THESIS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

Production and investigation of highly thermophilic multi-domain carbohydrate-active enzymes

Daniel Krska

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Biology and Biological Engineering

CHALMERS UNIVERSITY OF TECHNOLOGY

Gothenburg, Sweden 2021

Production and investigation of highly thermophilic multi-domain carbohydrate-active enzymes
Daniel Krska
ISBN 978-91-7905-569-1

Doktorsavhandlingar vid Chalmers tekniska högskola
Ny serie nr 5036
ISSN 0346-718X

Division of Industrial Biotechnology
Department of Biology and Biological Engineering
Chalmers University of Technology
SE-412 96 Gothenburg
Sweden
Telephone + 46 (0)31-772 1000

Cover:
Model of *Ck*Xyn10C-GE15A, as predicted by AlphaFold2

Printed by Chalmers Reproservice
Gothenburg, Sweden 2021

"You've got to be very careful if you don't know where you are going, because you might not get there."

- Lawrence Peter "Yogi" Berra

Preface

This dissertation serves as partial fulfillment of the requirements for obtaining the degree of Doctor of Philosophy at the Department of Biology and Biological Engineering at Chalmers University of Technology. The work was supported by grant awarded to Johan Larsbrink from the following funders: the Swedish Research Council Formas (Future Research Leader Grant, Dnr 2016-01065), the Swedish Energy Agency (Biofuel program – Biochemical methods, Dnr 2016-011207), and the EU Interreg program fund via the MAX4ESSFUN project (Ref. no. CTH-010). The PhD studies were carried out between May 2017 and October 2021 under the supervision of Assoc. Prof. Johan Larsbrink and co-supervision of Prof. Lisbeth Olsson. The thesis was examined by Prof. Pernilla Wittung-Stafshede.

The majority of the work in this thesis was carried out at the Division of Industrial Biotechnology (IndBio) at Chalmers University of Technology. Crystallization experiments were performed at Copenhagen University and x-ray diffraction data were collected at the MAX IV Laboratory, the ESRF, and the PETRA III Beamline Facilities by Jens-Christian Navarro Poulsen. Small angle x-ray scattering data were collected at the ESRF by Dr. Kim Krighaar Rasmussen. Isothermal titration calorimetry experiments were performed at the University of Michigan by Haley A. Brown, Adeline L. Morris, and Nicole M. Koropatkin. Differential scanning fluorimetry experiments were performed at Copenhagen University by Yusuf Theibich.

Daniel Krska
October 2021

Production and investigation of highly thermophilic multi-domain carbohydrate-active enzymes

Daniel Krska

Division of Industrial Biotechnology
Department of Biology and Biological Engineering
Chalmers University of Technology

Abstract

With the looming threat of climate change caused largely by an excess of carbon dioxide in the atmosphere, recent scientific efforts have focused on the substitution of fossil fuels and other polluting compounds with more environmentally conscious choices. To this end, the investigation of biomass as both a renewable source of energy and as a chemical basis to produce high-value products is being extensively investigated. Although plant biomass is complex, it is also an extremely rich carbon source, and microorganisms in a plethora of environments have evolved to exploit it. These microorganisms produce carbohydrate-active enzymes (CAZymes) to degrade the plant biomass into components that can be utilized for their growth. The deeper study of these enzymes, especially those containing multiple enzyme domains, can elucidate their mechanisms of action, and guide their exploitation for industrial purposes.

This thesis consists of the characterization of two different multicatalytic CAZymes from different bacteria found in extremely different environments. The enzymes both contain CE15 (carbohydrate esterase family 15) domains, which have not previously been studied in a multicatalytic context. *Ck*Xyn10C-GE15A from the hyperthermophilic *Caldicellulosiruptor kristjanssonii* consists of a GH10 (glycoside hydrolase family 10) xylanase linked to a CE15 enzyme, and additionally contains two CBM22 (carbohydrate binding module family 22) and three CBM9 domains. A second enzyme, *Be*CE15A-Rex8A from the gut bacterium *Bacteroides eggerthii*, consisting of a GH8 xylan-targeting domain and a CE15 domain was also investigated. Although the catalytic domains in both enzymes were active, no synergy was seen between them, respectively. As these enzymes were difficult to produce recombinantly, a new technique using split intein-mediated fusions to produce multicatalytic enzymes was investigated, with results showing that the produced enzymes remain catalytically active after the fusion event.

The work presented in this thesis contributes to the understanding of multidomain enzymes and the synergy (or lack thereof) of xylanases in combination with CE15 domains. It also provides structural insights into a number of highly thermophilic CAZyme domains, and has implications for industrial biorefinery applications.

Keywords: *Caldicellulosiruptor*, carbohydrate-active enzymes, multidomain enzymes, carbohydrate esterase, xylanase, thermostable enzymes, plant biomass degradation, protein structure

List of Publications

This thesis is based on the following papers, which are referred to in the text by their Roman numerals.

I **Krska D** and Larsbrink J (2020). Investigation of a thermostable multi-domain xylanase-glucuronoyl esterase enzyme from *Caldicellulosiruptor kristjanssonii* incorporating multiple carbohydrate-binding modules. *Biotechnology for biofuels*, 13, 68.
DOI: https://doi.org/10.1186/s13068-020-01709-9

II **Krska D**, Mazurkewich S, Brown HA, Theibich Y, Poulsen JCN, Morris AL, Koropatkin NM, Lo Leggio L, and Larsbrink J (2021). Structural and functional analysis of a multimodular hyperthermostable xylanase-glucuronoyl esterase from *Caldicellulosiruptor kristjansonii*. *Biochemistry,* 60, 27, 2206-2220.
DOI: https://doi.org/10.1021/acs.biochem.1c00305

III Kmezik C*, **Krska D***, Mazurkewich S and Larsbrink J (2021). Characterization of a novel multidomain CE15-GH8 enzyme encoded by the human gut bacterium *Bacteroides eggerthii*. *Sci Rep*, Sep 3;11(1):17662. DOI: 10.1038/s41598-021-96659-z.

IV **Krska D** and Larsbrink J (2021). Creation of large multidomain proteins using split intein technology. *Manuscript.*

*These authors contributed equally.
Reprints were made with permission from the respective publisher

Contribution Summary

I First author. Shared in study conception. Performed all experimental work and drafted the manuscript. Shared in editing of the manuscript.

II First author. Shared in study conception. Performed sequence alignments, produced and purified proteins, performed affinity PAGE, drafted the manuscript, and prepared the majority of figures. Shared in structural work. Shared in manuscript editing.

III First author (shared). Shared cloning and production of enzymes. Biochemically characterized enzymes on model substrates and various xylans. Generated model structures. Shared in writing and editing of the manuscript.

IV First author. Shared in study conception. Performed all experimental work and drafted the manuscript. Shared in editing of the manuscript.

Table of Contents

Abbreviations

AA – Auxiliary Activity
CASP – Critical Assessment of protein Structure Prediction
CAZy – Carbohydrate Active enZyme Database
CAZyme – Carbohydrate Active enZyme
CBM – Carbohydrate Binding Module
CBM1 – Carbohydrate Binding Module Family 1
CBM3 – Carbohydrate Binding Module Family 3
CBM7 – Carbohydrate Binding Module Family 7
CBM9 – Carbohydrate Binding Module Family 9
CBM22 – Carbohydrate Binding Module Family 22
CBM33 – Carbohydrate Binding Module Family 33
CE – Carbohydrate Esterase
CE15 – Carbohydrate Esterase Family 15
Cryo-EM – Cryogenic Electron Microscopy
ESRF – European Synchrotron Radiation Facility
GGM – Galactoglucomannan
GH – Glycoside Hydrolase
GH8 – Glycoside Hydrolase Family 8
GH9 – Glycoside Hydrolase Family 9
GH10 – Glycoside Hydrolase Family 10
GH11 – Glycoside Hydrolase Family 11
GH48 – Glycoside Hydrolase Family 48
GT – Glycosyltransferase
LCC – Lignin-Carbohydrate Complex
LPMO – Lytic Polysaccharide MonoOxygenase
MeGlcA – (4-*O*-methyl)-glucuronic acid
NMR – Nuclear Magnetic Resonance
PDB – Protein Data Bank
Phyre – Protein Homology/analogY Recognition Engine
PL – Polysaccharide Lyase
PUL – Polysaccharide Utilization Locus
Rex – Reducing-end-xylose releasing exo-oligoxylanase
SAXS – Small Angle X-ray Scattering

Chapter 1: Introduction

Traditional energy and chemical production processes have historically been focused around the use of petroleum products and other fossil resources (1). However, these sources have several drawbacks, in that they are non-renewable, and inherently damaging to the climate (1,2). With the evidence of these drawbacks becoming increasingly apparent, the search for more sustainable and climate-friendly alternatives has led to several promising possibilities. According to the Intergovernmental Panel on Climate Change, in order to be able to meet the goals of the Paris Agreement and limit global warming to 1.5°C, biofuels and other bioenergy sources are needed to play a large part (3,4).

In addition to bioenergy being required to replace fossil energy to meet the targets set out in the Paris Agreement, other sources of greenhouse gas emissions from petroleum products must be considered. Emissions from plastic lifecycles are a significant contribution to overall greenhouse gas emissions, and must also be lowered to meet climate targets (5). Although plastics only saw large scale production from the 1950s, they have quickly grown to become more than 20% of solid waste produced annually (6,7). Plastics account for over a gigaton of greenhouse gas emissions yearly (more than 2% of total emissions), and microplastics significant hamper the carbon fixation capacity of the natural environment (5). Production of plastic from non-fossil sources could provide a significant tool in meeting Paris Agreement targets.

A potential way to produce plastics (and other products) from non-fossil sources that will be expanded on in the next chapter is through the use of enzymes. Enzymes are, in simplest terms, biological catalysts capable of greatly increasing the rate of a chemical reaction (8). They are produced by every living organism and are an essential component to life on Earth. Enzymes generally only act on one substrate, or at most, a small range of similar substrates, meaning a different enzyme is generally needed for each reaction that requires catalysis. Different enzymes have evolved various properties to make them more suited for specific tasks, such as substrate specificity, reaction rate, thermostability, or ionic tolerance (8). The tremendous variety and functionality of enzymes produced by evolution is unparalleled by almost any other organic process, with over 8000 different enzyme classes and countless enzymes within those classes known today, and research is just starting to produce artificial enzymes that can rival natural efficiency (9,10).

1.1 Aims and Structure of this Thesis

Before work began on the research contained in this thesis, three overall aims were identified. The first was to study multicatalytic enzymes, specifically those which act on plant biomass, to greater understand how they function and what the relationship between the catalytic domains is. These types of enzymes, containing two or more catalytic domains within the same polypeptide chain, are relatively common amongst plant biomass degrading organisms, but they remain fairly unstudied compared to their single-domain counterparts. This investigation of multicatalytic enzyme activities is conducted in **Paper I**, **Paper II**, and **Paper III**.

A second aim of this thesis work was to further investigate and add to the understanding of carbohydrate esterase family 15 (CE15) enzymes. CE15 is thought to be responsible for breaking ester bonds between hemicellulose and lignin, although it has only recently become a more researched enzyme family. The work in this thesis aimed to expand the knowledge of this family, and hopefully give an indication of whether this family of enzymes could be useful in an industrial context. The investigation of CE15 enzymes is conducted in **Paper I**, **Paper II**, and **Paper III**.

The final aim of this thesis was the construction of a library of non-natural multicatalytic enzymes, in order to determine if rationally designed multicatalytic enzymes could be useful for industrial purposes. Using knowledge learned from the first thesis goal, a DNA-based library of different possible combinations of catalytic domains was to be constructed and tested. This aim was explored in **Paper IV**.

The overall thesis is structured into two major parts. The first part of the thesis is designed to summarize the research results of the contained work, and to provide a background for the research contained in the second part. The next chapter will discuss the societal context of the research, including potential applications. The third chapter discusses lignocellulose, and its structural diversity. Lignocellulose-degrading enzymes are discussed in more detail in chapter four, along with associated protein modules. Chapter five contains information on multicatalytic enzymes, their producers, and how they differ from other enzyme architectures. In chapter six, a brief discussion on designer enzymes and methods of assembling multicatalytic enzymes is found. A summary of the conclusions of the work is found after chapter six, along with an outlook of where the work may lead in the future.

