar. Interestingly, *E. coli* has recently

also been demonstrated to have both types of transporters for ribose, suggesting that recognition of both furan and pyran forms is not unique to galactose (Bagaria *et al.*, 2011). In the context of L-arabinose transport, YtfQ was also shown to bind the L-arabinose at the low μM concentration range similarly to galactose. It is predicted to bind the furanose form of L-arabinose as *E. coli* has a well structurally and characterized ABC system dedicated to L-arabinopyranose transport, *ie*. AraFGH (Quiocho *et al.*, 1974; Quiocho and Vyas, 1984). Therefore, in a similar manner to galactose, two transporters are likely responsible for scavenging the full content of arabinose found in the extracellular habitat of *E. coli*. This discovery has direct biotechnological implications as such transport systems are of vital importance in optimization of the transport of xylan-released arabinose by the fermenting bacteria and achieving higher growth rates and biofuel production. The expression of both of pyranose and furanose transport systems by the fermenting strain will allow uptake of the total pool of arabinose. This could be investigated in *E. coli* as it is more amenable in the lab so there are prospects of studying the co-expression of both systems (*ie*. *ytfQRTyjfF* and *araFGH*), since there is no evidence to prove that *gafABCD*Ec operon is induced in the presence of arabinose (Liu *et al*., 2005; Zheng *et al*., 2004).

2. 1. 4 Study aims

Given that there is experimental data which demonstrates that furanose transporters have evolved in nature and that the YtfQ binds arabinofuranose, as well as D-galactofuranose the study set out to investigate the orthologues of this SBP from a range of bacterial phyla. The aim was to identify how diverse this group of SBPs is, and focus on bacteria that thrive in close proximity to arabinose-rich polysaccharides such as rhizobial symbionts and plant pathogens. The YtfQ orthologues from such bacteria are expected to exhibit different ligand specificities and affinities, therefore we aimed to find one with selectivity for L-arabinofuranose over D-galactofuranose. Following expansion of this group by identification of the orthologous systems, we suggest the name of galactoarabinofuranose (*ie*. Gaf and complete operon: *gafABCD*) and the SBPs are referred to as GafAs. Since the genomic neighborhood of the *gaf* locus in *E. coli*

has no genes related to arabinose or galactose utilization, the search was focused on analysing the genomic context of the *gaf* genes and establishing any relevance to the uptake and catabolism of these sugars. Additionally, predictions of transcription regulation of the *gaf* operons, along with reporter assays from previous studies were taken into account in the selection of three GafAs for further experimental characterisation. The collected indications for the selected candidate GafAs provided enough evidence to point towards a potentially enhanced affinity for arabinose and a greater ligand specificity than YtfQ. To study these proteins experimentally, they were overproduced into *E. coli* and purified. Their downstream analysis required high levels of yield and purity and therefore following purification, the proteins were analysed by FT-ICR-MS and also overproduced in a range of rich media.

2. 2 Materials and Methods

2. 2. 1 Computational analysis

2. 2. 1. 1 Phylogram construction and synteny analysis

Identification of the orthologs and gene neighbourhood analysis were obtained from the Basic Local Alignment Search Tool (BlastP), the BioCyc database collection (https://biocyc.org) and the SEED viewer (http://core.theseed.org/FIG/seedviewer.cgi) (Altschul, 1990; Keseler *et al.*, 2012; Overbeek, 2005). Multiple sequence alignments and the homology were performed using the online sequence analysis software MAFFT (Katoh, 2002) using the L-iNS-I refinement method or Clustal Omega (Sievers *et al.*, 2014). The sequence alignments were curated in the Gblocks 0.91b tool of the phylogeny.fr (Dereeper *et al.*, 2008). The alignnments were inputted in the PhyML 3.0 for automatic model selection using the Akaike Information criterion (AIC). The AIC estimates the relative quality of statistical models and chooses one based on the quantity of information lost when a given method is used to represent the process which generated the data. The substitution model parameters as calculated by the AIC are described in the Table 2. 3 and their use in the phylotrees constructed is designated. The construction of the phylograms was performed in the PhyML 3.0 using the maximum – likelihood method (Guindon *et al.*, 2005) based on the substitution model and the parameters calculated by the AIC. The confidence of the branches was inferred using 500 bootstrap replications. The produced newick files were inputted into Interactive Tree Of Life (iTOL) to display and annotate the trees (http://itol.embl.de) (Letunic *et al.*, 2006). Further annotations were made in the scalable vector graphics editor, Boxy SVG. Functional annotations of the genes involved in the arabinose, galactose and xylose metabolism and related pathways were derived from the SEED comparative genomic database, PATRIC 3. 4. 2 (https://www.patricbrc.org), Biocyc and the Microbial Genomic context Viewer (MGcV, http://mgcv.cmbi.ru.nl) (Keseler *et al.*, 2012; Wattam *et al.*, 2016; Overbeek, 2005; Overmars *et al.*, 2013).

2. 2. 1. 2 Regulon predictions and fitness browser

The predictions for the transcriptional regulation of the *gaf* operons were obtained from RegPrecise online database (http://regprecise.lbl.gov/RegPrecise/) (Novichkov *et al.*, 2013). The Fitness browser (http://fit.genomics.lbl.gov/cgi-bin/myFrontPage.cgi), a depository online database of genome – wide fitness experiments for bacteria, was used to detect the effects on growth phenotypes of bacteria when *gafA* is deleted (Wetmore *et al.*, 2015; Rubin *et al.*, 2015).

Table 2. 3. Substitution model parameters for maximum - likelihood phylotree construction as calculated by the AIC.

Parameter	Description	Set at	Used in
LG	Substitution model for amino – acid replacement matrices	-	Extended phylotree & Phylogenomic tree
G	Gamma type of the substitution rate categories (*ie.* the likelihood of the phylogeny at one site is averaged over 8 conditional likelihoods corresponding to 8 rates)	8	Extended phylotree & Phylogenomic tree
I	Proportion of invariable sites	Estimated	Extended phylotree only
F	Equilibrium amino acid frequencies	Empirical	Extended phylotree & Phylogenomic tree
Bootstrap	Assessment of confidence levels for each branch support	500	Extended phylotree & Phylogenomic tree

2. 2. 1. 3 Online tools for examination of the gafAs coding sequence

The gene sequences of the *gafASw*, *gafASm*, *gafACb* and *araFCb* were codon optimized for overexpression in the *E. coli* using the online Jcat tool (http://www.jcat.de) (Grote *et al.*, 2005). The relative adaptiveness values, W, for each codon in the coding sequences, prior- and post- optimisation, were measured in the Graphical Codon Usage Analyser (GCUA, http://gcua.schoedl.de) (Fuhrmann *et. al.*, 2004). These were plotted in GraphPad Prism 7.0 and presented in the Figure A2. 1 of the Appendix. The figure includes the codon adaptiveness index (CAI) for each coding sequence and shows the increase in favourability for expression in *E. coli*. The CAI values were calculated in the GenScript Rare Codon Analysis Tool

(https://www.genscript.com/tools/rare-codon-analysis) and their reproducibility was tested by recalculation in the CAIcal server (http://genomes.urv.es/CAIcal/) (Puigbò *et al.*, 2008).

The nucleotide sequences had their identified signal peptides removed, as predicted by the SignalP 4.1 Server for all of the chosen *gafAs* (Figure A2. 2). The above signal peptide predictions were enhanced by predictions of the lipopeptides for the genes of the gram – positive *Clostridium beijerinckii* NCIMB 8052. The analysis happened in PRED – LIPO (http://biophysics.biol.uoa.gr/PRED-LIPO/) (Bagos *et al.*, 2008) and LipoP 1. 0 Server (http://www.cbs.dtu.dk/services/LipoP/) (Juncker *et al.*, 2003) (Figure A2. 2A&B).

2. 2. 2 Strains, media and Reagents

2. 2. 2. 1 Lysogeny broth and agar

The lysogeny broth (LB) was made in distilled water using 10 g/ L tryptone, 5 g/ L yeast extract and 10 g/ L NaCl. LB agar was made in smaller volumes of 150 ml using 1. 5 g tryptone, 0. 75 g yeast extract, 1. 5 g NaCl and 2. 25 g of 1.5 % (w/w) agar No.1 bacteriological. The pH of the liquid agar was balanced at 7. 5 prior to autoclaving. All the reagents were sourced from OXOID of Thermo Scientific Ltd.

2. 2. 2. 2 Overproduction media

Terrific broth (TB) was prepared in 900 ml of distilled water using 12 g/ L tryptone, 24 g/ L yeast extract and 4 ml glycerol. Following autoclaving at 121 °C for 20 mins, the volume was adjusted to 1 L by adding 100 ml of a sterilized salt mixture containing 170 mM KH_2PO_4 and 720 mM K_2HPO_4. Power broth (PB) was purchased from Molecular Dimensions Ltd. Each 100 g of the powder were dissolved in 1 L of distilled water and autoclaved as described above. The Autoinduction (AI) media powder, including the trace elements, was purchased from Formedium Ltd. Each 34.85 g of the powder were dissolved in 1 L of distilled water and autoclaved as described above. The chemical

constitution of each media is summarised in Table 2. 4 below.

Table 2. 4.Chemical constitution of the media used for the GafAs overproduction.

Type of ingredient (units)	Ingredients	Type of Growth Media		
		Lysogeny Broth (LB)	Autoinduction media (AIM)	Terrific Broth (TB)
Amino acids & vitamins (g/L)	Tryptone	10	10	12
	Yeast Extract	10	5	24
Salts (mM)	NaCl	170	-	-
	$(NH_4)_2SO_4$	-	25	-
	KH_2PO_4	-	50	17
	Na_2HPO_4 or K_2HPO_4	-	50	72
	$MgSO_4$	-	1.25	-
	Trace elements*	-	1x (*)	-
Carbon Source (mM)	Glucose	-	2.7	-
	α - Lactose	-	5.8	-
	Glycerol	-	70	54

*1000x Trace elements stock: $CaCl_2$ (20mM), $MnSO_4$ (10mM), $ZnSO_4$ (10mM), $CoCl_2$ (2mM), , $CuSO_4$ (2mM), $NiCl_2$ (2mM), NaMoO4 (2mM), Na_2SeO_3 (2mM) and $FeCl_3$ (50mM)

2. 2. 2. 3 Antibiotics and expression inducing agents

Antibiotics were added to the media or agar where appropriate; ampicillin was used at 100 µg/ml, from a 1000-fold stock which was sterilized with passage through a 0.22 µm filter and stored at 4 °C. Isopropyl-beta-D-thiogalactopyranoside (IPTG) was weighed and dissolved in distilled water to 10 ml. This was then filter sterilized using a 0.22 nm filter. All aliquots were stored in sterile Universal tubes, wrapped in aluminium foil to prevent photodegredation and stored at 4 °C. Large scale protein expression cultures needed larger amounts of IPTG and thus 1 M IPTG was prepared for that as described above. IPTG was added to cells such that its final concentration was 0.5 - 1 mM. Ampicillin and IPTG was sourced from Melford Laboratories Ltd. Arabinose induction of the pBADcLIC vector was achieved by adding appropriate amounts of a 1000-fold stock to the culture, at a final concentration of 20 mM. Arabinose was sourced from Sigma – Aldrich and prepared similarly to ampicillin and IPTG.

2. 2. 2. 4 Strains and plasmids

The genotypes of bacterial strains that were used for the work of this chapter are described in Table 2. 5. The plasmids used and constructed during this work are listed in Table 2. 6.

***Table 2. 5.* Strains used in the work of this chapter.**

Strains	Genotype	Source
DH5α	K-12 F ϕ80d*lacZ*ΔM15 *recA1 endA1 gyrA26 thi-1 supE44 relA1 deoR* Δ(*lacZYA-argF*)*U169*	Invitrogen
BL21 (DE3)	F⁻ *ompT gal dcm lon hsdS*$_B$($r_B^- m_B^-$) λ(DE3 [*lacI lacUV5-T7p07 ind1 sam7 nin5*]) [*malB*$^+$]$_{K\text{-}12}$(λ^S)	Novagen
BL21 Star (DE3)pLysS	BL21 (DE3) pLysS[*T7p20 ori*$_{p15A}$](Cm^R)	Thermo Scientific
BL21 Tuner (DE3)	BL21 (DE3) *lacY1*	Novagen

***Table 2. 6.* Plasmids used in the work of this chapter**

Plasmids	Description	Resistance	Source
pET20b	Overexpression vector with N – terminal *pelB* tag for periplasmic localisation	Amp	Novagen
pBADnLIC 2005	pBAD vector modified with LIC cassette	Amp	Dr. Eric Geertsma
pMALp5x	Overexpression vector for C – terminal fusions to *malE*	Amp	New England Biolabs
pCD01	*gafACb* in pET20b with N– terminal hexahistidine	Amp	This study
pCD02	*araFCb* in pET20b with N– terminal hexahistidine	Amp	This study
pCD03	*gafACb* in pBADnLIC2005 with N– terminal hexahistidine	Amp	This study
pCD04	*gafACb* in pMALp5x fused to *malE* (at C-terminal)	Amp	This study
pCD05	*gafASm* in pET20b with N– terminal hexahistidine	Amp	This study
pCD06	*gafASw* in pET20b with N – terminal hexahistidine	Amp	This study

2. 2. 3 Molecular Biology techniques

2. 2. 3. 1 Preparation of plasmid DNA

Plasmid DNA was isolated from 10 – 20 ml of overnight grown bacterial cultures using the Wizard Plus SV miniprep kit from Promega that employs the alkaline lysis method for plasmid recovery. The manufacturers recommended protocol was used.

2. 2. 3. 3 Agarose gel electrophoresis

The DNA fragments were resolved by electrophoresis using 1% (w/v) agarose gel in TBE buffer. TBE buffer was made up with 1.62 g/l Tris (Invitrogen), 2.75 g/l boric acid (Fisher Scientific) and 0.95 g/l EDTA (Fisher Scientific). The 1 % agarose gels were prepared by pouring molten agarose directly into a horizontal electrophoresis tank, including a comb to create appropriately sized wells and then allowing the gel to set. Ethidium bromide (EtBr) from Invitrogen, was used to stain the DNA and was added at 7 µl per 100 ml of molten agarose before pouring into the tank. Following drying, the comb and the tape were removed and the gel was submerged in TBE buffer. Normally, 5 to 10 µl of sample was loaded in each well; this was prepared by mixing the DNA with 1- 2 µl of 6x stock of sample loading buffer (NEB). Each gel contained at least one DNA ladder (Bioline Hyperladder or NEB) to act as a marker. DNA separation was performed by applying a potential difference of 80 Volts for 50 minutes. EtBr stained DNA was visualised on a transilluminator (Syngene Imaging System).

2. 2. 3. 4 Table of primers

The primers used in this work are shown in Table 2. 7. All primers were sourced from either Integrated DNA Technologies UK Ltd (IDT) or Sigma.

Table 2. 7. Primers used in the work of this chapter

Name of Primer	Sequence of primer (5'→3')
gafASmF	CATGCCATGGCCGAACTCGTCGTCGGCTTT
gafASmR	CCGCTCGAGGTAGCCGAGGCCTTTCTTTTCTTCG
gafACb-pBADF	ATGGGTGGTGGATTTGCTATGAAATACCTGCTGCCGACCGCT
gafACb-pBADR	TTGGAAGTATAAATTTTCGTATTTACGGTTCGGCAGTTCAGCAGCAGC
gafACb-pMALF	TCTAAAAAAAATCATCGGTTTC
gafACb-pMALR	GAACTGCCGAACCGTAAATAC

2. 2. 3. 5 Polymerase Chain Reaction (PCR)

PCR was used for the amplification of genes for cloning and screening. High fidelity

DNA polymerase (Phusion from NEB) was used for cloning. The reaction setup for a typical Phusion PCR reaction consisted of Phusion HF or GC buffer (5x), 200 µM of each dNTP, 0.5 µM of each forward and reverse oligonucleotide primers, 1 unit Phusion polymerase and either genomic or plasmid DNA template (< 250 ng). The standard reaction volume was 50 µl and the final volume was made up with sterile distilled deionised water. The whole process was performed on ice in 0.5 ml or 0.2 ml PCR tubes. Following a 2 minute of initial denaturation step (95 °C), the target DNA was amplified by 35 cycles of 98 °C for 10 secs, 55 - 65 °C for 30 secs and 72 °C for 30 secs per kilobase of target amplicon. The run finished with a final extension step of 10 minutes at 72 °C, and the PCR product was removed from the thermal cycler (Techne TC-312 or T100 Bio-Rad Thermocyclers) and kept in the fridge prior to running on the gel or stored in the – 20 °C.

Successful cloning of genes was verified by PCR screening using the GoTaq polymerase from Promega. The GoTaq Green Master mix used, included Taq DNA polymerase, 2X Green GoTaq reaction buffer, 400 µM of each dNTP and 3 mM $MgCl_2$. The reaction set up for a standard GoTaq 50 µl reaction included mixing 25 µl of the master mix, with 2. 5 µl of each primer (10 M) and 1 – 5 µl of genomic or plasmid DNA template (< 250 ng). The reaction volume was topped up to 50 µl using nuclease - free water. Following preparation of the samples on ice, these were transferred in the thermal cycler to undergo PCR set as follows: initial denaturation of 2 mins at 95 °C, 35 cycles of 95 °C for 45 secs, 55 – 65 °C for 45 secs and 72 °C for 1 kilobase of the target amplicon. Following the final extension step at 72 °C for 5 mins, the PCR product was resolved on DNA agarose gel.

2. 2. 3. 6 PCR clean up

The PCR clean up ensured isolation of the amplicon from the rest of the PCR reaction compoments, in case the PCR product was intended to be used in a subsequent cloning reaction. The Monarch PCR clean up kit from NEB was used and the protocol provided with it was followed.

2. 2. 3. 7 Gel extraction

DNA components were purified by running on a 0.7 % agarose gel electrophoresis and extracted from the gel using a Promega Wizard SV gel and PCR Clean- Up system kit, before switching to one provided by NEB (Monarch). The gel extraction happened following the instructions on the protocols provided with each kit.

2. 2. 3. 9 Conventional Ligation

The DNA sequence of the *gafASm* was PCR amplified from the genomic DNA of *Sinorhizobium meliloti* 1021 using the primers *gafASm*F and *gafASm*R (Table 2. 6). The PCR product was cloned into the overproducing vector using restriction enzymes from NEB which created overhangs. The pET20b vector and the PCR amplicon were digested in separate reactions of double digestion (*ie*. NcoI and XhoI) by mixing 1 unit of each enzyme with 10X NEB buffer and 1000 ng of DNA. The final volume of the reaction was up to 50 µl using nuclease-free water. The same procedure was followed for the subcloning of *gafACb* into pMALp5X. The *gafACb* was PCR out the pET20b vector with a 5' blunt end and an SbfI site introduced at the 3' end. The pMALp5X vector and the produced amplicon were digested separately in double digest reactions with XmnI and SbfI.

Following PCR clean up of the digestions, the restricted insert and vector were mixed in a ligation reaction to allow cloning. The molar ratio of the two fragments in the ligation reaction was 1:3 vector to insert. T4 DNA ligase, T4 DNA ligase buffer from NEB and nuclease – free water were added in the ligation mixture to a final volume of 20 µl. The approach taken was incubation of the ligation mixture at 16 °C, kept in the thermocycler overnight. For the semi – blunt end ligations into pMALp5X vector, the ligation mixture was incubated for an extra four hours at 4 °C. The ligated plasmids were kept in – 20 °C prior to host transformation.

2. 2. 3. 10 Gibson assembly

The *gafASw*, *gafACb* and *araFCb* sequences were further modified by the addition of 50 bps of upstream and downstream homology to the overproducing vector. These

were then synthetically produced by the GeneArt services of Invitrogen or IDT gene block services. The Geneart gene strings were resuspended in distilled water to reach final concentration of 100 ng/µl and the IDT blocks to a final concentration of 10 ng/µl. These were then inserted into pET20b vector by cloning using the Gibson Assembly Master Mix kit from NEB. The vector was digested with EcoRI to allow linearization and PCR cleaned up (Section 2. 2. 3). The reaction mixture included 70 ng of the EcoRI restricted pET20b with 3x molar excess of the Geneart gene strings calculated using the formula:

$$pmols = (weight\ in\ ng)\ x\ 1{,}000\ /\ (base\ pairs\ x\ 650\ daltons)$$

Nuclease – free water was added to a final volume of 20 µl in each reaction tube. The samples were incubated in a thermocycler at 50°C for 15 to 30 minutes (suggested in the assembly of 2 or 3 fragments). Following incubation, the samples were kept on ice or stored at –20°C for subsequent transformation.

2. 2. 3. 11 LIC independent cloning (LIC)

Ligation independent cloning was used to ligate PCR products of the *gafACb* with pBADcLIC2005. pBADcLIC2005 is a modified version of the commercially available pBAD vectors with the inclusion of a LIC cassette at the site of cloning (modified by Dr. Eric Geertsma).

Primers were designed such that a dedicated LIC tail was added to the PCR product, which was designed to be complementary to the LIC cassette in vectors. The gene of interest was amplified using PCR and purified using gel extraction. The vectors were linearised by digesting with the SwaI. The linearised vectors were purified by gel extraction. Both the insert and vector DNA were treated with T4 DNA polymerase which was supplied from Novagen. The 3' → 5' exonuclease activity of T4 DNA polymerase produces long, complementary overhangs on the vector and the insert. To do this, 200 ng of linearised vector was made up to 10 µl with sterile water and treated with T4 DNA polymerase in the presence of dCTP for pBAD vectors (1.5 µl of 25 mM

dCTP, 3 µl 5x buffer and 0.5 µl T4 DNA polymerase). The reaction was incubated at room temperature for 30 minutes, after which T4 DNA polymerase was deactivated by incubation at 75 °C for 20 minutes. A volume of insert DNA corresponding to a 1:1 and molar ratio with vector was treated with T4 DNA polymerase but this time in the presence of dGTP. Both the LIC ready insert and vectors can be stored at -20 °C for several months.

The LIC-ready components were mixed in a 1:3 ratio of vector to insert and incubated at room temperature for 20 min. 1.5 µl of 25 mM EDTA was added and the reaction was incubated at room temperature for another 15 min. 2 µl of the reaction was used to transform into competent cells using the heat shock method. Transformation of pBAD based constructs was done into competent cells of DH5α strain of *E. coli*.

2. 2. 3. 12 Heat - shock transformation of bacterial cells

The plasmids were transformed in chemically competent *E. coli* cells by heat – shock. The chemically competent cells were prepared from overnight *E. coli* cultures. These were inoculated into 50 – 100 ml LB in 250 ml conical flasks at a starting OD_{600} of 0.1. The culture was grown to exponential phase (*ie.* $O.D_{600}$ = 0.4 – 0.6) in 37 °C at a rotor with constant 220 rpm shaking. All steps followed occurred on ice. After the culture reached exponential phase, it was centrifuged at 4500 rpm and 4 °C for 10 mins. The collected pellet was washed with 10 ml 0.1 M $CaCl_2$ solution. Following two additional centrifugation – washes steps, the pellet was resuspended in 1 ml of 0.1 M $CaCl_2$/ 15% glycerol mixture. The 100 µl aliquots were kept at -80 °C until transformation.

For the transformation, one 100 µl aliquot of the E. coli strain was mixed with 2-4 µl of plasmid and kept on ice for 30 mins incubation. Thereafter, the eppendorf tube containing the mixture was placed on a heat block set at 42 °C, for 2 mins to allow heat shock of the cells and uptake of the plasmid DNA. The cells were then incubated at 37 °C for 1 to 2 hours with constant shaking at 200 rpm. Following outgrowth, 100 - 150 µl of the culture was plated on selective agar media.

2. 2. 3. 13 Restriction

Restriction of plasmids was performed prior to cloning of the constructs, and following cloning for screening purposes. All restriction enzymes were supplied from New England Biolabs Inc. One unit of the restriction enzyme/s, was/were mixed with 1 µg total plasmid DNA and 5 µl of 1 x CutSmart buffer. Nuclease – free water was added in the mixture to reach a final volume of 50 µl. Where allowed, the reaction was terminated by heat inactivation at 65 °C or 80 °C for 20 mins. The restricted plasmids were resolved on a 1 % DNA agarose gel.

2. 2. 4 Expression of recombinant proteins

2. 2. 4. 1 Small scale production in E. coli (BL21 DE3 or Star or Tuner)

The expression of the cloned constructs for production in *E. coli* was IPTG dependent for all vectors used in this study, apart from pBADcLIC which was induced by arabinose. For the initial small scale trial expressions, falcon tubes containing 20 ml of cell culture in LB were grown at 37 °C, with appropriate antibiotic and shaking was set 220 rpm. Once the OD_{600} reached mid-exponential phase (0.4 – 0.6), the cells were cooled and 1 mM of IPTG was added in the culture. The growth was then resumed at 37 °C or 20 °C. To check expression, the cells were collected at appropriate timepoints and spun down (13000 rpm, 10 mins). Then the pellets were mixed with sample buffer and cells were lysed by a 7-mins incubation at 95 °C. The cells were spun down to remove insoluble cellular debris. The lysates were resolved on a 12 % SDS-PAGE gel to assess protein production.

2. 2. 4. 2 Large scale production in E. coli (BL21 DE3 or Star or Tuner)

For the large scale expression 10 ml cells were grown overnight at 37 °C, with shaking at 220 rpm and used to inoculate up to 2 L LB containing ampicillin. Cells were grown to an OD_{600} of 0.4 – 0.6 and IPTG added to a concentration of 1 mM. Once induced constructs of p150 were grown for 5 hours at 20 °C, with shaking at 220 rpm in Rosetta pLysS and p60 for various 60 times, as detailed in the results section. Once

grown cells are spun down at 5000 rpm for 20 min at 4 °C and the pellet was treated appropriately to release the periplasmic contents.

2. 2. 5 Periplasmic extraction

The collected pellet from the large-scale production was subjected to periplasmic extraction following the ICOS (Ice cold osmotic shock), STE (Sucrose – Tris – EDTA) or TSE (Tris – Sucrose – EDTA) procedure. The ICOS method included a first step of resuspending each gram of the pellet in a 4 ml of sucrose mixture which included 750 Mm sucrose, 50 mM Tris- HCl (pH = 8) and 5 mM EDTA. The resuspended pellets were incubated on ice for 15 mins and then centrifuged down at 8500 xg for 20 mins at 4 °C. The supernatant was removed (sucrose fraction) and kept on ice. Each gram of the remaining pellet was resuspended in 4-5 ml of a solution containing 5 mM $MgCl_2$ and half a tablet of protease inhibitor. Following, 40 µl of 15 mg/ml of lysozyme (*ie*. 0.15 mg/ml final concentration) were added in the mixture and the sample was incubated on ice for 30 mins. The samples were spin down at 8500 xg for 20 mins and the supernatant (lysozyme – treated fraction) was kept on ice. The collected fractions were dialysed against Phosphate Buffered Saline or Tris- HCl/ NaCl buffer (pH=7.5).

The TSE method is essentially a modification of the first step of ICOS. The latter uses a sucrose mixture containing the following: 500 mM sucrose, 200 mM Tris- HCl and 1 mM EDTA. 1 ml of sucrose mixture was added in the centrifuged pellet, for each 100 ml of the collected culture with an OD_{600} equal to 1. The resuspended culture was incubated on ice for 30 mins and centrifuged down at 16000 xg for 30 mins, at 4 °C. The supernatant was dialysed overnight at the same buffers mentioned in the ICOS procedure. The pellet was discarded.

The STE method is similar to the ICOS, however the collected pellet is treated with solid lysozyme instead of liquid and the treatment happens at the same step with the sucrose at 30 °C. Each pellet from the culture was resuspended in 15 ml of STE buffer (*ie*. 500 mM sucrose, 50 mM Tris- HCl and 5 mM EDTA). 13 mg of lysozyme were

added and the sample was incubated at 30 °C for 1 – 2 hours. Following, the sample was centrifuged at 12000 rpm for 15 mins at 4 °C. The supernatant was collected and dialysed to remove EDTA prior to the purification.

2. 2. 6 Protein purification techniques

2. 2. 6. 1 Nickel-affinity chromatography using gravity flow columns

The periplasmic extract containing the protein of interest was mixed with Ni^{2+}-NTA resin (Qiagen) equilibrated with buffer A (50 mM Tris-HCl, pH 7.8, 200 mM NaCl, 20% glycerol) supplemented with 20 mM imidazole. The mixture was incubated for 1 h at room temperature in a disposable polystyrene column (Pierce). Subsequently, the resin was washed with buffer A containing 25 mM imidazole, for 20 column volumes (CV) to remove weakly binding contaminants. Bound His-tagged protein was eluted with elution buffer A containing 250 or 500 mM imidazole.

2. 2. 6. 2 Nickel-affinity chromatography using HisTrap HP columns

The periplasmic extract containing recombinant protein was filtered through a 0.2 μm filter. A 5 ml HisTrap column (GE Healthcare) was equilibrated with buffer A (50 mM Tris-HCl pH 8, 200 mM NaCl and 20 mM imidazole) and connected to an AKTA Purifier P-900. Filtered supernatant was then injected into a pre-equilibrated column. The column was washed with 7 CV of wash buffer (50 mM Tris-HCl, pH 8, 200 mm NaCl and 25 mM imidazole) to remove weakly bound contaminants. A gradient was applied over 20 CV from 10 to 500 mM imidazole to elute His-tagged bound protein. Purification was monitored by detecting the absorbance of the column eluent at 280 nm (A_{280}). Fractions were collected with an automated F-950 Fraction Collector (GE Healthcare) and were run on SDS-PAGE gel for analysis.

2. 2. 7 Protein concentration determination

2. 2. 7. 1 Bradford assay

A set of protein standards were produced by serially diluting a stock solution of bovine

serum albumin (BSA) from 200 µg/ml to 6.25 µg/ml. The protein of unknown concentration was serially diluted by at least 2 times. 50 µl of each of dilutions of BSA and unknown sample was added to 1.5 ml of Bradford reagent (Bio-Rad) in a cuvette. The absorbance was measured at 595 nm wavelength. A graph of protein concentration versus absorbance (A595) was plotted and the concentrations of the unknown samples were calculated using the standard curve.

2. 2. 7. 2 Using A280/extinction coefficient

A NanoDrop ND-1000 spectrometer was used to confirm the results from the Bradford assay by measuring the absorbance at a wavelength of 280nm. The molecular weight and the extinction coefficient of purified proteins were calculated using the peptide property calculator tool available on Northwestern University website (http://www.basic.northwestern.edu/biotools/proteincalc.html). Protein concentration was determined using the Beer-Lambert equation:

$$c = \frac{\varepsilon\, l}{A}$$

Where:

A = absorbance
ε = extinction coefficient
l = length of wavelength path
c = protein concentration

2. 2. 8 Dialysis, concentrating and storage of protein

Dialysis facilitated buffer exchange of the periplasmic extracts and the purified fractions. Dialysis was performed using 0.33 or 3.33 ml/cm dialysis tubing (sourced from Spectrum Labs) that is activated by rinsing with distilled water. Mediclips were used to seal the end of the tubing and then the sample pipetted into the dialysis tubing which was placed in a beaker containing the dialysis buffer at least 100x the volume to be dialysed. The first dialysis step lasted approximately five hours and the second dialysis step was done overnight, with all dialysis performed at 4 °C. Protein

concentration was increased using a MWCO 5000 (Da) Vivaspin columns, which were supplied from GE Healthcare. These columns use centrifugation to reduce the amount of solvent while preventing the protein from passing through the membrane. These were occasionally used for rapid buffer exchange. The protein was normally stored at –80 °C for long periods, or at 4 °C for routine use.

2. 2. 9 Fourier-transform ion cyclotron resonance mass spectrometry (FT-ICR-MS / FT-MS)

A CoEMS‟s Burker apex ultra 9.4 Tesla spectrometer was used for Fourier transform ion cyclotron resonace mass spectrometry of protein complexes. Purified protein after removal of the His-tag was dialysed into water and concentrated to 200 µM and given to Adam Dowle in the Technology Facilty, University of York, for analysis using the CoEMS‟s mass spectrometry equipment. Samples were diluted to between 10 and 100 µM to give a final concentration of 50 % (v;v) aqueous methanol, 1% (v;v) formic acid, before introduction into the mass spectrometer.FT-ICR-MS was performed using an Apex ultra Fourier Transform Ion Cyclotron Resonance mass spectrometer (Bruker Daltonics, Bremen, Germany) equipped with a 9.4 T superconducting magnet, interfaced to a Triversa NanoMate nanoESI source (Advion BioSciences, Ithaca, NY). Instrument control and data acquisition were performed using Compass 1.3 for Apex software (Bruker Daltonics). Nitrogen served as both the nebulizer gas and the dry gas for nanoESI. Observed precursors were manually selected for sustained off-resonance irradiation (SORI) fragmentation in the infinity ICR cell, with argon used as both the cooling and collision gas. Bruker data Analysis software (version 4.0) was used to perform the spectral processing and peak picking for both MS and SORI fragmentation spectra. Monoisotopic masses of proteins were calculated using the SNAP 2.0 algorithm in dataAnalysis (repetitive building block: C 4.9384, N 1.3577, O 1.4773, S 0.0417, H 7.7583).

2. 3 Results

2. 3. 1 Phylogenetic analysis shows that GafAs are widespread in bacteria.

The first step in the detection of GafAs which could be more arabinose specific involved identification of bacterial orthologues of YtfQ from *E. coli*. These were used to construct a circular phylogram which was intended to be used for explorative and informative purposes and to investigate the extend of occurrence of the orthologs in different bacteria phyla. Also, due to the abundancy of the GafAs used, which were derived from a variety of phyla, the extensive tree produced was used to detect potential horizontally transferred *gafAs*.

The initial searches for this purpose were performed in the SEED viewer and targeted orthologs from bio-industrially important bacteria and plant symbionts with representatives from various phyla and classes. These resulted in the identification of orthologs from *Clostridium beijerincki* NCIMB 8052 (Firmicutes), a saccharolytic mesophile with a great number of fermentation products including butanol and acetone (Johnson *et al.*, 1997); *Saccharophagus degradans* (γ–proteobacteria), a unique bacterium able to degrade 10 different complex polysaccharides (Ekborg, 2005), *Sinorhizobium meliloti* 1021 (α-proteobacteria), a rhizosphere bacterium considered an agriculturally important nitrogen fixer (Honeycutt, McClelland and Sobral, 1993), *Sorangium cellulosum* So ce56 (δ–proteobacteria); a prominent producer of fungicides and bactericides which thrives on cellulose and xylan polysaccharides (Huntley *et al.*, 2013) and Variovorax paradoxus S110 (β-proteobacteria), an endophytic bacterium especially important because of its use in the biodegradation of biogenic compounds and anthropogenic contaminants (Han *et al.*, 2010). The repertoire of GafA proteins was then expanded by BLASTP analysis using the protein sequence of YtfQ and the GafAs mentioned above as the starting query for each individual search. These focused on identifying orthologues with the highest identity from each bacterial phylum by restricting the search. Thereafter, the lowest scoring GafAs from α, β, γ and δ proteobacteria as well as Firmicutes, Spirochaetes and Actinobacteria, which still presented at least 45 % or higher identity to the query protein sequence were inputted

on new BlastP searches. The fresh searches intended to collect more GafAs and expand the collection by restricting the search for each bacterial orthologue to its respective phylum or class. Additionally, the analyses were also performed without restriction. The process was repeated until previous results were observed again, so that the sequence space was fully explored and a non–redundant set of GafAs was produced. To avoid creating large and impractical clades, only one bacterium from each species was included in the downstream phylogenetic analysis. The identified protein sequences, which amounted to 110, were used to produce a phylogram as described in Materials and Methods Section 2. 2. 1. 1. The produced tree is shown in Figure 2. 2.

The phylogenetic tree revealed that the orthologues are widespread within the bacterial phyla and classes tested. The γ-proteobacteria occupied the majority of the tree, however this is expected to be an artifact because of the unequal number of sequenced bacterial genomes present to-date between the different phyla/ classes. The γ-proteobacteria form three distinct clades, with the one governed by *Enterobacteriales* being larger. The phyla/ classes form unique clades which separate themselves from eachother, with only a few exceptions where some candidates cluster with distantly - related rather than cognate GafAs. Some exceptions include the GafAs from γ-proteobacteria *Cellvibrio japonicus* Ueda107 and *Marinomonas* sp. MWYL1; these are found in the α – proteobacteria which could indicate horizontal gene transfer has occurred between them. The use of non–orthologous protein sequences ensured the formation of an outgroup which clustered itself away from the orthologous GafA clades. The outgroup was consisted of SBPs of ABC transport systems from *E. coli* including the D-galactopyranose SBP, *ie*. MglB, the L- arabinopyranose ABC SBP, *ie*. AraF, the D-ribopyranose SBP, *ie*. RsbB and the predicted L-arabinopyranose SBP from *C. beijerinckii* NCIMB 8052. Within this outgroup, the most evolutionary related SBP to the GafAs is the RsbB as it shows the lowest branch length.

2. 3. 2 Further '*in silico*' analysis of gafA genes suggests functions in arabinose transport

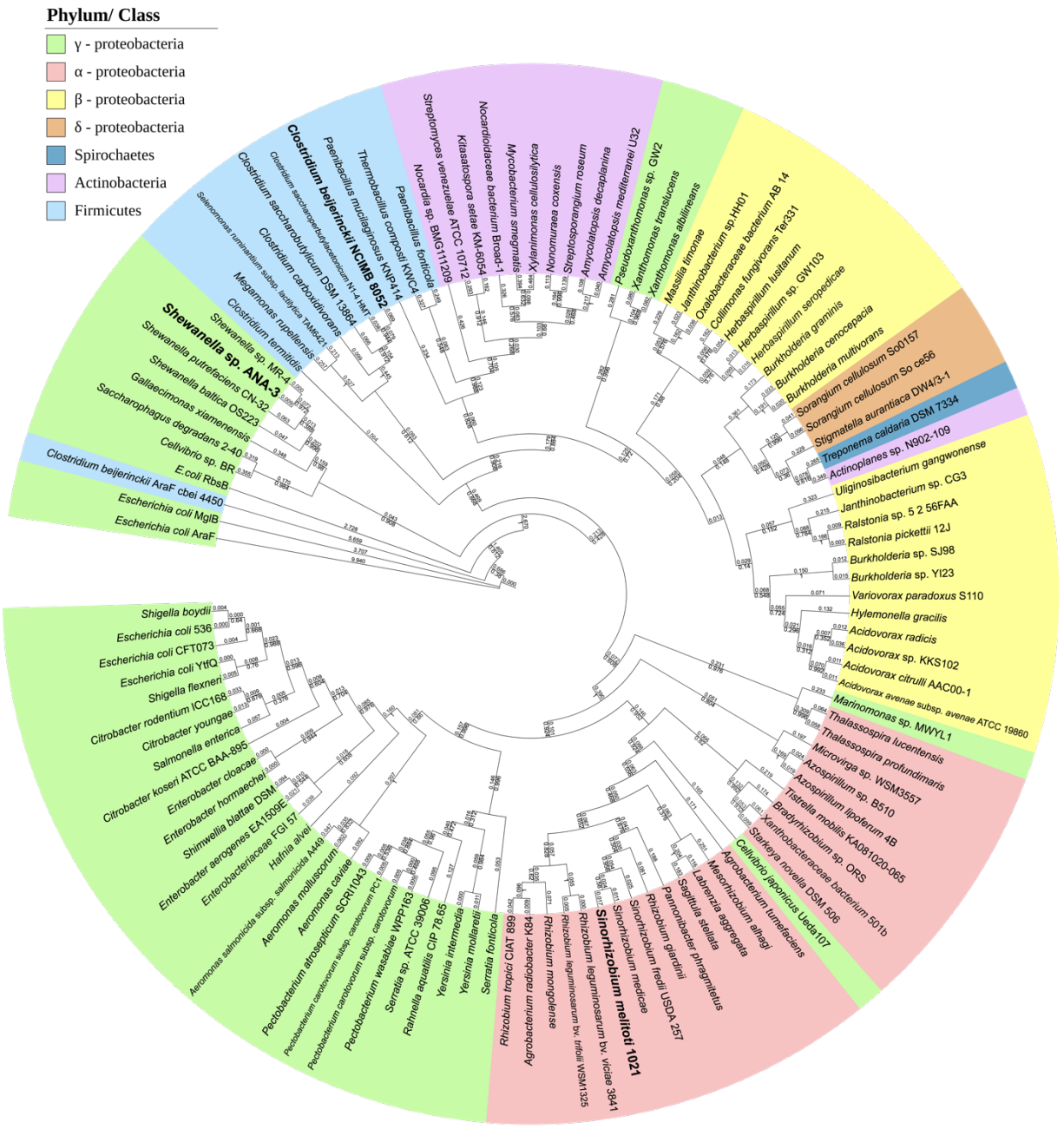


Figure 2. 2. Phylogram reveals GafAs are widespread in bacterial phyla/ classes.
Phylogenetic tree from maximum likelihood analysis of the GafAs alignment. Parameter values of the LG+ G + I + Γ model were used as calculated by AIC. The GafAs are found widespread in many bacterial phyla and classes.
The most notable GafAs from this analysis are the ones belonging to *Cellvibrio japonicus* Ueda107 and *Marinomonas* sp. MWYL1 which clustered with α - proteobacteria rather than the γ - proteobacteria clades. This could be indicative that these *gafA* genes were laterally transferred.
The values annotated on top of each node correspond to the evolutionary distance, which is measured in amino acid substitutions per site. The bootstrap values, present at the bottom each node, are provided as a fraction of 1.

'In silico' analyses were employed to gather further indications for GafA candidates which could point towards arabinose binding. The GafAs identified above by BlastP and Biocyc searches were patterns of gene analysis for synteny. The function of the neighboring genes of the *gafAs* was taken into consideration to establish a potential 'bias' towards arabinose binding versus galactose. Further, RegPrecise, an online database for the prediction of transcriptional regulatory networks in prokaryotes, was used to obtain indications on the transcription factors regulating the expression of the *gafABCD* operons and how these could be responsive to arabinose or galactose presence. RegPrecise performs manual curation taxonomic-specific reconstructions of bacterial regulons using comparative genomics by analysing regulons of a particular TF inferred in a set of closely related genomes. The database takes into account the functional classification of each TF based on the role of the genes in a particular metabolic pathway or biological process (Novichkov *et al*., 2007).

2. 3. 2. 1 Phylogenomic synteny analysis reveals arabinose - related gene clusters in many GafA clades

The GafA candidates identified from the phylogenetic analysis in the Section 2. 3. 1, had their genomic neighborhood analysed (20 kb upstream and downstream the *gafABCD* operon) using the following databases: Biocyc, Microbial Genomic context Viewer (MGcV), SEED viewer and PATRIC. The analysis showed that many of these bacterial operons are in proximity to genes encoding products involved in arabinan/ arabino-oligosaccharides hydrolysis and arabinose catabolism as well as uptake. Taking this into consideration, further investigation involved correlating the genomic context with the positioning of the GafA proteins on the phylogenetic tree. This approach was an attempt to detect whether there are any apparent clades relevant to arabinose utilization.

The subset of GafAs included in this analysis was considerably reduced compared to the phylogenetic analysis of *Section 2. 3. 1.* This way the phylogenomic analysis was limited to the GafA candidates which presented an interesting genomic context

relevant to arabinose or galactose. However, a few GafAs with low or no genetic context relevance to arabinose or galactose utilisation retained in the analysis in order to allow for construction of clades with true relations and avoid the rise of 'artifacts' or unrealistic clustering. The protein sequences of the candidates were aligned into MAFFT online alignment program using the L-INS-i iterative refinement method, which is normally used in the analysis with less than 200 protein sequences which contain one conserved domain and long gaps (Katoh, 2002). The choice was based on the fact that sequences with low identity to GafAs *ie*. the arabinopyranose dedicated transporters AraF from *E. coli* and *C. beijerinckii* NCIMB 8052 and the galactopyranose transporter MglB from *E. coli* were also included in the analysis. These sequences acted as the outgroup in the phylogenetic analysis, which is expected to position itself outside the clades formed by the GafAs *ie*. ingroup. This way one can examine how the GafAs relate to the evolutionary history of these SBPs, and also examine whether they form a clade which would suggest a tighter evolutionary and possibly functional relationship. Following curation of the alignment and as described in Materials and Methods *Section 2. 2. 1. 1*, a model for the construction of the phylogram was automatically selected using the AIC function (Table 2. 5). Similarly, to the extended tree shown in Figure 2. 2, the maximum – likelihood method was also used to create the current phylotree. The branches were inferred with high accuracy as the bootstrap was set at high value (*ie*. 500).

The unrooted phylogram produced is shown in Figure 2. 3. The latter was generated to assist in categorizing the different orthologs into clades and help study the relatedness between them. Subsequently, the above phylogram was made into rooted and was annotated with the genomic context of the GafAs to produce a phylogenomic tree to relate their inferred ligand specificities with their evolutionary distances (Figure 2. 4).

Group A is formed by a single clade and it's the smallest one in the tree. It is only comprised of sequences from γ-proteobacteria (Figure 2. 4). The clade in question is predicted to be functionally distinct from the γ proteobacteria clade of the group A.

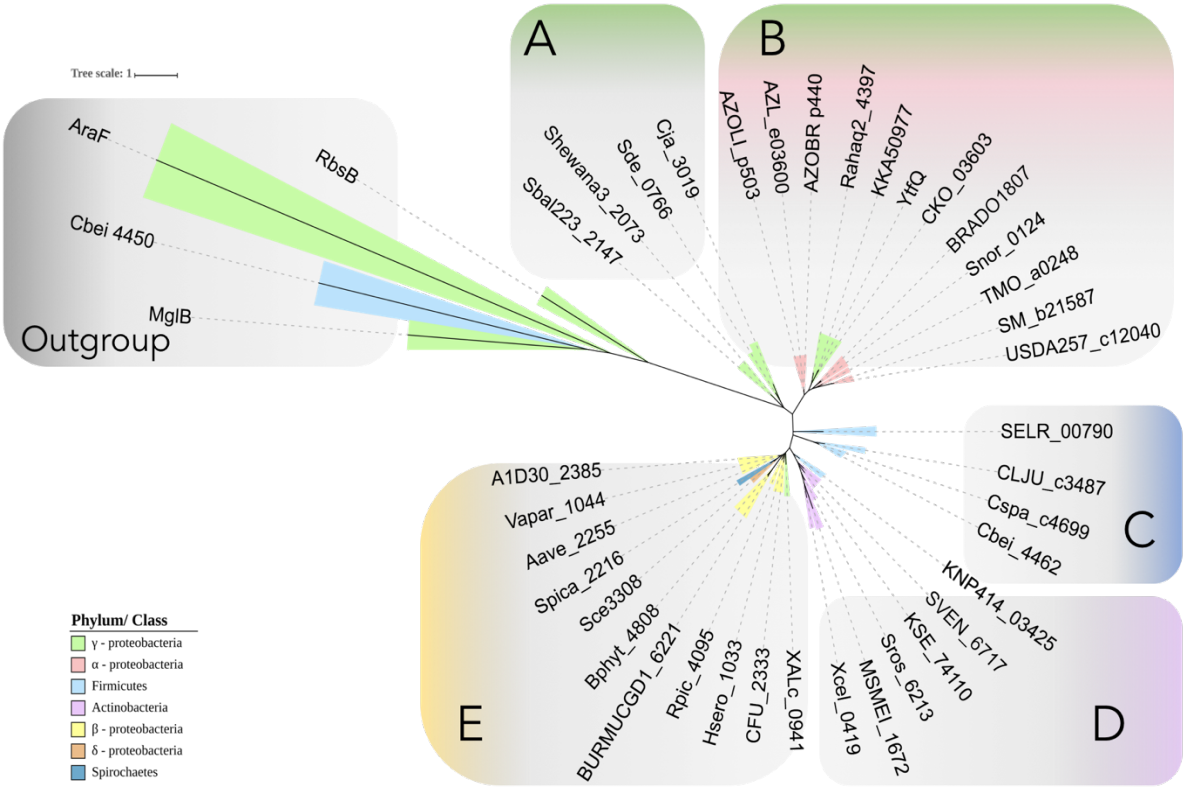


Figure 2. 3. Unrooted tree reveals the clade classification of the GafAs.
The unrooted tree was produced using the maximum - likelihood method, by using the LG + G + I + Γ parameters as calculated by AIC. The classification of the GafAs into separate 4 groups is apparent.

Names of the bacteria each GafA (and the non-orthologous proteins) belong to (by phylum/class):

γ - proteobacteria - MglB/ AraF/ RsbB/ YtfQ= *Escherichia coli* K-12 MG1655, Rahaq2_4397= *Rahnella aquatilis* CIP 78.65, KKA50977= *Salmonella enterica* subsp. *salamae,* CKO_03603= *Citrobacter koseri* ATCC BAA-895, Sbal223_2147= *Shewanella baltica* OS223, Shewana3_2073 (ie. GafASw)= *Shewanella* sp. ANA-3, Sde_0766= *Saccharophagus degradans* 2-40, Cja_3019= *Cellvibrio japonicus* Ueda10, XALc_0941= *Xanthomonas albilineans* GPEPC7 / **a - proteobacteria -** AZOLI_p50374= *Azospirillum lipoferum* 4B, AZL_e03600= *Azospirillium* sp. BSD, AZOBR_p440158= *Azospirillum brasilense* Sp245, BRADO1807= *Bradyrhizobium* sp. ORS 278, Snor_0124= *Starkeya novella* DSM 506, TMO_a0248= *Tistrella mobilis* KA081020-065, SM_b21587(*ie*. GafASm)= *Sinorhizobium meliloti* 1021, USDA257_c12040= *Sinorhizobium fredii* USDA 257 / **Firmicutes -** SELR_00790= *Selenomonas ruminantium* subsp. *lactilytica* TAM6421, CLJU_c34870= *Clostridium ljungdahlii* DSM 13528, Cspa_c46990= *Clostridium saccharoperbutylacetonicum* N1-4(HMT), *Cbei_4462(ie*. GafACb) / *Cbei_4450(ie*. AraFCb)= *Clostridium beijerinckii* NCIMB 8052, KNP414_03425= *Paenibacillus mucilaginosus KNP414* / **Actinobacteria -** AMED_4397= , SVEN_6717= *Streptomyces venezuelae*, KSE_74110= *Kitasatospora setae* ATCC 33774, Sros_6213= *Streptosporangium roseum* DSM43021, MSMEI_1672= *Mycobacterium smegmatis* MC2 155, Xcel_0419 = *Xylanimonas cellulosilytica* DSM 15894 / **β - proteobacteria -** CFU_2333= *Collimonas fungivorans* Ter331, Hsero_1033= *Herbaspirillum seropedicae* SmR1, Rpic_4095= *Ralstonia picketti* 12JR , BURMUCGD1_6221= *Burkholderia multivorans* CGD1, Bphyt_4808= *Burkholderia phytofirmans* PsJN, Aave_2255= *Acidovorax citrulli* AAC00-1, Vapar_1044= *Variovorax paradoxus* S110, A1D30_23855= *Acidovorax* sp. GW101-3H11 / **δ - proteobacteria -** Sce3308= *Sorangium cellulosum* So ce56 / **Spirochaetes -** Spica_2216= *Treponema caldaria* DSM 7334

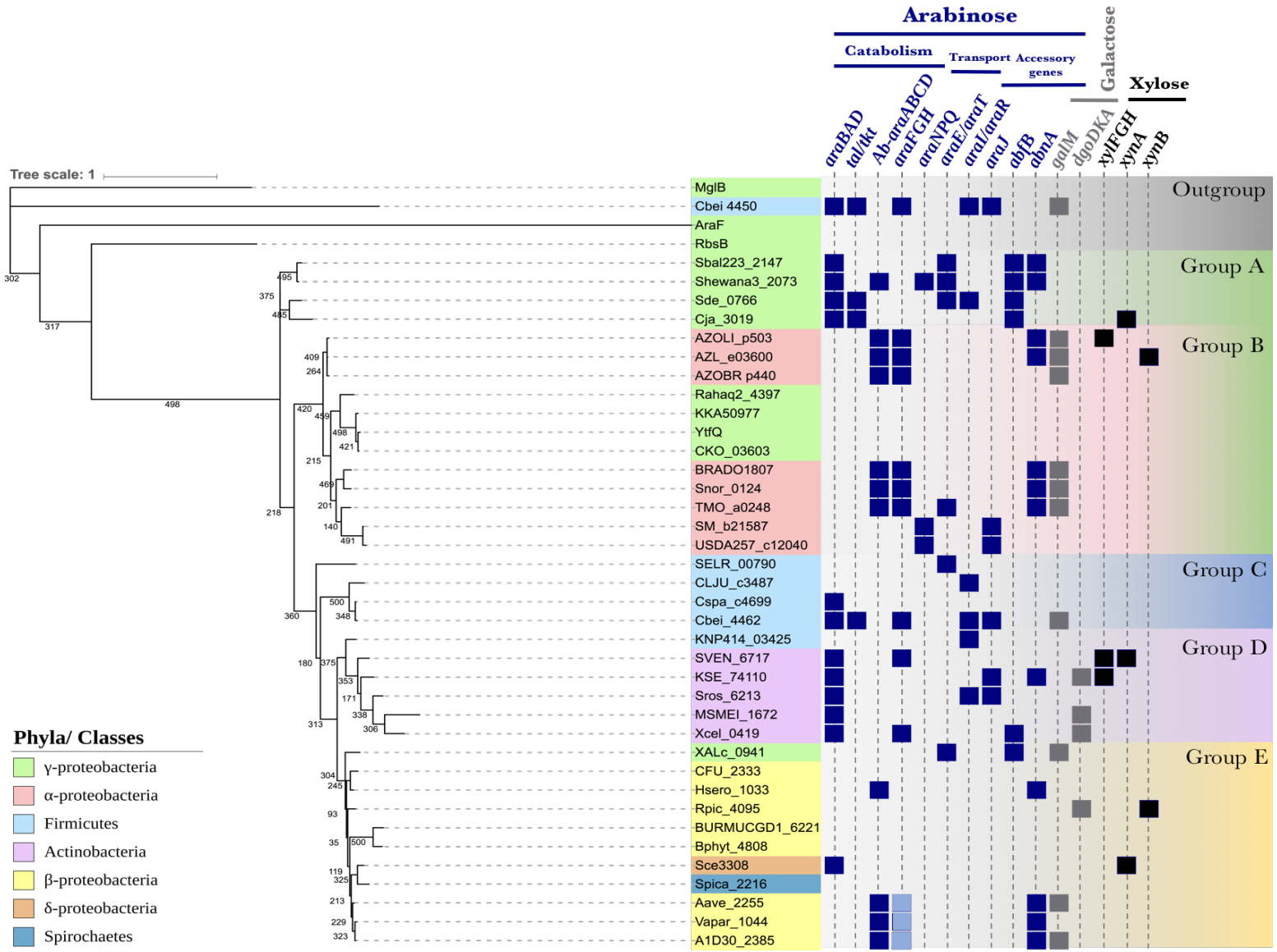

Arabinose
Catabolism
Transport
Accessory genes
Galactose
Xylose
araBAD
tal/tkt
Ab-araABCD
araFGH
araNPQ
araE/araT
araI/araR
araJ
dbfB
abnA
galM
dgoDKA
xylFGH
xynA
xynB
Tree scale: 1
MglB
Cbei 4450
AraF
RbsB
Sbal223_2147
Shewana3_2073
Sde_0766
Cja_3019
AZOLI_p503
AZL_e03600
AZOBR p440
Rahaq2_4397
KKA50977
YtfQ
CKO_03603
BRADO1807
Snor_0124
TMO_a0248
SM_b21587
USDA257_c12040
SELR_00790
CLJU_c3487
Cspa_c4699
Cbei_4462
KNP414_03425
SVEN_6717
KSE_74110
Sros_6213
MSMEI_1672
Xcel_0419
XALc_0941
CFU_2333
Hsero_1033
Rpic_4095
BURMUCGD1_6221
Bphyt_4808
Sce3308
Spica_2216
Aave_2255
Vapar_1044
A1D30_2385
Outgroup
Group A
Group B
Group C
Group D
Group E
Phyla/ Classes
γ-proteobacteria
α-proteobacteria
Firmicutes
Actinobacteria
β-proteobacteria
δ-proteobacteria
Spirochaetes

Figure 2. 4. Phylogenomic analysis reveals presence of arabinose related genes in the genomic context of many GafAs.
The phylogram was constructed using the maximum – likelihood method and the calculated parameters (LG + Γ + F) were applied. 500 bootstrap iterations were included in the construction to allow for high accuracy in the inference of the brances. The genomic context of each candidate *gafA* was investigated in the online microbial databases of PATRIC 3.4.2, Biocyc, SEED viewer and MgCV to help predict of the function of the neighbouring genes.
The genomic neighborhood of most of the candidate *gafAs* is governed by genes involved in the sensing, the transport and the catabolism of arabinose (blue). Some include genes encoding for enzymes involved in the hydrolysis of L-arabinofuranose residues from xylan. Notable examples include the candidates from Group B which is comprised solely by γ – proteobacteria with extensive clusters of arabinose utilisation found in close proximity to their *gafABCD*. Also, apart from arabinose relevant genes, some α proteobacteria from group A have genes for hydrolysis and transport of xylose (black) near *gafABCD* which indicates involvement in xylan utilisation. Interestingly, *aave_2255*, *vapar_1044* and *a1d30_2385* (light blue) in Group E are expressed under the same promoter with their respective *araFGH* operons. The two SBPs, one from each system, are predicted to be expressed first, preceding the expression of the membrane domain and the ATPase.
The presence of the respective genes/ operons is designated with a blue box next to each GafA ortholog. The light blue box indicates co-expression of the *araFGH* operon with *gafABCD* under the same promoter. The groups classification determined by the unrooted tree in Figure 2. 3 is also indicated in the current tree.
Names of the bacteria each GafA (and the non-orthologous proteins) belong to (by phylum/class):
γ - proteobacteria - MglB/ AraF/ RsbB/ YtfQ= *Escherichia coli* K-12 MG1655, Rahaq2_4397= *Rahnella aquatilis* CIP 78.65, KKA50977= *Salmonella enterica* subsp. *salamae,* CKO_03603= *Citrobacter koseri* ATCC BAA-895, Sbal223_2147= *Shewanella baltica* OS223, Shewana3_2073 (ie. GafASw)= *Shewanella* sp. ANA-3, Sde_0766= *Saccharophagus degradans* 2-40, Cja_3019= *Cellvibrio japonicus* Ueda10, XALc_0941= *Xanthomonas albilineans* GPEPC7 / **α - proteobacteria -** AZOLI_p50374= *Azospirillum lipoferum* 4B, AZL_e03600= *Azospirillium* sp. BSD, AZOBR_p440158= *Azospirillum brasilense* Sp245, BRADO1807= *Bradyrhizobium* sp. ORS 278, Snor_0124= *Starkeya novella* DSM 506, TMO_a0248= *Tistrella mobilis* KA081020-065, SM_b21587(*ie*. GafASm)= *Sinorhizobium meliloti* 1021, USDA257_c12040= *Sinorhizobium fredii* USDA 257/ **Firmicutes -** SELR_00790= *Selenomonas ruminantium* subsp. *lactilytica* TAM6421, CLJU_c34870= *Clostridium ljungdahlii* DSM 13528, Cspa_c46990= *Clostridium saccharoperbutylacetonicum* N1-4(HMT), *Cbei_4462(ie*. GafACb) / *Cbei_4450(ie*. AraFCb)= *Clostridium beijerinckii* NCIMB 8052, KNP414_03425= *Paenibacillus mucilaginosus KNP414* / **Actinobacteria -** AMED_4397= , SVEN_6717= *Streptomyces venezuelae,* KSE_74110= *Kitasatospora setae* ATCC 33774, Sros_6213= *Streptosporangium roseum* DSM43021, MSMEI_1672= *Mycobacterium smegmatis* MC2 155, Xcel_0419 = *Xylanimonas cellulosilytica* DSM 15894 / **β - proteobacteria -** CFU_2333= *Collimonas fungivorans* Ter331, Hsero_1033= *Herbaspirillum seropedicae* SmR1, Rpic_4095= *Ralstonia picketti* 12JR , BURMUCGD1_6221= *Burkholderia multivorans* CGD1, Bphyt_4808= *Burkholderia phytofirmans* PsJN, Aave_2255= *Acidovorax citrulli* AAC00-1, Vapar_1044= *Variovorax paradoxus* S110, A1D30_23855= *Acidovorax* sp. GW101-3H11 / **δ - proteobacteria -** Sce3308= *Sorangium cellulosum* So ce56 / **Spirochaetes -** Spica_2216= *Treponema caldaria* DSM 7334
Genes/Operons encode for : *araBAD*= primary arabinose catabolism, *tal*/*tkt*: Pentose Phosphate Pathway, *Ab-araABCD*: arabinose catabolic pathway II, *araFGH*: arabinopyranose ABC transport, *araE*/*araT*: arabinose secondary transport, *araI*/*araR*: arabinose, *araNPQ*: arabinosides ABC transport, *abfB*: α-L- arabinofuranosidase, *abnA*: arabinan endo- 1,5 (or 1,4)- α- arabinofuranosidase, *galM*: arabinose/galactose mutarotase, *dgoDKA*: galactose catabolic pathway II (via galactonate), *xylFGH*: xylose ABC transport, xynA: endo 1,4- xylanase, *xynB*: xylan 1,4- β- xylosidase

The genomic context of these GafAs is governed by many genes encoding for transport systems, as well as catabolic and hydrolytic enzymes involved in the xylan utilization. The most intriguing candidate from group A, is the Shewana3_2073 from *Shewanella* ANA sp. 3, as its genomic context is governed by a huge cluster for xylan utilization (Figure 2. 4 and 2. 5). The Shewana3_2086 protein is predicted to have xylanolytic activities whereas the Shewana3_2077 and Shewana3_2078 are predicted arabinofuranosidases, which are all actively involved in xylan deconstruction. Along with the GafABCD, the uptake of the released hydrolysed products is facilitated by Shewana3_2081 which is a predicted symporter for arabinosides and belongs in the Glycoside– Pentoside- Hexuronide: cation symporter family of the MFS. The strongest aspect of this cluster is the presence of two catabolic pathways: the primary degradation pathway for arabinose (*ie*. *araBAD*) and the degradation pathway III mentioned above. This ensures that the entire pool of arabinose released is efficiently shuttled into the intracellular catabolism with the arabinofuranose fraction catalysed into L-arabinono – 1, 4 – lactone by L-arabinose 1-dehydrogenase (Ab-AraA) and the arabinopyranose conversion into L- ribulose, by the first enzyme of the primary degradation pathway, L- arabinose isomerase (AraA).

The orthologs occupying the group B are from α and γ proteobacteria (Figure 2. 3). The evolutionary relatedness between the three subgroups of group B is high as they share a common ancestor (bootstrap value: 399). Examination of the phylograms shows that the GafAs from the γ proteobacteria (*eg*. YtfQ), present in the B2 subgroup, do not neighbor with genes related to arabinose utilisation. The absence of arabinose related cluster of genes denotes that the *gafABCD* operon from these bacteria is very unlikely to form part of a rapid response to arabinose – rich environments. This is inferred by the fact that genes whose products form part of a series of events, such as arabinose hydrolysis, signaling, acquisition and intracellular catabolism, are normally found colocalized. As a result of colocalization, a transcription factor's search for its regulatory sites is expedited; thus, allowing for quick transcription and translation as those are coupled spatially and temporally. Thus, the transcription factors synthesized near their genes can rapidly bind colocalized sites (Warren and ten Wolde, 2004;

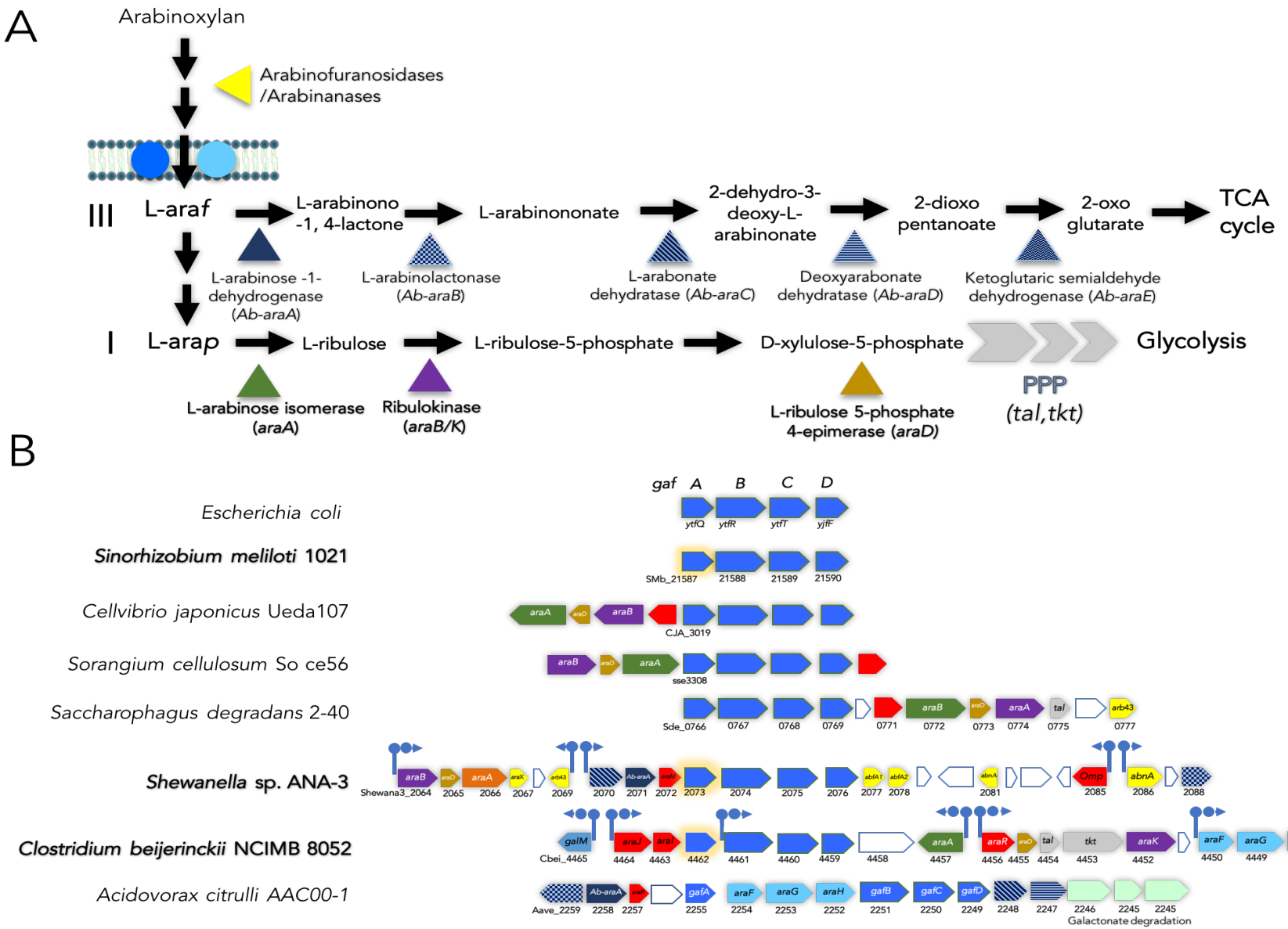

A
Arabinoxylan
Arabinofuranosidases
/Arabinanases
III
L-araf
L-arabinono
-1, 4-lactone
L-arabinononate
2-dehydro-3-
deoxy-L-
arabinonate
2-dioxo
pentanoate
2-oxo
glutarate
TCA
cycle
L-arabinose -1-
dehydrogenase
(Ab-araA)
L-arabinolactonase
(Ab-araB)
L-arabonate
dehydratase (Ab-araC)
Deoxyarabonate
dehydratase (Ab-araD)
Ketoglutaric semialdehyde
dehydrogenase (Ab-araE)
I
L-arap
L-ribulose
L-ribulose-5-phosphate
D-xylulose-5-phosphate
Glycolysis
PPP
(tal,tkt)
L-arabinose isomerase
(araA)
Ribulokinase
(araB/K)
L-ribulose 5-phosphate
4-epimerase (araD)
B
Escherichia coli
Sinorhizobium meliloti 1021
Cellvibrio japonicus Ueda107
Sorangium cellulosum So ce56
Saccharophagus degradans 2-40
Shewanella sp. ANA-3
Clostridium beijerinckii NCIMB 8052
Acidovorax citrulli AAC00-1
gaf A B C D
ytfQ ytfR ytfT yjfF
SMb_21587 21588 21589 21590
CJA_3019
sse3308
Sde_0766 0767 0768 0769 0771 0772 0773 0774 0775 0777
Shewana3_2064 2065 2066 2067 2069 2070 2071 2072 2073 2074 2075 2076 2077 2078 2081 2085 2086 2088
Cbei_4465 4464 4463 4462 4461 4460 4459 4458 4457 4456 4455 4454 4453 4452 4450 4449 4448
Aave_2259 2258 2257 2255 2254 2253 2252 2251 2250 2249 2248 2247 2246 2245 2245
Galactonate degradation

Figure 2. 5. Genomic context of selected gafAs with or without the presence of arabinose related gene operons

A) The first step in the utilisation of arabinoxylan (and arabinan) is the hydrolysis of the L-arabinofuranose by the action of arabinofuranosidases and arabinanases. The released L-arabinose is transported in the bacterial cytoplasm by dedicated transport systems such as the GafABCD and AraFGH. Once inside the cytoplasm the L-arabinopyranose is shuttled into the primary (I) catabolic pathway which includes the L-arabinose isomerase (AraA), Ribulokinase (AraB/K*)*, L-ribulose 5-phosphate (AraD) and 4-epimerase mentioned in the order of enzymatic activity. The products of the primary catabolic pathway enter the pentose phosphate phosphate and contribute to the central metabolism of the cell. An alternative pathway for the catabolism of the L-arabinose, specific to its furanose form, is the L-arabinose degradation pathway III (*Ab-araA-E*). The products of this pathway are shuttled into the TriCarboxylic Acid cycle (TCA) of the central bacterial metabolism.

B) The genomic context of *gafABCD* from selected bacteria is presented. The *cbei_4462* exhibit the most interesting genomic context and present high relevance to arabinose including genes for sensing (*araIJ* and *araR*), catabolism (*araBAD*), The designated transcription sites were obtained from RegPrecise preditions. Notable is the absence of any arabinose related genes in the genomic context of *ytfQ* and *sm_b21587*. Also important is the genomic operon from *Acidovorax citrulli*, which exhibits a co-transcribed operon including both pyranose and furanose-specific arabinose systems. With the exception of Ab-araE, the enzymes of degradation pathway III are all included in its cluster. It also includes the galactonate degradation enzymes which are involved in the bacterial galactose degradation pathway II.

The enzymes involved in the arabinose degradation pathway III are denoted with patterned fill. The galactonate utilisation enzymes are shown in light green. The chosen GafAs are highlighted in orange and their bacteria have their names in bold.

Names of the bacteria each GafA (and the non-orthologous proteins) belong to are indicated next to the respective operons. The colour and pattern fill of the arrows in (A) indicates the function of the products of the genes shown in the genomic context of (B).

Genes/Operons encode for : *araBAD*= primary arabinose catabolism, *Ab-araA-E*= arabinose degradation pathway III, *tal*/*tkt*: Pentose Phosphate Pathway, *araFGH*: arabinopyranose ABC transport, *araE*/*araT*: arabinose secondary transport, *araI*/*araJ*: two component system for arabinose sensing, araR: arabinose repressor, *abfB*: α- L- arabinofuranosidase, *abnA*: arabinan endo- 1,5 (or 1,4)- α- arabinofuranosidase, araX= arabinose phosphoglucomutase, araM/galM= arabinose/galactose mutarotase, Omp(Ara)= arabinose outer membrane receptor.

Hershberg *et al.*, 2005; Golding and Cox, 2006). Additionally, all of the four candidates comprising this node are from bacterial pathogens/ symbionts of animal hosts, ie. *Escherichia coli*, *Citrobacter koseri* ATCC BAA-895, *Salmonella enterica* subsp. *salamae* and *Rahnella aquatilis* CIP 78.65. Herbivorous mammals rely on resident gut bacteria to gain energy from the main food sources as they lack enzymes for degradation of polysaccharides. Analysis of the microbiota of the rumen and the human large intenstine, the main primary degraders of cellulose and hemicellulose where *Clostridium* spp., *Ruminococcus* spp., *Eubacterium* spp., *Prerotella* spp., *Bacteroides* spp. and *Roseburia* spp. (Flint *et al*., 2012). These bacteria are known to have extensive operons for hemicellulotic utilisation which would be taken into account in our synteny analysis, if these had a direct ortholog for GafA. Apart from the *Clostridium* spp., and the bacteria from this node, we failed to identify direct GafA orthologs in the genome of the rest of the aforesaid gut species and therefore focused our analysis towards the bacterial plant pathogens and rhizosphere bacteria.

The rest of the group B is consisted of plant symbionts which thrive in the rhizosphere and some of them induce nodule production, such as *Bradyrhizobium* sp. ORS 278 and *Sinorhizobium fredii* USDA 257. The corresponding genomic context of these bacteria is governed by genes responsible for the production of the arabinopyranose ABC transporter, *ie*. AraFGH and the L-arabinose degradation pathway III, *ie*. Ab-AraABCD. The presence of the pyranose transporter indicates that these bacteria are utilizing the full 100 % of L-arabinose present in the extracellular space. This is not surprising as the root exudates *ie*. mucilage of many plants such as pea, cowpea, maize and rice are rich in galactose and arabinose. Specifically, the concentrations of these sugars in the mucilage range from 14% to 34% for arabinose and 20.3% to 34% for galactose (Knee *et al.*, 2001; Moody *et al.*, 1988; Basic *et al.*, 1986; Chaboud and Rougier, 1984). Interestingly, arabinose is found primarily in the furanose form whereas only galactopyranose subunits are present in these sugar – rich exudates (Basic *et al.*, 1986; Moody *et al.*, 1988). This explains the fact that the starting substrate of the degradation pathway III is the furanose form of arabinose and not the pyranose. Even though the first enzyme of this pathway, *ie*. Ab-AraA (L- arabinose 1-dehydrogenase), has not

been crystallised bound to arabinofuranose, the recognition of the latter is inferred by the fact that the product of the reaction is the L-arabino- 1, 4 – lactone (Watanabe *et al.*, 2006), which couldn't be produced by the oxidation of the pyranose form. This is expected to serve for rapid shuttling of the substrate into intracellular catabolism following transport, as the requirement for conversion into the pyranose form is circumvented. In the crowded bacterial environment of the rhizosphere, this is offering a competitive edge to the bacterium and accelerates colonization of the niche. Further, the two last candidates from *Sinorhizobium meliloti* 1021 and *Sinorhizobium fredii* USDA 257, neighbor with an arabinofuranosidase and a predicted ABC transporter for arabinosides. The presence of these two enzymes points towards effective utilization of arabinan which is a frequent side chain of the pectic rhamnogalactan I (Mohnen, 2008).

The group C is comprised solely by Firmicutes (Figure 2. 3). Careful examination of the genomic context of the *gafAs* from this subgroup (Figure 2. 4), shows that the most interesting candidate is the Cbei_4462 from *Clostridium beijerinckii* NCIMB 8052. The synteny analysis revealed the presence of the *araFGH* operon in its arabinose gene cluster which encodes for the pyranose dedicated ABC transport system, similarly to the α - proteobacteria from group A. The genomic neighborhood resembles the one from *shewana3_2073*, in that it contains the *araBAD* operon which encodes the catabolic enzymes for arabinose utilization. It also exhibits a feature not so common across the genomic context of the GafAs studied here; it contains a predicted two – component system for sensing extracellular arabinose. A typical two - component system is comprised of a histidine kinase sensor which senses the extracellular cue and an intracellular response regulator which mediates the cellular response (Bourret, 1995). In this case the roles are predicted to be carried out by AraJ and AraI respectively. The operon also includes a gene predicted to encode for AraR; an arabinose regulated repressor, *ie*. *cbei_4456* (Mota, Sarmento and de Sa-Nogueira, 2001; Zhang *et al.,* 2011). A comparative genomic study which reconstructed the AraR regulons in *Clostridium* spp. detected AraR binding sites upstream the *araBAD*, *araFGH* as well as itself in *C. beijerinckii* NCBI 8052 (Zhang *et al.*, 2011). The study also

showed that the regulon extended to the pentose phosphate pathway of the central carbon metabolism, as binding sites were identified in the regulatory region of *tkt* (*cbei_4454*) and *tal* (*cbei_4455*). The close proximity of the latter genes and the common regulation exerted in this arabinose cluster, underlines the importance of arabinose utilization to *C. beijerinckii* NCBI 8052; this serves for rapid shuttling of the products of the L-arabinose primary catabolic pathway into pentose phosphate pathway for NADPH generation and subsequently the aromatic amino acid production via the chorismate biosynthesis I. Additionally, the presence of *gafABCD* operon in this huge cluster denotes the importance of the arabinofuranose uptake in utilizing the full context of arabinose found in the environmental niche of this bacterium.

The group D is comprised mainly by Actinobacteria, with many containing arabinose related genes in their genomic context (Figure 2. 3 and 2. 4). Their genomic context comprises the *araBAD* genes which are present in all of the operons of these bacteria as well as an abundancy of genes involved in the xylan hydrolysis. These encode for enzymes with xylosidase (*xynB)* and xylanase (*xynA)* as their predicted functions. The xylanases are glycosidases which catalyse the endohydrolysis of 1,4 –β -D- xylosidic linkages present in xylan (Whistler and Masak, 1955). The produced xylo-oligosaccharides of the previous reaction are the substrate of the xylosidases which hydrolyse the non-reducing end xylose residues from them. In the case of the *Xylanimonas cellulosilytica* DSM 15894, *abfB* encodes for arabinofuranosidase which hydrolyse the non-reducing end of arabinofuranose residues in the xylan and arabinosides. The synergistic action of these hydrolases creates a pool of transportable xylo-oligosaccharides and monosaccharides, including xylotriose, xylobiose, xylose and arabinose. The presence of the operons for AraFGH and the xylopyranose ABC transporter *ie*. XylFGH, in the cluster of *Xylanimonas cellulosilytica* DSM 15894 ensures that the arabinose and xylose released is efficiently transported inside the cell.

The last group, denoted as Group E, is governed by β - proteobacteria. The genomic contexts from the *Acidovorax* species (Aave_2255 and A1D30_23855) and *Variovorax*

paradoxus S110 (Vapar_1044), contain unique features which haven't been observed in the previous gene clusters examined. These are predicted to be in the same genomic operon with the *araFGH* and are very likely to be co-expressed as the first in the gene expression order are the SBP encoding genes *ie*. *araF* and *ytfQ*. Such organisation shuffles the gene arrangement of both systems which seemed fixed during the present analysis, prior to examination of these genomes from the Comamonadaceae genus. One apparent reason for this permutation in the genomic order, is that it potentially serves for an even speedier response to the arabinose pool present in the extracellular space, compared to the α - proteobacteria of the group A which exhibit the *araFGH* in the same cluster but not the in same operon as *gafABCD*.

To summarise the analysis a more holistic approach was followed by mapping arabinose related functions present in each cell of selected bacteria (Figure 2. 6). The schematic depiction includes the loci names of the *gafABCD* neighboring genes near the respective extracellular or subcellular function each gene encodes for (Figure 2. 6). Inspection of the diagram, shows that the most arabinose - relavant genomic neighbourhood of *gafABCD* are the ones from: *Shewanella* sp. ANA – 3 (Shewana3_2073), *Shewanella baltica* OS223 (Sbal223_2147), *Clostridium beijerinckii* NCBI 8052 (Cbei_4462), *Saccharophagus degradans* 2-40 (Sde_0766) and *Xanthomonas albilineans* GPEPC73 (XALc_0945). The choice of the above was based on whether their arabinose - related clusters could act individually to utilise efficiently arabinose by containing at least one (or none) gene encoding for functions of the arabinose initial utilisation stages *ie*. hydrolysis, sensing **and** at least one gene for the late utilisation stages *ie*. transport and catabolism. However, the above choice was irrespective of the amount of genes products in each cluster which participate in the individual stages.

2. 3. 2. 2 RegPrecise predictions show that arabinose - responsive transcription factors regulate GafA expresssion

Even though the examination of the genomic context is a good indication of what the

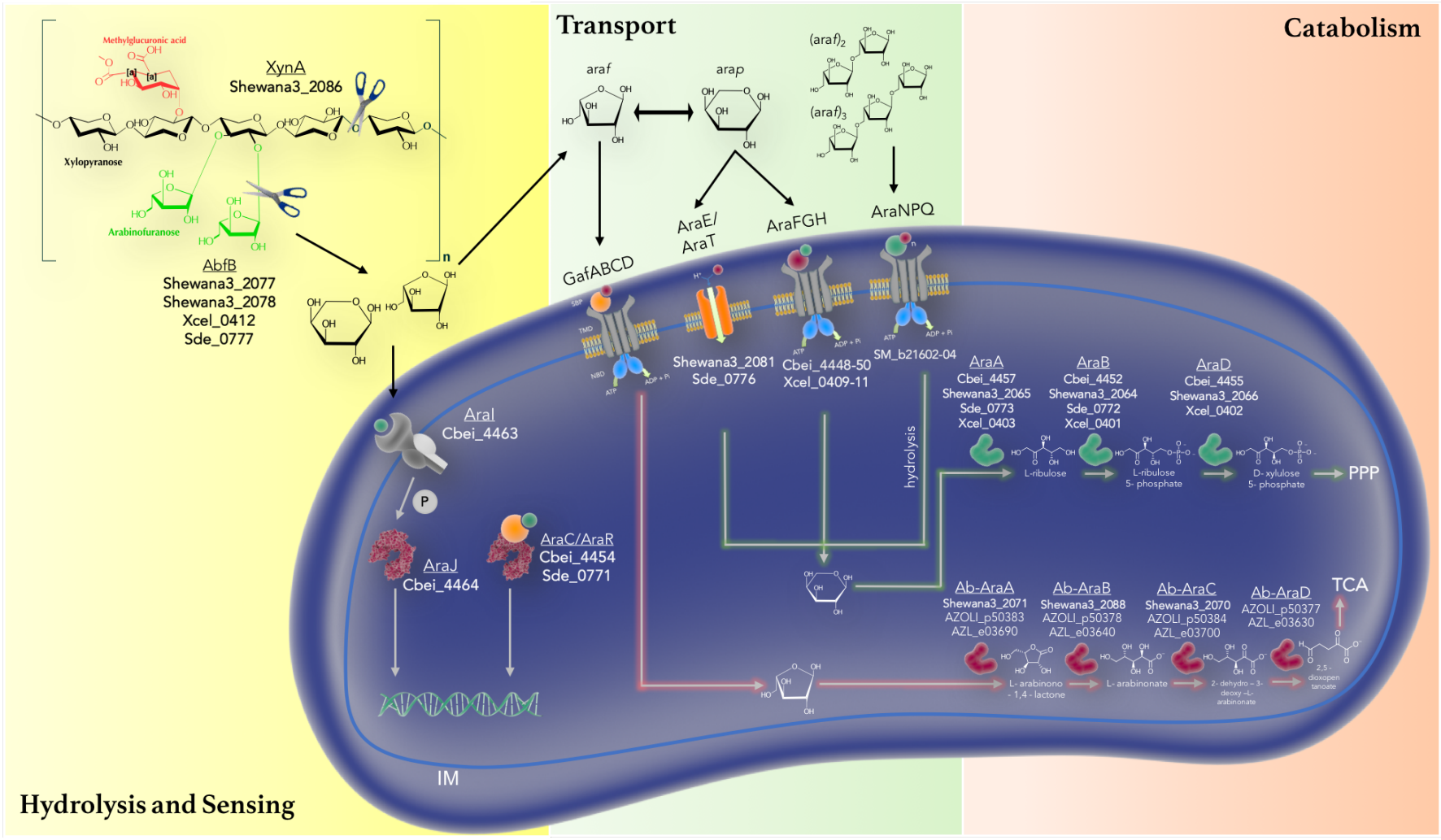

Transport
Catabolism
Hydrolysis and Sensing
Methylglucuronic acid
XynA
Shewana3_2086
Xylopyranose
Arabinofuranose
AbfB
Shewana3_2077
Shewana3_2078
Xcel_0412
Sde_0777
araf
arap
(araf)2
(araf)3
GafABCD
AraE/
AraT
AraFGH
AraNPQ
Shewana3_2081
Sde_0776
Cbei_4448-50
Xcel_0409-11
SM_b21602-04
AraI
Cbei_4463
P
AraJ
Cbei_4464
AraC/AraR
Cbei_4454
Sde_0771
hydrolysis
AraA
Cbei_4457
Shewana3_2065
Sde_0773
Xcel_0403
AraB
Cbei_4452
Shewana3_2064
Sde_0772
Xcel_0401
AraD
Cbei_4455
Shewana3_2066
Xcel_0402
L-ribulose
L-ribulose
5- phosphate
D- xylulose
5- phosphate
PPP
Ab-AraA
Shewana3_2071
AZOLI_p50383
AZL_e03690
Ab-AraB
Shewana3_2088
AZOLI_p50378
AZL_e03640
Ab-AraC
Shewana3_2070
AZOLI_p50384
AZL_e03700
Ab-AraD
AZOLI_p50377
AZL_e03630
TCA
L- arabinono
- 1,4 - lactone
L- arabinonate
2- dehydro – 3-
deoxy –L-
arabinonate
2,5 -
dioxopen
tanoate
IM

Figure 2. 6. Schematic diagram depicting the extracellular and subcellular functions of neighboring genes of selected gafAs

The bacterial model diagram summarises the functions and localisation of the products of the genes neighboring the studied *gafABCD* operons. The first step in the diagram is the hydrolysis of the xylan which requires the concerted activity of xylanases/ xylosidases and arabinofuranosidases. Two - component system, AraIJ, senses the extracellular released arabinose and induces the expression of arabinose transport systems and catabolic enzymes. In some bacterial systems, the intracellular transcription factor, AraC, is responsible for induction of the expression of arabinose utilisation genes and increase their basal levels which normally exist during glucose metabolism. AraE/AraT secondary transport systems and AraFGH, GafABCD ABC scavenging transport systems are transporting arabinose and amplifying the intracellular levels of arabinose, thus energising a positive feedback loop which leads to rise of expression levels. Outside the context of xylan utilisation, the AraNPQ system is responsible for the arabinoside uptake which are normally released by the function of arabinan arabinofuranosidases which act on long chains of arabinofuranose branching off pectin. Once inside the cell, the arabinopyranose is shuttled into the AraBAD catabolic pathway (green), which precedes the pentose phosphate pathway. In the case of *Azospirillum* spp. and cognate bacteria, prior to its conversion to the pyranose form, the arabinofuranose, is rapidly shuttled into the L-arabinose degradation pathway III therefore serving for rapid catabolism.

Most notable examples are the *Shewanella* sp. ANA – 3 (Shewana3_2073), *Clostridium beijerinckii* NCBI 8052 (Cbei_4462), *Saccharophagus degradans* 2-40 (Sde_0766) and *Xylanimonas cellulosilytica* DSM 15894 (Xcel_0419) with apparent ability to fully utilise arabinose following induction of the *gafA* - containing gene clusters. The exclusion of any proteins from the γ-proteobacteria of the subgroup B2 (*eg*. YtfQ) denotes the absence of relevance to arabinose utilisation.

Names of the bacteria each protein belongs to (by phylum/class):

γ - proteobacteria - Shewana3_20.. = *Shewanella* sp. ANA-3, Sde_07.. = *Saccharophagus degradans* 2-40 / **a - proteobacteria -** AZOLI_p503.. = *Azospirillum lipoferum* 4B, AZL_e036.. = *Azospirillium* sp. BSD, SM_b215.. = *Sinorhizobium meliloti* 1021/ **Firmicutes -** Cbei_44.. = *Clostridium beijerinckii* NCIMB 8052 / **Actinobacteria -**Xcel_04.. = *Xylanimonas cellulosilytica* DSM 15894

Genes/Operons encode for : AraBAD= primary arabinose catabolism, Ab-AraABCD: arabinose catabolic pathway II, GafABCD: arabinofuranose ABC transport, AraFGH: arabinopyranose ABC transport, AraE/AraT: arabinose secondary transport, AraI/AraR: arabinose, AraNPQ: arabinosides ABC transport, AbfB: α- L- arabinofuranosidase, XynA: endo 1,4- xylanase.

All the chemical structures were produced in Chem Doodle. The green arrow and red arrow represent catabolic pathways of arabinofuranose and arabinopyranose respectively.

substrate of the transporter might be, it cannot distinguish sufficiently between arabinose and galactose transport cause the clustering is species-dependent and there aren't enough galactose-related proteins to allow for confident predictions. Therefore, the search was refocused towards predictions on the regulation of the expression of the *gafABCD* operons. Regarding the characterized GafAEc, the *ytfQRTyjfF* (or *gafABCDEc*) operon was predicted to contain two binding sites for the galactose repressor GalR (and GalS) upstream of the promoter at positions -97.5 and +57.5 relative to the predicted transcriptional start (Horler *et al.*, 2009). The regulation was experimentally confirmed by a recent study contacted by Qian *et al.* (2016) which employed ChIP-chip assays to detect GalR target sequences *in vivo* and further validate the analysis using ChIP-qPCR. The analysis confirmed that the predicted regulatory site positioned at +57.5 (*ie*. GTGGAAACGCTTACT) is a genuine site of GalR binding. Notably, the most significant signal obtained from the ChIP-qPCR, as inferred by the occupancy units (*ie*. background – substracted enrichment relative to a transcriptionally inert gene) was produced by the binding at the regulatory site of *gafAEc* mentioned above (Qian *et al.*, 2016). Despite the evidence provided above, it is still unclear whether the regulation exerted by GalR is galactose dependent. Nevertheless, global gene expression analysis of a Δ*galR* mutant using DNA tiling microarrays did not detect any effect in *gafABCDEc* expression (Qian *et al.*, 2016). As the experiment was performed in the absence of D-galactose, it could indicate that the *gafABCDEc* is subjected to regulation by GalR solely in the presence of D-galactose, thereby activating rather than repressing the expression. Furthermore, the operon in question was also shown to be regulated by the arabinose activator/repressor AraC in ChIP-chip and ChIP-qPCR assays, in a similar manner to the study by Qian and his coworkers (Stringer *et al.*, 2014). However, the repression was shown to be arabinose independent as the expression of *gafAEc* wasn't affected by the presence of this sugar. Though its expression was slightly upregulated in a Δ*araC* mutant background, revealing a weak repression exerted by AraC which was shown to be irrespective of arabinose presence (Stringer *et al.*, 2014). This non-canonical regulation of *ytfQRTyjfF* expression by AraC and the fact that the WT levels

of expression are not significantly lower than that of the Δ*araC* mutant might suggest that the regulation exerted by this TF is not the sole one. It is evident that GalR is a potential candidate for the additional transcriptional control, which could very likely exert a stronger effect than AraC. One could also speculate that the binding of AraC leading to the weak repression of the *galABCDEc* operon is actually an artifact of vertical gene transfer during evolution. Regardless of the accurate regulation applied on the operon, the binding of both AraC and GalR indicates that this transport system is not arabinose neither galactose specific, as confirmed by their similar K_D values (Horler *et al.*, 2009). Also, proteins from 2D gels of the periplasm from *E. coli* which grew in LB media or glucose MOPS minimal media were N-terminally sequenced and revealed the presence of YtfQ at both conditions (Richard Horler thesis, 2009; Corbin *et al*., 2003; Lopez-Campistrous *et al*., 2005).

The RegPrecise online software was used to gain further insight on the regulation of *gafABCD* expression from other bacteria (Novichkov *et al.*, 2013). This software is a web resource for collection, visualization and analysis of transcriptional regulons reconstructed by comparative genomics. The search revealed that a homolog of the transcription factor AraR, likely regulates the expression of the *gafABCD* operon (cbei4459-4462) in *Clostridium beijerinckii* NCIMB 8052. Studying of the transcriptional regulation of *the* L- arabinose catabolic pathway in *Bacillus subtilis*, has provided insights in the structure and function of AraR of Gram (+*ve*) bacteria. This transcription factor consists of a GntR-type DNA-binding domain in its N-terminus, and an effector recognition domain which is shared with regulators from the GalR/LacI family (Mota, Tavares and S'a-Nogueira, 1999). Experimental data obtained by DNase I footprinting, showed that in the absence of the effector arabinose, AraR represses the expression of 13 genes, including the *araR* gene, the genes involved in arabinose catabolism (*araABD* and *araE*) and arabinoside degradation (*araNPQ*, *abfA*, *abnA*, and *abfA2*), and two genes with unknown functions (*araLM*) (Raposo *et al.*, 2004). In the case of the cognate bacterium, *Clostridium beijerinckii* NCIMB 8052, the RegPrecise database provides predictions for not just the transcriptional regulation of the *gafABCD* operon but for all of its arabinose – related

genomic neighbourhood (See Section 2. 3. 2. 1), which apparently constitutes an AraR regulon (Figure 2. 5). In the predicted regulon, the AraR binding sites are found in the regulatory regions upstream the *araI*, *galM*, *gafB*, *araA*, *araR* and *araF* genes. Interestingly, the *gafACb* (*ie*. *cbei_4462*) is not predicted to be directly subjected to regulation by AraR, which is an unusual example of transcriptional order for ABC systems, where the SBP – encoding gene is typically expressed first in copious numbers. Instead, the *gafACb* is predicted to be co-transcribed as a single unit with the genes encoding for the arabinose two - component regulatory system (*ie*. *araI* and *araJ, cbei_4463* and *cbei_4464* respectively).

Furthermore, the search in the RegPrecise database revealed further indications for the regulation of *gafASw* expression in *Shewanella* sp. ANA-3 (Shewana3_2073). This is predicted to be regulated by AraR ortholog of Shewanella, similarly to *gafACb*. This further enhances the indications that this PBP might recognize and transport the arabinofuranose over galactofuranose.

2. 3. 3 Experimental evidence to correlate SMb_21587 and Shewana3_2073 with arabinose uptake

The predictions obtained from the RegPrecise provided satisfactory indications to point towards substrate specificities exhibited by the GafAs. The collection of the indications was further strengthened by literature search which identified studies with implications on the induction of the *gafABCD* expression and the effect of its deletion in the bacterial fitness. Furthermore, the inability of specific candidate bacteria to grow on certain carbon sources of interest, such as arabinose and galactose, provided strong evidence for the incapacity of the respective GafAs in transporting these ligands.

2. 3. 3. 1 The expression of SM_b21587-90 is induced by L-arabinose, D-fucose and talose

Experimental evidence for the induction of GafA expression from *Sinorhizobium meliloti* 1021 were provided by a study conducted by Mauchline and his colleagues who

mapped the solute-binding protein-dependent transportome of the bacterium. The study employed reporter assays by cloning the putative promoter regions for the ABC operons of *S. meliloti* upstream of reporter genes (*ie*. *lacZ*, *gusA* and *gfp*) (Mauchline *et al.*, 2006). They showed that the expression of GafA (*SM_b21587*), and the rest of the operon, was heavily induced by D-fucose (40-fold), L-arabinose (10–33-fold), and talose (11-fold) (Table 2. 8, *see next page*) (Mauchline *et al.*, 2006). Considering that the induction profile corresponds to the specificity of the solute transported and that D-fucose (5-methyl-L-arabinose) is a structural derivative of the L-arabinose, it is not surprising it induces heavily the expression of SMb21587. These gene expression assays were the first experimental evidence found in the search to manage successfully to indirectly correlate SM_b21587 with its presumptive ligands. Another approach used to correlate the GafAs with specific ligands was performed by examination of the Fitness browser which provided the change in bacterial fitness following deletion of the GafAs.

2. 3. 3. 2 Deletion of Shewana3_2073 decreases growth fitness of Shewanella3_2073 on arabinose

The Fitness browser of MicrobesOnline provided the phenotype fitness of certain strains identified in this study following disruption of *gafAs* by transposon insertion (Wetmore *et al*., 2015). The Fitness browser is an online database which displays thousands of genome-wide fitness assays. The data in this database were collected using randomly barcoded transposons, each one inserted at a random location in the genome of a single strain. The uniqueness of the technique lies in the fact that each transposon includes a random barcode therefore the insertion location and the identity of the linked DNA bar code are detected by PCR using transposon specific primers followed by Illumina sequencing (TnSeq) (Wetmore *et al.*, 2015). The mutant fitness profiling was achieved by Barseq, during which the barcode was amplified by PCR and quantified by Illumina sequencing. The Barseq was performed twice for each barcode, once prior to the experiment in selective media *ie*. barcode abundancy of strain pregrown in rich media, and following termination of growth in selective media

Table 2. 8. Summary of the experimental evidence for GafAs to correlate to their ligand specificities

Protein name	RegPrecise	*Fitness values for growth phenotypes (t-scores) L-Arabinose	D-Galactose	(D- or L-) Fucose	D-Xylose	Experimental data
YtfQ	Crp/GalR/AraC	n/a	- 0.2	(L-) + 0.1	+ 0.1	ChIP- chip & ChIP- qPCR: AraR regulation[1]
Shewana3_2073	AraR	- 7.1 (-4.2)	n/a	n/a	n/a	n/a
CKO_03603	Crp/GalR/AraC	n/a	n/a	n/a	n/a	n/a
Cbei_4462	AraR	n/a	n/a	n/a	n/a	n/a
SM_b21587	n/a	- 0.4	- 0.1	(L-) 0.0	- 0.1	Reporter assays: induced by arabinose, fucose and talose[2]
A1D30_23855	n/a	- 3.7 (-2.5)	- 3.4 (-4.5)	0.0	0.0	n/a
Bphyt_4808	n/a	+ 0.4	0.0	(D-) - 0.2	0.2	n/a
Hsero_1033	n/a	- 0.3	- 0.5	(L-) + 0.3	- 0.5	n/a

1. Mauchline,T. et al. (2006) Mapping the Sinorhizobium meliloti 1021 solute-binding protein-dependent transportome. *Proceedings of the National Academy of Sciences*, 103, 17933-17938.
2. Stringer,A. et al. (2013) Genome-Scale Analyses of Escherichia coli and Salmonella enterica AraC Reveal Noncanonical Targets and an Expanded Core Regulon. *Journal of Bacteriology*, 196, 660-671

*Fitness = 0 means that mutants in this gene grew well as other mutants and probably about as well as wild type strains / Fitness < 0 means that the gene was important for fitness and the mutants were less abundant at the end of the experiment than at the beginning / Fitness > 0 means that the gene was detrimental to fitness and that mutants had a growth advantage.

(Wetmore *et al.*, 2015). This way the fitness of the mutant strains was equal to the $\log_2$ change in abundance during growth (typically 4 to 6 generations). The gene fitness was roughly the average of the fitness of the strains that have mutations within that gene. The values of the fitness of the genes were a measure of how well the mutant grew in comparison to itself prior the experiment initiation, to other mutants and to the wild-type (WT) strain. For example, a fitness equal to 0 meant that the gene had no effect on the growth, and that the mutant probably grew well as other mutants or similarly to WT. However, when the fitness was negative, that meant that the gene is important to the fitness of the mutant and that the mutants were more abundant at the start of the experiment that at the end. For example, fitness = -1 meant that mutants in the gene were half as abundant at the end of the experiment, compared to the start. Conversely, a positive fitness value designated that the gene was a burden to the fitness of the strain and its absence is benefiting growth at the given media. The gene loss was considered to exert a strong phenotype to the mutant strain if its fitness value were higher than +2 and smaller than -2. A second layer of confirmation was the statistical test of the reproducibility of the data obtained by the fitness values of all the mutants of a given gene. The test produced t scores, which indicated how reliably a gene fitness value is different from zero. Strong fitness phenotypes (-2> |fit| > +2) with t scores smaller than 4 for positive fitness values or higher than -4 for negative fitness values, were considered unreliable.

Essentially, any considerable decrease in the *gafA* fitness of a particular Δ*gafA*::Tn5 mutant strain when grown in a given carbon source would reflect how essential is the GafA in transporting the respective sugar. All the selective media included in the study, out of which a large percentage are based on minimal media supplemented with carbon source, are outlined in the study by Price and his colleagues (Price *et al.*, 2016). The fitness effects of the *gafAs* included in the Fitness browser are summarised in Table 2. No considerable fitness effects were observed when *gafAs* were disrupted in *E. coli* (ytfQ), *Sinorhizobium meliloti 1021* (sm_b21587), *Burkholderia phytofirmans PsJN* (bphyt_4808) and *Herbaspirillum seropedicae SmR1* (hsero_1033). These are expected results when *E. coli* is taken as a paradigm, as it possesses well characterised

ABC systems for the transport D-galactopyranose and L-arabinopyranose, MglABC and AraFGH respectively, as well as secondary transport systems. Therefore, if any negative effects on fitness were to be exerted by the loss of YtfQ, these are compensated by the function of the aforesaid systems. In the case of *Sinorhizobium meliloti* 1021, an ABC transporter system *ie*. AraABC (systematic identifiers SM_b20895, SM_b20894 and SM_b20893), has been previously shown to be required for intracellular accumulation of radiolabelled L-arabinose (Poysti *et al.*, 2007). This ABC system is confirmed to be participating in the L-arabinose uptake as SM_b20895 deletion causes a reduction in fitness of -4. 2 (t- score = -9. 9). Additionally, the study by Mauchline and his coworkers mentioned in Section *2. 3. 3. 2*, has identified gene operons for ABC transport systems inducible by galactose, including the *sma0203*-05 (induced 19-fold by galactose) and *smb21343*-45 (induced 6-fold by galactose). Even though none of the above systems caused a reduction in fitness of the *Sinorhizobium meliloti* 1021 on growth with D-galactose, it could be that growth deficiency is only presented in a double mutant or that a secondary transporter is compensating for the loss of either ABC systems. An ambiguous result was obtained in the case of *Acidovorax* sp. GW101-3H11; the deletion of A1D30_23855 (*ie*. GafA) had an apparent effect on the fitness of the bacterium when grown in L-arabinose albeit with somewhat unreliable t score (-2. 5). As mentioned in the Section *2. 3. 2. 1*, the predicted AraF SBP (*ie*. A1D30_23860) is co-transcribed with the *gafA* in the same operon (*ie*. *araFGH*), therefore we speculate that the presence of this arabinopyranose–specific SBP compensates for the loss of A1D30_23855 in the aforementioned mutant background and growth conditions. The participation of A1D30_23860 in the L-arabinose transport was verified by the Fitness browser, as it is shown to cause a reduction in fitness (*ie*. -4.5) with a highly reliable t- score (*ie*. -8.7). The A1D30_23855 is expected to be involved in D-galactose transport (fitness = -3.4, t- score = -4.5) and most notably, the A1D30_23860 is also active with an undisputed reduction in fitness (*ie*. -5.6, t score = -8.5) of the respective *Acidovorax* sp. GW101-3H11 mutant when grown in D-galactose. The negative fitness effect is not unforeseen, as the AraF ortholog from *E. coli* shows promiscuous activity in that it is also able to bind D-galactose, though with a 2-fold

weaker K_D than L-arabinose (Miller *et al.*, 1983). Such result points towards the high importance of this system in transporting arabinose as well as galactose and tempts one to speculate that this bacterium lacks the D-galactopyranose system, *ie*. MglABC. This wasn't further pursue as it falls out of the scope of the study. The most important result relevant to the scope of this study is concerned with the deletion of *shewana3_2073* which caused a notable decrease in gene fitness of *Shewanella* sp. ANA 3 when grown on arabinose. The t score (-4.2) of the experiments ascertained the above result, therefore underlining the importance of GafABCD from *Shewanella* sp. ANA 3 in transporting L-arabinose. Further search in list of the experiments failed to detect an effect of this deletion on growth with D-galactose. As a result the search was refocused in finding literature regarding the growth of *Shewanella* sp. ANA 3 on sugars including D-galactose.

2. 3. 3. 3 Shewanella sp. ANA 3 inability to grow on galactose implies that GafA doesn't transport galactofuranose

An additional evidence in the search for GafAs with higher likelihood to bind arabinofuranose over galactofuranose came from growth assays of the study contacted by Rodionov and his coworkers (Rodionov *et al*. 2010). Their analysis, similar to the current study, employed comparative genomic techniques, such as the analysis of conserved operons and regulons, and subsystems based approach to predict and experimentally verify carbohydrate utilization pathways in a group of 19 species from the genus *Shewanella*. They predict that the Shewana3_2073-76, identified here, is actively partaking in arabinose uptake. Most notably, they showed using growth experiments (*ie*. shaking flasks) that *Shewanella* sp. ANA-3 failed to grow on minimal media supplemented with D-galactose. They proceeded to confirm the latter phenotype using a microplate – based growth assay, which produced the same result. The inability of the strain to grow on D-galactose could be accounted on the absence of the catabolic pathway responsible for D-galactose breakdown, however these are suggested to be conserved in all *Shewanella* spp. examined. As far as the transport of is concerned, the GalP which is the secondary system for D-galactose uptake, was only

predicted to be present in a subset of species and *Shewanella* sp. ANA-3 wasn't one of them. They attributed this to gene loss throughout evolutionary history of the bacteria due to the features of the ecological niches they colonize (Rodionov *et al.*, 2010). Therefore, based on these data the Shewana3_2073 SBP is unlikely to recognise and bind D-galactofuranose.

2. 3. 4 Selection of GafA candidates for further characterization

The flowchart presented in Figure 2. 7 outlines the methodology of the process followed in the selection of GafAs which could potentially show smaller ligand binding specificity, and preferential binding of L- arabinofuranose over D- galactofuranose. The analysis showed that the **Shewana3_2073, Cbei_4462** and **SM_b21587** are strong candidates to satisfy the aims of the study. These will be referred to as **GafASm**, **GafACb** and **GafASw**, respectively, throughout the rest of the dissertation. Furthermore, the predicted arabinopyranose SBP from *Clostridium beijerinckii* NCIMB 8052, *ie*. Cbei_4450, was also included in the candidates as a negative control of arabinofuranose binding. The latter will be referred to as **AraFCb**.

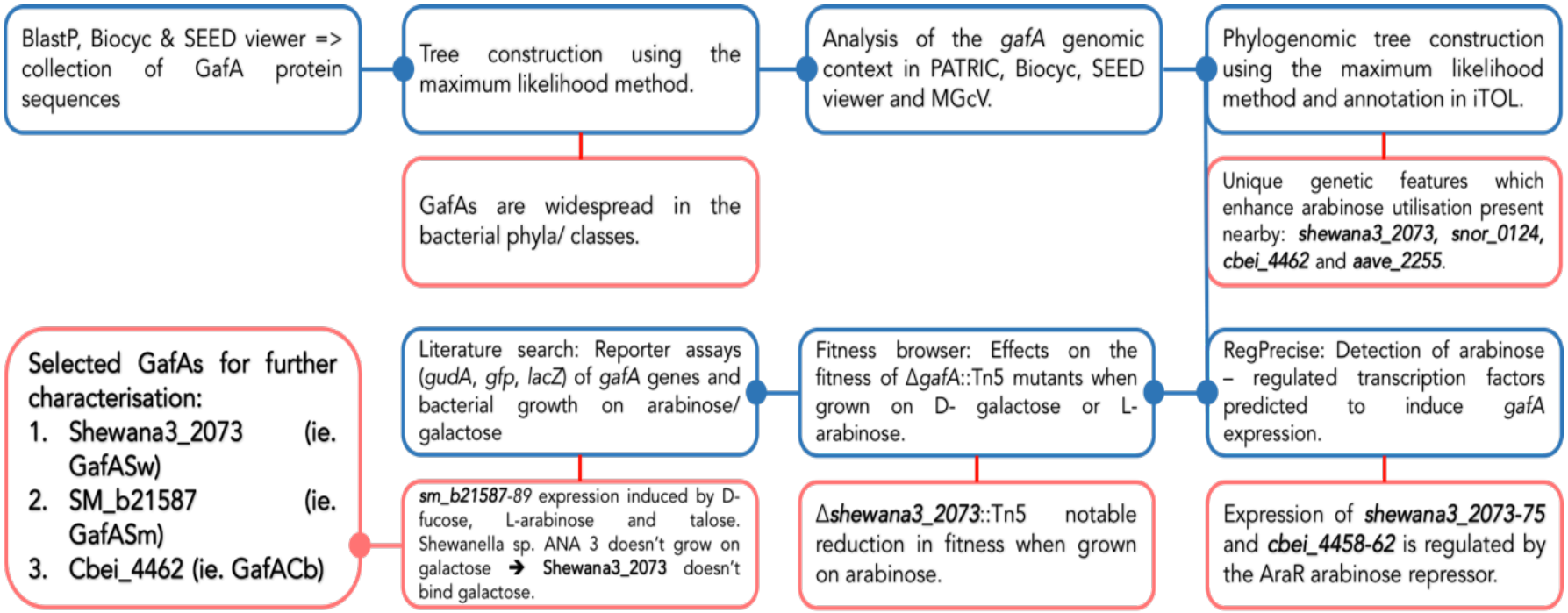


Figure 2. 7. Flowchart diagram to outline the methodology and the evidence collected which led to the selection of the GafAs.

The analysis performed in each step is described in the blue box, whereas the results of the analysis are outlined in the red boxes.

2. 3. 5 Cloning of *gafACb* and *araFCb* genes into overproducing vectors and expression trials

The first step followed in the methodology of characterizing SBPs is the cloning of the respective protein encoding genes in overexpression vectors. The genes are cloned in - frame with tags like hexahistidine to facilitate purification of the proteins. In the current study, the *gafACb* and *araFCb* sequences had their endogenous lipoprotein signal peptides removed and were codon – optimized for *E. coli*. They were initially cloned into the pET20b(+) vector to direct overproduction in the periplasm using the pelB leader sequence. Further efforts to achieve overproduction of GafACb, included subcloning of its coding sequence into the pBADcLIC and pMALp5X vectors. All plasmid constructs produced were tested in small scale expression trials for overproduction of the GafACb and AraFCb in various *E. coli* BL21 (DE3) strains including BL21 (DE3) Star and BL21 (DE3) Tuner cells.

2. 3. 5. 1 Strategy for periplasmic expression of gafACb and araFCb

The *gafACb* and *araFCb* were cloned in the pET20b vector first which belongs in the most commonly used protein overexpression systems *ie*. pET. The vectors of the pET expression systems drive successful overproduction via a T7 promoter, which is induced by the chromosomally expressed T7 RNA polymerase (Giordano *et al.*, 1989). More specifically, the pET20b (+) vector carries an N-terminal *pelB* sequence from *Erwinia carotova*, which serves for periplasmic localization of the overproduced proteins via the Sec- translocon system (Lei *et al.*, 1987). The choice to use this vector instead of the cognate pET21b (+), which allows for cytoplasmic expression, was based on the fact that both GafACb and AraFCb contain more than one cysteine residues so there is a chance that their tertiary structure features a disulphide bond. This speculation was enhanced by the fact that the crystallised structure of the orthologous GafAEc contains a disulphide bond (Horler *et al.*, 2009). As disulphide bond formation requires the oxidative conditions of the periplasm, the overproduction in the reductive cytoplasmic compartment is deemed unsuitable because it will lead to

protein misfolding and aggregation. The favourable oxidative conditions of the periplasm are not alone sufficient in mediating disulphide bond formation as the spontaneous protein oxidation is extremely slow and incompatible with cell growth and activity. Therefore, the oxidation of the thiol groups of the cysteine residues is catalysed enzymatically by the disulphide bond (Dsb) system present in the periplasmic compartment of *E. coli* (Kadokura, Katzen and Beckwith, 2003; Messens and Collet 2006). The enzymes of this system (*ie*. DsbA, B, C, D and G) facilitate the formation and rearrangement of the disulphide bonds in the mature polypeptide allowing the overproduced proteins to adapt their functional conformation in the periplasm (De Marco, 2009). Additional to the disulphide bond facilitation, the periplasm only contains a small proportion of the total cellular protein; therefore, the expression into this compartment results in a substantial purification upon periplasmic extraction.

As the periplasmic accumulation was ensured by including the *pelB* signal sequence, the design of the coding sequences to be synthesized involved detection and removal of the endogenous lipoprotein signal peptide as described in Materials and Methods Section 2. 2. 1. 3. The SBPs from the Gram (+) bacteria are also referred to as lipoproteins as the absence of a periplasmic space doesn't allow the existence of soluble periplasmic proteins. The lipoproteins are normally found anchored to the outer leaflet of the plasma membrane via a diacylglyceride. During the early stages of their production, they are localized in the cytoplasm as preprolipoproteins, which contain an N-terminal signal peptide. The N-terminus is recognized by the Sec or the Twin-arginine translocation (Tat) pathways which in turn facilitate translocation across the cytoplasmic membrane (Hutchings *et al.*, 2009; Thompson *et al.*, 2010). Following translocation, the –SH group of a conserved cysteine residue is modified by the transfer of a diacylglyceryl moiety from a glycerophospholipid; a reaction catalyzed by a prolipoprotein diacylglyceryltransferase (Braun and Wu, 1994). The cysteine residue is normally found immediately downstream of the signal peptide and is part of a conserved N-terminal lipoprotein sequence referred to as the lipobox motif (von Heijne, 1989). Later stages in the process include further reduction in the protein

length by a type II signal peptidase which cleaves the precursor immediately upstream of the lipid – modified cysteine. In the case of the *gafACb*, the first 22 amino acid residues were predicted to be a lipoprotein signal peptide with the conserved lipobox identified spanning from the 20th to 23rd residue (*ie*. LSG↓C) (Sutcliffe and Harrington, 2002) (Figure A2. 2A). The conserved Cys residue at the 23rd position is anticipated to be acting as the lipid anchor following the addition of the diacylglyceryl moiety. However, the prediction by the SignalP detected a longer signal peptide, and positioned the cleavage site between residues 28 and 29 (*ie*. SGA↓GG), and for that reason an additional 4 residues downstream the predicted lipoprotein were excluded (Figure A2. 2A). Regarding the *araFCb* sequence, the first 20 residues were predicted to be the lipoprotein signal peptide with the lipobox motif sequence predicted to be ^{18}IGG↓C^{21} (Cys_{21} predicted to be the conserved cys residue) (Figure A2. 2B). As the SignalP software predicted with high confidence the presence of a signal peptide formed by residues 1 – 32, an additional 11 residues downstream the predicted lipoprotein were also omitted from the coding sequence of the *araFCb* (Figure A2. 2B).

The resulting coding sequences were codon – optimized in Jcat for overexpression in *E. coli* (Grote *et al.*, 2005). The codon optimized sequences were then chemically synthesized and cloned into pET20b by Gibson assembly so that the coding sequence was in frame with a 3' nucleotide sequence encoding for a hexahistidine tag (Materials and Methods Section 2. 2. 3. 10). This resulted in two different vectors: pCD01 and pCD02, expressing an N-terminally *pelB* tagged and C-terminally hexahistidine tagged *gafACb* and *araFCb*. The successful cloning of the constructs in pET20b was verified by PCR screening and digestion by BsaI (Figure 2. 8A). The sequence of the inserts cloned were confirmed by sequencing. Expasy was used to calculate the expected monoisotopic mass of the proteins for the overproduction trials on the section below; these are 39.77 kDa for AraFCb and 34.89 kDa, including the pelB peptide.

2. 3. 5. 2 Small scale expression trials of gafACb and araFCb in pET20b

The produced pCD01 and pCD02 vectors were transformed into the expression strain *E. coli* BL21 (DE3) and used in small scale expression trials. Small-scale test expression is widely used as a predictive tool to determine which of the derivative clones actually produces soluble protein and to establish the optimal scale for the large-scale growth. The expression trials were performed initially at 37 °C growth temperature, and at later trials at 20 °C. Also, two other variants of the BL21 (DE3) strain, *ie*. Star and Tuner strains, were used in the expression trials in an attempt to achieve overproduction. The collected cell fractions were lysed and the protein content was resolved by SDS-PAGE. The expression trials at 37 °C were done as described at Materials and Methods Section 2. 2. 4. 1. Sampling of the culture occurred hourly during an incubation course of 4 hours after addition of 1 mM IPTG, including a final sampling at 24 hours. The proteins were not overproduced in the aforementioned conditions. However, the BL21 (DE3) cells overexpressing the constructs grew very well, reaching an OD_{600} of approximately 2.5 at 4 hours incubation following induction by IPTG. Therefore, the failure could not be attributed to toxicity caused by overproduction.

The second overproduction trial involved lowering the incubation temperature from 37 °C to 20 °C. This is a commonly used approach in the initial troubleshooting steps for protein overproduction, as lower growth temperature decreases the rate of protein production. Due to the reduced growth rate, the starting OD_{600} of the culture was doubled as compared to the 37 °C trials (*ie*. OD_{600} = 0.1). The culture was sampled at 0, 15 and 24 hours after induction with IPTG. The GafACb and AraFCb showed no accumulation throughout the incubation course (not shown).

In order to circumvent the inherent obstacles associated with overexpression of exogenous proteins in *E. coli*, the subsequent trials involved using different backgrounds of BL21 strains to optimise and maximise overproduction. The first trial attempted overexpression of the *gafACb* and *araFCb* in the BL21 (DE3) Star cells, which lack the RNAse E therefore stabilising the mRNA constructs of the *gafAs*. The overproduction was performed at 20 °C and the sampling occurred at the same

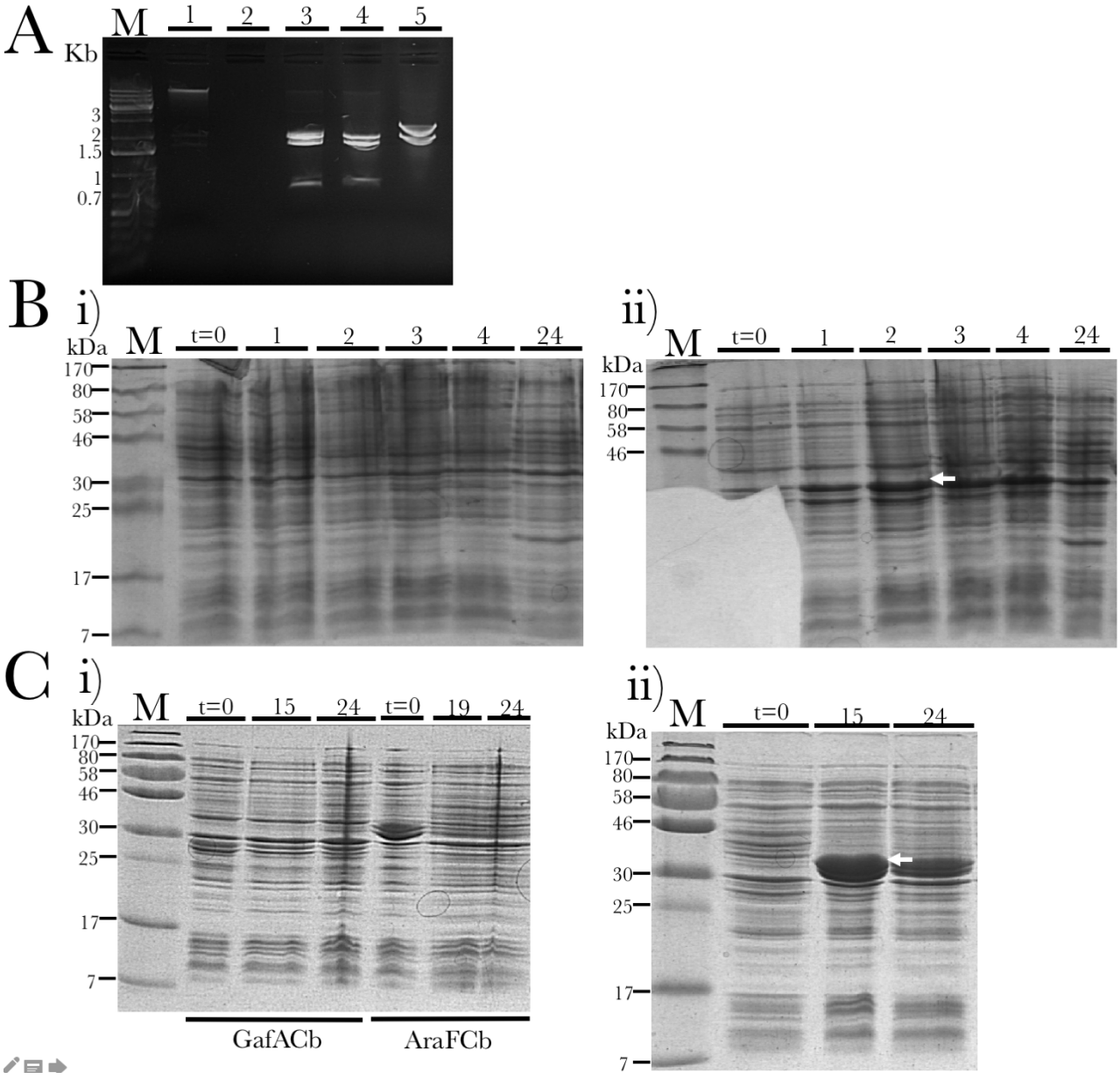

A
M
1
2
3
4
5
Kb
3
2
1.5
1
0.7
B
i)
kDa
M
t=0
1
2
3
4
24
170
80
58
46
30
25
17
7
ii)
M
t=0
1
2
3
4
24
kDa
170
80
58
46
C
i)
kDa
M
t=0
15
24
t=0
19
24
170
80
58
46
30
25
17
7
GafACb
AraFCb
ii)
M
t=0
15
24
kDa
170
80
58
46
30
25
17
7

Figure 2. 8. Screening for cloning and overexpression of gafACb and araFCb in pET20b

A) Agarose gel to verify recombinant pET20b plasmid with *gafACb* (pCD01) and *araFCb* (pCD02). The DNA agarose gel (1 %) stained with ethidium bromide, shows the products of restriction digestion of pCD01 (lanes 3-4) and pCD02 (lane 5). The plasmids, including the empty vector at lane 1, were digested with BsaI. The arrows represent the DNA fragments which contain full or partial sequence of the inserted gene. The BsaI restriction produced the following bands in the respective plasmids: pET20b/ Top band= 2041 bps, Bottom band= 1675 bps, pCD01/ 1st band = 2041 bps, 2nd band= 1822 bps and 3rd band= 818 bps, pCD02/ Top band= 2754 bps, bottom band= 2041 bps. M (Molecular marker) = Generuler plus 1 Kb.

B) i) SDS-PAGE of GafACb overproduction trial in BL21 DE3 at 37 °C. The collected fractions were sampled hourly on a 4 hours post-induction timecourse, including a final sampling at 24 hours. The cell lysates were resolved by Coomassie-stained 12% SDS-PAGE. The analysis showed that GafACb is not overproduced during the incubation course. ii) SDS-PAGE of AraFCb overproduction trial in BL21 DE3 at 37 °C. All conditions are the same as described above. There are low level of AraFCb overproduction, peaking at 2 hours following indution.

C) i) SDS-PAGE of the overproduction trials of GafACb and AraFCb in BL21 (DE3) Star cells at 20 °C. The collected fractions were sampled at 0, 15 and 24 hours following induction. The cell lysates were resolved by Coomassie-stained 12% SDS-PAGE. The analysis showed that neither GafACb nor AraFCb overproduced during the incubation course. ii) SDS-PAGE of the AraFCb overproduction trials in BL21 (DE3) Tuner at 20 °C. The sampling occurred at the same timepoints as above. The AraFCb showed high levels of overproduction at 15 hours following induction by IPTG.

The time the respective sample was collected in each instance is indicated above. M (Molecular marker) = Prestained Protein Marker, Broad Range NEB (7-175 kDa). The bands of the overproduced proteins are indicated with an arrow.

interval times mentioned above. The proteins didn't overproduced as they presented the same protein profile with the control (data not shown). Additionally, overexpression was also tested in BL21 (DE3) Tuner cells at the same conditions tested with the BL21 (DE3) Star cells. The Tuner cells are *lacZY* deletion mutants of BL21. The lac permease (LacY) mutation allows uniform entry of IPTG into in the population of the overproducing host cells, which produces a concentration-dependent and homogeneous level of induction. By adjusting the concentration of IPTG, expression can be regulated, thus a lower-level expression may enhance the solubility and activity of difficult target proteins. The AraFCb was overproduced at high levels at 15 hours following induction (Figure 2. 8Cii). However, the GafACb did not show any considerable level of overproduction (not shown).

2. 3. 5. 3 Cloning of gafACb into pBADcLIC and small scale expression trials

Following the unsuccessful overproduction trials of GafACb when expressed in pET20b, further efforts aimed to achieve overproduction of GafACb included subcloning of the coding sequence into the pBADcLIC vector. The pBAD system has been used successfully used previously to obtain high yield of soluble SBPs (Mulligan *et al.*, 2009). It allows addition of the decahistidine tag to either the N or C-terminus of the protein. The expression of the cloned construct is under the control of the P_{BAD} promoter. The P_{BAD} promoter, is derived from the *araBAD* operon, and is inducible by arabinose which lifts the AraC - mediated repression. The *araC* gene is ectopically expressed by pBADcLIC, ensuring that the uninduced levels are kept at a minimum. Therefore, the tight control of expression levels exerted by the *araC* - P_{BAD} promoter system helps to avoid the deleterious effects due to the uninduced expression of toxic genes (Guzman *et al.*, 1995). As GafACb is predicted to bind arabinofuranose, the use of pBADcLIC is not coincidental, since the presence of arabinose as an inducer could potentially stabilise the production of GafACb by accumulation of the bound form. The sequence was amplified from the pCD03 by PCR using the primers *gafACb*-pBADF and *gafACb*-pBADR to produce an amplicon which included upstream and downstream homologous sequences to pBADcLIC. The amplicon also included the

pelB sequence of the pET20b upstream the *gafACb* to target the overproduction in the periplasm, as pBADcLIC lacks a signal peptide N-terminal tag. Cloning of the produced amplicon into pBADcLIC by ligation independent cloning (see Materials and Methods, Section 2. 2. 3. 11) resulted in the production of the pCD03 expressing a *gafACb* tagged with a pelB sequence on its N-terminus and a decahistidine on its C-terminus. The insertion of the construct in the vector was confirmed by restriction digestion using NdeI (Figure 2. 9A).

The expression trials were performed at 20 °C growth temperature (Materials and Methods Section 2. 2. 3. 11). The trials included screening overproduction by BL21 (DE3), BL21 BL21 (DE3) Star cells. Sampling of the culture occurred at 0, 17 and 24 hours after addition of 10 mM arabinose. The unusually high concentration of arabinose used is to maintain the levels of inducer throughout the incubation course as part of it would inevitably be shuttled into the intracellular catabolism of the BL21 DE3 strains. The lysed fractions were resolved on 12% SDS-PAGE gel (Figure 2. 9B). The expression profiles showed that there wasn't any overproduction of the GafACb in any of the three strains tested (Figure 2. 9Bii). The similarity of the aforementioned expression profiles with the ones from the overproducing strains with the empty vector (Figure 2. 9Bi) was incredibly high. The only dissimilarity being an enlarged band presented at 17 hours following induction in BL21 Star cells. This is unlikely to be the GafACb as it seemingly has a lower size (<30 kDa).

2. 3. 5. 1 Cloning of gafACb into pMALp5X and small scale expression trials

Final efforts to overproduce GafACb included subcloning of the coding sequence into the pMALp5X vector. The *gafACb* was subcloned without the *pelB* sequence and a start codon as the cloning into pMALp5X creates N-terminal translational fusions with a natural affinity tag *ie*. *malE*. The *malE* encodes for the Maltose Binding Protein (MBP- 43.4 kDa), therefore there is no need for tagging the GafACb with a signal peptide as the fusion is directed in the periplasm by the MBP's leader peptide (Riggs, 2000). The subcloning was undertaken as described in the Materials and Methods

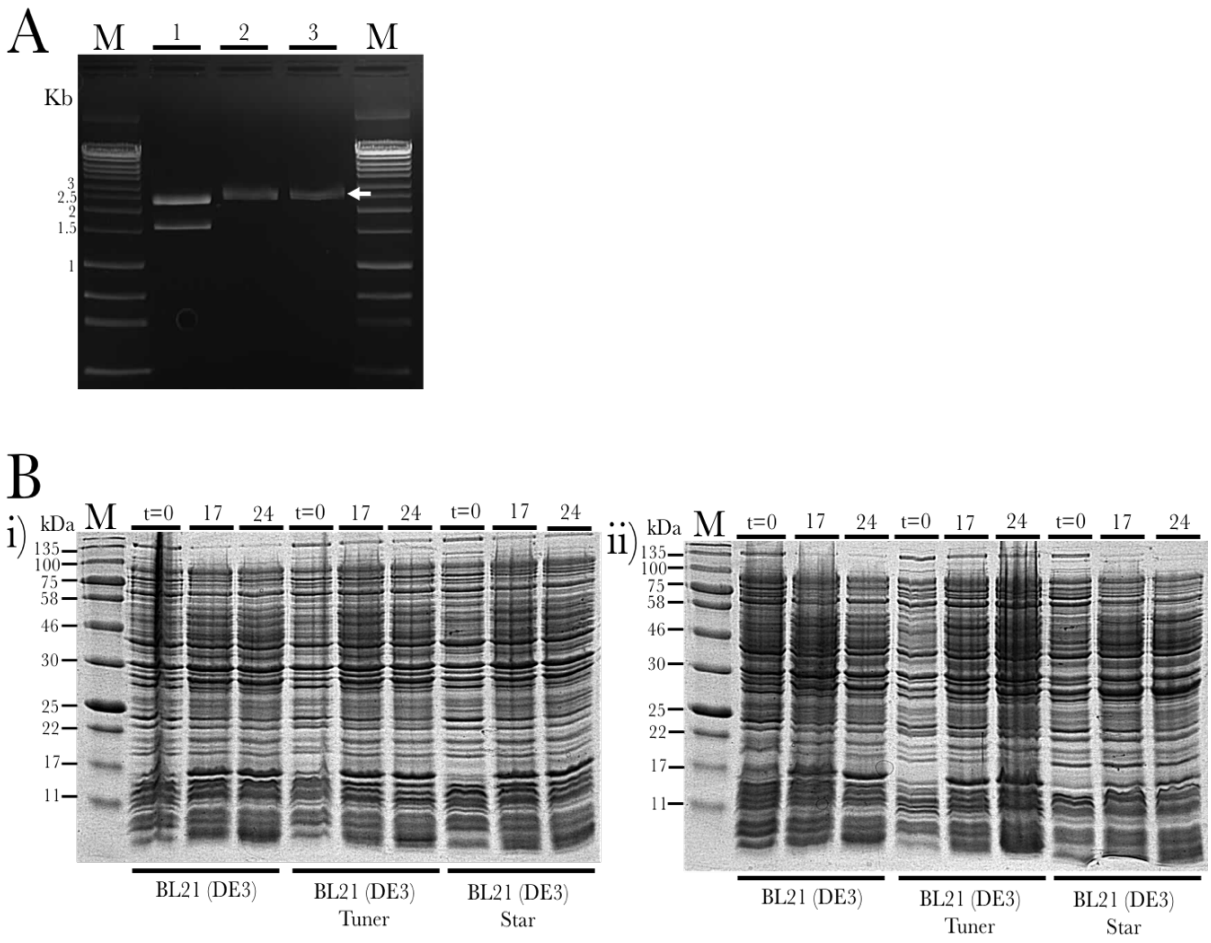


Figure 2. 9. Screening for cloning and overexpression of gafACb in pBADcLIC

A) Agarose gel to verify recombinant pBADcLIC plasmid with *gafACb* (pCD03). The DNA agarose gel (1 %) stained with ethidium bromide, shows the products of restriction digestion of pCD03 (lanes 2-3). The plasmids, including the empty vector at lane 1, were digested with NdeI. The arrows represent the DNA fragments which contain full or partial sequence of the inserted gene. The NdeI restriction produced the following bands in the respective plasmids: pBADcLIC/ Top band= 2481 bps, Bottom band= 1658 bps, pCD03/ Top band = 2642 bps, Bottom band= 2481 bps. M (Molecular marker) = Hyperladder I

B) i) SDS-PAGE of control expression trial of pBADcLIC in BL21 DE3, (DE3) Star and (DE3) Tuner at 37 °C. The collected fractions were sampled at 0, 17 and 24 hours following induction with 10 mM arabinose. The cell lysates were resolved by Coomassie-stained 12% SDS-PAGE. ii) SDS-PAGE of GafACb overproduction trial in BL21 DE3, (DE3) Star and (DE3) Tuner at 37 °C. All conditions are the same as described above.

The time the respective sample was collected in each instance is indicated above each gel. The BL21 DE3 strain tested at each occasion is indicated below the gels. 10 µl of sample were loaded in each well. M (Molecular marker) = Blue Prestained Standard, NEB (11-190 kDa). The bands of the overproduced proteins are indicated with an arrow.

Section 2. 2. 3. 9 and verified by restriction analysis using BsaI (data not shown). The expression trials were performed at 37 °C growth temperature (Materials and Methods *Section 2. 2. 4. 1*). The trials included screening overproduction by BL21 (DE3) and BL21 (DE3) Tuner cells. Sampling of the culture occurred hourly during an incubation course of 3 hours after addition of IPTG, including a final sampling at 24 hours. The lysed fractions were resolved on SDS-PAGE gels (Figure 2. 10). The *malE*::*gafACb* gene fusion produced large amounts of recombinant fusion protein (72 kDa) in both strains tested, reaching maximum yield at 3 hours following induction by IPTG at both BL21 DE3 (Figure 2. 10A) and Tuner cells (Figure 2. 10B). Unlike the typical overproduction by pET vectors, the high levels of the fusion protein are retained at 24 hours following induction. This is attributable to the fact that MBP offers extra protection to its downstream passenger protein from proteolytic degradation. Dissimilar to its intrinsic chaperone activity, its protective properties haven't been experimentally determined however they have been confirmed by studies which employed this tag (Rosano and Ceccarelli, 2014; Nallamsetty and Waugh, 2006).

2. 3. 6 Cloning of GafASw and GafASm and expression trials

The remaining selected GafAs, *ie*. GafASw and GafASm, were cloned in pET20b vector in a similar fashion to the GafACb and AraFCb. The optimization of the overproducing conditions involved screening their expression in the *E. coli* BL21 DE3 cells and the Star and Tuner aforesaid derivative strains.

2. 3. 6. 1 Cloning of GafASw and GafASm into pET20b

GafASm was cloned into pET20b using the traditional restriction-ligation technique (Materials and Methods Section 2. 2. 3. 9) as its sequence was amplified by PCR from the *Sinorhizobium meliloti* 1021 genome (supplied by Prof. Peter Young, University of York). The primers were designed in such way that the native signal peptide as predicted by SignalP (Figure A2. 2D) was excluded from the PCR amplification. As the gene wasn't synthetically produced the coding sequence wasn't codon optimized for

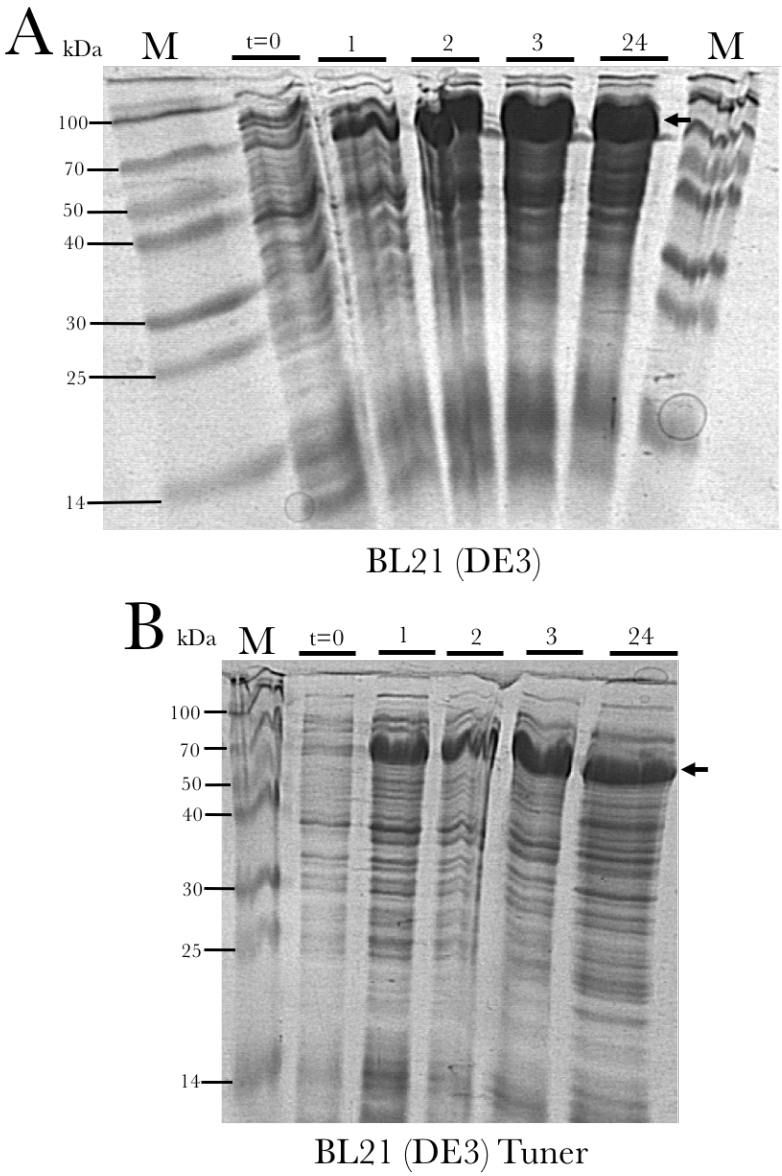


Figure 2. 10. Successful overproduction of GafACb fused to MBP

A) SDS-PAGE of the *malE*::*gafACb* overexpression trials in BL21 DE3 cells at 37 °C. The fractions were collected every one hour throughout a 4 hours post-induction timecourse, including a final sampling at 24 hours.. The cell lysates were resolved by Coomassie-stained 12% SDS-PAGE. The analysis showed that the MBP-GafACb fusion protein (72 kDa) reaches high overproduction levels at 3 hours, which are retained at 24 hours possibly because of the protection against proteolysis gained by the MBP fusion.

B) SDS-PAGE of the *malE*::*gafACb* overexpression trials in BL21 (DE3) Tuner cells at 37 °C. All conditions are the same as described above. The analysis showed that the MBP-GafACb fusion protein (72 kDa) reaches high overproduction levels at 3 hours following induction with IPTG.

The time the respective sample was collected in each instance is indicated above each gel. 10 µl of sample were loaded in each well. M (Molecular marker) = Smart Multi Color Pre-stained Protein Standard (14-120 kDa). The bands of the overproduced proteins are indicated with an arrow.

expression in *E. coli*, conversely to the strategy followed with the cloning of *gafACb* and *araFCb*. Despite the fact that the overproducing strain and the *S. meliloti 1021* are distantly related bacteria which belong to different branches of the Proteobacteria, GafAEc and GafASm show very high sequence identity (75%). Furthermore, *E. coli* has been previously used multiple times with great success in the overproduction of ABC SBPs from *Sinorhizobium* spp. without the requirement for codon optimization (Jebbar *et al.*, 2005; Oswald *et al.*, 2008). The obtained positive transformants were screened for the insertion of the constructs using restriction by BsaI (Figure 2. 11A). The cloning of *gafASw* followed the exact same strategy as the one with the *gafACb* and *araFCb* as described in Section 2. 3. 5. 1. Only difference was the prediction of the native signal peptide which happened in SignalP only (Figure A2. 2C), as the *Shewanella* spp. are Gram (-) bacteria and thus there wasn't a need to use the LipoP or PRED-LIPO softwares. Verification of the obtained plasmid was done with BsaI restriction (Figure 2. 11A). Following verification of the cloned constructs by sequencing, the produced plasmids were referred to as pCD05 and pCD06, which expressed an N-terminally *pelB* tagged and C-terminally hexahistidine tagged versions of *gafASm* and *gafASw,* respectively.

2. 3. 6. 2 Small scale expression trials of GafASw and GafASm

The produced pCD05 and pCD06 vectors were transformed into the expression strains *E. coli* BL21 (DE3), Star and Tuner and used in small scale expression trials. The conditions used in the expression trials are as described for *gafACb* (Section 2. 3. 5. 3). The collected cell fractions were lysed and resolved by 12% SDS-PAGE. The analysis showed that the GafASw was overproduced in BL21 Star cells prior to induction with IPTG, possibly because of the leakiness of the T7 promoter (Figure 2. 11Bi). However, in the case of the BL21 Tuner cells the overproduction levels are healthier for both of the proteins as they are reaching high levels at 15 to 24 hours following induction (Figure 2. 11Bii).

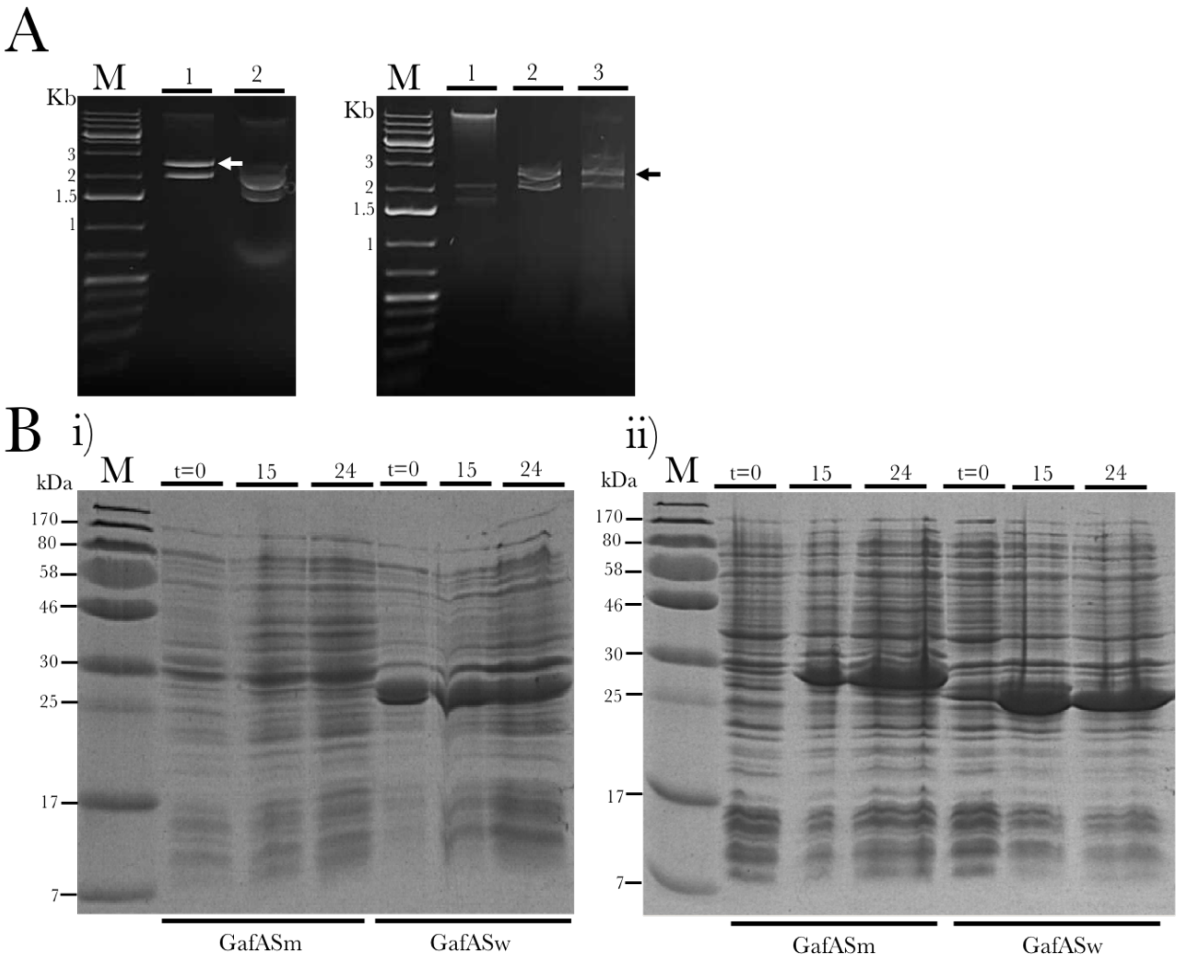


Figure 2. 11. Screening for cloning and overexpression of gafASm and gafASw in pET20b

A) DNA agarose gel (1%) to verify recombinant pET20b plasmid with *gafASm* (left) and *gafASw* (right). The gels stained with ethidium bromide, shows the products of restriction digestion of pCD05 (lanes 1, left gel) and pCD06 (lane 2 and 3, right gel). The plasmids, including the empty vector at lane 1 of the left gel and lane 2 of the right gel, were digested with BsaI. The arrows represent the DNA fragments which contain full or partial sequence of the inserted gene. The BsaI restriction produces the following bands in the respective plasmids: pET20b/ Top band= 2041 bps, Bottom band= 1675 bps, pCD05/ Top band = 2605 bps, Bottom band= 2041 bps, pCD06/ Top band= 2595 bps, Bottom band= 2041 bps

B) i) SDS-PAGE of the overproduction trials of GafASm and GafASw in BL21 (DE3) Star cells at 20 °C. The collected fractions were sampled at 0, 15 and 24 hours following induction. The cell lysates were resolved by Coomassie-stained 12% SDS-PAGE. The analysis showed that only GafASw shows overproduction albeit at low levels. ii) SDS-PAGE of the overproduction trials of GafASm and GafASw in BL21 (DE3) Tuner at 20 °C. The sampling occurred at the same timepoints as above. Both of the trialed proteins showed overproduced at moderate levels at 15 and 24 hours following induction by IPTG.

The time the respective sample was collected in each instance is indicated above. The proteins overproducted at each instance are indicated below the gels. 10 µl of sample were loaded in each well. M (Molecular marker) = Prestained Protein Marker, Broad Range NEB (7-175 kDa). The bands of the overproduced proteins are indicated with an arrow.

2. 3. 7 Large scale expression trials of GafASm and GafASw

Following the identification of optimal conditions for the overproduction of GafASm and GafASw, their production was scaled up in larger volume of media (Materials and Methods Section 2. 2. 4. 2) so as to obtain adequate amounts for the biochemical and structural characterisation of the proteins. The periplasmic extraction from the harvested cellular mass occurred by ice cold osmotic shock, with or without lysozyme. Thereafter, the overproduced proteins in the obtained periplasmic extracts were purified using gravity flow columns or an ÄKTA FPLC purifier machine.

2. 3. 7. 1 Identification of two protein species following the purification of GafASm

The GafASm production was scaled up in the optimised conditions as predetermined by the small scale expression trials (Section 2. 3. 6. 2). The large scale expression was performed using a volume ranging from 0.8 to 1L of LB media. By the end of the incubation course (*ie*. 17 – 20 hours following induction with IPTG), the culture was centrifuged and the cellular mass was collected. The periplasmic content including the overproduced proteins were released by the osmotic shock treatment. The principle behind the periplasmic extraction is the application of a chemical or mechanical pressure to cause the disruption of the outer membrane and the peptidoglycan cell wall and consequently release of the periplasmic contents. The products of the process are a) the supernatant which contains the periplasmic extracts and b) the remaining pelleted fraction which is enriched in vesicles with intact cytoplasmic membrane *ie*. spheroplasts. Excessive care is taken following peptidoglycan removal as the spheroplasts are susceptible to rupturing due to mechanical pressure *ie*. centrifugation. Some of the previously employed methods include the osmotic shock (Neu and Heppel, 1965), lysozyme – EDTA treatment (Malamy and Horecker, 1964), polymyxin digestion (Cerny and Teuber, 1971) and the chloroform extraction (Ames *et al.*, 1984) which all have successfully shown to reach up to 90% recovery of the target proteins (Donohue-Rolfe and Keusch, 1983; Chen *et al.*, 2004; Chen *et al.*, 2005). The polymyxin digestion technique was avoided, as it has

previously been shown to release cytoplasmic proteins (Teuber and Bader, 1976; Dixon and Chopra, 1986). Since the cytoplasmic proteins outnumber the ones residing in the periplasm by approximately a 10–fold (Ishihama *et al.*, 2008), even a low level of release would still cause substantial contamination. In the case of the chloroform treatment, a study has shown that it releases the higher total quantity of proteins from *E. coli* in comparison to cold osmotic shock (Lall, Eriho and Jay, 1989). Also, the activity of 6 periplasmic localized proteins in the extract was assessed and the results were assumed to be proportional to their relative abundancy in the periplasmic extract. The enzymes released from the chloroform treatment retained the highest activity in *E. coli* as compared to the cold osmotic shock. However, the same study never assessed the level of contamination by cytoplasmic proteins and therefore the clarity of the technique cannot be ascertained (Lall, Eriho and Jay, 1989). Additionally, the presence of chloroform in the periplasmic extract could potentially cause denaturation of the overproduced protein, as this organic solvent has been previously proved to be able to do so (Asakura, Ada and Schwartz, 1987). Since retaining the quality of the overproduced proteins was essential to their downstream characterisation studies, the chloroform treatment was also avoided. This study focused on using the osmotic shock method instead, on its own or in combination with lysozyme treatment. More specifically, the periplasmic extractions were done following either the Ice Cold Osmotic Shock (ICOS) or the Tris-Sucrose-EDTA (TSE) method. The first step of the ICOS is very similar to TSE, as they both include treatment of the harvested cells with Tris-HCl (pH=8), sucrose and EDTA during incubation on ice. Only difference is the additional step in ICOS which involves treatment of the pellet, obtained from the first step, with lysozyme in the presence of divalent cations (*ie*. Mg^{2+}). These methods are extensively explained and compared in the *Section 2. 3. 8* below.

Following periplasmic extraction of the GafASm with ICOS or TSE, the extract was dialysed in Phosphate Buffered Saline (pH= 7.5) or NaCl-Tris (pH= 7.5) and purified using an ÄKTA FPLC purifier or by gravity flow columns as described in the Materials and Methods Section 2. 2. 6. SDS-PAGE analysis of the purified fraction of GafASm which was extracted with ICOS, contained considerable level of contaminants and two

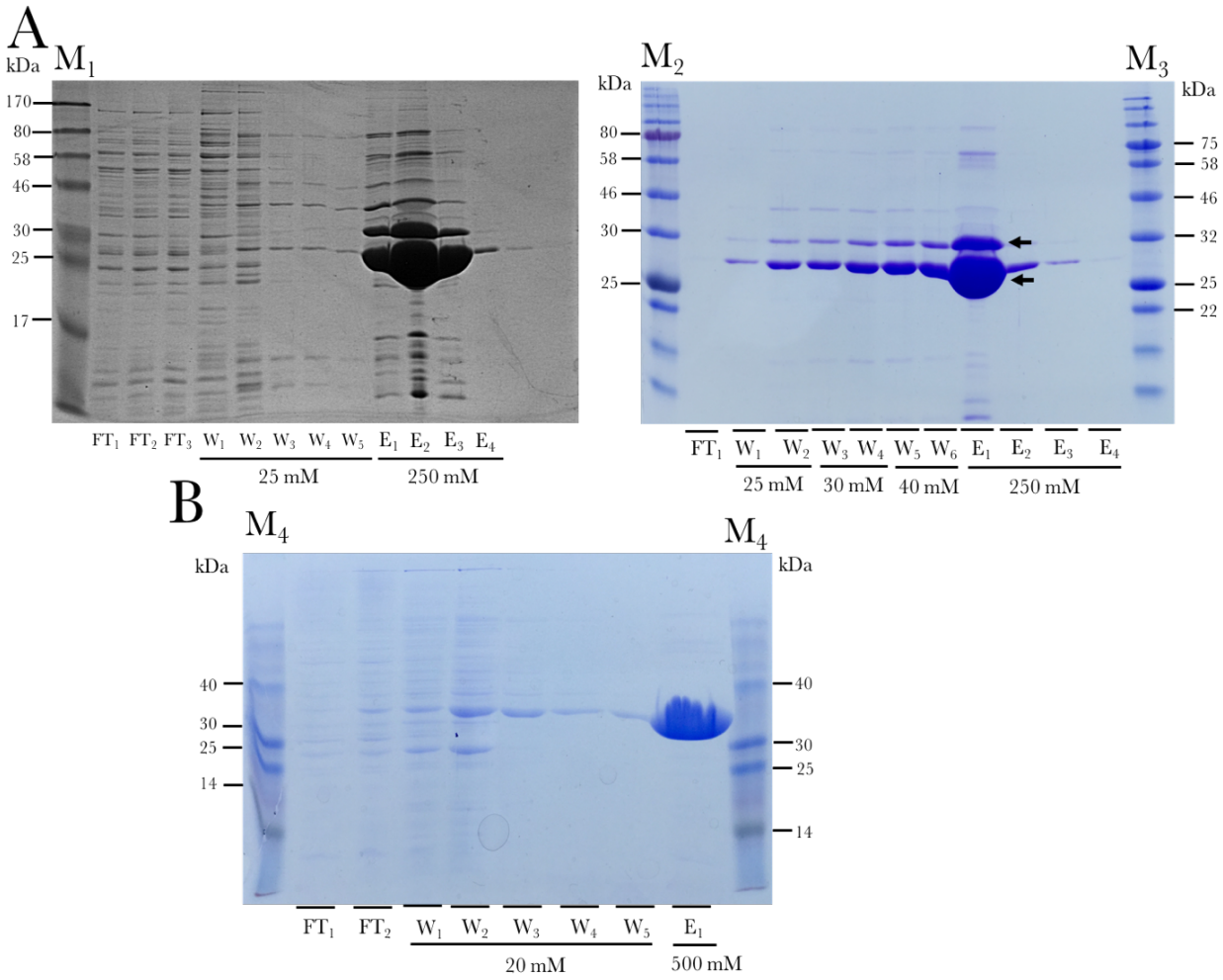


Figure 2. 12 Purification of GafASm from cultures of BL21 (DE3) Star cells

A) (Left) Coomassie stained SDS-PAGE of the first purification run of ICOS - extracted GafASm by Ni^{2+} affinity chromatography showing the fractions from the flow through, the washing and the elution steps. The high level of contaminants co-purified with the GafASm necessitated a second purification run. (Right) Coomassie stained SDS-PAGE of the second successive purification run of GafASm by Ni^{2+} affinity chromatography showing the fractions from the flow through, the washing and the elution steps. The run increased the purity of the elute. Two protein species are present in the purified fraction (indicated with arrow) which are hypothesized to be a pool of fully processed GafASm and its counterpart with partially cleaved or fully unprocessed pelB sequence.

B) SDS-PAGE gel of the purification run of TSE - extracted GafASm by Ni^{2+} affinity chromatography. The collected fractions from the flow through, the washing and elution steps are indicated. The level of contamination in the purified fraction is minimal and considerably lower compared to the ICOS - extracted fraction.

FT = flow through, W = washing steps, E = elution steps. The concentration of imidazole used is designated below the washing and elution steps.

Molecular markers (M) used: M1 = , M2 = Color protein standard (broad range, 11 - 245 kDa), M3 = Blue protein standard (broad range, 11 - 190 kDa) , M4 = Smart Multi Color Pre-Stained Protein Standard (14 - 120 kDa)

overproduced species are clearly distinguished (Figure 2. 12A). The acquired fraction underwent an additional purification run with increasing concentrations of imidazole in the washing buffer so as to eliminate any non-specific proteins binding to Ni-NTA agarose and subsequently decrease the level of contaminants. Indeed, the contamination was reduced, however part of the GafASm overproduced protein was eluted during the washing steps with 30 – 40 mM of imidazole (Figure 2. 12A). Therefore, in the future purification runs the imidazole concentration of the wash buffer was kept at the lowest value tested (*ie.* 25 mM), which wasn't sufficient to compete with the overproduced protein for binding to the Ni^{2+} ions. Notably, both of the bands with the highest intensity from the first purification run were retained in the eluted fraction of the second purification run. Therefore, there is a strong possibility that both species are His-tagged as they were co-purified at both successive runs. The predominant explanation to this observation would be the presence of an uncleaved form of GafASm with a larger mass than the processed form. This hypothesis is enhanced by previous studies which showed that a specific group of cytoplasmic proteins (Jacobson and Rosenbusch, 1976; Ajouz *et al.*, 1998; Yaagoubi *et al.*, 1994; Berrier *et al.*, 2000) and recombinant proteins targeted in the cytoplasm were released following treatment with the osmotic shock method (Thorstenson *et al*., 1997; LaVallie *et al.*, 1993). Additionally, the first step of the ICOS renders the peptidoglycan more susceptible to lysozyme activity as the EDTA acts on the lipopolysaccharide and increases the permeability of the outer membrane. Strikingly, treatment with the same concentration of lysozyme from egg white has been shown to be able to permeabilise the outer membrane of *E. coli* completely and it also exerted a time-dependent permeabilisation effect on the inner membrane. The presence of two protein species was confirmed by FT- ICR –MS (Section 2. 3. 8) which led to screening for an appropriate variation of cold osmotic shock method at later stages in the study. Unlike in ICOS, when the extract derived from the TSE method was purified it was much cleaner and didn't require a second purification run; this could be potentially attributed to the absence of lysozyme.

2. 3. 7. 2 Large scale overproduction of GafASw

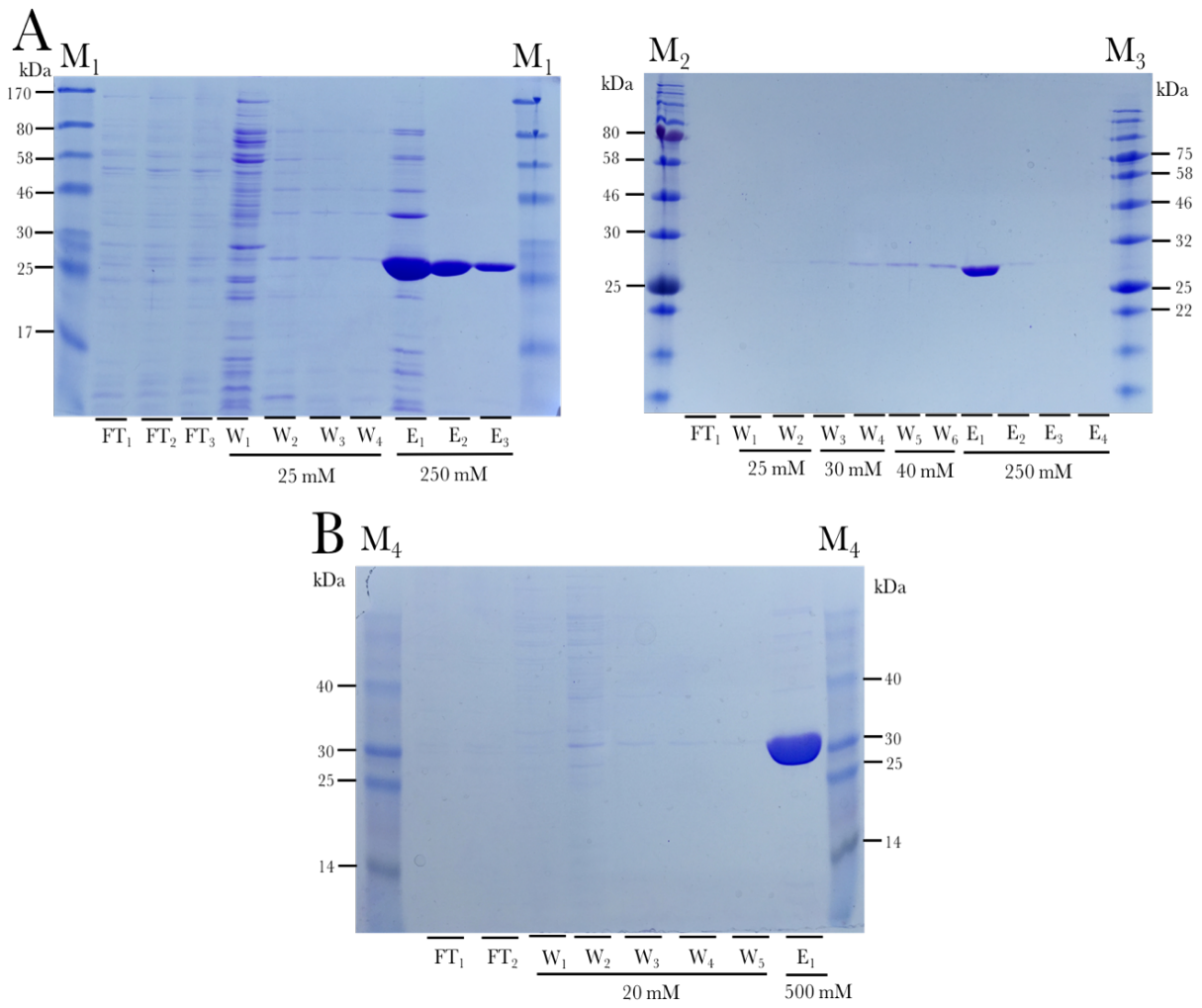


Figure 2. 13 Purification of GafASw from cultures of BL21 (DE3) Star cells

A) (Left) Coomassie stained SDS-PAGE of the first Ni^{2+} - based purification run of ICOS - extracted GafASw. The fractions obtained from the flow through, the washing and the elution steps are indicated. The ICOS periplasmic extraction releases high level of contaminants which were co-purified with the GafASw and led to a second purification run. (Right) Coomassie stained SDS-PAGE of the second successive purification run of GafASw by Ni^{2+} affinity chromatography showing the fractions from the flow through, the washing and the elution steps. The run increased the purity of the elute considerably with no contaminants present.

B) SDS-PAGE gel of the purification run of TSE - extracted GafASw. The collected fractions from the flow through, the washing and elution steps are indicated. The level of contamination in the purified fraction is negligible and considerably lower compared to the first purification run of the ICOS - extracted fraction.

FT = flow through, W = washing steps, E = elution steps. The concentration of imidazole used is designated below the washing and elution steps.

Molecular markers (M) used: M1 = Prestained Protein Marker, Broad Range NEB (7-175 kDa), M2 = Color protein standard (broad range, 11 - 245 kDa), M3 = Blue protein standard (broad range, 11 - 190 kDa) , M4 = Smart Multi Color Pre-Stained Protein Standard (14 - 120 kDa)

The GafASw was overproduced at the same conditions as GafASm described above. Following periplasmic extraction by ICOS or TSE the acquired fraction was purified by on a Ni^{2+} based column as described at Materials and Methods Section 2. 2. 6. The purification of the ICOS – derived extract (Figure 2. 13Ai) was shown to contain higher contamination than the TSE (Figure 2. 13B) as similarly to the purified fractions of GafASm (Figure 2. 12A,B). Thus, the purified fraction from the ICOS extract underwent an additional Ni^{2+} based purification step, which increased the purity (Figure 2. 11Aii). As this was also the case for the GafASm, it became a standard that the excessive amounts of contaminants in the eluted fractions were removed by extra runs of Ni^{2+} - based purifications.

2. 3. 8 FT- ICR - MS analysis of purified SBPs: Detection of the undefined form of GafASm in the purified fraction.

Fourier-transform ion cyclotron resonance (FT-ICR-MS) was used to ensure that the recombinant protein produced and purified is the one expected and that is not modified in any way during synthesis. The SDS-PAGE runs of the purified fraction of GafASm showed that there are two species present with very similar size (Figure 2. 14A); FT-ICR-MS analysis was used to confirm existence of both protein species in the purified fraction. The accuracy of this technique results from the extremely high resolution achievable (mass resolving power of up to $m/\Delta m_{50\%} \approx 400{,}000$ at full width half maximum) (Marshall and Hendrickson, 2008). At these resolutions it is possible to measure the monoisotopic masses of proteins to well within half a Dalton.

The technique was firstly used to study the purified fraction of GafASm. As shown in Figure 2. 10, there seem to be two protein species co-purified following Ni^{2+} based column purification. The accurately measured mass of the sample verified the existence of the three protein species in the fraction. As the protein species present are predicted to be the same protein due to the similarity in between this assures a low response factor and therefore the abundance of the species can be approximated

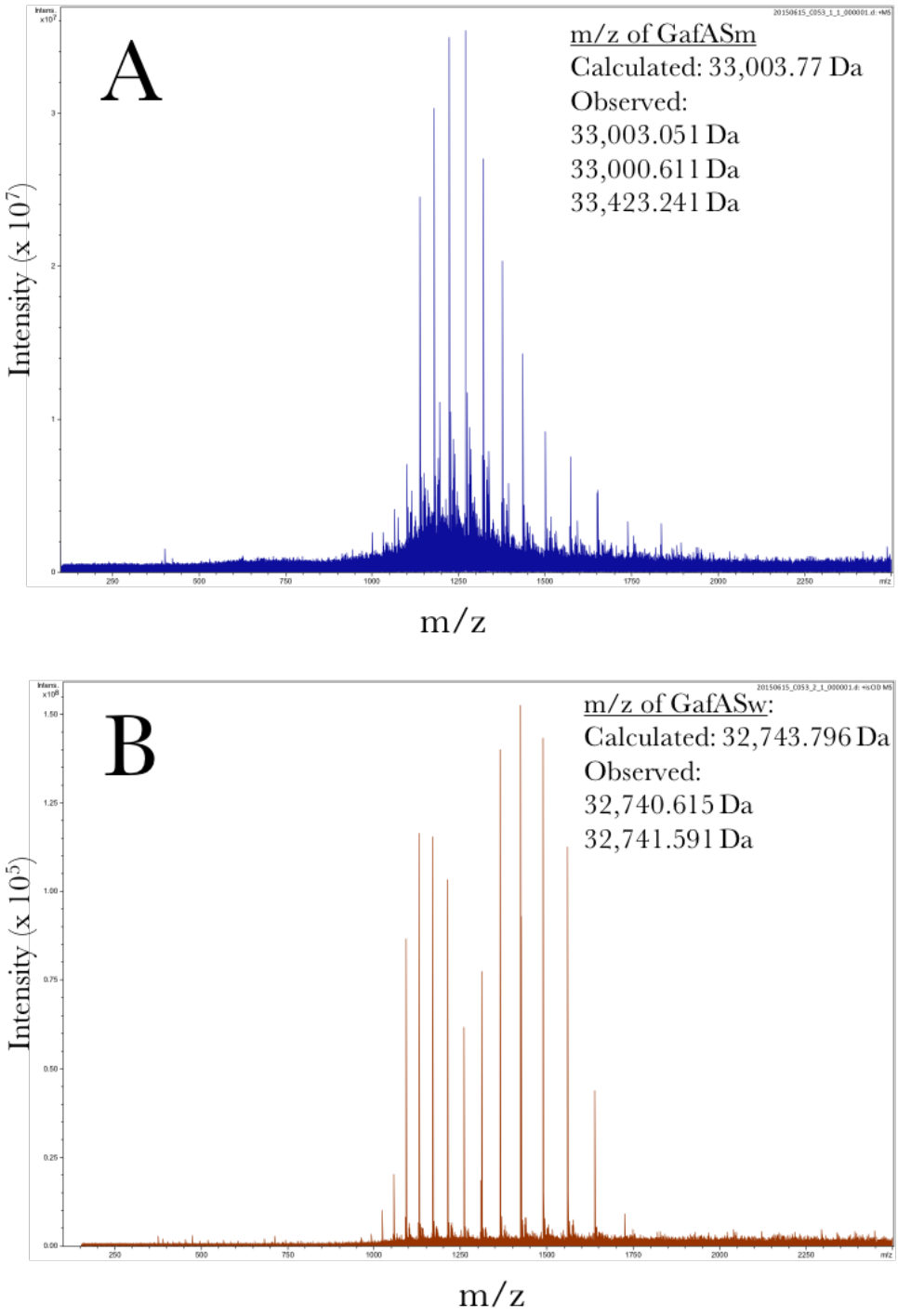


Figure 2. 14. FT-ICR-MS analysis of purified GafASm and GafASw.

A) FT-MS spectrum of 60 µM GafASm. The presence of two protein species (*ie*. 33,003.05 and 33,423.241 *m/z*, mass errors = 0.719 and 419.47 Da respectively) was detected.

B) FT-MS spectrum of 60 µM GafASw. The analysis confirmed the presence of a single protein species (ie. 32,740 and 32,741.591 m/z, mass errors = -3.18 and -2.2 Da respectively). The protein sequence used to calculate the monoisotopic masses of the two proteins are shown.

by their relative percentages. The obtained mass of the most abundant species *ie*. 33003.051 Da was in agreement with the expected monoisotopic mass of the pelB - cleaved GafASm (Figure 2. 14A). There was also an additional signal detected for a species with a mass of 33423.241 *m*/*z*. The large mass difference between expected and this mass could point towards the presence of a large adduct. Searches in the Unimod online database of protein modifications for mass spectrometry failed to detect an adduct modification whose size would justify the difference. A more logical explanation is the presence of a fraction of unprocessed GafASm with longer sequence *ie*. pelB not fully processed. We anticipate this to be the result of accumulation of GafASm with partially cleaved or completely uncleaved pelB sequence. To support this hypothesis, the search for potential uncleaved species with calculated monoisotopic mass similar to the observed one (*ie*. 33423.241 Da), identified the ↓*QPAMA*-GafASm with a mass of 33430.8 Da. The latter species represents the mature protein with 5 amino acids from pelB retained attached to the N-terminus due to non-canonical cleavage. This is investigated below by screening different periplasmic extraction methods of the overproduced proteins on the basis that unprocessed species could be accumulating in the cytoplasm and therefore a more sensitive periplasmic extraction method would isolate only periplasmic components. The same analysis with GafASw confirmed the presence of the correct protein species with a mass error equal to 3.15 for the major component (Figure 2. 14B).

2. 3. 9 Screening of periplasmic extraction methods to avoid purification of the undefined form of GafASm.

The existence of two protein species following large scale production and purification of GafASm was confirmed using FT-ICR-MS. The protein species with the larger size (33423.241 Da) is expected to be GafASm with partially uncleaved pelB sequence. This is pointing towards leakage of cytoplasmic components during a supposing periplasmic extraction process, and thus these species appear up following purification. To solely isolate the processed form of GafASm, and avoid any implications on functionality caused by the retained pelB sequence, the next

experiments focused on screening different periplasmic extraction methods aiming to detect a sensitive one which does not disrupt the bacterial inner membrane.

2. 3. 9. 1 Comparison of three variations of the osmotic shock method for the periplasmic extraction of GafAEc

The periplasmic extraction of the overproduced proteins was performed using the ice cold osmotic shock method, with or without the lysozyme treatment. Three different variations of the method were tested and compared for their ability to release the protein of interest in a relatively clean extract. The initial periplasmic extraction runs involved using the STE extraction buffer following cell culture harvesting by centrifugation. STE buffer is a Sucrose rich solution containing Tris buffer and EDTA. The role of sucrose is mainly to rise the extracellular osmolality which causes the cells to shrink and release water along with periplasmic contents into the surrounding medium. Additionally, sucrose can go through the outer membrane to enter the periplasmic space but is unable to penetrate the inner membrane. The presence of sucrose at the water–inner membrane interface presumably helps to stabilize the membrane (Strauss and Hauser, 1986) and therefore prevents cell lysis. Sucrose is also a well-known protein stabilizer, protecting the released proteins. EDTA facilitates periplasmic extraction by chelating divalent ions, which normally stabilize the lipopolysaccharide (LPS) in the outer membrane, resulting in LPS release and increased permeability of the outer membrane (Leive, 1965). Also, chicken egg lysozyme powder is added in the mixture to degrade the thin peptidoglycan wall of the *E. coli*. The above method was applied at multiple instances for isolation of the periplasmic fraction following overproduction. However, the yield of the released protein wasn't adequate at some occasions or the purity wasn't as high as expected, probably because of the release of cytoplasmic contents from a fraction of the cell mass. This would be a plausible explanation for the release of the unprocessed GafASm form with the uncleaved pelB sequence retained. If indeed leakage of cytoplasmic contents is feasible followed periplasmic extraction, it would be attributed to the use of lysozyme which disrupts the peptidoglycan and renders the inner membrane more

susceptible to lysis. This propagated further investigation of different variations of osmotic shock methods for isolation of the periplasmic fraction. Two additional methods were investigated out of which one excludes the lysozyme treatment:

a) Ice-cold osmotic shock (ICOS): the method is very similar to the one described above, however with increased levels of sucrose (0.5 M to 0.75 M). The EDTA concentration remains the same (5 mM) whereas the Tris-HCl concentration is decreased (50 mM to 20 mM). In contrast to the STE method, the pellet is treated firstly with the extraction buffer and following an extra centrifugation step, 40 µl of lysozyme (15 mg/ml) is added per gram of culture (approximate final concentration is 0.15 mg/ml). The treatment with lysozyme happens in presence of MgCl2 (5 mM) as cations decrease the efficacy of lytic activity thus avoiding complete cellular lysis (Boland *et al.*, 2004). Another discrepancy to the STE method is that the lysozyme treatment occurs on ice for a shorter time as opposed to 30 °C. The ICOS was firstly described with minor modifications to its current protocol by Witholt and his colleagues to produce spheroplasts from *E. coli* at various stages of growth (Witholt *et al.*, 1976).

b) TSE method: a modified method of the ice-cold osmotic shock used by Quan and colleagues. The method was claimed to give the purest and cleanest periplasmic fraction. They used mass spectrometry to analyse the contents of the periplasmic fraction following TSE extraction and the results of this analysis showed that the average signal coming from cytoplasmic and inner membrane proteins is 13-fold lower than the average signal of outer membrane proteins and 23-fold lower than that of periplasmic proteins (Quan *et al.*, 2013). The group also showed that amongst the 100 most abundant proteins, only eight were cytoplasmic; the rest are either outer membrane proteins (21%) or periplasmic proteins (71%). The method uses a similar extraction buffer to the above two methods though, Tris is added at 5 to 10-fold increased concentration and EDTA at lower levels. A unique feature of this method is that lysozyme is not used to treat the cells, which might explain the high purity

of the recovered periplasmic fraction.

The three methods were compared for the periplasmic extraction of a control protein *ie*. GafAEc. GafAEc was previously characterised in GHT lab therefore it has defined overproduction conditions (*ie*. BL21 DE3, harvesting 4 hours following induction). The pellet of the overproducing BL21 cells from the same biological replicate was divided into three portions of equal mass. Each one was treated with one of the three different methods described above and the whole cell, spheroplast and periplasmic extracts where run on an SDS-PAGE gel. The whole cell and spheroplast samples were dissolved in sample buffer after standardisation according to the OD_{600} (divided by 3) at time of harvesting. The results show that the periplasmic fraction of the ICOS method has the highest yield of protein, albeit it shows the highest contamination with membrane and cytoplasmic proteins (Figure 2. 12A). The TSE method recovered the lowest yield of protein compared to the rest three, however it has the lowest level of contaminants. Lastly, the STE method showed moderate levels of recovery of the periplasmic overproduced protein as compared to the other two methods. Also, the level of contamination was halfway between ICOS and TSE (Figure 2. 15A).

Subsequent, the periplasmic fractions were purified using Ni^{2+}- affinity chromatography in separate columns. Each column was washed 3 times with wash buffer containing 20 mM imidazole and eluted using elution buffer with 150-250 mM of imidazole. 10 ul from each 2 - 3 ml fraction (flow through, washing steps, elution steps) were run on an SDS-PAGE gel and stained on Coomassie blue. The results show that the highest yield of GafAEc was retrieved from the ICOS, however the protein was co-purified with a high level of contaminants (Figure 2.15B). The purified protein from TSE method presented the lowest yield but the highest purity. Therefore, the comparison showed that the ICOS is a good method for recovering high yield of the protein however if one aims to achieve a purer fraction of overproduced protein TSE periplasmic extraction is the best method out of the ones tested (Figure 2. 15B). A second purification step might be necessary for the ICOS as the contaminants will give high background error in ESI-MS and native MS at the following steps in the

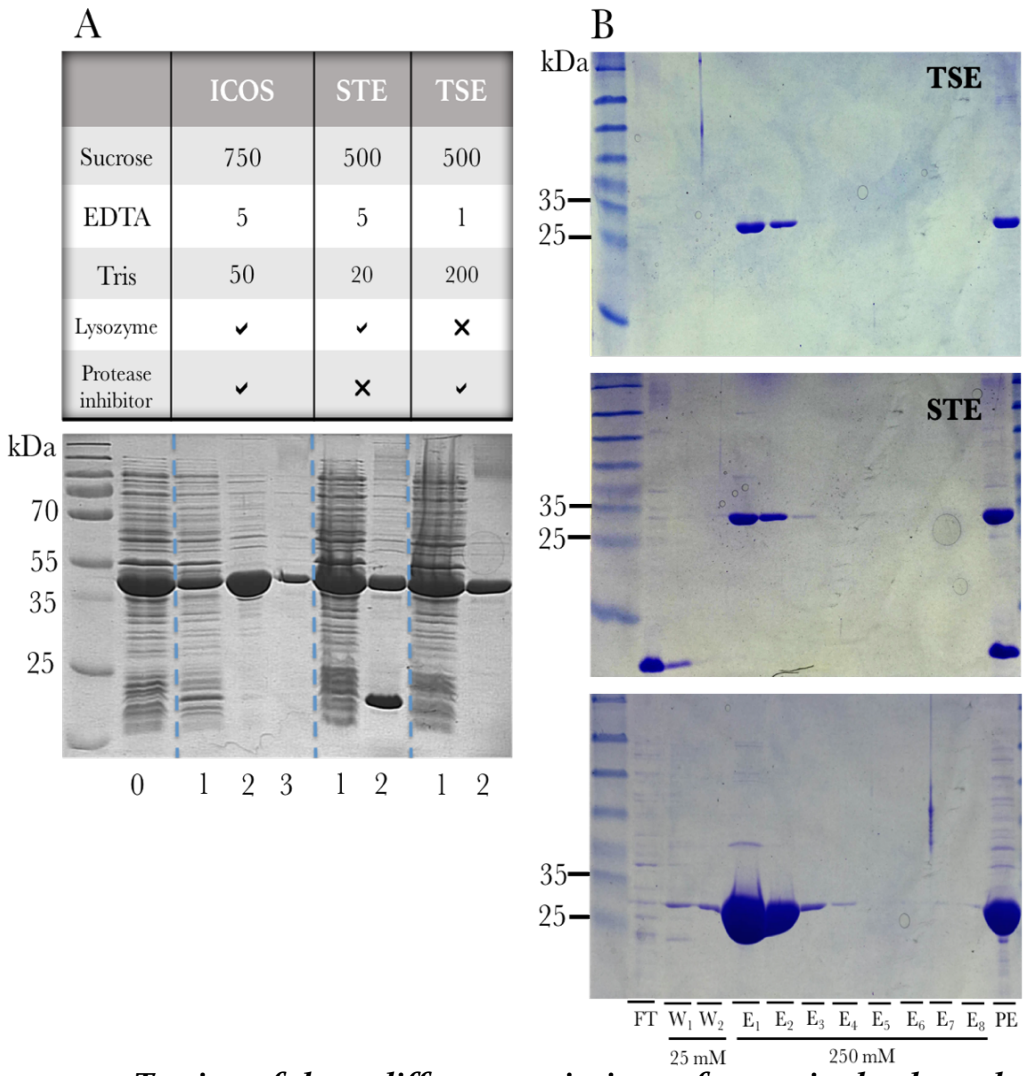


Figure 2. 15. Testing of three different variations of osmotic shock method for for periplasmic extraction of YtfQ.