The second part of this thesis contains the scientific publications that represent the bulk of the research that was performed for this thesis. **Paper I** investigates a novel xylanase-glucuronoyl esterase multicatalytic enzyme from *Caldicellulosiruptor kristjanssonii* (Figure 1.1 A), and characterizes the behaviour of the catalytic domains. **Paper II** expands further on the investigation of this enzyme, focusing largely on structural characterization and investigation of associated non-catalytic

domains. In **Paper III**, a different multicatalytic enzyme from *Bacteroides eggerthii* is characterized (Figure 1.1 B). **Paper IV** contains work on the construction of a multicatalytic enzyme library.

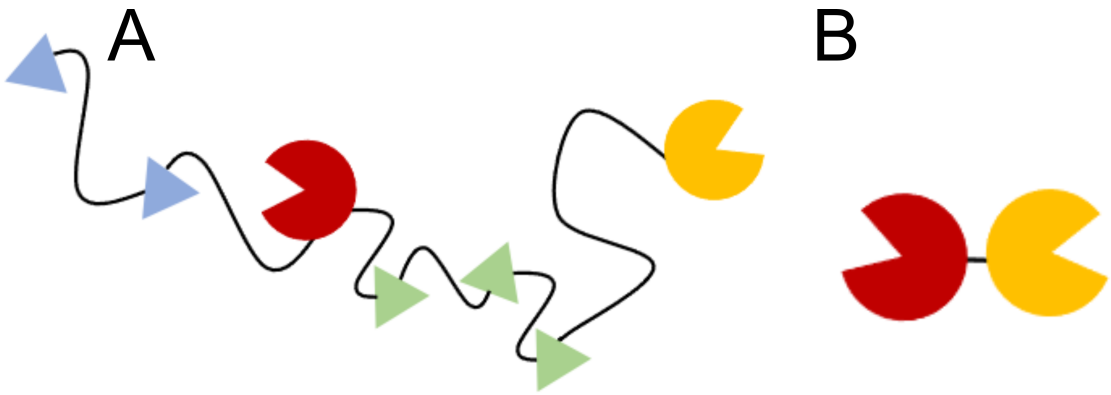

Figure 1.1: Schematic diagrams of the enzymes that are the primary focus of this thesis. A) *Ck*Xyn10C-GE15A from *C. kristjanssonii*, studied in **Paper I** and **Paper II**. B) BeCE15A-Rex8A from *B. eggerthii*, studied in **Paper III**. Blue triangles represent carbohydrate binding module family 22 (CBM22) domains, and green triangles represent carbohydrate binding module family 9 (CBM9) domains. Red circles with a missing wedge are representative of xylanases (although they belong to different families in A and B), and yellow circles with a missing wedge represent carbohydrate esterase family 15 (CE15) domains.

Chapter 2: The Bioeconomy

An important step in moving away from a fossil-based economy is the conversion to a bioeconomy. The bioeconomy is defined by the European Union as "the production of renewable biological resources and the conversion of these resources and waste streams into value added products, such as food, feed, bio-based products and bioenergy" (11). It is important to establish an economy in which resources are not extracted to depletion, but instead harvested sustainably and, ideally, in a carbon-neutral way. The idea of a "circular bioeconomy" expands on these concepts and has come into greater focus recently by focusing on the entire lifecycle of products produced in a bioeconomy setting, with emphasis on the reduction of waste generation and improvements in product longevity (12). The European Commission now considers circularity to be one of the most important aspects of an overall bioeconomy (13), indicating a greater need for sustainability in a bioeconomy.

There are several important aspects to consider during conversion to a bioeconomy. Arguably the most important is the consideration of input. In the framework of the bioeconomy, biomass is the input which is consumed to create new products. However, humans already consume a large amount of biomass produced in the form of food (both directly and indirectly through animal feed), as well as wood used to produce e.g. paper, furniture, housing, and heating (unless otherwise specified, in this thesis, "biomass" refers to biological material derived from plant sources) (14). The need for food production is expected to increase in the coming decades, and existing farmland is already beginning to struggle with issues such as soil degradation, drought, and other impacts of climate change (15,16). While the needs of industry and agriculture could be balanced, it must be done carefully to avoid famine – food security is considered a cornerstone of the bioeconomy (14).

Along with issues surrounding alternate uses of feedstock, there are more fundamental economic issues to consider in a bioeconomy (17). In general, the price of a feedstock must be high enough that it is economically advantageous to produce, but low enough that the end product is price-competitive with alternative sources (in this case referring to fossil resources, which unfortunately have a decades-long head start in optimisation of production) (17,18). Governmental policy decisions play a role in this, whether in mandating bio-based products, or removing or lowering the extremely large subsidies currently given to fossil fuel products (14,17). While the in-depth economics of this situation is not the focus of this thesis, it is important to emphasize that increasing efficiencies and lowering costs at all stages is essential to the success of the bioeconomy.

2.1 Biorefineries

The production of biofuels and bioplastics (along with other bio-based materials) can be conducted in a facility known as a biorefinery (19). A biorefinery is defined by the US National Renewable Energy Laboratory as "a facility that integrates biomass conversion processes and equipment to produce fuels, power, and chemicals from biomass" (19,20). Alternative definitions, such as the one from the International Energy Agency, restrict the concept to a facility utilizing biomass in a sustainable way (19,21). In general, a biorefinery uses input biomass, from e.g. forestry, agriculture, marine sources, or industrial and municipal waste, and converts it into useable end products (22). Biorefineries can be used to replace existing, high carbon output processes, or be used for novel production of high-value products, and do so while being both energy and material efficient (19).

Many functional biorefineries are currently in operation throughout the world, producing a variety of end products from various different feed stocks. The Spanish company Abengoa has constructed and operates a number of biorefineries for energy production, as well as bioethanol and bio-based jet fuel production. Lenzing, a company originally from Austria, produces multiple end-products from their biorefineries in Europe, including acetic acid, sodium sulphate, and xylose. In Norway, Boregaard uses wood to produce bioethanol, vanillin, biopolymers, and other products. Within Sweden, the Domsjö Fabriker, produces products such as cellulose and bioethanol from wood. These facilities are just some of the many worldwide that are already working to produce materials utilizing the biorefinery concept.

2.2 Biorefineries: A History

Throughout the development of biorefineries, challenges with process efficiency and final product yield have hampered movement toward a fully realized bioeconomy. Arguably, the first evidence of a biorefinery can be traced back as far as 9000 BCE, when the first indications of an ethanol distillation process can be found in China (23). Since then, improvements in biomass utilization have led to the production of many different end products, and the use of many types of input materials. Industrial scale biorefineries have existed since at least the late 1800's (23). In 1895, Germany saw the opening of the first industrial-scale lactic acid production plant by C. H. Boehringer Sohn, a company that is still in operation as Boehringer Ingelheim (20,23,24). Around the same time, ethanol was produced at an industrial scale as a fuel for internal combustion engines (eventually out-competed by petroleum-based fuels) (20). Ethanol was not the only biofuel used at the time, as some early diesel engines were demonstrated (by Rudolf Diesel himself) to run on vegetable oils (25).

One of the first large-scale processes for the conversion of woody biomass (biomass that can not alternatively be used as food) was developed and patented in 1909, which aimed to turn sawdust into ethanol (26). However, poor yields at the time proved the process to be non-viable. Further development of the idea was done during the period between World War I and World War II, by Nobel laureate Dr. Friedrich Bergius (23,26). His process was capable of nearly 50% yields of ethanol from wood waste. Dr. Bergius also utilized parts of the wood not converted into ethanol to produce other compounds (26). As the exploitation of fossil resources for both energy and petroleum-derived chemical products became cheaper and easier, the research and development focus shifted, and biorefinery advances lessened (20).

2.3 Biorefineries: The Future

With increasing concerns about pollution, greenhouse gas emissions, and the evident effects of climate change, focus has shifted back towards biorefineries to meet societal energy and chemical product needs. Governments around the world have and continue to provide various incentives towards the development and operations of biorefineries (27-29). However, many improvements are still necessary to compete with the scale, dominance, and cost-effectiveness of the petroleum industry, and biorefinery research now aims to make the processes more efficient and cost effective (30). While the petroleum industry has a significant head start in establishing supply chains and processes to keep costs low, technology to help process and degrade biomass has advanced considerably in the interim (17). While many older biorefinery methods utilized harsh chemical processes and could only produce limited end products, biotechnological advances have greatly changed what is possible to do with biomass. The capability to manipulate the genomes of different organisms has opened vast new possibilities for the utilization of biomass, both in increasing utilization of all components, as well as in the creation of new valuable chemicals from the different biomass feedstocks (31-33).

As we have come to better understand the chemical structure of plant biomass, new methods have been developed to better degrade and utilize the components. Biorefineries can now replace harsh chemicals with enzymatic catalysis to achieve the same or better results in more sustainable ways (34,35). The use of enzymes greatly reduces environmentally harmful waste products produced by traditional methods, and can lower the volume of non-biomass input needed to obtain the same output amounts (34,35). Greater understanding and advances in molecular biology techniques have enabled production of specific enzymes that can be used to target desired chemical bonds within plant biomass (35-37). Different enzymes can be combined into enzyme cocktails, which allow for the degradation of specific types of biomass into base components (35,38). Developing these cocktails is an ongoing process, as new discoveries that lead to potential improvements are constantly being made (35,38). The work presented within this thesis is primarily focused on the

investigation of novel enzymes and protein domains which have the potential for increasing the efficiency of such cocktails, as well as novel ways of constructing enzymes for use in these cocktails.

2.4 Enzymes Within Biorefineries

Enzymes used in biorefineries originally come from natural sources, and the organisms which evolved these enzymes use them to operate successfully in various environments. Although many enzymes exist intracellularly, and are expected to be exposed to a relatively constant environment, others are secreted into the extracellular environment (8). This has led to the need for these enzymes to be able to withstand and function in a variety of different conditions. Consequently, enzymes from more extreme environments are ideal to explore for industrial applications, as many less robust enzymes are simply unable to function in harsh industrial processes (39). This section will focus on different extreme conditions faced by enzymes, and the adaptations that have consequently evolved.

2.4.1 Temperature

Enzymes that can withstand temperature extremes are highly sought after in an industrial context, as many processes occur at temperatures outside of normal physiological range (40). Higher temperatures are important in industrial processes as the increase in temperature often leads to corresponding increases in substrate solubility and diffusion rates of substrates and products, as well as a decreased viscosity of the reaction medium (41). Perhaps most importantly, performing processes at high temperature greatly reduces the risk of contamination from environmental microorganisms (42,43).

The overall thermostability of an enzyme can be measured using three parameters: enzyme half-life at a given temperature ($t_{½}$), the free energy of stabilization of the enzyme (ΔG_{stab}) and the melting temperature of the enzyme (T_m) (44,45). These three factors are intrinsically related, and correlations can be noted among the three, e.g. increased ΔG_{stab} correlates with increased T_m (46). Despite these relatively simplistic measures for thermostability, the actual mechanisms behind it can be complex, and there is no universal set of enzyme properties that is present in every thermostable enzyme (47).

Several factors influence enzyme thermostability. In the literature, there are many examples of disulfide bridges leading to increased thermostability, especially at the N-terminus of proteins (48-50). However, not all thermostable enzymes contain disulfide bridges (51). A lack of disulfide bridges can be seen in the structures of protein domains determined in **Paper II**, despite the apparent thermostability of the overall enzyme. Disulfide bridges influence enzyme rigidity, which itself is a major factor impacting thermostability (52,53). Several other factors can contribute to

protein rigidity, and the most important is arguably the number of short helices in the protein (54). These helices decrease the protein flexibility when they replace connecting loops existing in related mesophilic proteins (54-56). The formation of salt bridges between charged amino acid residues also has a significant impact on increasing protein thermostability (54,57). Though these bridges make little contribution to stability at room temperature, they are significant in stability at higher temperatures (54,57). Thermostable enzymes show several other advantageous properties, outside of their thermostability. Industrial processes can contain harsh denaturants, detergents, and organic solvents, and thermostable enzymes often show a positive correlation with the ability to withstand such conditions (58). Thermostable enzymes also show a positive correlation between temperature stability and resistance to proteolysis, suggesting that they are more stable overall than their mesophilic counterparts (40,59).