A) A single biological replicate of BL21 DE3 pET21b (ytfQ) was induced with 1 mM IPTG at OD_{600} = 0.5. The cellular mass was collected by centrifugation 4 hours following induction and separated into three fractions of equal mass. Each fraction was treated separately with one of the following methods: ICOS, STE or TSE, to release the periplasmic contents of the cells. The reagents used at each one of the three methods are designated on the table. SDS- PAGE analysis followed by staining using Coomasie blue showed that the highest yield is retrieved by using ICOS whereas the TSE recovered the cleanest periplasmic preparation. (0= whole cells extract, 1= spheroplast extract, 2= periplasmic fraction and 3=supernatant following first centrifugation step at ICOS method).

B) Purification of the isolated periplasmic fractions from the three different methods using a Ni^{2+} - based column. The column was equilibrated and washed with 20 mM imidazole following binding of the YtfQ at the column. The YtfQ fraction was eluted off the column using 250 mM imidazole at 8 successive steps of 2 ml each. The results show that a trade off exists between purity and the yield of the protein with TSE retrieving the purest fraction albeit with the lowest concentration whereas the opposite occurs when the cells are treated with ICOS. (FT= flow through, W= washing steps, E= elution steps, Lysate= periplasmic fraction prior to purification).

FT = flow through, W = washing steps, E = elution steps. The concentration of imidazole used is designated below the washing and elution steps. Molecular markers used: Page Ruler Prestained (10 to 250 kDa).All concentrations mentioned are in mM.

characterisation of the GafAs.

2. 3. 8. 2 Lysozyme treatment during ICOS causes release of the uncleaved form of GafASm

Given that the ICOS method increases the yields but decreases the purity achieves we sought out to analyse the two steps of the process separately. The fractions from the first step (treatment with sucrose mix), and from the second step (lysozyme addition) of the ICOS extraction were purified separately to identify the stage/s at which the unprocessed form of GafASm is accumulated. The purification process occured as described in Materials and Methods Section 2. 2. 5. The eluted fractions were resolved on 12% SDS-PAGE gels (Figure 2. 16). Treatment with the sucrose fraction led to release of the processed GafASm, which should be localised only in the periplasm (Figure 2. 16Ai). The treatment with lysozyme in the presence of Mg^{2+} led to accumulation of the processed and the unprocessed forms of GafASm, therefore confirming that lysozyme is not only dissolving the peptidoglycan wall, but could potentially exert a destabilision effect on the inner membrane to a yet unknown extend, however enough to release the unprocessed forms (Figure 2. 16Aii). Nonetheless a study carried out by Derde *et al.* (2013), which studied the effects of hen egg white lysozyme on the outer and inner membranes of *E. coli*, showed that the same concentration tested here was able to increase the permeability of the inner membrane (Derde *et al.*, 2013). Also provided that the initial step of the ICOS treatment applies high osmotic pressure on the bacterial cell as well as high centrifugal forces, it is likely that the cells are rendered more susceptible to lysozyme - induced permeability which would be sufficient to cause lysis and release of cytoplasmic contents.

The high resemblance of the first step of the ICOS extraction to the TSE method, suggested a final experiment. The next experiment sought out to examine the effect of replacing the sucrose mix of ICOS with the TSE one. As described above the TSE sucrose mix contains 250 mM less sucrose, 4x less EDTA and 10x more Tris-HCl (pH=8). Essentially the experiment tested that TSE won't lead to release of the

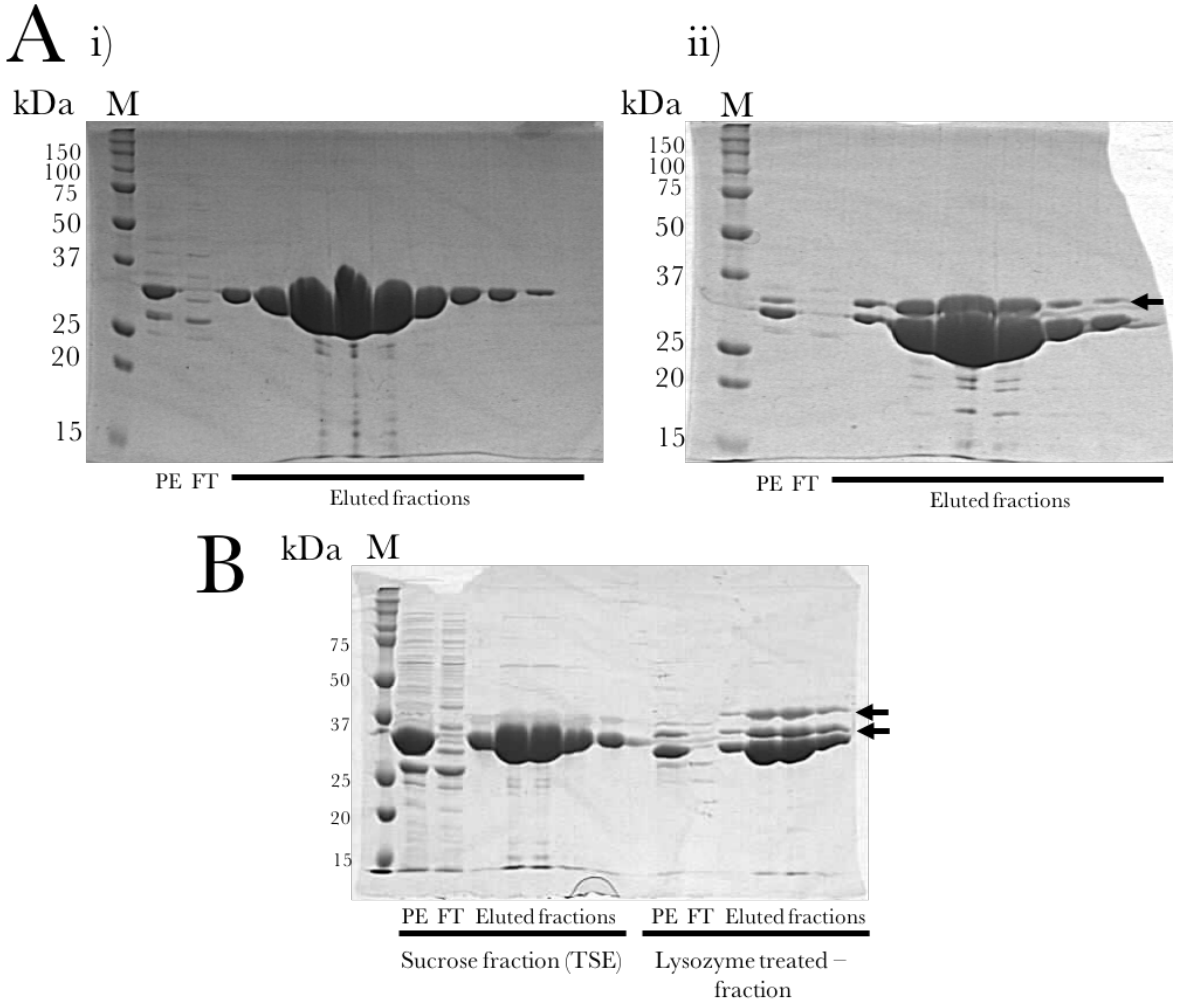


Figure 2. 16. Lysozyme treatment releases undefined GafASm species during ICOS.

A) SDS-PAGE gel of purified GafASm following extraction by ICOS. A single biological replicate of BL21 (DE3) Star pCD05 was induced with 1 mM IPTG at OD_{600} = 0.5. The cellular mass was collected by centrifugation at 18 hours following induction. The fractions collected from the two successive steps of ICOS were purified using a Ni^{2+} based column. SDS- PAGE analysis followed by staining using Coomasie blue revealed that the unprocessed form of GafASm is retained in the pellet and not released following treatment with the sucrose mix (i). The lysozyme inclusion in the treatment at step 2 of ICOS is possible to have caused the release of the unprocessed forms (ii).

B) SDS-PAGE gel of purified GafASm following extraction by ICOS with a modified first step to resemble TSE. All the conditions are as described above, apart from the step 1 of ICOS which utilised a TSE - like sucrose mix. The SDS-PAGE analysis showed that again the presence of the unprocessed species occurs after addition of lysozyme. The use of only TSE or the first step of ICOS for periplasmic extraction of GafASm prevents co-purification of the unprocessed forms with the fully pelB cleaved species.

PE= periplasmic extract, FT= flow through, Eluted fractions = Elution using 500 mM imidazole. The bands of the unprocessed forms of GafASm are indicated with an arrow. M = Biorad precision Plus protein Std.

putative unprocessed forms of GafASm. Indeed, this was the case as it failed to do so, whereas the additional step with lysozyme treatment released a heterogeneous pool of GafASm with two of the major bands predicted to be the partially processed forms of the protein (Figure 2. 16B).

2. 3. 10 Screening of overproduction media to increase production yields of the SBPs

Owing to the low yield of protein released by TSE as compared to ICOS (Figure 2. 10, 11, 13) the levels of the proteins were optimised by screening different media to achieve higher final cell density. Assuming that the levels of accumulation of recombinant proteins are retained unchanged and independent of media composition, then the increase in biomass yields alone will serve to increase the production of the recombinant protein. To examine the effect of media composition on the production of GafASm and GafASw, the relative level of expression was analyzed using three different media formulations which are richer than LB. The media tested included the Power Broth (PB) (Broedel E., M. Papciak and R. Jones, 2001), Terrific Broth (TB) (Tartoff and Hobbs, 1987) and Autoinduction (AI) (Studier, 2005) media. More information on the preparation of the aforementioned media is found in the Materials and Methods *Section 2. 2. 2. 2*. The BL21 (DE3) Star cells were inoculated in the respective media and used in small scale expression trials in an identical manner as explained in the *Section 2. 3. 4. 3*. The induction of expression by IPTG occurred at the start of the exponential phase in all cases apart from the AI media, where induction is not required as it contains lactose (allolactose); a natural inducer of P_{lac}. At the early stages of growth in AI media, the limited concentration of glucose present in the AI media prevents uptake of lactose. During mid to late exponential phase, once glucose reaches near - depletion levels, the lactose is taken up by the cells and converted into allolactose by β-galactosidase. Thereby, it lifts repression of the T7 lac by binding on LacI repressor, and consequently allows T7 RNA polymerase to induce expression of the gene of interest. The sampling occurred just before addition of IPTG and 20 hours following induction. In the case of AI media, the sampling happened at 18 hours following initiation of incubation at 20 °C. The collected fractions were lysed and run

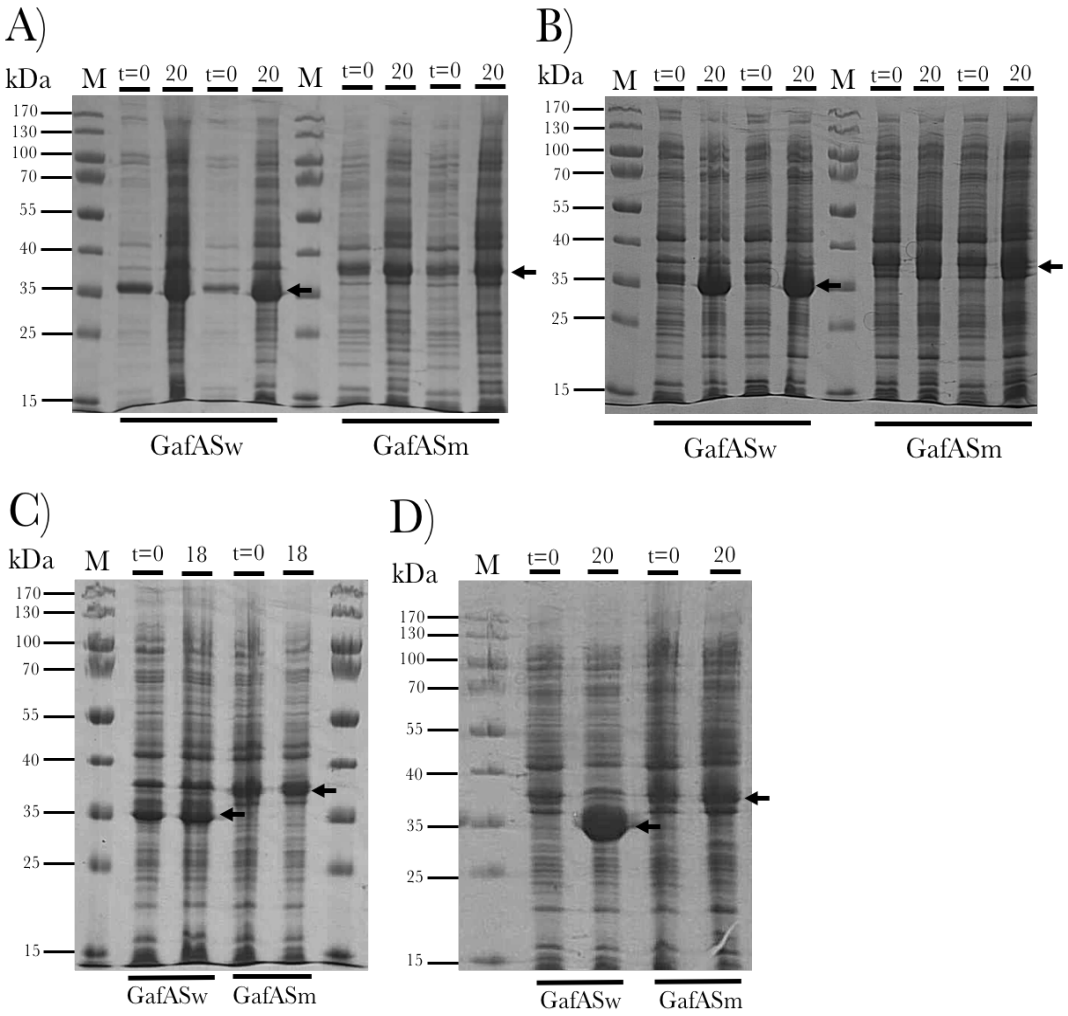


Figure 2. 17. Screening of different media for overproduction of GafASm and GafASw.

Coomassie stained SDS-PAGE gels from expression trials using three different rich media for optimisation of the GafASw and GafASm yields.

A. Expression profiles of the GafASw and GafASm when overproduced in LB media.
B. Expression profiles of the GafASw and GafASm when BL21 Star cells are grown in Power Browth (PB) media. The final OD_{600} was 1.5x higher than when LB was used. High overproduction levels are achieved for GafASw and moderate for GafASm.
C. Expression profiles of the GafASw and GafASm when BL21 Star cells are grown in AI media. The final OD_{600} was the same with the final OD_{600} when overproduction happened in LB. Both proteins are overproduced at moderate levels.
D. Expression profiles of the GafASw and GafASm when BL21 Star cells are grown in Terrific Browth (TB) media. The final OD_{600} was 2x higher than when LB was used. High overproduction levels are achieved for GafASw and moderate in the case of GafASm.

The predetermined overproducing conditions tested in the trials above were: *gafASw* and *gafASm* expression from pET20b in BL21(DE3) Star cells at 20 °C. The sampling occurred at 0 and 20 hours (timepoint at which the protein yields reached near - maximum in LB) following induction of expression with IPTG. The sampling for the Autoinduction (AI) media occurred at 18 hours following initiation of growth at 20 °C.

Time (t) is equal to the number of hours post induction (not applicable to AI media) and arrows indicate the position of an overproduced protein.

on a 12% SDS-PAGE resolving gel. The analysis showed that GafASw and GafASm were successfully overproduced in the PB and TB media, respectively (Figure 2. 17 B, D). As the final cell biomass was doubled in PB as compared to LB, the first replaced the latter in the overproduction of GafASw.

2. 4 Discussion

The furanose form of arabinose is a frequent sidechain present in the arabinoglucuronoxylan of the bioenergy grasses. The high quantities of this sugar released following chemical and enzymatic pretreatments of the grasses, raise its importance as a carbon source for the fermenting bacteria at downstream stages of the biofuel production. Its complete release from the xylan polysaccharide following chemical pretreatment is considered necessary despite many studies use industrial cocktails of hydrolases which do not include L-arabinofuranosidases (). A more correct methodology should aim for efficient removal of L-arabinofuranose instead as this would free more sites for xylanases to bind on the backbone of xylan (Li *et al.*, 2014; Sakamoto *et al.*, 2011). Therefore, the high titer of L-arabinofuranose in the fermenting media demands optimisation of the transport machinery for its uptake which will increase growth rates and consequently lead to higher titre of biofuel. The work described here, identified bacterial orthologue SBPs of an established ABC GalactoArabinofuranose (Gaf) SBP; *ie*. YtfQ from *E. coli* (Horler *et al.*, 2009), with potentially preferential binding of arabinose over galactose. The investigation was prompted assuming that the various ecological niches the bacteria reside in, would act as an evolutionary pressure and ultimately give rise to different subsets of binding specificities for the GafAs. Particular focus was given to the SBPs from bacterial symbionts and pathogens, as the efficient uptake and utilisation of arabinofuranose, present in the arabinoxylan and the root exudates, is expected to provide a competitive edge in the colonization of the crowded soil environment (Basic *et al.*, 1986; Aguedo *et al.*, 2013; Ji *et al.*, 2015).

Initial phylogenetic analysis, including all the GafAs, showed that these are widespread across a range of bacterial phyla and classes thus underlining that a necessity exists for the transport of the furanose form of sugars, and particularly for arabinofuranose. When the phylogenetic analysis was performed including a reduced number of GafAs which exhibit interesting genomic context related to arabinose or galactose utilisation, the separation into individual clades became apparent. The

general trend of clustering was dictated by the phylum/class, in such way that GafAs of the same class clustered with their most cognate GafAs from the same phyla. Only notable difference is the clustering of γ-proteobacterial GafAs (*Cellvibrio japonicus* Ueda107 and *Marinomonas* sp. Mw1) within the clade of α-proteobacterial ones which could be due to horizontal gene transfer. Preliminary analysis using the online database HGTree gave more insights to the presumptive gene transfer event. HGTree (Jeong *et al*., 2016) provides putative genome-wide horizontal gene transfer events for prokaryotic genomes by reconstruction of maximum likelihood trees for each orthologous gene and for respective 16S rRNA. The two trees are reconciled under parsimony framework so they give rise to reliable predictions. Using only a phylogenetic tree to infer vertical evolution could be unreliable because statistically significant sequence is not necessarily the result of vertical evolution (Koski and Golding, 2001; Ravenhall *et al*., 2015). Therefore, any predictions made by HGTree will differ to our trees in that they are supported by 16S rRNA reference trees. The database predicts that GafAs from *Cellvibrio japonicus* Ueda107 and *Marinomonas* sp. Mw1 are both the product of horizontal gene transfer from α-proteobacteria of mainly *Agrobacterium*, *Rhizobium* species. This would provide an explanation of why they cluster in the α-proteobacteria clade. However, the same applies for some γ-proteobacteria from the large cluster which includes GafAEc, *rahaq2_4397* (*Rahnella aqualitis*) and *asa_2140* (*Aeromonas salmonicida* subsp. *salmonicida* A449). This would explain why the Rahaq2_4397 and KKA50977 from γ-proteobacteria cluster with α's in the smaller tree constructed. When α-proteobacteria, *eg*. *snov_0124* (*Starkeya novella*) and *atu3222* (*Agrobacterium fabrum* str. C58), are the starting query in HGTree they still represent donors of GafAs to γ-proteobacteria. Therefore, we expect that the GafAs from γ-proteobacteria were derived from α-proteobacteria by either horizontal or vertical gene transfer. Most notably, the *shewana3_2073* (*ie*. GafASw) and *sbal223_2147* (*Shewanella baltica* OS223) are predicted to be derived from a horizontal event from either of *sde_0766* (*Saccharophagus degradans*), *cja_3019* (*Cellvibrio japonicus*) and *mmwyl1_3535* (*Marinomonas* sp. Mw1). Therefore, we predict that the GafASw could have been horizontally transferred from cognate

bacteria of the same phyla, who themselves obtained *gafAs* from α-proteobacteria. However, more evidence is required to conclude on definite horizontal gene transfer events, such as examination of genomic signature (eg. GC bias, nucleotide composition and codon usage) to corroborate the above predictions. Since the current study is oriented towards biochemical analysis of SBPs, we don't pursue this further.

The genomic context of GafASw and GafACb shows high relevance to arabinan/arabinose utilisation. Most of the arabinofuranosidases found in close proximity to the *gaf* operons in the bacteria analysed, are classified in CAZy as members of the GH43 and GH62 family. These enzymes normally catalyse hydrolysis by inversion of the anomeric carbon, which releases the **β anomer** of L-arabinofuranose. The latter is implied from examination of a GH62 arabinofuranosidase from *S. coelicolor*; its resolved structure was co-crystallised with β-L-arabinofuranose bound, representing the product state of inverted catalytic process which retained the furan form (Maehara *et al*., 2013). Most GH43 arabinofuranosidases with crystal structures are bound to arabino-xylooligomers and therefore we are unable to infer the same conclusion as above. XALc_0946 (*Xanthomonas albilineans*), Xcel_0412 (*Xylonimonas cellulolytica*) and AMED_3994 (*Amycolatopsis mediterranei*) all neighbor with *gaf* operons and are arabinofuranosidases from GH51 and GH54 families. These hydrolyse by retention of the anomeric carbon (Bouraoui *et al*., 2016) and thus are often co-crystallised with **α**-L-arabino**furanose** as their catalytic product (Hoevel *et al*., 2003). The above analysis combined with the fact that the aforesaid gene operons are both predicted to be regulated by arabinose-controlled TFs, are solid indications for a preference towards arabinose released from plant polysaccharides or root exudates. Unlike *gafASw* and *gafACb*, *gafAEc* is likely to the induced by GalR as suggested by Horler *et al*., 2009. Also, galactose is present in LB and MOPS minimal media as GafAEc was co-purified with it bound at both conditions. Assuming that the galactose concentrations are present at induction levels, and considering that: **1.** GafAEc is present in both of the aforesaid conditions and **2.** AraC exerts a non-canonical regulation to the GafABCDEc, then the expression is most likely to be regulated by GalR/S. Deletion of GafASw

causes a decrease in fitness of *Shewanella* sp. ANA3 when grown on minimal media with arabinose (Wetmore *et al*., 2015) and the bacterium in question cannot grow on galactose (Rodionov *et al.,* 2010). The first proves the essentiality of the system in uptake of L-arabinose and raises questions about the existence of an arabinopyranose transporter; the second finding points out that in case the GafABCD is present when galactose is in proximity to *Shewanella* sp. ANA3 then it is unable to recognise this ligand. These indications led to us to choose the above two candidates for further characterisation along with GafASm. The operon expressing the latter is induced in the presence of L-arabinose and D-fucose which are prominent monosaccharides of the root mucilage (Nguimbou *et al*., 2012; Tyler, 1965).

Due to consecutive failures to overproduce the GafACb and only achieving this at a late stage in the project the characterisation of this ortholog wasn't possible. GafASw and GafASm both overproduced at moderate levels when the incubation temperature was lowered to 20 °C and the RnaE was not produced in the cell *ie*. BL21 Star cells. These conditions slow down the production rate (Schein and Noteborn, 1988; Vasina and Baneyx, 1997) and also decrease the protein aggregation which is favoured at higher temperatures (Baldwin, 1986; Schellman, 1997). Further they promote its stabilisation at the RNA level. However, when working at the lower end of the temperature range, slower growth and reduced synthesis rates can result in lower protein yields. Therefore, we used different overproducing media to achieve higher yields; indeed, TB and PB sustained overproduction and allowed higher cell density by the end of the incubation.

The periplasmic extraction of overproduced GafASm using ICOS, resulted in isolation of two protein species. The analysis with FT-ICR-MS showed the presence of a larger species which was assumed to be GafASm with partially processed PelB sequence. The observed size of the undefined species doesn't fully match the predicted size of the possible combinations for partially cleaved GafASm. N-terminal sequencing would be a suggestion to resolve this issue, albeit the protein was extracted using TSE, which was shown not to release the undefined version of the protein. Other studies have

reported such incident when attempted to overproduce exogenous periplasmic in *E. coli*, including the pelB tagged Vpu protein from HIV-1 (Deb *et al*., 2017) and pelB tagged alkaline phosphatase IV from *Bacillus subtilus* (Koksharov *et al*., 2013). Assuming the unprocessed version of the protein is still anchored on the inner membrane or localised in the cytoplasm then treatment with TSE is not efficient in releasing cytoplasmic contents. Interestingly, a band with almost double the molecular weight than AraFCb was identified by peptide mass fingerprinting to be DnaK (data not shown), which is a cytoplasmically localised chaperone. The fraction for AraFCb extraction was treated with ICOS rather than TSE, verifying the potency of the first.

Chapter 3

Biochemical and structural analysis of GafASm from *Sinorhizobium meliloti* 1021 and GafASw of *Shewanella* sp. ANA – 3

The work presented in this chapter aims to provide biochemical insights into the function of both of the selected GafAs, *ie*. GafASm and GafASw, and structural information about the binding cavity of GafASw. The introduction emphasizes the importance of SBPs in functional transport engineering as they determine the substrate specificity of their ABC transport systems. Different classes of SBP are introduced with examples from characterised ABC systems, such as MglB and GafAEc. Thereafter, the reader is introduced to the techniques involved in the characterisation of the SBPs in this study. The first part of the Results concentrates on the efforts to characterize the binding specificity of the GafASm and draw comparison to GafAEc. Further, as GafASm its proven to bind galactose, its presence in the overproducing media leads to the next experiments which intended to purify ligand – free protein. The following sections include the biochemical analyses used to determine the binding constants of GafASm ligands. The second part of the chapter focuses on the determination of the binding specificity of GafASw, followed by a buffer assessment, similarly to GafASm. Subsequently, the binding constants are determined and the inability of GafASw to bind galactose but still bind L-arabinose is verified. As this is a desired feature of a GafA protein, the final part of the chapter describes the efforts to crystallise the GafASw and obtain its structure. The structure is compared to the GafAEc, and the unique features of the binding cavity of GafASw which render D-galactose binding unfavourable while retaining L-ara*f* binding, are identified and explained.

3. 1 Introduction

This study is originated following the discovery of the furanose-specific GafABCDEc system from *E. coli* made by Horler *et al*. (2009). The SBP of this system is able to bind L-ara*f*; an abundant branching unit of the plant polysaccharide xylan (Horler *et al*., 2009). Horler *et al*.'s discovery has underlined the importance of ABC transport in scavenging the full content of sugars released in the extracellular environment of bacteria, including both pyranoses and furanoses. As an effort to gain more information about L-arabinofuranose-specific transport we identified potential candidate SBPs, *ie*. GafASw and GafASm, which based on their genomic context, predictions and experimental data about their expression as well as growth assays (*see* Chapter 2) suggested that they seem to be 'biased' towards arabinose over galactose binding. Further analysis of how these GafAs bind the sugar in question and how this differs to the *E. coli* ortholog, *ie*. GafAEc, will ultimately lead to direct engineering of the GafAs in an effort to increase the transport efficiency of this xylan-derived sugar. The above approach is essentially targeting solely the SBPs of the ABC systems for metabolic engineering, avoiding any modifications on the membrane-bound components such as the transmembrane domain (TMD) and the nucleotide-binding domain (NBD). This is a logical strategy as it has been suggested that, as far as binding protein-dependent ABC transporters are concerned, the substrate specificity is exclusively determined by their SBPs and the transmembrane components act only as a nonspecific pore to allow substrate translocation across the inner membrane (Dawson, Hollenstein and Locher, 2007). The subsequent sections will introduce the reader to the different classes of the bacterial SBPs that exist and also provide notable examples of carbohydrate binding ones.

3. 1. 1 Structural classification of SBPs

The first SBP protein to be crystallised was the L-arabinopyranose binding protein (AraF) (Quiocho *et al*., 1974), and due to rapid advances in protein crystallisation and structure elucidation, many more structures have been deposited in the Protein Bank

Database (PBD) since then. The general structure of SBPs includes two α/β domains, featuring a central β-sheet formed by five β-strands which are surrounded by α-helices (Figure 3. 1). A hinge region connects the two domains, with the binding site normally located buried in between the two domains. The hinge allows for flexibility and free rotation of the domains in the absence of the ligand (Tang *et al*., 2007). This structural state of the protein is known as the apo- or open-unliganded (Quiocho *et al*., 1996), from which transitions to the closed-liganded state by trapping the ligand in between its two domains. The binding process is known as the "Venus Fly-trap" mechanism because of its apparent resemblance to how the Venus flytrap captures an insect.

Fukami-Kobayashi *et al*. (1999) have attempted to classify the SBPs based on the connectivity of their secondary structural elements and the topology of the β-sheets in the center of the structure (Fukami-Kobayashi *et al*., 1999). The group has defined two Classes based on the criteria above, and a third one was later introduced by Lee *et al*. (1999) following the discovery of a zinc-binding protein, *ie*. TroA (Lee *et al*., 1999). Generally, although with some clear exceptions, SBPs with the same type of substrates are grouped in the same class. Conversely, the size of the protein does not correlate with the class it belongs to (Berntsson *et al*., 2010). The unique structural elements distinguishing the SBPs into the three classes are mentioned in the sections below.

3. 1. 1. 1 Class I SBPs

The members of the Class I are characterised by a unique hinge region formed by three distinct strands connecting the lobes (highlighted in Figure 3. 1), thus positioning the N- and C-terminus within the same domain. Further, the β-sheet topology of the overall structure is β2β1β3β4β5 (Fukami-Kobayashi *et al*., 1999). The present class is the most relevant to this study, as its most prominent members are SBPs binding carbohydrates (*ie*. D-ribose, D-glucose, D-galactose, D-xylose and L-arabinose). Other members include SBPs responsible for binding of natriuretic peptides, autoinducer-2

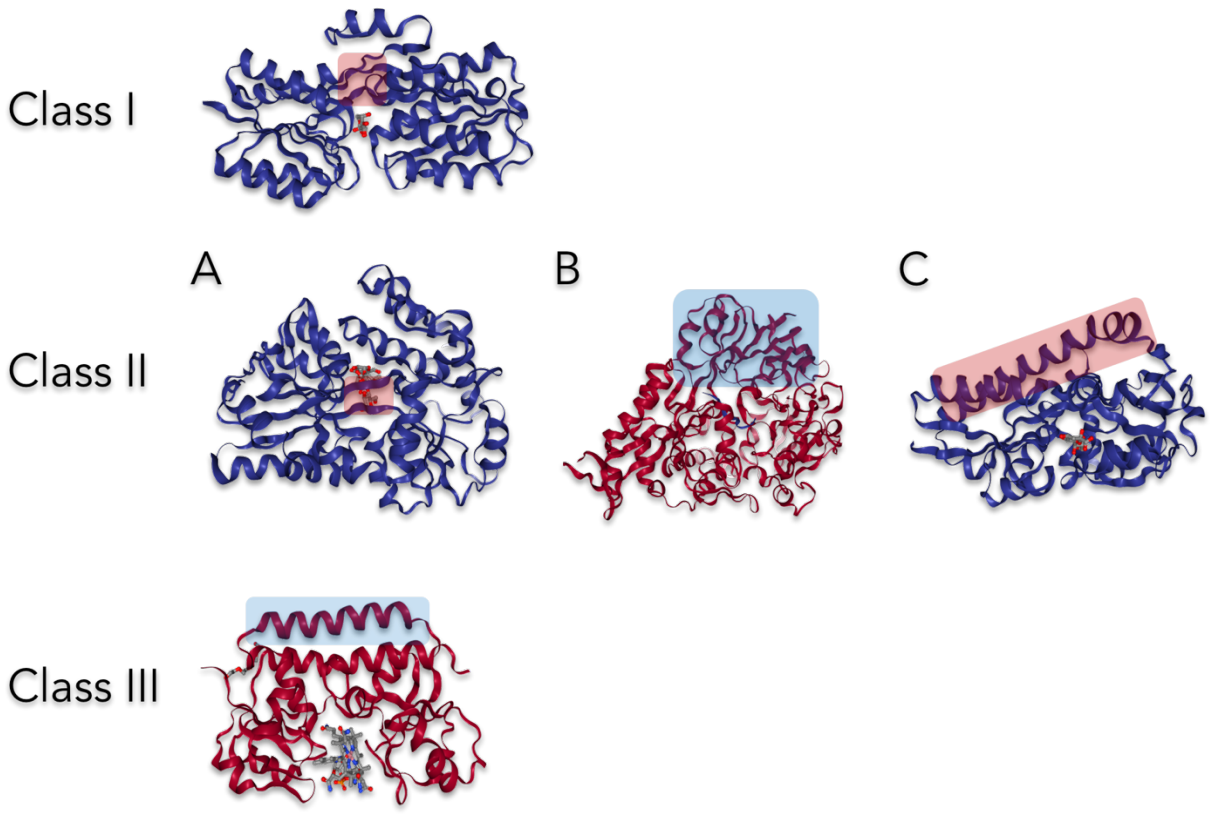


Figure 3. 1. Distinct features of the SBPs classes and examples of their members. A distinct feature of most of the SBPs members from Class I are the three interconnecting segments between their two domains. This class includes the SBPs for L-arabinopyranose, D-allopyranose, D-glucose, D-galactopyranose and D-ribopyranose.
The Class II is comprised by members with various features: The SBP shown at (A) contains two relative short hinges connecting the two domains. This feature is encountered in the structures of maltodextrin (shown here), molydbate, thiamine, and SBPs for iron binding. In (B) the SBPs structure contains an additional domain which increases its size considerably compared to other structures. Such feature is commonly detectable in SBPs for di and tripeptides, cellobiose, arginine and nickel. The group C of the Class II is comprised by members typically found in TRAP systems, such as SBPs which bind sialic acid, ectoine, glucuronic acid and galacturonic acid. These entail a large helix which functions as a hinge region.
A unique characteristic found in members of the Class III is a single connection between the two domains in the form of a rigid helix. Some examples from this class include the vitamin B12, zinc, heme and siderophore SBPs.
The proteins used to illustrate the features are: Class I = AraF (*E. coli*, 1ABE), Class II (A) = MaltodextrinBP (*E. coli*, 4MBP), (B) = OppA (*Lactococcus lactis*, 3DRF), (C) = SiaP (*Haemophilus influenzae*, 2WYP), Class III = BtuF (*E. coli* ,1N2Z). The classification presented here is based on Fukami-Kobayashi *et al.*, 1999. The distinct features of each class are highlighted with blue or red.

and branched chain amino acids such as the leucine/isoleucine/valine binding protein (Sack *et al*., 1989). The SBPs of this class mainly interact with Type I ABC transporters, with some of the SBPs showing high primary sequence identity to ligand binding domains in *lac*-repressor type transcription factors (Berntsson *et al.*, 2010); as a result, such incorrectly annotated proteins are often retrieved following BlastP searches of Class I SBPs.

3. 1. 1. 2 Class II SBPs

The Class II is the most variable in terms of the distinguishing features of its members. The β-sheet topology of this class is β2β1β3β*n*β4, with n indicating the domain swapped strand after the initial cross-over from the N-terminal to the C-terminal domain (Fukami-Kobayashi *et al*., 1999). A subset of the members of this class form a unique group in that they contain an extra domain (highlighted in Figure 3. 1B) therefore, extending their size which typically ranges from 55 to 70 kDa (Berntsson *et al.*, 2010). These SBPs interact with Type I ABC importers and recognise diverse ligands including di- and tri-peptides, nickel ions, cellobiose and arginine (Berntsson *et al.*, 2010). The extra domain in some of these peptide-specific SBPs lengthens their binding cavity to accommodate large ligands, *ie*. comprised of more than 20 peptide residues.

A larger group within the Class II, includes members which bind carbohydrates, putrescine, thiamine as well as ferric or ferrous ions. The distinct feature of this group is that they have two short strands which form the hinge region (Figure 3. 1A). A well-studied example belonging of this category is the Maltose binding protein (MBP, MalE) from *E. coli* (*See General Introduction 1. 4. 3*) (Spurlino, Lu and Quiocho, 1991).

Another example of unique structural elements exhibited by a subset of this class, are the SBPs from the TRAP transporter systems (tripartite ATP-independent periplasmic transporter). In contrast to the ABC transporters, the TRAP systems employ an electrochemical current (of H^+ or Na^+) to achieve the uphill translocation of their substrates. The SBPs of TRAP systems are also known as ESRs (extracellular solute

receptors) to distinguish them from their ABC counterparts. A distinguishing feature of the TRAP ESRs is the presence of an extra helix that spans across both domains (highlighted in Figure 3. 1C) (Berntsson *et al*., 2010). All of the ESR structures elucidated to date contain this long helix; examples of TRAP systems are ectoine (*ie*. UehA) (Lecher *et al*., 2009), lactate (Gonin *et al*., 2007), sialic acid (Severi *et al*., 2005) and glucuronic acid (Vetting *et al*., 2015). The conserved interaction between an arginine residue in the ESRs binding cavity and the carboxylate group of the substrates has been postulated to restrict the substrate range of TRAP systems to carboxylate-containing substrates (Fischer *et al*., 2015).

3. 1. 1. 3 Class III of SBPs

This class encompasses members with an α-helix serving as the interconnecting hinge region between the two domains (Berntsson *et al*., 2010) (Figure 3. 1). The helix is thought to increase the inflexibility of the structure, as reflected by the minimal movement of the domains upon ligand binding. An example of this rigid structure is the BtuF, the SBP responsible for the vitamin B_{12} uptake in *E. coli*, which rotates a mere 4° upon substrate binding (Hvorup *et al.*, 2007).

3. 1. 2 Bacterial SBPs involved in binding of xylan-derived sugars

Some of the first monosaccharide-specific SBPs to be characterised were from *E. coli.* These included the ribopyranose-binding protein (RbsB) (Mowbray and Cole, 1992), the galactopyranose/glucose-binding protein (MglB) (Vyas *et al*., 1983) and the arabinopyranose-binding protein (AraF) (Figure 3. 2). Since then a huge progress has been made using *E. coli* as the workforce and overproducing host to identify and characterise novel SBPs. One of the SBPs identified with pivotal importance to this study was the furanose-specific SBP *ie*. YtfQ (GafAEc) (Horler *et al*., 2009). As already mentioned, the furanose forms of sugars are highly unfavoured in solution, with the 6-membered ring form *ie*. pyranose being the predominant form. However, such concentrations still appear to be adequate for bacteria that are scavenging carbon and competing with other species in the colonisation of niches including the cell wall of

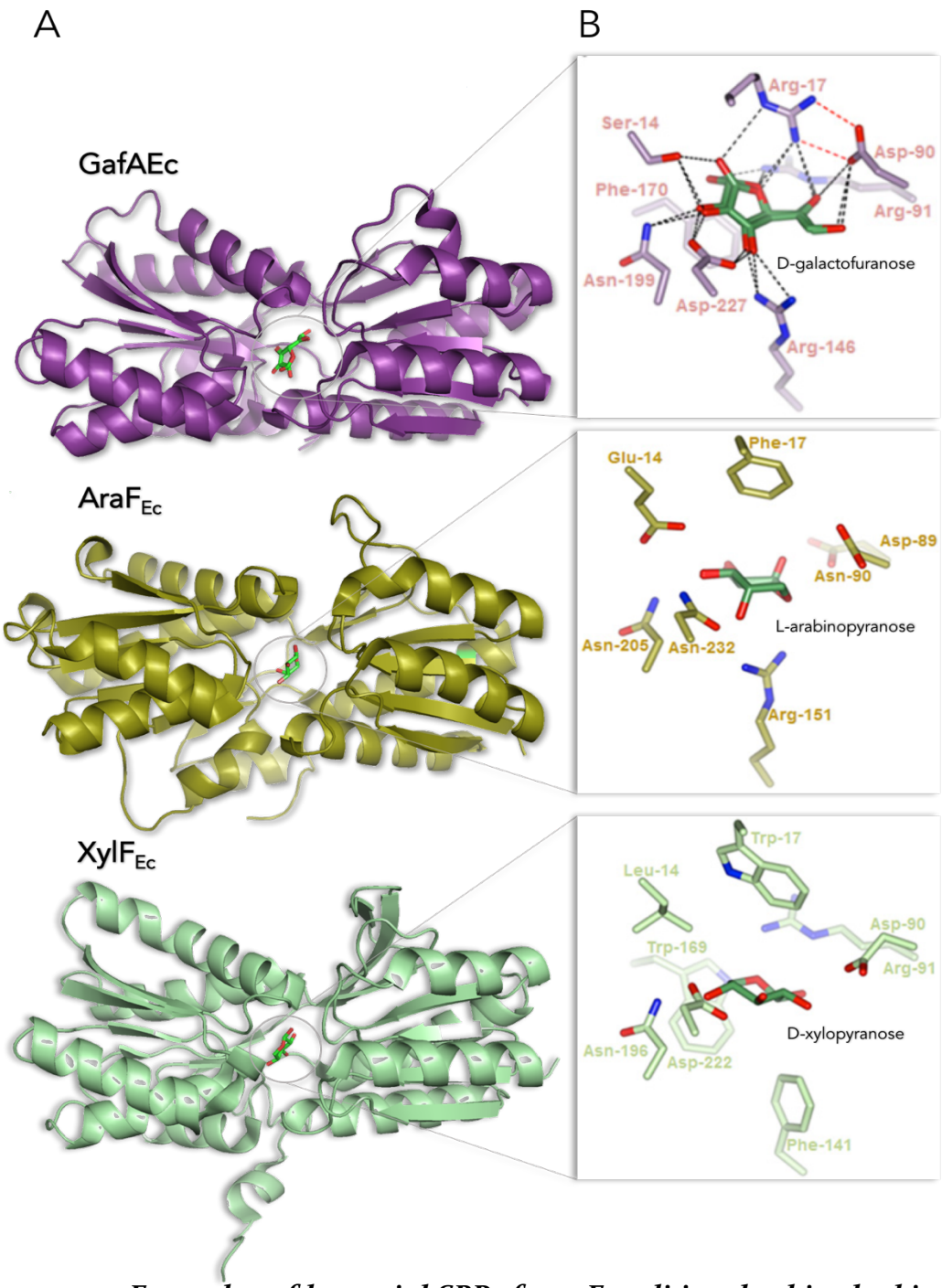


Figure 3. 2. Examples of bacterial SBPs from E. coli involved in the binding of xylan-derived monosaccharides.

(A) The 3D structures of the GafAEc (purple, 2VK2), AraF (olive green, 1ABE) and XylF (green, 3MAo) from the ABC transport systems responsible for uptake of galacto-arabinofuranose, arabinopyranose and xylopyranose, respectively. The structures are shown in the same orientation to demonstrate the conservation of fold in these SBPs. The ligands are shown in the stick format (green carbons).

(B) The binding sites of the 3 SBPs to illustrate the conservation of the triad Asn, Asp (or Asn) and Arg at the bottom of the binding site. The triad is also part of binding sites from other sugar SBPs including MglB (D-galactopyranose), RsbB (D-ribopyranose) and AlsB (D-allopyranose). Hydrogen bonds/ salt bridges at GafAEc binding site are indicated with dashed lines.

The annotated binding sites were derived from Maqpool *et al.*, 2015.

plants and the rhizosphere, to justify an extra dedicated transporter. Combining the function of the arabinopyranose-specific transport system (*ie*. AraFGH) and the arabinofuranose-specific system (*ie*. GafABCD) will enable a bacterium to rapidly and fully utilise the 100% of the L-arabinose present in the solution. Another important SBP for such application is the XylF from the xylopyranose-specific ABC system *ie*. XylFGH; this system is capable of scavenging be responsible for the transport of the released monosaccharides from the backbone of the xylan feedstock.

3. 1. 2. 1 Binding cavity of xylan-relevant carbohydrate SBPs

The SBP of interest in this study, *ie*. GafAEc, binds L-arabinose and D-galactose with similar binding affinity which was found to be 1.3 μM and 1.7 μM respectively (Horler *et al*., 2009). When the relative furanose concentrations in solution are taken into account, the K_Ds are even lower and equal to 0.24 and 0.30 μM for L-arabinofuranose and D-galactofuranose, respectively (Horler *et al*., 2009). The crystal structure of GafAEc was obtained with D-galactofuranose bound and upon closer examination the structure exhibits features commonly found in monosaccharide SBPs. It shares a very similar fold to XylF and AraF (Figure 3. 2A), as they all three belong to Class I SBP and therefore harbour three short peptide stretches as their hinge region which connects their N-terminal to C-terminal domain. The bottom side of the binding site contains conserved arginine, asparagine and aspartate residues with some also seen in AraF and XylF cavity (Figure 3. 2B). The triad is also conserved in other monosaccharide SBPs not presented here including RbsB, AlsB (D-allopyranose) and MglB (D-galactopyranose) (Maqpool *et al*., 2016). When GafAEc binding cavity is compared to its closest sequence and structural homologue, the D-ribopyranose binding protein RbsB, a specific adaptation in GafAEc is detected, where a phenylalanine in RbsB is now an arginine (Arg17) in GafAEc (Figure 3. 2B). The same observation is made when GafAEc is compared to its L-arabinopyranose counterpart, *ie*. AraF. The addition of a positive charge supports the formation of a salt-bridge, formed with Asp90, which changes its conformation in such way that it now points away from the binding pocket. This extra 'space' generated is to accommodate the extended form of the

furanose sugars (Maqpool *et al.*, 2016). Characteristics of the binding sites similar to the aforesaid give rise to unique ligand specificity, therefore we are prompted to gain more structural information about any of the two selected GafAs with distinctive and interesting ligand binding capabilities.

3. 1. 3 Techniques employed in the characterisation of SBPs

The biochemical characterisation of SBPs is essential when one would like to investigate the ligand binding specificity and perform thermodynamic analysis so that comparison can be drawn between orthologs. An extensive biophysical analysis is normally preceding the determination of the thermodynamic parameters, *ie*. binding affinities, enthalpy of reaction (ΔH°) and Gibbs free energy (ΔG), to ensure that the calculated values of the aforesaid parameters are accurate. The proteins in the current study had their stability and activity assessed in different acidic and alkaline buffers using differential scanning fluorimetry (DSF). Following purification of GafASm with Guanidine hydrochloride (GnHCl) to remove potential prebound ligand, the refolding of the protein back to its active state was confirmed using Circular Dichroism (CD). Finally, the monomeric state of GafASw in buffer was confirmed using Size Exclusion Chomatography - Multi-Angle Laser Light Scattering (SEC-MALLS). The initial thermodynamic analysis of the proteins proceeded with a DSF qualitative assessment of binding to determine the ligand specificities. The binding of the ligands was confirmed with intrinsic fluorescence experiments which some leading to determination of K_D values. Lastly, the calculated K_D values were confirmed using Isothermal Titration Calorimetry (ITC). The following section introduces the basic principles of each of the aforementioned techniques.

3. 1. 3. 1 Differential scanning fluorimetry (DSF)

The ability of the SBP to maintain its activity and stability at different temperatures and buffer conditions is a direct outcome of its inherent structure and conformation. A reliable and inexpensive method which can quickly provide such important information about the SBPs is the Differential scanning fluorimetry (DSF) or the

Thermal Shift assay. While known for a long time (Weber and Laurence, 1954), thermal shift analysis has recently become a widely established method for efficient screening of ligand binding or conditions which improve stability of the protein. As buffers, salts and other nonspecific detergents normally exert a stabilising effect on the proteins, DSF is a convenient assay used for screening such ingredients to facilitating selection of conditions which lead to a stable and fully active protein. The plethora of these conditions are tested in a fast and easy manner as the assay requires low protein concentration and amount of dye. Although the idea that the fluorescence increases when dye molecules accumulate to the hydrophobic regions of a protein has been known for more than 60 years (Weber and Laurence, 1954), it is only recently that this robust method has been put into application as a protein – ligand interaction screening method (Cummings *et al.*, 2006, Sorrell *et al.*, 2010). As firstly decribed by Koshland (1958) and Linderstrøm-Lang (1959), the binding of ligands stabilises the protein and increases its thermal stability (Koshland, 1958; Linderstrøm-Lang and Schellman, 1959). The site-specific binding of ligands influences the stability of the protein and usually increases its melting temperature; thus, DSF is a frequently employed in elucidating information concerning function studies, including allosteric effectors or in the case of the current study for the determination of the substrate specificity. The stabilizing effect of the ligands on the protein is thought to be the outcome of the coupling between two mutual processes under equilibrium: binding and unfolding (Fukada *et al.,* 1983, Sturtevant 1987). The extra free energy is believed to be accounted for the shift in the unfolding temperature of the protein when bound to a ligand (Shrake and Ross, 1990; 1992). Further, conformational changes could also be contributing to this effect, as mutations which changed the packing of the inner core modified the protein stability (Richards, 1997).

The method is based on the non-specific binding of the environmentally sensitive dye, SYPRO orange, to the protein hydrophobic surfaces. SYPRO orange is considered the dye with the most favourable properties, owning to its high signal-to-noise ratio. Other dyes, such as 1,8- ANS, can be tested in case no unfolding transition is observed with SYPRO orange. When the protein unfolds during thermal denaturation, the

exposed hydrophobic surfaces bind the dye, resulting in increase in fluorescence by displacing water (Figure 3. 3A). As mentioned above, an increase in the melting temperature *ie*. Tm (also known as the temperature of hydrophobic exposure, *ie*. T_h) of the protein + ligand samples compared to protein- only indicates that the protein binds the ligand. The Tm for each condition is equal to the midpoint of the sigmoidal curve which plots the raw fluorescence signal against the temperature (Figure 3. 3A). In the present study, the Tm of the proteins was calculated using the curves of the first derivative of the fluorescence emission as a function of temperature plotted against temperature. In such cases, the Tm is equal to the x axis value at the minimum of each curve (Figure 3. 3B).

3. 1. 3. 2 Circular dichroism (CD)

The unique tertiary structure of each protein is stabilized by multiple weak interactions, which includes hydrophobic interactions, hydrogen bonds and ionic interaction. This complex network of interactions determines the function of proteins (Rawn, 1989; Nelson and Cox, 2000). Nelson and Cox (2000) have suggested multiple treatments with reagents that can disrupt the intramolecular interactions and severely impact or cause collapse of the three-dimensional structure of the proteins, leading to complete loss of function. This illustrates the relationship between the conformation of the mature protein and its function.

The CD technique measures the unequal absorption of right- and left-handed circularly polarized light by the molecule. Proteins are optically active molecules because of the asymmetric carbon atom of their amino acids (*ie*. chiral chromophores), present in the L or D configurations. More importantly, the dextrorotary and levorotary nature of the secondary elements of a protein, such as α-helices and β-barrels, impart a distinct CD and produce unique CD spectral signatures making the method suitable for assessment of the structural stability (Figure 3. 4C). The obtained CD spectra can be readily used to estimate the fraction of a molecule found in the β-sheet, the α-helix or the β-turn conformation (Whitmore and Wallace,

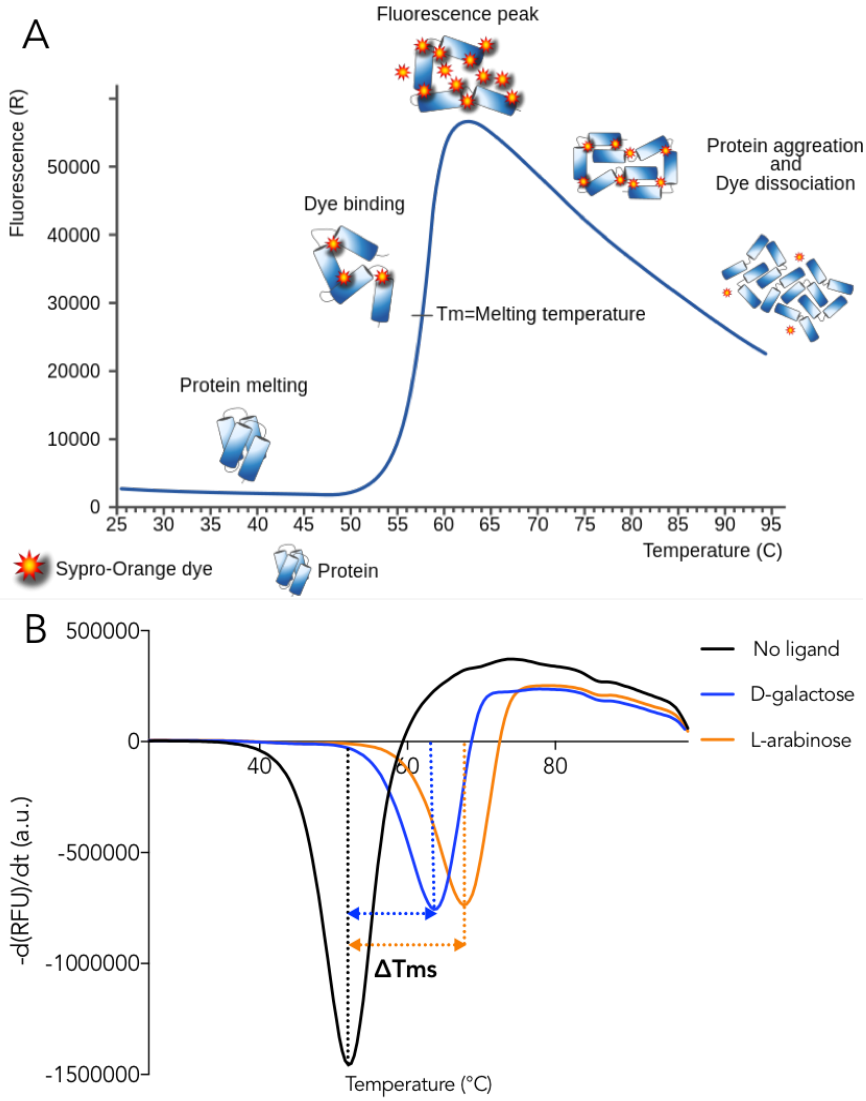


Figure 3. 3. Principle of Differential Scanning Fluorimetry.

A) The thermal shift assay is normally performed in a real-time qPCR machine and the protein is subjected to increasing temperatures ranging from 25 ◦C to 95 ◦C, in the presence of a dye (*eg*. SYPRO orange). As the protein begins to unfold the dye binds to the hydrophobic areas, which were previously unexposed, thus displacing water and causing a surge in the fluorescence signal. Thereafter the signal begins to decrease as the protein denatures completely which leads to dye dissociation. The whole process gives rise to a sigmoidal curve which represents cooperative unfolding of the protein. The Tm corresponds to the midpoint of the denaturation curve which can be calculated by non-linear fitting of the sigmoidal curves to a Boltzmann Equation.

B) The DSF data can also be presented with the first derivative of the fluorescence emission as a function of temperature (-dF/dT) plotted against temperature. The Tm in these melt derivative plots, is equal to the x value at the minimum of the curve. The example shown, confirms that the presence of ligands stabilises the protein and increases the Tm.

The figure at (A) was produced by Goran tek-en as part of the Creative Commons organisation. The melt derivative plot of figure (B) is derived from this study and presents the change in the melting temperature of GafASm in the presence of D-galactose or L-arabinose. ΔTm corresponds to the change in the Tm of the unbound protein (black) compared to the bound forms (blue and orange).

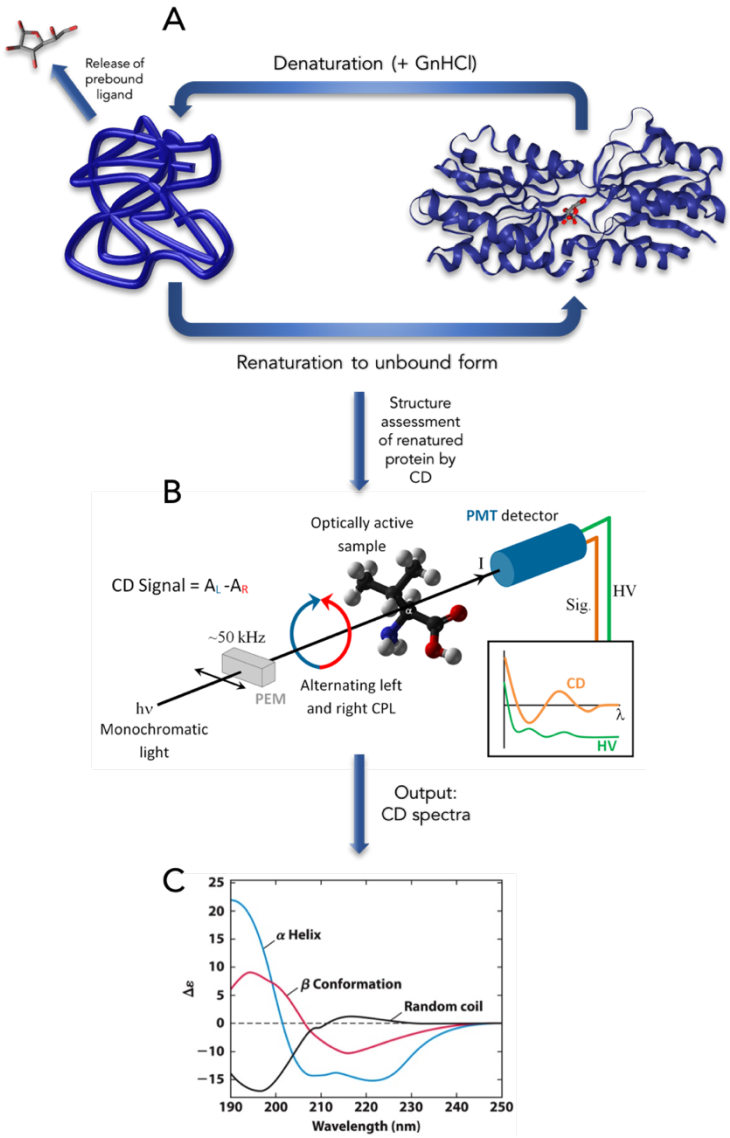


Figure 3. 4. Use of circular Dichroism for confirmation of renatured protein structure.

A) Treatment with GnHCl can release prebound ligands from purified proteins.

B) Function of Circular Dichromator: Monochromatised light goes through a Photo Elastic Modulator (PEM) which converts the linear polarised light into alternating (circularised) left and right handed polarised light. The two polarisations are differently absorbed when a chiral chromophore is present in their pathway. The difference in absorption is then detected by a Photo Multiplier Tube (PMT) and produces the CD signal (A_L-A_R).

C) Characteristic CD spectral signatures for α-helices (blue), β-sheets (red) and random-coil (black).

The structure of GafAEc is shown at (A) bound to D-galactofuranose. The figure at (B) was obtained from the ASTRID2 webpage of ISA facilities (AARHUS University). The CD spectra of the secondary elements was obtained from Lehninger *Principles of Biochemistry* (6th Edition 2013, H. Freeman and Company).

2008); therefore, the CD is a valuable tool to analyse changes in the protein conformation.

A circular dichroism spectrophotometer is also commonly termed as a circular dichroism spectropolarimeter or a circular dichrograph (Greenfield *et al.*, 2007). Inside the spectrophotometer there is a source of monochromatic linearly polarised light which becomes either left- or right- circularly polarised light (CPL) by passing it through an optical element known as the photo-elastic modulator (*ie*. PEM) (Figure 3.4B). This is a piezoelectric element which when in oscillation (typically around 50 kHz), induces stress to the attached silica component which in turn becomes a quarter-wave plate (Greenfield *et al.*, 2007). This in turn retards the linearly polarised light to produce left and right – polarised light at the drive frequency. On the other side of the sample position there is a light detector (*ie*. PMT) (Figure 3.4B). When there is a sample in the light path then the recorded light intensity is not constant, but instead it reflects the chirality of the molecule as the left- and right- CPL absorbed differ. By using an amplifier tuned to the frequency of the PEM, it is feasible to quantify the difference in intensity between the two CPLs (vAC). Also, the average total light intensity of PEM oscillations is taken into account to calculate the vDC, which accounts for any variations in the total light level. These two parameters, vAC and vDC, are then used to calculate the circular dichroism signal (Greenfield *et al.*, 2007).

3. 1. 3. 3 Intrinsic fluorescence of proteins

Fluorescence is comprised of two steps, the first one includes the absorption between electronic energy levels and the second one is the emission at a longer wavelength because of the vibrational energy disseminated to the surroundings (Lakowicz, 1999; Möller and Denicola, 2002). Since the absorption and emission have different detection wavelengths, the sensitivity of fluorescence assays is high and one can measure the fluorescence in the nanomolar range. Although remarkable progress has been achieved in the fluorescent labelling of proteins and the application thereof, several shortcomings still exist. The tedious and time-consuming procedures

encountered during labelling of the targets may lead to unpredicted complications during the experimental runs. It is generally accepted that some of the labels act by forming covalent interactions with the biomolecules, which can lead to structural and functional changes of the target protein. In order to avoid this drawback, one can utilize the intrinsic fluorescence of the proteins originating from its aromatic acid constituents to explore conformational changes upon ligand addition (Ghisaidoobe and Chung, 2014). The amino acids which exhibit characteristic ultraviolet fluorescence are the aromatics *ie*. tryptophan (Trp), tyrosine (Tyr) and phenylalanine (Phe) (Teale and Weber, 1957). Trp is the most abundant and is present at concentrations near 1 and close to 3 mol % in membrane proteins (Lakowicz, 1999). The phenyalanine contribution to the overall intrinsic fluorescence of a protein is insignificant, by virtue of its low absorptivity and very low quantum yield (Table 3. 1). Despite the similar quantum yield of Tyr and Trp, the latter features an indole group which is held responsible for its high molar absorption coefficient and slightly higher quantum yield (Table 3. 1). Therefore, the fluorescence spectrum of a protein, which contains all three amino acids, will usually resemble that of Trp *ie*. absorbance at 280 nm, and emission at around 350 nm (Teale and Weber, 1957). Exciting the protein at 280 nm ensures that both Trp and Tyr are excited and therefore the fluorescence spectrum of the protein could also be the product of the combined Trp and Tyr absorption, depending on the position and abundancy of these amino acids in the polypeptide. However, Trp absorbs 5 times more than Tyr at 280 nm, therefore the contribution of the former to the overall signal produced is substantial compared to the latter (Chen, 1972; Fasman, 1976). If the researcher wishes to obtain a signal mostly derived from the Trp residues then the protein can be excited at 295 nm; at such wavelength Trp absorbs almost 30 times more than the Tyr (Chen, 1972; Fasman, 1976). Excitation at 297 nm ensures exclusion of tyrosine contribution in the fluorescence signal (Chen, 1972). The intrinsic fluorescence of the protein becomes a useful tool in quantifying the binding affinity for ligands, only when Trp (and Tyr) residues are exposed on the surface and near the bind ing cavity of the SBPs. In such incidents, ligand addition shifts the positions of these residues significantly and gives

rise to quantifiable change in the fluorescence which can be titrated. The fluorescence intensity is linearly related to the concentration of the fluorophore only when dilute solutions of the fluorophore are included in the experiments; thus, the protein concentration is advised to be kept at 0.5 to 1.5 μM or at concentrations with less than 0.5 absorbance at the excitation wavelength (Möller and Denicola, 2001). In GHT lab, the determination of ligand affinities of GafAEc occurred using its intrinsic fluorescence at protein concentrations from 0.5 to 1 μM.

Table 3. 1 Fluorescence properties of aromatic amino acids in water (pH= 7).

	Lifetime (τ) (ns)	Absorption		Fluorescence (emission)	
		λ (nm)	Absorptivity (∈)	λ (nm)	Quantum Yield* (Φ_F)
Tryptophan	3.1	280	5400	350	0.2
Tyrosine	3.6	274	1400	303	0.14
Phenylalanine	6.4	257	200	282	0.04

*The quantum yield is defined as the ratio of photons absorbed to photons emitted through fluorescence, and gives the propability the excited state was deactivated by fluorescence and not by a non-radiative mechanism. The quantum yield of the aromatic amino acids was calculated as follows: the intensity of the fluorescence emitted at right angles to the direction of excitation is compared with the intensity of the light scattered in the same direction by a glycogen solution (quantum yield =1) (Teale and Weber, 1957).

3. 1. 3. 4 Isothermal Titration Calorimetry

The interaction between a protein and a ligand, causes heat release or absorption. The only technique able to measure this heat exchange is the Isothermal Titration Calorimetry (ITC). This technique is suitable for the investigation of the energetics of ligand binding as the analysis results in important thermodynamic information about the binding process.

The instrument is consisted of two identical cells, one known as the reference and the other is the sample cell. These are made of a highly efficient thermally conducting and inert material (*eg*. gold). The two cells are found embedded in an adiabatic jacket (Figure 3. 5A) (Wiseman *et al*., 1989). The ligand is placed in a syringe device, which injects the ligand into the sample cell during the titration and it acts as a stirrer to allow mixing of the protein with the ligand (Figure 3. 5A). During the titration, the heat released by the binding of the ligand causes a temperature change. This gives rise

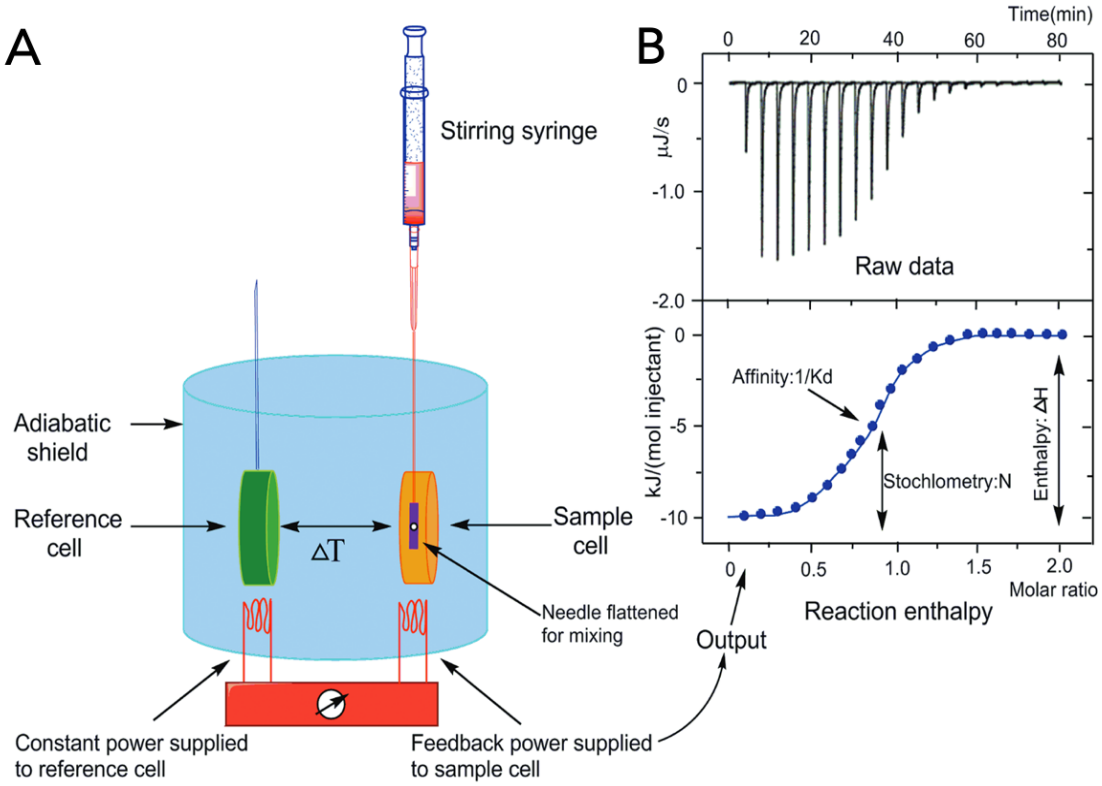


Figure 3. 5. Basic principle of Isothermal Titration Calorimetry (ITC).
(A) Schematic diagram of the components that constitute a isothermal titration calorimeter.
(B) The experiment produces an isotherm which includes the titration thermogram (upper section). The thermogram presents the heat per unit of time released following each injection of the ligand into the protein. An exothermic reaction releases heat and negative peaks are obtained. The individual peaks are intergrated by the software which produces a Wiseman plot (lower section). The heat released per mole of injectant is plotted against the molar ratio of the two reactants. An appropriate model is chosen and the isotherm is fitted to calculate the binding enthalpy ΔH, the binding constant Ka and the stochiometry (N). These data are then used to calculate the Gibb's free energy (ΔG) and entropy (ΔS).
The figure was taken from Song *et al.*, 2015.

to a difference in temperature between the reference and sample cell. The difference in temperature is measured and power is used to compensate and calibrate the sample cell back to the initial temperature in between the injections (Wiseman *et al.*, 1989). The differential power applied to the sample cells compared to the reference cell is monitored and the power is plotted as a function of time to produce a titration thermogram (Figure 3. 5B). The data analysis software produces a baseline and integrates the area occupied by each peak. The peak area is directly analogous to the heat liberated or absorbed upon ligand binding. The heat change is normalised to account for one mole of injectant and an ITC binding isotherm is produced by plotting the normalised heat against the molar ratio (Figure 3. 5B).

The binding isotherm is used to calculate the following:

a) stoichiometry (*n*): is equal to the molar ration at the inflection point of the isotherm.
b) affinity constant (K_a): K_a is calculated using the slope at the inflection point of the isotherm; depends on the chosen model.
c) enthalpy change (ΔH): defines the heat emitted (*ie*. negative ΔH) or absorbed (*ie*. positive ΔH) by a reaction at a constant pressure. This is equal to the y axis value at the lower plateau, provided the upper plateau (*ie*. 0 kj per mol) is reached.
d) Gibbs free energy (ΔG) and change in entropy (ΔS): ΔG defines the change in free energy between the reactants and the ligand bound complex produced, and measure the capacity of the system to do work. Negative ΔG means the reaction proceeds spontaneously whereas positive ΔG reactions require energy input. The ΔS defines the change between the reactants and the formed complex. A positive ΔS indicates that the disorder of the system is increasing. These two values are calculated by the Gibbs–Helmholtz equation, using the thermodynamic parameters mentioned above.

3. 1. 4 Objectives of the current chapter

The current study adapts a strategy aiming to obtain more information about L-arabinofuranose binding, the way itself has developed through natural evolution. The aim will be to gain insights into the selected candidates described in Chapter 2 (*ie*. GafASm and GafASw), by using thermodynamic and structural analysis. The kinetic analysis will be employed to identify whether the ligand binding specificity of GafASm or GafASw differs to the one GafAEc exhibits. Following identification of unique ligand binding specificities, the exact binding affinities (K_D) will be quantified for the ligands which presented the highest affinities. Next, a structural analysis will be employed to allow for a structure-function relationship study. The purpose will be to identify the unique features of the binding cavity for the GafA candidate with the most relevant binding specificity to the scope of this study. We hope that the search for an SBP which presents a higher substrate specificity and renders L-arabinofuranose binding, will ultimately pave the way to direct engineering of GafAEc for a more efficient utilization of the released L-arabinofuranose. Such engineering might decrease the ligand specificity and prevent competitive inhibition of L-arabinofuranose binding by sugars found in the lignocellulosic mixtures *eg*. D-galactose.

3. 2 Materials and Methods

3. 2. 1 Purification of ligand free protein using Guanidine hydrochloride (GnHCl)

3. 2. 1. 1 Buffers preparation

The wash buffer contained the following: 50 mM Tris-HCl buffer pH 7.8, 150 mM NaCl, 20% glycerol and 20 mM imidazole. Appropriate volume of 6 M Guanidine hydrochloride (GnHCl) stock was added in separate aliquots of the wash buffer, so that the final concentrations of the different aliquots were 2 M, 1.5 M, 1 M and 0.5 M GnHCl. The elution buffer was made up of 50 mM Tris-HCl buffer pH 7.8, 150 mM NaCl, 20% glycerol and 500 mM imidazole. The refolding buffer contained the same reagents with the wash buffer, mixed with 500 mM arginine monohydrochloride, 4 mM glutathione reduced and 0.4 mM glutathione oxidised. All reagents mentioned were sourced from Sigma-Aldrich Ltd.

3. 2. 1. 2 In-column denaturation and refolding

The periplasmic extract containing recombinant protein was filtered through a 0.2 μm filter. A 5 ml HisTrap column (GE Healthcare) was equilibrated with buffer A (50 mM Tris pH 8, 200 mM NaCl and 20 mM imidazole) and connected to an AKTA Purifier P-900. Filtered supernatant was then injected into a pre-equilibrated column. The column was washed with 2 CV of wash buffer (50 mM Tris, pH 8, 200 mm NaCl and 25 mM imidazole) to remove weakly bound contaminants. The bound protein was denatured with 30 CV of buffer containing 2 M GnHCl, followed by 4 CV of each of the wash buffer, mentioned in Section 3. 2. 1. 1, containing decreasing GnHCl concentrations. Finally, the column was washed with 5 CV reforlding buffer, followed by wash buffer protein and the protein was eluted with 50 mM KPi buffer, 200 mM NaCl, 20% glycerol, 500 mM imidazole pH 7.8.

3. 2. 1. 3 Denaturation by dialysis

For the purpose of comparing the different methods of GnHCl denaturation, the protein was dialysed overnight in 500 ml of denaturation buffer containing 50 mM Tris-HCl buffer pH 7.8, 150 mM NaCl and 2 M GnHCl, following its purification in the absence of GnHCl (see Chapter 2, Section 2. 2. 6. 2). Subsequently, one of the samples contained in the dialysis tube was transferred into a fresh dialysis buffer containing 50 mM Tris-HCl buffer pH 7.8 and 150 mM NaCl and a different one was placed in a fresh dialysis buffer containing refolding buffer and Tris-HCl/NaCl at the aforesaid concentrations.

3. 2. 1. 4 Denaturation by dilution

A single fraction obtained following dialysis in the presence of GnHCl, was refolded by addition of equal volume of refolding buffer. The protein sample was stored in -80 °C.

3. 2. 2 Biochemical and biophysical techniques

3. 2. 2. 1 Differential scanning fluorimetry

The thermal unfolding of GafASw and GafASm proteins in different buffers with or without the presence of ligand, was monitored using DSF. Separate fractions of the proteins were dialysed in the buffers described in the Table 3. 2.

Table 3. 2. Salts used in buffer and pH screening of GafAs

Buffer	Abbreviation	[Buffer] (mM)	[NaCl] (mM)	pH	pKa (at 25 °C)	pHed with
Tris sodium citrate	TrisNaCit	50	150	4.5	3.14, 4.76 and 6.4[a]	Citric acid**
Tris sodium citrate	TrisNaCit	50	150	5.5	3.14, 4.76 and 6.4[a]	Citric acid**
Bis-tris methane	BisTrisMeth	25	150	6.6	6.46[b]	HCl
HEPES, sodium salt	HEPES	50	150	7.2	7.5[b]	HCl
Phosphate buffered saline*	PBS	-	137	7.5	2.16, 7.21 and 12,32[c]	HCl
Trizma (HCl)	Tris	50	150	7.5	8.06[d]	NaOH
Bis-tris propane	BisTrisProp	25	150	8.2	6.81 to 9.2[b]	HCl
Glycine (sodium hydroxide)	Gly(NaOH)	100	150	9.5	9.78[e]	NaOH

*PBS contained 137 mM NaCl, 2.7 mM KCl, 10 mM Na_2HPO_4 and 1.8 mM KH_2PO_4.
**0.1 M solution of citric acid was used to pH the Tris sodium citrate. 500 ml were added in 1200 ml of 75 mM buffer to decrease pH to 4.5. 200 ml were added to 1000 ml of 60 mM buffer to reach pH = 5.5.
[a] The Merck Index 12th *ed.*, (Entries# 2387 and 8746);
[b] Dawson *et al.*, 1969;
[c] Gerhardt *et al.*, 1981;
[d] Good *et al.*, 1966; [e] Pearse, 1980.

The proteins were dyed with the fluorescent SYPRO orange supplied with the Protein Thermal Shift kit from Thermo Fisher Scientific (Applied Biosystems). The experiment was performed in a real - time PCR instrument *ie*. StepOnePlusTM Real Time PCR Systems. Protein - ligand solutions were dispensed into the wells of a 96 - well thin wall PCR plate (*ie*. Genomic Fast Optical 0.1 ml plates) which were supplied from Fisher Scientific. Each well contained 3 μM protein, 0.1 × (2.5 μl of 0.8 ×) SYPRO orange and appropriate volumes of ligands. Different stocks of the ligands were prepared, so that 2 μl were transferred in each well to reach desired concentrations of 0.6, 6, 60, 600 and 1200 μM. A final well volume of 20 μl was filled up by the protein's buffer. For each run, there were 8 reference wells where ligands were excluded. Also, 4 reference wells were included which did not contain protein or ligand but only buffer. The plates were sealed with optical sealing tape (Bio-Rad). The instrument run was set as instructed in the manual by the Applied Biosystems. The plate was heated up to 99 °C in increments of 0.5 °C.

The data analysis was undertaken using Microsoft Excel. Firstly, the representative temperatures applied to the wells throughout the timecourse were obtained by averaging the melt region temperature for each timepoint, *ie*. the temperatures applied on the wells of the plate were marginally different for each timepoint. This marginal variability in temperatures for each well was not taken into account in the Tm calculations, as the differences were down to the third decimal point, and the nature of the experiment was qualitative. The temperature averages for each timepoint were matched to the respective –d(RFU)/dt values from the melt derivative data. The minimum –d(RFU)/dt value for each replicate of the samples containing protein only was detected and the temperature at that point was noted. The average of the temperatures at the collected minima points was equal to the Tm of the protein without the presence of any ligand. The standard deviation for the average was calculated using the respective function in Microsoft Excel. The same analysis was applied for each replicate of the protein with added ligands at the various concentrations mentioned above. The ΔTm was calculated by subtracting the average Tm for the protein-only sample from each replicate of the individual ligand

concentrations. This resulted in 8 (or 4) ΔTm values for each concentration of the ligand tested, which were thereafter averaged to obtain the plotted ΔTm values. A t-test was performed to assess whether the mean values calculated for the ΔTms are equal. The ΔTm mean values were always compared to the one condition that caused the highest ΔTm in presence of the respective ligands. The analysis was performed in Microsoft Excel using the respective function by applying a two – tailed distribution with two – sample unequal variance. The percentage in decrease of stability was calculated by using the difference between the highest Tm (or ΔTm for the activity) and the condition in question. The difference was divided by: a) the Tm of the condition in question, for the stability or b) the ΔTm at highest activity and thereafter multiplied by 100 to calculate the percentage. The data were plotted in GraphPad Prism 7.0.

3. 2. 2. 2 Circular Dichroism (CD) spectroscopy

CD spectra analysis was used to assess the stability of GafASm fractions which were treated with GnHCl during purification, or in succeeding dialysis steps. The spectra were obtained using a J-810 spectropolarimeter (Jasco) controlled by the supplied software SpectraManager version 1.53.00 (Jasco). Protein was dialysed into 50 mM NaF and 20 mM Tris-HCl pH 7.5 and diluted to a concentration of 6 μM. The spectrum was recorded at 20 °C (Peltier temperature controller) in a 1 mm pathlength quartz cuvette (Starna) between 180 – 240 nm at 100 nm/minute with 1 nm pitch. The molar elipticity (θ) obtained from the values was corrected by subtracting the buffer control, and were plotted against the wavelength (nm) in GraphPad Prism 7.0.

3. 2. 2. 3 Fluorescence spectroscopy

Tryptophan and tyrosine fluorescence spectroscopy was performed using a FluoroMax 4 fluorescence spectrometer (Horiba Jobin-Yvon) with connected water bath to maintain temperature. The maximum emission for each protein was determined by spectra analysis *ie*. excitation at wavelengths 280, 295 and 297 nm at slit widths equal to 3 nm. This is further refined once a ligand which binds is found to maximise the

change in fluorescence upon binding.

The kinetic experiments for quantification of the binding affinities were performed in 3 ml quartz cuvettes (Starna) and purified protein was used at the concentration of 1.5 μM or 1 μM or 0.5 μM in PBS or Tris – HCl, pH = 7.5. Total volume of the sample was 3 ml, which was excited at 280 nm for GafASm and 297 nm for GafASw with slit widths of 3 nm. Emission was monitored at 330 (GafASw) or 342 nm (GafASm) with slit widths of 3 nm. The fluorometer operated in a time-based acquisition mode with a run time between 360-500 seconds and an integration time of 1 second. Increasing concentrations of ligand were added to the protein solution at intervals of 25-50 secs to ensure stabilisation of fluorescence levels following each ligand injection. The last injections of each run ensured that protein reached saturation as there was no further change in the fluorescence signal; therefore, the cumulative ligand concentration depended on the K_D of each protein for it. Any quantifiable fluorescence change was noted and transferred to Microsoft Excel for data analysis. The cumulative fluorescence change for each timepoint was plotted against the cumulative concentration of ligand in SigmaPlot 11. The data were fitted to a single rectangular hyperbola, using the equation below and non-linear regression (Michaelis-Menten model) was used to calculate the dissociation constant (K_D).

$$y = \frac{\alpha x}{b + x}$$

At the equation above the change in fluorescence is represented by *y* and *x* represents the cumulative concentration of ligand. The *b* is the dissociation constant (K_D) and a represents fluorescence when curve flattens at high concentration of *x*. The presented binding saturation graphs were plotted in GraphPad Prism 7.0. Representative table of the data used to calculate the dissociation constant (K_D) of arabinose to GafASw is given below in Table 3. 2 as an example.

Table 3. 3 Representative Fluorescence data for binding of L-arabinose to GafASw.

[L-arabinose] added (µM)	Cumulative [L-arabinose] added (µM)	Average fluorescence at each injection (cps)	Change in fluorescence at each injection (10000 cps = 1)	Cumulative change in fluorescence for each injection (10000 cps = 1)
0	0	22729194	0	0
2.5	2.5	23083231	35.4	35.4
2.5	5	23297661	21.4	56.8
2.5	7.5	23409830	11.2	68.1
2.5	10	23479764	6.99	75.1
2.5	12.5	23525496	4.57	76.9
2.5	15	23555160	2.97	82.6
2.5	17.5	23573586	1.84	84.4

3. 2. 2. 4 Isothermal Titration Calorimetry (ITC)

Calorimetry experiments were performed in the VP-ITC instrument (MicroCal Inc., GE Health Sciences). During a typical ITC run, the calorimeter cell was loaded with 1.4 ml of protein, and the syringe contained 300 µl of ligand. The ligand concentration in the syringe was in 10- or 7- times excess compared to the protein in the cell. The concentration of the protein in the cell was chosen according to c value, where c = [protein]/ (predicted) K_D. Experiments were carried out in PBS or Tris-HCl (NaCl), pH 7.5 at 25°C. The solutions in the cell and syringe were both degassed at 20 °C for 10 min before use. A typical run included 27 titrations which were performed as follows: the first preliminary injection of 2 µl of ligand solution was followed by 26 injections of 10 µl each, delivered at an injection speed of 10 µl s^{-1}. Each injection was added at 3 min intervals with a stirring speed of 307 rpm. The acquired raw titration data were analysed in MicroCal Origin 7 software where binding isotherms were fitted by an iteration process using the one-set of sites model. Baseline adjustment is done manually for each measurement. The goodness of the fit was determined by an algorithm in the software (least square method). The binding isotherm is then observed by integrating the obtained peaks and by plotting the resulting heats of each injection against the ratio of titrant to the protein concentration. The values for K_a and ΔH can be determined using the equations for single set of identical sites as

explained in the "MicroCal Tutorial Guide for fitting ITC Data" booklet.

3. 2. 2. 5 Size Exclusion Chromatography Multi-Angle Laser Light Scattering (SEC-MALLS)

Size exclusion chromatography (SEC) coupled with multi angle laser light scattering is a technique, which can determine the molecular weight of proteins and their complexes in solution (Folta-Stogniew and Williams, 1999). The apparatus includes a fractionation device, a multiangle light scattering (LS) detector and a concentration-calculating device (*ie*. refractive index detector) connected in series. Light passing through solvent interacts with the macromolecule (protein), and the interactions leads to scattering off the axis of the incident light beam. The LS detector uptakes the signal and determines the molecular mass, molecular root mean square (rms) radius size and aggregation state of the protein. The refractive index (RI) detector measures the concentration of the protein which is then used in the estimation of the protein molar mass. The masses are determined for each fraction eluted separately.

The experiment was conducted at room temperature (20 ±2°C). Solvent was 0.2 µm filtered before use and a further 0.1 µm filter was present in the flow path. The Superdex 200 10/30 GL SEC column (supplied from GE Healthcare) was equilibrated with at least 2 column volumes of solvent before use and flow was continued at the working flow rate until baselines for UV, light scattering and refractive index detectors were all stable. The samples of the GafASm (120 µl of each 0.5 mg/ml, 1 mg/ml, 2 mg/ml and 4 mg/ml present in 50 mM Tris, 150 mM NaCl at pH=7.5) were injected and separated on the Superdex column of the Shimadzu HPLC system (at 0.5 ml/min). Wyatt Dawn Heleos LS detector recorded the light scattering data and the Wyatt Optilab rEX refractive index detector with an SPD-20A UV detector assessed the concentration of the sample. Shimadzu LC Solutions software was used to control the HPLC and Astra V software for the HELEOS-II and rEX detectors. The Astra data collection was 1 minute shorter than the LC solutions run to maintain synchronisation.

Blank buffer injections were used as appropriate to check for carry-over between sample runs. Data were analysed using the Astra V software.

MWs were estimated using the Zimm fit method with degree 1. A value of 0.172 ml/g was used for refractive index increment (dn/dc) based on a control sample of BSA.

3. 2. 2. 6 Protein crystallisation and structure determination

The Hydra-96 Microdispenser (Robbins Scientific Corporation) and the Mosquito Crystal (TTP Labtech) robots were used to dispense commercially sourced crystallisation solutions and the protein-ligand solution containing GafASw with L-arabinose. The crystallisation screen prepared was in the vapour-diffusion sitting–drop method containing 150nL of crystallisation solution and 150nL of GafASw and L-arabinose at the final concentrations of 8mg/mL and 1.25mM respectively. The crystallisation condition which produced the crystal used to solve the structure of this protein was 0.2M Ammonium Nitrate, pH 6.2, 20% PEG 3,350 (B7 of the PEG/ION HT tray, Hampton Research). This crystal (Figure 3.6A) was harvested from the sitting–drop and coated in a solution of the aforementioned crystallisation solution supplemented with glycerol as the cryo-protectant, to a final concentration of 20% (v/v). Reflection data were collected to a resolution of 1.7 Å using the macromolecular crystallography (MX) Beamline I03 at the Diamond Light Source, Hartwell Science and Innovation Campus and processed with DIALS (Figure 3.6B). A Matthews coefficient of 2.1 Å3 /Da was calculated with a solvent content of 42.04%. The CCP4i2 suite of programs was used to scale the data and solve the structure. The processed DIALS data was scaled using AIMLESS, POINTLESS, Ctruncate and FreeRflag (Evans, 2011; Evans and Murshudov, 2013). Indexing permitted the determination of the crystal system, the unit cell dimensions and assignment to each spot on the image an index, quoted as three integers: h, k, and l. Data collection and processing statistics are summarised in Table 3.4. The structure was solved by molecular replacement using MOLREP (Vagin and Teplyakov, 2010; Lebedev *et al.*, 2008) using the coordinates for

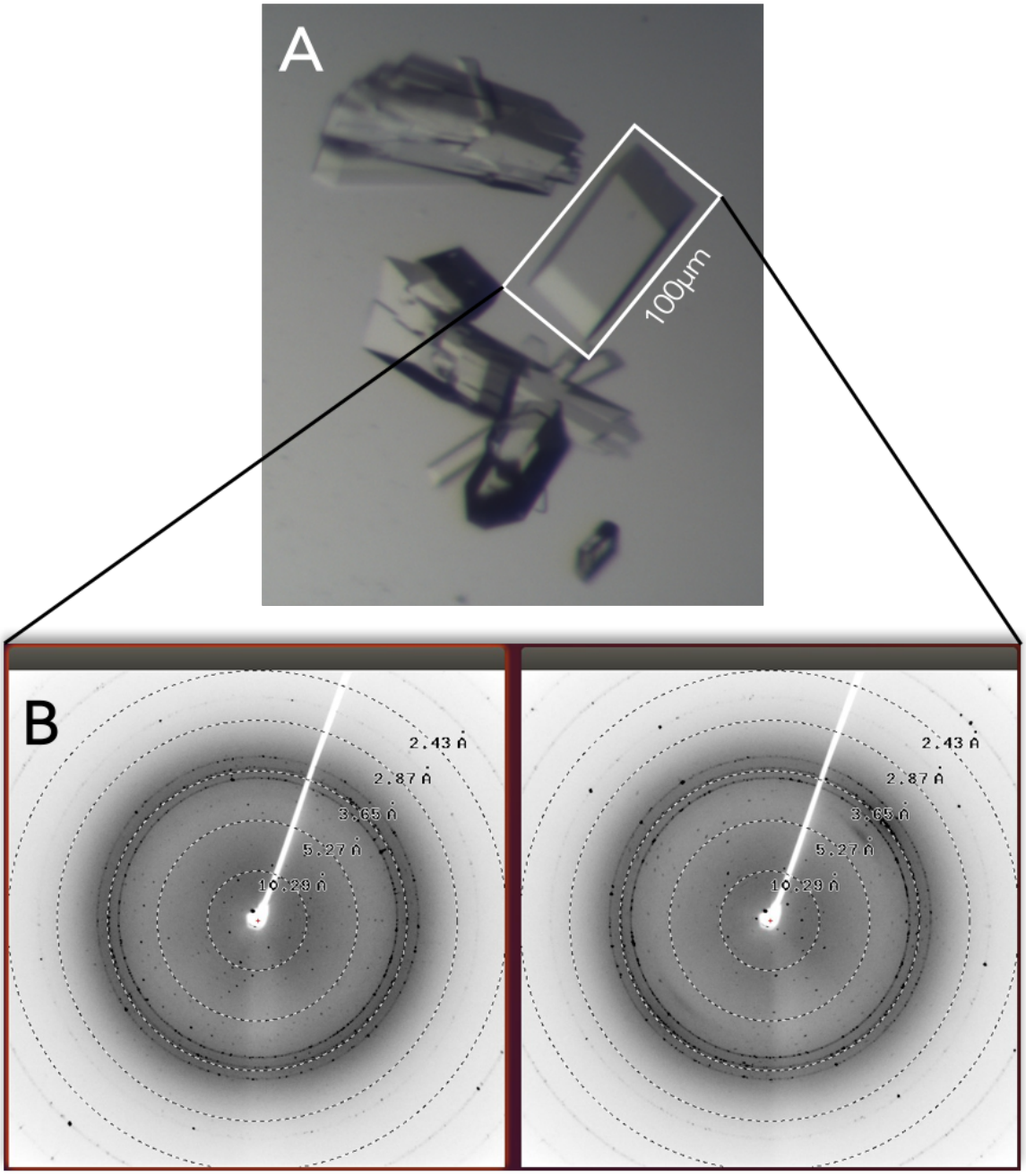


Figure 3. 6. Crystallisation of GafASw.