A major concern in industrial lignocellulosic biorefineries is microbial contamination, and thermostable enzymes help to reduce this risk (60,61). Thermophilic enzymes can also be seen, in general, to have higher rates of reaction than their lower-temperature counterparts, due to increased specific activity and lower fluid viscosity of the medium at high substrate concentrations (60,61). Interestingly, thermostable enzymes appear to have less susceptibility to inhibition by lignin present in the reaction mixture than their mesophilic counterparts, even though the impact of inhibition has been seen to increase with an increase in temperature (62).

Although less studied, enzymes at the opposite end of the temperature tolerance spectrum can also be useful for industry. These psychrophilic enzymes, mainly produced by deep sea organisms, possess many advantageous properties for industry (63). While the high- and low-temperature enzyme utilization strategies obviously cannot be used at the same time in a biorefinery, it seems likely that the industrial niche for both exists.

2.4.2 pH

Many microorganisms survive and thrive at very low pH levels, as extreme as below pH 1 (64). As one of the most common pre-treatment methods for lignocellulosic biomass is an acid hydrolysis step, acidophilic enzymes could be a highly useful addition to a lignocellulosic biorefinery (65). Acidophilic enzymes are still a relatively less studied class of enzymes, however, several acidophilic xylanases and cellulases have been discovered thus far (64,66,67). Many of these acid-stable enzymes are also thermostable, providing an added advantage for their use in industrial processes (64). Also enzyme activity at high pH has been documented in the scientific literature (68). Like acidophilic enzymes, not much focus has been directed towards these alkaliphilic enzymes. Many of the commercialized alkaliphilic enzymes are utilized in detergents, including commercial cellulases (69).

The investigation of alkaliphilic enzymes for usage in industrial biorefineries has thus far been fairly limited, however.

2.4.3 Salinity

Halophilic enzymes are those which operate best with high concentrations of salt in solution, some of which can function at 5M or higher concentrations (39). This halophilicity is often conferred via an increase in percentage of acidic amino acids within the protein, and a corresponding decrease in basic ones compared to less halotolerant counterparts (70). In addition to halophilicity, many halophilic enzymes show tolerance to a wide range of pH values and temperatures, suggesting that these are highly stable enzymes (71,72). The potential drawbacks of halophilic enzymes, including their requirements for specific salts (rather than a general high-ionic strength medium), as well as often displaying low solubility in aqueous media, have led to them not being highly utilized industrially at this point in time (73-75).

2.4.4 Inhibitors

The rate of enzyme activity is a function of many properties of the environment in which an enzyme finds itself, not the least of which is the presence of inhibitors (76). Inhibitors can function through several different mechanisms, but all involve the inhibitor molecule binding to the enzyme being inhibited (77). Often this occurs through binding of the inhibitor at the catalytic site, but this is not always the case (77). Inhibitors are of great relevance to biorefinery enzymes. Firstly, inhibitors can aid in the study of enzyme mechanisms by helping to lock the enzyme in a mid-reaction conformation long enough to obtain structural information (78). This allows researchers to gain insight into the mechanism of action of the enzyme and can possibly inform later enzyme engineering efforts (78). Secondly, and more relevant to applications, lignocellulose pretreatment often produces by-products that inhibit enzymatic reactions, and the ability of an enzyme to work despite the presence of these inhibitory products is a key factor that influences its use in a biorefinery (79).

Chapter 2: Summary

- A transition to a bioeconomy is necessary for a sustainable, climate friendly future
- Biorefineries aim to produce useful products from renewable, biological sources
- Enzymes are used within biorefineries to degrade the input material into its component sugars, which can then be used to construct the end products
- There are a variety of enzyme properties that can be beneficial in this context, including thermostability, pH tolerance, halotolerance, and inhibitor tolerance

Chapter 3: Lignocellulose

In order to most efficiently utilize plant biomass (in both the biorefineries discussed in chapter two, and in any other applications), it is essential to understand what, exactly, plant biomass is. There are three major components that make up the majority of lignocellulosic plant biomass: polysaccharides, lignin, and extractives and ash (80). While the exact amounts of each component vary depending on the initial source (Table 1), these major components are present in almost every source available, and major types are introduced in this chapter.

3.1 The Cell Wall

In plants, the cell wall is a structure which encloses each cell (81). It is primarily made up of polysaccharides, lignin, and glycoproteins, with different plant cell types each showing different compositions, proportions, and structures (82,83). The cell wall structure can be divided into two separate types: the primary and secondary cell walls (81,82). All plant cells are surrounded by a primary cell wall, which is thin and extensible (81,83). The secondary cell wall is formed in some plant cell types after the cell stops growing, and consists of new layers of material deposited inside the primary cell wall, making the overall cell wall more rigid (81,83). This secondary cell wall is necessary for terrestrial plants to grow upright (83). The exact composition of both of these cell walls is determined by a number of factors, including plant species, cell type, and light exposure, among others (84).

The primary cell wall consists mainly of polysaccharides, most notably cellulose, as well as glycoproteins (85,86). Additionally, it generally contains some or all of the following polysaccharides: xylan, xyloglucan, and β(1→3, 1→4)-D-glucan (mixed linkage glucan). There may also be lignin and several other minor non-carbohydrate compounds incorporated into the primary cell wall (85). All of these components are combined in an intricate and complex structure through both covalent and non-covalent bonding (85,87). Modern cell wall models posit that xyloglucan can "glue" portions of the cellulose fibrils together into bundles, as well as fill some of the space in between (88). Within these models, large amounts of pectin exist throughout the entire primary cell well (88). Recent work has shown that this leads to a highly heterogeneous distribution of cell wall components and cell wall strength, the reasons for which are not entirely clear (89).

Secondary cell walls are again based largely on polysaccharides, mainly cellulose, accompanied by greater quantities of xylan, lignin, and glucomannan, with the lignin fraction showing the most significant increase, and less xyloglucan and pectin (88).

Table 1: Composition of various lignocellulose sources. A dash symbol indicates that a value was not provided by the sou

Source	Sugarcane Bagasse	Sugarcane Straw	Poplar	Willow	Eucalyptus	Pine	Switchgrass (Whole Plant)	Switchgrass (Leaves)
Polysaccharides (w/w %)	51.6-72.5	52.8-64.4	58-72	64.9-80	60.76	69	59.07-59.94	51.91-56.7
Cellulose (w/w %)	28.4-45.5	33.3-36.1	42-49	34.8-45	48.07	42	33.11-33.65	28.24-31.66
Hemicellulose (w/w %)	22.7-27.0	18.4-28.9	16-23	30.1-35	12.69	27	26.1-26.96	23.67-25.04
Lignin (w/w %)	19.1-32.4	25.8-40.7	15.4-29.1	20-29	25.9-33.2	26.8-28	17.35-18.36	15.46-17.29
Extractives & Ash (w/w %)	1.5-10.1	2.5-23.2	–	–	–	–	–	–
References	(90)	(90)	(91)	(91)	(91)	(91)	(92)	(92)

In this structure, the cellulose fibrils form loosely bound bundles which are coated with a xylan-lignin complex. Glucomannan chains hydrogen bond with the cellulose bundles, linking them together (88). The secondary cell wall (when present) is generally the larger of the two cell wall structures, and comprises the bulk of cell wall material (86).

3.2 Polysaccharides

Polysaccharides, the main components of plant cell walls discussed above, are chains of at least ten sugar molecules in length (chains of two sugars are known as disaccharides, and 3-9 sugar chains are oligosaccharides) (95). The linkages between monosaccharides (individual sugar molecules) within a larger molecule are generally designated in the format (1→4), indicating that the glycosidic linkage is formed between the hydroxyl group of the anomeric carbon of one monosaccharide and the hydroxyl group of the fourth carbon of the second monosaccharide (96). Glycosidic linkages will always follow the format in (X→Y) when written in this thesis (and often outside this thesis as well)(97). By convention, a polysaccharide is named by replacing the -ose suffix of the monosaccharide with an -an suffix, for example, a purely xylose-containing polysaccharide becomes xylan (98). There do exist many exceptions to this rule of polysaccharides that were discovered and named before the convention was adopted, for example, cellulose, starch, and pectin (98).

The ends of a polysaccharide are referred to as reducing and non-reducing ends (99). Reducing ends comprise an anomeric carbon not involved in a glycosidic bond, leaving it free to be oxidized by an appropriate oxidizing agent. This does not occur with the cyclic form of the sugar residue, only the linear form (which exists in equilibrium with the cyclic form in solution if not confined in a glycosidic bond), and does not occur in every polysaccharide. Consequently, in a non-reducing end of a polysaccharide, the anomeric carbon of the sugar residue is involved in a glycosidic bond (99). In addition to the linear backbone of polysaccharides, they can also have "branches" off the main chain, that is, sugar side chains that are not part of the main backbone (100). The sugars making up the branches can be different to those making up the backbone, so theoretically countless variations are possible, although far fewer have been observed as naturally occurring (97,101). These branches serve a number of different functions biologically, depending on the specific polysaccharide backbone and branches, although an important feature is a general increase in solubility of branched polysaccharides compared to unbranched (97). Some of the most biorefinery-relevant polysaccharides are described in the following subsections.

3.2.1 Cellulose

Cellulose is the most abundant naturally occurring polymer on earth (102), and can comprise up to 90% of plant biomass (103), although it is more commonly 40-50%

of the dry weight of lignocellulosic feedstocks (22). It can also be considered to be a fairly simple component of plant biomass, as it is composed exclusively of D-glucose units connected through β(1→4)-glycosidic bonds (22,99). Despite being composed solely of straight glucan chains, cellulose forms highly crystalline and insoluble crystals (crystalline fibers), making it insoluble in water and challenging to hydrolyze. However, successful hydrolysis of cellulose results in individual glucose monomers, which can easily be utilized in later steps of a biorefinery process (22).

3.2.2 Hemicellulose

Hemicellulose is the name given to a wider variety of polysaccharides, although the exact definition as to what constitutes a hemicellulose polymer can be unclear (104). Traditionally, hemicelluloses were defined based on extractability with an alkaline treatment, however, this definition does not fully encapsulate some compounds which are considered to be hemicelluloses, and includes others which are not (104). Indeed, what constitutes a hemicellulose is under some debate in scientific literature. In general, hemicelluloses can be defined as equatorial β(1→4)-linked polysaccharides that are not cellulose (104). Regardless of the definition, hemicelluloses are an important and often underutilized source of sugar for biorefineries. Some of the more commonly utilized hemicelluloses include various mannans, mixed-linkage glucan, xylans, and xyloglucans (105). These polymers consist of a variety of sugars including, but not limited to, the pentoses arabinose and xylose, as well as the hexoses fucose, galactose, glucose, mannose and rhamnose (105). Due to this, hemicellulose utilization is more difficult than cellulose utilization. Additionally, the composition and proportions of hemicelluloses vary greatly between different plants and even different plant tissues (104,106,107).

3.2.2.1 Xylan

Xylan is the most abundant hemicellulose in both hardwood trees and grasses, and accounts for approximately one third of the renewable carbon on earth (105,108-111). Given its abundance, xylan is the most important hemicellulose for biorefineries, and the most important feedstock for lignocellulosic biorefineries after cellulose. Xylan is traditionally thought to have a backbone consisting of β(1→4) linked xylose sugars, which can additionally have a variety of appended carbohydrate and non-carbohydrate moieties (Figure 3.1) (105,112,113). However, β(1→3) xylan is also known, existing mostly in algal cell walls (114). In either case, xylan has also been known to cover cellulose fibrils in plants (exactly how it interacts with and covers cellulose is determined by the specific pattern of substitutions on the backbone), meaning that in order to degrade the cellulose, one must first remove or degrade xylan (110). Chemical methods to remove xylan often leave it in an unusable state, so enzymatic methods are preferred in order to most efficiently use the input material (115).