A) Image of the crystal which produced the GafASw structure during crystallisation trials. The well shown is B7 of the PEG/ION HT tray which contained 0.2M Ammonium Nitrate, pH 6.2, 20% PEG 3,350.

B) Diffraction of crystal shown at (A) obtained at Beamline I03 at the Diamond Light Source.

an *Escherichia coli* galactofuranose binding protein, YtfQ (PDB: 2VK2) as the search model. The initial model built by MOLREP was then refined using continuous cycles of REFMAC5 (Murshudov *et al.*, 2011; Vagin *et al.*, 2004; Nicholls *et al.*, 2012) followed by manual model building in COOT (Emsley *et al.*, 2010) until the final structure was obtained. The final structure was then deposited with the PDB ID 5OCP via wwPDB (Berman, Henrick and Nakamura, 2003).

Table 3. 4. Parameters for data collection

Data collection for GafASw (co-crystallised with L-arabinofuranose/ PDB ID: 5OCP)	
Space group	$P2_12_12_1$
Cell constants:	
a, b, c (Å)	73.92, 86.33, 87.28
a, b, γ (°)	90.00, 90.00, 90.00
Resolution (Å)	43.64-1.70 1.73-1.70
% Data completeness (in resolution range)	100 (43.64-1.70) 99.9 (1.73-1.70)
R_{merge}	0.08
I/σ(I)	3.77 (at 1.7Å)
Data redundancy	7.4
Refinement	
Refinement program	REFMAC 5.8.0158
Resolution (Å)	43.45 – 1.70
No. of reflections (for free R test set)	58975 (3081)
R_{work}/ R_{free}	0.17/ 0.20
Wilson B-factor (A^2)	18.0
Total number of atoms	5171
RMSZ – Protein	
Bond lengths	1.01
Bond angles	0.98
RMSZ – Ligand (Average)	
Bond lengths	1.11
Bond angles	1.38

3. 3 Results

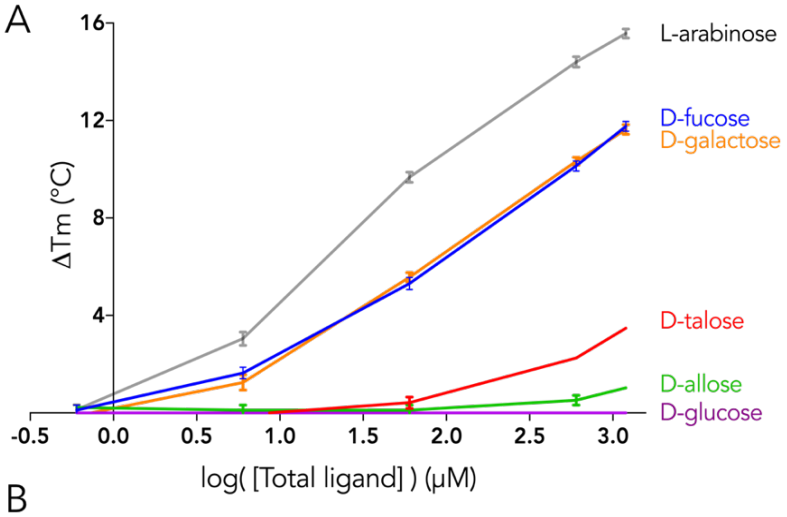
3. 3. 1 GafASm shows similar ligand specificity to GafAEc

The first step in the biochemical analysis of the GafASm was the qualitative assessment of binding to determine the ligand specificity of the protein. This was performed using Differential Scanning Fluorimetry (DSF); a rapid and inexpensive screening method to identify the ligands that bind and stabilise the purified protein (See Introduction *Section 3. 1. 4. 1)*.

The experiment was performed as described in the Methods and Materials *Section 3. 2. 2. 1.* As the data would be solely used in a qualitative manner to analyse the ligand binding, the choice of the protein buffer was not taken into consideration. Instead, the GafASm was dialysed in a commonly used phosphate buffer *ie*. Phosphate buffered saline (PBS- 137 mM NaCl, 2.7 mM KCl, 10 mM Na_2HPO_4 and 1.8 mM KH_2PO_4), maintained at near neutral pH (7.5). For each ligand concentration, 8 replicates were included in an effort to increase statistical significance of the results. As the number of the wells in the microplates was inadequate for the number of replicates included, two separate runs were performed to test the full spectrum of target ligands.

The results showed that the melting temperature of the unbound form of GafASm was 52.4 °C at the first run and 56 °C at the second (Figure A3. 1). The presence of L-arabinose caused the largest Tm change as it increased the thermal stability of the protein by 27.4 % (*ie*. increase of 14.4 °C), at 1.2 mM total sugar concentration (Figure 3. 7). The extent of the temperature shift is believed to be proportional to the affinity of the ligand for a given protein (Matulis *et al.*, 2015; Bullock *et al.*, 2015), provided that the ligand concentration exceeds its K_D value (Matulis *et al.*, 2015). Thus, this result suggests that GafASm binds L-arabinose with the highest affinity compared to the rest of the ligands tested. Furthermore, the presence of D-fucose increased the thermal stability of GafASm by 11.7 °C, a change very similar to the one induced by D-galactose (*ie*. 11.6 °C at 1.2 mM [sugar]) (Figure 3. 7, S3. 1), suggesting that the two ligands bind with similar affinities. The D-talose and D-allose caused considerably lower ΔTm

A



B

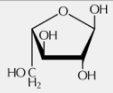
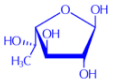
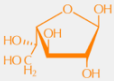
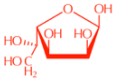
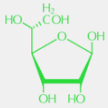
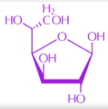
Monosaccharide	Structure	ΔTm (°C)	[Furanose] (μM) present at 1.2 mM total sugar	K_D (μM) for GafAEc
β-L-arabinofuranose		14.44	150	1.3 ± 0.24
β-D-fucofuranose		11.76	60	6.2 ± 0.54
β-D-galactofuranose		11.6	84	1.7 ± 0.3
β-D-talofuranose		3.47	348	25.7 ± 2.5
β-D-allofuranose		1.02	102	30.3 ± 4.4
β-D-glucofuranose		NC	1.68	NB

Figure 3. 7. The ligand binding specificity of GafASm resembles that of GafAEc, recognising both arabinose and galactose with high affinity.

(A) The DSF analysis of GafASm was perfomed in a 96-well plate format using a real-time PCR system. The binding of the ligands was tested in PBS buffer at tenfold range of dilutions, apart from the highest concentration. Each concentration included 8 replicates. The data were plotted as the change in the Tm of the unbound form of GafASm compared to the Tm of the bound forms. The $\log_{10}$ of the concentrations tested is plotted on the x axis.

(B) The ΔTms were tabulated next to the respective structure of the sugars tested. The order of ligands presented is as determined by the ΔTm at 1200 μM of total sugar concentration. The respective furanose concentrations at 1200 μM are also shown; the calculated values are based on their relative percentages in aqueous solution at 25 to 31 ◦C. The K_D of GafAEc for the ligands, which was previously calculated using intrinsic fluorescence, is also provided for comparison.

All structures were produced in ChemDoodle. NC = There was no change observed in the Tm of the protein supplied with ligand/ NB = No binding detected due to no apparent change in fluorescence upon addition of ligand.

compared to the aforementioned ligands, suggesting that GafASm exhibits lower affinities towards them (Figure 3.7). Similar to GafAEc, GafASm doesn't bind glucose as there is no change in its thermal stability in the presence of this sugar, even up to 1. 2 mM. The furanose form of glucose is found in negligible concentrations (<0.1%) in aqueous solution due to its unstable nature caused by the *cis* interaction between the 3OH and the 4-substituent (Figure 3.7) (Sinnott, 2007); the same principle applies for the xylofuranose (not shown). However, when arabinose and galactose are considered, their pyranose forms are destabilised by a single axial –OH at position 4, whereas the furanose forms have all the substituents (*ie*. positions 3 and 4 of the ring) *trans* to each other (Figure 3.7); therefore, they present higher levels of furanose compared to xylose and glucose (Sinnot, 2007). The profound scarcity of D-glucofuranose form in solution confirms the selectivity of the GafAs towards the furanose form over the pyranose form of sugars, as both GafAEc and GafASm remain inert in the presence of D-glucose (Horler *et al.*, 2009). Such assumption is further enhanced by the bacterial D-galactopyranose-specific SBP, *ie*. MglB, which is able to recognise and bind D-glucopyranose (Zukin *et al.*, 1977); firstly because of the similarity between the pyranose forms of the galactose and glucose ligands and secondly due to the mere fact that the pyranose form of glucose is seen in great abundancy by the bacteria.

The values for the ΔTms calculated here suggest the order of ligand preference of GafASm, which shows great resemblance to that known for GafAEc (Figure 3.7) (Horler *et al.*, 2009). The order of ligand preference for GafAEc is derived from the binding affinities of the abovementioned ligands as calculated by intrinsic fluorescence experiments (Horler *et al.*, 2009). The only exception in this observation is the binding of D-galactose and D-fucose, as the latter exhibits 6 times lower affinity for GafAEc, when compared to the first. In contrast, the change in the denaturation Tm of GafASm upon presence of D-galactose is very similar to D-fucose.

3. 3. 2 CD analysis of GnHCl-pretreated GafASm

Since SBPs have high affinity to their respective ligands, and generally exhibit a very low off-rate for the ligand-binding equilibrium (Miller *et al*., 1980), highly purified proteins very often contain a bound ligand (Miller *et al*., 1980; Severi *et al*., 2005). For any downstream analysis including biochemical and structural studies the prebound ligand needs to be removed. Because the SBPs are typically very stable and easily refolded, reversible denaturation with GnHCl (Miller *et al*., 1983) is used followed by extensive washes during purification and thorough dialysis following purification.

Purified fractions of GafAEc contained 40% of the protein prebound to D-galactose (Horler *et al*. 2009). The hypothesis that D-galactose could be present in the rich overproduction media (*ie*. LB) was proved wrong as the co-purification of the GafAEc with D-galactose persisted when the protein was overproduced in M9 minimal media supplemented with glucose (Horler *et al.*, 2009). Therefore, the D-galactose is expected to be sourced endogenously potentially via a novel metabolic route. It is established that in sugar starvation conditions, *E. coli* produces minimal amounts of galactose intracellularly (Death and Ferenci, 1994). The endogenously accumulated pool of galactose (around 200 μM) is sufficient to act as an inducer and activate the expression of *mgl*/*gal* operons which in turn actively assimilate and catabolise glucose (Death and Ferenci, 1994). One can assume that glucose, unless exogenously supplied, is not found in LB media, as described by Miller (1987). However, a study which examined the physiology of *E. coli* during growth in LB broth, has concluded that this media contains less than 100 μM fermentable sugars which are utilizable by *E. coli* (Sezonov, Joseleau-Petit and D'Ari, 2007). Therefore, it is possible that the scarcity in sugars, and potentially D-glucose, would cause the accumulation of D-galactose as described by Death and Ferenci (1994). Although the assumptions mentioned require concrete investigation, the fact that GafAEc was purified prebound to D-galactose is an adequate indication for us to seek to purify ligand – free GafASm, as the latter was also shown to bind D-galactose (Section 3. 3. 1).

Ligand-free proteins can be obtained by reversible denaturation with guanidine-hydrochloride (GnHCl), followed by dialysis against a large volume of ligand-free

buffer (Miller *et al*., 1983). This approach is commonly used in the purification of ligand/free di-/tri-peptide-specific SBPs (*eg*. OppA, MppA, DppA), as these peptides are found in ample amounts in tryptone which is added in LB media as a carbon source (Maqbool *et al*., 2011; 2012). The GnHCl is thought to act by engaging in transient stacking interactions with planar π-systems of the side chains of the aromatic amino acids which causes displacement of water from the hydration surface of the protein (Mehrnejad *et al*., 2010). Molecular dynamic simulations have shown that GnHCl is more efficient than urea in inducing denaturation of the polypeptides by destabilizing the α-helices (Camilloni *et al*., 2008), therefore GnHCl is considered a stronger reagent for the denaturation of α-helix rich protein, such as SBPs. Dissimilar to GnHCl, the simulation proved that urea acts by mainly destabilizing the β-sheets of the tertiary structure of the proteins (Camilloni *et al*., 2008). This was confirmed by the study of Mehrnejad and his coworkers (2010), which showed that GnHCl was almost three times faster than urea in eradicating the helicity of the peptide (Mehrnejad *et al*., 2010). As the *koff* rate of D-galactose from either GafAEc or GafASm are unknown, the treatment with GnHCl was performed during purification of the protein or during overnight dialysis following purification, which both provide enough time to facilitate displacement of the ligand (Materials and Methods, *Section 3. 2. 1*). Treatment by dilution is considered a faster denaturation technique and therefore was avoided as it could cause release of the ligand only for part of the co-purified fraction. Following successive washes with decreasing concentrations of GnHCl (*ie*. 2 M, 1.5 M, 1 M and 0.5 M) inside the Ni^{2+} purification column, the protein was also washed with refolding buffer. Likewise, the purified protein fraction which was dialysed in the presence of GnHCl (and PBS or Tris/HCl, pH=7.5), was subjected to an extra dialysis run in refolding buffer (and PBS or Tris/HCl, pH=7.5). The refolding buffer included L-arginine which is thought to have a stabilizing effect on the purified protein and prevents its aggregation during renaturation back to its native conformation (Shiraki *et al*., 2002; Reddy, 2005; Taneja and Ahmad, 1994). The positive effects of arginine in avoiding aggregation is a product of its characteristics which resemble the action of GnHCl in protein denaturation. The similarities in the properties of the above

reagents don't lead to the same abilities, as binding of L-arginine on the protein surface is far more limited compared to GnHCl, therefore the former only suppresses intermolecular interactions rather than destabilizing the protein in a fashion the latter operates in (Arakawa *et al*., 2007). Additionally, the refolding buffer included reduced glutathione in combination with its oxidised form which was added at 10 times less concentration than its reduced counterpart. Such mixtures facilitate the reformation of the disulphide bonds during refolding (Okumura *et al*., 2011). GafASm has both residues which form a disulphide bond in GafAEc (*ie*. between Cys 127 and 191) conserved and therefore it is predicted to form one itself. The protein was not treated with reducing agents such as β-mercaptoethanol or DTT, so the disulphide bond is expected to remain intact during protein denaturation by GnHCl. However, the glutathione mixture was still included in the refolding buffer to reverse any spontaneous reduction of the disulphide bonds, as these were shown to be susceptible to reduction by mild alkaline conditions (Florence, 1980).

All the GnHCl – treated fractions, were analysed with circular dichroism to detect whether they refolded properly (Figure 3. 8). Following normalisation of the obtained data to account for minor differences in the concentration of the samples, the spectra of all of the fractions presented the same profile to the untreated purified fraction (Figure 3. 8). This is indicative that the fractions have all reverted back to their native conformations and can be used in the downstream thermodynamic analysis. Such analysis proved valuable in verifying that the protein is present in its native conformation prior to undertaking extensive and expensive experiments with it. Additional to the investigation of renaturation of the GafASm, its stability and activity in different buffers and pH was also examined.

3. 3. 3 Screening of GafASm stability and activity in acidic and alkaline buffers.

The cytoplasm has been shown to be able to rapidly recover its pH back to physiological levels when the acidity of the extracellular space was intentionally decreased by addition of HCl (Wilks and Slonczewski, 2007). However, in the same

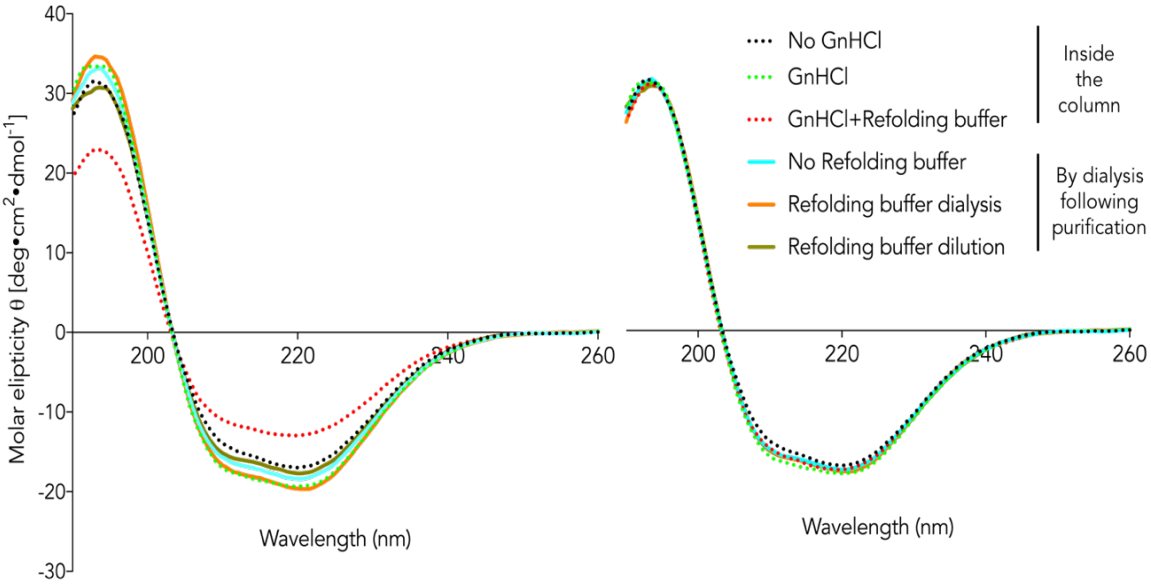


Figure 3. 8. Far UV circular dichroism spectra of GnHCl pretreated GafASm.
CD analysis was performed on GafASm following treatment with GnHCl denaturant for the release of any prebound galactose. The GnHCl treatment happened either during the Ni^{2+}-based purification run, with or without the presence of refolding buffer (dotted line); or by overnight dialysis with 2M GnHCl (continuous line), followed by dialysis (orange) or dilution (olive green) in refolding buffer. The blank-corrected data are plotted in the spectra shown on the **left**, whereas the normalized data correcting for protein concentration are shown on the **right**. The analysis shows that all fractions retain their secondary structure elements intact, indicating that the full fraction of GafASm refolds back to its mature conformation following treatment by GnHCl. The obtained spectra are similar to CD plots of proteins rich in α-helices *ie*. a common feature of SBPs.
The run was performed on J-810 spectropolarimeter with 6 µM of GafASm in 50 mM NaF and 20 mM Tris-HCl (pH=7.5). Graph was plotted in GraphPad Prism 7.0.

conditions periplasmic acidification persisted and never recovered, indicating that its pH is primarily determined by the pH of the extracellular environment (Wilks and Slonczewski, 2007). In a biotechnological context, the acidification or alkalinisation of the periplasm of bacteria during fermentation is very likely as the lignocellulosic biomass is pre-treated with H_2SO_4 or NaOH. A robust sugar-binding SBP for such application will be required to retain its activity in fluctuating pHs which are defined by the extracellular environment (Wilks and Slonczewski, 2007). Thus, the search for efficient transport systems of xylan-derived sugars should take into account the stability and activity of SBPs in acidic or alkaline conditions. For this reason, we employed DSF to study the stability and activity of GafASm across the pH spectrum, in an effort to identify the conditions were the protein is retaining full binding capacity of L-arabinose and D-galactose. The screen also served for detection of a suitable buffer since it is known that the ionic strength of the buffer can modulate the magnitude of the electrostatic interactions and consequently has the capacity to influence the binding of the ligands (Papaneophytou *et al*., 2014). Such information is therefore useful for the binding assays to follow, so that the determination of the binding affinities is deemed trustworthy.

The stability of the GafASm was reflected by how high its Tm was in the conditions tested. This is a common indicator of protein stability as it has been demonstrated multiple times that augmented thermal stability is correlated with high structural order and low protein flexibility (Ericsson *et al*., 2006; Vedadi *et al*., 2006; Nettleship *et al*. 2008). Also, this is frequently an indicator of increased conformational homogeneity of the protein, therefore the assay can give indications about the monodispersity of the sample (Vedadi et al., 2006; Nettleship *et al*. 2008;).

3. 3. 3. 1 GafASm partially retains stability but loses activity at highly acidic conditions

Separate fractions of GafASm were dialysed in buffers with acidic pH range (Table 3. 2, *Section 3. 2. 2. 1*). The choice of the buffers was based on their pK_a which defines their effective pH range *ie*. the range where the buffering properties that maintain the

pH of the solution are active (Blackwell, 1954; Joseph, 1958). The buffers used were Tris sodium citrate (TrisNaCit) for pHs 4.5 and 5.5 and Bis-Tris methane (BisTrismeth) at pH 6.6. HEPES buffer (pH = 7.2) was also included as a representative of near-neutral conditions. Conditions near the isoelectric point of the GafASm (*ie*. pI = 4.93) were avoided as they could cause precipitation of the protein. All buffers were supplemented with 150 mM NaCl. The rhizosphere normally features higher salinity than bulk soil, as the roots actively remove water and trace elements which are essential to their survival but exclude sodium salts which are toxic to their growth, therefore causing salinization (Miller and Wood, 1996). It has been suggested that the rhizobia species have adapted to a salt tolerance near 300 mM, which is the minimal level of tolerance that can support symbiosis in the root (Roumiantseva *et al*., 2011). Specifically, *Sinorhizobium meliloti* 1021 is able to grow on LB (*ie*. ≈ 84 mM NaCl) and at NaCl concentrations near 300 mM; however, exhibiting poor growth rates at 400 mM (Miller-Williams, 2006). Nonetheless, the periplasm of *S. meliloti* 1021, might not see the elevated levels of the extracellular NaCl in the rhizosphere due to the transport activity of BetC. This transporter accumulates high levels of glycine betaine inside the cell which in turn acts as an osmoprotectant (Roumiantseva and Muntyan, 2015). The *bet* genes are active at basal levels in normal salinity levels but their expression is increased as a response to osmotic stress induced by elevated levels of NaCl (Roumiantseva and Muntyan, 2015). Taking these into consideration, we decided to include 150 mM NaCl in the buffers, a value halfway between the NaCl concentration found in LB and in a halophilic environment (*ie*. 340 -400 mM) (Olliveier *et al.*, 1994), such as rhizospheres.

The investigation was performed by DSF analysis as described in Materials and Methods, *Section 3. 2. 2. 1.* Eight replicates were included for each buffer condition and each ligand. Also, wells without any protein added and having only buffer were included in the run as a negative control. The results of the data analysis are shown in Figure 3. 9. The protein was most stable at buffer TrisNaCit (pH = 5.5) as it presented the highest Tm compared to the rest three buffers (Figure 3. 9A). Similarly, high Tms are attained from GafASm when present in moderately acidic conditions, *ie*.

BisTrisMeth and HEPES (Figure 3. 9A). Only in highly acidic conditions, *ie*. pH = 4.5, did the protein show a somewhat decrease, in Tm (*ie*. 7.26 °C) and in its stability compared to pH = 5.5. The negative impact of highly acidic conditions seen here is supported by the fact that increased hydrogen-ion concentration has a strong viability effect and reduces the symbiosis capabilities of *Sinorhizobium meliloti* (Lowendorf, 1981; Graham *et al*., 1982).

The activity of the GafASm was heavily reduced in very acidic conditions at pH = 4.5 (Figure 3. 9B, S3. 2). The difference between the ΔTm at pH = 4.5 and pH = 6.6 was shown to be highly significant when compared by t-test ($p = 2.74e^{-15}$). There was approximately a loss of 12 °C on the ΔTm for both arabinose and galactose, which amounts to a notable 32 % decrease in activity and therefore underlines the extend of the consequences of acidic conditions. However, less acidic conditions are better representatives of the pH changes that occur in the rhizosphere; particularly in the rhizoplane of the nitrogen fixing leguminous host plants. As these plants fix nitrogen in the form of ammonia ions (*ie*. NH_4^+), they release protons which in combination with secreted organic acids decrease the pH of the rhizosphere (Hellweg, Pühler and Weidner, 2009). One study measured the pH in the leguminous plants rhizosphere following acidification and found that it ranges from 5.3 to 5.5 (Nye, 1981). At such conditions, the activity of GafASm appears decreased albeit not in the same extend as in pH = 4.5. In agreement with this observation is the reduced growth of *S. meliloti* 1021 observed at pH 5.75 compared to neutral conditions (Hellweg *et al*., 2009). Therefore, in the event of acidification, the GafASm is still able to bind arabinose present in the released mucilage, albeit with less efficiency which is agreeable with the decline in its ability to form symbiosis at lower pH. The activity of the GafASm appears unaffected at higher pH, as the ΔTms for L-arabinose are insignificantly different (p = 0.201) between HEPES and BisTrisMeth (Figure 3. 9B).

A

Buffer	pH	Tm	S. D.
TrisNaCit	4.5	45.47	± 0.20
TrisNaCit	5.5	**52.73**	± 0.20
BisTrisMeth	6.6	49.89	± 0.18
HEPES	7.2	49.93	± 0.14

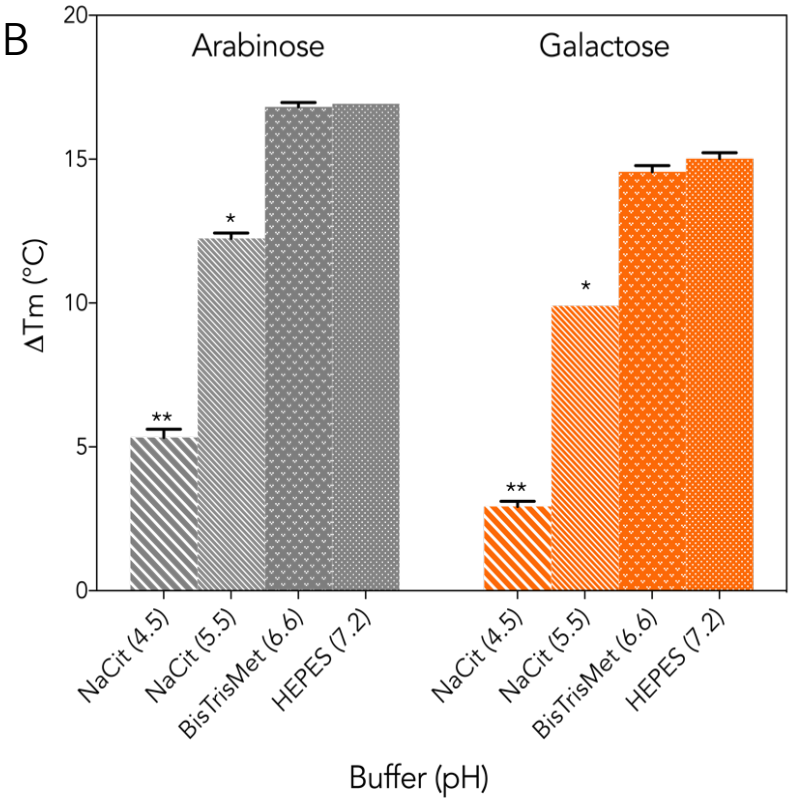


Figure 3. 9. GafASm shows partial loss of activity but retains stability at low pH. The GafASm was dialysed in different acidic buffers and its stability was tested using DSF. Its ability to bind L-arabinose and D-galactose was also tested as a measure of its activity in acidic environment.

(A) Tabulated data from the DSF analysis of GafASm present in acidic and neutral buffers ranging from pH 4.5 to 7.2. The stability of GafASm in the different buffers was assessed based on its Tm. Notably, the stability at the lowest pH tested (*ie*. Tris sodium citrate, pH = 4.5) is lowered compared to less acidic buffers.

(B) The change in the Tm upon binding of arabinose and galactose was used as a measure of the protein activity. In very acidic pH (*ie*. Tris sodium citrate, pH = 4.5) the activity is heavily impacted as the ΔTm is only one third for arabinose, and one fourth for galactose when compared to the ones measured in HEPES.

Eight replicates were included for each of: protein only, protein + D-galactose, protein + Larabinose, buffer only. The ligands were supplied at the highest previously tested concentration *ie*. 1200 µM. All buffers were supplemented with 150 mM NaCl. Asterisks represent the significance of difference of the data when compared to HEPES buffer, *$p < 0.05$, **$p < 0.001$.

3. 3. 3. 2 GafASm maintains activity but slightly loses stability at highly alkaline conditions

The DSF analysis of stability and activity in alkaline buffers proceeded in a similar manner as described above in *Section 3. 3. 3. 1.* The buffers used were PBS and Tris (base) at pH = 7.5, Bis-Tris propane (BisTrisProp) at pH 8.5 and Glycine (NaOH) at pH =9.5 to cover the lower alkaline spectrum. All buffers were supplemented with 150 mM NaCl as explained above in *Section 3. 3. 3. 1.*

GafASw retained its stability across the pH tested, however it presented a partial loss (13%) at pH = 9.5 (Figure 3. 10A, S3. 3). The activity was the highest at BisTrisProp (pH = 8.5) however it was reasonably maintained across all basic pH tested. When compared to BisTrisProp (pH = 7.5), the protein lost activity ranging from 10% to 12% in the rest of the pHs examined.

The ability of this SBP to retain good stability and activity across alkaline conditions could be attributed to its requirement in colonization of the plant root. Even though, the extent to which the presence of L-arabinose and D-galactose promote maintenance of growth of *S. meliloti* in natural soil varies, based on the microbial competition and availability of these sugars, D-galactose presence was indeed shown to enhance the growth of this bacterium (Bringhurst, Cardon and Gage, 2001). An *S. meliloti* biosensor strain, which constitutively expressed DsRed, was able to utilize α-galactosides (*eg*. melibiose and stachyose) and increase its growth in alfalfa, clover and barrel medic seed wash (Bringhurst, Cardon and Gage, 2001). The presence of these galactosides is expected to increase the pool of monomeric D-galactose in the root mucilage. Nevertheless, root exudates from a variety of plants are known to contain D-galactose (Nguimbou *et al*., 2012; Koroney *et al*., 2016; Aulakh *et al*., 2001). It is clear from the DSF analysis that D-galactose binding is maintained at high pH; such result underlines the importance of D-galactose import in alkaline conditions.

A

Buffer	pH	Tm	S. D.
PBS	7.5	46.93	± 0.14
Tris	7.5	**48.9**	± 0.18
BisTrisProp	8.5	47.28	± 0.0035
Gly(NaOH)	9.5	43.314	± 0.14

B

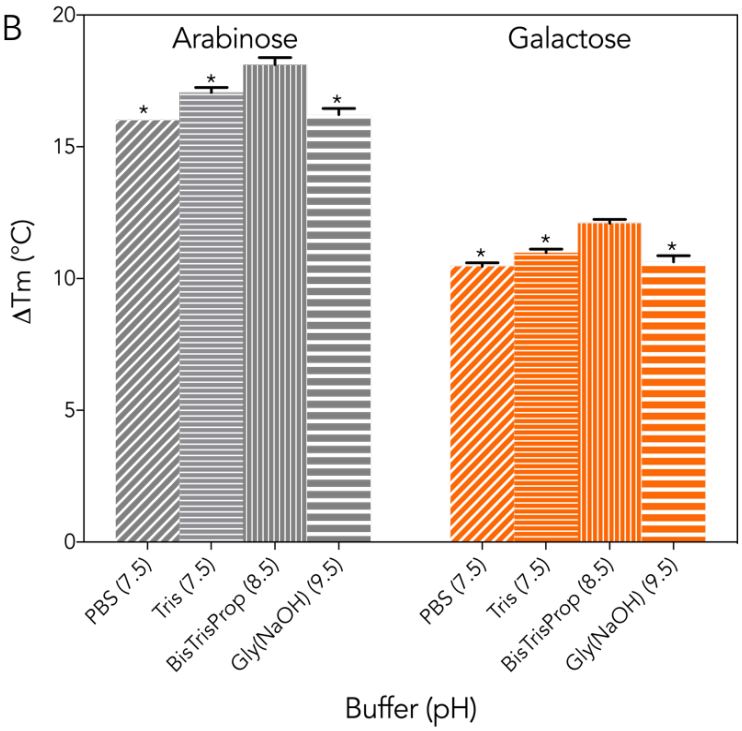


Figure 3. 10. GafASm shows partial loss of stability at highly alkaline conditions.

The GafASm was dialysed in different alkaline buffers and its stability was tested using DSF. Its ability to bind arabinose and galactose was also tested and used as an indicator of its activity in basic environment.

(A) Tabulated data from the DSF analysis of GafASm present in alkaline ranging from pH 7.5 to 9.5. The stability of GafASm in the different buffers was assessed based on its Tm. The SBP exhibits the highest Tm in Tris buffer (pH = 7.5). However, the Tm is kept at similar levels in PBS (pH = 7.5) and Bis-Tris propane (pH = 8.5).

(B) The change in Tm upon binding of arabinose and galactose assessed the activity of the protein. Even though, all of the ΔTms are significantly different to Bis-Tris propane, high ΔTms are observed for all alkaline buffers, indicating good activity across the lower alkaline pH range.

8 replicates were included for each of: protein only, protein + D-galactose, protein + L-arabinose, buffer only. The ligands were supplied at the highest previously tested concentration *ie.* 1200 µM. All buffers were supplemented with 150 mM NaCl. Asterisks represent the significance of difference of the data when compared to Bis-Tris propane buffer, *$p < 0.05$.

3. 3. 4 Intrinsic fluorescence spectroscopy of GafASm

The intrinsic fluorescence (see *Section 3. 1. 4. 3* for introductory notes) of GafASm was exploited to determine the binding affinities for the abovementioned ligands. As different excitation wavelengths affect different aromatic amino acids, the number of these in the mature form of GafASm was taken into consideration. This was calculated to be four tryptophan (Trp), five tyrosine (Tyr) and nine phenylalanine (Phe) residues. As GafAEc produced respectable amounts (4 - 6.5%) of fluorescence quench upon D-galactose binding (Horler *et al*., 2009), we then seek out to check the position of GafASm aromatic residues by mapping them on GafAEc structure. The two structures share 3 Trp, 4 Tyr and 5 Phe at conserved positions (Figure 3. 11A). Mapping the conserved residues to the structure of GafAEc showed that 2 of the Trp residues are near the binding cavity, out of which one (*ie*. W16) participates in pi-stacking interaction to stabilise D-galactofuranose binding (Figure 3. 11B). Further, two Tyr residues (*ie*. Y99, Y194) are predicted to contribute to the fluorescence emission as they are positioned in proximity to the binding site (Figure 3. 11B).

The intrinsic fluorescence emission experiments were performed as described at Materials and Methods, Section 3. 2. 2. 3. Firstly, the spectra analysis detected the excitation wavelength of the protein at which the change in fluorescence emission was monitored during the kinetic analysis.

3. 3. 4. 1 Spectra analysis for determination of the excitation wavelength

Before fluorescence spectroscopy can be used, the technique needs to be optimised for the protein under investigation. Initially, the best excitation wavelength is investigated through a series of scans which measure the emission intensity between 300 and 450 nm at two excitation wavelengths, *ie*. 280 nm and 295 nm, with the buffer only and then with 1 μM of GafAEc.

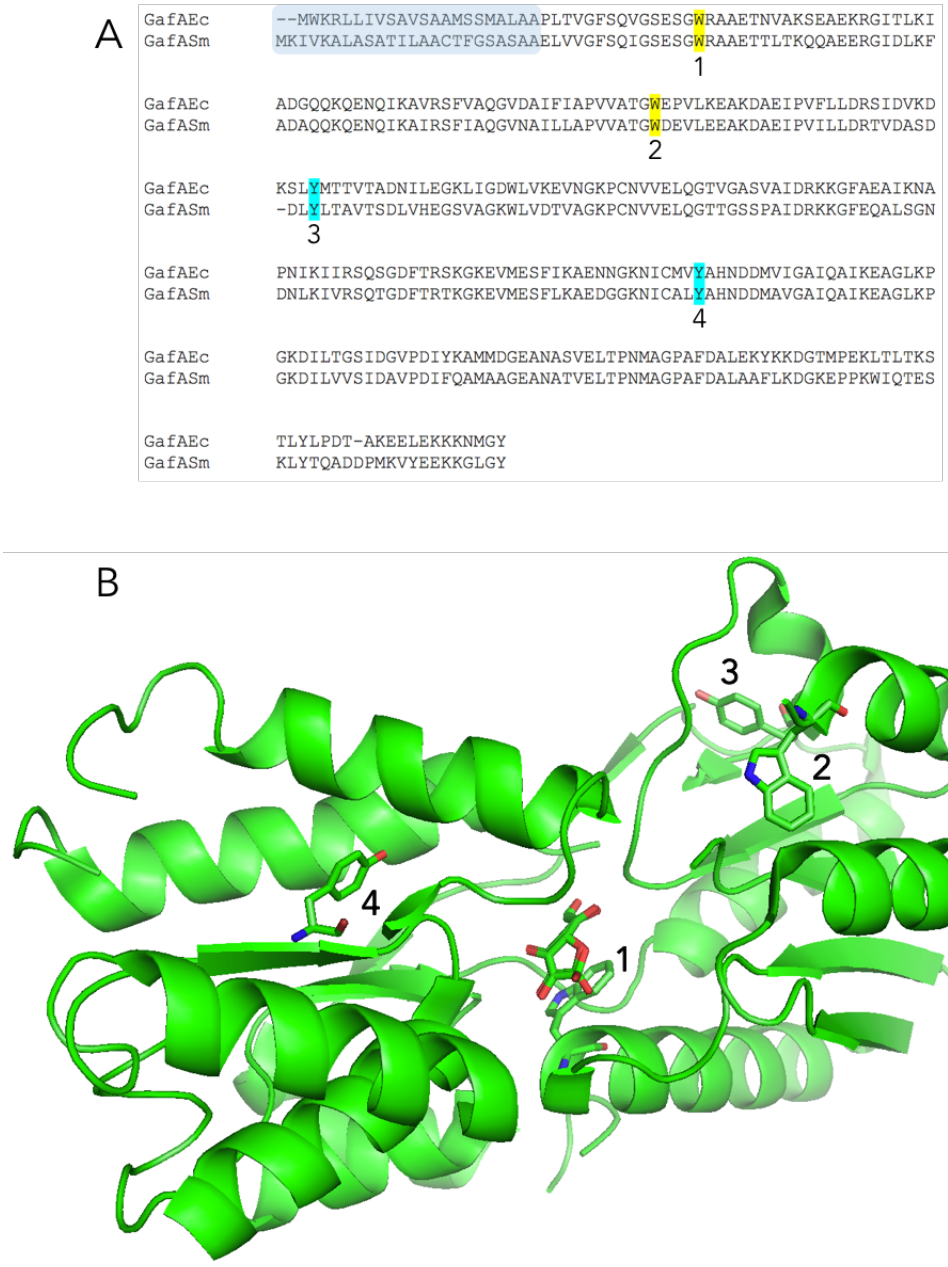


Figure 3. 11. Conserved aromatic residues of GafASm are located near the GafAEc binding cavity

A) Primary sequence alignment of the GafASm with GafAEc to detect conserved aromatic amino acids. The conserved tryptophan and tyrosine residues detected close to the binding site are highlighted with yellow and light blue, respectively.

B) Position of the conserved aromatic residues from GafASm which reside close to the binding site of GafAEc. The respective aromatics are referred to the primary sequence of (A) by numbering. The analysis is essential for the intrinsic fluorimetry assays of GafASm, which will quantify the change in binding based on the emission of the aromatics present near the binding cavity of the SBP and will be responsive to conformational changes.

The signal peptide is shaded in blue. The numbering corresponds to the position of the aromatic residues in the tertiary structure of GafAEc.

W: tryptophan, Y: tyrosine, α-D-Gal*f*: α-D-galactofuranose, 1: W16, 2: W71, 3: Y99, 4: Y194. Two β-sheets near Y194 (4) were omitted from the structure to allow for unobstructed view.

The results of the analysis are shown in Figure 3. 12. To detect any shifts in the fluorescence emission 100 μM were added in the protein mixture, as it has already been shown by DSF to bind GafASm (Section 3. 3. 1). Both excitation wavelengths produced similar fluorescence emission spectra of GafASm (Figure 3. 12A), and therefore the choice was based on the signal difference upon ligand addition. When GafASm was excited at 280 nm in the presence of 200 μM L-arabinose the emission signal showed a 3.8% decrease and a minor blue shift. Excitation at 280 nm is thought to produce emission due to the absorbance of both Trp and Tyr residues (Lakowicz, 1999); such possibility would be in agreement with the positioning of these residues near the binding site as detected in GafAEc structure (Figure 3. 12B). At wavelengths longer than 295 nm, the absorption is due primarily to tryptophan (Lakowicz, 1999). Therefore, the change in fluorescence emission at 295 nm was also assessed by addition of 200 μM L-arabinose. There was no signal change observed at 295 nm excitation, and therefore we can assume that the tyrosine residues are responsible for the majority of the signal produced at 280 nm.

The reliability of the spectra analysis at 280 nm was assured by addition of 300 μM of D-galactose, which produced a signal change compared to the protein only sample (Figure 3. 12B). The negative control used was D-glucose which doesn't bind as it failed to produce the hypochromic and hypsochromic changes in the signal seen with the two ligands above (Figure 3. 12C).

3. 3. 4. 2 Equilibrium ligand titrations for determination of the K_Ds

The measured parameters obtained from the spectra analysis, *ie*. λmax_{Em}= 341 at 280 nm excitation, were used in the titration runs which aimed to determine the K_D of L-arabinose and D-galactose. The experiment was performed as described in Materials and Methods, *Section 3. 2. 2. 3.*

The analysis produced a persistent drift in the fluorescence signal that persisted and never stabilised, following each ligand addition (data not shown). Since the K_D of the ligand binding determines at which magnitude the concentrations of the protein and

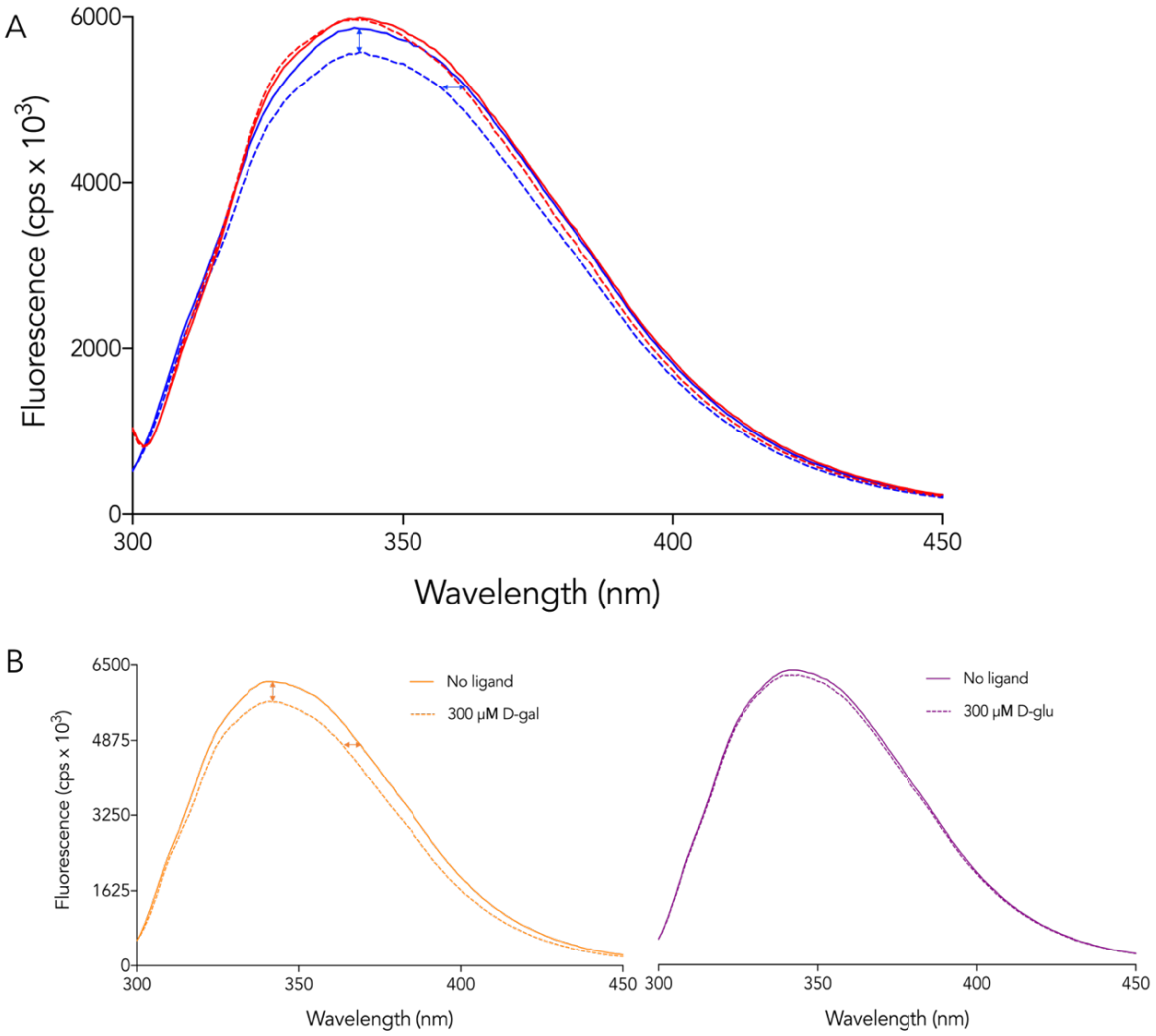


Figure 3. 12. Fluorescence spectra analysis of GafASm.

A) Emission scan between 300 - 450 nm when excited at 280 nm (blue), 295 nm (red). The solid lines represent spectra of GafASm only and dashed lines indicate spectra obtained following addition of 100 µM L-arabinose. A change in signal is only produced by excitation at 280 nm, indicating that Tyr residues might be the primary source of the side chains who's environment change upon ligand binding.

B) Emission scan between 300 - 450 nm of GafASm when excited at 280 nm in the presence of D-galactose. The solid lines represent spectra of GafASm only and dashed lines indicate spectra obtained following addition of 300 µM D-galactose. There is a hypochromic and a blue shift in signal produced, verifying reproducibility of emission obtained at 280 nm excitation in the presence of a bound ligand.

C) Same spectra analysis as in A and B but in the presence of D-glucose which was used as a negative control. The signal shows a negligible decrease which probably accounts for buffer dilution.

ligand should be, it might be that D-galactose and L-arabinose have a K_D in the nanomolar range and therefore require less amount of protein in the titration. Decreasing the concentration of the protein by half, from 1 to 0.5 μM, still produced a drift during the titration time-course with L-arabinose and it also decreased the size of the quench between the titrations (data not shown). Horler *et al*. (2009) did not come across this observation when they titrated the binding of the aforementioned ligands to GafAEc. This could be attributed to the kon rate difference between the GafAEc and GafASm. As these parameters are unknown the scenario where GafASm has a faster *k*on rate of binding than GafAEc is examined; in such case the GafASm is much faster than GafAEc in forming an association with the L-arabinose/ D-galactose leading to quicker depletion of the furanose form of the sugars. Therefore, the drift produced in the signal could be attributable to the rates of tautomerisation between the pyranose to furanose forms of the sugars, *ie*. how fast the furanose form is replenished in solution following its binding by the SBP. Anderson and Garver (1973) calculated the half-time for the approach of each sugar tautomers to its equilibrium level using gas-liquid chromatography and showed that these are: 16.2 minutes for galactofuranose, 6.4 minutes for α-arabinofuranose and 7.2 minutes for β-arabinofuranose (Anderson and Garver, 1973). As the longest titration in this experiment lasted 6 mins there wouldn't be adequate time for the furanose form of L-arabinose to equilibrate back to its starting concentration; however, some amount would still slowly accumulate during the time-course, possibly accounting for the progressive decrease in the fluorescence signal. To further investigate whether this observation is attributable to the slow tautomerisation of the furanose forms, the interconversion rate was increased by performing the experiment at higher temperature, *ie*. 35 °C instead of 25 °C (Isbell and Pigman, 1938). Addition of dimethyl sulfoxide (DMSO) to increase the furanose concentration (Franks *et al*., 1989) was not an option as this reagent is a potent protein denaturant (Arakawa, Kita and Timasheff, 2007). The drifting pattern in the quench of the signal was still observable at the higher temperature, however the protein reached saturation (*ie*. stabilisation of the fluorescence signal) approximately 10 seconds faster than it did at 25 °C (Figure A3. 4).

Due to the failure of the experiment to produce reliable quench in the signal, ITC was employed instead to calculate the K_D of the sugars.

3. 3. 5 ITC analysis of L-arabinose binding to GafASm

Due to the failure of intrinsic fluorescence of GafASm to produce a reliable quench the measurement of the K_D for L-arabinose binding was attempted using ITC (see *Section 3. 1. 4. 4* for introductory notes). The experiment was performed as explained in *Section 3. 2. 2. 4* using 70 μM of protein titrated with 27 consecutive injections of L-arabinose at 10 times higher concentration.

The analysis showed that the GafASm binds arabinose in the high nanomolar range (*ie*. K_D = 438 nM) which corresponds to 3 times tighter binding than GafAEc (Figure 3. 13). The number (*n*) of binding sites was more than a unity (1.32 per mol of sites) which may be due to underestimation of the protein concentration. The result confirms the necessity for higher scavenging capacity of GafASm to achieve faster accumulation of the L-arabinofuranose present in the root exudates. Furthermore, a high affinity system will be suited in the case that the bacterium is utilising the carbon source as a chemoattractant. Chemotaxis is the response of motile bacteria to environmental chemical gradients by moving towards chemoattractants or away from repellents (Cui and Davidson, 2011). Arabinose, as already mentioned, makes up a large proportion of the sugars in the pea plant mucilage and is found at physiologically relevant concentrations which were enough to allow growth of *Rhizobium leguminosarum* (Knee *et al*. 2001). Both *S. meliloti* and *Rhizobium leguminosarum* presented low performance in forming symbiosis and nodules when they lost the capacity to utilize certain carbon sources (Poysti *et al*., 2007). The correlation between carbon assimilation and rhizosphere occupancy underlines how important it is for the bacterium to be able to detect the L-arabinofuranose at a distance from the roots and move towards them. This is demonstrated by the high affinity verified here which allows for scavenging capacities.

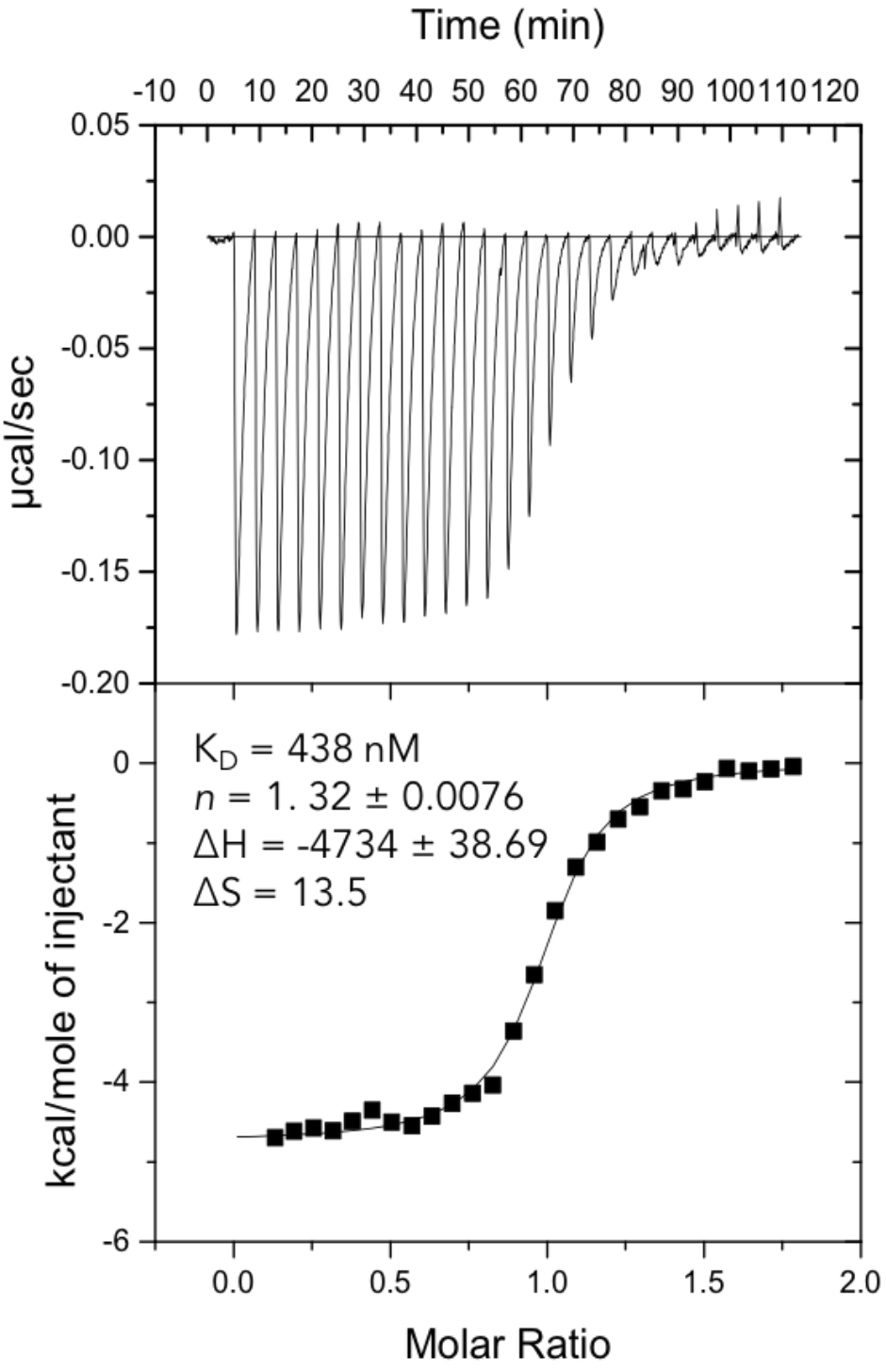


Figure 3. 13. Representative binding isotherm for the interaction of GafASm with L-arabinose.

The top panel (thermogram) represent heat differences upon each injection of ligand and the lower panels (isotherm) show integrated heats of injection (■) and the best fit (solid line) to a one-site binding model using Microcal Origin software.

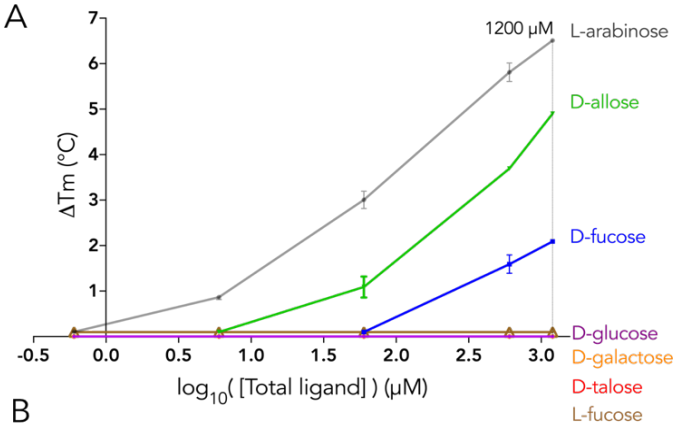
3. 3. 6 The ligand specificity of GafASw differs from GafAEc and GafASm

As the ligand specificity of GafASm appears to be similar to GafAEc, the attention was refocused to characterisation of GafASw. The initial steps in its characterisation were concerned with the biochemical characterisation of GafASw. The first step in the biochemical analysis of the GafASw was to determine its ligand specificity using DSF, similarly to the approach followed with GafASm (*Section 3. 3. 1*). The experiment was performed as described in the Methods and Materials *Section 3. 2. 2. 1.* The analysis was done in a similar manner to GafASm; the buffer used was PBS and two runs were performed to cover the whole range of ligands tested.

The results of the analysis are presented in Figure 3. 14. The melting temperature of the unbound form of GafASw was 47 °C at the first run and 43 °C at the second run (Figure A3. 5). Similarly, to GafASm (*Section 3. 3. 1*) the L-arabinose causes the highest Tm shift and stabilises the protein by 14.4 %, indicating that it could bind with the highest affinity (Figure 3. 14, S3. 5). The ligand which produced the second highest shift in the analysis was D-allose, and not D-galactose as observed with GafASm. This ligand caused an increase in stability of around 11 % which amounts to a 4.9 °C increase in the Tm of the GafASw. A shift in the Tm was also caused by the presence of D-fucose; the presence of this ligand induced an approximate 4.6 % increase in stability of the GafASw and therefore it's predicted to bind with the third highest affinity. A direct comparison of the binding affinities with GafAEc or GafASm cannot be drawn at this stage as the order of magnitude at which the binding occurs could be different for each protein.

Upon further examination of the results and when all of the assessed ligands are taken into consideration, notable discrepancies exist between the apparent ligand specificity of GafASw and of GafASm and GafAEc. GafASw is unable to bind D-galactose and D-talose (Figure 3. 14) as they both fail to produce a change in Tm (Figure A3. 5) and therefore, the protein appears to exhibit a narrower ligand

A



B

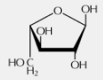
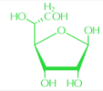
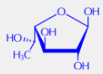
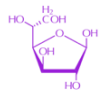
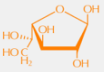
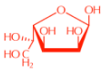
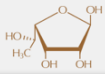
Monosaccharide	Structure	ΔTm (◦C)*	[Furanose] (µM) present at 1.2 mM total sugar	K_D (µM) for GafAEc
β-L-arabinofuranose		6.51	150	1.3 ± 0.24
β-D-allofuranose		4.9	102	30.3 ± 4.4
β-D-fucofuranose		2.09	60	6.2 ± 0.54
β-D-glucofuranose		NC	1.48	NB
β-D-galactofuranose		NC	84	1.7 ± 0.3
β-D-talofuranose		NC	348	25.7 ± 2.5
β-L-fucofuranose		NC	-	NT

Figure 3. 14. GafASw shows different ligand specificity to GafAEc, as it doesn't bind D-galactose or D-talose.

(A) The DSF analysis of GafASw was perfomed in a 96-well plate format using a real-time PCR system. The binding of the ligands was tested in PBS buffer at tenfold range of dilutions, apart from the highest concentration (*ie*. 1200 µM). Each concentration included 8 replicates. The data were plotted as the change in the Tm of the unbound form of GafASw compared to the Tm of the bound forms (or the Tm of the ligand -protein mixtures). The **$\log_{10}$** value of the concentrations tested is plotted on the x axis against the ΔTm.

(B) The ΔTms were tabulated next to the respective structure of the sugars tested. The order of ligands presented is as determined by the ΔTm at 1200 µM of total sugar concentration. The respective furanose concentrations at 1200 µM are also shown; the calculated values are based on their relative percentages in aqueous solution (from 25 to 31 ◦C). The K_D of GafAEc for the ligands, is also provided for comparison (Horler *et al*., 2009).

*The ΔTm values indicate the Tm difference between of the samples of protein with ligand (at1200µM) and protein only/ NC = There was no change observed in the Tm of the protein supplied with ligand/ NB = No binding detected due to no apparent change in fluorescence upon addition of ligand/ NT = Not tested in the intrinsic fluorescence assays performed by Horler *et al*. (2009)/ All structures were produced in ChemDoodle.

specificity compared to GafAEc and GafASm. Assuming that *Shewanella* sp. ANA-3 encounters D-galactose in its natural environment and that the *gafABCD*Sw is expressed during such encounter, the inability of GafASw to bind this sugar is the first indication to confirm the hypothesis that failure of *Shewanella* sp. ANA-3 to grow on D-galactose (Rodionov *et al.*, 2010) could be attributed to the lack of dedicated transport systems (see *Section 2. 3. 3. 3*). The structural similarities between D-galactose and D-talose could provide the first indications about the evident discrepancy in binding. Both D-talose and D-galactose are hexoses and therefore their furanose forms exhibit a longer hemiacetal linkage with two carbon atoms facing away from the sugar ring (Figure 3. 14). Unlike hexoses, the furanose form of pentose sugars, *ie*. L-arabinofuranose, have a shorter hemiacetal linkage by one carbon and therefore could be accommodated to a potentially smaller binding cavity, leading to exclusion of the furanose forms of hexoses. However, this is unlikely as D-fucofuranose, which has a similar hemiacetal linkage to D-galactofuranose, is still able to bind (Figure 3. 13). Examination of D-fucofuranose structure shows that the C_6 is a methyl and not an alcohol group (*ie*. $-CH_2OH$), so the extra hydroxyl group in D-galactofuranose and D-talofuranose could be the determining factor in the abolishment of their binding.

An additional ligand included in the analysis was the L-fucofuranose which failed to cause a shift. We anticipate that this is attributed to its low concentration in solution similarly to D-glucofuranose. Due to the C_6 hemiacetal linkage and the hydroxyl constituent in C_4 facing the same plane of the sugar ring (Figure 3. 13), the furanose form of L-fucose is expected to less favoured in solution compared to its D-anomer.

3. 3. 7 Screening of the activity and stability of GafASw in acidic and alkaline conditions

The activity and stability of the GafASw in both acidic and basic buffers was assessed by DSF in an identical manner to GafASm in terms of the methodology applied (see *Section 3. 3. 3*).

3. 3. 7. 1 The activity and stability of GafASw is severely decreased at high acidity

The results of the screening in acidic conditions are shown in Figure 3. 14. The GafASw was the most stable at HEPES (pH= 7.2) and it also showed maximal activity at this condition (Figure 3. 15A, B). In highly acidic conditions (pH= 4.5) the SBP presented a more than 40% decrease in stability (Figure 3. 15A, S3. 6B) and complete loss of activity when compared to HEPES, demonstrating how the binding is pH-dependent (Figure 3. 14B). Despite the relatively high stability of the GafASw in less acidic conditions (*ie*. pH= 5.5), GafASw retained half its maximal activity at such conditions.

This is in discordance to the acid-tolerant nature of the L-arabinose isomerase from *Shewanella* sp. ANA-3 (Rhimi *et al*., 2011). The enzyme in question was highly active at pH = 5.5 to 6.5 and retained more than 80% of its activity from pH= 4.0 to 8.5. Remarkably, this protein preserved more than 60% of its relative activity at pH 3.0 and 3.5 (Rhimi *et al*., 2011), which are conditions that haven't been tested at the present study. This raises a 'pH optima anomaly', as the SBPs are more likely to be affected by the changes in the extracellular pH as compared to intracellular enzymes, cause of the lack of the extra barrier the inner membrane offers. In fact, the cytoplasm of *E. coli* is capable of recovering its physiological pH after a drop in the external pH; unlike the periplasmic space which instead adopted the acidity (Wilks and Slonczewski, 2007). The oxymoron observation here could be attributed to the presence of acidic patches present in the intracellular space of the bacteria (Morimoto *et al*., 2016, 2017). In such conclusion came a group which studied a selection of enzymes from the acidophile *Ferroplasma acidiphilum* (Golyshina *et al*., 2006). These enzymes presented optimal activities and stabilities at much lower pH than the mean cytoplasmic pH (*ie*. 5.6) of this bacterium (Macalady *et al*., 2004). Therefore, the L-arabinose isomerase robustness in acidic environments could be due to the variable pH occurring in the cytoplasm and not necessarily because of a potential acidic habitat, which would be unsupported by the complete loss of activity of GafASw seen here. Nonetheless, *Shewanella* sp. ANA-3 thrives in estuarine and aquatic environments which are primarily alkaline (Saltikov *et al*., 2003).

A

Buffer	pH	Tm	S. D.
TrisNaCit	4.5	33.91	± 0.15
TrisNaCit	5.5	46.44	± 0.14
BisTrisMeth	6.6	44.83	± 0.003
HEPES	7.2	48.40	± 0.003

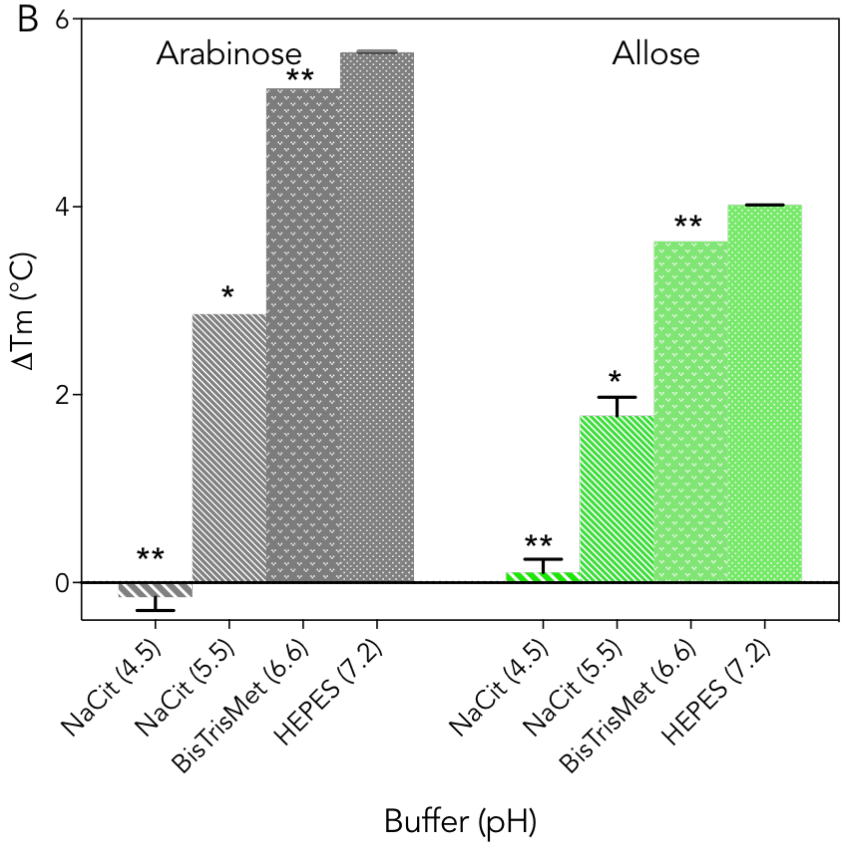


Figure 3. 15. GafASw shows complete loss of activity and stability at very low pH.

The GafASw was dialysed in different acidic buffers and its stability was tested using DSF. Its ability to bind L-arabinose and D-allose was also tested as a measure of its activity in acidic environment.

(A) Tabulated data from the DSF analysis of GafASm present in acidic and neutral buffers ranging from pH 4.5 to 7.2. The stability of GafASm in the different buffers was assessed based on its Tm. Notably, the stability at the lowest pH tested (*ie*. Tris sodium citrate, pH = 4.5) appears severely impacted compared to less acidic buffers.

(B) The change in the Tm upon binding of L-arabinose and D-allose was used as a measure of the protein activity. In very acidic pH (*ie*. Tris sodium citrate, pH = 4.5) the activity is completely lost as there is no ΔTm.

Eight replicates were included for each of: protein only, protein + D-allose , protein + L-arabinose, buffer only. The ligands were supplied at the highest previously tested concentration *ie*. 1200 µM. All buffers were supplemented with 150 mM NaCl. Asterisks represent the significance of difference of the data when compared to HEPES buffer, *$p < 1e^{-5}$, **$p < 1e^{-10}$.

3. 3. 7. 2 GafASw retains full activity and stability at alkaline buffers

Unlike in acidic conditions, the GafASw preserved full activity across the alkaline conditions tested (Figure 3. 15B, S3. 4). The differences between the conditions were found to be insignificant ($=0.018 < p < 0.031$) showing that the SBP retains maximal capacity for binding of L-arabinose and D-galactose. Also, the SBP was unaffected at high pH, as its stability only showed a mere 2.45 % drop (Figure 3. 15A).

The robustness of this SBP across the basic spectrum seen here, could be an evolutionary result of *Shewanella* sp. ANA-3 thriving in environments which are primarily alkaline. The strain sp. ANA-3 was originally isolated from a brackish estuarine environment and is known to thrive in aquatic or marine environments (Saltikov *et al*., 2003). In fact, most of the *Shewanella* spp. have been isolated from such environments including Huon River estuary in Australia (*ie*. *Shewanella olleyana*; Skerrat, 2002), Antartic coastal marine (*ie*. *Shewanella frigidimarina* and *Shewanella livingstonensis*; Bozal *et al*., 2002), and the Baltic sea (*ie*. *Shewanella baltica, Shewanella hafniensis and Shewanella morhuae;* Satomi *et al.,* 2006*;* Hambright *et al.,* 2016). The pH of open-ocean seawater typically ranges between 7.8 and 8.4 when measured *in situ*, with estuarine environments having slightly lower pH, fluctuating from neutral to 7.5 (Ringwood and Keppler, 2002). Therefore, the periplasmic space of this bacterium is expected to be basic which would justify the preservation of activity and stability of GafASw at such conditions.

3. 3. 8 Thermodynamic analysis of L-arabinose, D-allose and D-galactose binding to GafASw using intrinsic fluorimetry and ITC.

The biochemical analysis for GafASw proceeded in a similar fashion to GafASm (*Sections 3. 3. 4, 3. 3. 5*) by using intrinsic fluorimetry and ITC to confirm the qualitative analysis done by DSF and calculate the binding affinities of the ligands. The two ligands, *ie*. L-arabinose and D-allose, which are predicted by DSF to show the highest affinity for GafASw were included in the analysis. Also, as the main difference

A

Buffer	pH	Tm	S. D.
PBS	7.5	44.78	± 0.012
Tris	7.5	**44.86**	± 0.004
BisTrisProp	8.5	44.60	± 0.083
Gly(NaOH)	9.5	43.76	± 0.003

B

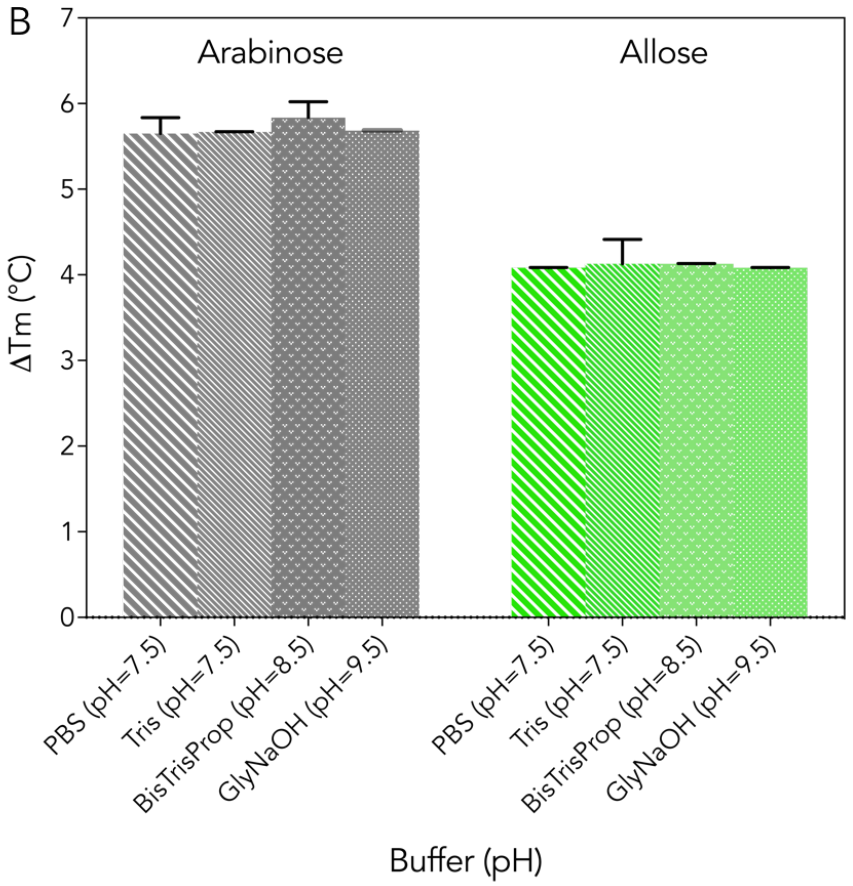


Figure 3. 16. GafASw retains almost full stability and activity across alkaline conditions.

The GafASw was dialysed in different alkaline buffers and its stability was tested using DSF. Its ability to bind L-arabinose and D-allose was also tested and used as an indicator of its activity in moderately basic buffers.

(A) Tabulated data from the DSF analysis of GafASw present in alkaline buffers ranging from pH 7.5 to 9.5. The stability of GafASm in the different buffers was assessed based on its Tm.

(B) The change in Tm upon binding of arabinose and galactose assessed the activity of the protein. The activity is fully retained across the lower alkaline pH spectrum as the ΔTm are insignificantly different.

8 replicates were included for each of: protein only, protein + D-galactose, protein + L-arabinose, buffer only. The ligands were supplied at the highest previously tested concentration ie. 1200 µM. All buffers were supplemented with 150 mM NaCl.

between GafASw and GafAEc is the inability of the first to bind D-galactose, this ligand was also included in the assays to verify it does not bind.

3. 3. 8. 1 Spectra analysis of GafASw

Initially, the SBP was analysed for changes in spectra upon L-arabinose addition to detect the most useful wavelength for emission intensity, from 300 to 450 nm in a similar manner to GafASm (see *Section 3. 3. 4. 1)*. When excited at 280 nm, the protein showed a somewhat hyperchromic shift, at around of 8.6 %, and a minor red shift of around 3 nm (Figure 3. 17A). This indicated that the Tyr residues of the protein are more than likely to be contributing to the fluorescence emission change upon binding.

Further, we attempted to check whether there is a collective contribution by both Tyr and Trp residues by exciting the GafASw at 295 nm. The results showed that such wavelength caused a larger hyperchromic change (*ie*. 12 %) and a small blue shift of around 5 nm (Figure 3. 17B). The hyperchromic change decreased to 9.5% when the protein was excited at 297 nm, with no shift observed. Since the fluorescence emission upon excitation at 297 nm is attributed to Trp residues only, we can assume that the effect seen at the 295 nm is caused by the conformational change of both Trp and Tyr residues. However, mapping the conserved Trp and Tyr residues of GafASw (Figure 3. 18A) to the GafAEc binding cavity reveals that the Tyr194 is replaced by a Trp residue in GafASw. Therefore, if this residue was to be accounted for the fluorescence change seen at 295 nm, its replacement by a Trp suggests otherwise.

Nonetheless, GafASw might contain Tyr residues close to its binding cavity which are not taken into consideration in the current assessment because they are not conserved in the GafAEc sequence.

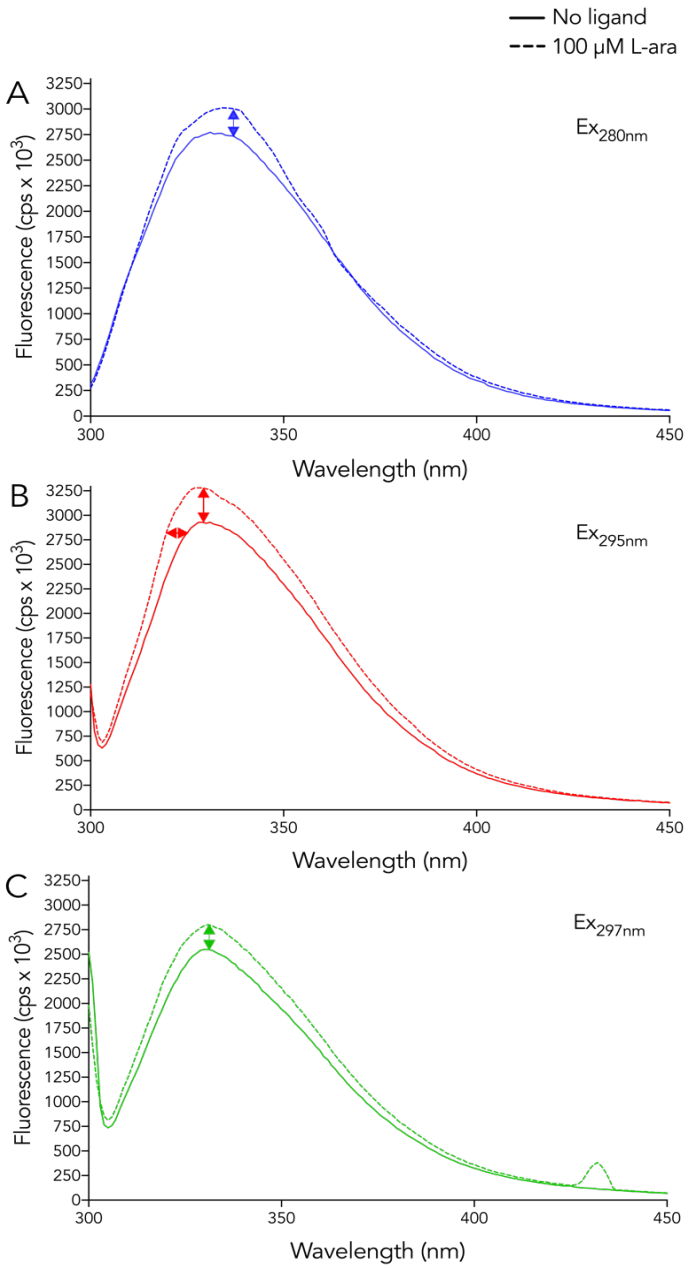


Figure 3. 17. Fluorescence spectra analysis of GafASw

A) Emission scan between 300 - 450 nm when excited at 280 nm. The solid lines represent spectra of GafASm only and dashed lines indicate spectra obtained following addition of 100 µM L-arabinose.

B) Emission scan between 300 - 450 nm of GafASw when excited at 295 nm in the presence of L-arabinose. The solid lines represent spectra of GafASw only and dashed lines indicate spectra obtained following addition of 100 µM L-arabinose. There is a hyperchromic and a minor blue shift in signal produced.

C) Emission scan between 300 - 450 nm of GafASw when excited at 297 nm in the presence of 100 µM L-arabinose. Only a minor increase in signal is observed upon addition of ligand.

The solid lines represent spectra of GafASw only and dashed lines indicate spectra obtained following addition of 100 µM L-arabinose.

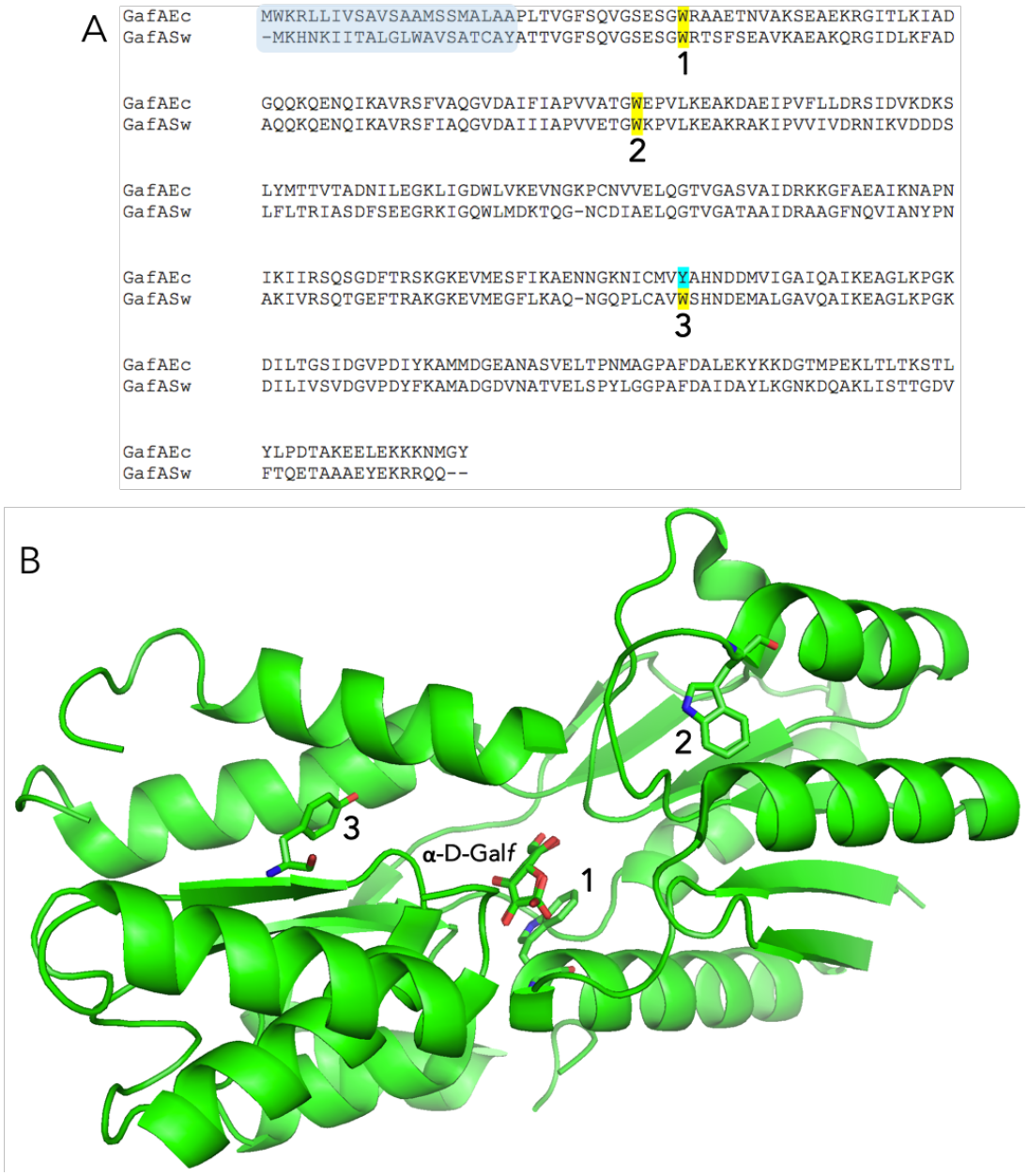


Figure 3. 18. Mapping of the conserved aromatic residues from GafASw on GafAEc structure.

A) Primary sequence alignment of the GafASw with GafAEc to detect conserved aromatic amino acids. The conserved tryptophan and tyrosine residues detected close to the binding site are highlighted with yellow and light blue, respectively. The tyrosine at position 194 of GafAEc is replaced by a tryptophan in GafASw, which could potentially contribute to the intrinsic fluorescence emitted upon ligand binding.

B) The position of the conserved aromatic residues designated at (A).

The signal peptide is indicated in blue. The numbering corresponds to the position of the aromatic residues in the tertiary structure of GafAEc.

W: tryptophan, Y: tyrosine, α-D-Gal*f*: α-D-galactofuranose, 1: W16, 2: W71, 4: Y194. Two β-sheets near Y194 (3) were omitted from the structure to allow for unobstructed view.

3. 3. 8. 2 Intrinsic fluorescence analysis and ITC confirm GafASw inability to bind D-galactose

The binding of D-galactose was assessed by excitation of GafASw at 297 nm as determined in the spectra analysis (*Session 3. 3. 9. 2*). Addition of 300 µM of D-galactose maintains the same fluorescence intensity and fails to produce a hyperchromic change or blue shift (Figure 3. 19A) as observed with L-arabinose (Figure 3. 16B). This was further corroborated by a time-based assay where the protein was excited at the aforesaid wavelength and titrated with D-galactose. Monitoring the fluorescence emission at 327 nm (λ_{max}) showed that the addition of D-galactose at 150 µM final cumulative concentration caused a decrease in the signal and not an enhancement, as it would be expected in hyperchromic shifts. The quenching in signal is not representative of true binding and is most likely to be cause by photobleaching of the protein's autofluorescence (Amar and Chung, 2014). This was exemplified by the striking similarity between the fluorescence profiles of adding only buffer in the protein and D-galactose addition (Figure 3. 19B).

As GafASw presents an inability to bind D-galactose in comparison to high affinity exhibited by its close orthologue GafAEc, this result required further confirmation. ITC analysis was employed to further support this considering its importance in the current study. The experiment was performed as described in *Section 3. 2. 2. 4* of the Materials and Methods, using 100 µM of protein titrated with 27 consecutive injections of 1 mM D-galactose. The addition of D-galactose failed to produce a substantial change in power as presented in heat changes of the isotherm in Figure 3. 20, thus assuring that binding does not occur. The minor fluctuations represented by the insignificant peaks are background noise most likely attributed to buffer additions.

3. 3. 8. 3 D-allose binds to GafASw with micromolar affinity

Similarly to D-galactose, the binding of D-allose to GafASw was assessed by intrinsic protein fluorescence and ITC. Measurements were performed either in PBS (*ie*. 10 mM

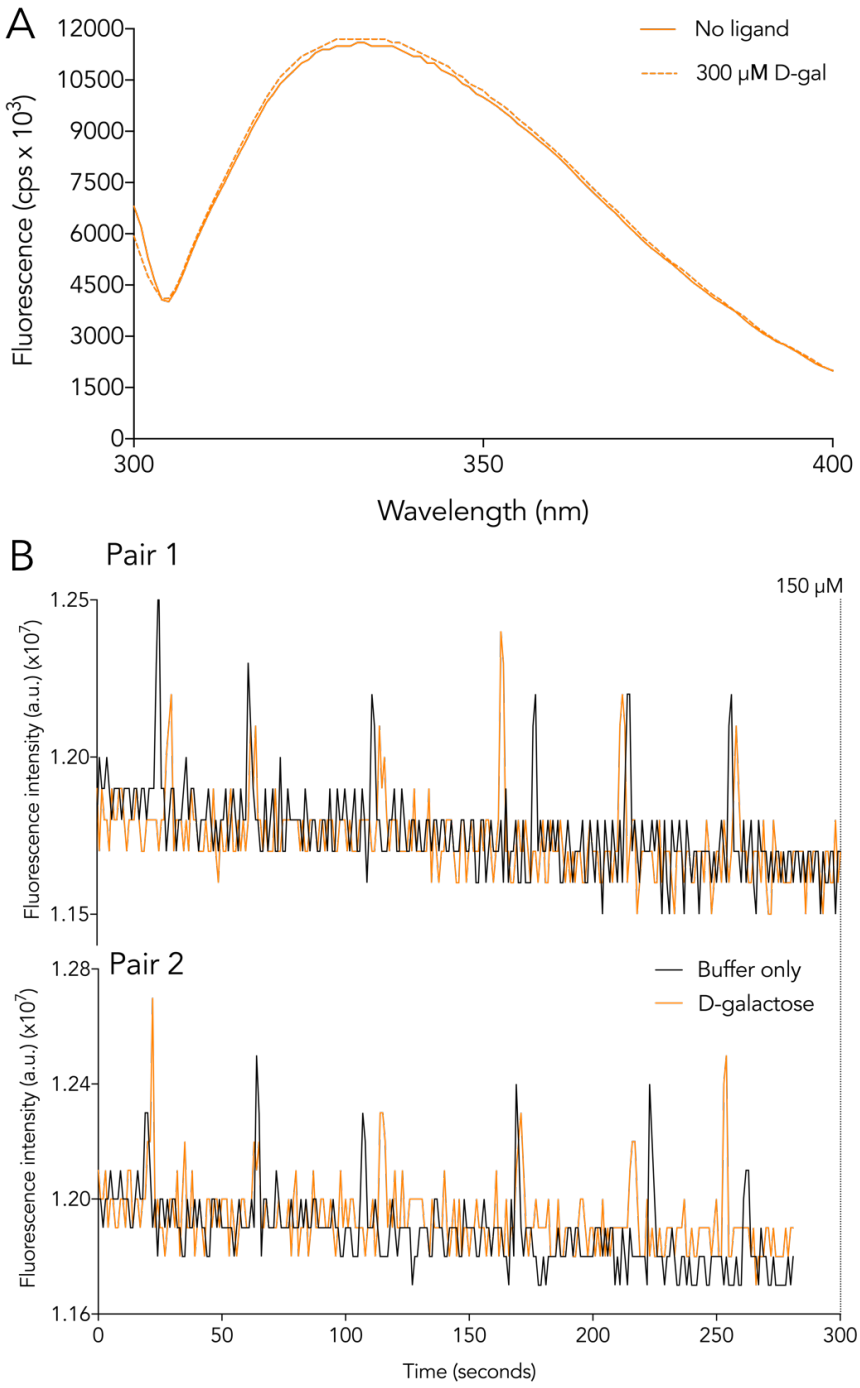


Figure 3. 19. Intrinsic fluorescence analysis of GafASw with D-galactose indicates no binding

A) Emission scan between 300-400 nm of 0.5 µM GafASw when excited at 295 nm. The solid lines represent spectra of GafASw only and dashed lines indicate spectra obtained following addition of 300 µM cumulative concentration of D-galactose.

B) Fluorescence change upon additions of D-galactose (orange) in the protein mixture resembles the titration during which only buffer was added. Each addition occurred at the timepoints corresponding to the peaks and added 25 µM of sugar. The cumulative concentration of D-galactose by the end of the run was 150 µM.