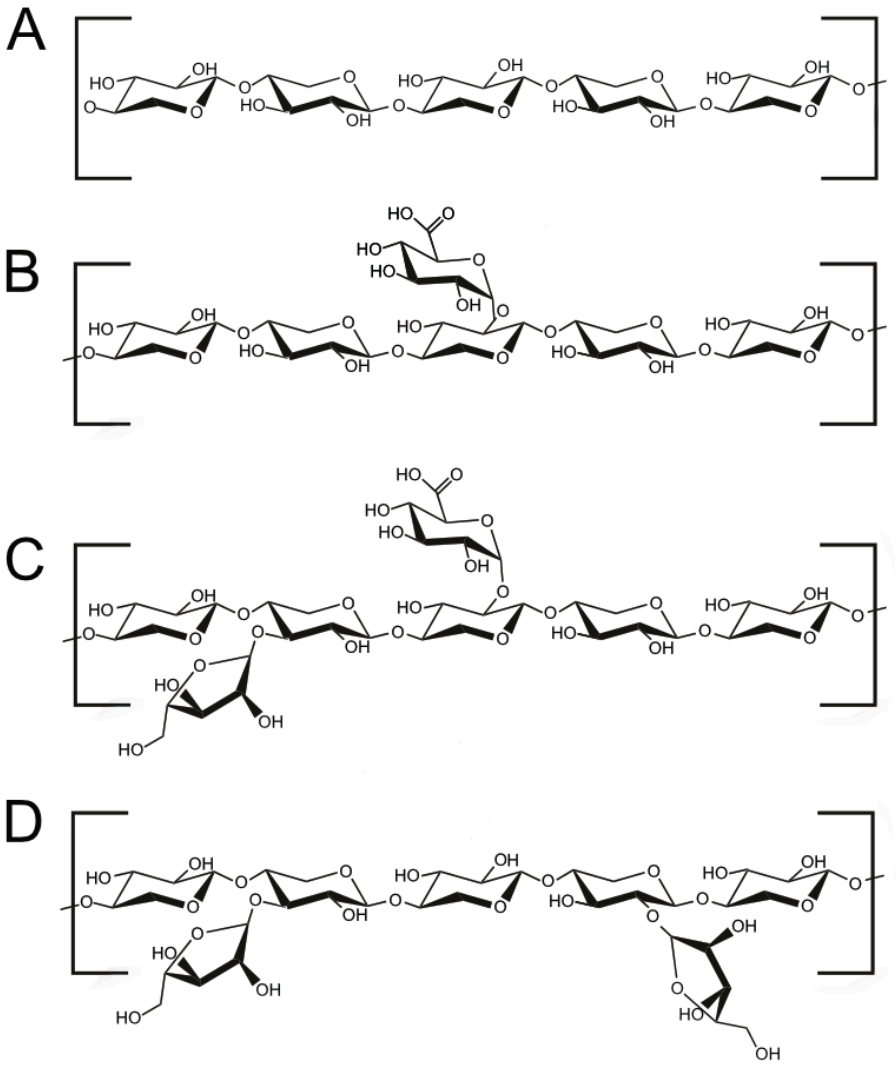

Figure 3.1: Xylan types. Homoxylan (A), glucuronoxylan (B), glucuronoarabinoxylan (C), and arabinoxylan (D) are depicted as chemical structures. Arabinose is shown linked to the (1→4) xylan backbone in a (1→3) configuration (C, D) and a (1→2) configuration (D). Glucuronic acid is shown linked in a (1→2) configuration (B, C).

On a structural level, xylans from different sources can be very different (116). Depending on its substitutions, xylans can be subdivided into arabinoglucuronoxylan, arabinoxylan, glucuronoarabinoxylan, glucuronoxylan, heteroxylan, and homoxylan (117). The various branching patterns and substitutions are too numerous to mention individually in this thesis, and leads to xylans from different sources having diverse compositions and properties (116). This can lead to experimental difficulties in comparing xylan-acting enzymes, as both commercial and non-commercial xylan sources may not be directly comparable (116).

Substitutions on the xylan backbone are known to have a significant impact on xylan behaviour, as well as its interactions with cellulose (118). Major xylan substitutions include acetyl groups linked at the 2 or 3 carbon position, α(1→2)- and α(1→3)-L-

arabanose, and α(1→2)-(4-*O*-methyl)-glucuronic acid (MeGlcA). In interactions with cellulose, the arabinose residues stabilize the interaction of individual xylan chains with cellulose, with the (1→2) linked arabinose providing much greater stabilizing effects (although the (1→3) linked arabinose seems to form more contacts with cellulose overall). MeGlcA moieties can also cross-link chains via Ca^{2+} ions, providing a strong stabilizing effect. At high temperatures, MeGlcA also stabilizes the xylan-cellulose interaction through hydrophobic effects (118). Additionally, MeGlcA is thought to be directly involved in the linkage between xylan and lignin, contributing to lignocellulose recalcitrance (119) (discussed in 2.4, below).

3.2.2.2 Mannan

Mannans, polysaccharides consisting of a mannose-containing β(1→4) linked backbone, can be divided into four major categories: linear mannan, galactomannan, glucomannan and galactoglucomannan (GGM) (120). The first two of these has a backbone consisting solely of mannan, while the latter two have a backbone consisting of glucose and mannose in a non-repeating pattern (104,120). Mannans are found in almost all plants, but are major components of softwood plants (120). Similar to xylan, mannan can also be found coating cellulose in some cases (121). In a lignocellulosic biorefinery context, the most relevant form of mannan is GGM, as it is the most abundant form of mannan, and the most abundant hemicellulose in softwood (122). Overall, degradation of all mannan forms can be considered important for more advanced biorefineries (120,123).

3.2.2.3 Xyloglucan

As mentioned above, xyloglucan is a hemicellulose which is thought to function as a glue to hold cellulose fibrils together (88). It is found in every terrestrial plant, as well as many algal species, and in many species is the most abundant hemicellulose in the primary cell wall (104). Xyloglucan , like cellulose, of a backbone of D-glucose units connected through β(1→4)-glycosidic bonds (124). This backbone is regularly substituted with α-D-xylosyl residues through an α(1→6) linkage (104,124,125). The canonical xyloglucan structure features a repeating motif of three substituted glucose units followed by one unsubstituted (124,125). However, this exact pattern is not observed in all plant species, and xyloglucan can be as little as 30% substituted with xylose (124). As well, additional sugars can be attached to these xylose substitutions, making potential xyloglucan structures extremely complex (124,125).

3.2.2.4 Mixed-Linkage Glucan

Glucans are polysaccharides in which the backbone is comprised primarily of glucose (126,127). Following this definition, it becomes clear that cellulose can be considered a glucan. Focusing solely on hemicellulosic glucans, the glucans of highest interest within plant biomass have been shown to be β(1→3, 1→4)-glucan (found primarily in grasses) (104).

3.2.3 Pectin

Pectin is the final major polysaccharide component in lignocellulosic biomass (128). It is different from hemicellulose and cellulose in that it is relatively easily extracted utilizing acid treatment or chelators, and has a high galacturonic acid content, adding a significant negative charge (128). It can be up to 35% of the primary cell wall of a plant, with the highest amounts of pectin being found in dicots (129). Important pectic polysaccharides include apigalacturonan, homogalacturonan, rhamnogalacturonan I and II, and xylogalacturonan (128,129). The most abundant of these, homogalacturonan, is typically responsible for the strengthening of cell walls (128). Pectin has far more functions than simply strengthening cell walls – it has been noted to be involved in plant cell morphogenesis, defense, signaling, cell-cell adhesion, seed hydration, and fruit development, as well as other roles (129).

3.3 Lignin

The fourth major component of lignocellulosic biomass is lignin. Unlike the others, it is not a polysaccharide, and not a carbohydrate at all (130). After cellulose, it is often reported as the most abundant natural substance in nature (131). From a biorefinery perspective, however, lignin is a little used component (131,132). Most biorefinery concepts consider lignin as nothing more than a source of energy (nearly 98% of lignin is burned for energy), and produce more lignin than they use (132). Lignin can also be a hinderance to the biorefinery, as it can inhibit enzymatic degradation of other components. The major reason for this is that lignin is incredibly complex. At least thirty five different major lignin monomers from eleven different classes of metabolites are used as the basic building blocks of lignin (although there are many others that do not occur as frequently), and its exact composition varies greatly between different species (133). The majority of these building blocks are aromatic compounds, and are linked together either through cross-linking reactions or polymer-polymer coupling (131). The most frequent bonds within lignin are carbon-carbon and carbon-oxygen bonds, although others are known to exist (131).

Lignin is generally formed by the plant cell after the polysaccharide network is established (134). Monolignols are synthesized in the cytoplasm of the producing cell, and then exported to the cell wall (135). These are oxidized, which leads to a polymerization cascade, driven by oxidative coupling rather than an enzymatic process (135). The oxidation of monolignols is, however, catalyzed by a laccase enzyme (although peroxidases can also be involved) (135,136). The binding of the growing lignin network to the polysaccharides of the cell wall is a complex process that involves aggregation of lignin and hemicelluloses, dehydrating the local environment, and allowing nucleophilic reactions to take place between the hemicellulose and lignin (137).

Despite its complex structure, if processed correctly, lignin can be an excellent source of value-added chemicals in a lignocellulosic biorefinery (131,132). Lignin can be a source of chemicals that are currently produced using fossil fuels – production of these would both increase the profitability of a lignocellulosic biorefinery, as well as decrease overall dependence on fossil fuels (138). Some of the major compounds that have been successfully produced from lignin include phenol, vanillin, and ethylbenzene (building block for polystyrene) (138,139). Conversion of lignin to biofuels has also been demonstrated, including conversion into energy-dense jet fuel (140). As well, bioplastic precursors have been successfully derived from lignin fermentation (141-143). Despite its complexity, lignin remains a promising source of many important and valuable chemicals as biorefinery output.

3.4 Lignin-Carbohydrate Complex

Lignin is often covalently linked to the cellulose or hemicellulose fractions of lignocellulose in what is known as a lignin-carbohydrate complex (LCC) (144,145). The majority of lignin in hardwood plants, and all lignin in softwood plants, is suggested to exist in these complexes (144,146,147). These LCCs add stability to plant biomass, as well as increase its recalcitrance, and the covalent bonds between the lignin and polysaccharides impede its removal from the biomass (144,147). When the cellulose or hemicellulose portion of the LCC is acted on by the appropriate enzymes, the enzyme efficiency is greatly decreased when compared to their actions on pure cellulose or hemicellulose (148). Additionally, pieces of the polysaccharide will remain attached to the lignin even after treatment of LCCs with polysaccharide-degrading enzymes, increasing the difficulty of utilizing the lignin, and decreasing the yield of sugars from polysaccharide degradation (119,149).

LCCs have several major types of bonds linking the lignin component to the carbohydrates: ester-, ether-, and glycosidic bonds (Figure 3.2) (144). These connections exist in various amounts (and the amounts differ between different species), and can thus link the lignin to various carbohydrates in the material. Of these, benzyl ether, benzyl ester, and phenyl glycosidic linkages have been observed to be the most common (144). The exact nature of LCC bonds is however difficult to determine, and requires the use of advanced techniques, such as 2- or 3D NMR, along with highly optimized extraction protocols (150).

3.5 Extractives, Ash, and Other Cell Wall Components

The remainder of the dry biomass weight consists of extractives and ash, which can make up over 20% of the plant biomass, with the amount greatly varying between different parts of the plant (151-153). This includes both organic (extractives) and inorganic (ash) components (153). The extractives consist of a variety of secondary metabolites including alkaloids, aromatic compounds, fatty acids, phenols, protein,

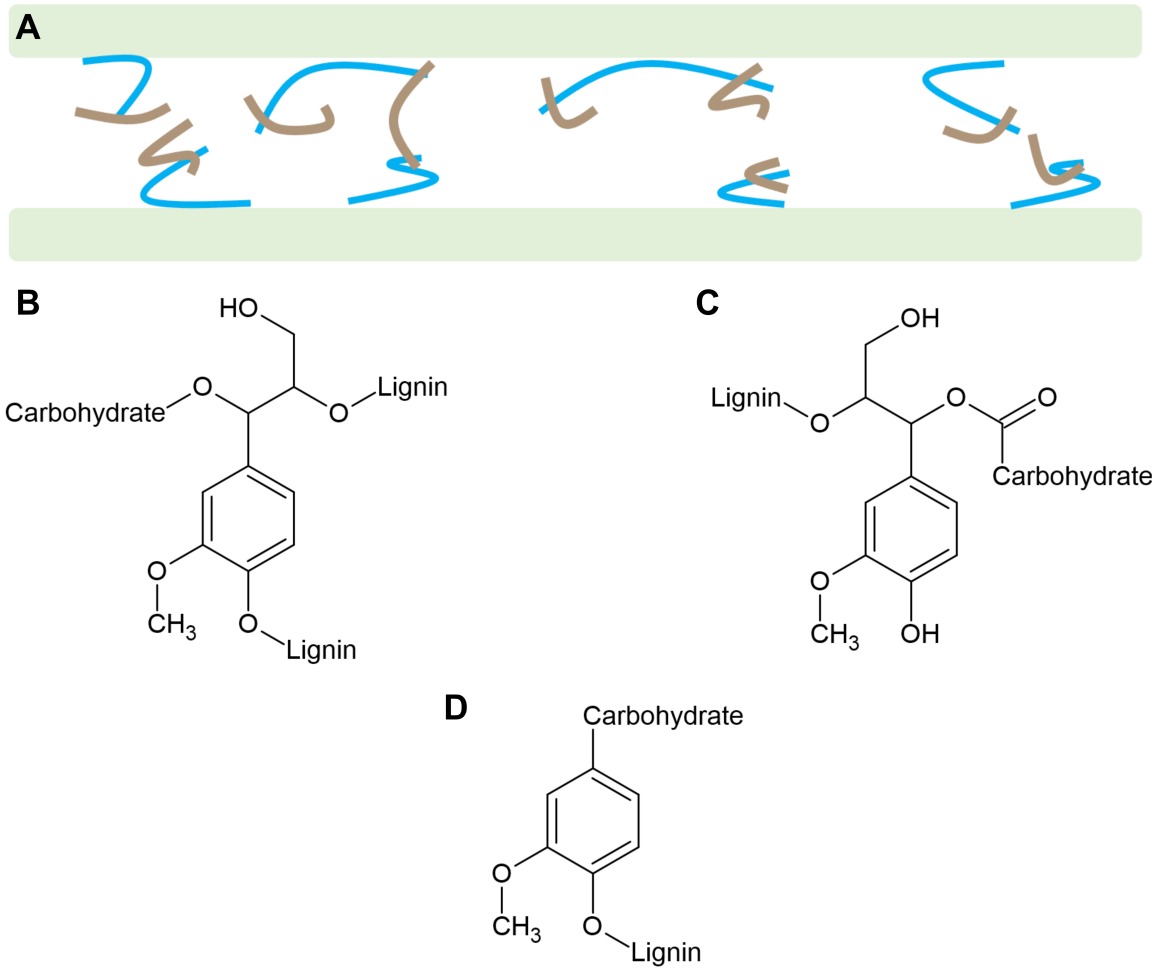

Figure 3.2: Schematic representation of the LCC (A) (150). Green rectangles represent cellulose microfibrils, blue lines represent hemicellulose, and the brown lines represent lignin. Examples of an LCC ether bond (B), ester bond (C) and glycosidic bond (D) (144).

terpenes and wax (154,155). These exist in the plant mainly for protection, and are typically in higher concentrations in the bark than in other parts of the plant (156,157). The ash in biomass consists of three main components: soil and sand contamination (from handling of the biomass), inherent vascular ash, and structural ash (158). While the first type can simply be washed away, the second and third type consist of minerals incorporated into cell walls, and are not as easily removed (158,159). Although it is possible these minerals may be useful in other contexts and can be extracted and sold, no such process currently operates at a large scale (153).

Ultimately, removal of these compounds is preferred prior to the use of the lignocellulosic biomass as a feedstock, as they may act as inhibitors towards microbial fermentation and hydrolysis (153). Ash can often cause physical problems for equipment used in biorefineries (160,161). While ash removal is well studied, significantly less research has been conducted into the impacts of extractive removal (162). Recent work has however shown that several extractive compounds negatively impact product yields, and regardless, as the organic extractives can be valuable end products in their own right, removal is considered preferable (163-165).

While this section has covered the most important cell wall polysaccharides, a vast number of others exist in varying quantities in the cell walls of different plant species. One example is callose, consisting of β(1→3)-linked glucose residues, and which is produced as a stress response and can be present in cell walls of various plant tissues (166). The polysaccharides carrageenan and alginate, special classes of galactans, are present in high amounts in different algal species (167). Plants produce many other polysaccharides, such as starch and gums, though those are generally not incorporated into the cell wall (167,168).

Chapter 3: Summary

- Plant biomass is an extremely complex substance
- Polysaccharides can be linear or branched
- Major polysaccharides in plant cell walls include cellulose, xylan, mannan, and pectin
- Lignin is not a carbohydrate, but contains a variety of potentially useful structures
- Lignin can be linked to hemicelluloses via LCCs

Chapter 4: Carbohydrate-Active Enzymes

With the complexities of the plant cell wall discussed in chapter three, the system of enzymes required to fully degrade all plant cell polymers must also be complex. This task falls to a group of enzymes known as carbohydrate-active enzymes (CAZymes). CAZymes, in simplest terms, are enzymes which facilitate the assembly or degradation of oligosaccharides or polysaccharides (169). They are extremely abundant in nature, with the largest database of such enzymes currently housing several million enzyme sequences (170). CAZymes are classified into five distinct classes: Glycoside Hydrolases (GH), Carbohydrate Esterases (CE), Polysaccharide Lyases (PL), GlycosylTransferases (GT) , as well as the more recently added Auxiliary Activities (AA) class (169,170). The AA class includes enzymes targeted towards lignin degradation, as well as the recently discovered lytic polysaccharide monooxygenases (LPMO) (171). While many of these AA enzymes do not act directly on polysaccharides (LPMOs being the exception), they assist other CAZymes by degrading and removing lignin from the polysaccharides being targeted (171). The work in this thesis is focused largely on enzymes from the GH and CE classes, and the most relevant families from these classes will be discussed in greater detail in this chapter.

CAZymes are, in general, an extremely useful type of enzymes with numerous applications in industry (169). Apart from biorefineries, CAZymes are used in animal feed in order to increase its nutritional availability, in the food industry to help with sugar extraction (and other purposes), in the pharmaceutical industry, the textile industry, the paper industry, the brewing industry, and in waste management, to name only a few applications (172,173).

4.1 Glycoside Hydrolases

Out of all the classes of CAZymes currently documented, GHs are by far the most abundant, making up almost half of the CAZy (Carbohydrate-Active enZyme) database (170). This large group contains enzymes with a huge variety of different functions, however, all of them are defined by their ability to catalyze glycosidic bond hydrolysis (174). Apart from the crucial role they play in lignocellulose degradation, they are involved in a diverse range of functions; for example, lysozyme and neuraminidase enzymes are glycoside hydrolases (10).

At the moment, glycoside hydrolases are sub-categorized by sequence identity in CAZy into 171 different families, although many GH sequences have been identified that do not belong to an existing family, and new families are being

discovered every year (170). Because the system of classification is based on sequence rather than function (activities), many families have multiple identified functions, and many functions exist in multiple families (170). GH enzymes are essential for every organism that degrades lignocellulose as a carbon source – without them, degradation of lignocellulose is not possible.

CAZymes (and the majority of known enzymes) are characterized using the Michaelis-Menten equation and the constants which derive from it (175). However, it can be argued that this is not an entirely accurate way to characterize these enzymes, even if it is the best way currently available. In a traditional Michaelis-Menten enzyme reaction, the substrate is assumed to be uniform and is consumed when the product is produced by the enzyme (175). This assumption does not hold true with polysaccharides, as they are inevitably of wildly varying lengths, and are not instantly consumed as enzymes act on them, but rather shortened sequentially. While differences in polysaccharide length may not impact the enzymes acting on them, it is cannot be said for certain, as isolating polysaccharides of a specific length is incredibly difficult. Purchased polysaccharides can also vary greatly between suppliers and even between batches, in terms of mixture of polysaccharide lengths as well as in terms of ash and other contaminants. Finally, as mentioned, polysaccharides can contain a large number of backbone substitutions, which can have a large impact on enzyme binding and activity (176). All of these factors combined strongly suggest that Michaelis-Menten kinetic constants are more of a "best guess" approach for investigating the activity of these enzymes, rather than numbers derived from effective measurement of enzyme kinetic rate.

4.1.1 Xylanases

As discussed above, xylan is the most abundant and most industrially relevant polysaccharide found in lignocellulose, after cellulose itself. Enzymes which degrade xylan are fittingly known as xylanases, and can be found in GH families 5, 7, 8, 10, 11, 12, 30, 43, 98, and 141 (170,177). Xylanases, categorized by function, can be described as either endo- or exo- acting (note that the endo-/exo- categorization can be applied to many types of GH enzymes – cellulases, chitinases, mannanases, etc.) (177). Endo-acting xylanases are enzymes which bind along a xylan chain and cleave the glycosidic bonds, resulting in long-chain xylooligomers (178). Exo-acting xylanases, on the other hand, cleave the xylan chain from either the reducing end or non-reducing end, often releasing xylobiose (although other short xylooligomers are possible) (178,179). Interestingly, although the first xylanase was first reported in 1955, the first exo-xylanase did not appear in the scientific literature until 1989 (109,178). Xylanases can also be processive, meaning that they can perform several successive cuts on the same chain before letting go (180).

Xylanases are highly relevant enzymes in an industrial context. Currently, they are commercially produced for bakeries, the paper industry, as feed additives, and more

(181). Within the paper industry, xylanases are used for bleaching of wood pulp, as well as de-inking recycled paper, resulting in significantly lowered usage of harsh chemicals that would be traditionally used for these processes (182,183). In the animal feed industry, xylanases are applied to certain cereal plants before they are fed to animals, increasing digestibility and energy availability from the crops (184). In bakeries, xylanases are used to produce better quality and more consistent doughs (185). Xylobiose produced by xylanases has shown promise as a prebiotic, and is therefore attractive to the pharmaceutical industry (186,187). Finally, the use of xylanases in lignocellulosic biorefineries and for biofuel production can increase yields dramatically, as lignocellulosic biomass can be more than a third xylan by dry weight (181,188).

4.1.1.1 Glycoside Hydrolase Family 8 Enzymes

Glycoside hydrolase family 8 (GH8) is an exclusively bacterial family which contains enzymes of a variety of different functions, including chitosanase, cellulase, licheninase, endo-β(1→4)-xylanase, and reducing-end-xylose releasing exo-oligoxylanase (Rex) activities (170). Of these, no activity is exclusive to GH8 enzymes, however, very few Rex enzymes have been identified within other families (170,189). Interestingly, aside from polysaccharide degradation, GH8 enzymes are also involved in bacterial cellulose synthesis, although they have not been implicated in the same processes in plants (190). Compared to other xylanase families, GH8 xylanases remain relatively little-studied (191). These enzymes are exclusively endo-xylanases, and show a variety of different product profiles, although all produce xylooligosaccharides of between two and four xylose units (170,192). The majority of known GH8 xylanases are single domain enzymes, with few exceptions (192,193). This is in contrast to the broader group of xylanase enzymes, which are often multidomain proteins (191). **Paper III** is partially focused on the study of a Rex domain from the multicatalytic *B. eggerthii* enzyme *Be*CE15A-Rex8A.