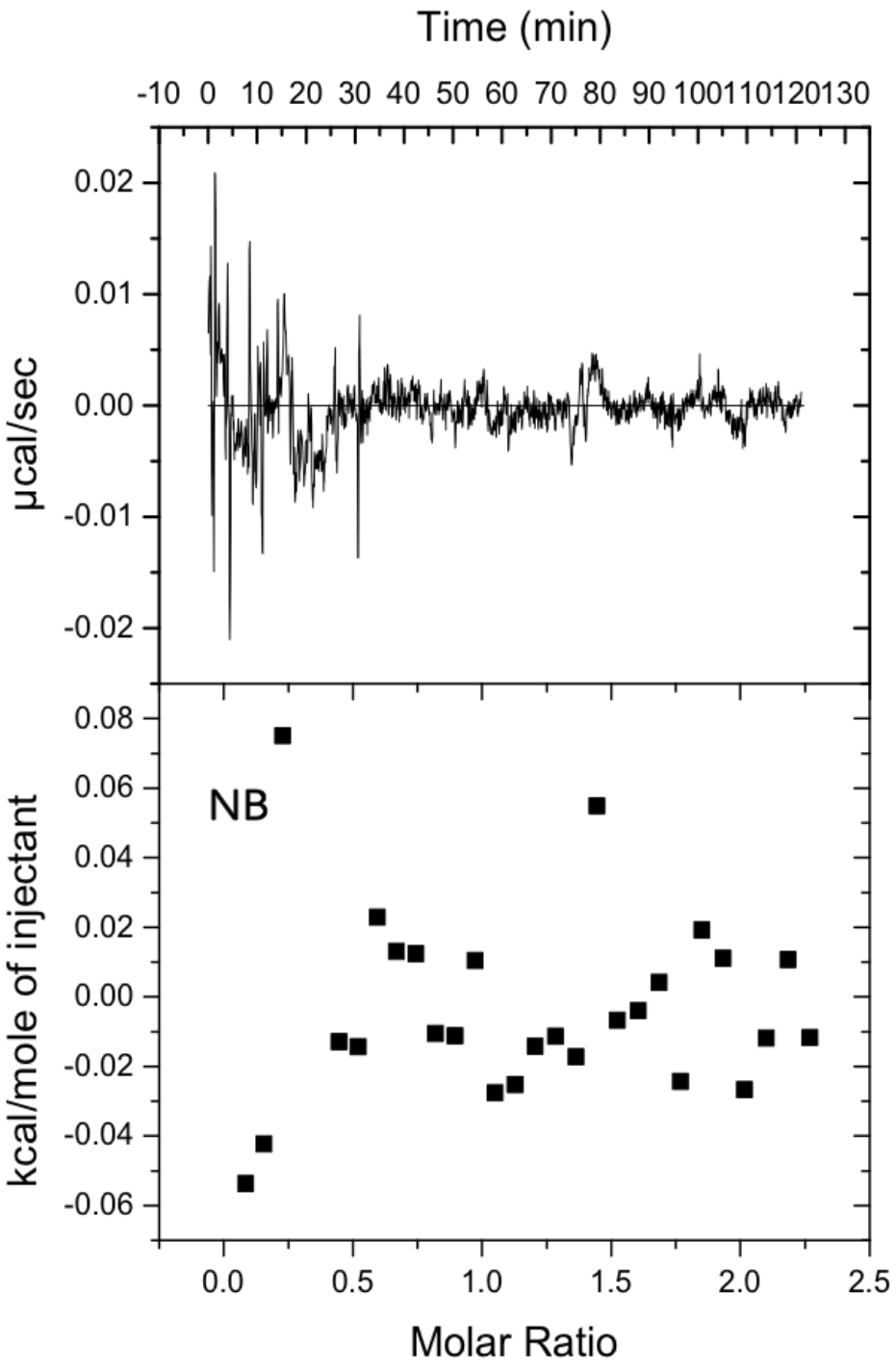


Figure 3. 20. ITC analysis to confirm that GafASw doesn't bind D-galactose

The top panel (thermogram) represent heat differences upon each injection of ligand and the lower panels (isotherm) show integrated heats of injection (■).

final [PO_4^{3-}]) or in 50 mM Tris (with 150 mM NaCl), as the protein was shown to retain its full activity at these conditions (see *Section 3. 3. 7. 2*). The ionic strength of the buffers was calculated using the Biomol buffer calculator (Beynon, 1988) and found to be very similar (*ie*. 0.025 M for PBS and 0.04 M for Tris), therefore this variable is not expected to hinder the binding of D-allose.

The intrinsic fluorescence of GafASw at 327 (*ie*. Emλ_{max}) was decreased by the addition of D-allose until saturation was achieved (Figure 3. 21A). The average maximal fluorescence change was about 4% (Figure 3. 21A) and the dissociation constant calculated using the one-site model for specific binding from triplicate measurements was 24.8 ± 0.22 μM (Figure 3. 21B). ITC analysis was performed using 120 μM of GafASw titrated by 27 consecutive additions of L-arabinose at 10 times higher concentration. The average measured K_D from triplicate runs was 27.4 ± 0.43 μM which is very near the value obtained from intrinsic protein fluorescence (Figure 3. 22). GafAEc exhibits a similar binding affinity for D-allose (*ie*. 30.3 ± 4.4 μM), which falls within the range of the GafASw when the large stand ard deviation is considered (Horler *et al*., 2009).

The similarity in the calculated K_Ds between the two proteins indicates that D-allose features a structure which is accommodated in the binding cavity of GafASw, therefore the same residue modifications which perturb binding of D-galactose and D-talose (see *Section* 3. 3. 6) are unlikely to effect the recognition of D-allose as the magnitude of binding is the same with GafAEc. The D-allofuranose hemiacetal linkage at C_5 substituent is pointing up in the Haworth projection of the structure (Figure 3. 14) as opposed to D-galactofuranose and D-talofuranose, where C_5 is facing the opposite direction. We predict that the residues near this side of D-allofuranose will interact with the C_5 substituent and promote ligand binding. We wouldn't be able to verify this without the crystal structure of GafASw as a similar sugar structure, *ie*. D-glucofuranose, is not favoured in solution as already discussed above.

To confirm the order of ligand preference exhibited by GafASw, an intrinsic fluorescence-based competition assay was performed. The protein was firstly

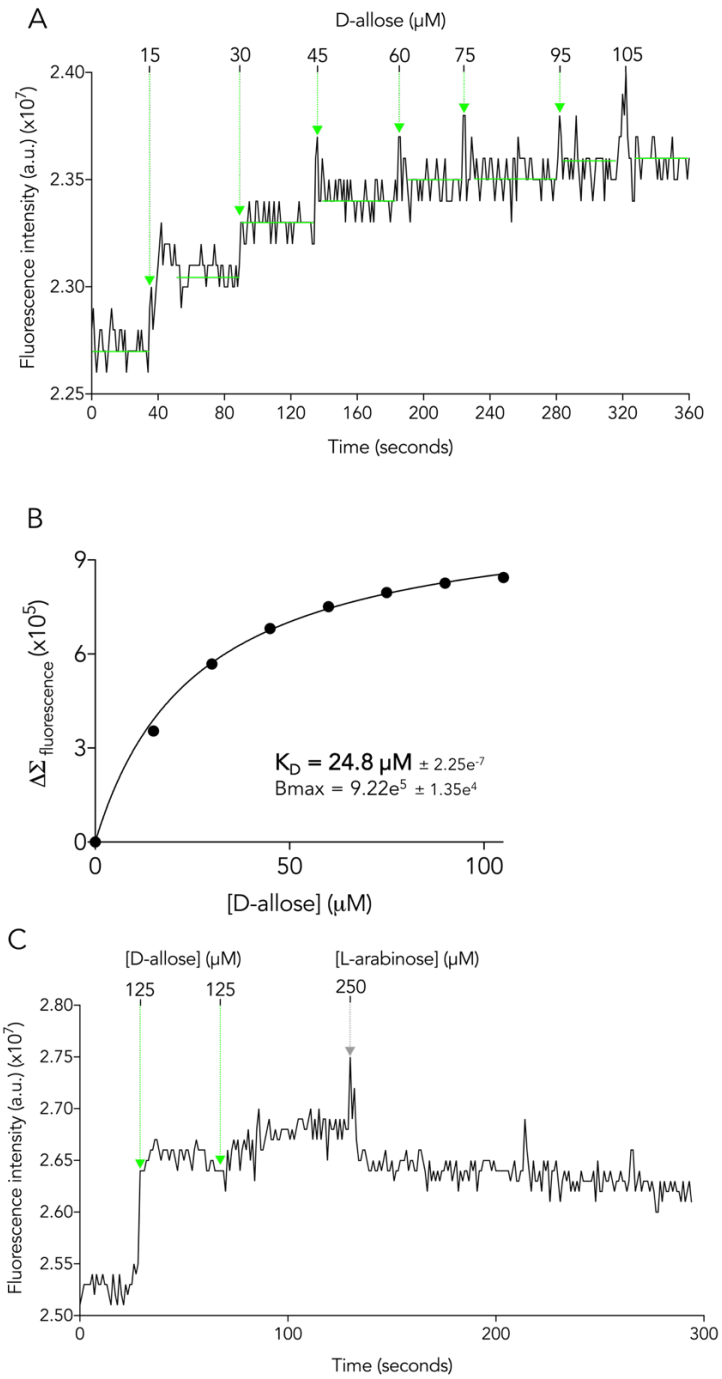


Figure 3. 21. Intrinsic fluorescence analysis of D-allose binding on GafASw.

(A) Fluorescence change in 1 μM GafASw upon multiple additions of D-allose, at 297 nm excitation wavelength and 321 nm emission. The averages values of the fluorescence signal as designated by the green lines were used to calculate the difference in fluorescence ($\Delta\Sigma_{fluorescence}$) in between the additions.

(B) Graphpad plot of the $\Delta\Sigma_{fluorescence}$ against cumulative concentration of D-allose. The model for one-site specific binding was used to calculate the binding affinity (K_D) and maximal density of the binding site (Bmax)

(C) Displacement of D-allose from 3 μM of GafASw upon addition of equimolar concentration of L-arabinose, as indicated by a quench in the fluorescence signal.

The concentrations added at each titration are indicated by the green (D-allose) and grey (L-arabinose) arrows.

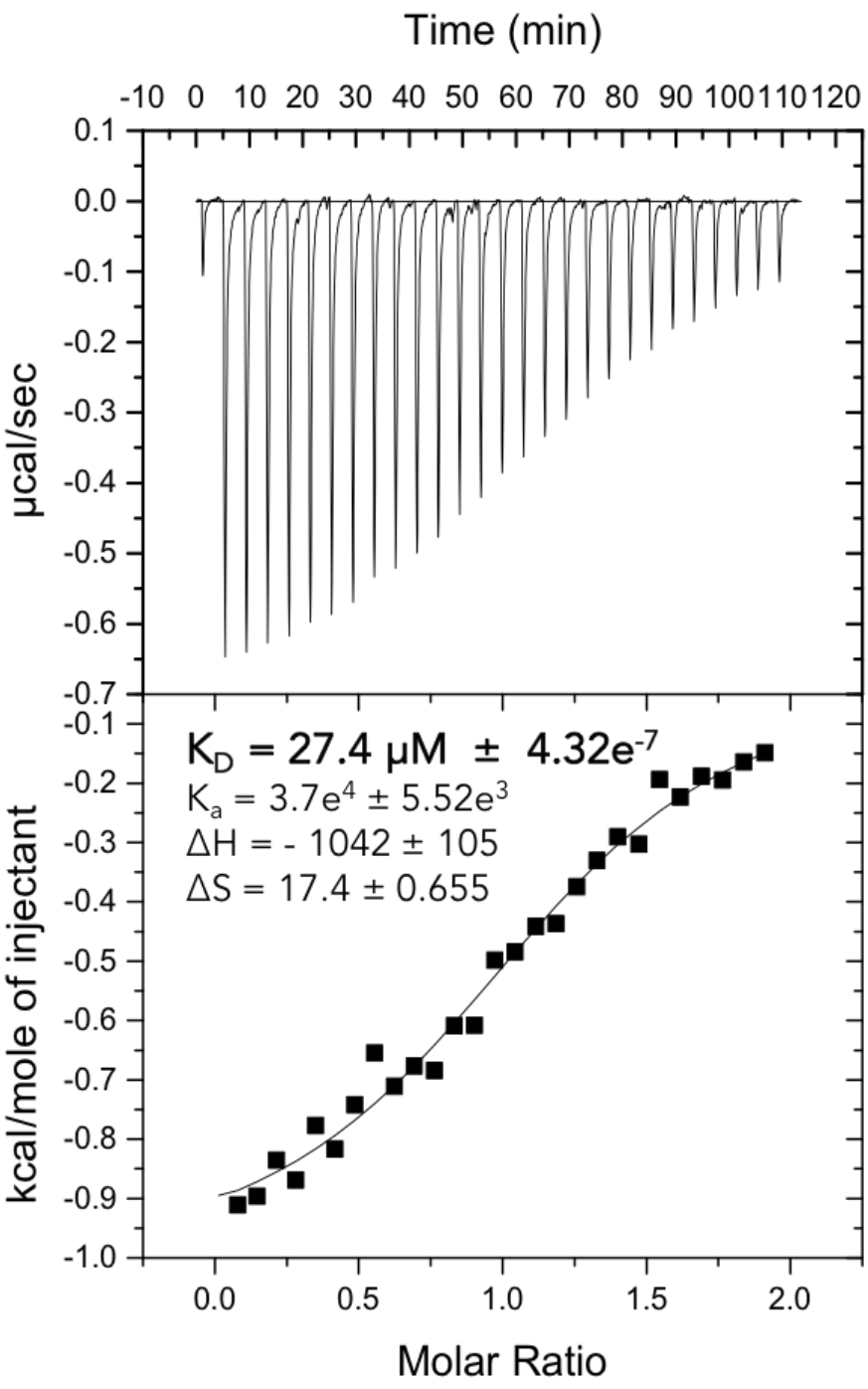


Figure 3. 22. ITC analysis shows D-allose binds to GafASw at micromolar range
The top panel (thermogram) represent heat differences upon each injection of ligand and the lower panels (isotherm) show integrated heats of injection (■). the lower panels (isotherm) show integrated heats of injection (■) and the best fit (solid line) to a one-site binding model using Microcal Origin software. The parameters designated were calculated from three replicates.

saturated by addition of 250 μM D-allose which caused a 5.5 % enhancement in the fluorescent signal. Following, an equimolar concentration of L-arabinose was added which provoked a 1.6% of quench in fluorescence, verifying that it binds stronger than D-allose (Figure 3. 21C).

3. 3. 9. 4 L-arabinose binds to GafASw weaker than GafAEc

The same strategy with D-allose was followed for the quantification of the binding affinity of L-arabinose to GafASw. To obtain a reliable saturation curve the concentration of GafASw was kept within the low micromolar range (0.5 μM), similar to the K_D of GafAEc for L-arabinose. Higher concentrations of protein (*ie*. 1.0 and 1.5 μM) produced shallow binding curves (data not shown) and therefore were avoided. Keeping the protein concentration low though led to small enhancement (~1.5%) in the fluorescent signal upon addition of L-arabinose (Figure 3. 23A). The K_D was still calculated using the one-site specific binding model and shown to be 6.64 ± 0.66 μM (Figure 3. 23B). To corroborate and ensure the result above is reliable, ITC was performed using 120 – 170 μM GafASw titrated with 27 successive injection of 1.2 – 1.7 mM L-arabinose. The binding affinity as calculated by ITC was 1.5 times lower (*ie*. 4.19 ± 0.54 μM) than the intrinsic protein fluorescence (Figure 3.23B, 3. 24). Considering the latter produced small fluorescence change, we are inclined to accept the results from ITC analysis. Therefore, GafASw was shown to have the ability to bind L-arabinose as predicted by the arabinose-controlled transcriptional regulation exerted in its operon and the heavy relevance of its genomic neighborhood to arabinose utilisation (see Chapter 2). Albeit, the binding affinity appears around 3.5 times and 10 times less tight than GafAEc (Horler *et al*., 2009) and GafASw (Section 3. 3. 5), respectively. This could reflect the abundancy of L-arabinose each of the respective bacteria encounters in its close proximity. A potentially higher pool of arabinofuranose near *Shewanella* sp. ANA-3 would imply it doesn't require the strong scavenging activity of an SBP like GafASm to accumulate equal concentrations of the sugar.

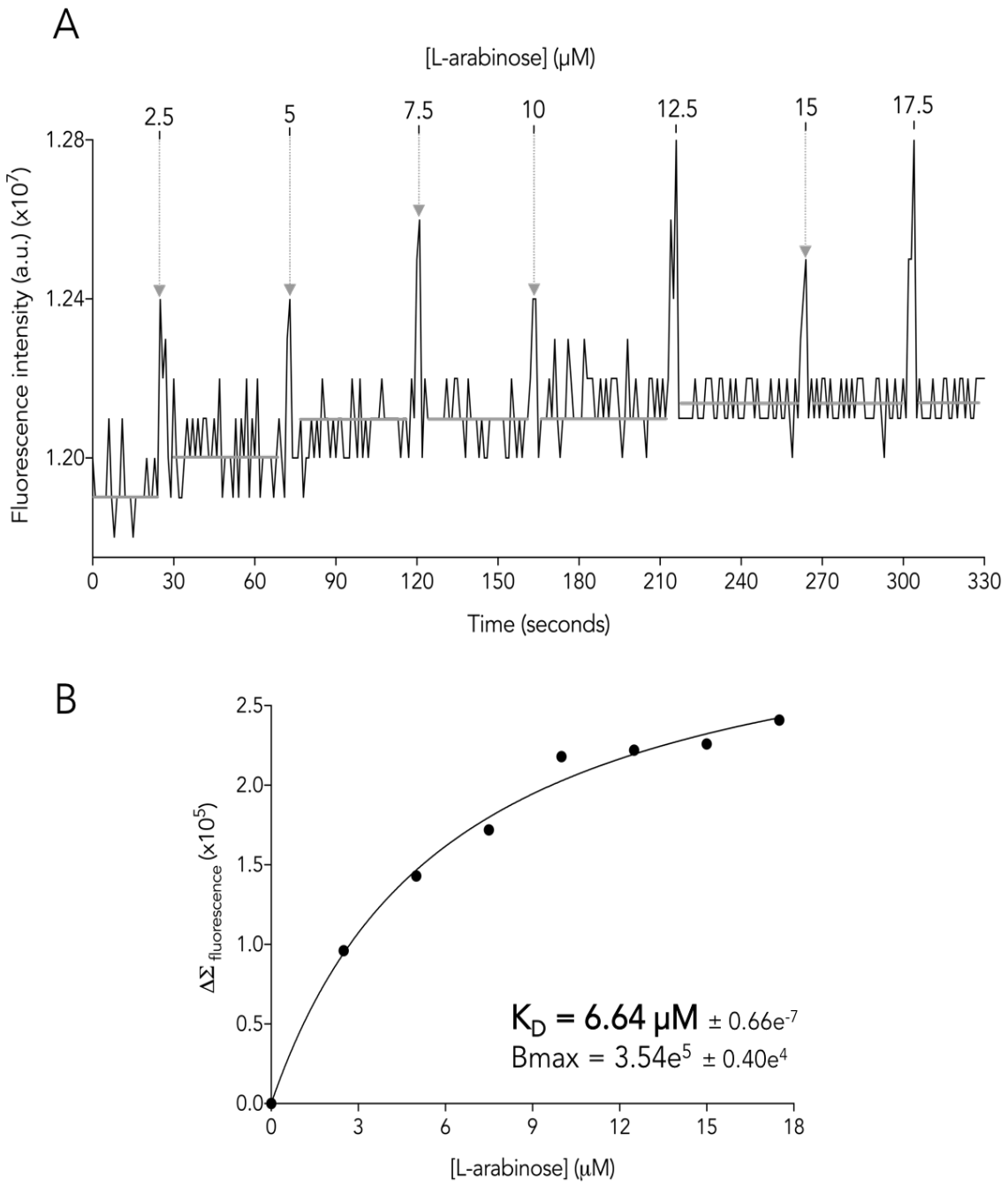


Figure 3. 23. Intrinsic fluorescence analysis of L-arabinose binding to GafASw.

(A) Fluorescence change in 1 µM GafASw upon multiple additions of L-arabinose.

(B) Graphpad plot of fluorescence change against cumulative concentration of L-arabinose. The model for one-site specific binding in GraphPad Prism 7.0 was used to calculate the binding affinity (K_D) and maximal density of the binding site (Bmax). The concentrations added at each injection are indicated on top of the grey arrows.

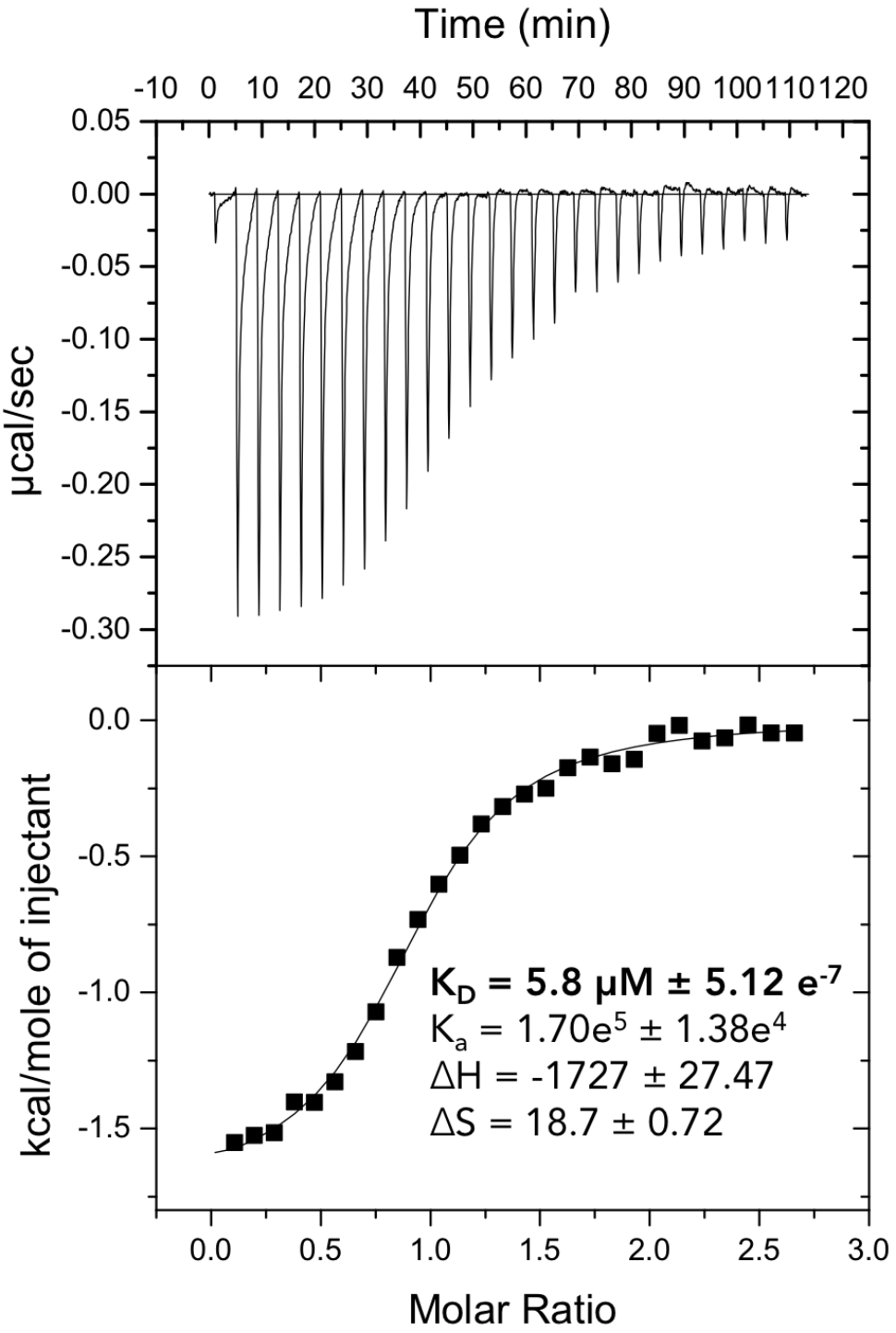


Figure 3. 24. ITC analysis shows L-arabinose binds to GafASw at low macromolar range.

The top panel (thermogram) represent heat differences upon each injection of ligand and the lower panels (isotherm) show integrated heats of injection (■). The lower panels (isotherm) show integrated heats of injection (■) and the best fit (solid line) to a one-site binding model using Microcal Origin software. The parameters designated were calculated from three replicates.

3. 3. 9 Solving the structure of GafASw

In addition to the biochemical characterisation undertaken during the project, a collaboration was set up with the York Structural Biology Laboratory (YSBL), in which they would attempt to crystallise the GafASw bound to L-arabinose, so that the molecular determinants which lead to abolishment of D-galactose binding were exemplified. Reyme Herman solved the crystal structure of GafASw in collaboration with Prof. Anthony Wilkinson of YSBL. The crystallographers were provided with 3 ml of 125 µM and 3 ml of 250 µM GafASw purified from Power Broth cultures together with L-arabinose, which was added at double the concentration of the protein. Crystal growth trials were set up with GafASw and exogenous L-arabinose as described in the Materials and Methods section 3. 2. 2. 6. Following incubation for a number of weeks, the PACT and JSCG+ plates were examined under the microscope and crystals were discovered to have grown in some of the conditions assessed (Figure 3. 6). The crystals were diffracted to near 3 Å with a low X-ray yield source. The arabinose-GafASw were then diffracted with a thinner Xray beam using the macromolecular crystallography (MX) Beamline I03 at the Diamond Light Source, Hartwell Science and Innovation Campus, Oxford, Uk. Reflection data were collected to a resolution of 1.7 Å and following data collection and post processing, a PDB file was produced which allowed analysis of the structure.

3. 3. 9. 1 Characteristics of the GafASw 3D structure

The structure was crystallised as a dimer which led to further investigation of the possibility that the SBP is forming a homodimer. We determined the oligomeric state of isolated GafASw by static light scattering coupled to size-exclusion chromatography (SEC-MALLS), testing different concentrations (*ie*. 0.5-3 mg/ml) in the event where the dimer formation is favoured at higher concentrations. GafASw (Figure 3. 25) was found to have a molecular mass near 32.2 kDa, which matches the monomeric state of the protein, proving that its domains do not form stable dimers. Thus, the elucidated structure is a crystallisation dimer due to the way the protein packs itself within the salt lattice. Traces of the dimer were indeed

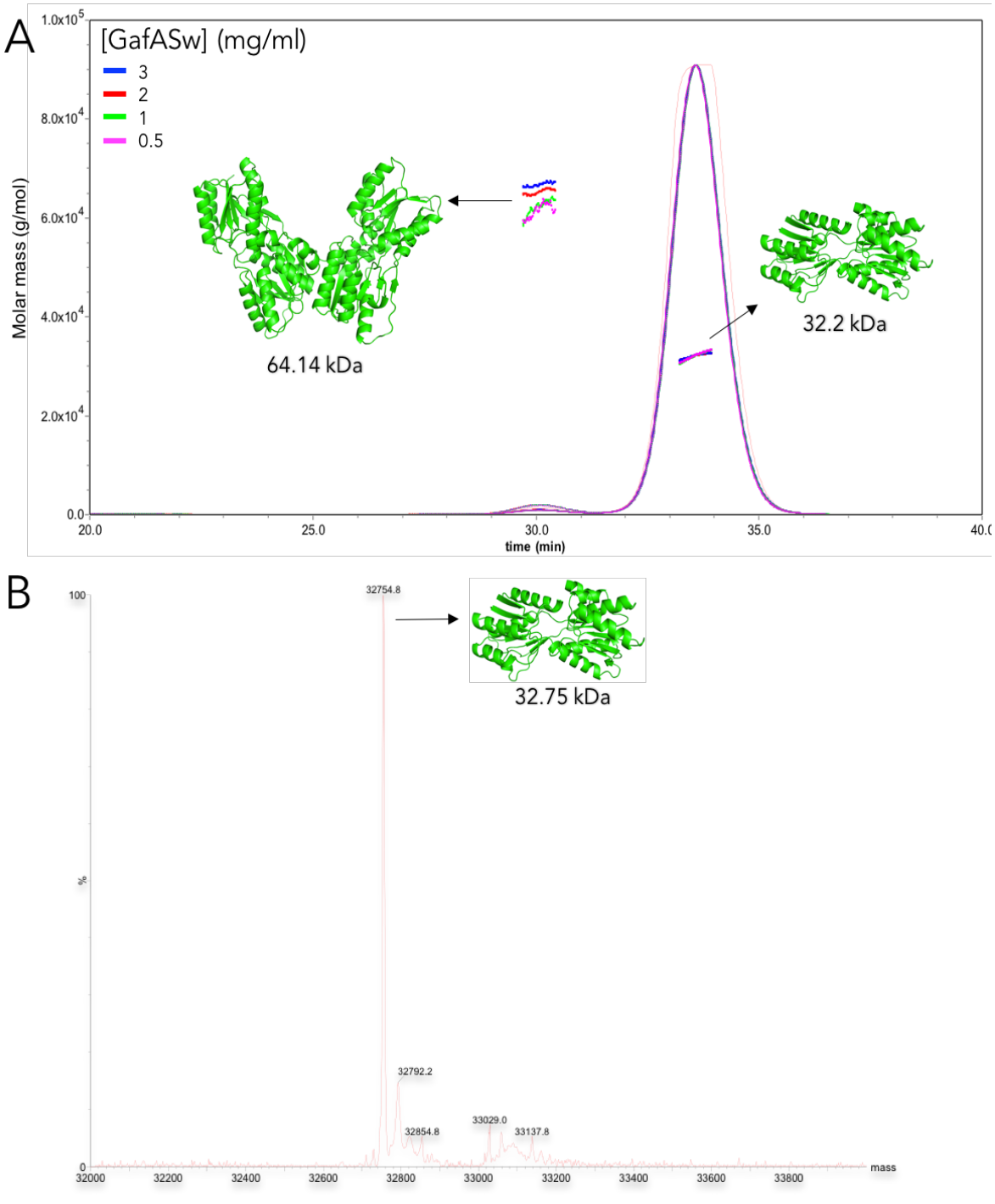


Figure 3. 25. GafASw exists in monomeric form in solution.

A) SEC-MALLS analysis of GafASw at four different concentrations, *ie*. 0.5, 1, 2 and 3 mg/ml, present in 50 mM Tris/ 150 mM NaCl. The monomer (*ie*. 32.2 kDa) is the prevalent form in solution as only traces of the dimer (*ie*. 32.2 kDa) are present.

B) Native ESI-MS analysis of GafASw (500-5000 m/z)). Protein was dialysed in 10 mM ammonium acetate and dissolved in 0.001 % formic acid/ 3 % methanol prior to the run.

GafASw structures were prepared on MacPymol.

detectable in solution (Figure 3. 24A), however not at high enough concentrations to justify a physiological homodimer. Native Electrospray Ionization mass spectrometry (ESI-MS) was used to corroborate the results above (Figure 3. 25B). The ESI-MS is an excellent technique for the study of intact biomolecular structure in the near-native state as present in the gas phase and has been successful previously in detecting protein complexes (Loo *et al*., 1997; Zhang *et al*., 2013; Benesch *et al*., 2007). Upon ESI-MS analysis of the protein (35 μM) under native conditions (85.5 : 11.5 : 3 : 0.001, water:ammonium acetate:methanol:formic acid), the only detectable protein species in the *m*/*z* space weighed 32.75 kDa verifying the monomeric state of the GafASw (Figure 3. 25B).

The structure shows residues 1 to 287 (+6) of mature GafASw protein with the His-Tag packed against the structure (data not shown), apparently without perturbing folding or ligand binding. There are two domains connected by a hinge region created from three peptide stretches (highlighted in red at Figure 3. 26, 3.27A) with binding cavity buried in between them, verifying that is a member of the Class I SBPs. The N-terminal domain is consisted of the residues 1-102 and 245-269 and contains six parallel β-strands, with five of them being surrounded by four α-helices; two found on top and two on bottom of the β-sheet (Figure 3. 27A). The order of the β-strands in the N-terminal domain is β2-β1-β3-β4-β5-β11 (Figure 3. 27B). The C-terminal domain encompasses the remaining residues (*ie*. 103-244 and 270-287). It features five β-strands aligned horizontally, in a perpendicular fashion, against the β-sheet of the N-terminal domain. The order of the strands is β7-β6-β8-β9-β10-β12 (Figure 3. 27B), with β12 positioned anti-parallel to the rest. A single disulphide bridge was detected formed between the residues Cys124 and Cys187 (Figure 3. 26, 27A). All the aforementioned features are conserved in GafAEc, confirming their ortholog relationship. The most important resemblance between them is binding of the furanose form of sugars; as the GafASw was crystallised with α-L-arabinofuranose bound.

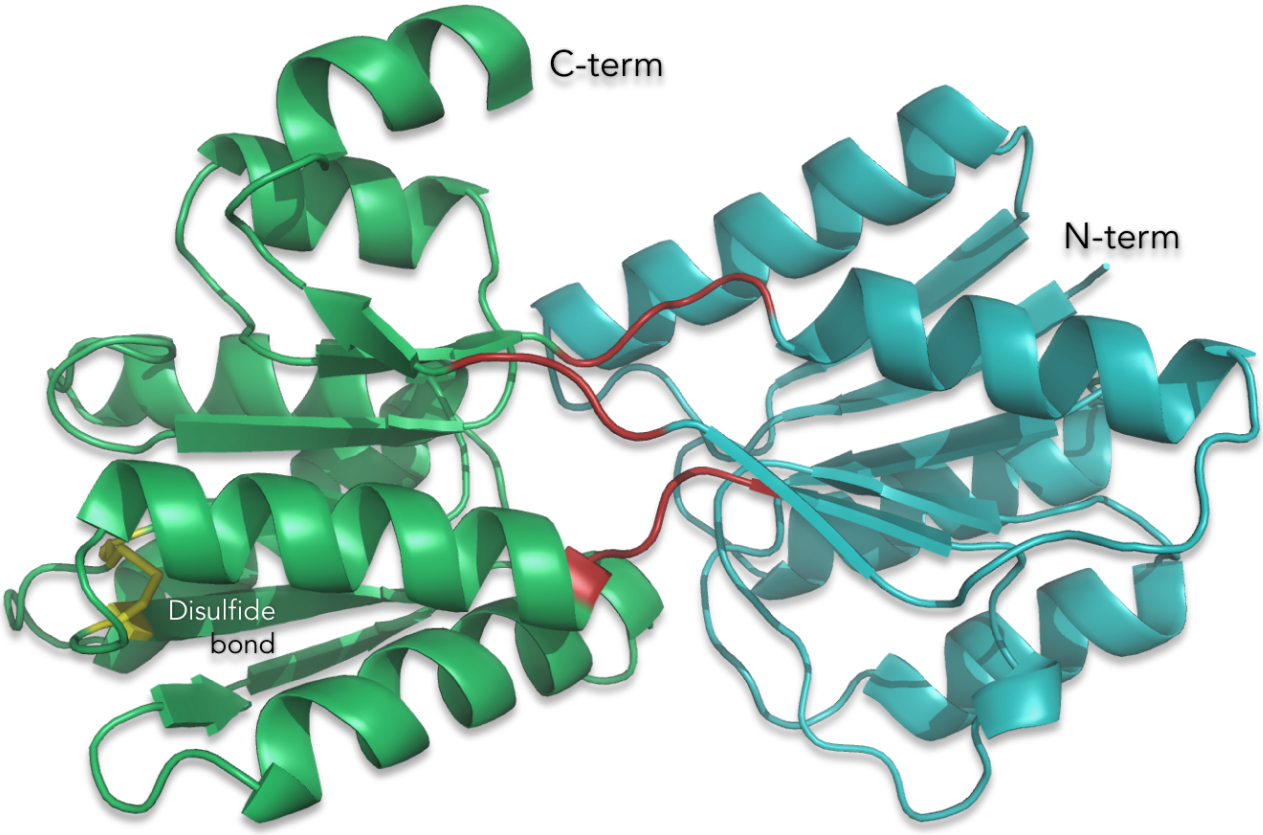


Figure 3. 26. GafASw belongs to Class I of SBPs.
Back view of GafASw structure to show the three connecting strands (red) which constitute the hinge region between N-terminal domain (blue) and C-terminal domain (green). The disulphide bond is presented in yellow.

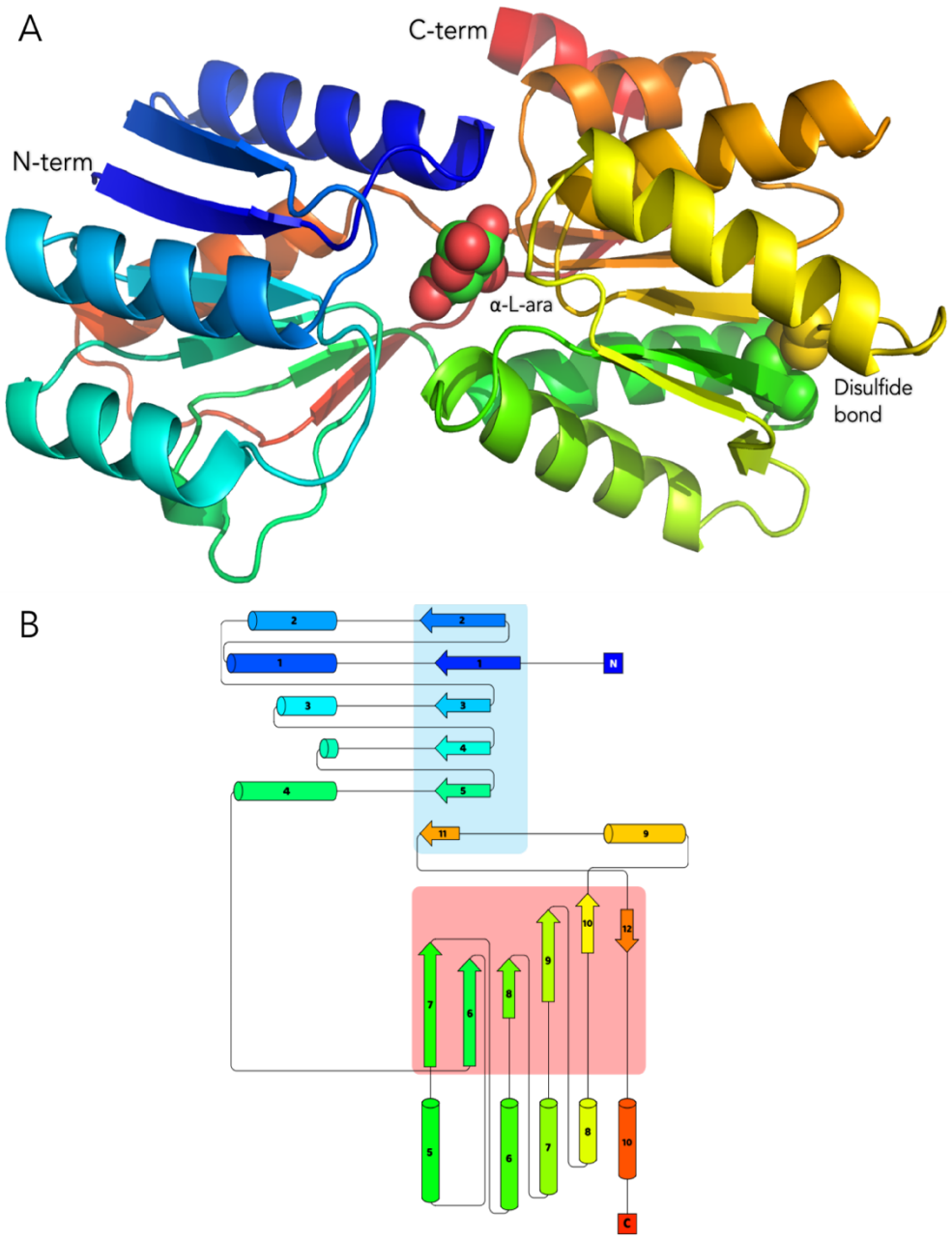


Figure 3. 27. Structure of GafASw in the closed-liganded state.

A) Tertiary structure of GafASw to show the α-helices and β-sheets. The bound ligand, *ie*. α-L-arabinofuranose is shown in the spheric format (green/red balloons) buried in between the two domains. The disulphide bond is designated in the spheric format as formed between the residues Cys124 (green) and Cys187 (yellow).

B) Topology diagram of GafASw produced in Pro-origami software. The β-strands are shown as arrows and the α-helices as cylinders. The β-sheet of the N-terminus domain is highlighted in blue and the C-terminus one in red. The blue to red colouring scheme presented corresponds to the structure in (A) and represents the direction of secondary elements from the N- to the C-terminus. The structure follows the same topology with GafAEc which is distinct to Class I SBPs; however, with additional β-sheets (β-11 and 12) and α helix (α10).

3. 3. 9. 2 Binding cavity of GafASw

The structure of GafASw was isolated as a monomer in Pymol to facilitate the study of its binding cavity. The stacking of L-arabinofuranose in the binding pocket of GafASw is stabilized by the contribution of the aromatic side chains of tryptophan 14 and phenylalanine 167, which act as a binding surface above and below the plane of the furan ring (Figure 3. 28A). Excluding these aromatic residues allows the hydrogen bonding interactions of the remaining residues with the ligand to be clearly examined (Figure 3. 28B). The oxygen atom of the furan ring is forming a hydrogen bond with the guanidino group of arginine 89. The oxygen atom of C1 is H-bonded with the glutamate 11 and the one at C2 is bound by Ser12, Asn195 and Asp223 (Figure 3. 28B). Asp223 is also participating in hydrogen bonding with C3 oxygen, as well as Arg143 does. The C4 oxygen is not bound by any residue. The C5 oxygen forms a hydrogen bond with the Asp88 and the Arg89. GafASw, same to GafAEc, harbours the three residues commonly found in monosaccharide SBPs at the bottom of their binding pocket, *ie*. asparagine, aspartate and arginine (see *Section 3. 1. 2*) (Figure 3. 28B).

3. 3. 9. 3 Comparison of GafASw binding site to GafAEc

The high identity of the primary protein sequences of GafASw and GafASm (ie. 67.57%) also applies for their tertiary structure. DaliLite server (Holm and Rosenström, 2010) was used to compare the 3D structure of GafASw against those in the Protein Data Bank (PDB), which returned a 61% identity similarity to GafAEc; that is the highest compared to the rest of the PBD structures, with second result being 30% less similar *ie*. ribopyranose SBP (ie. RsbB) from *Thermoanarobacter tengcongensis*. The two structures were superimposed in Pymol to compare their binding sites. Most of the interactions mentioned above are conserved (Figure 3. 29A) apart from some 'peculiarities' of the GafASw which explain its inability to bind D-galactofuranose. The salt bridge between Arg17 and Asp90 of GafAEc (*ie*. Arg15 and Asp88 in GafASw) is not present in GafASw and instead the Asp88 appears tilted by around 45° with respect to its counterpart, *ie*. Asp90, in GafAEc (Figure 3. 29B).

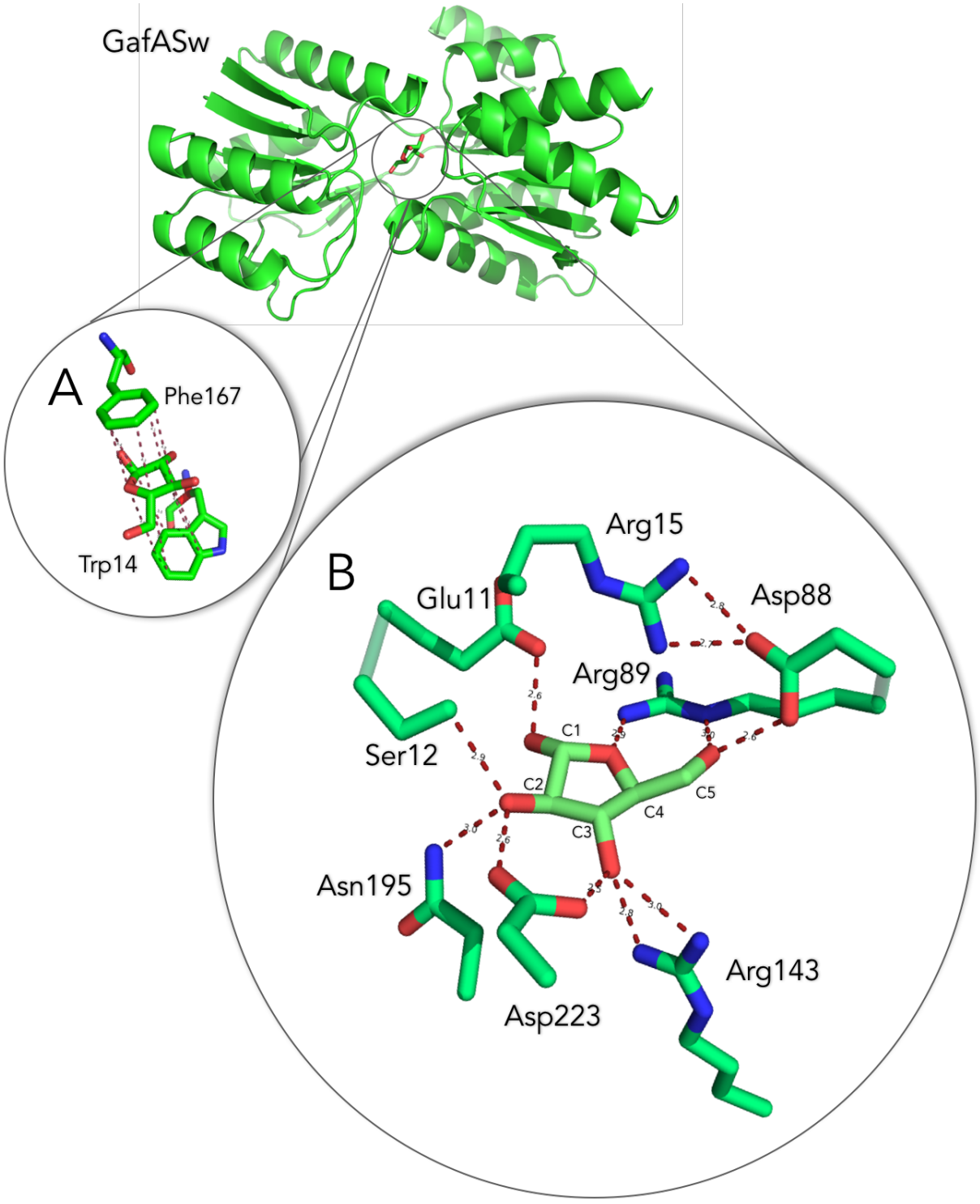


Figure 3. 28. Binding pocket of GafASw with the ligand β-L-arabinofuranose.

A) L-arabinofuranose and residues within 5.9 Å to represent the stacking interactions between the aromatic residues (Trp14 and Phe167) that act as the binding surface above and below the plane of the carbon ring.

B) The binding cavity of GafASw showing residues within 3.5 Å without the aromatic residues from (A). The carbon backbone of the ligand is numbered C1 to C5.

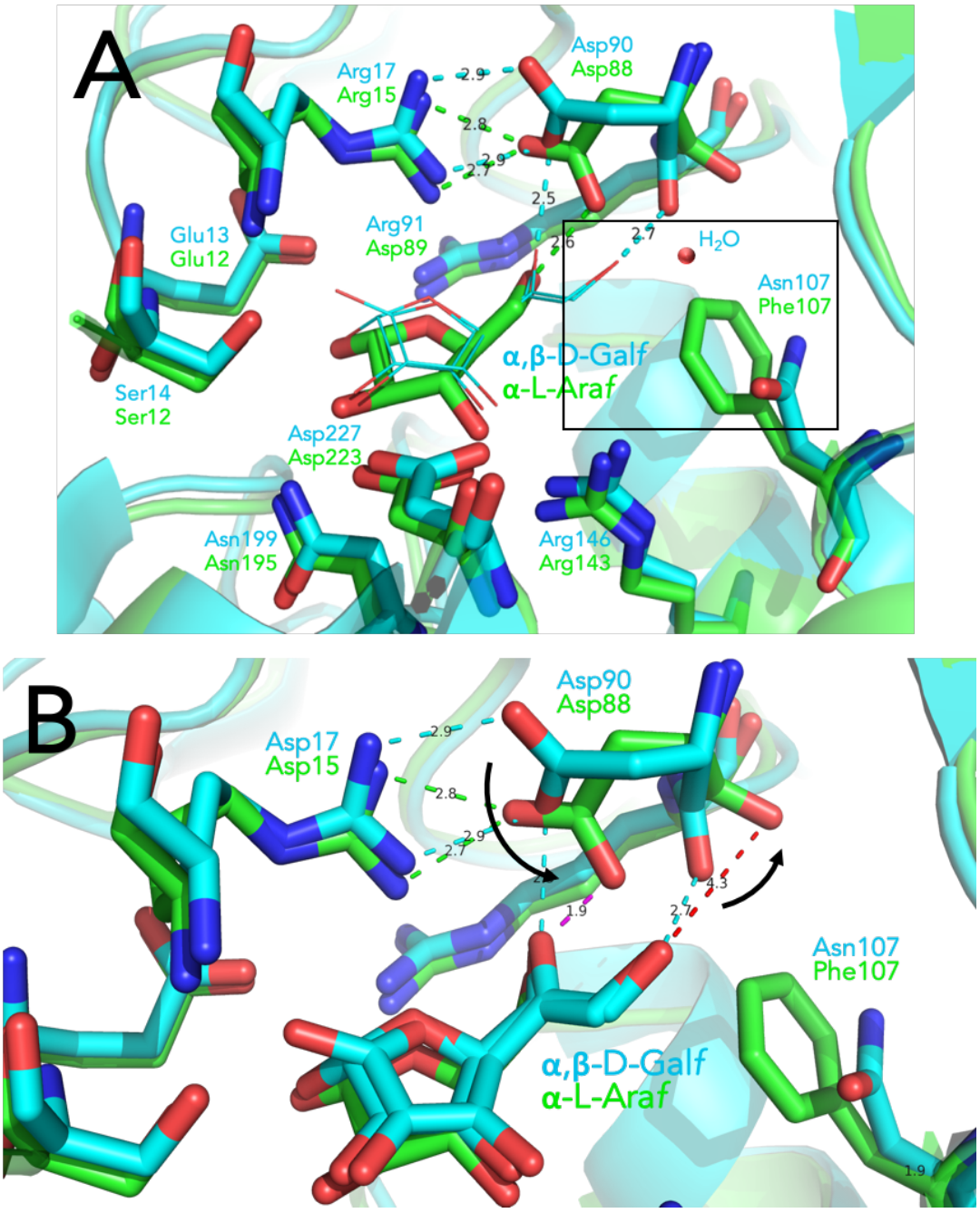


Figure 3. 29. Tilting of Asp88 hinders D-galactofuranose binding in GafASw.
A) The amino acid chains involved in coordinating D-galactofuranose by GafAEc (cyan) or L-arabinofuranose by GafASw (green) are labelled by their positions in the PDB files 2VK2 and 5OCP, respectively. The presence of Phe at position 107 could perturb the interaction of C6 with the H_2O molecule (designated in the box).

B) Closer view of the compared binding sites to present the tilting of Asp88 in GafASw. The distance between the C5 oxygen of a hypothetical D-galactofuranose and the carboxylate side chain of Asp88 is 1.9 Å, whereas between the C6 oxygen and the carboxylate of the backbone is 4.3 Å. The tilting in GafASw is shown using arrows which represent an approximate 45° shift of the Asp88 compared to Asp90, to face the bound L-arabinofuranose.

This leads to direct interaction of Asp88 with the C5 oxygen atom of L-arabinofuranose, while still maintaining hydrogen bonding with Asp90 (Figure 3. 29B). The tilting of Asp88 translates to inability of the SBP to bind sugars
similar in structure to D-galactofuranose, as such sugars will have their hemiacetal group (C5 and C6) pointing towards the direction of Asp88, as seen in GafAEc. If we were to fit the D-galactofuranose to the binding cavity of GafASw and assuming this would place its hemiacetal group at an identical position to the C5 linkage of the L-arabinofuranose, it's binding appears disrupted (Figure 3. 28B). In such case, the hydrogen bond between the oxygen of C5 from D-galactofuranose and the carboxylate group of the Asp88 (or 90) side chain shortens from 2.5 Å to 1.9 Å (bond shown in pink at Figure 3. 29B). Also, the size of the hydrogen bonding between the oxygen of C6 and the carboxylic group of Asp88 backbone, extends from 2.7 Å to 4.3 Å, demonstrating the unfavourability of the binding. The mean donor-acceptor distances in protein secondary structure elements are close to 3.0 Å (Jeffrey, 1997). The hydrogen bonds have been categorised based to their relative distances between donor-acceptor: (i) 'strong, mostly covalent' = 2.2 - 2.5 Å (energy of bond = 40-14 kcal/mol), (ii) 'moderate, mostly electrostatic' (energy of bond = 15-4 kcal/mol) = 2.5 – 3.2 Å and (iii) 'weak, electrostatic' = 3.2 – 4 Å (energy of bond 4 kcal/mol) (Jeffrey, 1997). Indeed, in GafASw the binding of L-arabinofuranose involves hydrogen bonding from the category (i) and (ii). However, in the case of D-galactofuranose the hypothetical bonds are either too weak (4.3 Å) or too strong (1.9 Å). Further, a residue which could potentially lead to perturbation of D-galactofuranose binding is the phenyalanine 107 (Figure 3. 29A). This is an asparangine in GafAEc, which due to its charged nature permits the presence of a water molecule and its interaction with the oxygen of C6 (Horler *et al*., 2009). We predict that the hydrophobicity of the phenylalanine would dislocate the water molecule and hinder stabilisation of the C6 at this position (Figure 3. 29A).

3. 3. 9. 4 Comparison of GafASw with ribopyranose and ribofuranose SBPs

The binding cavity of GafASw was compared to the D-ribopyranose SBP from *E. coli*

(*ie*. $RsbB_{Ec}$) (Mowbray and Cole, 1992) because of their high homology (*ie*. most homologous non-GafA SBP to GafASw with sequence identity equal to 33%). The most notable feature is the conservation of Asp88 in $RsbB_{Ec}$, albeit at a further ~30° tilting compared to GafASw (Figure 3. 30A). This places its Asp88 at an approximately overall 75° tilted position when compared to GafAEc. Such shift facilitates hydrogen bonding with the C1 and C2 oxygen atoms of the ribopyranose. The Arg at position 17 is replaced by a phenyalanine residue as the formation of a salt bridge (*ie*. GafAEc) or hydrogen bonding (*ie*. GafASw) is not supported due to the Asp88 shifted positioning (Figure 3. 30A). This observation led us to assume that the Asp88 is essential for stabilisation of the bound ligand, with its degree of tilting determined by the length and the substituent of the C5 and/or C6 of the furan ring (Figure 3. 30B). The long hemiacetal group at C5 of the D-galactofuranose 'forces' the Asp88 away and facilitates the bridge formation. When it comes to GafASw, the hemiacetal group of the pentose L-arabinofuranose is shorter and therefore allows the carboxylate group of the Asp88 side chain to move closer to it and form a hydrogen bond with C5 substituent. In the case of $RsbB_{Ec}$, the Asp88 interacts with the oxygen atoms of the anomeric carbon (C1) and C2 in the ribopyranose and therefore causes the residue to be almost perpendicularly positioned against the sugar ring (Figure 3. 3B).

We predict that because of the non-variable substituents at the C1 and C2 positions (*ie*. typically a hydroxyl group), the position of the Asp88 will be conserved if the plane of the sugar ring is sat in such way that these substituents face the aspartate in question, similarly to $RsbB_{Ec}$ binding cavity (Figure 3. 3B, C). This is indeed the case when the binding cavity of other pyranose binding proteins are examined; for example, allopyranose SBP (*ie*. $AlsB_{Ec,}$ 1RPJ) (Chaudhuri *et al.*, 1999), arabinopyranose SBP (*ie*. $AraF_{Ec,}$ 1ABE) (Quiocho *et al*., 1984), xylopyranose SBP (*ie*. $XylF_{Ec}$, 3M9X) (Sooriyaarachchi *et al.*, 2010) and galactopyranose SBP (*ie*. $MglB_{Ec}$, 2HPH) (Cuneo *et al*., 2007) all have the Asp residue in question perpendicularly positioned and interacting with both C1 and C2 (or just the C1) substituents. Notably, this relationship observed between the C5 and C6 of furanose rings and C1 (, C2) substituents from pyranose rings with Asp88 is confirmed by examination of the binding cavity of a

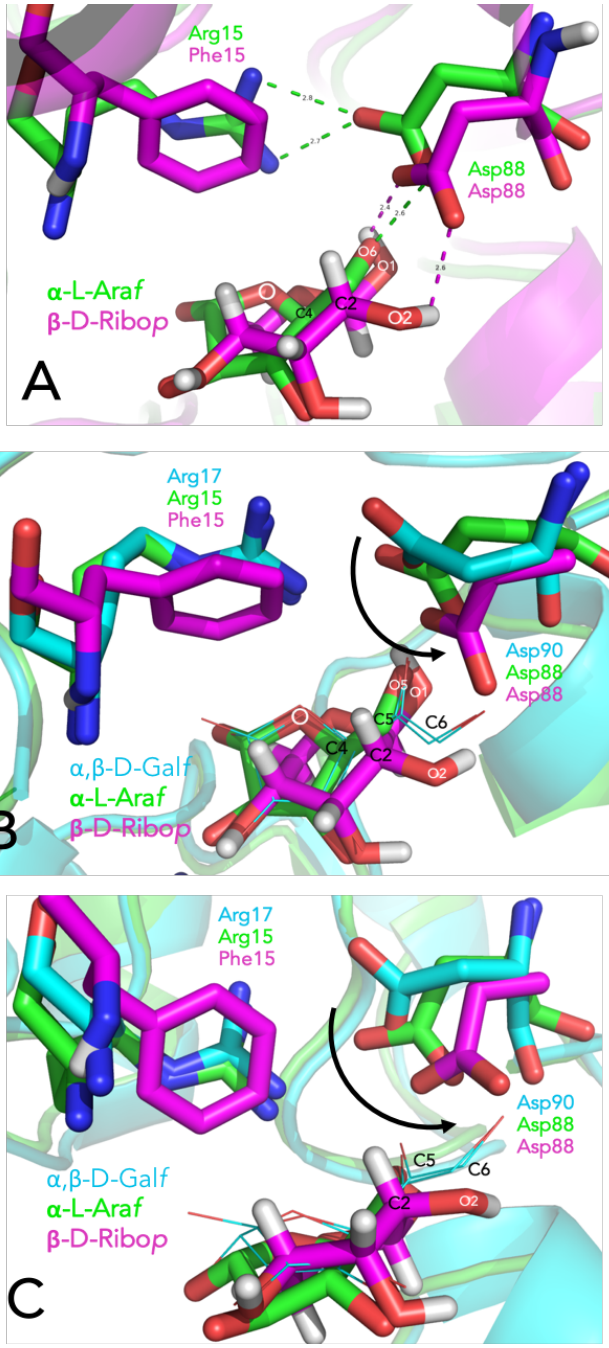


Figure 3. 30. Tilting of Asp88 is determined by the length and substituent of the nearest carbon in the sugar ring

A) The amino acid chains involved in coordinating D-ribopyranose by $RsbB_{Ec}$ (magenta) or L-arabinofuranose by GafASw (green) are labelled by their positions in the PDB files 2DRI and 2VK2, respectively.

B) Upper view of the superimposed binding sites of GafASw (green) with GafAEc (cyan) and $RsbB_{Ec}$ (magenta) to show the proximity of C6 and C1 constituents of the sugars to Asp88.

C) Side view of the superimposed binding sites of GafASw (green) with GafAEc (cyan) and $RsbB_{Ec}$ (magenta) to show the overall tilting of the Asp88 residue.

ribofuranose SBP from *Hahella chejuensis* (*ie*. $R_f$bp, 3KSM) (Bagaria *et al*., 2011). Based on this discrepancy observed between the furanose and pyranose binding proteins regarding interactions with the Asp88 of their respective SBPs, we would expect that the oxygens of C5 and C6 substituents in the ribofuranose ring to hydrogen bond with the Asp88 and provoke a tilted position, similar to D-galactofuranose since they are both hexoses with equally elongated hemiacetal group. Surprisingly, the structural equivalent to Asp88 in R_fBP_{Hc}, *ie*. Asp127 was instead perpendicularly positioned, like the pyranose forms mentioned above, with the oxygen atoms of C3 and C2 forming hydrogen bonds with its carboxylate group, and the C6 constituent interacting with an asparagine residue instead at the opposite side of the ring (Figure 3. 31A, B). This confirmed the notion that the Asp88 remains facing directly the sugar ring, unless an elongated C5 (*ie*. L-arabinofuranose) or an even larger C6 (*ie*. D-galactopyranose) hemiacetal groups from the furan forms need to be accommodated.

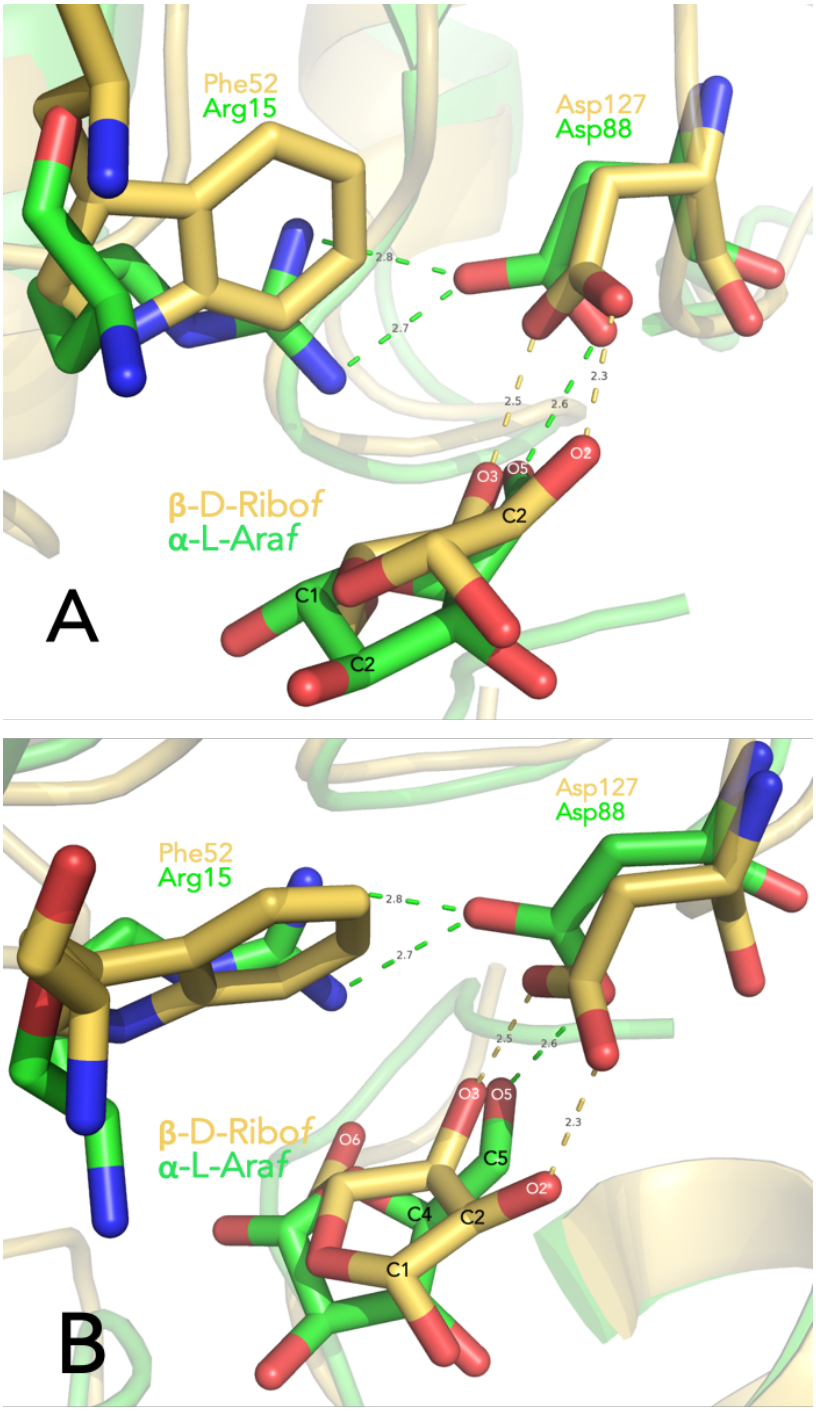


Figure 3. 31. Asp127 interacts with C3 and C2 of ribofuranose in R_fbp_{Hc}

A) Side view of the amino acid chains involved in coordinating D-ribofuranose by R_fbp_{Hc} (light brown) or L-arabinofuranose by GafASw (green) are labelled by their positions in the PDB files 3KSM and 2VK2, respectively. The structureal equivalent of Asp88, *ie*. Asp127, interacts with C2 and C3 of the ribofuranose ring.

B) Upper view of the superimposed binding sites of GafASw (green) with GafAEc (cyan) and $RsbB_{Ec}$ (magenta) to show the proximity of C6 and C1 constituents of the sugars to Asp88.

Oxygen atoms are shown in white and the carbon atoms in black font.

3. 4 Discussion

The discovery that furanose-specific (Horler *et al.*, 2009) ABC systems exist has raised their importance in strategies for optimisation of the transport capacity of fermenting bacteria for biofuel production. Here we focus in the transport of the furanose form of arabinose, an abundant sugar which exists as a branching unit of the xylan backbone. Bioinformatics analysis of the biochemically and structurally characterised SBP of GafAEc (Horler *et al.*, 2009), detected GafASw and GafASm as potential candidates for further characterisation because of their strong bias towards arabinose utilisation. Following overproduction of GafASw and GafASm in E. coli (see Chapter 2), we sought out to biochemically and structurally characterise the candidates using DSF, intrinsic protein fluorescence and ITC.

3. 4. 1 The role of GafASm in *S. meliloti* 1021

Different to other GafAs, the genomic neighbourhood of *gafASm* doesn't have a strong relevance to monosaccharide arabinose, but it does to arabinosides instead. It neighbors with arabinoside ABC transport system, AraNPQ and a putative arabinofuranosidase, *ie*. AbfB. This denotes its importance in the uptake of arabinofuranose released from xylan and arabinan. The absence of genes encoding for catabolic enzymes, more hydrolases, transcription factors and two-component systems for sensing indicates that these are possibly in close proximity with other arabinose-related transport systems. Such assumption is highly possible, considering the significant number of genes encoding for ABC transport systems in *Sinorhizobium* meliloti and in Rhizobia species ranging from 40-70% of all of the transporter proteins encoded in the genome (Mauchline *et al*., 2006; Giulani *et al*., 2011). There are 200 ABC genes in *Sinorhizobium meliloti*, 216 in *Mesorhizobium loti*, 269 in *Rhizobium leguminosarum*, and 240 in *Bradyrhizobium japonicum* compared with 67 in *Escherichia coli* and 124 in *Pseudomonas aeruginosa* (Mauchline *et al*., 2006). BlastP searches of the arabinopyranose binding protein $AraF_{Ec}$ against the *Sinorhizobium meliloti* 1021 genome detected a predicted orthologue, *ie*. SM0020_05120. However,

examination of its genomic cluster didn't detect any arabinose-relevant genes. Nonetheless, we would not be able to draw relevance to arabinose by the catabolic genes for the arabinose degradation pathway III, *Ab-araABCDEF*, as these were detected to be carried on a megaplasmid, *ie*. pSymB megaplasmid, and not on the chromosome (Poysti *et al*., 2007).

More information about *gafABCDSm* was derived from the study done by Mauchline *et al.*, (2006), which mapped the SBP-dependent transportome of *S. meliloti* 1021. The fact that the expression of *gaf* system is induced by the presence of L-arabinose, D-fucose and talose but not by D-galactose was a good indication that this transporter might be unable to bind D-galactose (Mauchline *et al.*, 2006). Conversely to assumption D-galactose binds GafASm, though the system is unlikely to be a dedicated D-galactofuranose transporter since it doesn't induce its expression. Even though we weren't able to measure the binding affinity for D-galactose, Bourdès *et al*. (2012) developed a FRET-based biosensor to rapidly screen the binding affinities of the plethora of ABC systems found in *Sinorhizobium* and *Rhizobium* species. The direct ortholog of GafASm from *Rhizobium leguminosarum* bound D-galactose with 5. 3 μM binding affinity. Due to the incredibly high sequence identity (87 %) between the two orthologues, we expect the GafASm to bind galactose with similar affinity. Bourdès and his coworkers tested whether arabinose binds to RL2376-79, prompt by a different study conducted by Ramachandran *et al.*, (2011) which showed by microarray analysis that presence of L-arabinose causes upregulation of the aforesaid operon. However, Bourdès report incorrectly that the D conformer caused the upregulation in the Ramachandran *et al.* (2011) study, and actually proceeded in testing this conformer for binding to RL2376. When D-arabinose was tested there was no change in FRET-ratio and therefore the group concluded that it doesn't bind. Assessing the hypothetical scenario of D-arabinofuranose binding, this would resemble D-allofuranose in that they both have their hemiacetal group pointing upwards in respect to the furan ring. However, this group in D-arabinofuranose is smaller by a carbon molecule, since it's a pentose, and thus its hydroxyl group might not be extending long enough to reach the residue which interacts with O_6 from D-allofuranose. The rest of the presumptive

interactions are predicted to be conserved as they are the same with L-arabinofuranose. Nonetheless, the D-arabinose is not as abundant in nature and most importantly is not found in xylan (Saha and Bothast, 1999), pectin (Prade *et al*., 1999) or root mucilage (Tyler, 1965; Wise, 1960) therefore, due to its low physiological importance for adaptation of *R. leguminosarum* in the rhizosphere is not suprising the D conformer doesn't bind.

The Tm of GafASm in the presence of D-fucose was stabilized to the same extend with D-galactose, therefore they are thought to be binding with similar affinities. However, GafAEc binds D-galactose with a 6 times higher affinity than D-fucose (Horler *et al*., 2009). A potentially enhanced affinity for D-fucose compared to GafAEc, could have been the result of evolutionary adaptation in rhizosphere as this sugar is a prominent component of the plant root mucilage. D-fucopyranose accounts for 20% of the total sugar content of the mucilage released by the root cap of maize (*Zea mays L.*) (Chaboud, 1983), 9% in cowpea (Moody *et al*., 1988) and approximately 8% in rice (*Oryza sativa L.*) mucilage (Chaboud and Rougier, 1984). An environmental niche rich in fucose would potentially act as an evolutionary force and shape the ligand binding specificity of GafASm differently to its orthologs. Even though the furanose form accounts only for 5% of the total fucose in aqueous solution (Angyal, 1984), the highly competitive environment of the rhizosphere would mean that species equipped with the necessary transport systems for utilisation of all or most furanose forms of sugars, would exhibit increased fitness compared to the rest and therefore gain a competitive edge.

Furthermore, a high affinity system will be suited in the case that the bacterium is utilising the carbon source as a chemoattractant. For example, in *Azospirillum brasilense*, SbpA is involved in chemotaxis towards D-galactose, L-arabinose and D-fucose and is the SBP for a high affinity ABC transporter for D-galactose (Van Dommelen et al., 1997). The similarity of GafASm to SbpA binding specificity is underlying the importance these sugars have in the chemotaxis and success of symbiosis between bacteria and plant roots. The examples are not limited to SbpA; a

more extensively studied SBP is ChvE from *Agrobacterium tumefaciens*. This SBP was shown to participate in chemotaxis and virulence gene induction by working with the VirA/VirG two component system (Kemmer, Liang and Nester; 1997) and respond to sugar binding during infection of plants. Its multiple functionalities are activated following binding of a broad set of sugars including glucose, galactose, galacturonic acid and glucuronic acid (He et al., 2009; Hu et al., 2012), thus highlighting the crucial role a high affinity sugar SBP can play in the fitness of a bacterium; in this case by promoting virulence signalling. Such capacity will potentially offer a competitive advantage to the bacterium. An interesting arabinose-related ABC transporter is the AraABC transporter which is encoded by the p*Sym* plasmid; this was shown to be necessary for arabinose uptake but its expression was unresponsive to galactose (Poysti *et al*., 2007). AraBCD was hypothesised to carry out a similar role to ChvE since the level of its expression is drastically increased in the presence of seed exudates (Poysti *et al*., 2007) and therefore could offer an increased competition phenotype. This transporter is more than likely to be specific for the pyranose form of arabinose, as there is already one chromosomal copy for the furanose form and is also able to bind glucose which is not a property of the GafAs as demonstrated with GafAEc (Horler *et al*., 2009), and further established in the current study. When the *araA* was conjugated into a Δ*chvE* mutant of *A. tumefaciens* with a *virE* reporter plasmid, it failed to induce *virE* therefore is unable to complement the virulence gene induction in the absence of ChvE (Geddes and Oresnik, 2012). If a hypothetical solution containing only L-arabinofuranose is assumed then GafASm would have a binding affinity of 52.5 nM(*ie*. 12% of overall L-arabinose content) (Table 3. 5), demonstrating the high scavenging capabilities of the system. Therefore, there is a possibility that GafA could carry out a similar function to ChvE, as the arabinopyranose-specific ABC transporter from *S. meliloti* 1021 fails to do so. Further, for *Sinorhizobium meliloti* species, the uptake of disaccharides such as trehalose and sucrose was found to increase the competitiveness of the species for nodule occupancy in the root of alfalfa legume. This was exemplified by mutating the ABC transport systems responsible for for transport of trehalose/ sucrose which in turn led to poor nodulation phenotype

Table 3. 5. Binding of various ligands to purified GafAs using DSF, intrinsic fluorescence and ITC.

Ligand	GafA orthologue		
	GafAEc	GafASw	GafASm
L-arabinose	1.3 ± 0.24 µM	6.64 ± 0.66 µM (4.19 ± 0.54 µM)*	0.43 µM
D-galactose	1.7 ± 0.3 µM	NB*	+
D-allose	30.3 ± 4.4 µM	24.8 ± 0.22 µM (27.4 ± 0.43 µM)*	+
D-fucose	6.2 ± 0.54 µM	+	+
D-talose	25.7 ± 2.5 µM	+	NB
D-glucose	NB*	NB	NB

All quantitative assessment of ligand binding was done by intrinsic protein fluorescence apart from:
*Binding affinity as calculated by ITC,
(+), Indicates binding as assessed qualitatively by DSF/ NB, No binding of the ligand as confirmed by DSF
NB*, No binding of the ligand as confirmed by DSF and/or intrinsic fluorescence and/or ITC.

(Jensen, Peters and Bhuvaneswari, 2002). To confirm the importance of the GafABCD system in nodule occupancy and utilisation of seed exudates, the respective operon can be deleted and it's the effect on growth of *S. meliloti* 1021 in presence of seed exudates and individual sugars can be assessed. Further, its potential as an inducer can be assessed using the *A. tumefaciens virE* reporter strains AT11043 similarly to the experiment performed with araA by Geddes and Oresnik (2012).

3. 4. 2 The role of GafASw in *Shewanella* sp. ANA3

The expression of *gafABCD* system from *Shewanella* sp. ANA3 is predicted to be regulated by the arabinose controlled AraR TF. Along with its arabinose related expression, the operon is found in a huge cluster for arabinan utilisation suggesting that this bacterium is encountering this polysaccharide frequently and is probably importing the L-arabinofuranose monomers by employing the GafABCD transporter. Along with the above indications, the bacterium is unable to grow on D-galactose (Rodionov *et al*., 2010). This was not attributed to the absence of metabolic enzymes

for galactose catabolism as these are present in all *Shewanella* genomes analysed by Rodionov *et al*., 2010. Instead, the growth phenotype of the *Shewanella* species is consistent with the distribution of the galactose secondary transporter, *ie*. GalP, as the strains which lacked the transporter including *Shewanella* sp. ANA 3 failed to grow on D-galactose. Using BlastP we found that the mosaic distribution of *galP* relation to galactose transport doesn't apply to *gafABCD* distribution. Only one of the galactose 'growers', *ie*. *S. baltica,* harbours the *gaf* operon whereas *S. loihica* PV-4 and *S. woodyi* do not. Therefore, there isn't high confidence in relating the positive phenotype to the presence of GafABCD. If all strains had it, then it might be that the system contributes to galactose import in the cell. However, since GafAEc binds D-galactose and Rodionov *et al.,* (2010) established a correlation between transport and inability of *Shewanella* spp. (including sp. ANA 3) to grow on galactose, we are prone to believe that GafASw is: a) unable to bind galactose or b) since its regulated by AraR, is not responsive to galactose thus cannot enable growth in such conditions.

Biochemical analysis of GafASw revealed its ligand specificity and showed that indeed this SBP is unable to bind D-galactose. This unique feature is within the few cases arabinose related proteins lack specificity for galactose. The arabinopyranose $AraF_{Ec}$ from *E. coli* (Declerck and Abelson, 1994), GafAEc (Horler *et al*., 2009) and AraE from *Bacillus subtilis* (Krispin and Allmansberger, 1998). This principle extends to catabolic enzymes as AraA, the first enzyme in the arabinose catabolism, binds D-galactose (Wallace *et al*., 1978). The structural analysis attributed this inability to tilting of Asp88 disrupting a salt bridge with a nearby Asn residue and the presence of a phenylalanine residue instead of an asparagine, compared to GafAEc. The phenylalanine residue is predicted to disrupt interaction of the O_6 of galactofuranose with a water molecule and thus destabilizing its hypothetical binding. Such observation increases the bioindustrial importance of this transporter as the binding of L-arabinofuranose is not inhibited by D-galactofuranose.

The binding affinity for L-arabinofuranose is 4 times and 12 times weaker than GafAEc and GafASm (Table 3. 5), potentially indicating that *Shewanella* sp. ANA3 encounters

high arabinofuranose concentrations in its close proximity which would excuse a lower affinity. The high capacity of the bacterium to increase the arabinofuranose concentration when grown near arabinan is implied by the large arabinose-utilisation cluster. RegPrecise and Rodionov *et al*., (2010) have assigned roles to all of the genes in close proximity to *gafABCD* and showed that there is a plethora of arabinosidases and arabinofuranosidases. Specifically, they identified 2 arabinosidases, one predicted to reside extracellularly and one in the periplasm. Also, 3 arabinofuranosidases are thought to be present in the cluster, with each localized in separate compartments *ie*. one in the cytoplasm, one in the periplasm and the last found extracellularly (Rodionov *et al*., 2010). Based on localization of these enzymes in different compartments we designed a model for arabinan utilization shown in Figure 3. 32. The model predicts that the hydrolysis of arabinan produces arabinosides and arabinofuranose which make their way into the periplasmic space through the outer membrane porin (*ie*. $OmpA^{Ara}$, also expressed in the cluster). In the periplasm, the arabinofuranosidase (AbfA) hydrolyses terminal non-reducing arabinofuranose which in turn binds to GafASw and gets translocated into the cytoplasm (Figure 3. 32). The endohydrolytic activity of arabinosidases (*ie*. AbnA and Arb43) further decreases the arabinosides to a transportable size (*ie*. arabinobiose to arabinotetraose) which enter the cytoplasm through the arabinosides (Figure 3. 32). In the cytoplasm, the arabinofuranosidase AbfA releases arabinofuranoses and further fuels the primary catabolism of arabinose. The model increases the importance of GafASw in arabinose transport in the furanose form, as its positioned very close to its substrate and therefore allows for rapid uptake avoiding equilibration to the pyranose form. Nonetheless, a BlastP search of the arabinopyranose SBP from *E. coli*, $AraF_{Ec,}$ yielded no match in *Shewanella* sp. ANA-3. Combined with the notable decrease in fitness upon deletion of *gafASw*, as reported by the Fitness browser (Wetmore *et al*., 2015), we are prompt to speculate that the GafABCDSw is the only dedicated transporter for L-arabinose monomers. The absence of an arabinopyranose dedicated system may suggest that arabinan and xylan are the only source of arabinose easily accessed by this bacterium. Since the estuarine environments exhibit high level of biodiversity, we

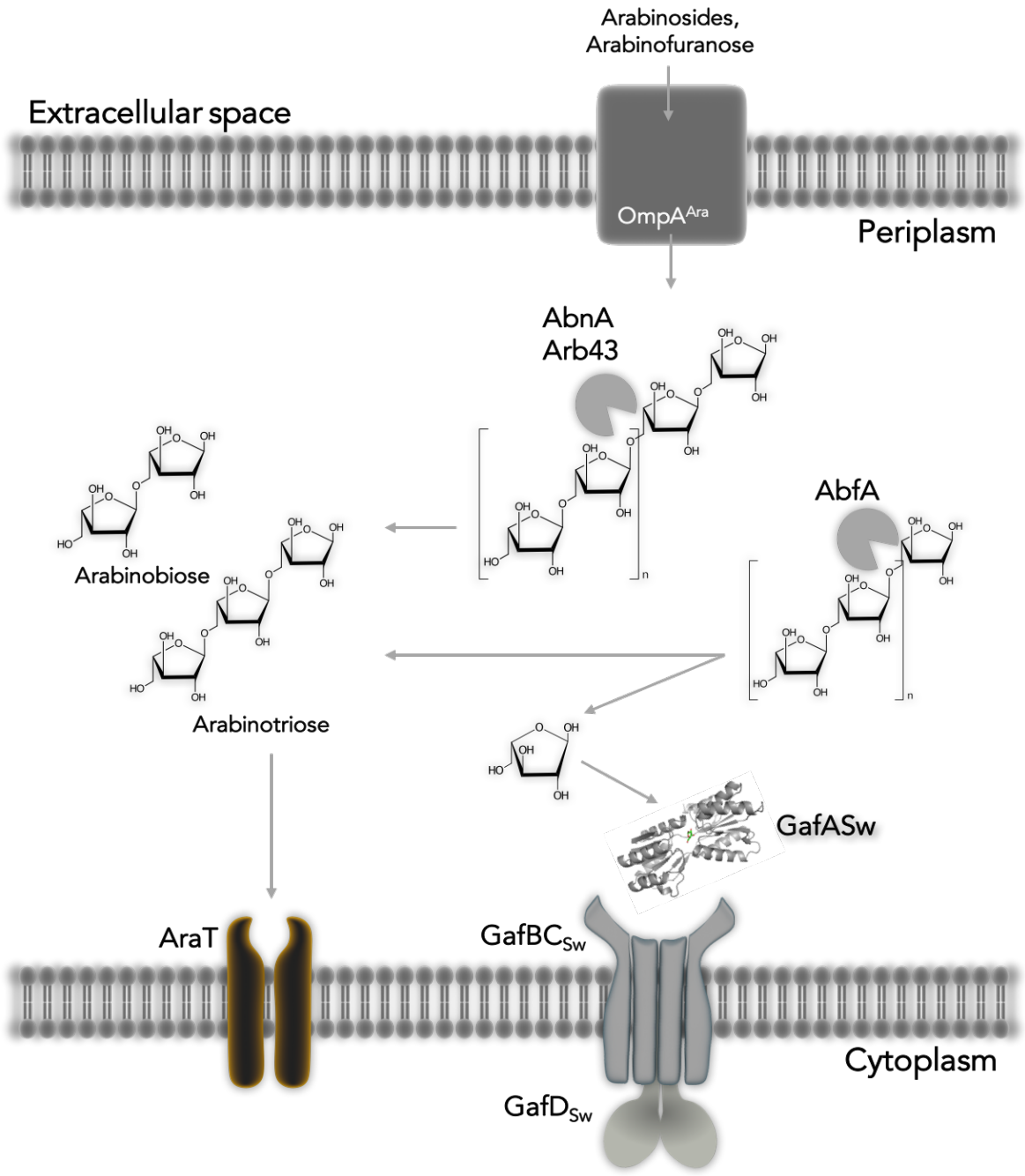


Figure 3. 32. Model for arabinan utilisation by Shewanella sp. ANA 3, MR4 and MR7

The released arabinosides and arabinose from the activity of the extracellular arabinosidase and arabinofuranosidase make their way into periplasm through the outer membrane porin ($OmpA^{Ara}$). In the periplasm, the arabinosidases (AbnA and Arb43) cleave the arabinosides into smaller products which travel through the AraT secondary transporter into the cytoplasm. The periplasmic arabinofuranosidase (AbfA) catalyses the production of arabinofuranose which binds on GafASw and translocates through into the cytoplasm via the transmembrane domain of the $GafBCD_{Sw.}$ The model is based on the predictions of Rodionov *et al.*, 2010.

cannot completely exclude the presence of galactose in such environment, or be definite about its enrichment by arabinan or xylan polymers. Nonetheless, arabinan utilisation genes have previously been identified in bacteria from estuarine and mangrove environments including *Mangrovimonas*-like strains (Dinesh *et al*., 2016), *Strepromyces xiamenensis* (Xu *et al*., 2009), *Mangrovibacterium di*azotrophicum (Huang *et al*., 2014) and *Halomonas avicenniae* (Soto-Ramírez *et al*., 2007).

3. 4. 3 Structure of GafASw

The structure of GafASw follows the pattern of Class I SBPs with an additional α-helix and two β-strands that are also found in GafAEc and MglB and align perfectly until the penultimate residue. The binding of L-arabinofuranose is similar to other SBPs and involves two aromatic residues that stabilise the binding by stacking the plane of the sugar ring in between then. It also contains a disulphide bond which is seen in LivJ and LivK SBPs but not in other solved structures. The location of the bond outside the binding cavity would suggest that is not implicated in the function of the protein. The superimposition of GafAEc, GafASw and Rsb_{Ec} showed that the tilting of Asp88 is deduced by the extend of the hydroxymethyl group at the C_5 or C_6 of the ring. Notably, this is confirmed by analysis of the binding cavity of the ribofuranose binding SBP, *ie*. $R_fBP_{Hc,}$ with the aspartate residue positioned perpendicularly against the sugar as there is no extended C_6 group, and it interacts with the O2 and O3 instead.

We weren't able to predict whether the GafA ortholog we failed to characterise, *ie*. GafACb, is unable to bind D-galactofuranose merely by aligning the sequence as all of the residues of the binding cavity are conserved (Figure 3. 33). However, there is a methionine present in the phenylalanine position of GafASw (*designated by a blue circle*, Figure 3. 33) which is not predicted to cause water exclusion. The neutral and nonpolar methionine is considerably smaller than asparagine found in GafAEc, and therefore we don't expect it to stabilise the water molecule at such position. Most importantly, we wouldn't be able to predict the extend of tilting the respective Asp residue has in GafACb. Nevertheless, *Clostridium beijerinckii* NCBI 8052 has been

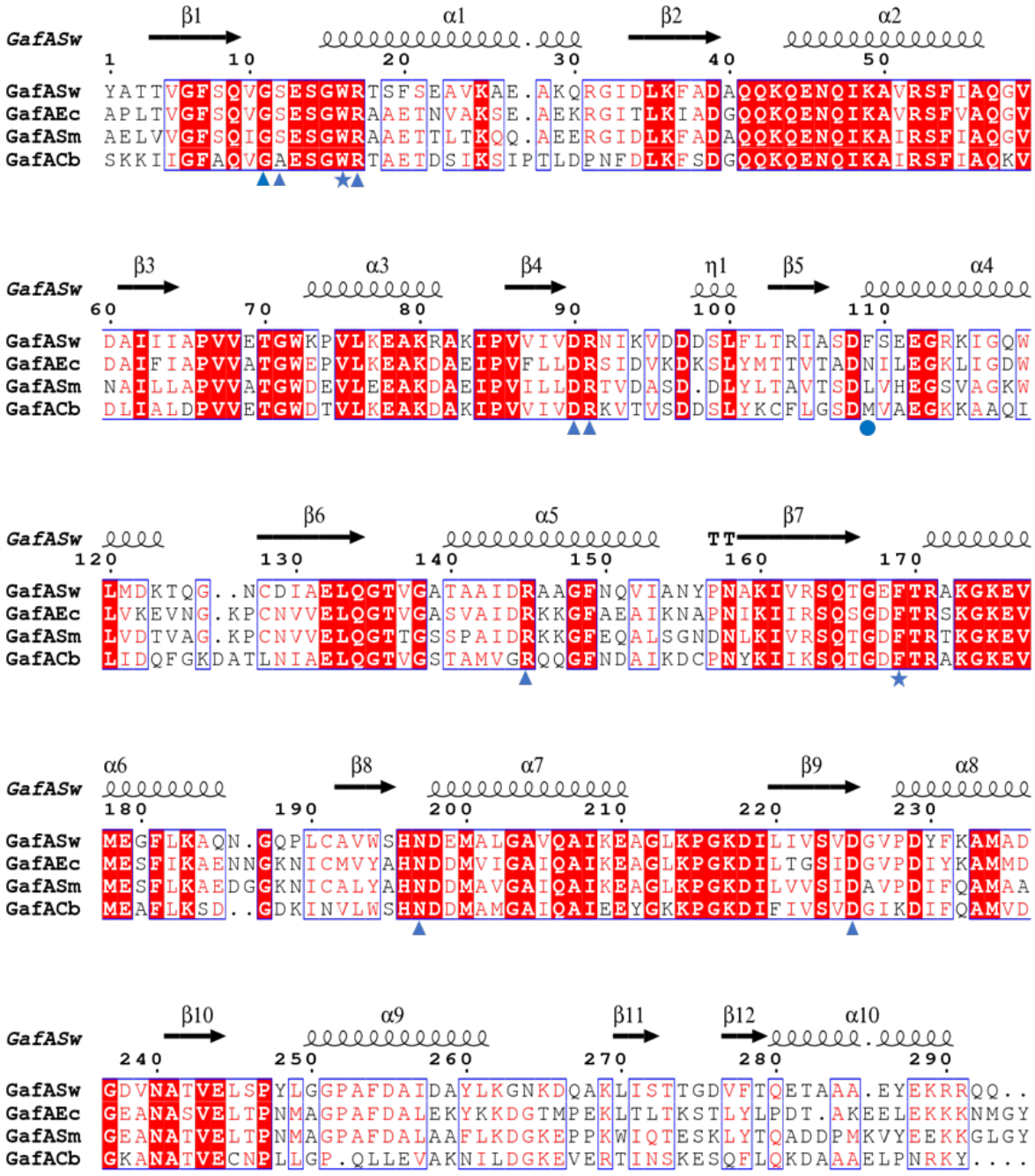


Figure 3. 33. Multiple structure-based alignment of the amino acid sequences of GafASw, GafAEc, GafASm and GafACb.