As mentioned, a unique activity among some GH8 enzymes is the ability to release xylose and xylobiose from the reducing end of relatively short xylooligosaccharides (170,194,195). The term “Rex” can be somewhat confusing in the scientific literature surrounding CAZymes, as it is used both for reducing-end-xylose releasing exo-oligoxylanase enzymes, the activity found in the GH8 family, and for reducing-end xylose-releasing exoxylanase enzymes (and the terms are sometimes incorrectly used interchangeably) (178,189,194-196). Although there is overlap between the two, it is more correct to use the “exo-oligoxylanase” term for GH8 Rex enzymes, as they show low or no activity on polymeric xylan (194). To date, the only known enzyme which could truly be referred to as a reducing-end xylose-releasing exoxylanase is the GH30 XYN IV from *Trichoderma reesei* (197). The exact xylooligosaccharide size requirements which steer Rex function is unknown (though it is likely to be different on a case-by-case basis), as xylooligosaccharides of greater size than xylohexose are prohibitively expensive, and not routinely used for testing

(194,198,199). The major feature thought to be responsible for limiting the size of the substrate is a small Leu-His-Pro loop present in most Rex enzymes, although in two of the currently characterized enzymes, this is substituted with an Arg-His-Ser loop instead (200). This loop does not appear to be entirely responsible for substrate length determination however; some enzymes in each loop category are active on xylan, and some are not (194).

To date, structures of 17 different GH8 enzymes have been deposited in the Protein Data Bank (PDB) (170,201). The structures share a common $(\alpha/\alpha)_6$ fold (Figure 4.1) (195). GH8 structures typically show a binding cleft, in which the substrate rests and can be acted on by catalytic residues (195,202). The exact position and nature of the catalytic residues can vary between three different sub-classes of GH8 (GH8a, GH8b, and GH8c) (203). While the catalytic acid is conserved throughout the family (a glutamate residue around 100 amino acids in to the sequence), the catalytic base is different in the different subfamilies (203). The catalytic base is an aspartate in GH8a, a glutamate in GH8b, and currently unknown in GH8c (195,203).

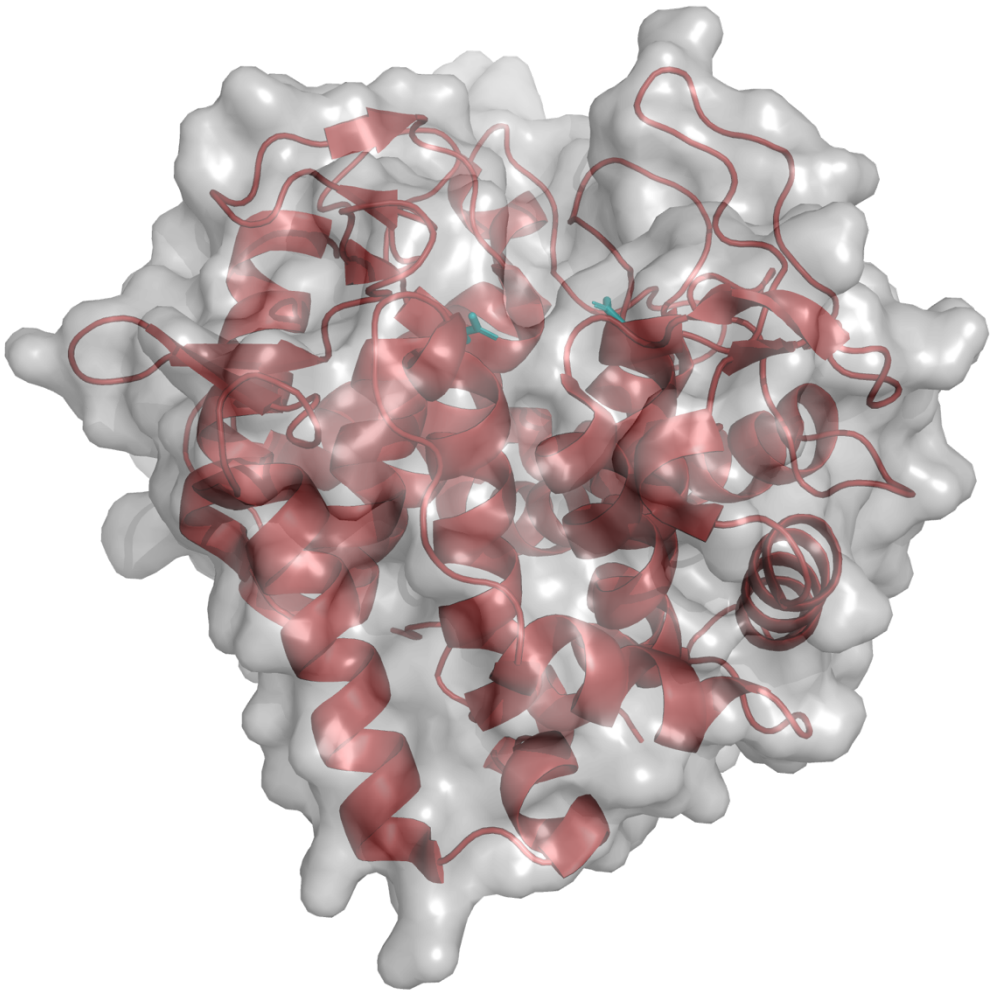

Figure 4.1: Three-dimensional structure of a representative GH8 enzyme (PDB ID: 6SHY), Rex8A from *Paenibacillus barcinonensis* (195). Catalytic residues are highlighted in cyan. The $(\alpha/\alpha)_6$ fold can be seen to be the main component of the structure.

4.1.1.2 Glycoside Hydrolase Family 10 Enzymes

A large percentage of enzymes within the glycoside hydrolase family 10 (GH10) group of enzymes are bacterial in origin, although eukaryotic and archaeal enzymes in this family do exist (170). Activities within this family are somewhat limited, with the vast majority of enzymes having endo-xylanase activity, although other activities have been documented (170). An important feature of GH10 xylanases is that they are seemingly only slightly impacted by substitutions on the xylan backbone, allowing them to effectively degrade a variety of different xylans (186). This flexibility to work around backbone substitutions comes with the limitation that GH10 enzymes are only slightly able to hydrolyze insoluble xylans (186). Along with xylanases from the glycoside hydrolase 11 family (GH11), GH10 xylanases are the most likely family of xylanase to be thermostable (204). **Paper I** and **Paper II** focus on a the multicatalytic enzymes *Ck*Xyn10C-GE15A from *C. kristjanssonii*, which contains a thermostable GH10 domain. The same domain is used as a in **Paper IV** for multicatalytic enzyme construction.

Many more GH10 structures are available as compared to GH8 – currently 55 different structures of GH10 enzymes are available in the PDB (170,201). All structurally solved members display a $(\beta/\alpha)_8$ fold, the most common protein fold among GH enzymes (Figure 4.2 A) (205). As these are endo-acting enzymes, they have a binding cleft going along the side of the enzyme which is generally capable of binding up to seven xylose residues (206). Two catalytic residues are present within the general GH10 sequence, both of them being glutamates (207).The first glutamate (present around residue 140) acts as a nucleophile, and the second

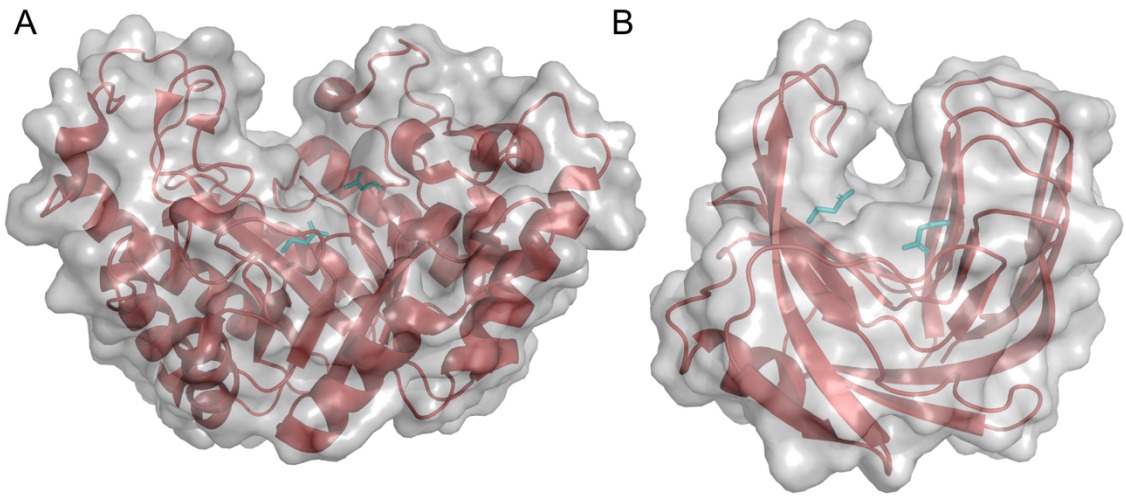

Figure 4.2: Three-dimensional structure of a representative GH10 enzyme (PDB ID:1W2P), Xyn10A from *Cellvibrio japonicus* (**A**) and three-dimensional structure of a representative GH11 enzyme (PDB ID: 1BCX) from *Bacillus circulans* (**B**) (208,209). The $(\beta/\alpha)_8$ fold can be seen to be the main component of the structure in **A**, and the jelly roll fold can be seen as the main component of the structure in **B**. In both cases, the catalytic residues are highlighted in cyan.

(around residue 250) acts as a catalytic base (207).

4.1.1.3 Glycoside Hydrolase Family 11 Enzymes

Like GH10 enzymes, the vast majority of glycoside hydrolase family 11 (GH11) enzymes are also bacterial in origin (170). The family only has two reported activities, endo-β(1→4)-xylanase and exo-β(1→4)-xylosidase, which makes it a highly specialized family compared to many others within CAZy (170). The GH11 family has been relatively well characterized structurally, with 36 different structures currently available in the PDB (170,201). Family members display a jelly roll fold (Figure 4.2 B) (209). In these enzymes, the catalytic nucleophile and the catalytic base are both generally glutamate residues, with the nucleophilic glutamate coming earlier in the protein sequence (209).

4.1.2 Cellulases

Cellulases are extremely important enzymes, both from a microbial and an industrial perspective. Cellulose is the most abundant organic compound on Earth, and its composition and relative simplicity make it an excellent carbon source (210). Although cellulose degradation is a relatively rare strategy, it can be found throughout various types of microbial life; it is found in bacteria, fungi, and archaea, and in both aerobic and anaerobic microorganisms (211). Within CAZy, cellulases are found in a large number of GH families; families 1, 3, 5, 6, 7, 8, 9, 12, 26, 44, 45, 48, 51, 74, 124, and 131 (170,212).

Similar to xylanases, cellulases can also be categorized as either endo- or exo- acting (213,214). These behave in the same way as their xylanase counterparts, with endo-enzymes cleaving cellulose chains randomly, and exo-cellulases cleaving chains from either the reducing or non-reducing end, and some enzymes are also processive (212). Like xylanases, cellulases are extensively used in industrial applications (much more so, in fact) (172). For an in-depth discussion of these applications, an excellent review by Kuhad *et al.* discusses applications in great detail, the contents of which will be briefly summarized here (172). Cellulases are, among many other applications, heavily used in a biorefinery context for the degradation of feedstocks to allow for the production of valuable bioproducts (215).

4.1.2.1 Glycoside Hydrolase Family 9

Glycoside hydrolase family 9 (GH9) almost exclusively consists of cellulases, with less than 1% of discovered enzymes in the family having primarily non-cellulase activity (170,216). Many of these enzymes do, however, display side activity on other polysaccharides, such as xylan. These enzymes are found in all kingdoms of life, and are a common feature among microorganisms that break down cellulose (and are also common in plants), although they are typically not found in aerobic fungi (216). Surprisingly, genes encoding GH9 cellulases have also been discovered in termites

and several other animals (217-220). This is in contrast to the traditional view that animals who consume cellulose rely solely on their gut microbiota for polysaccharide deconstruction (218).