Triangles indicate residues in direct contact with the respective ligand in both GafASw and GafAEc. The stars designate the aromatic residues that form part of the binding site. The circle is the position of the Phe107 in GafASw thought to cause water exclusion and destabilisation of D-galactofuranose binding.

shown to actively accumulate radioactive labelled galactose (Mitchell, 1996). The accumulation was credited to an ATP-driven system rather than the glucose phosphotransferase system which galactose failed to bind (Mitchell, 1996). and a BlastP search detected a direct homolog with 50 % identity to the galactopyranose $MglB_{Ec}$ denoting the importance of D-galactose uptake to this bacterium.

Two residues are thought to cause abolishment of D-galactofuranose binding by GafASw; ie. the first is Phe17 which disrupts the salt bridge and provokes the movement of Asp88 to face and interact with the bound sugar. The second important residue is Phe107 thought to cause water exclusion and destabilisation of D-galactofuranose binding. The structure alignment at Figure 3. 34, investigates the conservation of these two residues in various GafAs. The Phe17 is also present in GafA from *Shewanella baltica* os195 and is also expected to be unable to bind D-galactofuranose like its close ortholog *ie*. GafASw. Interestingly, the GafAs from *Xylanimonas cellulosilytica* DSM 15894 (*ie*. Xcel_0419), *Variovorax paradoxus* S110 (*ie*. Vapar_1044) and *Herbaspirillum seropidicae* SmR1 (*ie*. Hsero_1033) have an asparangine residue instead of glutamine or phenynalanine at position 17. Also, they all feature a phenylalanine at position 107, similarly to GafASw and unlike GafAEc which has an asparangine. These observations point towards a potentially different binding specificity.

3. 4. 4 Biotechnological importance of GafAs

The two SBP proteins tested here were both robust in alkaline buffers and maintained full capacity to bind L-arabinose. Therefore, they are perfect candidates for engineering into bacteria to uptake the arabinose released from NaOH pretreated biomass. Furthermore, the importance of GafAs for such purpose can be assessed by using *E. coli* in competition assays of GafABCDEc with AraFGH in actively released arabinofuranose. Ectopic expression of the two systems in separate low copy plasmids will ensure inducible expression at similar levels. The arabinofuranose can be released in a mixture of arabinan with exogenously supplemented L-arabinofuranosidase. In

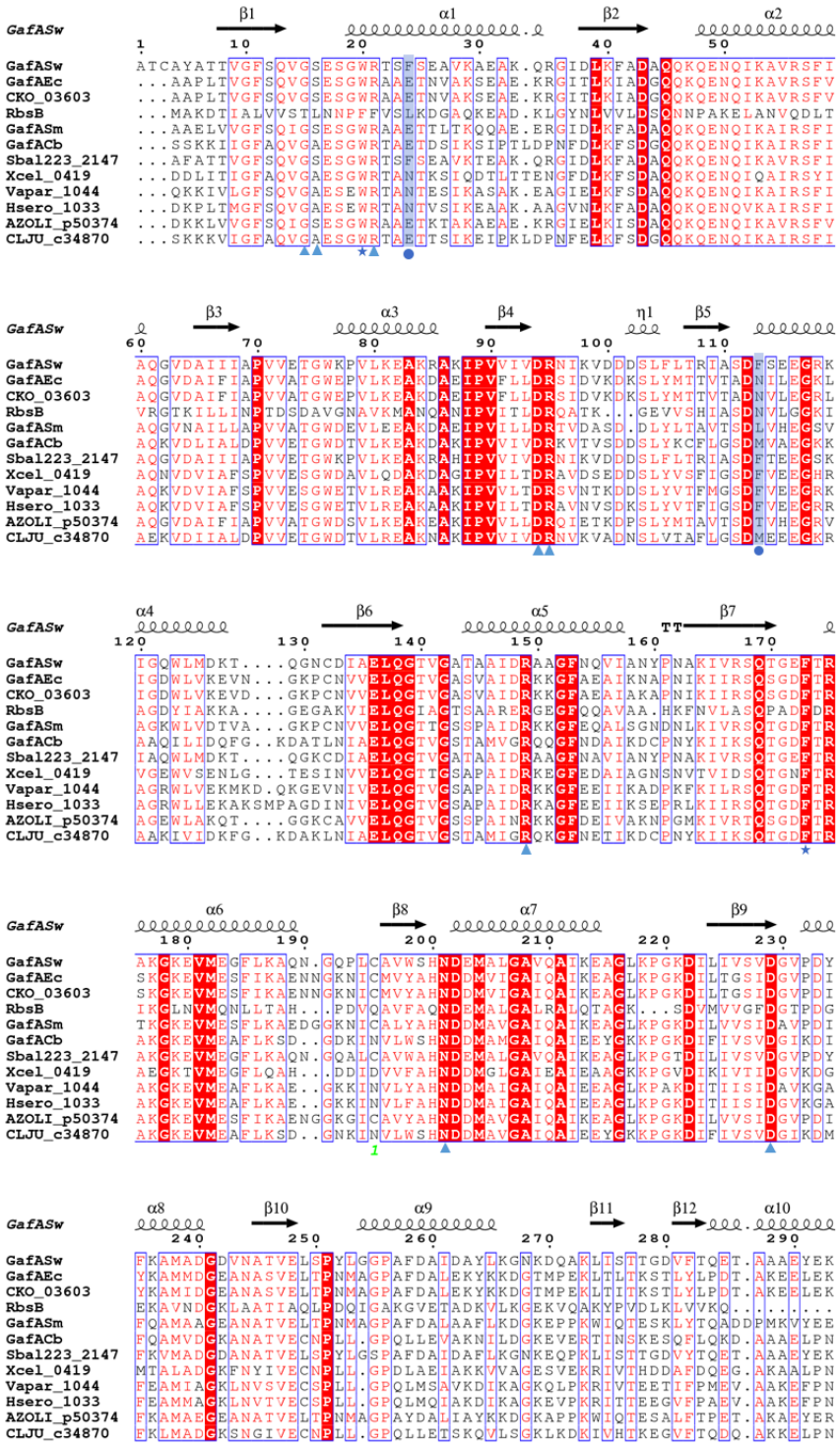


Figure 3. 34 Multiple structure-based alignment of the amino acid sequences of various GafAs.

Triangles indicate residues in direct contact with the respective ligand in both GafASw and GafAEc. The stars designate the aromatic residues that form part of the binding site. The circles (blue shade) designate the position of the Phe107 in GafASw thought to cause water exclusion and destabilisation of D-galactofuranose binding and Phe17 in GafASw which disrupts the salt bridge in the binding cavity.

such experimental set the *E. coli* strains expressing the two systems, should preferably be tagged with a fluorophore so their abundance is tractable with both a growth curve and a tag. Furthermore, the GafABCDSw can be expressed in *E. coli*, as *Shewanella* sp. ANA3 is a cognate γ-proteobacterium, and also assess its ability to compete with AraFGH for binding of L-arabinose derived from arabinan.

In a bioreactor setting, the abundance of sugars will be much higher as not only xylan will be present but cellulose and pectin-derived oligosaccharides as well. The galactose from pectin will compete with arabinose for binding to the GafAEc, however such issue will not be encountered if GafASw is used, either in *Shewanella* sp. ANA3 or engineered in the fermenting bacterium. This is the case for $AraF_{Ec}$ which is subjected to competitive inhibition by galactose with a *Ki*= 5. 5 x 10-4 mols/L (Hoggs, 1972). Similarly, AraABC arabinose transporter from pSymB of *S. meliloti* 1021 is also subjected to such inhibition as competitive assays using radiolabelled arabinose, showed that supplementation with galactose greatly reduced the accumulation of arabinose in *S. meliloti* 1021 (Geddes and Oresnik, 2012). Similar experiments can be performed with GafAEc and GafASw to demonstrate the inability of D-galactose to inhibit L-arabinose binding to GafASw.

Chapter 4

Creation of transport deletion (TD) strains in *E. coli* and their application in identification of an endogenous xylobiose transporter

This chapter focuses on the transport of the products of AGX enzymatic hydrolysis *ie*. D-glucuronic acid, L-arabinose, D-xylose and also xylo-oligomers produced during enzymatic hydrolysis of xylan. A prologue for xylo-oligomers and xylobiose transport is provided as the aims of this Chapter include identification of an endogenous xylobiose transporter from *E. coli*. Additionally, a brief introduction is provided to exemplify the use of engineered *E. coli* mutants with abolished transport for xylan-relevant sugars, the so-called Transport Deletion (TD) mutants. Initially, the efforts to create the TD strain for arabinose, *ie*. TDara, are described. Next, the experiments to test its functionality and robustness as a TD mutant by adding back *araE* MFS transporter, are explained. The same approach applied for the pre-defined TD mutants for D-xylose (TDxyl) and D-glucuronic acid (TDglcA) is described. Lastly, the importance of TD mutants in identifying transporters able to uptake xylan-derived sugars/sugar acids is established by the discovery of a cryptic gene in *E. coli* which encodes for a transporter able to uptake xylobiose.

4. 1 Introduction

4. 1. 1 Xylo-dextrins, transport and utilization.

The degradation of hemicellulose is an important part of the cycle of carbon and several organisms which reside near these polysaccharides have adopted in such niches by hydrolyzing and metabolizing hemicellulose (see General Introduction) (Shulami *et al*., 2007). The combined activity of endoxylanases and xylosidases on the

xylan backbone leads to release of a cocktail of soluble xylo-oligomers with varying length. These are transported inside the cell via membrane transporters; thus far, the identified ones belong to the ABC and the secondary transport systems. In 2012, Han *et al.* identified and described a cluster in the thermophile *Caldanaerobius polysaccharolyticus* which encodes for genes with putative xylanolytic activities (Han *et al.*, 2012). This cluster included an ABC transporter, a xylosidase, a two-component system and genes responsible for the metabolism of xylose liberated in the process. They purified the SBP and showed that it bound xylooligomers with the following binding preference based: xylotriose> xylobiose> xylotetraose. Shulami *et al*. (2011) described a xylo-oligomer-specific ABC transport system present in *Geobacillus stearothermophilus* and encoded by the operon *xynDCEFG*. Interestingly, this system is also responsive to regulation by XylR, similar to the xylose ABC transport system in *E. coli*. Additionally, Tsujibo *et al*. (2004), building on their discovery of an intracellular β-D-xylosidase (*ie*. BxlA) (Tsujibo *et al*., 2001) in *Streptomyces thermoviolaceus* OPC-520, attempted to characterise its genetic neighbouring operon, *bxlEFG*, present in a predicted xylo-oligosaccharide utilisation cluster (Tsujibo *et al*., 2004). This operon encoded for a predicted ABC transporter with an SBP protein (*ie*. BxlE) exhibiting a signature sequence which is unique to Class I carbohydrate SBPs. The recombinant BxlE bound xylobiose with the highest affinity (K_D = 8.75 x 10^{-9} M) followed by xylotriose (K_D = 8.42 x 10^{-8} M). The lowest affinity was measured for xylohexaose (K_D = 1.16 x 10^{-6} M). The calculated K_D values are similar to the ones SBPs exhibit for chitinobiose (Schlösser and Schrempf, 1996), cellobiose and cellotriose (Xiao *et al*., 2002) from cognate *Streptomyces* species.

The studies described above combined with the recent development of synthetic genetic constructs (Czar *et al*., 2009) have opened the opportunity to use these and other similar genes to engineer a bacterial strain able to uptake and ferment xylooligosaccharides to ethanol. Expressing and regulating these functionalities into *E. coli* can potentially enable its growth on xylan and xylodextrins.

4. 1. 2 Can *E. coli* grow on xylo-oligomers?

As already alluded to, *E. coli* can grow on monomeric xylose as the sole carbon source. The transport systems XylE from the MFS family and XylFGH from the ABC transport family import xylose into the cell, the first one with much lower affinity than the latter (Sumiya et al., 1995; Ahlem *et al*., 1982).

Despite the presence of xylose dedicated systems in *E. coli*, this bacterium cannot grow on xylo-oligomers as the sole carbon source (Qian *et al*., 2003; Rosanna Henessey, personal communication). Qian *et al*. (2003) attempted to change this by successfully introducing a plasmid expressing xylanolytic activities in *E. coli*. This plasmid included a novel xylooligosaccharide utilization operon from *Klebsiella oxytoca*, *ie*. *xynTB*. The gene *xynT* encodes a xylooligomer/cation symporter; and *xynB* encodes an intracellular β-xylosidase, which cleaves the linkage between the xylose monomers. The *E. coli* cells ectopically expressing $xynTB_{Ko}$ were assessed in their ability to utilise a mix of xylooligomers (*ie*. xylose to xylohexaose) using thin layer chromatography and densitometry. The results demonstrated that expression of these genes enabled utilisation of xylooligosaccharides with an efficiency inversely proportional to the oligomer length (Qian *et al*., 2003).

4. 1. 3 The need for xylo-oligomer transporters during fermentation

Cocktails of xylan degrading enzymes with negligible amounts of exo-xylanase or β-xylosidase are preferred as their activity limits the release xylo-oligosaccharides (Vázquez *et al*., 2002) and produces mainly xylose which is readily fermentable. Attempting to minimise the release of xylodextrins stems from the inhibitory activity they exert on cellulases which leads to lower glucose yields from cellulose (Selig *et al*., 2008; Qing *et al*., 2010; Zhang *et al*., 2012). Ultimately, one would exogenously supply β-xylosidase to lead to complete hydrolysis of the xylo-oligosaccharides into xylose and remove their negative effects, as suggested by Li *et al*., 2014. However, this would increase the cost of the process and therefore one might try to increase the production of the released xylanases and xylosidases by the fermenting bacterium. Though, the

high yield of such enzymes would speed the hydrolysis of xylo-oligosaccharides into monosaccharides, the excessive production could overwhelm the bacterial cells which are not natural overproducers of such enzymes. Another option to overcoming this issue, is by utilising the naturally occurring activities of the bacterial transport systems in the membrane. By identification of dedicated bacterial strains which produce transporters with high affinity for xylo-oligosaccharides, *eg*. xylobiose and xylotriose, would increase the rate of uptake of these sugars and their degradation in the bacterial cytoplasm. To circumvent the toxicity excess transporter monomers in the bacterial membrane can cause, these can be expressed ectopically in low copy bacterial vectors, *eg*. pWKS30 (Wang and Kushner, 1998).

4. 1. 4 Transport deletion (TD) mutants for identification of useful transporters

Metabolic engineering studies for xylan utilisation, similar to Qian *et al*., (2003) (see *Section 4. 1. 3*) have employed a strategy which considered more stages than just the extracellular or intracellular hydrolysis of the polysaccharide. The aforesaid group took into account the contribution of the transport systems in the uptake of xylobiose and xylodextrins, therefore emphasising their role in the efficient utilisation of xylan. A study conducted later to the above, followed a consolidated bioprocessing approach by modifying two *E. coli* strains to cooperatively accomplish enzyme production so that their engineered features led to higher efficacy in xylan hydrolysis and biofuel production (Shin *et al*., 2010). Despite the fact that the main focus is again the hydrolysis of the sugars, this group included the XynT transporter used by Qian and his co-workers, which potentially maximised the potential of the cell to uptake the released xylooligomers.

However, other studies which introduced chromosomally or ectopically expressed genes to enable *E. coli* growth on xylan, never included a transport system in their engineering efforts therefore neglecting the fact that the hydrolysed sugars need to cross the membrane barrier. For example, Bokinsky *et al*. (2011), engineered *E. coli* strain BW25113 to degrade xylan by introducing a plasmid which they named as

pXylan. The products of pXylan were a successfully secreted endoxylanase (Xyn10B) and a cytosolic β-xylosidase (Gly43F). The plasmid constitutively expressed the functionalities without causing toxicity to the cells (Bokinsky *et al*., 2011). The strain bearing the plasmid gained the ability to grow on birchwood xylan and switchgrass pretreated with ionic liquid (IL) (Bokinsky *et al*., 2011). They reported that the Xyn10B was able to hydrolyse pretreated switchgrass and mainly release xylotriose and xylobiose (*ie*. trisaccharide and disaccharide of xylose, respectively) (Bokinsky *et al*., 2011). However, since *E. coli* is not thought to transport xylobiose and xylotriose (Qian *et al*., 2003), it is unknown how the strain imported these xylooligomers. One can assume that mutation of endogenous xylose transporters increased their promiscuity and made them capable to import xylosides, which might explain the long lag phase, reaching 20 hours at times (Bokinsky *et al*., 2011). Also, IL pretreatment has been shown to release xylose in its monomeric form which could account for fuelling the initial growth stages during lag phase (Sun *et al*., 2013; Socha *et al*., 2014). Although, the strain gained the ability to grow on IL-pretreated switchgrass, the true capabilities of pXylan weren't fully assessed as the strain was not optimised for transport.

Useful tools have been created for studying and optimising the activity of glycolytic hydrolases. Alvira *et al.* (2010) have reported the use of enzymatic microassays run in 2 ml eppendorf tubes for rapid assessment of wheat straw hydrolysis. A different study conducted by Wolfrum and coworkers, has created a laboratory-scale assay which allowed multiple samples from various feedstocks to be screened simultaneously following a small-scale enzymatic hydrolysis step (Wolfrum, Ness and Scarlata 2013). The lack of interest on engineering of transport systems for biofuel production, means that not many tools are available for quick and reliable assessment of the bacterial transport capabilities. This could explain why some studies are discouraged from optimising bacterial transport systems.

The current study is building up on previous work carried out in the Thomas lab to create *E. coli* growth deficient strains by deletion of targeted transport systems. These strains, namely **Transport Deletion (TD) mutants**, will ideally show reduced or

abolished growth as compared to the WT when grown in minimal M9 media supplemented with the respective xylan-derived sugar or sugar acid. Previous Thomas lab workers have created TD strains for D-xylose (*ie*. TDxyl) and D-glucuronic acid (*ie*. TDglcA). The features of these strains are summarised in Table 4. 1. Functional redundancy conferred by other native transport systems is common (Desai and Rao, 2010; Hasoma *et al*., 2004) and therefore deletion of targeted systems often doesn't lead to complete growth abolishment. For example, in the case of TDxyl, the *araH* (*ie*. membrane subunit of the L-arabinopyranose) was also deleted as the double mutant Δ*xylExylH* was still able to grow on D-xylose due to the promiscuous activity of the AraFGH system (Henrique Neves, personal communication). The TD mutants are considered useful tools for identification of exogenous transporters when the deletion of the endogenous transport systems is sufficient to cause a near-abolishment growth. The TDxyl has already been used successfully in the Thomas lab to identify exogenous MFS transporters for xylooligomers (Rosanna Henessey, personal communication) by screening for restoration of growth when the respective transporter genes were ectopically expressed in the TDxyl strain.

Table 4. 1. **Transport deletion (TD) mutants created in GHT lab**

TD name	Abbreviation	*E. coli* strain	Genotype	Complementation attempt with:
TDglucuronate	TDglcA	BW25113	Δ*exuT*	pWKS30 (*exuT*)
TDxylose	TDxyl	BW25113	Δ*xylE* Δ*xylH* *araH*::kan^{R}	pWKS30 (*xylE*)

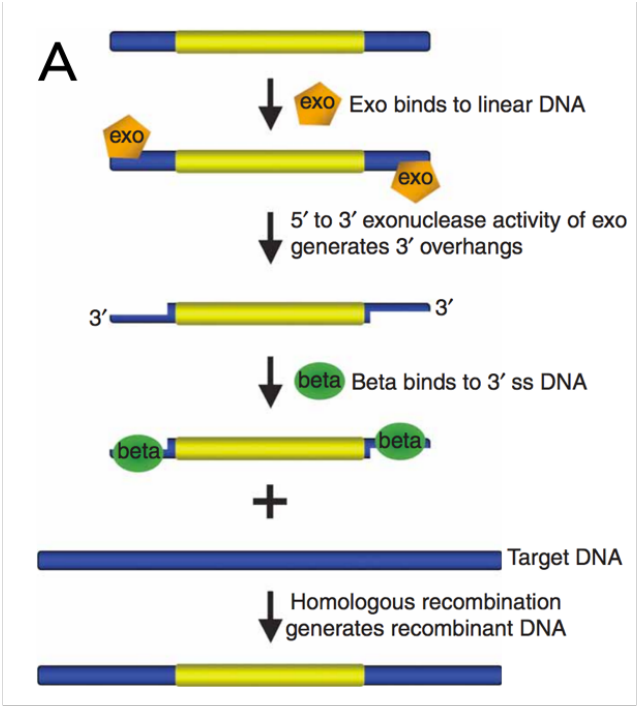
4. 1. 5 Lambda red recombineering and Flp recombinase

Genetic recombination is a process by which a molecule of nucleic acid is broken and then joined to a different DNA molecule. The λRed system is a tool commonly used by molecular biologists for cloning or genome engineering. The technique is based on homologous recombination and allows for direct modification of *E. coli* DNA, without the requirement for restriction sites (Daiguan et al., 2000). The naturally occurring recombination system of *E. coli* involves the enzymes RecABCD (Kowalczykowski *et*

al., 1994); even though the λRed system is not produced by *E. coli* but derived from the λ bacteriophage there is interplay between the two systems during transfection. The technique is also known as recombineering, as it is a **recomb**ination-mediated genetic **engineering** method.

Murphy *et al*. (1991), was the first to test the red system in *E. coli* and showed that it was able to promote recombination between the bacterial chromosome and linear dsDNA molecules which entered the cell via electroporation (Murphy *et al*., 1991). The system was proven to be far more efficient than other systems used at the time for making gene replacements as PCR generated species could be used in the reaction (Murphy *et al*., 1991). Recombineering plasmids (*eg*. pKD46 and pSIM8) carry the genes which encode the constituents of the red system *ie*. *exo*, *beta* and *gam*. Exo has a 5'- to 3'-dsDNA exonuclease activity, which can generate 3'-overhangs on linear DNA (Figure 4. 1A). Beta recognises and binds the single-stranded DNA (3'-overhangs) and promotes single stranded annealing which finally generates the recombinant DNA (Figure 4. 1A). The Gam, 'guards' the whole process as it prevents RecBCD nuclease from degrading the double-stranded linear DNA fragments (Sharan *et al*., 2009).

For gene replacements or deletions, usually a cassette encoding for a drug-resistance gene is amplified by PCR which incorporates at least 50 bases of flanking homology to the gene to be deleted (Figure 4. 1B). The double stranded DNA (dsDNA) amplified during PCR is electroporated into cells expressing the λRed genes. These genes are normally carried on plasmids, *eg*. pKD46 (Datsenko and Warren, 2000) and pSIM18 (Datta, Constantino and Court, 2006), which are allow production of the λRed system upon induction *eg*. with L-arabinose or caused by temperature shift. The recombinants are selected by growth on respective antibiotic. Following confirmation of the deletion, the antibiotic resistance cassette is typically removed, a process performed by the Flp recombinase or flippase introduced on a second plasmid (*ie*. pCP20).



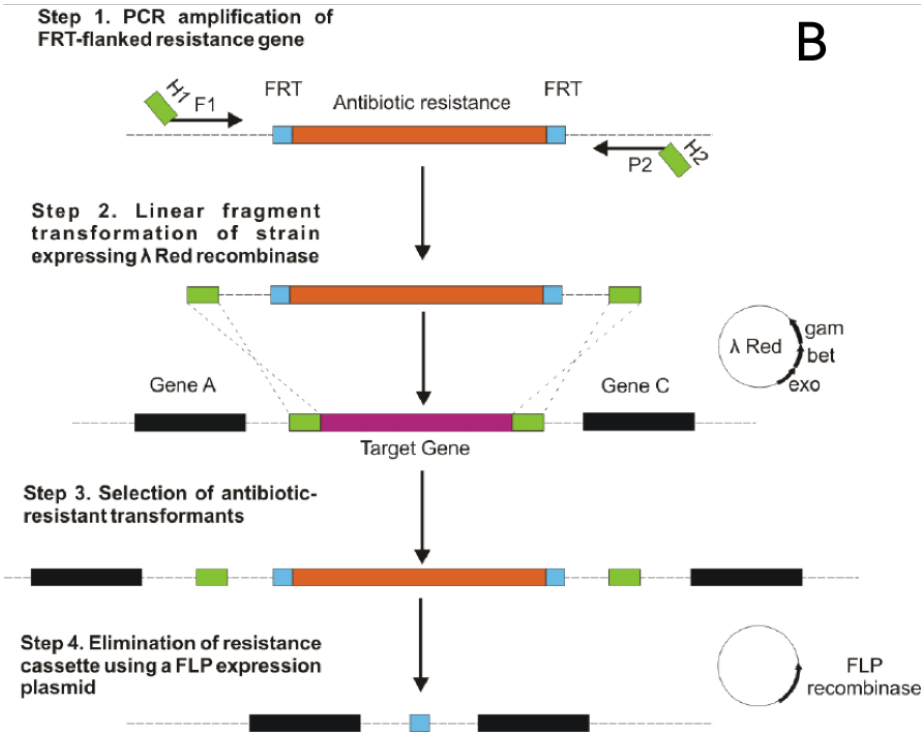


Figure 4. 1 Recombineering using the λRed system

A) Overview of bacteriophage λ recombination. The Exo enzyme is shown as an orange pentagon and beta. Gam, which is not shown here, is 'guarding' the process by preventing RecBCD system from degrading the dsDNA. Figure taken from Sharan *et al.*, 2009.

B) Process involved in deletion of genes using recombeeniring. Figure taken from Levarski *et al.*, (2011).

The Flp recognises both FRT sites which flank the inserted antibiotic sequence and catalyses recombination between the two sites causing removal of the cassette (Schlake and Bode, 1994) (Figure 4. 1B).

4. 1. 6 Aims of the current chapter

Here, we create a TD strain for arabinose (*ie*. **TDara**) by deletion of *araE* encoding the arabinopyranose MFS transporter and *araH,* of the AraFGH ABC system. We attempt to functionally complement and 'validate' all three TD strains (*ie*. **TDara, TDxyl and TDglcA**) by re-introducing ectopic expression of their MFS transporters. To demonstrate the application of the produced TD strains, we use the TDxyl strain to phenotype the function of a putative MFS transporter from *E. coli* potentially involved in the xylo-oligomer import, *ie*. YagG.

4.2 Materials and Methods

4. 2. 1 Bacterial strains, plasmids and growth media

All the strains and plasmids used in the study are listed in Table 4. 2 below.

Table 4. 2. Bacterial strains and plasmids used in the study

Strains	Genotype	Source
MG1655	K-12 F^- λ^- *ilvG*$^-$ *rfb-50 rph-1*	Blattner *et al.* (1997)
BW25113	F^-, DE(*araD-araB*)567, *lacZ*4787(del)::*rrnB*-3, LAM^-, *rph*-1, DE(*rhaD-rhaB*)568, *hsdR*514	Datsenko & Wanner (2000)
BW25113 Δ*araE*	BW25113 deletion mutant of *araE* (cured for kanamycin)	This study
BW25113 Δ*araH*	BW25113 deletion mutant of *araH* (cured for kanamycin)	This study
BW25113 Δ*araHaraE* (*ie.TDara*)	BW25113 deletion mutant of *araH* and *araE* (cured for kanamycin)	This study
BW25113 Δ*exuT* (*ie.*TDglcA)	BW25113 deletion mutant of *exuT* (cured for kanamycin)	GHT lab
BW25113 Δ*xylHxylEaraH* (*ie.* TDxyl)	BW25113 deletion mutant of *xylH*, *xylE* and *araH* (cured for kanamycin)	GHT lab
TDxyl Δ*arsB*::$xynB_{Kp}$	TDxyl with chromosomal insertion of xynB xylosidase from *Klebsiella pneumoniae*	GHT lab
Plasmids	**Description**	**Source**
pWKS30	Low copy number Amp^R cloning vector – IPTG inducible	Wang *et al.*, 1991
pBADnLIC2005	pBAD vector (high copy) modified with LIC cassette, Amp^R - Arabinose inducible	Dr. Eric Geertsma
pCD07	pWKS30 expressing *xylE* from *E. coli*	This study
pCD08	pWKS30 expressing *araE* from *E. coli*	This study
pCD09	pBADcLIC2005 expressing *araE* from *E. coli*	This study
pCD10	pWKS30 expressing *xylE* from *E. coli*	This study
pKD46	Tm sensitive plasmid used for recombineering (carries lambda Red genes: exo, bet and gam) – Arabinose inducible	Datsenko & Wanner (2000)

LB media used for the overnight inoculation of the strains was prepared as described in *Section 2. 2. 2. 1* of Chapter 2. The minimal M9 media used in the microplate-based growth assays was supplied from VWR International (*ie*. M9 medium broth powder biotech grade). 1 mM total concentration of $MgSO_4$ was added in the autoclaved M9 media prior to the growth assays. All the sugars used in the growth assays (*ie*. D-xylose,

L-arabinose and D-glucuronic acid) were supplied from Sigma-Aldrich, and xylobiose from Tokyo Chemistry Industry UK Ltd.

4. 2. 2 Molecular Biology techniques

All the techniques used in the study for this Chapter are described in Chapter 2 at the following sections: Plasmid DNA extraction (*2. 2. 3. 1*), Agarose gel electrophoresis (*2. 2. 3. 3*), PCR (*2. 2. 3. 5*), PCR clean up (*2. 2. 3. 6*), Gel extraction (*2. 2. 3. 7*), Conventional ligation of constructs into the low copy vector pWKS30 (*2. 2. 3. 9*), LIC independent cloning of constructs into pBADcLIC (2. 2. 3. 11), Restriction (*2. 2. 3. 13*) and heat-shock transformation of *E. coli* (*2. 2. 3. 12*). The pWKS30 plasmid was digested with XhoI and EcoRI for cloning of *araE* and *exuT,* and with EcoRI and BamHI for cloning of *xylE*. The cloning was confirmed by M13 PCR screen and DNA sequencing.

The list of primers used for the amplification of the MFS transporter sequences from the genomic DNA of *E. coli* are shown in Table 4. 2.

Table 4. 3. Primers used in the work of this chapter

Name of Primer	Sequence of primer (5'→3')
ExuTF	CCGCTCGAGG*TAA*AAG**AAGGAGA**TATACTCATGGCAACGTTCGGGG
ExuTR	CGGAATTCTTAATGTTGCGGTGCGGG
XylEF	CGGAATTCGTAAAAG**AAGGAGA**TATACATATGTTACATCAATA
XylER	AAGGATCCGGACAGGAAGATTACAGC
AraEF	CCGCTCGAGG*TAA*AAG**AAGGAGA**TATACTCATGGTTACTATCAATAC
AraER	CGGAATTCTCAGACGCCGATATTTCT
AraEpBADF	ATGGGTGGTGGATTTGCTATGGTTACTATCAATACGGAATCT
AraEpBADR	TTGGAAGTATAAATTTTCGACGCCGATATTTCTCAACTTCTCGCCTGC
AraEF (KO)	TTTTATGACCCTGCCGCATG
AraER (KO)	GGTTTTATCCCATTCCCCGC
AraHF (KO)	AATGCCGATCACCACATTCG
AraHR (KO)	CGTCAAGCTCCACCATTGAG

The emphasised sequences correspond to synthetic functional features added in the respective amplicons for cloning into the pWKS30 vector. The underlined sequences indicate restriction enzyme sites, italics are for the stop codons and bold font designates added Shine Dalgarno sequences.

4. 2. 2. 1 Electroporation of dsDNA into E. coli cells

An overnight culture was diluted one hundred times in fresh LB and grown to an OD_{600} ranging between 0.4 and 0.6. The cells were placed on ice for 20 mins and the all of the next steps took place on ice. The cells were centrifuged (4500 rpm for 10 mins) and the pellet was washed with 10% glycerol. The washing procedure was repeated 3 times. The cells were resuspended in 10% glycerol solution and distributed in 75 µl aliquots. 1 to 3 µl of dsDNA (*ie*. 200 – 500 total ng) was added to the aliquot prior to electroporation. The aliquot was transferred to a 0.2 cm gap eletroporation cuvette and was subjected to an electrical field of 12 kV/cm, 25 µF and 200 Ω for 5 milliseconds using the electroporator MicroPulser (Bio-Rad®). Immediately after the electrical pulse, the cells were mixed with 950µl of pre- warmed LB or SOC medium. The culture was incubated at 37 °C for an hour of outgrowth and then plated in LB plates with the appropriate antibiotic.

4. 2. 3 λRed-mediated gene disruption

Single gene mutations were inserted in strains using the λ-recombinase system (Datsenko and Wanner, 2000). The primers araEF and araER (KO) were used to amplify the kan^R cassette from the respective Keio collection mutant (*ie*. *E. coli* **BW25113** Δ*araE*). The PCR incorporated extra 300 bps upstream and downstream of the amplicon homologous to the start and end of *araE* gene, respectively. The same process was followed for the amplification of *araH::kan*^{R} cassette. The sequence of *E. coli* MG1655 strain is identical to BW25113 as far as these two sites are concerned.

The pKD46 plasmid (see *Section 4. 1. 5*) was transformed into *E. coli* MG1655 WT strain. A single transformant was inoculated into LB and grown to exponential phase at 30 °C, to maintain plasmid propagation, in the presence of arabinose for the induction of *exo*, *gam* and *beta* genes. As the MG1655 strain contains the catabolic genes for L-arabinose, as opposed to BW25113, a higher concentration of inducer was included, *ie*. 20 mM. The growth was stopped during the exponential phase at growth OD_{600} = 0.4

to 0.6. Electroporation competent cells were prepared which were electroporated with the respective dsDNA containing the *kan*^{R} insertions, as described at *Section 4. 2. 2. 1.* Following outgrowth the culture was plated on LB plates containing 30 µg/ ml kanamycin. The colonies obtained were replated on *kan*^{R} plates to confirm the phenotype and subsequently made chemically competent as described in *Section 2. 2. 3. 12.*

4. 2. 3. 1 Curing of kan^R *cassette using pCP20*

The chemically competent Δ(*gene of interest*)::*kan*^{R} strains were transformed with pCP20 as described in *Section 2. 2. 3. 12.* Single transformants were grown in LB at 42 °C overnight to allow expression of Flp recombinase. A loopful of culture was plated in LB plates without selection, and grown overnight at 37 °C to ensure complete removal of pCP20. Up to 25 colonies from each deletion genotype were tested for loss of *kan*^{R} cassette and pCP20 by re-plating individual colonies at separate plates containing 30 µg/ ml kanamycin, 100 µg/ ml ampicillin and no antibiotic. The curing of *kan*^{R} was also verified by PCR screening using the same primers as in *Section 4. 2. 3.*

4. 2. 4 Microplate-based growth assays in minimal M9 media

Prior to each growth assay 3 to 5 colonies (ie. biological replicates) for each genotype were inoculated for overnight growth into 1 ml of M9 media supplemented with glucose (20 mM), $MgSO_4$ (1 mM) and ampicillin (100 µg/ml) when required to maintain and propagate the plasmid constructs. Following overnight incubation, the cultures were diluted 10 times and their cell density (OD600) was measured using a spectrophotometer. To ensure initial cell density was equal for all cultures and any changes observed during the growth assay are attributed to the growth assay parameters, the cultures were standardized to an OD650 of 0.25 using fresh M9 media. For each microplate growth assay, a 96-well Corning Costar® plate was prepared by loading 144 µl of M9 media in each well. The media was supplemented with 1 Mm IPTG, 10 or 20 mM of sugar/ sugar acid, 1 mM $MgSO_4$ and 100 µg/ ml ampicillin where required. 6 µl of culture was added in each well for a starting OD650 of approximately

0.01. Blank controls of the M9 media lacking cultures were included to standardise the results during data analysis. Positive controls for the TD assays were the WT strains and in some assays glucose was included. The readings were taken in the Infinite 200 PRO plate reader from Tecan®.

Data analysis was performed using Microsoft Excel for all assays. The plotted data for each genotype, at each timepoint, are the average of 3 (or 5) biological replicates. The average for the blank controls was calculated, and substracted from each timepoint of the samples. The standard errors were calculated as the square root of the standard deviation divided by the number of biological replicates (3 or 5). The graphs presented in the Chapter were produced in GraphPad Prism 7.0, whereas the ones in Appendix were created in SigmaPlot.

4. 3 Results

4. 3. 1 Creation and functional complementation of TD mutants

4. 3. 1. 1 Ectopic expression of exuT restores growth of TDglcA

The TDglcA, is an *E. coli* BW25113 strain with the gene encoding for the hexuronate MFS transporter, *exuT*, deleted. The mutant strain was produced and tested by co-workers in GHT lab, who confirmed that TDglcA shows complete abolishment of growth on minimal media supplemented with D-glucuronic acid (data not shown). Here, the applicability of this strain as a TD mutant was tested by functional complementation with re-introduction of *exuT in trans*.

The *exuT* gene was cloned into the low-copy vector, pWKS30 as described in *Section 2. 2. 3. 9*. TDglcA pWKS30 (*exuT*) was tested for recovery of growth in M9 media supplemented with 10 mM or 20 mM D-glucuronic acid as the carbon and energy source, in a microplate-based assay. The pWKS series of vectors have been created especially for overproduction of proteins which are toxic in *E. coli* in high copy number (Wang and Kushner, 1991). The expression of *exuT*, which is under the control of the *lac* promoter, was induced by IPTG during the incubation course. The growth of the culture was monitored via absorbance measurements every 30 minutes, over a total timespan of 24 hours. The first assay performed, showed that TDglcA pWKS30 (*exuT*) presented partial complementation of phenotype at 20 mM D-glucuronic acid (Figure 4. 2). Therefore, in the second growth assay the above strains were re-tested and also included TDglcA pWKS30 (*exuT*) without addition of IPTG to assess the growth with leaky expression of *exuT*.

The deletion of *exuT* in TDglcA cells completely abolished growth during the course of incubation, confirming the results from previous Thomas lab workers and affirming its applicability as a TDglcA at 10 mM D-glucuronic acid. Also, the growth experiment

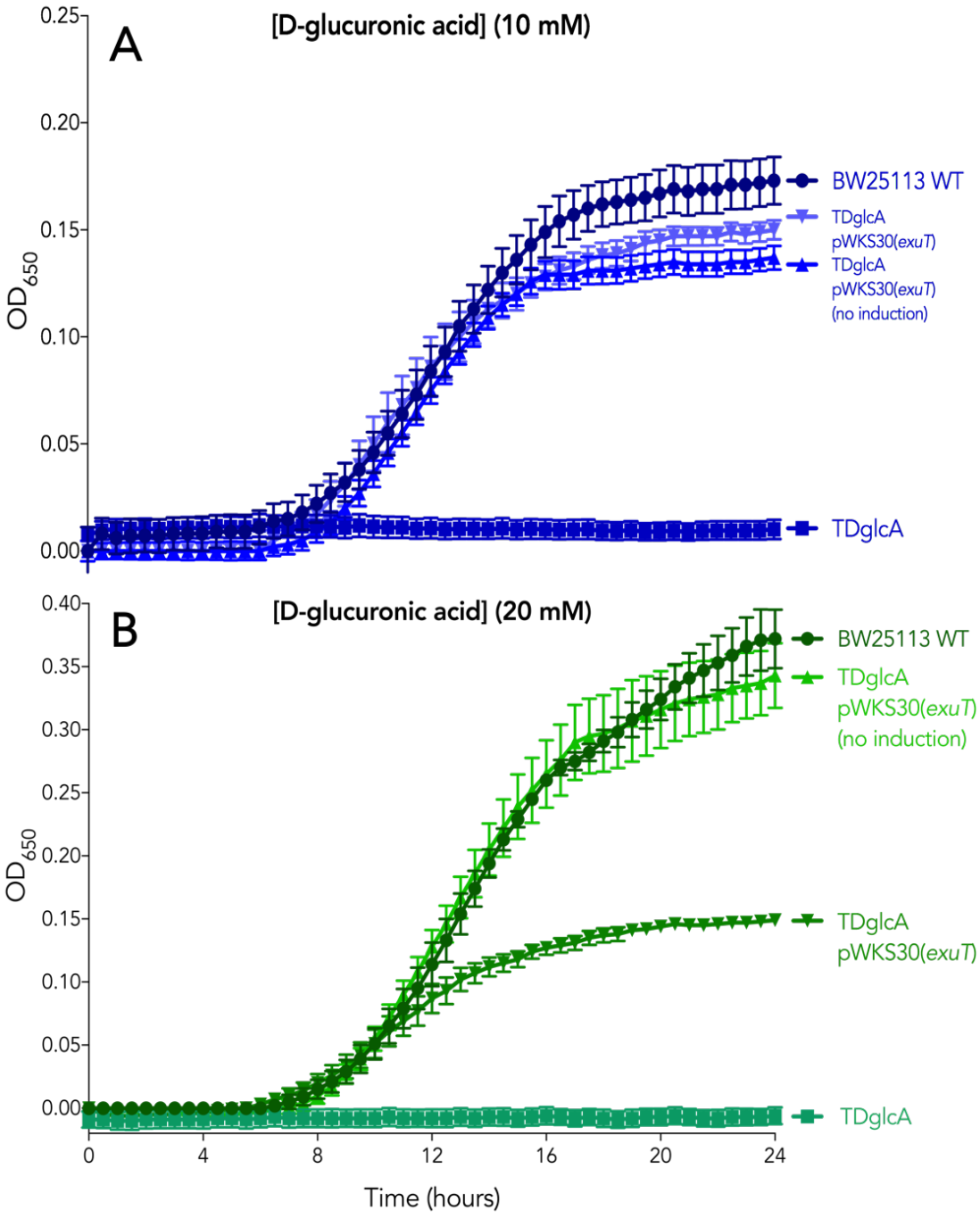


Figure 4. 2 Un-induced ectopic expression of exuT restores growth of TDglcA
A) Growth curve of TDglcA (*ie*. *E. coli* BW25113 Δ*exuT*) with *exuT* expressed *in trans,* in minimal media supplemented with 10 mM D-glucuronic acid. The *exuT* expression was induced using 1 mM IPTG (*nabla*) or kept un-induced (*triangles*). The WT (*circles*) and TDglcA (*squares*) are also shown for comparison. The cell density (OD_{650}) measurements were taken every 30 minutes across a 24-hours timecourse. Each timepoint is the average OD_{650} from 5 biological replicates and the error bars are standard errors calculated from the standard deviation.
B) Same growth assay with (A) in presence of 20 mM D-glucuronic acid.

suggested that the uninduced culture presented leaky expression of *exuT*, which was sufficient to restore the growth near WT-like levels at both 10 and 20 mM (Figure 4. 2A, B). As expected the culture supplemented with higher concentration of glucuronic acid grew to a higher biomass by the end of the incubation course (Figure 4. 2B). Conversely, the induced culture showed reduced growth at 20 mM concentration of substrate compared to the uninduced TDglcA pWKS30 (exuT) and the BW25113 WT (Figure 4. 2B). As the latter phenotype was not observed in the presence of 10 mM glucuronic acid, it could potentially be attributed to the high influx of the D-glucuronic acid and due to its acidic nature, it could have caused the intracellular pH to drop thus leading to cell lysis. Similar observations were made before with the accumulation of metabolites causing depression of growth such in the case of gluconic acid and lactic acid build up halting the growth of *Gluconacetobacter xylinus* (Liu *et al*., 2016).

4. 3. 1. 2 Deletion of araE and araH abolishes growth of E. coli on L-arabinose

The TDara strain was constructed by deletion of genes encoding for the characterised L-arabinose transport systems of *E. coli*, *ie* AraE (Daruwalla, Paxton and Henderson, 1981) and AraH from the AraFGH ABC transport system. The parental strain genetically modified here was the *E. coli* MG1655 as the BW25113 used to create TDglcA and TDxyl is deficient in the *araBAD* catabolic genes for arabinose. This BW25113 strain is normally used for induction of the λRed system which is under the control of the pBAD promoter (Datsenko and Wanner, 2000) and therefore retaining the catabolism of arabinose would deplete its inducing activities. The *araH* and *araE* were successively deleted in the order mentioned, using the λRed recombineering system as described in *Section 4. 2. 3*. The single mutants and the double mutant (Δ*araHaraE*) created, were tested for growth defects on a microplate growth assay including 20 mM L-arabinose as described in *Section 4. 2. 4*.

Deletion of *araE* extended the lag phase by 6 hours as compared to the WT (Figure 4. 3). This is in accordance with the study by Hanosa *et al*. (2004) which showed that the

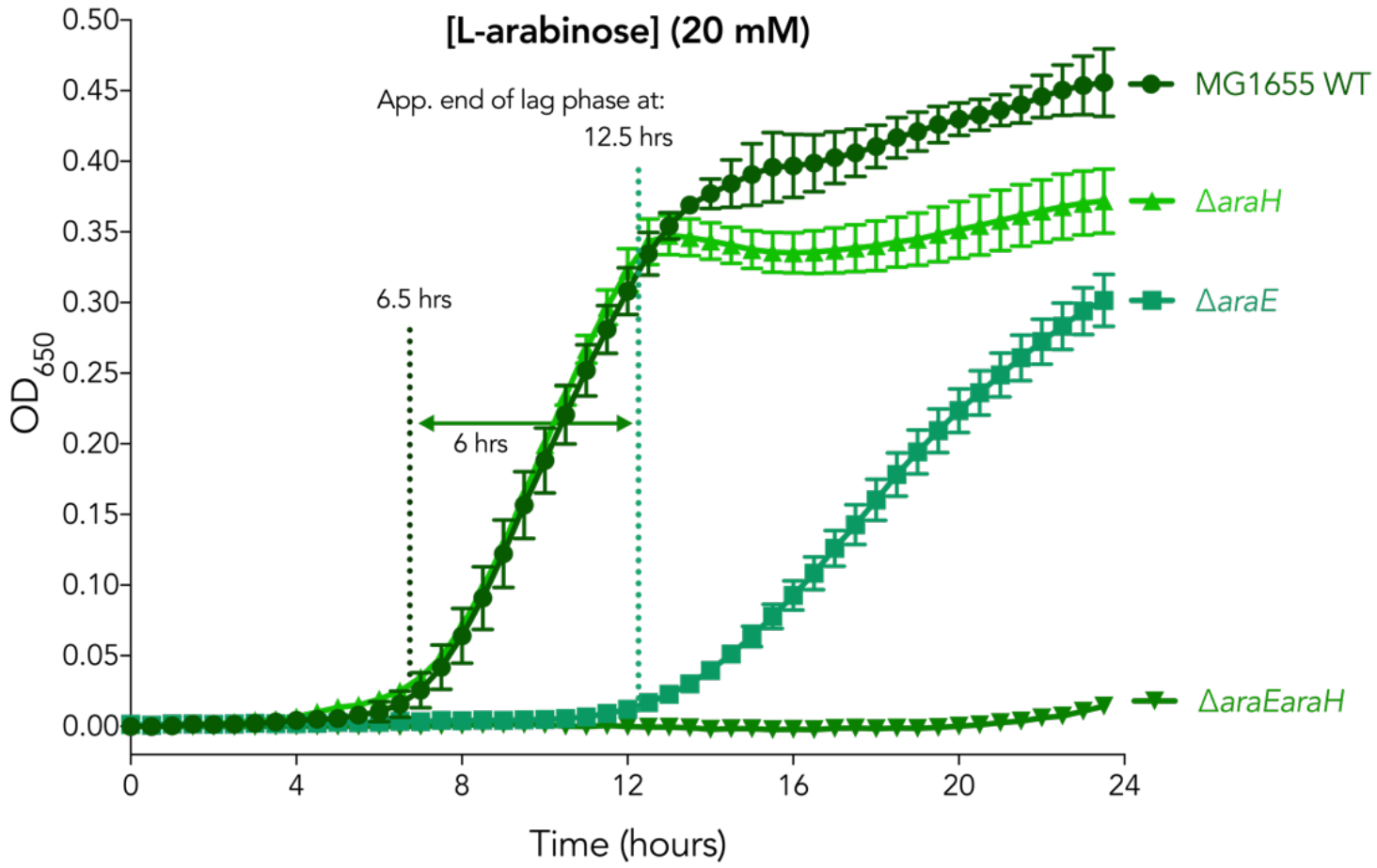


Figure 4. 3 E coli MG1655 ΔaraEaraH is a TDara

Growth curves of *E. coli* MG1655 single and double mutants for arabinose transport systems. The strains assessed for defective growth were Δ*araH* (*triangles*), Δ*araE* (*squares*) and Δ*araEaraH* (*triangles*). The WT strain (*circles*) is included for comparison. The cell density (OD_{650}) measurements were taken every 30 minutes across a 24-hours timecourse. Each timepoint is the average OD_{650} from 3 biological replicates and the error bars are standard errors calculated from the standard deviation. The starting concentration of L-arabinose in each well was 20 mM.

single mutant did not exit lag phase not until after 12 hours. The growth assay was distinctly different from ours, as this group assessed the Δ*araE* mutant of *E. coli* W3110 at a pH-stat under argon gas phase to ensure anaerobic conditions. Also, the minimal media used was enriched with more mineral salts than M9 and the strain was supplied with 3 times higher arabinose concentration (Hasona *et al.*, 2004). Despite the discrepancies in the set-up of the assay, the Δ*araE* mutant tested here also exhibited a 12-hours lag phase.

Growth of the Δ*araH* is less defective compared to Δ*araE*. The strain enters exponential phase at the same time with the WT (Figure 4. 3). However, the WT strain reached 1.2 times higher OD_{650} compared to Δ*araH*. Similar to Δ*araE,* this is consistent with the study conducted by Hanosa and coworkers (Hanosa *et al.*, 2004).

In the case of the double mutant, the deletion of both transport systems is sufficient to cause a strong defective phenotype on growth in L-arabinose. MG1655 Δ*araEaraH* strain only showed residual growth following 21 hours of incubation which could be partially attributed to facilitated diffusion (*ie*. spontaneous passive transport of solute via transmembrane integral proteins) of the sugar (Khankal *et al.*, 2008). Facilitated diffusion was shown to occur in LacY mutants which lacked both aforementioned transport systems (Morgan-Kiss, Wadler and Cronan, 2002; Goswitz and Brooker, 1993). Arabinose was able to diffuse homogeneously across the membrane of such mutants and induce the expression of a GFP reporter which was under the control of pBAD (Morgan-Kiss, Wadler and Cronan, 2002). Deletion of the endogenous transport systems led to the creation of **TDara**, which was tested for its applicability as a TD strain by functional complementation with *araE*.

4. 3. 1. 3 Ectopic expression of araE from pBADcLIC partially restores growth of TDara on L-arabinose.

The next experiment was performed to confirm the strain is able to regain ability to grow on arabinose, therefore verifying its applicability in testing exogenous transporters for reversing transport deficient phenotypes. We attempted this by re-

introducing the *araE* gene *in trans* in TDara, similarly to the strategy followed for TDglcA (*Section 4. 3. 1. 1*). The *araE* was cloned into pBADcLIC vector as described in *Section 2. 2. 3. 11*. The choice of this vector is not accidental, as the expression of the cloned constructs is under the promoter present in pBAD vector which is derived from the upstream region of *araBAD* catabolic genes of *E. coli* (Dunn *et al*., 1984). Therefore, this places the *araE* expression under the control of L-arabinose. In a bioindustrially scaled-up culture supplied with xylan-derived sugars the abundancy of L-arabinose would eliminate the need for an exogenously supplied inducer if pBAD-regulatable plasmids are used.

TDara was transformed with pBAD (*araE*) and was grown in a microplate assay as described in *Section 4. 2. 4*. To allow for statistical significance 3 different biological replicates were included for each genotype. The results are shown in Figure 4. 4.
In both concentrations tested (*ie*. 10 and 20 mM arabinose), pBAD(*araE*) only partially complements the TDara phenotype (Figure 4. 4). The final OD_{650} of the strain appears 2.5 times decreased than the WT, with a shallower log phase which is much longer in higher concentration of the sugar, *ie*. 20 mM (Figure 4. 4). This could be caused by toxicity of excess transporter units in the membrane (Laible *et al*., 2004); or it could potentially be attributed to insufficient transporter units in the membrane to allow for restoration back to WT levels. As the pBADcLIC is a high copy number plasmid is likely to be the outcome of the first. This is exemplified by the AraE overproduction in the single Δ*araH* mutant (Figure A4. 1B). The mutant grows similar to the WT strain (see *Section 4. 3. 1. 2*), however the presence of *araE* under the control of pBAD causes a growth defect at both 10 and 20 mM L-arabinose (Figure A4. 1B, A4. 2B). Despite the apparent toxicity in Δ*araH* mutant, the Δ*araE* pBAD(*araE*) presents shorter lag phase than Δ*araE,* but the latter reaches almost doubled cell density compared to the first by the end of the incubation when grown at 10 mM L-arabinose (Figure A4. 1A). Though, the shortened lag phase persisted when grown at 20 mM, both strains grew to the same OD_{650} by the end of the run (Figure A4. 2A). This could indicate at which growth stages each system is more important. It is apparent that AraE, as a low affinity but high capacity system, is participating in the early stages of growth as its deletion

extends the lag phase (Figure A4. 1A, Figure A4. 2A). At the start of the incubation, the carbon source is still present in abundance and the cells haven't had the chance to accumulate enough ATP; therefore, in order to allow for fast and energy inexpensive utilisation of L-arabinose the cells are likely to employ the AraE system. However, at the later stages of growth, during the transition from explonential phase to stationary phase, the L-arabinose is expected to be present at much lower concentrations due to consumption by the strain and therefore at such stages the high affinity transporter *ie*. AraFGH is expected to be actively transporting the sugar. Indeed, this is evident by the decreased OD_{650} during the stationary phase and at the end of the incubation course as compared to WT (Figure A4. 1B, Figure A4. 2B)

4. 3. 4 Ectopic expression of araE from pWKS30 partially restores growth of TDara on L-arabinose.

The *araE* was cloned into the pWKS30 vector to circumvent potential toxicity caused by the high copy number of pBADcLIC vector, which would explain the reduced complementation levels of TDara. The cloning happened as described in the Materials and Methods *Section 2. 2. 3. 9*. The plasmid was transformed into the aforementioned strains which were tested for complementation by a microplate assay as described in the *Sections* above.

The ectopic expression of araE from pWKS30 partially recovered the growth of TDara in both 10 mM and 20 mM (Figure 4. 5). However, in 10 mM arabinose the TDara pWKS30(*araE*) strain grew better than the respective one with pBAD(*araE*) (Figure 4. 4), when both are compared to the WT profiles. This could be because of the use of a low copy vector which would alleviate any toxicity caused by excess AraE units in the membrane. The pWKS30(*araE*) was not a burden to the single mutants as opposed to pBAD(*araE*) further endorsing the possibility that the AraE lower production is not toxic to the cells anymore. In fact, it successfully recovered the growth of the Δ*araE* strain (Figure A4. 3). However, this could also be attributed to the fact that the high concentration of the inducer used in this experiment is enough to overcome the "all

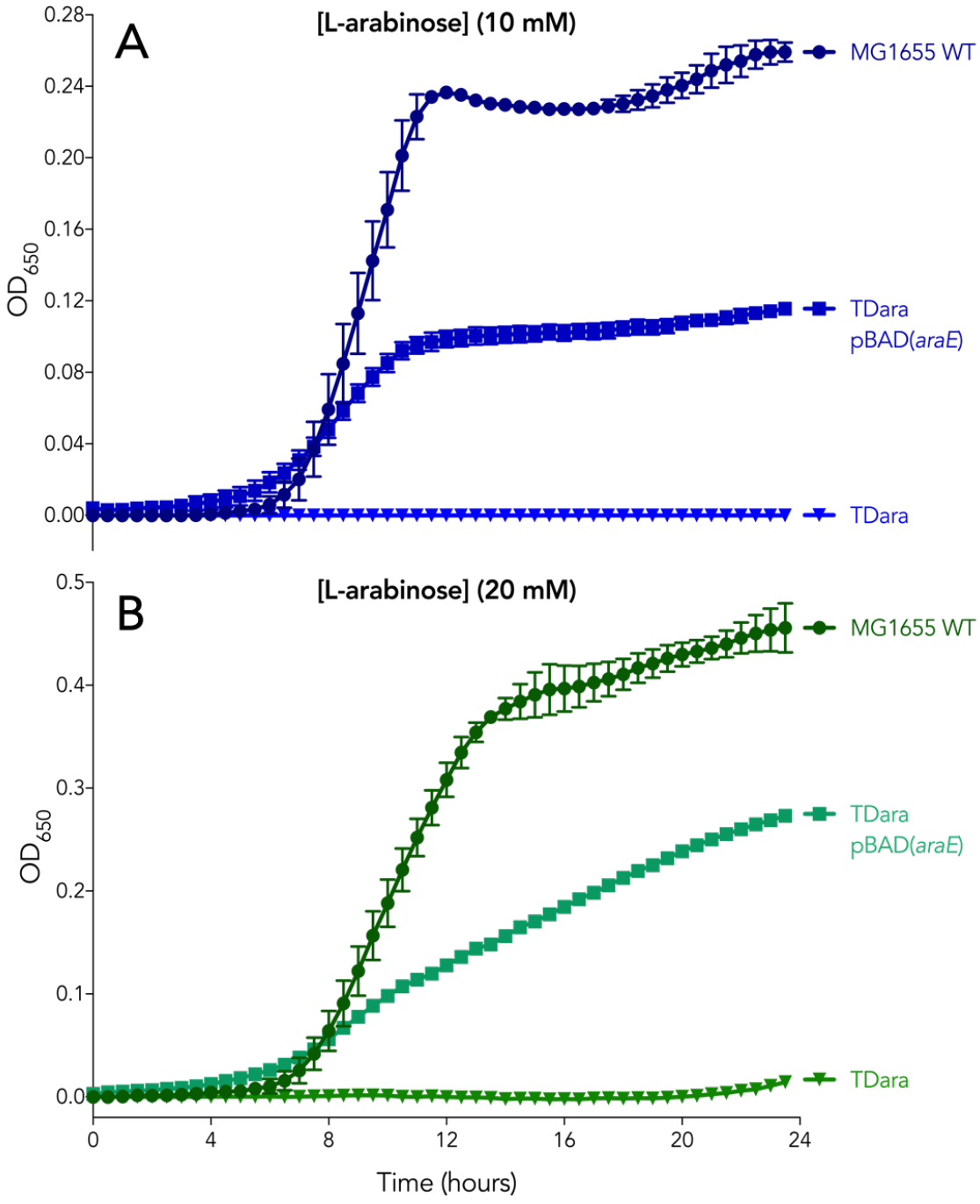


Figure 4. 4 Ectopic expression of araE from pBADcLIC partially restores growth of TDara.

A) Growth curve of TDara with pBAD(*araE) (squares)*. The *araE* expression was induced by 10 mM L-arabinose. The WT (*circles*) and TDara (*triangles*) are also shown for comparison. The cell density (OD_{650}) measurements were taken every 30 minutes across a 24-hours timecourse. Each timepoint is the average OD_{650} from 3 biological replicates and the error bars are standard errors calculated from the standard deviation.

B) Same growth assay with (A) in presence of 20 mM of L-arabinose.

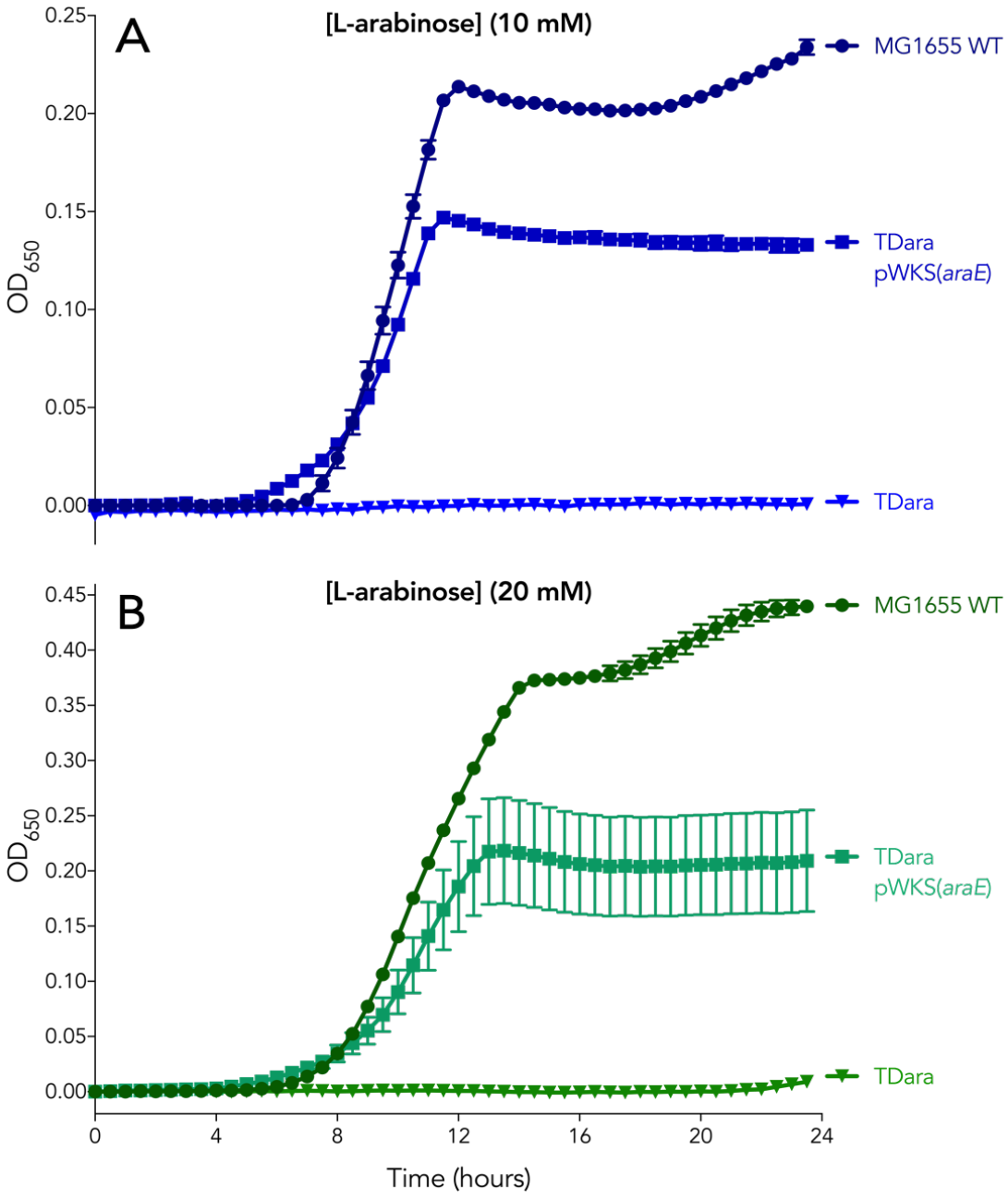


Figure 4. 5 Ectopic expression of araE from pWKS30 partially restores growth of TDara.

A) Growth curve of TDara with pWKS30(*araE) (squares)* in minimal media supplemented with 10 mM L-arabinose. The *araE* expression was induced with 1 mM IPTG. The WT (*circles*) and TDara (*triangles*) are also shown for comparison. The cell density (OD_{650}) measurements were taken every 30 minutes across a 24-hours timecourse. Each timepoint is the average OD_{650} from 3 biological replicates and the error bars are standard errors calculated from the standard deviation.

B) Same growth assay with (A) in presence of 20 mM of L-arabinose.

or none" effect of the Plac promoter in the pWKS30 previously observed (Novick and Weiner, 1957) as compared with low concentrations of L-arabinose used which would sustain the all or none induction of the pBAD vector (Siegele and Hu, 1997; Khlebnikov *et al*., 2000).

4. 3. 5 Functional complementation of TDxyl with pWKS30 (xylE)

The gene encoding for the xylose MFS transporter, *ie*. *xylE*, was cloned into a low-copy vector, *ie*. pWKS30, for ectopic expression in the TDxyl to test the level of the restoration of growth and the capacity of the strains to revert back into WT-like growth. The *xylE* sequence was firstly cloned into the pCR-II blunt vector as a blunt PCR product. The *xylE* insert was then subcloned into pWKS30. The pWKS30 (*xylE*) was transformed into chemically competent TDxyl and WT BW25113 cells for the subsequent growth assays.

The TDxyl pWKS30 (*xylE*) was tested for recovery of growth in M9 media supplemented with 10 mM or 20 mM of xylose as the sole carbon and energy source, as decribed at Materials and Methods *Section 4. 2. 4*. The expression of *xylE*, which is under the control of the lac promoter, was induced induced during the incubation course using 1 mM of IPTG. The growth of the culture was monitored via absorbance measurements every 30 minutes, over a total timespan of 24 hours, similar to *Sections 4. 3. 3* and *4. 3. 4*.

The results of the incubation showed that the ectopic expression of the *xylE*, restores the growth at higher levels than the WT (Figure 7). However, the complemented TDxyl presents a longer lag phase than the WT, irrespective of the xylose concentration. This could be attributed to the fact that the elevated XylE levels imposed toxicity to the cells and slowed their growth rate. As expected the WT BW25113 culture supplemented with higher concentration of xylose grew at higher biomass by the end of the incubation course compared to 10 mM of xylose. Likewise, the complemented strains showed growth levels dependent on sugar concentration. Interestingly, the BW25113 pWKS30 (*xylE*) presented a longer lag phase than the WT,

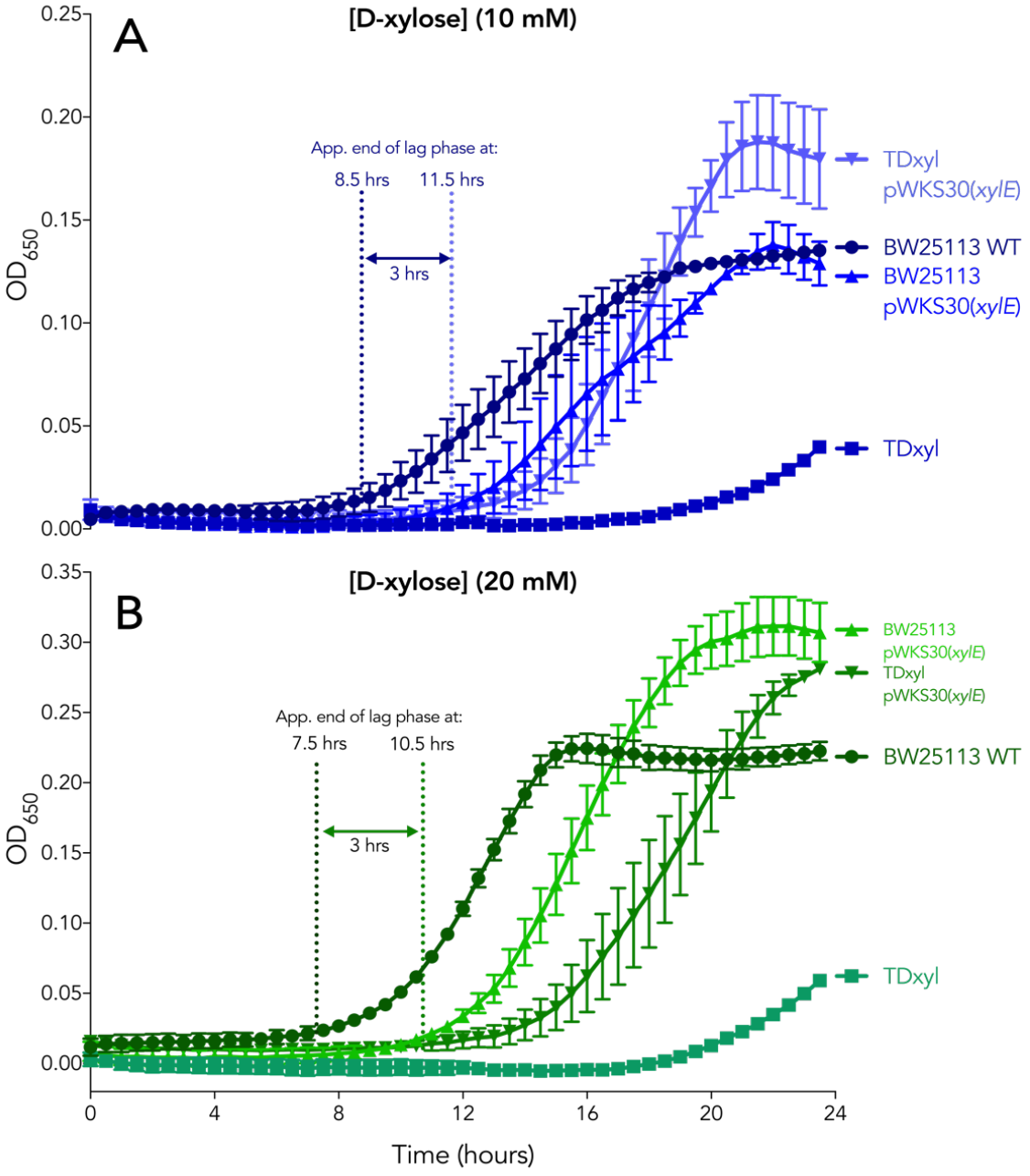


Figure 4. 6 Ectopic expression of xylE from pWKS30 restores growth of TDxyl.

A) Growth curve of TDxyl with pWKS30(*xylE) (nabla)* in minimal media supplemented with 10 mM D-xylose. The *xylE* expression was induced with 1 mM IPTG. The WT (*circles*) and TDxyl (*squares*) and are also shown for comparison. The cell density (OD_{650}) measurements were taken every 30 minutes across a 24-hours timecourse. Each timepoint is the average OD_{650} from 3 biological replicates and the error bars are standard errors calculated from the standard deviation.

B) Same growth assay with (A) in presence of 20 mM of D-xylose.

potentially for the same reason the complemented mutants did, but grew to a higher OD_{650} by the end of the incubation. This was not the case for the same strain that grew at 10 mM of substrate, which reached a similar OD_{650} to the WT by the end of the incubation (Figure 4. 6A). This could be because the secondary transporters are more efficient at higher substrate concentrations, and therefore there is higher accumulation of xylose inside the cells when grown at 20 mM. The deletion of XylE, XylH and AraH replacement in TDxyl cells completely abolished growth during the first 22 hours of incubation, at both xylose concentrations (Figure 4. 6B). Following 22 hours of incubation, some residual growth is observed which might be partly due to diffusion (Khankal *et al*., 2008).

4. 3. 2 Case study: Application of TDxyl in identification of a transporter which imports xylobiose in *E. coli*

4. 3. 2. 1 YagG from E. coli is a putative xylooligomer MFS transporter

Qian *et al*. (2003) (see *Section 4. 1. 3)* assessed the phylogenetic distribution of the XynT xyloside transporter from *Klebsiella oxytoca* and found that the conservation of the orthologues was not species dependent (Figure 4. 7A). The analysis found homologos of the XynTB operon in Gram (*-ve*) γ-proteobacteria *Klebsiella pneumoniae* and *E. coli* but also in Firmicutes such as *Clostridium acetobutilicum* (Figure 4. 7A). Interestingly, their phylogenetic tree includes five XynT orthologs in *E. coli* found in two of the groups (Figure 4. 7, Group II and IV). Group II contains the GusB (or UidB), YihP and YihO. None of them have been shown, or predicted to be involved in xyloside transport. GusB is a glucuronide-specific secondary transporter (Poolman *et al*., 1996) whereas YihO is a putative sulfoquinovose (*ie*. sulfonic acid derivative of glucose) symporter, as its deletion abolishes growth of *E. coli* in the aforesaid carbon source (Denger *et al*., 2014). The *yihP* is present in the same operon with the *yihO* and it is speculated to act as an efflux system for 3-sulfopropanediol (an end product of sulfoquinovose metabolism) (Denger *et al*., 2014). Nonetheless, the cluster encoding for the two aforementioned transport systems includes *yihQ* which produces an α-glucosidase and is active against α-glucosyl fluoride (Okuyama *et al*., 2004). The

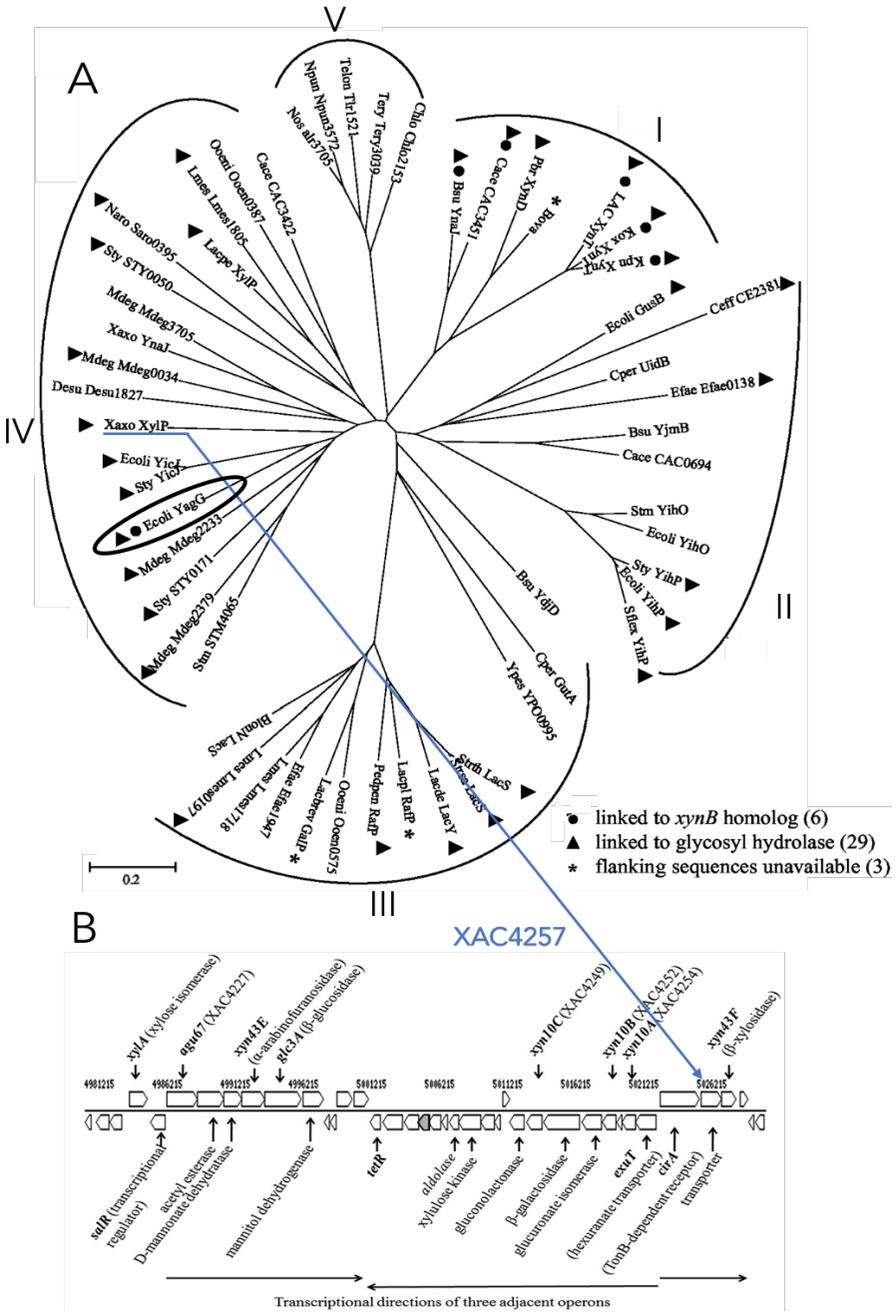


Figure 4. 7. YagG is a putative xyloside transporter

A) YagG is part of the unrooted phylogenetic tree produced by Qian *et al.* 2003 which demonstrated the relationship between the XynT orthologs. The abbreviated bacterial names are followed by the gene name. The transporters were proposed to form five separate Groups with Group I containing the starting query *ie*. $XynT_{Ko}$, which was characterised to be a xyloside transporter. YagG (*circled*) from *E. coli* neighbors with a XynB homolog and is found in group IV. The figure was rotated 90◦ clockwise to allow for better view of Group IV. Reprinted from Qian *et al.*, (2003).

B) XylP from *Xantomonas citri* pv. *citri* strain 306 (*ie*. Xaxo XylP, underlined) clusters itself in the same clade with YagG. XylP is part of a xylan utilsation operon, which contains β-xylosidase and xylanases for xylooligomers. Taken from Chow *et al.*, (2015).

confirmed function of this enzyme affirms the function predictions for the two transport systems found in proximity to it. The orthologs present in Group IV are more relevant to xylose uptake compared to Group II. The YicJ is a putative xyloside transporter which is present in the same operon with YicI. The operon is a predicted member of the XylR regulon (Garcia-Vallve and Romeu, 1999). Okuyama *et al.* (2004) have shown that the YicI is a α-xylosidase of the GH 31 family which is mostly active on isoprimeverose (*ie*. disaccharide comprised of xylose and glucose molecules), α-xylosyl fluoride and xyloglucan oligosaccharides. The enzyme was barely active on pNP α-xyloside (Okuyama *et al*., 2004), potentially indicating that YicI is failing to recognise xylooligomers. Okuyama *et al*. (2014) haven't reported the localisation of the hydrolase; in any case it is cytoplasmic then the cognate transporter, YicJ, is more likely to recognise and transport isoprimeverose and xyloglucans over xylooligomers. Use of online automated softwares for detection of the subcellular/ extracellular localisation of *E. coli* proteins, *ie*. STEP db 2.0 (Sub-cellular Topologies of *E. coli* Polypeptides) and EchoLOCATION, predict that YicI is localised in the cytoplasm and thus not secreted.

Another member of the Group IV, is the $XylP_{Xc}$ transporter from the *Xanthomonas citri* pv. *citri* strain 306. A study has shown that this bacterium owns a cluster of genes specific to the hydrolysis and utilisation of xylan (Chow *et al*., 2015) (Figure 4. 7B). The group employed a recombinant xylanase (Xyn10A) from the aforementioned cluster to hydrolyse xylan and used the hydrolysate as a carbon source for *Xantomonas citri* pv. *citri* strain 306. The presence of the oligosaccharides in the hydrolysate induced the expression of xylanases/xylosidases, *ie*. *xyn10A, B, C* and an α-glucuronidase *ie*. *agu67,* thus indicating the importance of the cluster in the utilisation of AGX. This was exemplified by the hydrolytic functions of Xyn10A and C which produced a mixture of MeGX3 (*ie*. methylglucuronoxylotriose), xylotriose and xylobiose (Chow *et al*., 2015). Additionally, the presence of xylotriose was able to increase the expression of *xyn10A*, *xyn10C* and *agu67*. BlastP search of $XynT_{Ko}$ restricted to *Xanthomonas citri* pv. *citri* strain 306 produced a highest identity match. Using the WP code of this protein in UniPrac (UniProt Archive), we were able to trace the match which was

found to be $XylP_{Xc}$ with XAC4257 as its locusID. The result confirmed their phylogenetic relationship reported by Qian *et al*. (2003) (Figure 4. 7B). XAC4257 is found immediately downstream the xylanases of the above study, therefore suggesting possible implication of the transporter in the import of xylobiose and xylotriose. Another interesting candidate present in Group IV is the YagG from *E. coli*. YagG is annotated as a putative xylosidase in Ecocyc, and TCDB (Transport Classification Database) classified it as a member of the glycoside-pentoside-hexuronide (GPH) transporters (Saier *et al.*, 2016). YagG clusters itself very closely to both $XylP_{Xc}$ and YciI (Figure 4. 7A) indicating evolutionary and potentially a functional relationship. It presents a 27 % sequence identity to $XynT_{Ko}$ and 36 % to $XylP_{Xc}$. Also YagG is found in the same operon with YagH which is a direct homolog of XynB from *Klebsiella oxytoca* (Qian *et al*., 2003). As XynB is active on xylodextrins, the YagH is predicted to have a β-xylosidase activity and has previously been classified as a GH 43 family hydrolase (Garcia-Vallve and Romeu, 1999). Since YicJ, as already mentioned above, is unlikely to be a xyloside transporter we chose to phenotype the function of YagG in an effort to validate the usability of TDxyl and potentially discover an endogenous *E. coli* xyloside-specific transporter.

4. 3. 2. 2 Ectopic expression of yagG restores growth of TDxyl xynB on xylobiose

To test the specificity of a putative xyloside transporter we require more than just the defective growth of TDxyl. To validate the xyloside-specific transport, an intracellular xylosidase is required to cleave the imported xyloside into xylose monomers which will subsequently be shuttled into the intracellular catabolism to fuel growth and produce a distinct phenotype during the growth assays. We therefore used a derivative of the TDxyl strain produced in Thomas lab, kindly provided by Rosanna Henessey. This derivative, *ie*. TDxyl *arsB::xynB*$_{Kp}$ (also referred here as TDxyl *xynB*) expresses an intracellular β-xylosidase from *Klebsiella pneumoniae* that is a direct homolog of the XynB from *Klebsiella oxytoca* used by Qian and his coworkers. Similar to its ortholog, $XynB_{Kp}$ was found to be active on xylobiose and xylotriose (Rosanna Henessey,

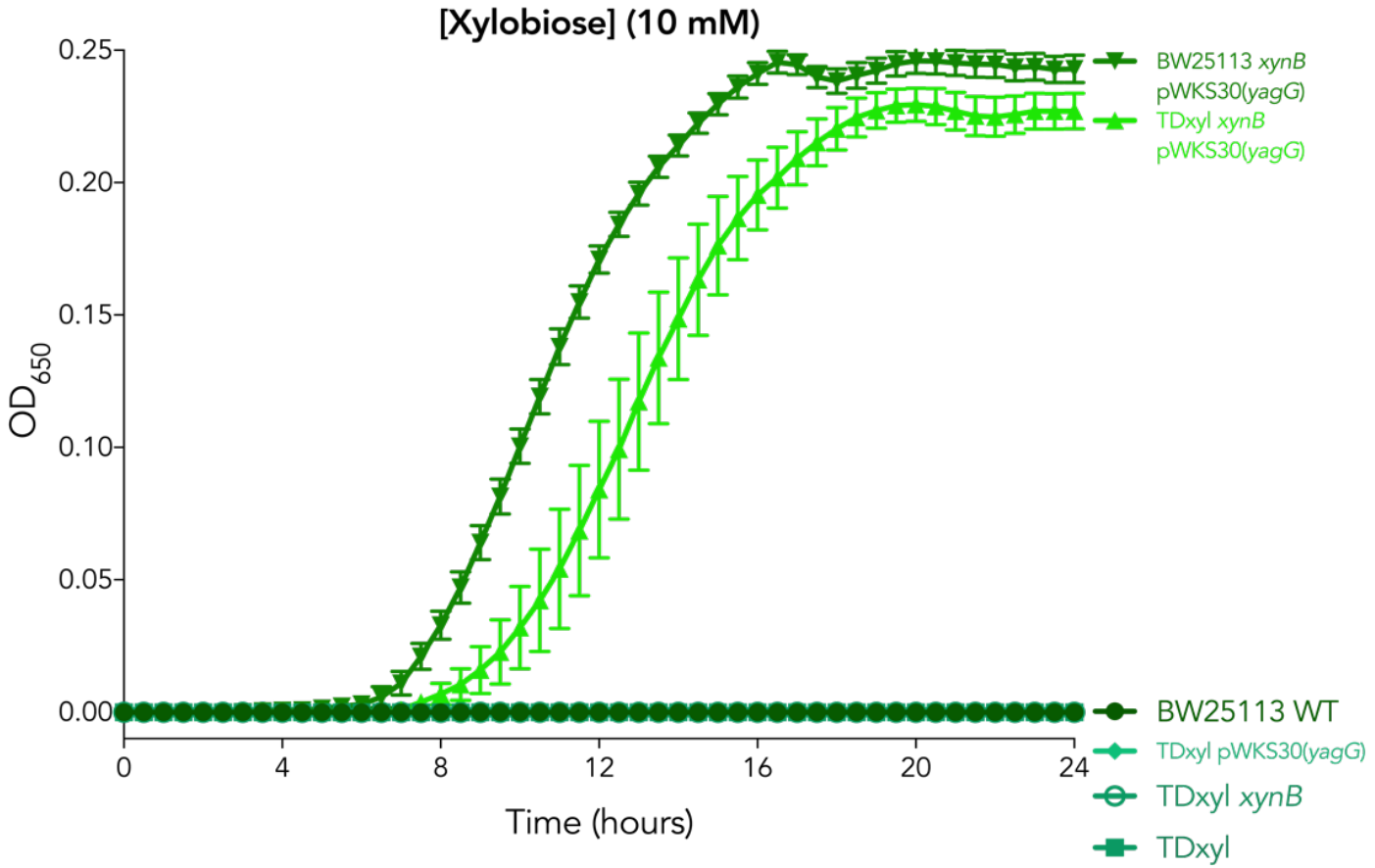


Figure 4. 8 Ectopic expression of yagG enables growth of E. coli on xylobiose
Growth curve of TDxyl *xynB pWKS30(yagG) (triangles)* in 10 mM xylobiose. The *yagG* expression was induced by 1 mM IPTG. The WT (*filled circles*), TDxyl (*squares*), Tdxyl xynB (*open circles*), WT *xynB* pWKS30(*yagG*) (*nabla*) and TDxyl pWKS30(*yagG*) (*rhombus*) were also included in the analysis. The cell density (OD_{650}) measurements were taken every 30 minutes across a 24-hours timecourse. Each timepoint is the average OD_{650} from 3 biological replicates and the error bars are standard errors calculated from the standard deviation. The error bars correspond to standard errors calculated as the square root of the standard deviation of 3 biological replicates.

personal communication). The TDxyl xynB has already been used with success to identify exogenous transporters for xylobiose and xylotriose (unpublished data). The *yagG* was cloned into the pWKS30 as described in the *Section 2. 2. 3. 9*. The resulting pWKS30 (*yagG*) was transformed into the strains to be tested in the microplate assay in a format similar to *Sections 4. 3. 3, 4, 5*. The results of the growth assay on 10 mM xylobiose are shown in Figure 4. 8. The growth assay showed that TDxyl *xynB* failed to grow confirming that the strain requires a dedicated transporter for the xylobiose (Figure 4. 8). Both, the BW25113 WT failed to grow underlining the inability of *E. coli* to naturally uptake and catabolise xylobiose. The ectopic expression of *yagG* in TDxyl *xynB* permitted growth on xylobiose which indicated that YagG can transport the disaccharide. Strikingly, the growth pattern of BW25113 *xynB* pWKS30(*yagG*) compared to TDxyl xynB pWKS30(*yagG*) resembles what is observed with *xylE* when overexpressed in the respective strains (Figure 4. 8 and 4. 4B). The xylose concentration is essentially the same between the two assays (*ie*. 20 mM), though it is potentially lower in this assay considering the β-xylosidase could become overwhelmed by the intracellular pool of xylobiose accumulated in the cytoplasm. This could explain for the higher OD_{650} reached in the assays which tested *xylE* compared to the assays assessed here with *yagG*. Overall, in both assays the BW25113 (or BW25113 *xynB*) expressing the constructs presented a shorter lag phase than the respective TDxyl (or TDxyl *xynB*). This could be explained by the presence of other transporters in the WT compared to TDxyl, which are able to transport xylose or adapted to transport xylobiose, leading to faster accumulation of the sugars and subsequently shorter lag phase.

4. 4. Discussion

Many tools have been produced for the fast screening of glycolytic hydrolases that act on various pre-treated feedstocks. Conversely, little to none attention has been given to the import of the hydrolysed xylan and as far as we are concerned no previous study has reported the creation of 'biotools' to serve identification of useful transport systems for xylan-derived sugars. The current study has produced and tested the function of strains useful in the engineering of membrane transport for biofuel production. These strains are called Transport Deletion (TD) mutants owing to their abolished or defected growth on the specified sugar/sugar acid. The TD mutants created are unable to grow on xylan-derived sugars/ sugar acids *ie*. D-xylose, L-arabinose and D-glucuronic acid. The growth of TD strains can revert back to partial or near-WT levels by re-introduction of their relevant MFS transporters. This validates their function as useful tools for identification of exogenous transporters by phenotyping. To further endorse their essentiality, an endogenous transporter from *E. coli*, *ie*. YagG, which is potentially cryptic is found to be able to transport xylobiose using the TDxyl strain.

4. 4. 1 'Validation' of TDglcA strain by recovering its growth

The deletion of *exuT* is sufficient to cause abolishment of growth on D-glucuronic acid. The study which mapped the hexuronate transport system in the *E. coli* chromosome obtained a similar observation (Mata-Gilsinger and Ritzenthaler, 1983); the various strains bearing mutated copies of *exuT* failed grow on D-galacturonic acid. As ExuT can transport both D-galacturonic acid and D-glucuronic acid, we anticipate that the growth defects extend to D-galacturonic acid as reported by Mata-Gilsinger and Ritzenthaler (1983). This is indeed the case (data not shown), which extend the function of TDglcA to a TDgalcA. This strain can be particularly useful for the transporter studies regarding pectin which is rich in D-galacturonic acid (Worth, 1967).

The TDglcA phenotype is able to be rescued by low copy production of ExuT. The basal expression of *exuT* from the plasmid appeared high enough to allow transport of D-glucuronic acid in amounts sufficient to propagate growth and induce the hexuronate catabolic genes. When induced the production of ExuT seems to cause a concentration-dependent phenotype, as at 20 mM the complemented strain grows to a lower final cell density compared to WT, while remaining unaffected in 10 mM. A possible explanation to this is the toxicity caused by the overproduction of membrane transporters which is a common obstacle in the crystallographic analysis of such systems. A different explanation, which is supported by the concentration dependent phenotype observed, is that the acidic nature of the substrate is lowering the intracellular pH thus leading to cell death. Similar observations were made before with the accumulation of metabolites causing depression of growth such in the case of gluconic acid and lactic acid build up halting the growth of *Gluconacetobacter xylinus* (Liu *et al*., 2016). These hypotheses are easily tested by growing the TD strain with or without ectopic expression of *exuT* in glucose. The absence of an acidic substrate will clarify the cause of the growth defect in the complemented strain.

4. 4. 2 Creation of a TD mutant strain for L-arabinose

The TDara is an *E. coli* MG1655 strain lacking the two arabinose transport systems, *ie*. AraE and AraFGH. The defective phenotype presented here has previously been reported but not in the form of growth assays. Horazdovrsky and Hogg (1989) investigated the three components of the AraFGH and how essential they were to growth in arabinose using complementation plasmids expressing combinations of the constructs. They reported that their respective TDara strain presented very low accumulation of L-arabinose (Horazdovrsky and Hogg, 1989). This was reversed by ectopic expression of the plasmid encoding for AraFGH, which increased intracellular L-arabinose levels by 100 times. The TDara strain reported in their study, lacked the whole *araFGH* operon whereas the present study reports that *araH* gene knockout, is

sufficient to deactivate the transport system. The phenotype is only evident when both transport systems are absent. As the deletion of *araE* led to elongation of the lag phase and the deletion of *araH* decreased the final OD_{650} we are prompt to speculate that the two systems are more active at different stages of the growth deduced by the concentration of L-arabinose present in the growth media. This speculation is supported by the inherent features of the systems, with the low affinity AraE being more active at high concentrations of arabinose *ie*. at the start of the incubation and AraFGH, as a high affinity system, being efficient at lower concentrations of the sugar *ie*. by the end of the growth assay. A study conducted by Mergele *et al*. (2008) investigated the timing and the dynamics of single cell expression for arabinose utilisation has concluded that the two systems could be orchestrated in such manner that they act like a single protein; this would explain the differing phenotypes observed here by the single mutants.

Regarding the inadequate complementation profile seen in TDara and single mutants with pBAD (*araE*), three presumptive events could be occurring:

(i) The expression of *araE* from a high copy number plasmid overwhelms the bacterial cells and therefore causes an early halt of growth and entry into the stationary phase. This is supported by the defective growth of Δ*araH* only observed in the presence of AraE when ectopically produced. To test this, the *araE* can be subcloned into a medium-copy vector and transformed into TDara and the individual single mutants to be re-tested in growth assays using the same sugar concentrations. Medium copy plasmids with pBAD promoters and various antibiotic resistance have recently been created (*ie*. pBR322) (Chakravartty and Cronan, 2015).

(ii) Excess arabinose inside the cell can be toxic when it leads to L-ribulose-5-phosphate (*ie*. intermediate metabolite of arabinose catabolism) accumulation (Koita and Rao, 2012). If the flux of arabinose is greater than the pentose phosphate pathway can accommodate it can lead to L-ribulose-5-

phosphate accumulation and increase toxicity (Koita and Rao, 2012). Even though this was suggested by Koita and Rao (2012) to excuse the presence of arabinose efflux pumps in *E. coli*, we identified no other study which attributed toxicity of arabinose due to its increased flux overwhelming intracellular metabolism. These arabinose efflux systems could become essential when higher concentrations of L-arabinose are encountered which are physiologically relevant. In fact, this phenomenon of L-arabinose-induced toxicity was only reported *in vitro* when the enzyme that removes L-ribulose-5-phosphate is deleted (*ie*. L-ribulose-5-phosphate-4-epimerase) (Englesberg *et al*., 1962). The possibility that this is the cause of the decreased growth is further diminished when the results are taken into consideration, as the TDara pBAD (*araE*) grew at higher OD_{650} at 20 mM as compared to 10 mM.

(iii) The last possibility is related to **all-or-none** induction of genes the L-arabinose exerts to pBAD promoter when *araE* expression is responsive to its concentration (Siegele and Hu, 1997). Siegele and Hu have used pBAD(*gfp*) constructs to track the response of *E. coli* populations to intermediate concentrations (*ie*. 13. 3 mM) of L-arabinose. The group analyzed population distributions and revealed that the pBAD system displays an all-or-nothing expression pattern similar to the lac system. They showed that in uninduced cells the background expression of the *ara* regulon leads to a wide distribution of *ara* uptake proteins at basal level. When they supplied inducer at the aforesaid concentrations, the population induced high production of the arabinose uptake systems in such way that led to a heterogenous timing of gene induction across the population, giving rise to a mixture of induced and uninduced cells. The first retain stable production of the uptake systems whereas the second go through stochastic events which maintain basal level of the transporters (Siegele and Hu, 1997). However, whether the uninduced cells will finally accumulate enough L-arabinose to induce expression (in our case of *araE*) depends on how high the inducer concentration is. In the case of lac promoter (Novick and Weiner, 1957), the uninduced cells actually

presented a growth advantage over the induced cells therefore outreplicating them as the growth assay (Novick and Weiner, 1957). Again, Siegele and Hu suggested the L-arabinose levels deduce whether the induced cells will outperform the uninduced (Siegele and Hu, 1997). This is agreeable with the results seen here as TDara pBAD(*araE*) strain, whose growth is solely dependent on the *araE* induction by arabinose, reaches stationary phase following 10 hours of incubation at 10 mM arabinose (Figure 4. 4). This could be attributed to uninduced cells fully outperforming the induced cells the therefore completely switching off expression of *araE* (*ie*. all or **none** induction), despite some arabinose inducer still being present. Whereas in the case of 20 mM there is a prolonged exponential phase, which is still active by the end of the 24 hours of incubation. Due to the higher concentration of arabinose in the media the rate of induction seems to outperform the rate at which uninduced cells outreplicate induced cells. To test whether the higher rates of induction would be fast enough to bring about higher numbers of induced cells and fix them as the prominent population in the culture (*ie*. **all** or none induction), higher concentration of L-arabinose can be supplemented in the media. In this case, we expect that the induced cells will become the predominant pool in the culture fast enough to decrease the prolonged log phase and grow similarly to WT. Another way to test this is by decoupling the *araE* expression from L-arabinose concentration and remove this autocatalytic phenomenon. This was attempted by Khlebnikov *et al*. (2010), by placing the *araE* expression under the Ptac promoter which is inducible by IPTG, on a low copy vector. The uncoupling led to homogeneous expression of the *araBAD* promoter across the system, as tracked by *gfp* under the control of the promoter in question (Khlebnikov *et al*. 2010). This was indeed attempted in the present study, however based of the (*i*) assumption a low copy vector (pWKS30) was used to circumvent any potential toxic effects caused by high copy number pBAD vector. The Ptac promoter is stronger than the Plac (Aman, Ochs and Abel, 1988) present in pWKS30 (Wang and

Kushner, 1991) so we didn't expect to see the homogeneous response to arabinose observed by Khlebnikov *et al* (2010). Nonetheless, the IPTG-inducible *araE* expression actually improved the phenotype of the TDara as in 10 Mm arabinose it led to 37.5 % difference in final OD_{650} compared to WT, whereas TDara pBAD(*araE*) grew to a 56 % lower final OD_{650} compared to WT (Figure 4. 5). Based on the results by Novick and Weiner (1957) the Plac is still subjected to all or none induction when induced with methyl-β-D-thiogalactoside (TMG). However, the group tested 111 to 166 times lower concentration of inducer than the IPTG concentration used in our experiments. Therefore, we expect that 1mM IPTG will be sufficient to allow for the induced population of the culture to outperform the uninduced. When the Δ*araE* single mutant is taken into consideration one should account for the positive feedback loop exerted by the basal expression of the chromosomal copy of *araFGH*, *ie*. when adequate amount of arabinose is transported inside the cytoplasm by the ectopically produced AraE, then the induction threshold is reached and more units of AraFGH act synergistically with the AraE to accumulate further arabinose in the cell. This is confirmed by the phenotype of the Δ*araE* mutant which is restored to almost WT levels (Figure A4. 3A). Therefore, the control of the two transport systems by two distinct promoters finally accomplishes complementation of the defected growth presented by the Δ*araE* mutant.

Despite the moderate complementation levels of TDara by pBAD(*araE*) and pWKS30 (*araE*), this strain can still be used for identification of exogenous transporters of arabinose or arabinosides on the basis that the assessment is qualitative. Nonetheless, one can use the Δ*araE* single mutant instead of the TDara and screen for decrease of lag phase duration by exogenous transporters expressed on pWKS30.

4. 4. 3 'Validation' of a TDxyl strain by recovering its growth

Similar to TDglcA and TDara, the phenotype of TDxyl is somewhat rescued when the MFS transporter XylE is expressed ectopically. Unlike TDglcA, the complemented TDxyl grows well in 20 mM xylose, possibly because of the neutral nature of the xylose compared to glucuronic acid. It is tempting to try and perform an interpretation of the results similar to the complementation of TDara with pWKS30, however this could lead to wrong conclusions because the single mutants weren't tested in such conditions and the *xylE* was only expressed in pWKS30 and not in higher copied plasmids. A comparison between the two experiments can still be drawn though, as the same vector was used for both constructs and the carbon source is present in the same concentrations. The TDara pWKS30(*araE*) failed to reach a final OD_{650} higher to the WT unlike the TDxyl pWKS30(*xylE*) which did on both sugar concentrations tested. This could be attributed to the binding affinities of the MFS transporters in question; XylE shows 3 times higher affinity than AraE (*ie*. 63 – 169 μM compared to 140-320 μM for the latter) (Sumiya et al., 1995; Daruwalla, Paxton and Henderson, 1981), therefore rendering it more active at lower sugar concentrations, normally present at the end of the exponential phase compared to the start of the incubation (Kazuki Iizuka, collaboration).

The production of XylE causes the lag phase to elongate at both TDxyl and BW25113 indicating that this could be the outcome of excess XylE units causing toxicity. This is not uncommon, as it has previously been shown that the production of the glycerol-3-phosphate MFS antiporter (*ie*. GlpT) in *E. coli*, when induced by IPTG during the incubation course, it causes a plateu which lasts for approximately 2.5 hours before the strain is able to continue growing (Gubellini *et al*., 2011). Possible toxicity can be tested by growing the strains on glucose and assess the XylE presence irrespective of its transport activities. Despite the extended lag phase, the TDxyl was still used to show the capacity of endogenous YagG to transport xylobiose with success.

4. 4. 3 Detection of a transporter from *E. coli* able to import xylobiose

Employing the derivative strain of TDxyl, TDxyl *xynB*, we identified a transport system in *E. coli* which can transport xylobiose. This is the first reported system in *E. coli* which can transport a xyloside. Currently we cannot be sure if the YagG is a dedicated xylobiose transporter as the TDxyl *xynB* pWKS30 (yagG) should be tested for growth recovery on more xylosides, to help define the specificity of the system. This experiment was not performed during the present study due to time limitation and the high cost attached with purchasing xylotriose and larger xylooligomers.

A very important question following this discovery is why is WT *E. coli* unable to grow on xylobiose if it possesses YagG. One possibility is that the *yagGH* operon is inactivated and therefore not expressed in presence of xylobiose. The operon indeed lies within a cryptic gene island derived from a prophage named as CP4-6 prophage by Blatnerr *et al*. (1997) (Casjens, 2003; Garcia-Vallvè and Romeu, 1999). Garcia-Vallvè and Romeu calculated the G+C content of *yagH* and found it was 64% which differed distinctly for the G+C of *E. coli* glycolytic hydrolases (*ie*. 52.6 ± 3.2% and from the average for the *E. coli* genome (*ie*. 50.8 %). Also, their study showed that the percentage of G+C at the third codon position for *yagH* was 87.9% which is much different than the general value for E. coli genes which was found to be 58.2 % (Garcia-Vallvè and Romeu, 1999). In their phylogenetic analysis using the UPGMA (Unweighted Pair Group Method with Arithmetic Mean) method they showed that the *yagH* positioned itself as an outgroup away from the *E. coli* GHs in the produced dendrogram (Garcia-Vallvè and Romeu, 1999). They suggest that the *yagH* was acquired through a transduction event based on the fact that *argF* gene, which is also present in the CP4-6 prophage, was acquired in such manner (Van Vliet, Boyen, and Glansdorff, 1988). These indications suggest that the *yagGH* operon is horizontally transferred. However, putative XylR and CRP binding sites have been detected upstream the *yagGH* operon in a computational study (Laikova, 2001). The regulation has not been verified experimentally and therefore the operon might not be responsive to xylose via XylR. This can be tested by growing WT and TDxyl *xynB* in a

mixture of sugars containing a small amount of xylose and a larger amount of xylobiose. The xylose will ensure activation of XylR, which in a hypothetical scenario will activate the expression of *yagGH* leading to growth. A positive phenotype can be attributed to the transport activity of YagG by RT-qPCR analysis. Also, a β-galactosidase reporter assay of the yagGH promoter region would clarify whether it is responsive to xylose. The experiment can be repeated with TDxyl to test the putative xylosidase activity of YagH. The current study attempted to clarify the activity of the YagH by cloning the whole operon in pWKS30 and use it to restore growth of TDxyl on xylobiose. Despite the strain failing to grow (data not shown), YagH could still be a xylosidase as the operon was cloned as found in the chromosome and therefore the *yagH* was placed in distance from the Plac. Irrespective of the regulation of *yagGH* expression by *E. coli,* the YagG is shown here to be able to transport xylobiose which raises its bioindustrial importance. The discovery is vital for engineering *E. coli* to utilise xylan derived sugars, as is shown here that despite yagG being a horizontally acquired gene, its sequence didn't require codon optimisation for production in *E. coli* and it didn't impose toxicity when overexpressed in the low copy vector pWKS30.

Chapter 5

Conclusion

Combustion of fossil-fuels increases the levels of atmospheric carbon dioxide (CO_2) and correspondingly becomes a great contributor to the anthropogenic climate change. Plants are a potential alternative to petroleum-based fuels, as they fix CO_2 via photosynthesis, leading to neutral carbon emission if used for combustion of biofuels. The hemicellulotic component of studied plant feedstocks, including bioenergy grasses, is made of a xylan polysaccharide. Xylan represents the most abundant hemicellulosic polysaccharide and is composed primarily of xylose, arabinofuranose, and glucuronic acid. Bacteria have evolved a plethora of enzymatic strategies for hydrolyzing xylan into its oligometic and monomeric sugars which are used to power growth in vicinities to such polysaccharides. These mechanisms have been extensively studied and many enzymes have reported for activity to different types of xylan and oligomers. An overlooked stage in the bioconversion of biomass into biofuel is the import of the released sugars and oligosaccharides. Given the propensity of bacterial organisms to control the transport of sugars in response to the nutritional status of their environment, an understanding of the mechanisms and identification of sugar uptake systems is critical to enable metabolic engineering strategies to be designed to achieve optimal substrate utilization and fermentation performance. The present study worked towards this purpose, to identify and characterise bacterial systems which are specific to L-arabinofuranose and xylobiose.

Chapter 2 used the galacto/ arabinofuranose binding protein from *E. coli* (YtfQ/GafAEc; Horler *et al.*, 2009) as the starting query to identify bacterial orthologs

(*ie*. GafAs) and by phylogenetic, phylogenomic and '*in silico*' searches to infer relevance to L-arabinose uptake. As *E. coli* does not encounter arabinofuranose as much as plant symbionts and pathogens do, we hypothesized that some of the orthologs from such bacteria could exhibit unique features that enhance L-arabinofuranose uptake. Unlike L-arabinofuranose, the furanose form of D-galactose is not a structural unit of plant polysaccharides *ie*. galactans of pectin, which suggest that GafAs from plant bacteria might be oriented towards arabinofuranose binding (Jones, Seymour and Knox, 1997; Bhattacharjee and Timell, 1964). GafAEc presents almost equimolar affinity for arabinose and galactose (Horler *et al*., 2009), and is also able to bind a wide range of substrates which can exist in their furanose forms from moderate to low amounts in solution. Therefore, one of the aims was to identify an ortholog with limited specificity compared to GafAEc and preferably enhanced affinity for L-arabinofuranose. The phylogenetic tree produced included 110 GafA orthologs and demonstrated that these are widespread in bacterial phyla including α, β, γ, δ-proteobacteria, Firmicutes and Spirochaetes. Further, the number of candidates was reduced to include *gafAs* with interesting genomic context which related to arabinan, xylan or galactose utilization. Many of the orthologs neighbored with arabinofuranosidases from families GH43, GH51, GH54 and GH62 as classified by CAZy. Irrespective of their mechanism of action, there are examples of resolved structures from all aforesaid families co-crystallised with the furan form of arabinose as their product, underlining the need for an arabinofuranose-specific transport system for both bacterial systems and engineering strategies (Maehara *et al*., 2013; Bouraoui *et al*., 2016 and Hoevel *et al*., 2003). The three selected GafAs were GafACb and GafASm from *Shewanella* sp. ANA-3 and *Clostridium beijerinckii* NCIMB 8052 respectively due to their arabinose-related genomic context, predictions of arabinose-regulated expression (Novichkov *et al*., 2013) and inability to grow on galactose (only presented by *Shewanella* sp. ANA-3) (Wetmore *et al*., 2015). GafASm from *S. meliloti* 1021 was also included in the downstream analysis, as its expression is induced in L-arabinose and D-fucose (Mauchline *et al*., 2006), two sugars that are prominent in the root mucilage (Nguimbou *et al*., 2012; Tyler, 1965). GafASm and GafASw,

overproduced in BL21 Star cells, and their presence was confirmed by FT-ICR-MS analysis.

Chapter 3 characterised the two GafAs by biochemical and structural analysis. GafASm is able to bind D-galactose despite its expression not being induced by it. The change in the thermal stability of the GafASm in the presence of D-fucose was similar to D-galactose, suggesting that the two ligands might bind with equimolar or similar affinities. In the case of GafAEc, D-fucose binds 6 times weaker than D-galactose (Horler *et al.*, 2009); therefore, a potentially enhanced affinity for D-fucose presented by GafASm suggests that this sugar might be a chemoattractant derived from mucilage. L-arabinose bound GafASm with 5 times higher affinity than GafAEc, supporting the previous assumption that GafABCD system might be involved in chemotaxis and also allowing for a competitive advantage in the rhizosphere.

Biochemical analysis of GafASw revealed its ligand specificity and showed that this SBP is unable to bind D-galactose. This inability is a unique feature of GafASw and places it amongst the few cases of arabinose-related bacterial proteins that lack specificity for galactose (Declerck and Abelson, 1994, Wallace *et al.*, 1978). The structural analysis attributed this inability to tilting of Asp88 disrupting a salt bridge with a nearby Asn residue and the presence of a phenylalanine residue instead of an asparagine, compared to GafAEc. The Asp88 is instead interacting with the C5 of arabinofuranose. The phenylalanine residue is predicted to disrupt interaction of the O_6 of galactofuranose with a water molecule and thus destabilizing its hypothetical binding. GafASw binds L-arabinofuranose with 4 times and 12 times weaker affinity than GafAEc and GafASm, potentially indicating that it encounters higher amounts of this sugar in its environment. This is suggested by the large arabinose utilization cluster which includes *gaf* operon and a range of extracellular, periplasmic and cytoplasmic arabinofuranosidases. Both GafASm and GafASw showed high stability and activity at highly alkaline buffers and therefore are excellent candidates to be considered import of sugar from alkaline pretreated biomass. Also, the narrow ligand specificity of GafASw, increases its bioindustrial importance as the binding of L-arabinofuranose is not inhibited by D-galactofuranose, normally found in galactans of

pectin (Motherway, Fitzerald and van Sinderen, 2010) and potentially in pretreated biomass.

Chapter 4 produced and tested the function of strains useful in the engineering of membrane transport for biofuel production. These strains are called Transport Deletion (TD) mutants owing to their abolished or defected growth on the specified sugar/sugar acid. Since the aim of the study is oriented towards utilization of xylan, we developed TD strains unable to grow on sugars/ sugar acids commonly found in xylan. The TD mutants created by Thomas lab co-workers are unable to grow on D-xylose and D-glucuronic acid. The first was named TDxyl, which doesn't produce the xylose secondary transporter, XylE and the membrane subunits, XylH and AraH, of the xylose and arabinose ABC systems, respectively. The TDglc shows a complete abolishment of growth in glucuronic acid because of not producing the ExuT MFS transporter of hexuronates. The present study added TDara in the collection of xylan TD mutants, a strain unable to grow on minimal media supplemented with arabinose. The deletion of *araE* and *araH* by λRed-assisted recombination was sufficient to abolish growth of *E. coli* MG1655 on arabinose.

The growth of TD strains was reverted back to partial or near-WT levels by re-introduction of their relevant MFS transporters. This validated their function as useful tools for identification of exogenous transporters by phenotyping. TDara overproducing *araE* from the high-copy vector pBADcLIC grew to moderate levels compared to the WT strain. This could be attributable to the all-or-none induction of genes when under the expression of P_{araBAD} promoter. However, *araE* expressed from the low copy vector pWKS30 shortened the lag phase of the Δ*araE* single mutant, and resembled the WT growth. Therefore, we conclude that Δ*araE* can be used for the same application as a TDara would.

To further endorse the essentiality of TD strains, an endogenous transporter from *E. coli*, *ie*. YagG, which is potentially cryptic is found to be able to transport xylobiose using the TDxyl strain. Employing the derivative strain of TDxyl, TDxyl *xynB* (encodes for an intracellular xylosidase from *Klebsiella pneumoniae*), we identified a transport

system in *E. coli* which can transport xylobiose. This is the first reported system in *E. coli* which can transport a xyloside. The ectopic expression of YagG from pWKS30 enabled growth of TDxyl *xynB* on xylobiose without apparent toxicity from overproduction. The discovery adds YagG in the repertoire of systems to be considered for engineering of bacteria in production of biofuels.

The ultimate envision for the study is to incorporate the above functionalities into *E. coli* to improve the uptake of xylan-derived sugars and oligosaccharides for biofuel production. To work towards this aim, we obtained an ethalogenic *E. coli* strain produced by Woodruff *et al.*, (2013) that bears chromosomal copies of the homoethanol pathway from *Zymomonas mobilis*, and has deleted genes encoding for endogenous fermenting pathways. The engineered strain, namely LW06, produced ethanol approaching the minimum inhibitory concentration for *E. coli* growth (*ie*. 30 g/L) which approximated 80% of theoretical yield for engineered *E. coli* strains (Woodruff *et al.*, 2013). Due to deletion of native competing pathways, the fermentation byproducts were kept at limited concentrations (*ie*. succinate, acetate, formate and lactate) (Woodruff *et al.*, 2013). As the strain also bears a chromosomal copy of *bla* gene it has ampicillin resistance. To enable downstream engineering of the strain we deleted *bla* gene using the λRed system (data not shown) and created CD01 strain. The modified CD01 can still produce ethanol as assessed qualitatively using aldehyde indicator plates (Conway *et al.*, 1987) (data not shown).

CD01 can be further engineered by introducing the following functionalities studied here:

(a) GafASw can be engineered to recognise the membrane components of the Gaf system in *E. coli* (*ie*. GafBCEc) (Figure 5. 1). Because of the high homology between GafAEc and GafASw, the latter might require minimal engineering to dock on the membrane proteins of the native Gaf system. If this fails, the whole system can be ectopically expressed or replace GafABCDEc in the chromosome.

(b) The *exuT* can be tested for uptake of methylglucuronic acid, or a demethylase (Shulami *et al*., 1999) can be targeted in the *E. coli* periplasm to remove the methyl group and allow import and catabolism of glucuronic acid.

(c) All engineered components should be under Pxyl promoter (Figure 5. 1) so that they are responsive to the sugar found in the highest abundance in xylan, *ie*. xylose. This will alleviate the phenomenon of sugar utilisation hierarchy (ie. glucose→arabinose→xylose) and also remove the need for exogenously supplied inducers *eg*. IPTG. Induction by xylose will also ensure production of *gafABCDEc* and *yagG* in the presence of xylan as these are not known to be responsive to arabinose or xylooligomers, respectively.

(d) The production of hydrolases is another stage to be considered in CBP of xylan by *E. coli*. These can be produced by a second strain to alleviate the burden of overproduction caused by multiple genes ectopically expressed in *E. coli*. The importance of arabinofuranosidases in utilisation of xylan is demonstrated by the synergistic action of these enzymes with endo-xylanases, which led to higher amounts of xylose liberated from xylan backbone (Goldbeck *et al*., 2014). An interesting candidate to be engineered into *E. coli* is a recently discovered bifunctional enzyme, *ie*. Ac-Abf51A, from *Alicyclobacillus* sp. A4. The enzyme is both an exo-α-L-arabinofuranosidase and endo-xylanase shown to be able to hydrolyse arabinose and xylose from wheat arabinoxylan (Yang e t al., 2015). The gene can be tagged with OsmY for secretion, a tag that has been used successfully to produce and secrete endo-xylanases in the extracellular environment of *E. coli* (Bokinsky *et al*., 2011).

Further to engineering a single or multiple *E. coli* strains as described above, additional work should focus on mutating the Gln19 and Asn107 residues of GafAEc to Phe so that it resembles GafASw. This way the worker can exemplify whether the abolishment of D-galactofuranose binding is solely dependent to the presence of aromatic amino acids at these two positions. Additionally, it would be useful

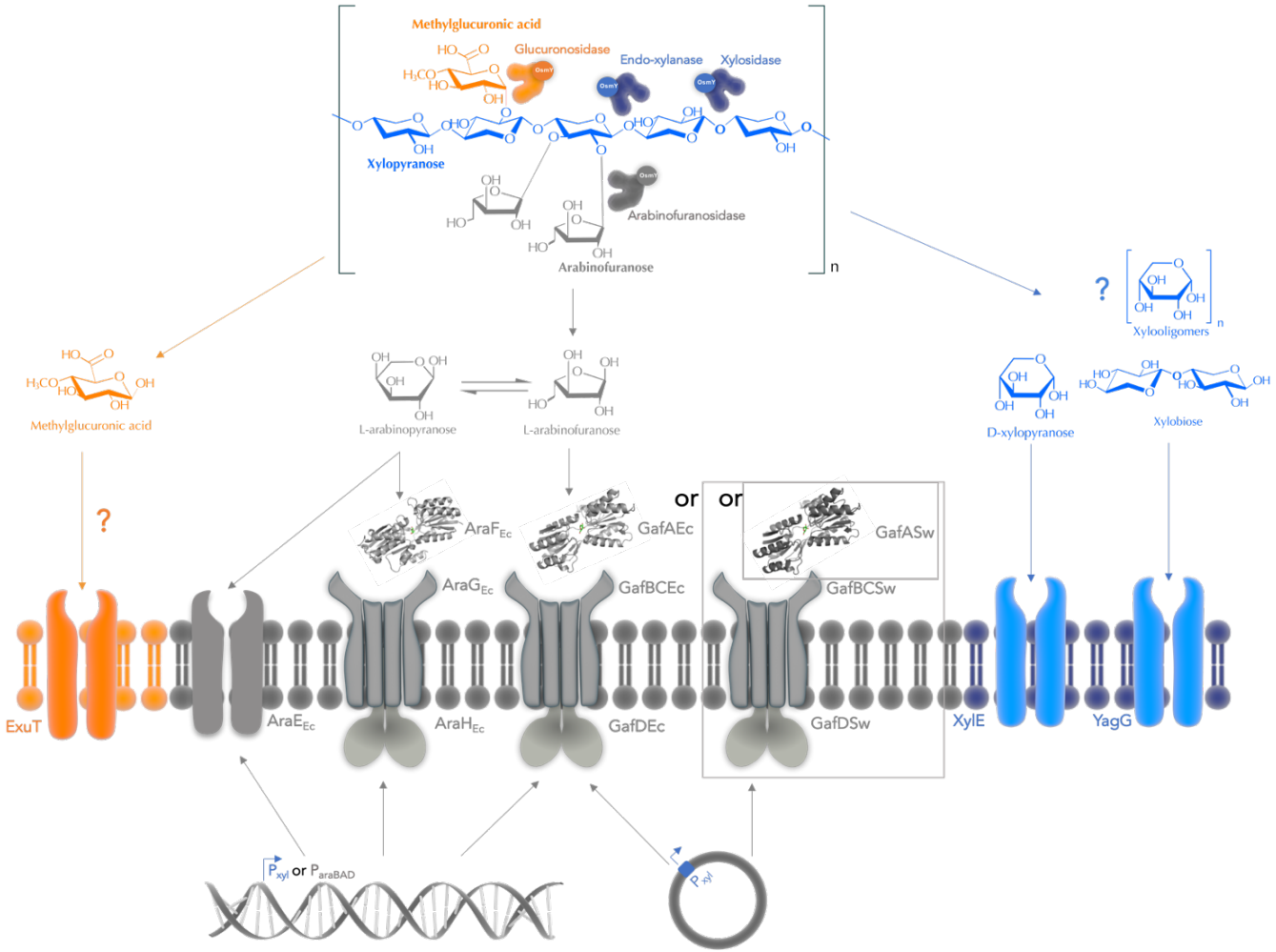


Figure 5. 1 Engineering of studied functionalities into ethalogenic strain CD01.
Further engineering of CD01 involves improving the uptake of the released sugars from xylan. GafASw can be engineered to recognise the GafBCEc membrane components or the whole system can be added into *E. coli* chromosome. All the components, including the MFS transporters for glucuronic acid (*exuT*), xylose (*xylE*) and xylobiose (*yagG*) should be expressed under the control of the Pxyl promoter so that the sugar utilisation hierarchy is alleviated.

to introduce mutations in the aforesaid sites so that engineered GafAEc resembles the GafAs *Xylanimonas cellulosilytica* DSM 15894 (*ie*. Xcel_0419), *Variovorax paradoxus* S110 (*ie*. Vapar_1044) and *Herbaspirillum seropidicae* SmR1 (*ie*. Hsero_1033) so that it features an Asn and Phe at positions 19 and 107, respectively. Binding assays using the engineered versions of GafAEc will shed light on potentially altered binding specificities and/or affinities of these orthologues.

Moreover, the TDara can be utilised in competition assays with actively released arabinofuranose from arabinan. We have already cloned the *gafABCD*Ec and *araFGH* in pWKS30 for these competition growth assays, where the cytoplasmic accumulation of the released arabinofuranose will be monitored using GFP/RFP under the control of pBAD promoter. Nonetheless, monitoring the growth of these two strains without the reporter would still provide useful information, since a potentially shorter lag phase for the TDara overproducing the Gaf system over the pyranose-specific system would imply faster more efficient uptake of the released carbon source. The engineered versions of GafAEc mentioned in the last paragraph can be included in these experiments to test how they differ to the endogenous version during the competition assays. The presence of D-galactose in competition assays of GafAEc vs. GafAEc (E19F, N107F) would give us an indication of the level of inhibition on uptake exerted.

Appendix

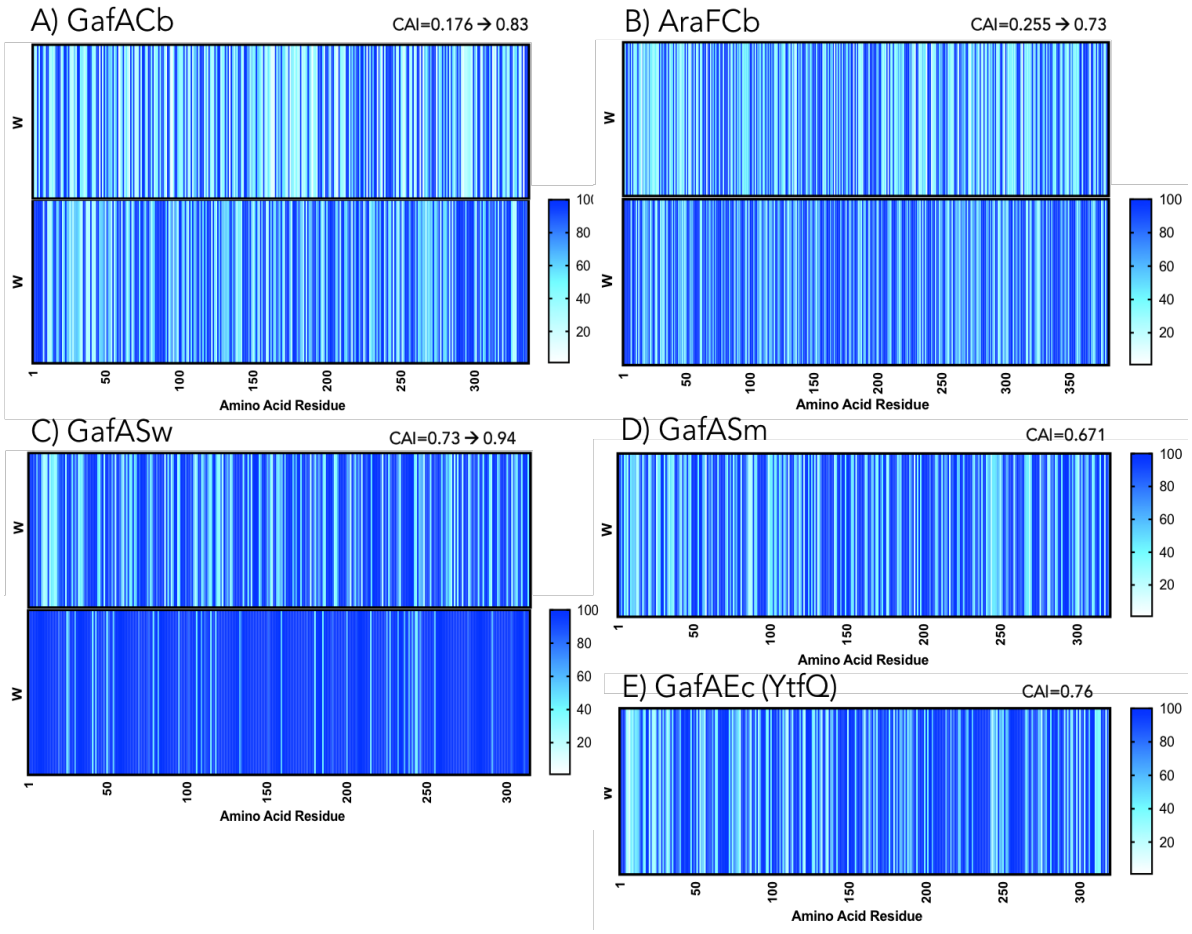


Figure A2. 1. Optimisation of the GafA and AraFCb coding sequences for expression in E. coli.

Codon usage is presented by the heat maps which plot the relative adaptiveness, W, values as a function of the residue number. The top heat maps for A, B and C correspond to the relative adaptiveness per amino acid of the coding sequence of GafACb, AraFCb and GafASw prior to codon optimization in Jcat, respectively. The bottom heat maps correspond to the coding sequence obtained after optimisation. The codon adaptiveness index (CAI) for each sequence, before and after optimisation, is designated at the top of each heat map. The heat maps of GafASm and the control protein, YtfQ, are shown in D and E respectively.

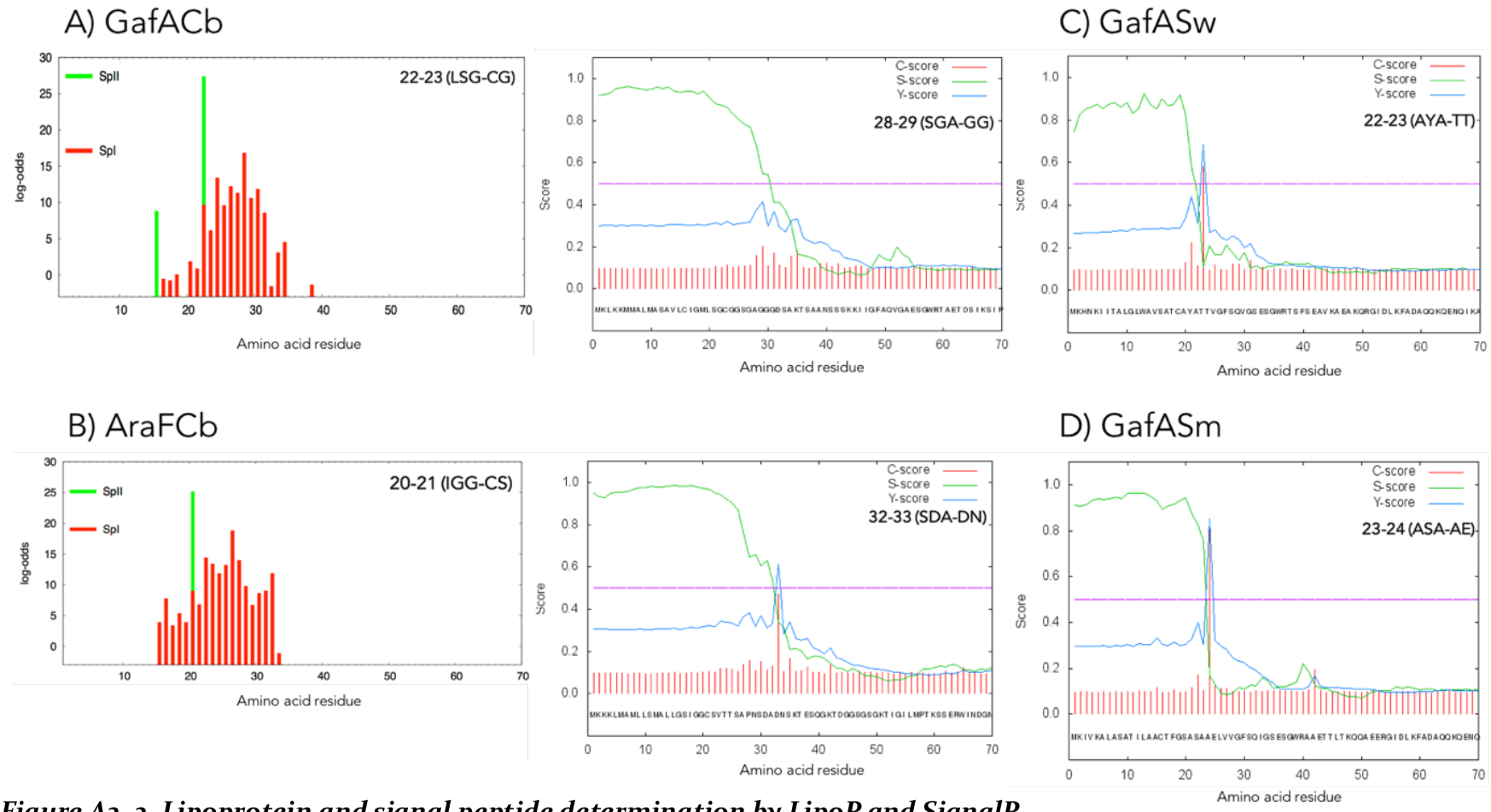


Figure A2. 2. Lipoprotein and signal peptide determination by LipoP and SignalP.
A) (Left) The results of LipoP analysis of the GafACb protein sequence showed that a potential cleavage site is present between residues 22-**23** (LSG-**C**G). (Right) The results were not reciprocated by SignalP, as the latter positioned the cleavage site between the amino acids 28-**29** (SGA-**G**C)**.** B) (Left) Similar analysis in LipoP for the detection of AraFCb lipopeptide showed a potential cleavage site predicted by LipoP present between 20-**21** (IGG-**C**S). (Right) However, the prediction from the SignalP placed the cleavage site 11 amino acid downstream the LipoP site. C & D) The SignalP predictions for the GafASw and GafASm detected the putative cleavage sites at 22-**23** (AYA-**T**T) and 23-**24** (ASA-**A**E), respectively.

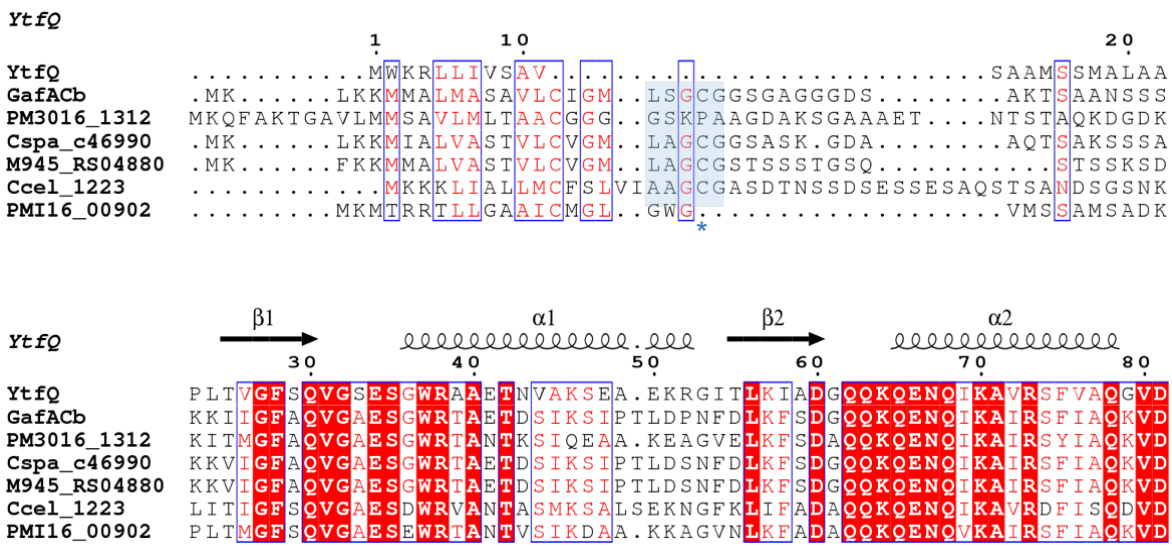


Figure A2. 3. Multiple structure-based alignment of YtfQ (GafAEc) against GafACb (Cbei_4462) and its closely related orthologues.

The orthologues of the GafACb with the highest identity score were obtained by BlastP search. The structure-based alignment was performed in ESpript3.0 using the YtfQ structure (PDB: 2VK2) as the template. The GafAs from Firmicutes contain from 45 (GafACb) to 57 (PM3016_1312) amino acids prior to the first secondary element of the YtfQ structure. The predicted lipobox for the Firmicutes GafAs is shaded in blue and the conserved Cys lipid anchor residue is designated with a blue asterisk. Following cleavage, there is an additional 20 - 27 amino acid residues retained in these GafAs when compared to YtfQ. Conversely, the GafAs from the Gram (-) bacteria (*ie*. YtfQ and PMI16_00902) retain only 3 amino acids upstream the mature protein following cleavage of the predicted signal peptide by SignalP. As the role of the extra amino acids in the GafAs from Firmicutes couldn't be predicted these were excluded from the coding sequence prior to cloning of *gafACb* into pMALp5X.

The name tags correspond to the GafAs from the following bacteria:

M945_RS04880 = *Clostridium saccharobutylicum* DSM 13864, Ccel_1223 = *Clostridium cellulolyticum* H10, Cspa_c46990 = *Clostridium saccharoperbutylacetonicum* N1-4(HMT), PM3016_1312 = *Paenibacillus mucilaginosus* 3016, PMI16_00902 = *Herbaspirillum* sp. CF444

Only the first 80 amino acids are shown as the signal peptide and lipoprotein sequences are included.

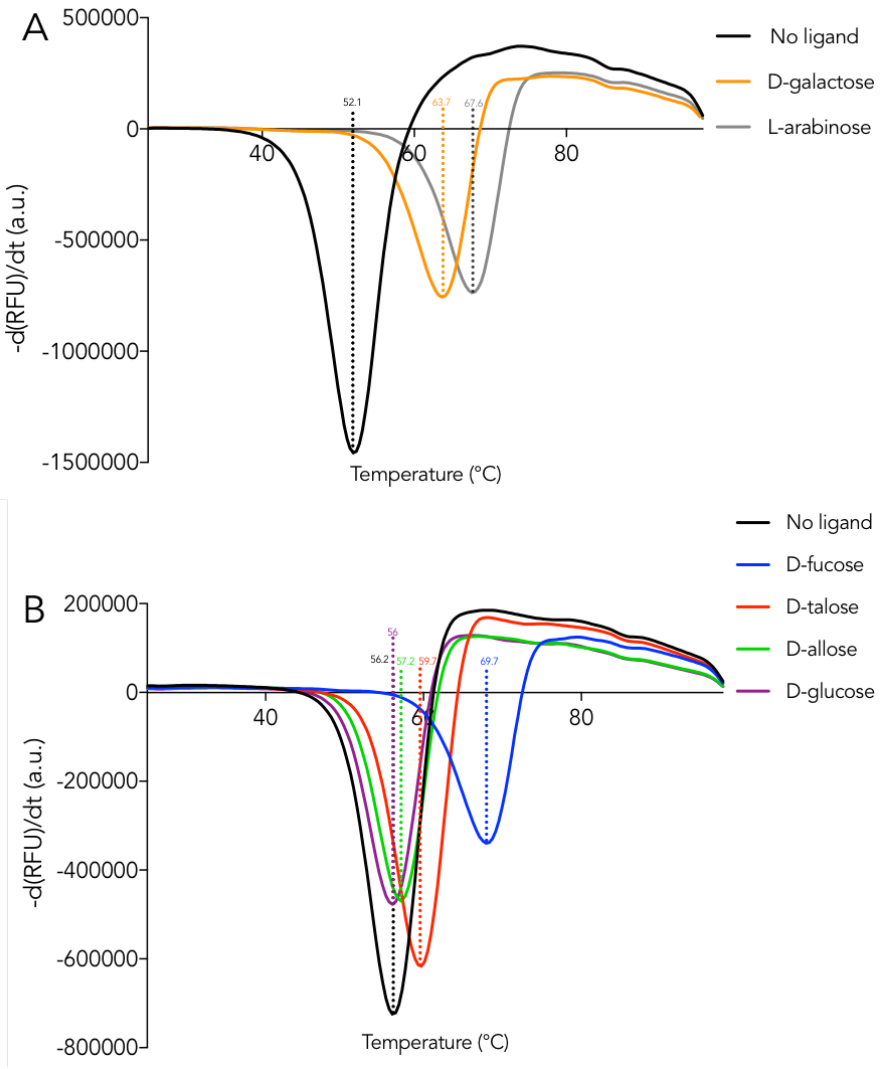


Figure A3. 1. Raw melt derivative data for the ligand specificity of GafASm

A) The raw derivative data of the DSF analysis to demonstrate the change in the Tm of the unbound form GafASm (black) compared to the Tm of the bound forms with D-galactose (orange) or L-arabinose (grey) are present in the buffer. The arabinose showed the largest Tm, which potentially indicates that it binds to the protein the strongest.

B) The raw derivative data of the DSF analysis to illustrate the change in the Tm of the unbound form GafASm (black) compared to when one of the following sugars is included in the denaturation runs: D-glucose (purple), D-allose (grey), D-talose (red) or D-fucose (blue). The highest Tm change is caused by D-fucose but when both runs are taken into account the GafASm shows the highest binding affinity for L-arabinose.

The plotted data of the thermal denaturation profiles correspond to the highest concentration of ligand tested for each sugar (*ie*. 1200 µM). The calculated Tms indicated are the average of 8 replicates for each sugar. For each replicate, the Tm is represented by the minimum value of the first derivative of the fluorescence emission plotted as a function of temperature (-dF/dt). Two different plots were produced in GraphPad Prism 7.0, as they were derived from two separate experiments due to the limited number of wells in each microplate.

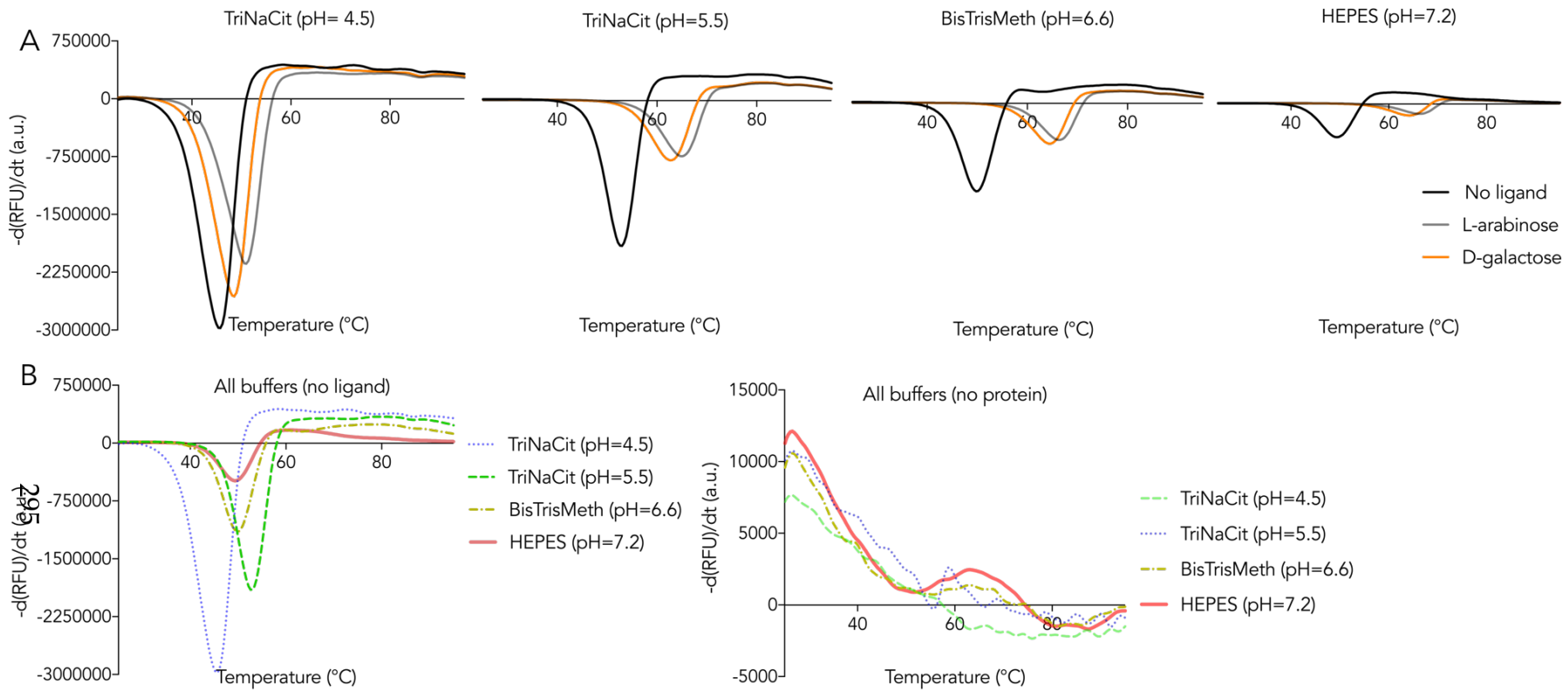


Figure A3. 2. Raw melt derivative data for screening stability and activity of GafASm in acidic conditions.

A) The raw derivative data of the DSF analysis to demonstrate the change in the Tm of the unbound form GafASm (black) compared to the Tm of the bound forms with D-galactose (orange) or L-arabinose (grey) present in the acidic buffers. The acidic conditions tested ranged from pH = 4. 5 to pH = 6. 6. The buffers used in the analysis were TrisNaCit (pH = 4.5 and 5.5) and BisTrisMeth (pH = 6.6); HEPES was used as a control to represent neutral conditions. The stability and activity of the GafASm is severely impacted at pH = 4.5, as the Tm of the protein appears decreased compared to neutral conditions.

B) (Left) The raw derivative data of the DSF analysis to illustrate the change in the Tm in the acidic pHs tested. (Right) The buffer does not have any effect on the fluorescent signal, as no sigmoidal curve is observed and the RFU units are very low.

Ligand concentration tested for each sugar is 1200 µM. The curves presented are derived from the average of 8 replicates for each temperature. For each replicate, the Tm is represented by the minimum value of the first derivative of the fluorescence emission plotted as a function of temperature (-dF/dt).

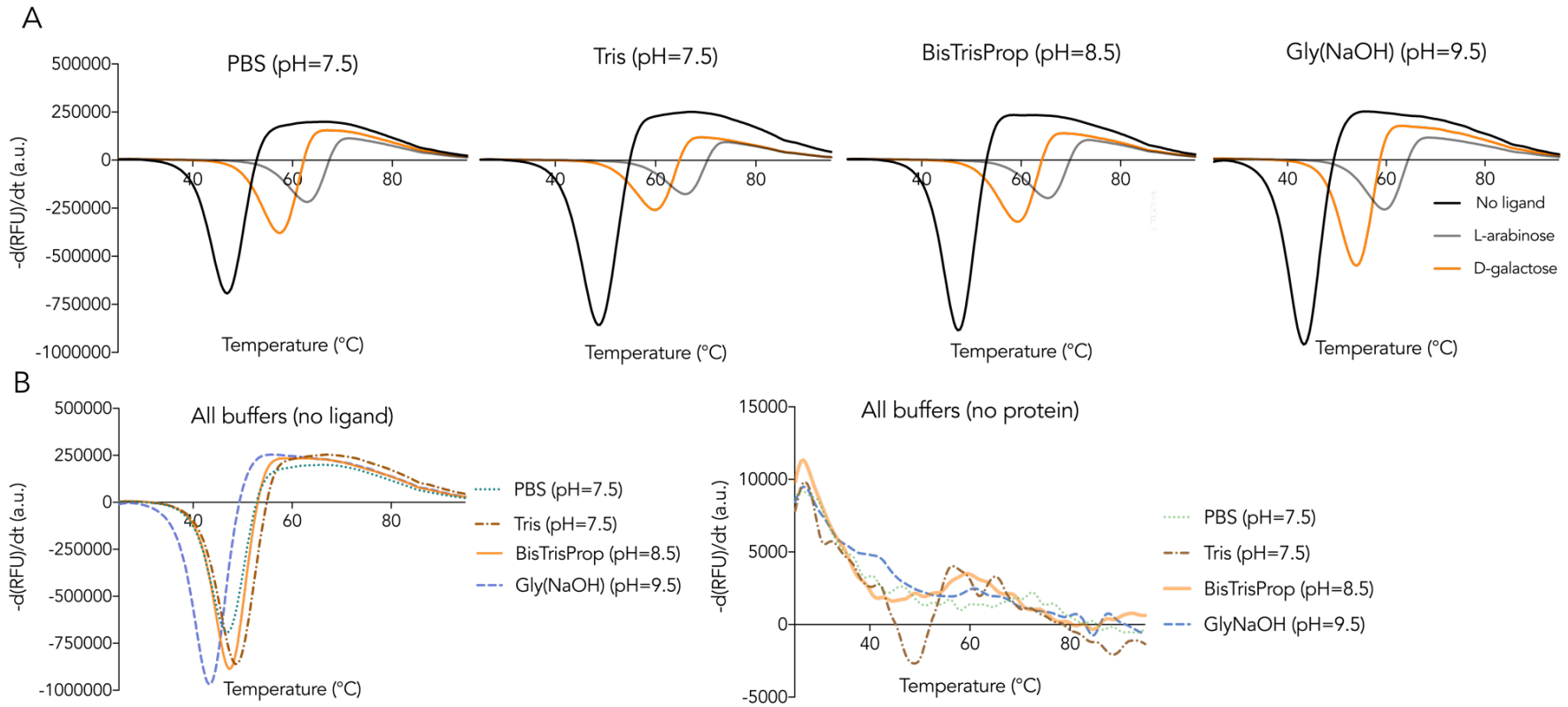


Figure S3. 3. Raw melt derivative data for screening stability and activity of GafASm in alkaline conditions.

A) The raw derivative data of the DSF analysis to demonstrate the change in the Tm of the unbound form GafASm (black) compared to the Tm of the bound forms with D-galactose (orange) or L-arabinose (grey) present in alkaline buffers. The alkaline conditions tested ranged from pH = 7. 5 to pH = 9.5. The buffers used in the analysis were PBS, Tris (both at pH = 7.5), BisTrisProp (pH = 8.5) and Glycine (NaOH) (pH = 9.5).

B) (Left) The raw derivative data of the DSF analysis to illustrate the change in the Tm in the alkaline pHs tested. (Right) Buffers don't affect the fluorescent signal, as no sigmoidal curve is observed and the RFU units are very low.

Ligand concentration tested for each sugar is 1200 µM. The curves presented are derived from the average of 8 replicates for each temperature. For each replicate, the Tm is represented by the minimum value of the first derivative of the fluorescence emission plotted as a function of temperature (-dF/dt).

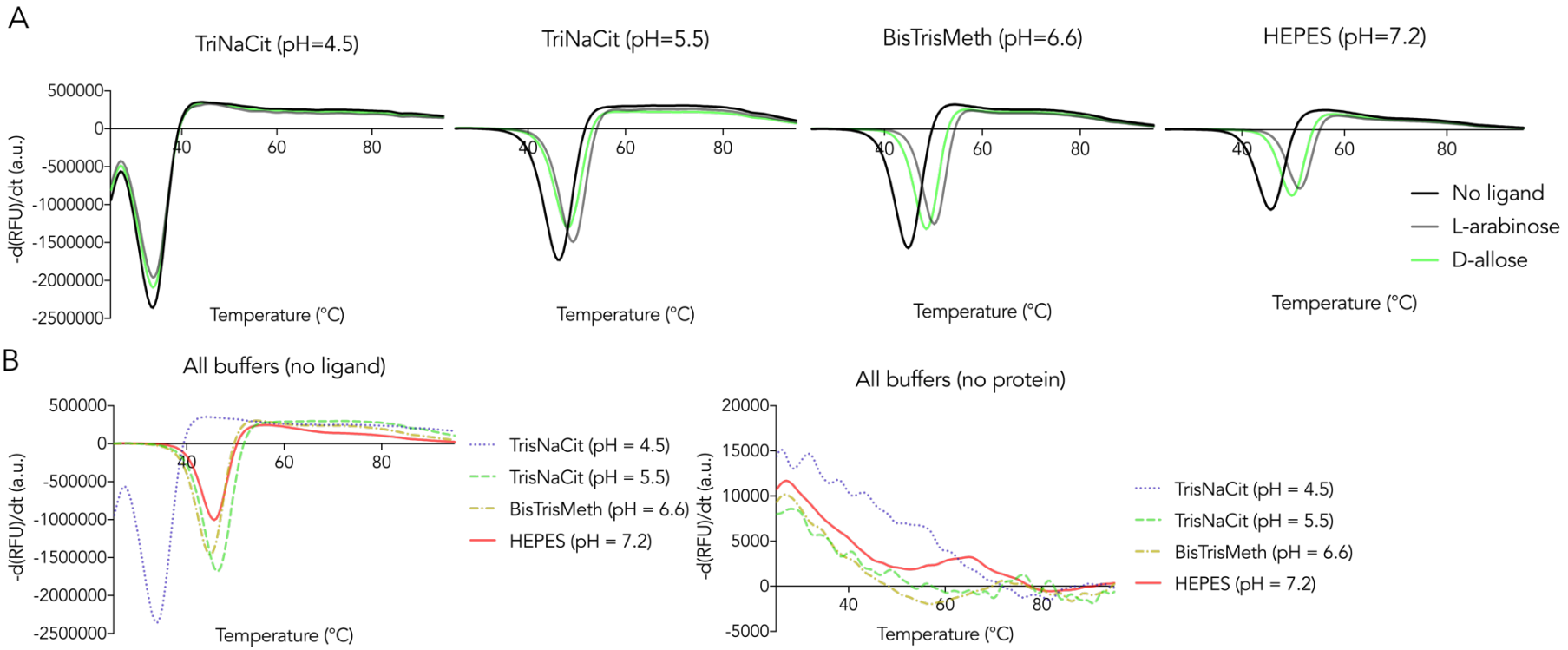


Figure A3. 4. Raw melt derivative data for screening stability and activity of GafASw in acidic conditions.

A) The raw derivative data of the DSF analysis to demonstrate the change in the Tm of the unbound form GafASw (black) compared to the Tm of the bound forms with D-allose (green) or L-arabinose (grey) present in the acidic buffers. The acidic conditions tested ranged from pH = 4. 5 to pH = 6. 6. The buffers used in the analysis were TrisNaCit (pH = 4.5 and 5.5) and BisTrisMeth (pH = 6.6); HEPES was used as a control to represent neutral conditions. The stability and activity of the GafASw is severely impacted at pH = 4.5, as the Tm of the protein appears decreased compared to neutral conditions and also shows no binding of the ligands tested.

B) (Left) The raw derivative data of the DSF analysis to illustrate the change in the Tm in the acidic pHs tested. (Right) The buffer does not have any effect on the fluorescent signal, as no sigmoidal curve is observed and the RFU units are negligible compared to when protein is supplied.

Ligand concentration tested for each sugar is 1200 µM. The curves presented are derived from the average of 8 replicates for each temperature. For each replicate, the Tm is represented by the minimum value of the first derivative of the fluorescence emission plotted as a function of temperature (-dF/dt).

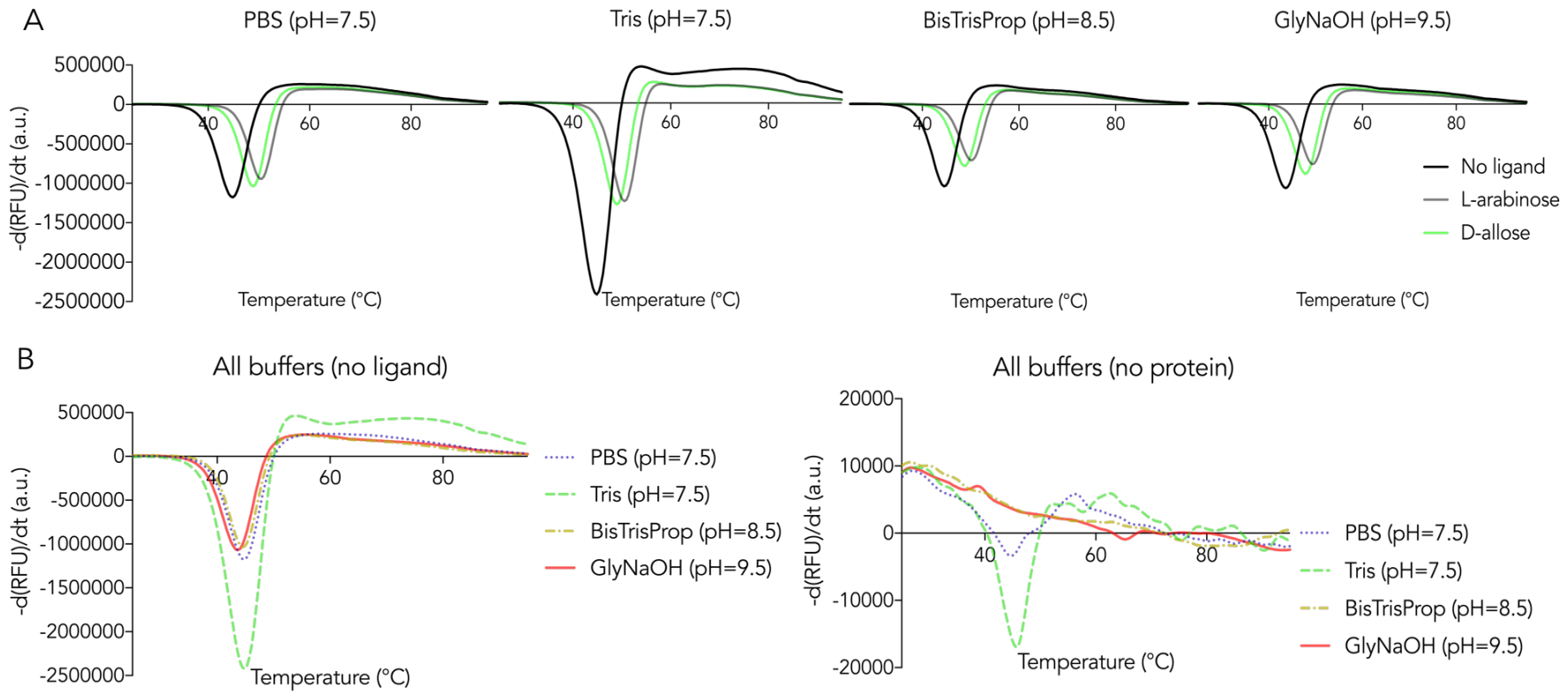


Figure A3. 5. Raw melt derivative data for screening stability and activity of GafASw in alkaline conditions.

A) The raw derivative data of the DSF analysis to demonstrate the change in the Tm of the unbound form GafASm (black) compared to the Tm of the bound forms with D-allose (green) or L-arabinose (grey) present in alkaline buffers. The alkaline conditions tested ranged from pH = 7. 5 to pH = 9.5. The buffers used in the analysis were PBS, Tris (both at pH = 7.5), BisTrisProp (pH = 8.5) and Glycine (NaOH) (pH = 9.5).

B) (Left) The raw derivative data of the DSF analysis to illustrate that there is no change in the Tm of the protein when present in the alkaline pHs tested. (Right) Buffers don't affect the fluorescent signal, as no sigmoidal curve is observed and the RFU units are negligible.

Ligand concentration tested for each sugar is 1200 µM. The curves presented are derived from the average of 8 replicates for each temperature. For each replicate, the Tm is represented by the minimum value of the first derivative of the fluorescence emission plotted as a function of temperature (-dF/dt).

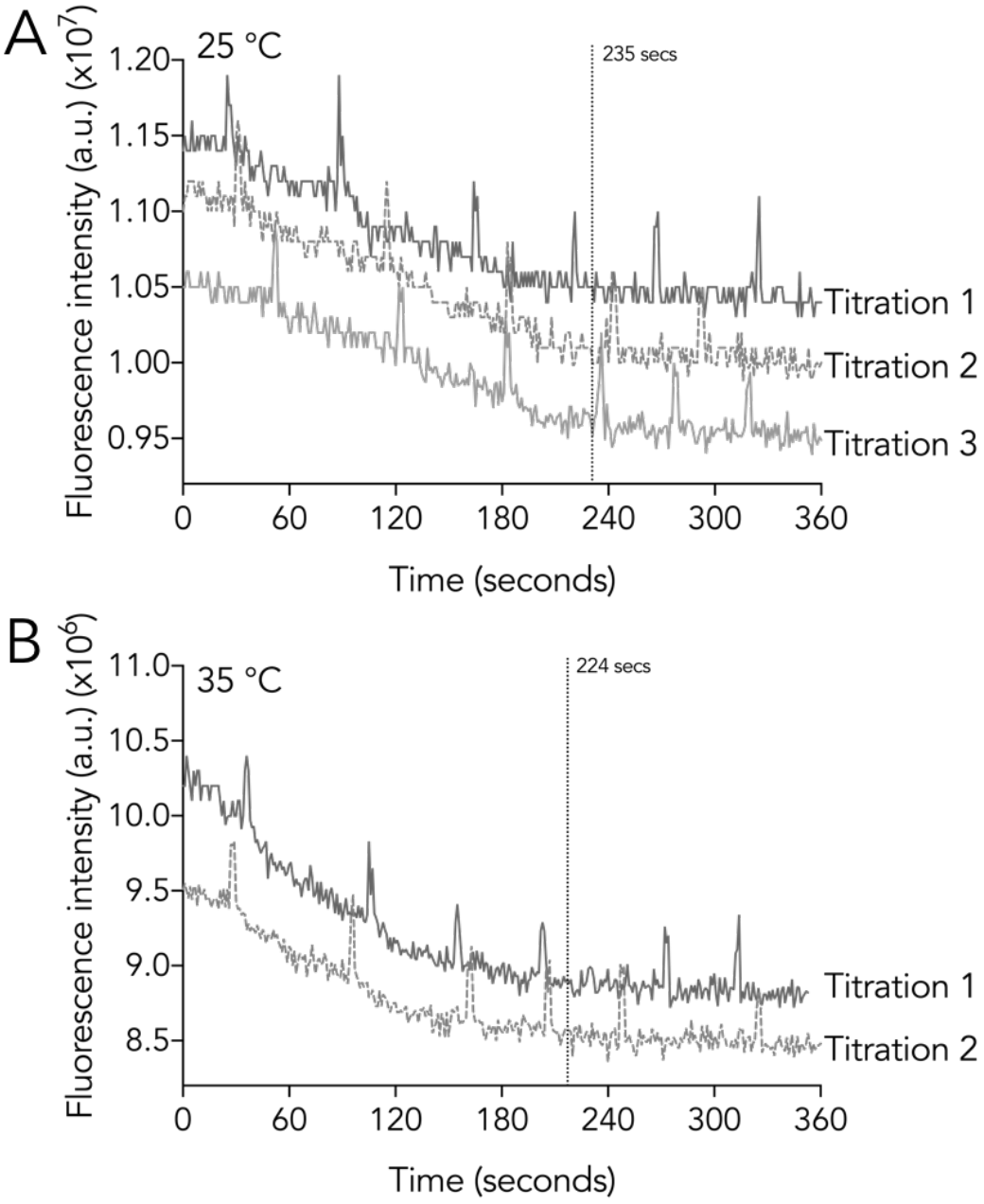


Figure A3. 6. Titration of GafASm with L-arabinose at 25 and 35 °C.

A) The intrinsic fluorescence signal produced by GafASm undergoes quenching upon addition of L-arabinofuranose at 25 °C, albeit exhibiting a drifting pattern.

B) The same run was performed at 35 ° C, which increases the furanose forms in the solution.

The concentration of the protein was 0.5 µM. Each addition added 1.25 µM L-arabinose accumulating 7.5 µM total sugar by the end of the run. The ligand additions are indicated by the peaks in signal.

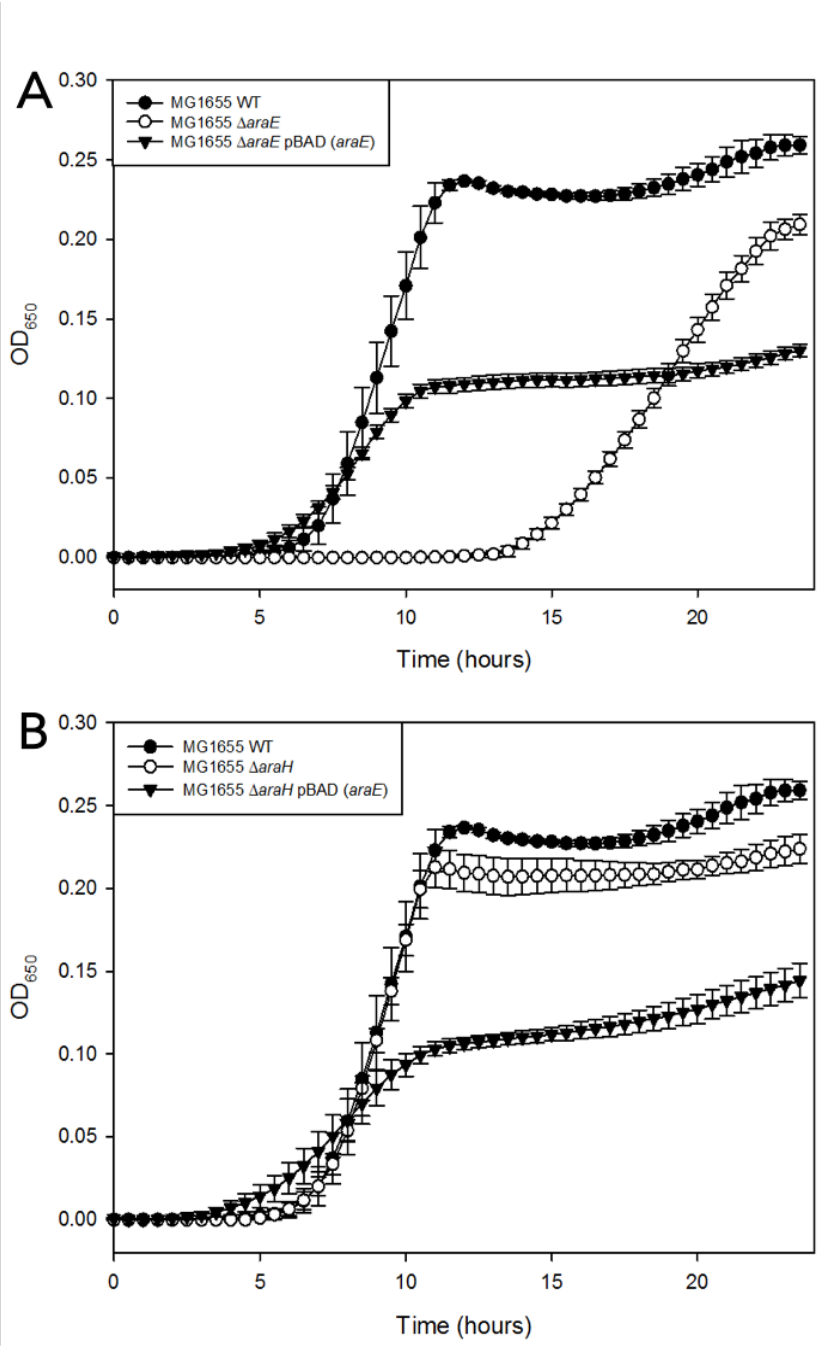


Figure A4. 1 Ectopic expression of araE from pBADcLIC in single mutants in presence of 10 mM L-arabinose.

A) Growth curve of Δ*araE* pBAD(*araE) (nabla)*. The *araE* expression was induced by 10 mM L-arabinose. The WT (*filled circles*) and Δ*araE* (*open circles*) are also shown for comparison. The cell density (OD_{650}) measurements were taken every 30 minutes across a 24-hours timecourse. Each timepoint is the average OD_{650} from 3 biological replicates and the error bars are standard errors calculated from the standard deviation.

B) The same growth assay as (A) for Δ*araH* strain. Shown in: *filled circles* = *E. coli* MG1655 WT, *nabla*= MG1655 Δ*araH* and *open circles* = MG1655 Δ*araH* pBAD (*araE*).

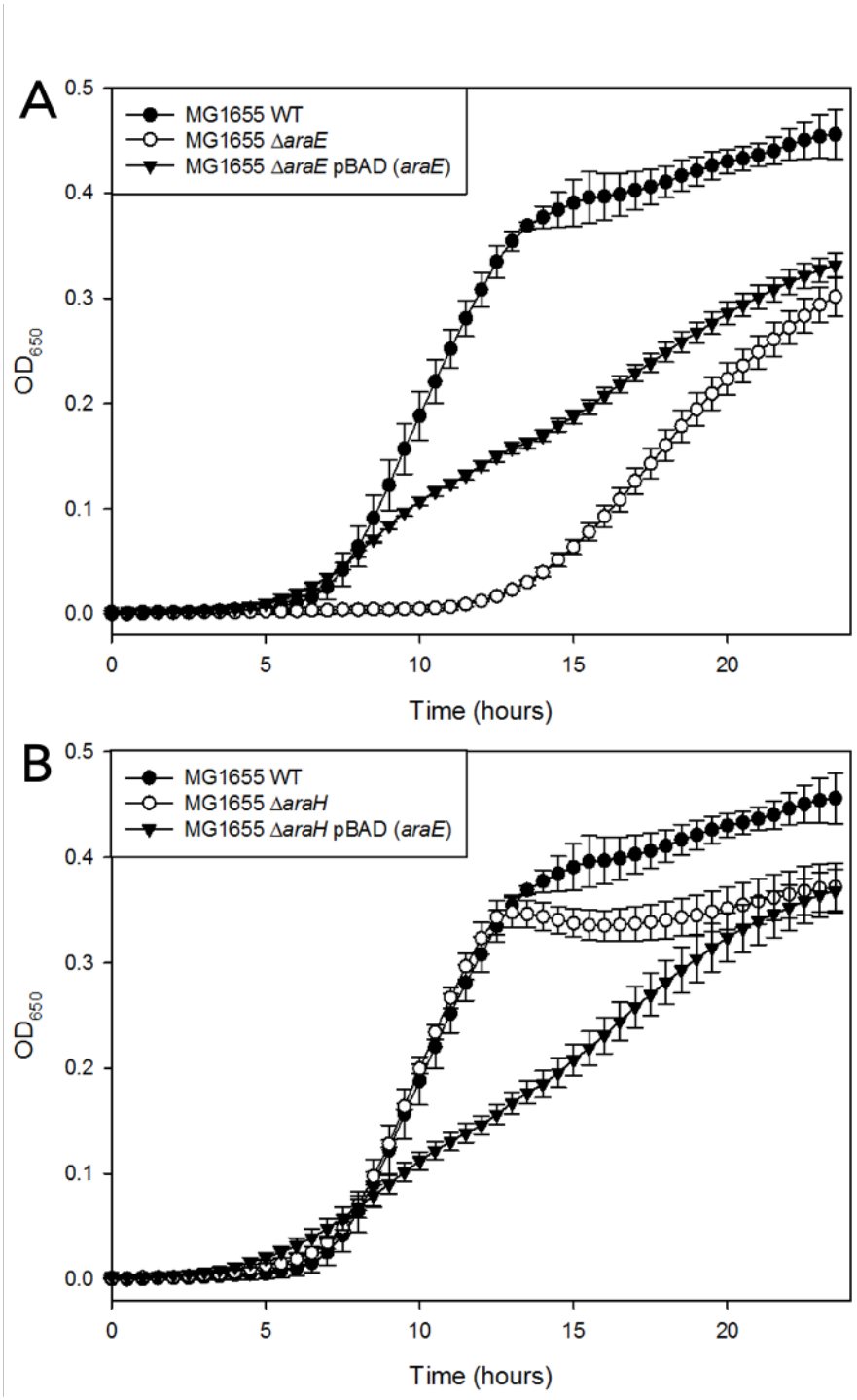


Figure A4. 2 Ectopic expression of araE from pBADcLIC in single mutants at 20 mM L-arabinose.

The growth curves presented here are the same strains and conditions presented in Figure A4. 1, however the starting concentration of L-arabinose in each well was doubled, *ie*. 20 mM.

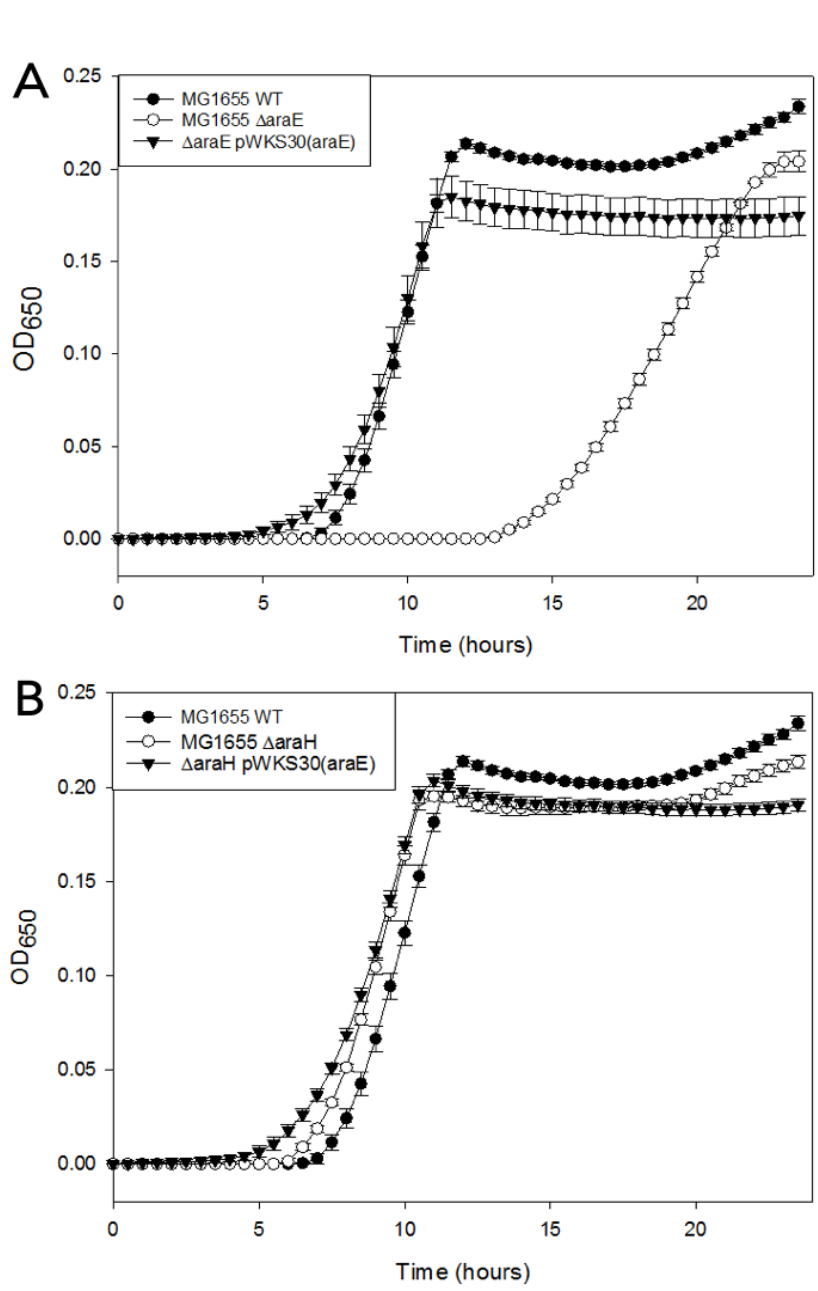


Figure A4. 3 Ectopic expression of araE from pWKS30 in single mutants at 10 mM L-arabinose.

A) Growth curve of Δ*araE* pWKS30(*araE) (nabla)*. The *araE* expression was induced with 1 mM IPTG. The WT (*filled circles*) and Δ*araE* (*open circles*) are also shown for comparison. The cell density (OD_{650}) measurements were taken every 30 minutes across a 24-hours timecourse. Each timepoint is the average OD_{650} from 3 biological replicates and the error bars are standard errors calculated from the standard deviation.

B) The same growth assay as (A) for Δ*araH* strain. Shown in: *filled circles* = *E. coli* MG1655 WT, *nabla*= MG1655 ΔaraH and *open circles* = MG1655 ΔaraH pBAD (*araE*).

Abbreviations

ABC	ATP binding cassette
ATP	Adenosine triphosphate
AG	Arabinoxylan
AGX	Arabinoglucuronoxylan
CBP	Consolidated BioProcessing
D-galf	D-galactofuranose
D-galp	D-galactopyranose
D-ribof	D-ribofuranose
D-ribop	D-ribopyranose
dsDNA	Double stranded DNA
ESI-MS	Electrospray ionization mass spectrometry
FT-ICR-MS	Fourier-transform ion cyclotron resonance mass spectrometry
IL	Ionic Liquid
L-araf	L-arabinofuranose
L-arap	L-arabinopyranose
MFS	Major Facilitator Superfamily
NBD	Nucleotide binding domain
SBP	Substrate binding protein
SEC-MALLS	Size Exclusion Chromatography -Multi-Angle Laser Light Scattering
TC	Transporter classification
TDara	Transport deletion mutant for L-arabinose
TDglcA	Transport deletion mutant for D-glucuronic acid
TDxyl	Transport deletion mutant for D-xylose
TF	Transcription factor
TMD	Transmembrane domain
TRAP	Tripartite ATP-independent periplasmic transporter
YSBL	York Structural Biology Laboratory

References

Adelsberger H, Hertel C, Glawischnig E, Zverlov VV, Schwarz WH. Enzyme system of *Clostridium stercorarium* for hydrolysis of arabinoxylan: reconstitution of the in vivo system from recombinant enzymes. *Microbiology*. 150. pp. 2257–66.

Aguedo, M., Vanderghem, C., Goffin, D., Richel, A. and Paquot, M. (2013). Fast and high yield recovery of arabinose from destarched wheat bran. Industrial Crops and Products, 43, pp.318-325.

Ahlem C., Huismal W, Neslund G and Dahmns AS. (1982). Purification and properties of a periplasmic D-xylose-binding protein from Escherichia coli K-12. J Biol Chem. 257 (6), pp. 2926-31.

Ajouz, B., Berrier, C., Garrigues, A., Besnard, M. and Ghazi, A. (1998). Release of thioredoxin via the mechanosensitive channel MscL during osmotic downshock of *Escherichia coli* cells. *J. Biol. Chem.* 273, pp. 26670–26674.

Alper, H. and Stephanopoulos, G. (2009). Engineering for biofuels: exploiting innate microbial capacity or importing biosynthetic potential?. *Nature Reviews Microbiology*, 7(10), pp.715-723.

Altschul, S. (1990). Basic Local Alignment Search Tool. Journal of Molecular Biology, 215(3), pp.403-410.

Amann, E., Ochs, B. and Abel, K. (1988). Tightly regulated tac promoter vectors useful for the expression of unfused and fused proteins in *Escherichia coli*. Gene, 69(2), pp.301-315.

Ames, G., Prody, C. and Kustu, S. (1984). Simple, Rapid, and Quantitative Release of Periplasmic Proteins by Chloroform. *Journal of Bacteriology*, 160(3), pp.1181-1183.

Anders, N., Wilkinson, M., Lovegrove, A., Freeman, J., Tryfona, T., Pellny, T., Weimar, T., Mortimer, J., Stott, K., Baker, J., Defoin-Platel, M., Shewry, P., Dupree, P. and Mitchell, R. (2012). Glycosyl transferases in family 61 mediate arabinofuranosyl transfer onto xylan in grasses. Proceedings of the National Academy of Sciences, 109(3), pp.989-993.

Andrerson L. and Garver C. John, (1973). Computer Modeling of the Kinetics of Tautomerization (Mutarotation) of Aldoses: Implications for the Mechanism of the Process. *Advances in Chemistry: Carbohydrates in Solution, Chapter 2.*, 117. pp. 20-38.

Angyal, S. (1972). Complexes of carbohydrates with metal cations. I. Determination of the extent of complexing by N.M.R. spectroscopy. *Australian Journal of Chemistry*, 25(9), p.1957.

Angyal, S. J., in R. S. Tipson and D. Horton (eds.), (1984). *Advances in Carbohydrate Chemistry and Biochemistry*, Vol. 42, Academic Press, New York, pp. 15-68.

Anisimova, M., Gascuel, O. and Sullivan, J. (2006). Approximate Likelihood-Ratio Test for Branches: A Fast, Accurate, and Powerful Alternative. Systematic Biology, 55(4), pp.539-552.

Arakawa, T., Ejima, D., Tsumoto, K., Obeyama, N., Tanaka, Y., Kita, Y. and Timasheff, S.(2007). Suppression of protein interactions by arginine: A proposed mechanism of the arginine effects. *Biophysical Chemistry*, 127(1-2), pp. 1-8.

Arakawa, T., Kita, Y. and Timasheff, S. (2007). Protein precipitation and denaturation by dimethyl sulfoxide. Biophysical Chemistry, 131(1-3), pp.62-70.

Asakura, T., Ada, K. and Schwartz, E. (1987). Stabilising effect of various organic solvents on protein. *The Journal of Biological Chemistry*, 253(18), pp.6423-6425.

Aulakh, M., Wassmann, R., Bueno, C., Kreuzwieser, J. and Rennenberg, H. (2001). Characterization of Root Exudates at Different Growth Stages of Ten Rice (*Oryza sativa L.*) Cultivars. *Plant Biology*, 3(3), pp.298-298.

Ashwell, G., Wahba, A. and Hickman, J. (1958). A new pathway of uronic acid metabolism. Biochimica et Biophysica Acta, 30(1), pp.186-187.

Bacic, A., Moody F. S., Clarke E. A. (1986) Structural Analysis of Secreted Root Slime from Maize (*Zea mays* L.). *Plant Physiology*, 80, 771-777.

Bagaria A., Kumaran D., Burley K. S., Swaminathan S., (2011). Structural basis for a ribofuranosyl binding protein: Insights into the furanose specific transport. *Proteins (Structure, Function and Bioinformatics),* 79(4), pp. 1352-1357.

Bagos, P., Tsirigos, K., Liakopoulos, T. and Hamodrakas, S. (2008). Prediction of Lipoprotein Signal Peptides in Gram-Positive Bacteria with a Hidden Markov Model. Journal of Proteome Research, 7(12), pp.5082-5093.

Bailey, J. (1991)9. Toward a science of metabolic engineering. *Science*, 252(5013), pp.1668-1675.

Bhattacharjee, S. and Timell, T. (1965). A STUDY OF THE PECTIN PRESENT IN THE BARK OF AMABILIS FIR (ABIES AMABILIS). *Canadian Journal of Chemistry*, 43(4), pp.758-765.

Balan, V. (2014). Current Challenges in Commercially Producing Biofuels from Lignocellulosic Biomass. *ISRN Biotechnology*, 2014, pp.1-31.

Baldwin R. L. (1986). Temperature dependence of the hydrophobic interaction in protein folding. *Proc. Natl. Acad. Sci. U.S.A.* 83, pp.8069–8072

Beg QK, Kapoor M, Mahajan L and Hoondal GS. (2001). Microbial xylanases and their industrial applications: a review. Appl Microbiol Biotechnol, 56, pp.326–338

Begley GS, Hansen DE, Jacobson GR, Knowles JR. (1982) Stereochemical course of the reactions catalyzed by the bacterial phosphoenolpyruvate:glucose phosphotransferase system. *Biochemistry*. 21, pp. 5552–5556.

Beisel, C. and Afroz, T. (2015). Rethinking the Hierarchy of Sugar Utilization in Bacteria. *Journal of Bacteriology*, 198(3), pp.374-376.

Berg, E. (2015). A New Spin On The Old Gram Stain. *Chemical and Engineering News*, (ISSN 0009-2347).

Berg, H. (1980). Concepts of biochemistry. *Bioelectrochemistry and Bioenergetics*, 7(4), p.823.