As the second-largest cellulase family, GH9 contains most plant and animal cellulases, as well as many bacterial cellulases, and its enzymes have found significant usage in industrial applications (221,222). The family contains most well-characterized processive cellulases, although both processive and non-processive activities are present within the family (222,223). Within bacterial systems, GH9 enzymes are among the most important and abundant cellulases (although the same is not true of fungal systems) (223). GH9 enzymes to date have twenty protein structures deposited in the PDB (170,201). Like the GH8 enzymes, GH9 enzymes display an $(\alpha/\alpha)_6$ fold (Figure 4.3 A) (224), with binding clefts containing a minimum of six sugar binding sites (224). GH9 enzymes have three important catalytic residues – a glutamate around residue 400 as the catalytic acid, and two aspartates near residue 55 that act as catalytic bases (225).

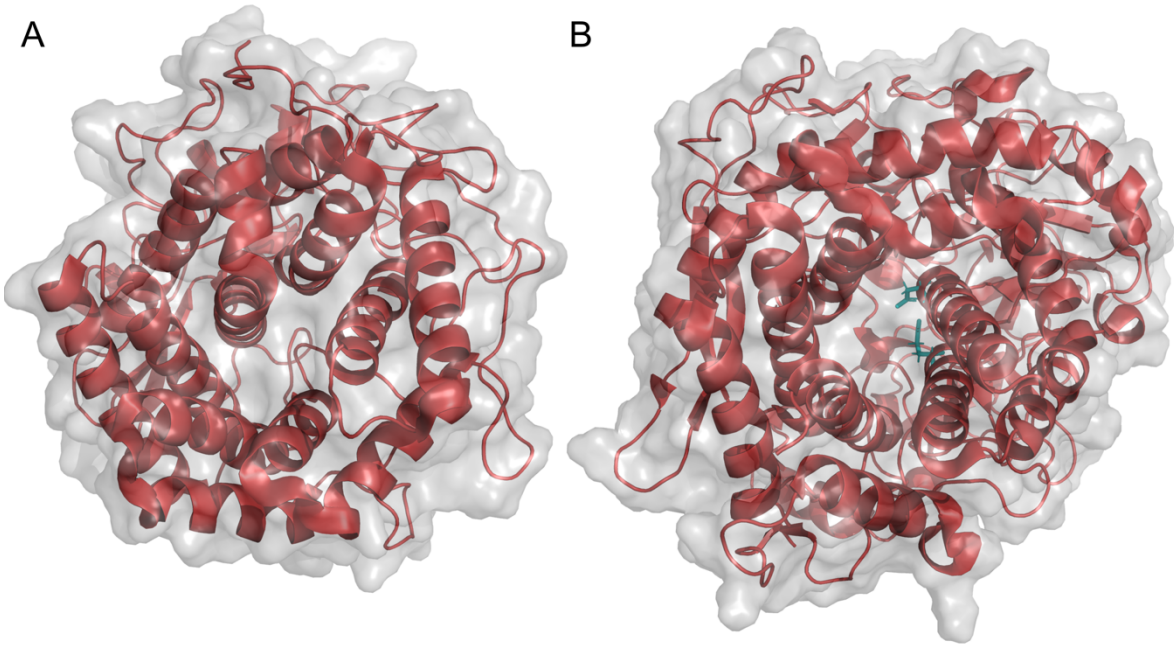

Figure 4.3: Three-dimensional structure of a representative GH9 enzyme (PDB ID: 4DOD), the CelA GH9 from *Caldicellulosiruptor bescii* (**A**) and a three-dimensional structure of a representative GH48 enzyme (PDB ID:5YJ6), CelS from *Clostridium thermocellum* (**B**) (226,227). The $(\alpha/\alpha)_6$ fold can be seen to be the main component of the structure in both structures. Catalytic residues are not shown for the GH9, as no information on experimental determination could be found in literature. Catalytic residues for the GH48 are shown in cyan (228).

4.1.1.2 Glycoside Hydrolase Family 48

Glycoside hydrolase family 48 (GH48) enzymes are almost exclusively cellulases, although glucanase and chitinase activity have been detected within the family (170).

They are considered to be the most important component of bacterial cellulose degradation systems, even more so than GH9 enzymes (223). Almost all GH48 enzymes are bacterial in origin, with only 40 of over 1000 sequences in CAZy being of non-bacterial origin (170,229). They can also be found as free enzymes, multicatalytic enzymes, and as part of a cellulosome, enzyme organizational systems that will be discussed in detail in the next chapter (226,230-232). The rare non-cellulolytic GH48 enzymes are typically non-bacterial in nature; for example, GH48 chitinases can be found within certain species of beetle (233). In **Paper IV**, a GH48 domain from the *C. bescii* CelA protein is used as an example domain for multicatalytic enzyme construction.

Unlike many other cellulases, GH48 enzymes are typically only found in a single copy in a genome (234). They are commonly found to work in conjunction with GH9 cellulases, and often show significant synergy when combined in a reaction (177). In fact, in what is perhaps the most efficient cellulase known to exist, CelA from *C. bescii*, both a GH9 and a GH48 catalytic domain are present within the same polypeptide in a multicatalytic configuration (226). To date, only ten structures of GH48 enzymes are available in the PDB (170,201). Like GH8 and GH9 enzymes, all GH48 enzymes have an $(\alpha/\alpha)_6$ fold (Figure 4.3 B) (229). Rather than a binding groove, GH48 enzymes have a binding tunnel, through which polysaccharide is fed and cleaved (229). The catalytic residues have been identified as a glutamate present around residue 55 (catalytic base), and an aspartate present around residue 225 (catalytic acid) (228).

4.2 Carbohydrate Esterases

While GHs represent the largest and most studied CAZyme class, also other classes are extremely important for the degradation of lignocellulose. Carbohydrate esterases are enzymes which remove *-N* or *-O* ester-bonded side chains and other modifications from mono-, oligo-, and polysaccharides and enable other CAZymes to reach their substrates (235). Currently, there are almost 100 000 putative CE sequences in CAZy, spread across 19 families (although family 10 has been removed, leaving 18 active families) (170). As with GH enzymes, many CE families have overlapping activities due to the system of classification based on sequence rather than function (170). These enzymes can act on a variety of different polysaccharide substrates, including chitin, chlorogenic acids, hemicelluloses, and pectin (235).

4.2.1 Carbohydrate Esterase Family 15

All characterized enzymes within carbohydrate esterase family 15 (CE15) characterized thus far have been glucuronoyl esterases, which cleave ester bonds between lignin and glucuronoxylan (236). These enzymes exist in both bacteria and fungi, although bacterial and fungal enzyme variants are structurally significantly different from each other (237). The removal of lignin from xylan is thought to be

very important for further efficient xylan degradation (238). Unfortunately, due to the difficulty in producing structurally consistent LCCs, no natural substrates for these enzymes are accessible to researchers, and they are typically characterized with model substrates, on which they may be significantly less efficient (238). **Paper I** and **Paper II** partially focus on the characterization of a CE15 domain from the *C. kristjanssonii* multicatalytic *Ck*Xyn10C-GE15A enzyme, and **Paper III** includes the characterization of a CE15 domain from the *B. eggerthii Be*CE15A-Rex8A enzyme. **Paper IV** uses the *Ck*Xyn10C-GE15A CE15 domain for multicatalytic enzyme construction.

Currently, there are only eight structures of CE15 enzymes in the PDB, although the number of deposited structures has increased greatly in recent years (170,201). These enzymes have an overall α/β-hydrolase fold, although with some extra features (additional N-terminal β-strands and α- and 3_{10}-helicies sandwiching the central fold) (Figure 4.4) (237). Bacterial CE15s generally differ from their fungal counterparts by the inclusion of three extra inserted regions, although they are not always present (**Paper II**)(51,239). These enzymes typically have a catalytic triad of a serine, a glutamate or an aspartate, and a histidine, typical of esterases (239). In

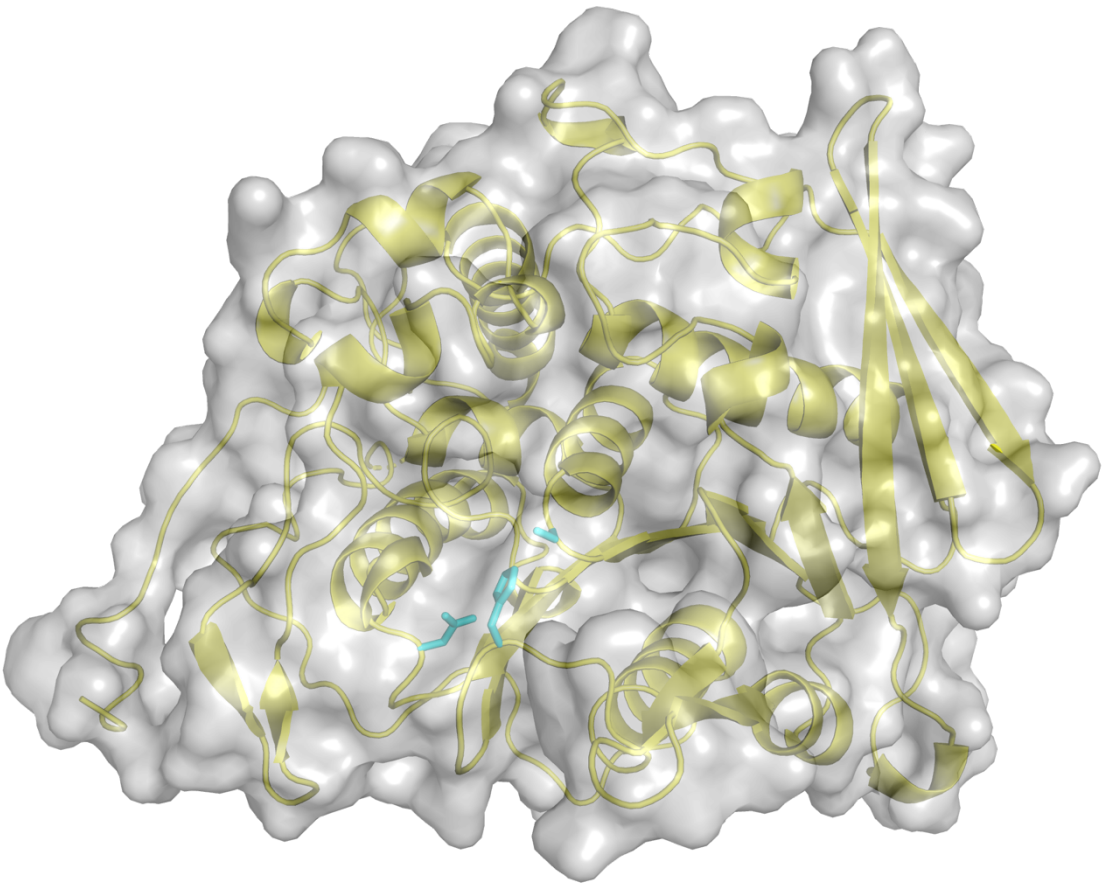

Figure 4.4: Three-dimensional structure of a representative CE15 enzyme (PDB ID: 7NN3), *Ck*GE15A from *Caldicellulosiruptor kristjanssonii* (**Paper II**). The α/β-hydrolase fold can be seen as the central component of the structure. Catalytic residues are highlighted in cyan.

addition to the catalytic residues, several other important amino acids are generally conserved within the family, the most important of which is a conserved arginine located directly after the catalytic serine which is suggested to form the oxyanion hole stabilizing the transition state intermediate (239). In a few cases, this arginine is substituted with an aromatic residue, and those enzymes display significantly less activity than their counterparts (**Paper III**)(200,239). Replacing the aromatic residue with an arginine does not restore activity, however, and can eliminate it altogether (200).