Berman, H., Henrick, K., & Nakamura, H. (2003). Announcing the worldwide protein data bank. Nature *Structural & Molecular Biology*, 10(12), 980-980.

Berntsson, R., Smits, S., Schmitt, L., Slotboom, D. and Poolman, B. (2010). A structural classification of substrate-binding proteins. *FEBS Letters*, 584(12), pp. 2606-2617.

Berrier, C., Garrigues, A., Richarme, G. and Ghazi, A. (2000). Elongation factor Tu and DnaK are transferred from the cytoplasm to the periplasm of *Escherichia coli* during osmotic downshock presumably via the mechanosensitive channel MscL. *J. Bacteriol*. 182, pp. 248–251.

Beynon, R. (1988). A Macintosh Hypercard stack for calculation of thermodynamically-corrected buffer recipes. *Bioinformatics*, 4(4), pp.487-490.

Biemans-Oldehinkel, E., Doeven, M. and Poolman, B. (2005). ABC transporter architecture and regulatory roles of accessory domains. *FEBS Letters*, 580(4), pp.1023-1035.

Blackwell, R. (1954). Verification of the Henderson-Hasselbalch equation. *Journal of Chemical Education*, 31(3), p.138.

Blattner, F. (1997). The Complete Genome Sequence of *Escherichia coli* K-12. *Science*, 277(5331), 1453-1462.

Blast.ncbi.nlm.nih.gov. (2017). Protein BLAST: search protein databases using a protein query. [online] Available at: https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE=Proteins

Bokinsky, G., Peralta-Yahya, P., George, A., Holmes, B., Steen, E., Dietrich, J., Soon Lee, T., Tullman-Ercek, D., Voigt, C., Simmons, B. and Keasling, J. (2011). Synthesis of three advanced biofuels from ionic liquid-pretreated switchgrass using engineered *Escherichia coli*. *Proceedings of the National Academy of Sciences*, 108(50), pp.19949-19954.

Boland, J., Davidson, P., Weiss, J. and Bruce, B. (2004). Cations Reduce Antimicrobial Efficacy of Lysozyme-Chelator Combinations. *Journal of Food Protection*, 67(2), pp.285–294.

Borths, E. L., Locher, K. P., Lee, A. T. & Rees, D. C. 2002. The structure of Escherichia coli BtuF and binding to its cognate ATP binding cassette transporter. *Proc. Natl. Acad. Sci.* USA, 99, 16642-7.

Bourret, R. (1995). Two-component signal transduction. Trends in Microbiology, 3(12), pp.490-491.

Braun, V. and Herrmann, C. (2007). Docking of the Periplasmic FecB Binding Protein to the FecCD Transmembrane Proteins in the Ferric Citrate Transport System of *Escherichia coli*. *Journal of Bacteriology*, 189(19), 6913-6918.

Braun, V. and Wu, H. (1994). Chapter 14 Lipoproteins, structure, function, biosynthesis and model for protein export. *New Comprehensive Biochemistry*, 27, pp.319 - 341.

Bringhurst, R., Cardon, Z. and Gage, D. (2001). Galactosides in the rhizosphere: Utilization by Sinorhizobium meliloti and development of a biosensor. *Proceedings of the National Academy of Sciences*, 98(8), pp.4540-4545.

Brodeur, G., Yau, E., Badal, K., Collier, J., Ramachandran, K. and Ramakrishnan, S. (2011). Chemical and Physicochemical Pretreatment of Lignocellulosic Biomass: A Review. *Enzyme Research*, 2011, pp.1-17.

Bullock, A., Debreczeni, J., Fedorov, O., Nelson, A., Marsden, B. and Knapp, S. (2005). Structural Basis of Inhibitor Specificity of the Human Protooncogene Proviral Insertion Site in Moloney Murine Leukemia Virus (PIM-1) Kinase. *Journal of Medicinal Chemistry*, 48(24), pp.7604-7614.

Camilloni, C., Guerini Rocco, A., Eberini, I., Gianazza, E., Broglia, R. and Tiana, G. (2008). Urea and Guanidinium Chloride Denature Protein L in Different Ways in Molecular Dynamics Simulations. *Biophysical Journal*, 94(12), pp.4654-4661.

Cantarel, B., Coutinho, P., Rancurel, C., Bernard, T., Lombard, V. and Henrissat, B. (2009). The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. *Nucleic Acids Research*, 37(Database), pp. D233-D238.

Carpita, N. (1996). STRUCTURE AND BIOGENESIS OF THE CELL WALLS OF GRASSES. *Annual Review of Plant Physiology and Plant Molecular Biology*, 47(1), pp.445-476.

Cerny, G. and Teuber, M. (1971). Differential release of periplasmic versus cytoplasmic enzymes from *Escherichia coli B* by polymyxin B. *Archiv for Mikrobiologie*, 78(2), pp.166-179.

Chaboud, A. (1983). Isolation, purification and chemical composition of maize root cap slime. *Plant and Soil*, 73(3), pp.395-402.

Chaboud, A. and Rougier, M. (1984). Identification and Localization of Sugar Components of Rice (Oryza sativa L.) Root Cap Mucilage. *Journal of Plant Physiology*, 116(4), pp.323-330.

Chakdar, H., Kumar, M., Pandiyan, K., Singh, A., Nanjappan, K., Kashyap, P. and Srivastava, A. (2016). Bacterial xylanases: biology to biotechnology. *3 Biotech*, 6(2).

Chakravartty, V. and Cronan, J. (2015). A series of medium and high copy number arabinose-inducible *Escherichia coli* expression vectors compatible with pBR322 and pACYC184. Plasmid, 81, pp.21-26.

Chang, A., Lin, R., Studley, W., Tran, C. and Saier, Jr, M. (2004). Phylogeny as a guide to structure and function of membrane transport proteins (Review). *Molecular Membrane Biology*, 21(3), pp.171-181.

Chaudhuri BN, Ko J, Park C, Jones TA, Mowbray SL. (1999). Structure of D-allose binding protein from *Escherichia coli* bound to D-allose at 1.8 Å resolution. *J Mol Biol*, 286, pp. 1519–1531.

Chen H., (2014). Chemical composition and structure of natural lignocellulose. *Biotechnology of Lignocellulose*, pp. 25-71.

Chen, R. F. (1972) Measurements of absolute values in biochemical fluorescence spectroscopy. J. *Research National Bureau Standards*. 76A (6), pp. 593-606.

Chen, Y. C., Chen, S. J., Chang, M. C., & Chen, T. L. (2005). Comparison of various methods for periplasmic release of recombinant creatinase from *Escherichia coli*. *Journal of the Chinese Institute of Chemical Engineers*, 36(5), 527-532.

Chen, Y., Chen, L., Chen, S., Chang, M. and Chen, T. (2004). A modified osmotic shock for periplasmic release of a recombinant creatinase from *Escherichia coli*. *Biochemical Engineering Journal*, 19(3), pp.211-215.

Chen, Y., Sharma-Shivappa, R., Keshwani, D. and Chen, C. (2007). Potential of Agricultural Residues and Hay for Bioethanol Production. Applied Biochemistry and Biotechnology, 142(3), pp.276-290.

Chow, V., Shantharaj, D., Guo, Y., Nong, G., Minsavage, G., Jones, J. and Preston, J. (2015). Xylan Utilization Regulon in Xanthomonas citri pv. citri Strain 306: Gene Expression and Utilization of Oligoxylosides. *Applied and Environmental Microbiology*, 81(6), pp.2163-2172.

Chuang, C., Chen, L., Fu, R., Chen, S., Ho, M., Huang, J., Hsu, C., Wang, C., Chen, M. and Tsai, R. (2014). Involvement of the Carboxyl-Terminal Region of the Yeast Peroxisomal Half ABC Transporter Pxa2p in Its Interaction with Pxa1p and in Transporter Function. *PLoS ONE*, 9(8), p.e104892.

Collins P., Ferrier R. (1995) Monosaccharides: Their Chemistry and Their Roles in Natural Products, Wiley

Collins T, Feller G, Gerday C, Meuwis MA (2012) Family 8 enzymes with xylanolytic activity. US Patent 8309336

Collins T, Hoyoux A, Dutron A, Georis J, Genot B, Dauvrin T, Arnaut F, Gerday C, Feller G (2006) Use of glycoside hydrolase family 8 xylanases in baking. *J Cereal Sci* 43, pp. 79–84

Cummings, M. D., Farnum, M. A. & Nelen, M. I. (2006). Universal screening methods and applications of ThermoFluor. *J Biomol Screen*, 11, pp. 854-63

Conway T, Sewell GW, Osman YA, Ingram LO. (1987). Cloning and sequencing of the alcohol dehydrogenase-II gene from *Zymomonas mobilis*. *J Bacteriol*. 169(6). Pp. 2591–2597.

D M Miller, 3rd, J S Olson, J W Pflugrath and F A Quiocho. (1983). Rates of ligand binding to periplasmic proteins involved in bacterial transport and chemotaxis. *J. Biol. Chem.* 258, pp. 13665-72.

Daiguan Yu, Hilary M. Ellis, E-Chiang Lee, Nancy A. Jenkins, Neal G. Copeland, and Donald L. Court. (2000). An efficient recombination system for chromosome engineering in *Escherichia coli*. *PNAS*, 97 (11), pp. 5978-5983

Dantas, G., Legey, L. and Mazzone, A. (2013). Energy from sugarcane bagasse in Brazil: An assessment of the productivity and cost of different technological routes. Renewable and Sustainable Energy Reviews, 21, pp.356-364.

Datta, S., Costantino, N. & Court, D.L. (2006). A set of recombineering plasmids for gram-negative bacteria. *Gene*, 379, pp. 109–115.

DAVIDSON, A. L., SHUMAN, H. A. & NIKAIDO, H. (1992). Mechanism of maltose transport in *Escherichia coli*: transmembrane signaling by periplasmic binding proteins. *Proc. Natl. Acad. Sci. USA*, 89, 2360-2364.

Davis, E.O., Henderson, P.J., 1987. The cloning and DNA sequence of the gene *xylE* for xylose-proton symport in *Escherichia coli* K12. *J. Biol. Chem*. 262, pp. 13928–13932.

Daus, M., Berendt, S., Wuttge, S., and Schneider, E. (2007). Maltose binding protein (MalE) interacts with periplasmic loops P2 and P1 respectively of the MalFG subunits of the maltose ATP binding cassette transporter (MalFGK2) from *Escherichia coli*/*Salmonella* during the transport cycle. *Molecular Microbiology*, 66(5), 1107-1122.

Dawson RJ, Hollenstein K, Locher KP. (2007). Uptake or extrusion: Crystal structures of full ABC transporters suggest a common mechanism. *Mol Microbiol*

de O. Buanafina, M. (2009). Feruloylation in Grasses: Current and Future Perspectives. Molecular Plant, 2(5), pp.861-872.

de Vries, R. P., and Visser, J. (2001). *Aspergillus* enzymes involved in degradation of plant cell wall polysaccharides, *Microbiol. Mol. Biol. Rev*. 65(4), pp. 497-522

DEAN, D. A., DAVIDSON, A. L. & NIKAIDO, H. (1989). Maltose transport in membrane vesicles of *Escherichia coli* is linked to ATP hydrolysis. *Proc. Natl. Acad. Sci*. USA, 86, 9134-9138.

Death, A. and Ferenci, T. (1994). Between feast and famine: endogenous inducer synthesis in the adaptation of *Escherichia coli* to growth with limiting carbohydrates. *Journal of Bacteriology*, 176(16), pp.5101-5107.

DeBoy, R., Mongodin, E., Fouts, D., Tailford, L., Khouri, H., Emerson, J., Mohamoud, Y., Watkins, K., Henrissat, B., Gilbert, H. and Nelson, K. (2008). Insights into Plant Cell Wall Degradation from the Genome Sequence of the Soil Bacterium Cellvibrio japonicus. Journal of Bacteriology, 190(15), pp.5455-5463.

Declerck, N. and Abelson, J. (1994). Novel substrate specificity engineered in the arabinose binding protein. "*Protein Engineering, Design and Selection*", 7(8), pp.997-1004.

Deb, A., Johnson, W., Kline, A., Scott, B., Meador, L., Srinivas, D., Martin-Garcia, J., Dörner, K., Borges, C., Misra, R., Hogue, B., Fromme, P. and Mor, T. (2017). Bacterial expression, correct membrane targeting and functional folding of the HIV-1 membrane protein Vpu using a periplasmic signal peptide. *PLOS ONE*, 12(2), p.e0172529.

Denger, K., Weiss, M., Felux, A., Schneider, A., Mayer, C., Spiteller, D., Huhn, T., Cook, A. and Schleheck, D. (2014). Sulphoglycolysis in Escherichia coli K-12 closes a gap in the biogeochemical sulphur cycle. *Nature*, 507(7490), pp.114-117.

Derde, M., Lechevalier, V., Guérin-Dubiard, C., Cochet, M., Jan, S., Baron, F., Gautier, M., Vié, V. and Nau, F. (2013). Hen Egg White Lysozyme Permeabilizes *Escherichia coli* Outer and Inner Membranes. *Journal of Agricultural and Food Chemistry*, 61(41), pp.9922-9929.

Dereeper, A., Guignon, V., Blanc, G., Audic, S., Buffet, S., Chevenet, F., Dufayard, J., Guindon, S., Lefort, V., Lescot, M., Claverie, J. and Gascuel, O. (2008). Phylogeny.fr: robust phylogenetic analysis for the non-specialist. Nucleic Acids Research, 36(Web Server), pp.W465-W469.

Dinesh, B., Lau, N., Furusawa, G., Kim, S., Taylor, T., Foong, S. and Shu-Chien, A. (2016). Comparative genome analyses of novel Mangrovimonas-like strains isolated from estuarine mangrove sediments reveal xylan and arabinan utilization genes. *Marine Genomics*, 25, pp.115-121.

Directive 2009/28/EC of the *European Parliament* and of the Council of 23 April 2009 on the promotion of the use of energy from renewable sources (amendment and subsequently repeal of Directives 2001/77/EC and 2003/30/EC)

Dixon, R. and Chopra, I. (1986). Leakage of periplasmic proteins from Escherichia coli mediated by polymyxin B nonapeptide. *Antimicrobial Agents and Chemotherapy*, 29(5), pp.781-788.

Donohue-Rolfe, A., & Keusch, G. T. (1983). Shigella dysenteriae 1 cytotoxin: periplasmic protein releasable by polymyxin B and osmotic shock. *Infection and Immunity*, 39(1), pp.270–274.

Dreyfus J., (1964). Characterisation of a Sulfate and Thiosulfate-transporting system in *Salmonella typhimurium*.

Dunn T.M., Schleif R. (1984). Deletion Analysis of the Escherichia coli ara PC and PBAD Promoters. *J.Mol. Biol*. 180, pp. 201-204.

Ebringerová A, Heinze T (2000) Xylan and xylan derivatives–biopolymers with valuable properties, 1. Naturally occurring xylans structures, isolation procedures and properties. *Macromol. Rapid Commun.* 21, pp. 542–556.

Ebringerová, A. and Heinze, T. (2000). Xylan and xylan derivatives – biopolymers with valuable properties, 1. Naturally occurring xylans structures, isolation procedures and properties. *Macromolecular Rapid Communications*, 21(9), pp.542-556.

Ekborg, N. (2005). Saccharophagus degradans gen. nov., sp. nov., a versatile marine degrader of complex polysaccharides. INTERNATIONAL JOURNAL OF SYSTEMATIC AND EVOLUTIONARY MICROBIOLOGY, 55(4), pp.1545-1549.

El Yaagoubi, A., Kohiyama, M. and Richarme, G. (1994). Localization of DnaK (Chaperone 70) from Escherichia coli in an osmotic-shock-sensitive compartment of the cytoplasm. J. Bacteriol. 176, pp. 7074–7078.

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

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

Englesberg, E., Anderson, R., Lee, N., Weinberg, R., Hoffee, P., Huttenhauer, G. and Boyer, H. (1962). L-Arabinose-sensitive, L-ribulose 5-phosphate 4-epimerase-deficient mutants of *Escherichia coli. Journal of Bacteriology*, 84, pp.137-146.

Ericsson, U., Hallberg, B., DeTitta, G., Dekker, N. and Nordlund, P. (2006). Thermofluor-based high-throughput stability optimization of proteins for structural studies. *Analytical Biochemistry*, 357(2), pp.289-298.

Evans, P. R. (2011). An introduction to data reduction: space-group determination, scaling and intensity statistics. *Acta Crystallographica Section D: Biological Crystallography*, 67(4), pp. 282-292.

Evans, P. R., & Murshudov, G. N. (2013). How good are my data and what is the resolution? *Acta Crystallographica Section D: Biological Crystallography*, 69(7), pp. 1204-1214.

Fargione, J., Hill, J., Tilman, D., Polasky, S. and Hawthorne, P. (2008). Land Clearing and the Biofuel Carbon Debt. Science, 319(5867), pp.1235-1238.

Fasman, G. D., Editor (1976) Handbook of Biochemistry and Molecular Biology, 3rd Edition, Proteins, Volume I, pp. 183-203, CRC Press.

Felsenstein, J. (1985). Confidence Limits on Phylogenies: An Approach Using the Bootstrap. Evolution, 39(4), p.783.

Fischer, M., Hopkins, A., Severi, E., Hawkhead, J., Bawdon, D., Watts, A., Hubbard, R. and Thomas, G. (2015). Tripartite ATP-independent Periplasmic (TRAP) Transporters Use an Arginine-mediated Selectivity Filter for High Affinity Substrate Binding. *Journal of Biological Chemistry*, 290(45), pp.27113-27123.

Flint, H., Scott, K., Duncan, S., Louis, P., and Forano, E. (2012). Microbial degradation of complex carbohydrates in the gut. *Gut Microbes*, 3(4), 289-306.

Florence, T. (1980). Degradation of protein disulphide bonds in dilute alkali. *Biochemical Journal*, 189(3), pp.507-520.

Forward J.A., Behrendt M. C., Wyborn N. R., Cross R and Kelly D. J., (1997) TRAP transporters: a new family of periplasmic solute transport systems encoded by the dctPQM genes of *Rhodobacter capsulatus* and by homologs in diverse Gram-negative bacteria. *J. Bacteriol*. 179, 5482–5493

Franks, F., Lillford, P. and Robinson, G. (1989). Isomeric equilibria of monosaccharides in solution. Influence of solvent and temperature. *Journal of the Chemical Society, Faraday Transactions 1: Physical Chemistry in Condensed Phases*, 85(8), p.2417.

Fuhrmann, M., Hausherr, A., Ferbitz, L., Schödl, T., Heitzer, M. and Hegemann, P. (2004). Monitoring dynamic expression of nuclear genes in *Chlamydomonas reinhardtii* by using a synthetic luciferase reporter gene. *Plant Molecular Biology*, 55(6), pp.869-881.

Fukada, H., Sturtevant, J., and Quiocho, F. 1983. Thermodynamics of the binding of L-arabinose and of D-galactose to the L-arabinose-banding protein of *Escherichia coli*. *J. Biol. Chem*., 258, pp. 13193-13198.

Fukami-Kobayashi, K., Tateno, Y. and Nishikawa, K. (1999) Domain dislocation: a change of core structure in periplasmic binding proteins in their evolutionary history. *J. Mol. Biol*. 286, pp. 279–290.

Gárdonyi, M., Jeppsson, M., Lidén, G., Gorwa-Grauslund, M. and Hahn-Hägerdal, B. (2003). Control of xylose consumption by xylose transport in recombinant *Saccharomyces cerevisiae*. *Biotechnology and Bioengineering*, 82(7), pp.818-824.

Geddes, B. and Oresnik, I. (2012). Inability To Catabolize Galactose Leads to Increased Ability To Compete for Nodule Occupancy in *Sinorhizobium meliloti*. *Journal of Bacteriology*, 194(18), pp.5044-5053.

Gerhardt, P. 1981. Manual of methods for general microbiology. American Society for Microbiology. Washington, DC, USA.

Ghisaidoobe, A. and Chung, S. (2014). Intrinsic Tryptophan Fluorescence in the Detection and Analysis of Proteins: A Focus on Förster Resonance Energy Transfer Techniques. *International Journal of Molecular Sciences*, 15(12), pp.22518-22538.

Giordano, T., Deuschle, U., Bujard, H. and McAllister, W. (1989). Regulation of coliphage T3 and T7 RNA polymerases by the lac represser-operator system. Gene, 84(2), pp.209-219.

Giuliani E., Frank M. A, Corgliano M. D., Seifert C., Loren H., Collar R. F., (2011) Environment sensing and response mediated by ABC transporters. *BMC Genomics*, 12(S8), pp. 1-14.

Goldbeck, R., Damásio, A., Gonçalves, T., Machado, C., Paixão, D., Wolf, L., Mandelli, F., Rocha, G., Ruller, R. and Squina, F. (2014). Development of hemicellulolytic enzyme mixtures for plant biomass deconstruction on target biotechnological applications. Applied Microbiology and Biotechnology, 98(20), pp.8513-8525.

Golding,I. and Cox,E. (2006) Physical Nature of Bacterial Cytoplasm. Physical Review Letters, 96.

Golyshina, O., Golyshin, P., Timmis, K. and Ferrer, M. (2006). The 'pH optimum anomaly' of intracellular enzymes of *Ferroplasma acidiphilum*. *Environmental Microbiology*, 8(3), pp.416-425.

Gonin, S., Arnoux, P., Pierru, B., Lavergne, J., Alonso, B., Sabaty, M. and Pignol, D. (2007). Crystal structures of an Extracytoplasmic Solute Receptor from a TRAP transporter in its open and closed forms reveal a helix-swapped dimer requiring a cation for α-keto acid binding. *BMC Structural Biology*, 7(1), p.11.

González, G., López-Santín, J., Caminal, G. and Solà, C. (1986). Dilute acid hydrolysis of wheat straw hemicellulose at moderate temperature: A simplified kinetic model. Biotechnology and Bioengineering, 28(2), pp.288-293.

Good, N., Winget, G., Winter, W., Connolly, T., Izawa, S. and Singh, R. (1966). Hydrogen Ion Buffers for Biological Research. *Biochemistry*, 5(2), pp.467-477.

Graham, P., Viteri, S., Mackie, F., Vargas, A. and Palacios, A. (1982). Variation in acid soil tolerance among strains of Rhizobium phaseoli. *Field Crops Research*, 5, pp.121-128.

Greenfield, N. (2007). Using circular dichroism spectra to estimate protein secondary structure. *Nature Protocols*, 1(6), pp.2876-2890.

Griffith, J., Baker, M., Rouch, D., Page, M., Skurray, R., Paulsen, I., Chater, K., Baldwin, S. and Henderson, P. (1992). Membrane transport proteins: implications of sequence comparisons. *Current Opinion in Cell Biology*, 4(4), pp.684-695.

Groeneveld, P., Stouthamer, A. and Westerhoff, H. (2008). Super life - how and why 'cell selection' leads to the fastest-growing eukaryote. *FEBS Journal*, 276(1), pp.254-270.

Grohmann, K. and Bothast, R. (1997). Saccharification of corn fibre by combined treatment with dilute sulphuric acid and enzymes. Process Biochemistry, 32(5), pp.405-415.

Grote, A., Hiller, K., Scheer, M., Munch, R., Nortemann, B., Hempel, D. and Jahn, D. (2005). JCat: a novel tool to adapt codon usage of a target gene to its potential expression host. Nucleic Acids Research, 33(Web Server), pp.W526-W531.

Gubellini, F., Verdon, G., Karpowich, N., Luff, J., Boël, G., Gauthier, N., Handelman, S., Ades, S. and Hunt, J. (2011). Physiological Response to Membrane Protein Overexpression in *E. coli*. *Molecular & Cellular Proteomics*, 10(10), pp.M111.007930.

Guindon, S., Lethiec, F., Duroux, P. and Gascuel, O. (2005). PHYML Online--a web server for fast maximum likelihood-based phylogenetic inference. Nucleic Acids Research, 33(Web Server), pp.W557-W559.

Guzman, L., Belin, D., Carson, M. and Beckwith, J. (1995). Tight regulation, modulation, and high-level expression by vectors containing the arabinose PBAD promoter. Journal of Bacteriology, 177(14), pp.4121-4130.

H. C. Neu, L. A Heppel. (1965). The release of enzymes from Escherichia coli by osmotic shock during the formation of spheroplasts. *J. Biol. Chem*, 240, pp. 3685-3692.

H. Conner, A. and Anderson, L. (1972). The tautomerization and mutarotation of β-L-arabinopyranose. Participation of both furanose anomers. *Carbohydrate Research*, 25(1), pp.107-116.

H. Zhang, W. Cui, M. L. Gross, R. E. Blankenship, Native mass spectrometry of photosynthetic pigment-protein complexes, *FEBS Lett*. 2013, 587:1012-20.

Haldar, D., Sen, D. and Gayen, K. (2016). A review on the production of fermentable sugars from lignocellulosic biomass through conventional and enzymatic route—a comparison. *International Journal of Green Energy*, 13(12), pp.1232-1253.

Hamacher, T., Boles, E., Gárdonyi, M., Hahn-Hägerdal, B. and Becker, J. (2002). Characterization of the xylose-transporting properties of yeast hexose transporters and their influence on xylose utilization. *Microbiology*, 148(9), pp.2783-2788.

Hambright, W., Deng, J., Tiedje, J., Brettar, I. and Rodrigues, J. (2016). *Shewanella baltica* Ecotypes Have Wide Transcriptional Variation under the Same Growth Conditions. mSphere, 1(5), pp. e00158-16.

Han, J., Choi, H., Lee, S., Orwin, P., Kim, J., LaRoe, S., Kim, T., O'Neil, J., Leadbetter, J., Lee, S., Hur, C., Spain, J., Ovchinnikova, G., Goodwin, L. and Han, C. (2010). Complete Genome Sequence of the Metabolically Versatile Plant Growth-Promoting Endophyte *Variovorax paradoxus* S110. *Journal of Bacteriology*, 193(5), pp.1183-1190.

Han, Y., Agarwal, V., Dodd, D., Kim, J., Bae, B., Mackie, R., Nair, S. and Cann, I. (2012). Biochemical and Structural Insights into Xylan Utilization by the Thermophilic Bacterium *Caldanaerobius polysaccharolyticus*. *Journal of Biological Chemistry*, 287(42), pp.34946-34960.

Hasona, A., Kim, Y., Healy, F., Ingram, L. and Shanmugam, K. (2004). Pyruvate Formate Lyase and Acetate Kinase Are Essential for Anaerobic Growth of Escherichia coli on Xylose. *Journal of Bacteriology*, 186(22), pp.7593-7600.

He, F., Nair, G., Soto, C., Chang, Y., Hsu, L., Ronzone, E., DeGrado, W. and Binns, A. (2009). Molecular Basis of ChvE Function in Sugar Binding, Sugar Utilization, and Virulence in *Agrobacterium tumefaciens*. *Journal of Bacteriology*, 191(18), pp.5802-5813.

Heinze T, Barsett, H., Ebringerová, A. (2005). Polysaccharides I: structure, characterisation and use. *Adv Polym Sci.*, 186, pp. 1-67

Hellweg, C., Pühler, A. and Weidner, S. (2009). The time course of the transcriptomic response of Sinorhizobium meliloti 1021 following a shift to acidic pH. *BMC Microbiology*, 9(1), p.37.

Henderson, P. and Maiden, M. (1990). Homologous Sugar Transport Proteins in *Escherichia coli* and Their Relatives in Both Prokaryotes and Eukaryotes. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 326(1236), pp.391-410.

HERSHBERG,R. et al. (2005) Chromosomal organization is shaped by the transcription regulatory network. *Trends in Genetics*, 21, 138-142.

Himmel, M., Baker, J. and Overend, R. (1996). A Review of:"Enzymatic Conversion of Biomass for Fuels Production". *Fuel Science and Technology Internati*onal, 14(1-2), pp.339-339.

Hogg, R. (1977). L-arabinose transport and the L-arabinose binding protein of *Escherichia coli*. *Journal of Supramolecular Structure*, 6(3), pp.411-417.

HOLLAND, I. B. (2003). ABC proteins: from bacteria to man, *Academic Press*.

Hovel, K. (2003). Crystal structure and snapshots along the reaction pathway of a family 51 -L-arabinofuranosidase. The EMBO Journal, 22(19), pp.4922-4932.

Honeycutt, R., McClelland, M. and Sobral, B. (1993). Physical map of the genome of *Rhizobium meliloti* 1021. *Journal of Bacteriology*, 175(21), pp.6945-6952.

Horazdovsky, B. and Hogg, R. (1989). Genetic reconstitution of the high-affinity L-arabinose transport system. *Journal of Bacteriology*, 171(6), pp.3053-3059.

Horler, R., Muller, A., Williamson, D., Potts, J., Wilson, K. and Thomas, G. (2009). Furanose-specific Sugar Transport: CHARACTERIZATION OF A BACTERIAL GALACTOFURANOSE-BINDING PROTEIN. *Journal of Biological Chemistry*, 284(45), pp.31156-31163.

Hough, L. (1972). Stereochemistry of Carbohydrates. *Carbohydrate Research*, 24(2), p.510.

Hu, X., Zhao, J., DeGrado, W. and Binns, A. (2012). Agrobacterium tumefaciens recognizes its host environment using ChvE to bind diverse plant sugars as virulence signals. *Proceedings of the National Academy of Sciences*, 110(2), pp.678-683.

Huang, X., Liu, Y., Dong, J., Qu, L., Zhang, Y., Wang, F., Tian, X. and Zhang, S. (2013). Mangrovibacterium diazotrophicum gen. nov., sp. nov., a nitrogen-fixing bacterium isolated from a mangrove sediment, and proposal of Prolixibacteraceae fam. nov. *INTERNATIONAL JOURNAL OF SYSTEMATIC AND EVOLUTIONARY MICROBIOLOGY*, 64(Pt 3), pp.875-881.

Hunke, S., Mourez, M., Jéhanno, M., Dassa, E. & Schneider, E. (2000). ATP modulates subunit-subunit interactions in an ATP-binding cassette transporter (MalFGK2) determined by site-directed chemical cross-linking. *J. Biol. Chem.*, 275, 15526.

Huntley, S., Kneip, S., Treuner-Lange, A. and Sogaard-Andersen, L. (2013). Complete Genome Sequence of Myxococcus stipitatus Strain DSM 14675, a Fruiting Myxobacterium. Genome Announcements, 1(2), pp.e00100-13-e00100-13.

Hutchings, M., Palmer, T., Harrington, D. and Sutcliffe, I. (2009). Lipoprotein biogenesis in Gram-positive bacteria: knowing when to hold 'em, knowing when to fold 'em. *Trends in Microbiology*, 17(1), pp.13-21.

Hyde, S., Emsley, P., Hartshorn, M., Mimmack, M., Gileadi, U., Pearce, S., Gallagher, M., Gill, D., Hubbard, R. and Higgins, C. (1990). Structural model of ATP-binding proteing associated with cystic fibrosis, multidrug resistance and bacterial transport. *Nature*, 346(6282), pp.362-365.

IPCC, Intergovernmental panel on climate change. *Climate Change 2013 – The Physical Science Basis* (2013)

Isbell S. Horace and Pigman W. William, (1938). Pyranose-Furanose interconversions with reference to the mutarotations of galactose, levulose, lactulose and turanose. *Journal of Research of the National Bureau of Standards*, 20, pp. 773-798.

Ishihama, Y., Schmidt, T., Rappsilber, J., Mann, M., Hartl, F., Kerner, M. and Frishman, D. (2008). Protein abundance profiling of the *Escherichia coli* cytosol. *BMC Genomics*, 9(1), p.102.

J. A. Loo, Studying noncovalent protein complexes by electrospray ionization mass spectrometry, *Mass Spectrom. Rev.* 1997, 16: 1-23.

J. L. Benesch, B. T. Ruotolo, D. A. Simmons, C. R. Robinson, Protein complexes in the gas phase: technology for structural genomics and proteomics, *Chem. Rev*. 2007, 107: 3544-67.

Jacobson, G.R. and Rosenbusch, J.P. (1976). Abundance and membrane association of elongation factor Tu in *E. coli*. Nature.261, pp.23–26

Jebbar, M., Sohn-Bosser, L., Bremer, E., Bernard, T. and Blanco, C. (2005). Ectoine-Induced Proteins in *Sinorhizobium meliloti* Include an Ectoine ABC-Type Transporter Involved in Osmoprotection and Ectoine Catabolism. *Journal of Bacteriology*, 187(4), pp.1293-1304.

Jeffrey, George A., (1997). An introduction to hydrogen bonding, *Oxford University Press*, 200, pp.191 and pp.14.

Jensen, J., Peters, N. and Bhuvaneswari, T. (2002). Redundancy in Periplasmic Binding Protein-Dependent Transport Systems for Trehalose, Sucrose, and Maltose in *Sinorhizobium meliloti*. *Journal of Bacteriology*, 184(11), pp.2978-2986.

JOHNSON, J., TOTH, J., SANTIWATANAKUL, S. and CHEN, J. (1997). Cultures of "Clostridium acetobutylicum" from Various Collections Comprise Clostridium acetobutylicum, Clostridium beijerinckii, and Two Other Distinct Types Based on DNA-DNA Reassociation. International Journal of Systematic Bacteriology, 47(2), pp.420-424.

JOSEPH, N. (1958). A pH Calculator Based on Linear Transformations of the Henderson-Hasselbalch Equation. *Science*, 128(3333), pp.1207-1208.

Joys ter Beek, J., Guskov, A. and Slotboom, D. (2014). Structural diversity of ABC transporters. *The Journal of General Physiology*, 143(4), pp.419-435.

Juncker, A., Willenbrock, H., von Heijne, G., Brunak, S., Nielsen, H. and Krogh, A. (2003). Prediction of lipoprotein signal peptides in Gram-negative bacteria. Protein Science, 12(8), pp.1652-1662.

Kaback HR. (1968) The role of the phosphoenolpyruvate-phosphotransferase system in the transport of sugars by isolated membrane preparations of *Escherichia coli*. The *Journal of biological chemistry*, 243, pp. 3711–3724.

Kaback, H. (2005). Structure and mechanism of the lactose permease. *Comptes Rendus Biologies*, 328(6), pp.557-567.

Kadokura, H., Katzen, F. and Beckwith, J. (2003). Protein Disulfide Bond Formation in Prokaryotes. *Annual Review of Biochemistry*, 72(1), pp.111-135.

Kaji, A. and Saheki, T. (1975). Endo-arabinanase from Bacillus subtilis F-11. *Biochimica et Biophysica Acta (BBA) - Enzymology*, 410(2), pp.354-360.

Kambam, P. and Henson, M. (2010). Engineering bacterial processes for cellulosic ethanol production. *Biofuels*, 1(5), pp.729-743.

Katoh, K. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Research, 30(14), pp.3059-3066.

Kellerman, O. & Szmelcman, S. (1974). Active transport of maltose in *Escherichia coli* K12. *Eur. J. Biochem.*, 47, pp. 139-149.

Kelly, D.J. and Thomas, G.H. (2001) The tripartite ATP-independent periplasmic (TRAP) transporters of bacteria and archaea. *FEMS Microbiol. Rev*. 25, 405–424

Kemner, J., Liang, X. and Nester, E. (1997). The *Agrobacterium tumefaciens* virulence gene *chvE* is part of a putative ABC-type sugar transport operon. *Journal of Bacteriology*, 179(7), pp.2452-2458.

Keseler, I., Mackie, A., Peralta-Gil, M., Santos-Zavaleta, A., Gama-Castro, S., Bonavides-Martínez, C., Fulcher, C., Huerta, A., Kothari, A., Krummenacker, M., Latendresse, M., Muñiz-Rascado, L., Ong, Q., Paley, S., Schröder, I., Shearer, A., Subhraveti, P., Travers, M., Weerasinghe, D., Weiss, V., Collado-Vides, J., Gunsalus, R., Paulsen, I. and Karp, P. (2012). EcoCyc: fusing model organism databases with systems biology. *Nucleic Acids Research*, 41(D1), pp.D605-D612.

Khankal, R., Chin, J. W. and Cirino, P. C. (2008). Role of xylose transporter in xylitol production from engineeres *Escherichia coli*. *Journal of Biotechnology* 134, 246–252.

Khlebnikov, A., Risa, O., Skaug, T., Carrier, T. and Keasling, J. (2000). Regulatable Arabinose-Inducible Gene Expression System with Consistent Control in All Cells of a Culture. *Journal of Bacteriology*, 182(24), pp.7029-7034.

Knee,E. et al. (2001) Root Mucilage from Pea and Its Utilization by Rhizosphere Bacteria as a Sole Carbon Source. Molecular Plant-Microbe Interactions, 14, 775-784.

Koroney, A., Plasson, C., Pawlak, B., Sidikou, R., Driouich, A., Menu-Bouaouiche, L. and Vicré-Gibouin, M. (2016). Root exudate of *Solanum tuberosumis* enriched in galactose-containing molecules and impacts the growth of *Pectobacterium atrosepticum*. *Annals of Botany*, 118(4), pp.797-808.

Koksharov, M., Lv, C., Zhai, X., Ugarova, N. and Huang, E. (2013). Bacillus subtilis alkaline phosphatase IV acquires activity only late at the stationary phase when produced in Escherichia coli. Overexpression and characterization of the recombinant enzyme. Protein Expression and Purification, 90(2), pp.186-194.

Koshland, D. (1958). Application of a Theory of Enzyme Specificity to Protein Synthesis. Proceedings of the National Academy of Sciences, 44(2), pp.98-104.

Koski, L. and Golding, G. (2001). The Closest BLAST Hit Is Often Not the Nearest Neighbor. *Journal of Molecular Evolution*, 52(6), pp.540-542.

Krämer, R. (1994). Functional principles of solute transport systems: concepts and perspectives. *Biochimica et Biophysica Acta (BBA) - Bioenergetics*, 1185(1), pp.1-34.

Krispin O. and Allmansberger B., (1998). The *Bacillus subtilis* AraE Protein Displays a Broad Substrate Specificity for Several Different Sugars. *Journal of Bacteriology*, 180(12), pp. 3250-52.

Kurakake, M., Takao, J., Asano, O., Tanimoto, H. and Komaki, T. (2011). Production of L-Arabinose from Corn Hull Arabinoxylan by Arthrobacter aurescens MK5 α-L-Arabinofuranosidase. *Journal of Food Science*, 76(2), pp.C231-C235.

Lagaert, S., Pollet, A., Courtin, C. and Volckaert, G. (2014). β-Xylosidases and α-l-arabinofuranosidases: Accessory enzymes for arabinoxylan degradation. Biotechnology Advances, 32(2), pp.316-332.

Lagaert, S., Pollet, A., Delcour, J., Lavigne, R., Courtin, C. and Volckaert, G. (2010). Substrate specificity of three recombinant α-l-arabinofuranosidases from Bifidobacterium adolescentis and their divergent action on arabinoxylan and arabinoxylan oligosaccharides. Biochemical and Biophysical Research Communications, 402(4), pp.644-650.

Laible, P., Scott, H., Henry, L., Hanson, D. (2004) Towards higher-throughput membrane protein production for structural genomics initiatives. *J. Struct. Funct. Genomics*. 5, pp. 167–172.

Laikova, O. (2001). Computational analysis of the transcriptional regulation of pentose utilization systems in the gamma subdivision of Proteobacteria. *FEMS Microbiology Letters,* 205(2), pp.315-322.

Lakowicz, J.R. Principles of Fluorescence Spectroscopy, (1999). 3rd ed.; Spinger: Berlin/Heidelberg, Germany.

Lall, S., Eriho, B. and Jay, J. (1989). Comparison for four methods for extracting periplasmic proteins. *Journal of Microbiological Methods*, 9(3), pp.195-199.

LaVallie, E.R., DiBlasio, E.A., Kovacic, S., Grant, K.L., Schendel, P.F. and McCoy, J.M. (1993). A thioredoxin gene fusion expression system that circumvents inclusion body formation in the *E. coli* cytoplasm. *Bio/Technology* 11, pp. 187–193.

Lebedev, A. A., Vagin, A. A., & Murshudov, G. N. (2008). Model preparation in MOLREP and examples of model improvement using X-ray data. *Acta Crystallographica Section D: Biological Crystallography*, 64(1), 33-39.

Lecher, J., Pittelkow, M., Zobel, S., Bursy, J., Bönig, T., Smits, S., Schmitt, L. and Bremer, E. (2009). The Crystal Structure of UehA in Complex with Ectoine—A Comparison with Other TRAP-T Binding Proteins. *Journal of Molecular Biology*, 389(1), pp.58-73.

Lee, Y. and Zeikus, J. (1993). Genetic organization, sequence and biochemical characterization of recombinant-xylosidase from *Thermoanaerobacterium saccharolyticum* strain B6A-RI. *Journal of General Microbiology*, 139(6), pp.1235-1243.

Lei, S., Lin, H., Wang, S., Callaway, J. and Wilcox, G. (1987). Characterization of the Erwinia carotovora pelB gene and its product pectate lyase. *Journal of Bacteriology*, 169(9), pp.4379-4383.

León-Barrios, M., Pérez-Yépez, J., Dorta, P., Garrido, A. and Jiménez, C. (2017). Alkalinity of Lanzarote soils is a factor shaping rhizobial populations with Sinorhizobium meliloti being the predominant microsymbiont of Lotus lancerottensis. *Systematic and Applied Microbiology*, 40(3), pp.171-178.

Letunic, I. and Bork, P. (2006). Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. Bioinformatics, 23(1), pp.127-128.

Li, F., Ren, S., Zhang, W., Xu, Z., Xie, G., Chen, Y., Tu, Y., Li, Q., Zhou, S., Li, Y., Tu, F., Liu, L., Wang, Y., Jiang, J., Qin, J., Li, S., Li, Q., Jing, H., Zhou, F., Gutterson, N. and Peng, L. (2013). Arabinose substitution degree in xylan positively affects lignocellulose enzymatic digestibility after various NaOH/H_2SO_4 pretreatments in Miscanthus. Bioresource Technology, 130, pp.629-637.

Li, H., Long, C., Zhou, J., Liu, J., Wu, X., and Long, M. (2013b). Rapid analysis of mono-saccharides and oligo-saccharides in hydrolysates of lignocellulosic biomass by HPLC, *Biotechnol. Lett*. 35(9), pp. 1405-1409.

Liu, M., Durfee, T., Cabrera, J., Zhao, K., Jin, D., and Blattner, F. (2005). Global Transcriptional Programs Reveal a Carbon Source Foraging Strategy by *Escherichia coli*. *Journal of Biological Chemistry*, 280(16), 15921-15927.

Liao, J., Mi, L., Pontrelli, S. and Luo, S. (2016). Fuelling the future: microbial engineering for the production of sustainable biofuels. *Nature Reviews Microbiology*, 14(5), pp.288-304.

Linderstrøm-Lang, K., Schellman, J. A. (1959). "Protein structure and enzyme activity". *The Enzymes*. 1 (2), pp.443–510.

Locher, K. P., Lee, A. T. & Rees, D. C. 2002. The *E. coli* BtuCD structure: a framework for ABC transporter architecture and mechanism. *Science*, 296, pp. 1091.

López-Maury, L., Marguerat, S., and Bähler, J. (2008). Tuning gene expression to changing environments: from rapid responses to evolutionary adaptation. *Nat. Rev. Genet*. 9, pp. 583–593.

Lowendorf, H. S., Baya, A. M. & Alexander, M. Survival of Rhizobium in Acid Soils. (1981). *Appl. Environ. Microbiol.* 42, 951–957.

Lu, G., Westbrooks, J. M., Davidson, A. L. & Chen, J. 2005. ATP hydrolysis is required to reset the ATP-binding cassette dimer into the resting-state conformation. *Proc. Natl. Acad. Sci. USA*, 102, pp. 17969.

Maehara, T., Fujimoto, Z., Ichinose, H., Michikawa, M., Harazono, K. and Kaneko, S. (2014). Crystal Structure and Characterization of the Glycoside Hydrolase Family 62 α-l-Arabinofuranosidase from *Streptomyces coelicolor*. *Journal of Biological Chemistry*, 289(11), pp.7962-7972.

Miljković M., Carbohydrates (2009). Chapter 2: Conformational Analysis of Monosaccharides. Springer Science+Business Media, LLC.

Macalady, J.L., Vestling, M.M., Baumler, D., Boekelheide, N., Kaspar, C.W., and Banfield, J.F. (2004) Tetraether-linked membrane monolayers in *Ferroplasma spp*: a key to survival in acid. *Extremophiles* 8: 411–419.

Madej, M., Sun, L., Yan, N. and Kaback, H. (2014). Functional architecture of MFS D-glucose transporters. *Proceedings of the National Academy of Sciences*, 111(7), pp.E719-E727.

Maloney P. C, (2002). Bacterial Membrane Transport: Superfamilies of Transport Proteins. eLS.

Maqbool, A., Hervé, M., Mengin-Lecreulx, D., Wilkinson, A. and Thomas, G. (2012). MpaA is a murein-tripeptide-specific zinc carboxypeptidase that functions as part of a catabolic pathway for peptidoglycan-derived peptides in γ-proteobacteria. *Biochemical Journal*, 448(3), pp.329-341.

Maqbool, A., Levdikov, V., Blagova, E., Hervé, M., Horler, R., Wilkinson, A. and Thomas, G. (2011). Compensating Stereochemical Changes Allow Murein Tripeptide to Be Accommodated in a Conventional Peptide-binding Protein. *Journal of Biological Chemistry*, 286(36), pp.31512-31521.

Marshall, A. and Hendrickson, C. (2008). High-Resolution Mass Spectrometers. *Annual Review of Analytical Chemistry*, 1(1), pp.579-599.

Matulis, D., Kranz, J., Salemme, F. and Todd, M. (2005). Thermodynamic Stability of Carbonic Anhydrase: Measurements of Binding Affinity and Stoichiometry Using ThermoFluor. *Biochemistry*, 44(13), pp.5258-5266.

Maurya, D., Singla, A., and Negi, S. (2015). An overview of key pretreatment processes for biological conversion of lignocellulosic biomass to bioethanol. *3 Biotech*, 5(5), 597-609.

Mauchline, T., Fowler, J., East, A., Sartor, A., Zaheer, R., Hosie, A., Poole, P. and Finan, T. (2006). Mapping the Sinorhizobium meliloti 1021 solute-binding protein-dependent transportome. Proceedings of the National Academy of Sciences, 103(47), pp.17933-17938.

McCarter, J. and Stephen Withers, G. (1994). Mechanisms of enzymatic glycoside hydrolysis. Current Opinion in Structural Biology, 4(6), pp.885-892.

McCoy, J., Levin, E. and Zhou, M. (2015). Structural insight into the PTS sugar transporter EIIC. *Biochimica et Biophysica Acta* (BBA) - General Subjects, 1850(3), pp.577-585.

McKee, L., Pena, M., Rogowski, A., Jackson, A., Lewis, R., York, W., Krogh, K., Vikso-Nielsen, A., Skjot, M., Gilbert, H. and Marles-Wright, J. (2012). Introducing endo-xylanase activity into an exo-acting arabinofuranosidase that targets side chains. Proceedings of the National Academy of Sciences, 109(17), pp.6537-6542.

Mehrnejad, F., Khadem-Maaref, M., Ghahremanpour, M. and Doustdar, F. (2010). Mechanisms of amphipathic helical peptide denaturation by guanidinium chloride and urea: a molecular dynamics simulation study. *Journal of Computer-Aided Molecular Design*, 24(10), pp.829-841.

Messens, J. and Collet, J. (2006). Pathways of disulfide bond formation in Escherichia coli. *The International Journal of Biochemistry & Cell Biology*, 38(7), pp.1050-1062.

Miller-Williams, M. (2006). Isolation of salt-sensitive mutants of *Sinorhizobium meliloti* strain Rm1021. *Microbiology*, 152(7), pp.2049-2059.

Miller, 3rd, D., Olson, J. and Quiocho, F. (1980). The mechanism of sugar binding to the periplasmic receptor for galactose chemotaxis and transport in Escherichia coli. *J. Biol. Chem*, 255, pp.2465-2471.

Miller, H., (1987). Practical aspects of preparing phage and plasmid DNA: Growth, maintenance and storage of bacteria and bacteriophage. *Methods Enzymol*ogy, 152, pp. 145-170.

Mitchell, W. (1996). Carbohydrate Uptake and Utilization by *Clostridium beijerinckii* NCIMB 8052. *Anaerobe*, 2(6), pp.379-384.

MOHNEN,D. (2008) Pectin structure and biosynthesis. Current Opinion in Plant Biology, 11, 266-277.

Moody, S., Clarke, A. and Bacic, A. (1988). Structural analysis of secreted slime from wheat and cowpea roots. *Phytochemistry*, 27(9), pp.2857-2861.

Moody,S. et al. (1988) Structural analysis of secreted slime from wheat and cowpea roots. Phytochemistry, 27, 2857-2861.

Morbach, S., Tebbe, S. & Schneider, E. (1993). The ATP-binding cassette (ABC) transporter for maltose/maltodextrins of *Salmonella typhimurium*. Characterization of the ATPase activity associated with the purified MalK subunit. *J. Biol. Chem*., 268, 18617-18621.

Morgan N Price, Kelly M Wetmore, Robert Jordan Waters, Mark Callaghan, Jayashree Ray, JenniferV Kuehl, RyanA Melnyk, JacobS Lamson, Yumi Suh, Zuelma Esquivel, Harini Sadeeshkumar, Romy Chakraborty, Benjamin E Rubin, James Bristow, Matthew J Blow, Adam P Arkin, Adam M Deutschbauer. (2016). Deep Annotation of Protein Function across Diverse Bacteria from Mutant Phenotypes. *BioRxiv*.

Morimoto, Y., Kami-ike, N., Miyata, T., Kawamoto, A., Kato, T., Namba, K. and Minamino, T. (2016). High-Resolution pH Imaging of Living Bacterial Cells To Detect Local pH Differences. mBio, 7(6), pp. e01911-16.

Morimoto, Y., Kami-ike, N., Namba, K. and Minamino, T. (2017). Determination of Local pH Differences within Living *Salmonella* Cells by High-resolution pH Imaging Based on pH-sensitive GFP Derivative, pHluorin(M153R). *BIO-PROTOCOL*, 7(17).

Mosbah, M. and Mars, M. (2006). Genotypic diversity of rhizobia isolated fromRetama raetam in arid regions of Tunisia. *Annals of Microbiology*, 56(4), pp.305-311.

Mota, L., Sarmento, L. and de Sa-Nogueira, I. (2001). Control of the Arabinose Regulon in Bacillus subtilis by AraR In Vivo: Crucial Roles of Operators, Cooperativity, and DNA Looping. Journal of Bacteriology, 183(14), pp.4190-4201.

Mota, L., Sarmento, L. and de Sa-Nogueira, I. (2001). Control of the Arabinose Regulon in Bacillus subtilis by AraR In Vivo: Crucial Roles of Operators, Cooperativity, and DNA Looping. *Journal of Bacteriology*, 183(14), pp.4190-4201.

Mota, L., Tavares, P. and Sa-Nogueira, I. (1999). Mode of action of AraR, the key regulator of l-arabinose metabolism in *Bacillus subtilis*. *Molecular Microbiology*, 33(3), pp.476-489.

Mourez, M., Hofnung, M. & Dassa, E. 1997. Subunit interactions in ABC transporters: a conserved sequence in hydrophobic membrane proteins of periplasmic permeases defines an important site of interaction with the ATPase subunits. EMBO J., 16, 3066-3077.

Moussatova, A., Kandt, C., O'Mara, M. L. & Tieleman, D. P. 2008. ATP-binding cassette transporters in Escherichia coli. Biochim. Biophys. Acta, 1778, 1757-1771.

Mowbray SL, Cole LB. 1.7 Å X-ray structure of the periplasmic ribose receptor from *Escherichia coli.* (1992). J Mol Biol, 225, pp. 155–175.

Mowbray, S. and Cole, L. (1992). 1.7 Å X-ray structure of the periplasmic ribose receptor from Escherichia coli. *Journal of Molecular Biology*, 225(1), pp. 155-175.

Mulligan C., Kelly D. J and Thomas G. H. (2007) Tripartite ATP-independent periplasmic transporters: application of a relational database for genome-wide analysis of transporter gene frequency and organization. *J. Mol. Microbiol. Biotechnol*. 12, 218–226

Mulligan, C., Geertsma, E., Severi, E., Kelly, D., Poolman, B. and Thomas, G. (2009). The substrate-binding protein imposes directionality on an electrochemical sodium gradient-driven TRAP transporter. Proceedings of the National Academy of Sciences, 106(6), pp.1778-1783.

Murphy, K.C. (1991). λ gam protein inhibits the helicase and M stimulated recombination activities of *Escherichia coli* RecBCD enzyme. *J. Bacteriol*. 173, pp. 5808-5821.

Murshudov, G. N., Skubák, P., Lebedev, A. A., Pannu, N. S., Steiner, R. A., Nicholls, R. A., ... & Vagin, A. A. (2011). REFMAC5 for the refinement of macromolecular crystal structures. Acta Crystallographica Section D: Biological Crystallography, 67(4), 355-367.

Naik, S., Goud, V., Rout, P. and Dalai, A. (2010). Production of first and second generation biofuels: A comprehensive review. *Renewable and Sustainable Energy Reviews*, 14(2), pp.578-597.

Nallamsetty, S. and Waugh, D. (2006). Solubility-enhancing proteins MBP and NusA play a passive role in the folding of their fusion partners. *Protein Expression and Purification*, 45(1), pp.175-182.

Nanmori, T., Watanabe, T., Shinke, R., Kohno, A. and Kawamura, Y. (1990). Purification and properties of thermostable xylanase and beta-xylosidase produced by a newly isolated *Bacillus stearothermophilus* strain. *Journal of Bacteriology*, 172(12), pp.6669-6672.

Nelson, D. and Cox, M. (2000). Principles of Biochemistry. *Macmillan Press*, London, UK.

Nettleship, J. E., Brown, J., Groves, M. R. & Geerlof, A. (2008). Methods for protein characterization by mass spectrometry, thermal shift (ThermoFluor) assay, and multiangle or static light scattering. *Methods Mol. Biol*. 426, pp. 299–318

Neu, H. and Heppel, L. (1965). The release of enzymes from *Escherichia coli* by osmotic shock and during the formation of spheroplasts. *The Journal of Biological Chemistry*, 240, pp.3685-3692.

Nguimbou, R., Boudjeko, T., Njintang, N., Himeda, M., Scher, J. and Mbofung, C. (2012). Mucilage chemical profile and antioxidant properties of giant swamp taro tubers. *Journal of Food Science and Technology*, 51(12), pp.3559-3567.

Nicholls, R. A., Long, F., & Murshudov, G. N. (2012). Low-resolution refinement tools in REFMAC5. *Acta Crystallographica Section D: Biological Crystallography*, *68*(4), 404-417.

Niesen, F., Berglund, H. and Vedadi, M. (2007). The use of differential scanning fluorimetry to detect ligand interactions that promote protein stability.

Novichkov, P., Kazakov, A., Ravcheev, D., Leyn, S., Kovaleva, G., Sutormin, R., Kazanov, M., Riehl, W., Arkin, A., Dubchak, I. and Rodionov, D. (2013). RegPrecise 3.0 – A resource for genome-scale exploration of transcriptional regulation in bacteria. BMC Genomics, 14(1), p.745.

Novick, A. and Weiner, M. (1957). ENZYME INDUCTION AS AN ALL-OR-NONE PHENOMENON. *Proceedings of the National Academy of Sciences*, 43(7), pp.553-566.

Nye, P. (1981). Changes of pH across the rhizosphere induced by roots. Plant and Soil, 61(1-2), pp.7-26.

O'Connell Motherway, M., Fitzgerald, G. and van Sinderen, D. (2010). Metabolism of a plant derived galactose-containing polysaccharide by Bifidobacterium breve UCC2003. Microbial Biotechnology, 4(3), pp.403-416.

Oh, B.H., Kang, C.H., Bondt, H.D., Kim, S.H., Nikaido, K., Joshi, A.K. and Ames, G.F. (1994). The bacterial periplasmic histidine-binding protein. Structure/ function analysis of the ligand-binding site and comparison with related proteins. *J. Biol. Chem*. 269, pp. 4135–4143.

Okie, J.G., Van Horn, D.J., Storch, D., Barrett, J.E., Gooseff, M.N., Kopsova, L., Takacs-Vesbach, C.D., 2015. Niche and metabolic principles explain patterns of diversity and distribution: theory and a case study with soil bacterial communities. *Proc Biol Sci* 282, pp. 60.

Okumura, M., Saiki, M., Yamaguchi, H. and Hidaka, Y. (2011). Acceleration of disulfide-coupled protein folding using glutathione derivatives. *FEBS Journal*, 278(7), pp.1137-1144.

Olliveier, Bernard, Pierre Caumette, Jean-Louis Garcia, and Robert A. Mah. (1994) Anaerobic Bacteria from Hypersaline Environments. *American Society for Microbiology*, *Microbiological Reviews*. 58.1, pp. 27-38.

Osborn, H. (2006). Carbohydrates. *Oxford: Academic Press*.

Oswald, C., Smits, S., Höing, M., Sohn-Bösser, L., Dupont, L., Le Rudulier, D., Schmitt, L. and Bremer, E. (2008). Crystal Structures of the Choline/Acetylcholine Substrate-binding Protein ChoX from *Sinorhizobium meliloti* in the Liganded and Unliganded-Closed States. *Journal of Biological Chemistry*, 283(47), pp.32848-32859.

Overbeek, R. (2005). The Subsystems Approach to Genome Annotation and its Use in the Project to Annotate 1000 Genomes. Nucleic Acids Research, 33(17), pp.5691-5702.

Overmars, L., Kerkhoven, R., Siezen, R. and Francke, C. (2013). MGcV: the microbial genomic context viewer for comparative genome analysis. *BMC Genomics*, 14(1), p.209.

Padan, E. (2009). Bacterial Membrane Transport: Superfamilies of Transport Proteins. eLS.

Pao, S.S., Paulsen, I.T., Saier, M.H., (1998). Major Facilitator Superfamily. *Microbiol Mol Biol Rev*, 62, pp. 1–34.

Papaneophytou, C., Grigoroudis, A., McInnes, C. and Kontopidis, G. (2014). Quantification of the Effects of Ionic Strength, Viscosity, and Hydrophobicity on Protein–Ligand Binding Affinity. *ACS Medicinal Chemistry Letters*, 5(8), pp.931-936.

Parachin, N., Bergdahl, B., van Niel, E. and Gorwa-Grauslund, M. (2011). Kinetic modelling reveals current limitations in the production of ethanol from xylose by recombinant *Saccharomyces cerevisiae*. *Metabolic Engineering*, 13(5), pp.508-517.

Pearse AGE. Histochemistry: theoretical and applied. 4. London: Churchill-Livingstone; 1980.

Petersen, T., Brunak, S., von Heijne, G. and Nielsen, H. (2011). SignalP 4.0: discriminating signal peptides from transmembrane regions. Nature Methods, 8(10), pp.785-786.

Pierce, J. (2008). Metabolic engineering for the production of renewable fuels and chemicals. *Journal of Biotechnology*, 136, pp. S3-S4.

Pollet, A., Delcour, J. and Courtin, C. (2010). Structural determinants of the substrate specificities of xylanases from different glycoside hydrolase families. *Critical Reviews in Biotechnology*, 30(3), pp.176-191.

Poolman, B., Knol, J., van der Does, C., Henderson, P., Liang, W., Leblanc, G., Pourcher, T. and Mus-Veteau, I. (1996). Cation and sugar selectivity determinants in a novel family of transport proteins. *Molecular Microbiology*, 19(5), pp.911-922.

Poysti, N., Loewen, E., Wang, Z. and Oresnik, I. (2007). *Sinorhizobium meliloti* pSymB carries genes necessary for arabinose transport and catabolism. Microbiology, 153(3), pp.727-736.

Prade, R., Zhan, D., Ayoubi, P. and Mort, A. (1999). Pectins, Pectinases and Plant-Microbe Interactions. *Biotechnology and Genetic Engineering Reviews*, 16(1), pp.361-392.

Puigbò, P., Bravo, I. and Garcia-Vallve, S. (2008). CAIcal: A combined set of tools to assess codon usage adaptation. *Biology Direct*, 3(1), p.38.

Qian, Y., Yomano, L. P., Preston, J. F., Aldrich, H. C. and Ingram, L. O. (2003). Cloning, Characterization and Functional Expression of the Klebsiella oxyotoca Xylodextrin Utilization Operon (*xynTB*) in *Escherichia coli*. *Applied and Environmental Microbiology*, 69, pp. 5957– 5967.

Qian, Z., Trostel, A., Lewis, D., Lee, S., He, X., Stringer, A., Wade, J., Schneider, T., Durfee, T. and Adhya, S. (2016). Genome-Wide Transcriptional Regulation and Chromosome Structural Arrangement by GalR in *E. coli*. *Frontiers in Molecular Biosciences*, 3.

Quan, S., Hiniker, A., Collet, J. and Bardwell, J. (2013). Isolation of Bacteria Envelope Proteins. *Methods in Molecular Biology*, 966, pp. 359-366.

Quiocho, F. and Vyas, N. (1984). Novel stereospecificity of the L-arabinose-binding protein. Nature, 310(5976), pp.381-386.

Quiocho, F., Phillips, G., Parsons, R. and Hogg, R. (1974). Crystallographic data of an l-arabinose-binding protein from *Escherichia coli*. *Journal of Molecular Biology*, 86(2), pp.491-493.

Quiocho, F.A. and Ledvina, P. (1996). Atomic structure and specificity of bacterial periplasmic receptors for active transport and chemotaxis: variation of common themes. *Mol. Microbiol*. 20, pp. 17–25.

Quistgaard, E.M., Löw, C., Moberg, P., Trésaugues, L., Nordlund, P., (2013). Structural basis for substrate transport in the GLUT-homology family of monosaccharide transporters. *Nat Struct Mol Biol* 20, pp. 766–768.

Ramachandran, V., East, A., Karunakaran, R., Downie, J. and Poole, P. (2011). Adaptation of Rhizobium leguminosarum to pea, alfalfa and sugar beet rhizospheres investigated by comparative transcriptomics. *Genome Biology*, 12(10), p.R106.

Rasmussen LE, Sørensen HR, Vind J, Viksø-Nielsen A. (2006). Mode of action and properties of the β-xylosidases from *Talaromyces emersonii* and *Trichoderma reesei*. *Biotechnol Bioeng*, 94, pp. 869–76.

Rawn, D. J. (1989). Proteins, energy and metabolism. *Biochemistry*. Neil Patterson Publishers, Burlington, N.C.

Reddy K., R. (2005). L-Arginine increases the solubility of unfolded species of hen egg white lysozyme. *Protein Science*, 14(4), pp.929-935.

Reddy, V., Shlykov, M., Castillo, R., Sun, E. and Saier, M. (2012). The major facilitator superfamily (MFS) revisited. *FEBS Journal*, 279(11), pp.2022-2035.

Reider Apel, A., Ouellet, M., Szmidt-Middleton, H., Keasling, J. and Mukhopadhyay, A. (2016). Evolved hexose transporter enhances xylose uptake and glucose/xylose co-utilization in *Saccharomyces cerevisiae*. *Scientific Reports*, 6(1).

Rhimi, M., Bajic, G., Ilhammami, R., Boudebbouze, S., Maguin, E., Haser, R. and Aghajari, N. (2011). The acid-tolerant L-arabinose isomerase from the mesophilic *Shewanella* sp. ANA-3 is highly active at low temperatures. *Microbial Cell Factories*, 10(1), p.96.

Ribeiro, G., Gruninger, R., Badhan, A. and McAllister, T. (2016). Mining the rumen for fibrolytic feed enzymes. *Animal Frontiers*, 6(2), p.20.

Richards, F. (1997). Protein stability: still an unsolved problem. *Cellular and Molecular Life Sciences CMLS,* 53(10), pp.790-802.

Riggs, P. (2000). Expression and Purification of Recombinant Proteins by Fusion to Maltose-Binding Protein. *Molecular Biotechnology*, 15(1), pp.51-63.

Ringwood, A. and Keppler, C. (2002). Water quality variation and clam growth: Is pH really a non-issue in estuaries? *Estuaries*, 25(5), pp.901-907.

Romano, A. (1986). Microbial sugar transport systems and their importance in biotechnology. *Trends in Biotechnology*, 4(8), pp.207-213.

Rong Fu Wang and Kushner, S. (1991). Construction of versatile low-copy-number vectors for cloning, sequencing and gene expression in *Escherichia coli*. Gene, 100, pp.195-199.

Rosano, G. and Ceccarelli, E. (2014). Recombinant protein expression in Escherichia coli: advances and challenges. *Frontiers in Microbiology*, 5.

Rodionov, D., Mironov, A. and Gelfand, M. (2001). Transcriptional regulation of pentose utilisation systems in the *Bacillus*/*Clostridium* group of bacteria. *FEMS Microbiology Letters*, 205(2), pp.305-314.

Rodionov, D., Yang, C., Li, X., Rodionova, I., Wang, Y., Obraztsova, A., Zagnitko, O., Overbeek, R., Romine, M., Reed, S., Fredrickson, J., Nealson, K. and Osterman, A. (2010). Genomic encyclopedia of sugar utilization pathways in the Shewanella genus. BMC Genomics, 11(1), p.494.

Roumiantseva, M., Onishchuk, O., Belova, V., Kurchak, O. and Simarov, B. (2011). Polymorphism of *Sinorhizobium meliloti* strains isolated from diversity centers of alfalfa in various soil and climatic conditions. *Russian Journal of Genetics: Applied Research*, 1(2), pp.97-102.

Rubin, B., Wetmore, K., Price, M., Diamond, S., Shultzaberger, R., Lowe, L., Curtin, G., Arkin, A., Deutschbauer, A. and Golden, S. (2015). The essential gene set of a photosynthetic organism. Proceedings of the National Academy of Sciences, 112(48), pp. E6634-E6643.

Saha, B. (2000). α-l-Arabinofuranosidases. Biotechnology Advances, 18(5), pp.403-423.

Sakamoto, T., Ogura, A., Inui, M., Tokuda, S., Hosokawa, S., Ihara, H., and Kasai, N. (2011). Identification of a GH62 alpha-L-arabinofuranosidase specific for arabinoxylan produced by *Penicillium chrysogenum*. *Appl. Microbiol. Biotechnol*. 90(1), pp. 137-146.

Saloheimo, A., Rauta, J., Stasyk, O., Sibirny, A., Penttilä, M. and Ruohonen, L. (2006). Xylose transport studies with xylose-utilizing *Saccharomyces cerevisiae* strains expressing heterologous and homologous permeases. *Applied Microbiology and Biotechnology*, 74(5), pp.1041-1052.

Sandermann, H. (1977). beta-D-Galactoside Transport in *Escherichia coli*: Substrate Recognition. *European Journal of Biochemistry*, 80(2), pp.507-515.

Schellman J. A. (1997). Temperature, stability, and the hydrophobic interaction. *Biophys. J.* 73 2960–2964.

Schlake T, Bode J. (1994). Use of mutated FLP recognition target (FRT) sites for the exchange of expression cassettes at defined chromosomal loci. *Biochemistry*. 33 (43), pp. 12746–12751.

Schlösser, A., and H. Schrempf. (1996). A lipid-anchored binding protein is a component of an ATP-dependent cellobiose/cellotriose-transport system from the cellulose degrader *Streptomyces reticuli*. *Eur. J. Biochem*. 242, pp. 332– 338.

Selig, M. J., Knoshaug, E. P., Adney, W.S., Himmel, M. E., and Decker, S. R. (2008). Synergistic enhancement of cellobiohydrolase performance on pretreated corn stover by addition of xylanase and esterase activities, *Bioresource Technol*. 99(11), pp. 4997-5005.

Severi, E., Randle, G., Kivlin, P., Whitfield, K., Young, R., Moxon, R., Kelly, D., Hood, D. and Thomas, G. (2005). Sialic acid transport inHaemophilus influenzaeis essential for lipopolysaccharide sialylation and serum

resistance and is dependent on a novel tripartite ATP-independent periplasmic transporter. *Molecular Microbiology*, 58(4), pp.1173-1185.

Sezonov, G., Joseleau-Petit, D. and D'Ari, R. (2007). Escherichia coli Physiology in Luria-Bertani Broth. *Journal of Bacteriology*, 189(23), pp.8746-8749.

Sharan, S., Thomason, L., Kuznetsov, S. and Court, D. (2009). Recombineering: a homologous recombination-based method of genetic engineering. *Nature Protocols*, 4(2), pp.206-223.

Sharff, A. J., Rodseth, L. E., Spurlino, J. C. & Quiocho, F. A. (1992). Crystallographic evidence of a large ligand-induced hinge-twist motion between the two domains of the maltodextrin binding protein involved in active transport and chemotaxis. *Biochemistry*, 31, pp. 10657-10663.

Shin, H., McClendon, S., Vo, T. and Chen, R. (2010). Escherichia coli Binary Culture Engineered for Direct Fermentation of Hemicellulose to a Biofuel. *Applied and Environmental Microbiology*, 76(24), pp.8150-8159.

Shiraki, K., Kudou, M., Fujiwara, S., Imanaka, T. and Takagi, M. (2002). Biophysical Effect of Amino Acids on the Prevention of Protein Aggregation. *Journal of Biochemistry*, 132(4), pp.591-595.

Shrake, A. and Ross, P. (1990). Ligand-induced biphasic protein denaturation. *J. Biol. Chem.*, 265, pp. 5055-5059.

Shrake, A. and Ross, P. (1992). Origins and consequences of ligand-induced multiphasic thermal protein denaturation. *Biopolymers*, 32(8), pp.925-940.

Shulami S, Gat O, Sonenshein AL & Shoham Y (1999) The glucuronic acid utilization gene cluster from Bacillus stearothermophilus T-6. J Bacteriol, 181, pp. 3695–3704.

Shulami, S., Zaide, G., Zolotnitsky, G., Langut, Y., Feld, G., Sonenshein, A. L. and Shoham, Y. (2007). A Two-Component System Regulates the Expression of an ABC Transporter for Xylo-Oligosaccharides in Geobacillus stearothermophilus. *Applied Environmental Microbiology*, 73, pp. 874–884.

Siegele, D. and Hu, J. (1997). Gene expression from plasmids containing the *araBAD* promoter at subsaturating inducer concentrations represents mixed populations. *Proceedings of the National Academy of Sciences*, 94(15), pp.8168-8172.

Sievers, F., Wilm, A., Dineen, D., Gibson, T., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Soding, J., Thompson, J. and Higgins, D. (2014). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Molecular Systems Biology, 7(1), pp.539-539.

Silhavy, T.J., Kahne, D., and Walker, S. (2010). The Bacterial Cell Envelope. *Cold Spring Harb Perspect Biol*, 2.

Sims, R., Mabee, W., Saddler, J. and Taylor, M. (2010). An overview of second generation biofuel technologies. *Bioresource Technology*, 101(6), pp.1570-1580.

Sinnott, M. (2007). Carbohydrate chemistry and biochemistry. 1st ed. RSC Publishing. *Chapter 2: Conformations of Monosaccharides*, pp. 58-62.

Skerratt, J. (2002). Shewanella olleyana sp. nov., a marine species isolated from a temperate estuary which produces high levels of polyunsaturated fatty acids. *International Journal Of Systematic And Evolutionary Microbiology*, 52(6), pp.2101-2106.

Slotboom, D. (2013). Structural and mechanistic insights into prokaryotic energy-coupling factor transporters. *Nature Reviews Microbiology*, 12(2), pp.79-87.

Sobetzko, P., Travers, A., Muskhelishvili, G., 2012. Gene order and chromosome dynamics coordinate spatiotemporal gene expression during the bacterial growth cycle. *PNAS*, 109, pp. 355–356.

Socha, A., Parthasarathi, R., Shi, J., Pattathil, S., Whyte, D., Bergeron, M., George, A., Tran, K.,Stavila, V., Venkatachalam, S., Hahn, M., Simmons, B. and Singh, S. (2014). Efficient biomass pretreatment using ionic liquids derived from lignin and hemicellulose. *Proceedings of the National Academy of Sciences*, 111(35), pp. E3587-E3595.

Somerville, C. (2007). Biofuels. *Current Biology*, (Vol. 17, No. 4), pp.R115 - R119.

Song, C., Zhang, S. and Huang, H. (2015). Choosing a suitable method for the identification of replication origins in microbial genomes. *Frontiers in MICROBIOLOGY*, 6.

Song, S., Park, C., 1997. Organization and regulation of the D-xylose operons in *Escherichia coli* K-12: XylR acts as a transcriptional activator. *J. Bacteriol*. 179, pp. 7025–7032.

Sørensen, H., Jørgensen, C., Hansen, C., Jørgensen, C., Pedersen, S. and Meyer, A. (2006). A novel GH43 α-l-arabinofuranosidase from Humicola insolens: mode of action and synergy with GH51 α-l-arabinofuranosidases on wheat arabinoxylan. Applied Microbiology and Biotechnology, 73(4), pp.850-861.

Sorensen, H., Pedersen, S., Jorgensen, C. and Meyer, A. (2007). Enzymatic Hydrolysis of Wheat Arabinoxylan by a Recombinant "Minimal" Enzyme Cocktail Containing β-Xylosidase and Novel endo-1,4-β-Xylanase and α-L-Arabinofuranosidase Activities. Biotechnology Progress, 23(1), pp.100-107.

Sorrell, F. J., Greenwood, G. K., Birchall, K. & Chen, B. (2010). Development of a differential scanning fluorimetry based high throughput screening assay for the discovery of affinity binders against an anthrax protein. *J Pharm Biomed Anal*, 52, pp. 802-8.

Spurlino J.C., Lu G.Y. and Quiocho F.A., (1991). The 2.3-A resolution structure of the maltose- or maltodextrin-binding protein, a primary receptor of bacterial active transport and chemotaxis. *J. Biol. Chem.,* 266 (8), pp. 5202-19.

Story, R. M. & Steitz, T. A. (1992). Structure of the recA protein–ADP complex. *Nature*, 355, 374-376.

Strauss, G. and Hauser, H. (1986). Stabilization of lipid bilayer vesicles by sucrose during freezing. *Proceedings of the National Academy of Sciences*, 83(8), pp.2422-2426.

Stringer, A., Currenti, S., Bonocora, R., Baranowski, C., Petrone, B., Palumbo, M., Reilly, A., Zhang, Z., Erill, I. and Wade, J. (2013). Genome-Scale Analyses of *Escherichia coli* and *Salmonella enterica* AraC Reveal Noncanonical Targets and an Expanded Core Regulon. *Journal of Bacteriology*, 196(3), pp.660-671.

Studier, F. (2005). Protein production by auto-induction in high-density shaking cultures. *Protein Expression and Purification*, 41(1), pp.207-234.

Sturtevant, J. (1987). Biochemical Applications Of Differential Scanning Calorimetry. *Annual Review of Physical Chemistry*, 38(1), pp.463-488.

Sumiya, M., Davis, E.O., Packman, L.C., McDonald, T.P., Henderson, P.J., (1995). Molecular genetics of a receptor protein for D-xylose, encoded by the gene *xylF*, in *Escherichia coli*. *Recept. Channels* 3, pp. 117–128.

Sun, N., Liu, H., Sathitsuksanoh, N., Stavila, V., Sawant, M., Bonito, A., Tran, K., George, A., Sale, K., Singh, S., Simmons, B. and Holmes, B. (2013). Production and extraction of sugars from switchgrass hydrolyzed in ionic liquids. *Biotechnology for Biofuels*, 6(1), p.39.

Sunna, A. and Antranikian, G. (1997). Xylanolytic Enzymes from Fungi and Bacteria. *Critical Reviews in Biotechnology*, 17(1), pp.39-67.

Sutcliffe, I. and Harrington, D. (2002). Pattern searches for the identification of putative lipoprotein genes in Gram-positive bacterial genomes. *Microbiology*, 148(7), pp.2065-2077.

Sweet, G., Gangor, C., Voegele, R., Wittenkindt, N., Beuerie, J., Truniger, V., Lin, E. & Boos, W. 1990. Glycerol facilitator of *Escherichia coli*: cloning of glpF and identification of the glpF product. *J. Bacteriol*., 172, pp. 424-430.

Tam, R. & Saier, M. H., JR. (1993). Structural, functional, and evolutionary relationships among extracellular solute-binding receptors of bacteria. *Microbiol. Rev*., 57, pp. 320-46.

Tame, J. R., Murshudov, G. N., Dodson, E. J., Neil, T. K., Dodson, G. G., Higgins, C. F. & Wilkinson, A. J. 1994. The structural basis of sequence-independent peptide binding by OppA protein. *Science*, 264, 1578-81.

Taneja, S. and Ahmad, F. (1994). Increased thermal stability of proteins in the presence of amino acids. *Biochemical Journal*, 303(1), pp.147-153.

Tang, C., Schwieters, C.D. and Clore, G.M. (2007). Open-to-closed transition in apo maltose-binding protein observed by paramagnetic NMR. *Nature.* 449, pp. 1078–1082.

Tanino, T., Ito, T., Ogino, C., Ohmura, N., Ohshima, T. and Kondo, A. (2012). Sugar consumption and ethanol fermentation by transporter-overexpressed xylose-metabolizing *Saccharomyces cerevisiae* harboring a xyloseisomerase pathway. *Journal of Bioscience and Bioengineering*, 114(2), pp.209-211.

Tartoff K. D. and Hobbs C. A., 1987, Improved Media for Growing Plasmid and Cosmid Clones, *Bethesda Res. Lab. Focus*. 9, pp.12.

Teale, F. and Weber, G. (1957). Ultraviolet fluorescence of the aromatic amino acids. *Biochemical Journal*, 65(3), pp.476-482.

Teuber, M. and Bader, J. (1976). Action of polymyxin B on bacterial membranes. *Archives of Microbiology*, 109(1-2), pp.51-58.

Thompson, B., Widdick, D., Hicks, M., Chandra, G., Sutcliffe, I., Palmer, T. and Hutchings, M. (2010). Investigating lipoprotein biogenesis and function in the model Gram-positive bacterium *Streptomyces coelicolor*. *Molecular Microbiology*, p.no-no.

Thorstenson, Y.R., Zhang, P., Olson, S. and Mascarenhas, D. (1997). Leaderless polypeptides efficiently extracted from whole cells by osmotic shock. *J. Bacteriol*. 179, pp. 5333–5339.

Timung R., Deshavath N. N., Goud V. V. and Dasu V. V., 2016. Effect of Subsequent Dilute Acid and Enzymatic Hydrolysis on Reducing Sugar Production from Sugarcane Bagasse and Spent Citronella Biomass, *Journal of Energy*, 2016. pp. 1-12.

Tomme, P., Driver, D., Amandoron, E., Miller, R., Antony, R., Warren, J. and Kilburn, D. (1995). Comparison of a fungal (family I) and bacterial (family II) cellulose-binding domain. *Journal of Bacteriology*, 177(15), pp.4356-4363.

Tsujibo, H., Kosaka, M., Ikenishi, S., Sato, T., Miyamoto, K. and Inamori, Y. (2004). Molecular Characterization of a High-Affinity Xylobiose Transporter of Streptomyces thermoviolaceus OPC-520 and Its Transcriptional Regulation. *Journal of Bacteriology*, 186(4), pp.1029-1037.

Tyler, J. (1965). 987. The seed mucilage of Lepidium sativum(cress). Part II. Products of hydrolysis of the methylated mucilage and the methylated degraded mucilage. *Journal of the Chemical Society (Resumed)*, p.5300.

Vagin, A. A., Steiner, R. A., Lebedev, A. A., Potterton, L., McNicholas, S., Long, F., & Murshudov, G. N. (2004). REFMAC5 dictionary: organization of prior chemical knowledge and guidelines for its use. *Acta Crystallographica Section D: Biological Crystallography*, *60*(12), 2184-2195.

Vagin, A., & Teplyakov, A. (2010). Molecular replacement with MOLREP. *Acta Crystallographica Section D: Biological Crystallography*, 66(1), pp. 22-25.

van der Heide, T. and Poolman, B. (2002). ABC transporters: one, two or four extracytoplasmic substrate-binding sites? *EMBO reports*, 3(10), pp.938-943.

Vedadi, M., Niesen, F., Allali-Hassani, A., Fedorov, O., Finerty, P., Wasney, G., Yeung, R., Arrowsmith, C., Ball, L., Berglund, H., Hui, R., Marsden, B., Nordlund, P., Sundstrom, M., Weigelt, J. and Edwards, A. (2006). Chemical screening methods to identify ligands that promote protein stability, protein crystallization, and structure determination. *Proceedings of the National Academy of Sciences*, 103(43), pp.15835-15840.

Vetting, M., Al-Obaidi, N., Zhao, S., San Francisco, B., Kim, J., Wichelecki, D., Bouvier, J., Solbiati, J., Vu, H., Zhang, X., Rodionov, D., Love, J., Hillerich, B., Seidel, R., Quinn, R., Osterman, A., Cronan, J., Jacobson, M., Gerlt, J. and Almo, S. (2014). Experimental Strategies for Functional Annotation and Metabolism Discovery: Targeted Screening of Solute Binding Proteins and Unbiased Panning of Metabolomes. *Biochemistry*, 54(3), pp.909-931.

Viitanen P, Garcia ML, Kaback HR. (1984) Purified reconstituted lac carrier protein from Escherichia coli is fully functional. *Proceedings of the National Academy of Sciences of the United States of America*. 81, pp. 1629-33.

VOGEL, J. (2008). Unique aspects of the grass cell wall. Current Opinion in Plant Biology, 11(3), pp.301-307.

von Heijne, G. (1989). The structure of signal peptides from bacterial lipoproteins. *"Protein Engineering, Design and Selection"*, 2(7), pp.531-534.

Wackett, L. (2011). Engineering microbes to produce biofuels. *Current Opinion in Biotechnology*, 22(3), pp.388-393.

Wagschal K, Franqui-Espiet D, Lee CC, Robertson GH, Wong DWS. (2005). Enzyme-coupled assay for β-xylosidase hydrolysis of natural substrates. *Appl Environ Microbiol*, 71, pp.5318–23.

Wagschal, K., Heng, C., Lee, C. and Wong, D. (2008). Biochemical characterization of a novel dual-function arabinofuranosidase/xylosidase isolated from a compost starter mixture. Applied Microbiology and Biotechnology, 81(5), pp.855-863.

Walker, J. E., Saraste, M., Runswick, M. J. & Gay, N. J. 1982. Distantly related sequences in the alpha-and beta-subunits of ATP synthase, myosin, kinases and other ATP-requiring enzymes and a common nucleotide binding fold. *EMBO J.*, 1, 945-951.

Wang, R. and Kushner, S. (1991). Construction of versatile low-copy-number vectors for cloning, sequencing and gene expression in Escherichia coli. Gene, 100, pp.195-199.

Wagner, S., Baars, L., Ytterberg, J., Klussmeier, A., Wagner, C.S., Nord, O., Nygren, P., Wijk, K. J. and Gier, J. (2007). Consequences of membrane protein overexpression in Escherichia coli. Molecular and Cellular Proteomics, 6, pp. 1527-1550.

Warren,P. and ten Wolde,P. (2004) Statistical Analysis of the Spatial Distribution of Operons in the Transcriptional Regulation Network of Escherichia coli. Journal of Molecular Biology, 342, 1379-1390.

Watanabe, S., Shimada, N., Tajima, K., Kodaki, T. and Makino, K. (2006). Identification and Characterization of l-Arabonate Dehydratase, l-2-Keto-3-deoxyarabonate Dehydratase, and l-Arabinolactonase Involved in an Alternative Pathway of l-Arabinose Metabolism. *Journal of Biological Chemistry*, 281(44), pp.33521-33536.

Wattam, A., Davis, J., Assaf, R., Boisvert, S., Brettin, T., Bun, C., Conrad, N., Dietrich, E., Disz, T., Gabbard, J., Gerdes, S., Henry, C., Kenyon, R., Machi, D., Mao, C., Nordberg, E., Olsen, G., Murphy-Olson, D., Olson, R., Overbeek, R., Parrello, B., Pusch, G., Shukla, M., Vonstein, V., Warren, A., Xia, F., Yoo, H. and Stevens, R. (2016). Improvements to PATRIC, the all-bacterial Bioinformatics Database and Analysis Resource Center. *Nucleic Acids Research*, 45(D1), pp.D535-D542.

Weber, G. & Laurence, D. J. (1954). Fluorescent indicators of adsorption in aqueous solution and on the solid phase. *Biochem J*, 56, xxxi.

Wetmore, K., Price, M., Waters, R., Lamson, J., He, J., Hoover, C., Blow, M., Bristow, J., Butland, G., Arkin, A. and Deutschbauer, A. (2015). Rapid Quantification of Mutant Fitness in Diverse Bacteria by Sequencing Randomly Bar-Coded Transposons. mBio, 6(3), pp.e00306-15.

Whistler, R. and Masak, E. (1955). Enzymatic Hydrolysis of Xylan1. Journal of the American Chemical Society, 77(5), pp.1241-1243.

Whitmore, L. and Wallace, B. (2008). Protein secondary structure analyses from circular dichroism spectroscopy: Methods and reference databases. *Biopolymers*, 89(5), pp.392-400.

Widenhorn K. A., Boos W., Somers J.M and Kay W. W. (1988) Cloning and properties of the *Salmonella typhimurium* tricarboxylate transport operon in *Escherichia coli*. *J. Bacteriol*. 170, 883–888

Wilkens, S. (2015). Structure and mechanism of ABC transporters. F1000Prime Reports, 7.

Wilks, J. and Slonczewski, J. (2007). pH of the Cytoplasm and Periplasm of Escherichia coli: Rapid Measurement by Green Fluorescent Protein Fluorimetry. *Journal of Bacteriology*, 189(15), pp.5601-5607.

Winn MD, Murshudov GN and Papiz MZ. (2003). Macromolecular TLS refinement in REFMAC at moderate resolutions. *Methods in Enzymology*, 374, pp.300-321.

Wise, L. I., (1960) Chemistry of natural gums and mucilages and some related polysaccharides. *Journal of the Franklin Institute*, 270(2), p.151.

Wiseman, T., Williston, S., Brandts, J. and Lin, L. (1989). Rapid measurement of binding constants and heats of binding using a new titration calorimeter. *Analytical Biochemistry*, 179(1), pp.131-137.

Witholt, B., Boekhout, M., Brock, M., Kingma, J., van Heerikhuizen, H. and de Leij, L. (1976). An efficient and reproducible procedure for the formation of spheroplasts from variously grown *Escherichia coli*. *Analytical Biochemistry*, 74(1), pp.160-170.

Woodruff, L., May, B., Warner, J. and Gill, R. (2013). Towards a metabolic engineering strain "commons": An *Escherichia coli* platform strain for ethanol production. *Biotechnology and Bioengineering*, 110(5), pp.1520-1526.

Wolfrum, E., Ness, R., Nagle, N., Peterson, D. and Scarlata, C. (2013). A laboratory-scale pretreatment and hydrolysis assay for determination of reactivity in cellulosic biomass feedstocks. *Biotechnology for Biofuels*, 6(1), p.162.

Worth, H. (1967). The Chemistry and Biochemistry of Pectic Substances. *Chemical Reviews*, 67(4), pp.465-473.

Xiao, H., Gu, Y., Ning, Y., Yang, Y., Mitchell, W., Jiang, W. and Yang, S. (2011). Confirmation and Elimination of Xylose Metabolism Bottlenecks in Glucose Phosphoenolpyruvate-Dependent Phosphotransferase System-Deficient Clostridium acetobutylicum for Simultaneous Utilization of Glucose, Xylose, and Arabinose. Applied and Environmental Microbiology, 77(22), pp.7886-7895.

Xiao, X., F. Wang, A. Saito, J. Majka, A. Schlösser, and H. Schrempf. (2002). The novel *Streptomyces olivaceoviridis* ABC transporter Ngc mediates uptake of N-acetylglucosamine and N,N'-diacetylchitobiose. *Mol. Genet. Genomics* 267, pp. 429–439.

Xu, J., Wang, Y., Xie, S., Xu, J., Xiao, J. and Ruan, J. (2009). Streptomyces xiamenensis sp. nov., isolated from mangrove sediment. *INTERNATIONAL JOURNAL OF SYSTEMATIC AND EVOLUTIONARY MICROBIOLOGY*, 59(3), pp.472-476.

York, A., (2017). Bacterial physiology: An inside job on metabolism. *Nat Rev Micro*. 15, pp.383.

Young, E., Comer, A., Huang, H. and Alper, H. (2012). A molecular transporter engineering approach to improving xylose catabolism in *Saccharomyces cerevisiae*. *Metabolic Engineering*, 14(4), pp.401-411.

Young, E., Tong, A., Bui, H., Spofford, C. and Alper, H. (2013). Rewiring yeast sugar transporter preference through modifying a conserved protein motif. *Proceedings of the National Academy of Sciences*, 111(1), pp.131-136.

Zhang, J., Tang, M., Viikari, L. (2012). Xylans inhibit enzymatic hydrolysis of lignocellulosic materials by cellulases, Bioresour. Technol. 121, pp.8-12.

Zhang,L. et al. (2011) Ribulokinase and Transcriptional Regulation of Arabinose Metabolism in Clostridium acetobutylicum. *Journal of Bacteriology*, 194, 1055-1064.

Zheng, D., Constantinidou C., Hobman, JL., Minchin, SD.. (2004) Identification of the CRP regulon using in vitro and in vivo transcriptional profiling. Nucleic Acids Res., 32. pp. 5874-93.

Zukin, R., Strange, P., Heavey, L. and Koshland, D. (1977). Properties of the galactose binding protein of Salmonella typhimurium and *Escherichia coli*. *Biochemistry*, 16(3), pp.381-386.