4.3 Carbohydrate Binding Modules

Carbohydrate Binding Modules (CBMs) are not enzymes like GHs and CEs, but are instead smaller non-catalytic protein domains which bind to various carbohydrates (240). Generally, they are found associated with a CAZyme (through being a part of the same polypeptide chain – multiple domains of the same protein), for which they assist in substrate recognition and binding (241). However, there are many examples where CBMs have been identified with no linked domains, seemingly existing on their own (240). CBMs are currently the only class of proteins categorized as "Associated Modules" in CAZy (170). As of this writing, there are 88 families (although CBM7 has been removed, and CBM33 has since been found to have catalytic activity and been reclassified as Auxiliary Activity Family 10) of CBMs in CAZy, with almost 275 000 proteins distributed across them (170). CBMs can recognize a broad range of carbohydrates – almost all known carbohydrates can be bound by CBMs from one or more families (170,241).

CBMs can be functionally classified into three main types, based on how they bind to carbohydrates (241). A type A CBM possesses a flat surface containing a high proportion of aromatic residues, which bind to the hydrophobic face of polysaccharides (242). Type A CBMs have not been observed to bind to mono- or oligosaccharides (241). Type B CBMs bind along a polysaccharide chain using a cavity in the CBM surface to accommodate the individual polysaccharide strands. A Type C CBM is one that binds to the end of polypeptide chains, generally through a binding site pocket (241).

There are five main functional roles that have been identified for CBMs: proximity effect (binding to the substrate and keeping the enzyme in close proximity), targeting function (affinity for a specific portion of the substrate, targeting the enzyme domain to that portion), disruption (loosening tightly-packed polysaccharides), adhesion (anchoring the attached enzyme domain to the surface of the producing cell), stabilization of the attached enzyme domain, and active site extension of the attached enzyme domain (243-247). Not every CBM will have all of

these functions; it is more common for them to only display one or two (243). In fact, it is difficult to envision how any CBM would perform all six roles.

CBMs are a structurally diverse set of proteins, and can be classified by structure into seven different groups (243). The groups are: β-sandwich fold, β-trefoil fold, cysteine knot, oligonucleotide/oligosaccharide binding fold, hevein fold, unique fold, and hevein-like unique fold. The β-sandwich fold is the most common CBM fold, which is present in more families than all other folds combined (243). Despite the common fold, there is no commonality in terms of metal binding (for structure or function), binding site location, or even number of binding sites, indicating that CBMs have evolved diverse ways to take advantage of this fold type (243,248).

4.3.1 Carbohydrate Binding Module Family 3

Carbohydrate binding module family 3 (CBM3) proteins are type A CBMs which generally bind crystalline cellulose (although chitin binding has also been reported, and recent findings have shown that some CBM3 proteins bind hemicellulose) (170,249,250). Currently in CAZy there are over 2 000 sequences identified, with 142 characterized, and 20 structures (these numbers may be overestimated, as CAZy will often count CBMs as characterized if they are attached to an enzymatic domain which has been characterized) (170). Four major subgroups of CBM3 have been identified so far, designated a-d, and grouped based on sequence similarity (250). Subgroup c is particularly interesting, as proteins in this group do not appear to bind cellulose on their own (251). Instead, they improve activity of their attached cellulase, and help it to act in a processive manner (249,251,252). **Paper IV** uses a CBM3 domain from the *C. bescii* CelA protein in multicatalytic enzyme construction.

Structurally, CBM3 domains share a β-sandwich fold (Figure 4.5 A) (170,243). Polysaccharide binding residues differ between the different CBM3 subgroups (253-256). For example, in subgroup b, a histidine, tryptophan, tyrosine, and arginine-aspartate ion pair form a planar hydrophobic surface (253-255). In other cases, the non-aromatic residues are replaced with more aromatic residues, making all of the important binding residues aromatic (256). In addition to the polysaccharide binding site, CBM3 domains often have a shallow groove with highly conserved amino acids, which has been proposed to be involved in binding proline-threonine-rich linker regions linking CBM3 modules to enzymatic domains (249,257).

4.3.2 Carbohydrate Binding Module Family 9

Carbohydrate binding module family 9 (CBM9) proteins are significantly less studied and less prevalent than CBM3 proteins, with just under 650 sequences in CAZy, 50 characterized and only two structures (170). They are reported to be exclusively coupled to xylanases, and bind cellulose (170). Both solved structures

display a β-sandwich fold and have three calcium-binding sites (although one is only partially occupied in *Ck*CBM9.3) (Figure 4.5 B) (**Paper II**)(51,170,258). CBM9 proteins are thought to function using a clamp mechanism, in which the polypeptide chain is held between two aromatic residues (51,258). However, not every putative CBM9 shows this binding clamp, indicating that there may be subfamilies of CBM9 modules that have yet to be properly classified (**Paper II**)(51). Due to the binding clamp having been observed to bind to polysaccharide chain ends, the CBM9 family is classified as type C CBMs, although more work is needed to determine if the CBM9 modules lacking the binding clamp residues still function as type C (51,258). Examples of CBM9 modules with a double tryptophan binding clamp, a tryptophan-tyrosine binding clamp, and a lack of traditional binding clamp are all explored in **Paper II**. These modules all show different specificities for various polysaccharides, although all do show at least some binding. Interestingly, the CBM9 domain that did

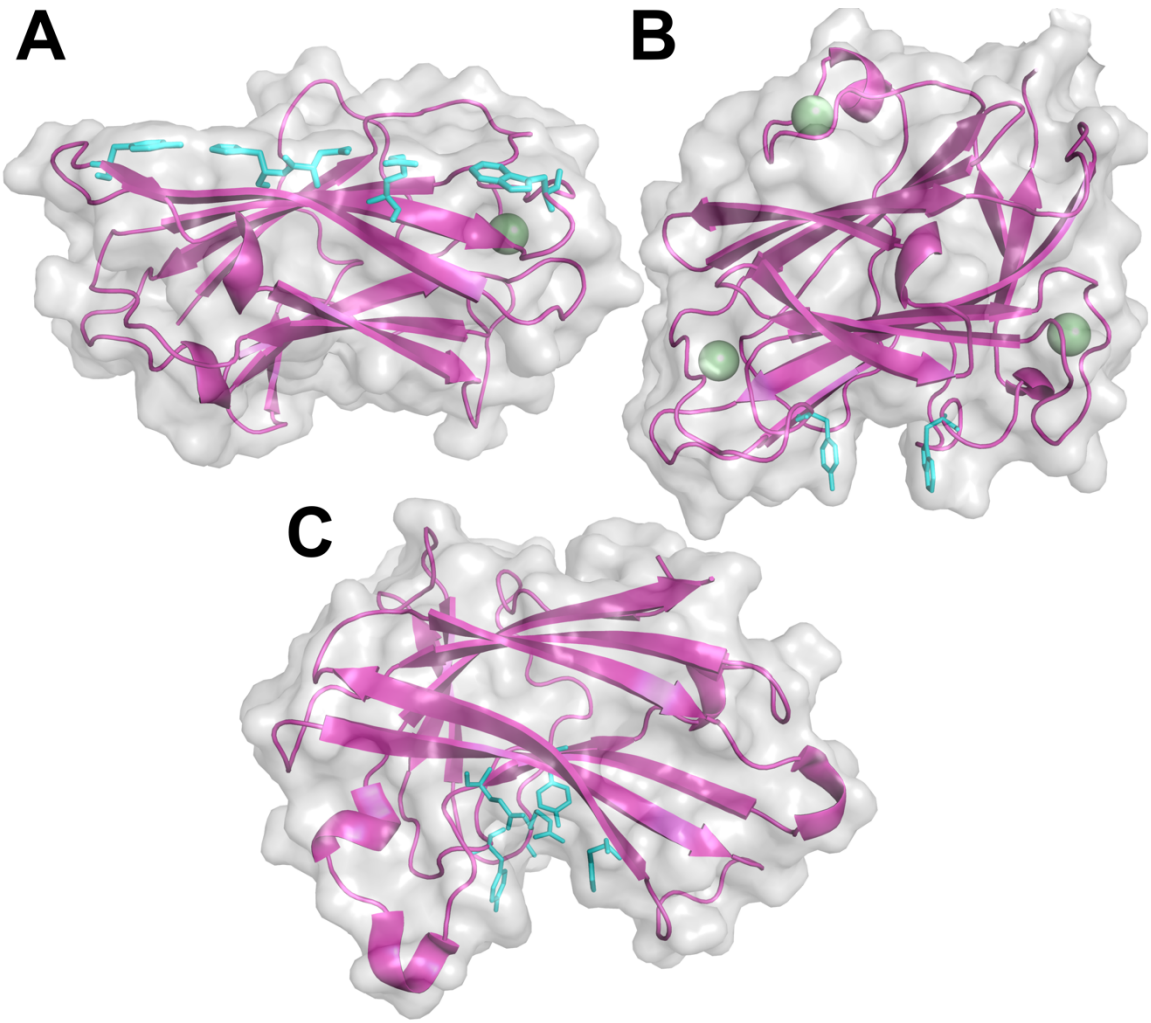

Figure 4.5: Three-dimensional structure of a representative CBM3 domain (PDB ID: 1NBC) from *Acetovibrio thermocellus* (**A**), a representative CBM9 domain (PDB ID: 7NWN) from *Caldicellulosiruptor kristjanssonii* (**B**), and a representative CBM22 domain (PDB ID: 4XUR) from *Paenibacillus barcinonesis* (**C**) (51,257,259). Calcium ions are shown in green. The typical β-sandwich fold can be seen as a central component of all three structures. Binding residues are shown in cyan.

not show the characteristic binding clamp appeared to severely limit the growth rate of *E. coli* during production, but a mechanistic explanation was not determined. This suggests there could be a larger overall role to the CBM9 family than simply assisting enzyme domains with binding to polysaccharides.

4.3.3 Carbohydrate Binding Module Family 22

Carbohydrate binding module family 22 (CBM22) proteins display, like the two previously mentioned modules, a β-sandwich fold (Figure 4.5 C) (170). CAZy currently lists just over 1500 known sequences, 111 of them characterized, and four with known structures (170). CBM22 domains are type B CBMs, showing several conserved aromatic residues which function as a clamp for the target polysaccharide chains (259). These aromatic residues, along with conserved nearby polar residues, have been shown to be essential for polysaccharide binding within CBM22 modules (260).

In addition to the carbohydrate-binding role of CBM22 modules, they have also shown in some cases to confer thermostability to their attached enzyme (usually a GH10 enzyme) (261-264). This thermostabilizing effect was seen for the CBM22 modules studied in **Paper I**. CBM22 modules often appear in duplicate or triplicate within a polypeptide, and in some cases displaying similar properties for ligand binding, and in other cases not (259,261,265-267). These modules have been shown to increase xylanase activity in some cases, but in other cases have been seen to decrease activity (261,266,268). For the CBM22 modules studied in **Paper I**, significant differences were observed in several properties, most notably in solubility – the first CBM22 module would not remain soluble on its own, and required at a minimum to be expressed in a polypeptide containing the second CBM22 module as well. There were also observed differences in binding between the protein containing both CBM22 modules and the second CBM22 module on its own, suggesting that the two modules have different binding capabilities. These CBM22 modules were also seen to decrease the activity of the attached GH10 domain, however, as they increased its thermostability, it became capable of functioning at the optimal growth and environmental temperature of the encoding organism, indicating an exchange of rate of activity for function. The CBM22 modules studied in **Paper I** were also used in multicatalytic enzyme construction in **Paper IV**.

4.4 Methods of Enzyme Study

The study of enzymes can take several forms depending on what information the researcher is looking for, however, it is common to explore several different properties in order to gain a more complete understanding of the studied enzyme. Enzymes can be characterized to determine the speed and efficiency of their

reactions, their properties, and their structure. The methods through which to determine these features are discussed in the sections below.

4.4.1 Enzyme Activity Measurements

Enzyme activity measurements are vital to the understanding of a specific enzyme. The activity of an enzyme can be measured in different ways, and the "correct" way often depends on the enzyme itself, the available substrates, and the available methods to measure product formation. Measurement of enzyme activity can be performed in either a continuous or discontinuous manner (269). In a continuous assay, the progress of the reaction can be monitored in real time, for example, by using spectroscopic or fluorescence methods to follow product formation. In contrast, a discontinuous assay is one in which product formation cannot be followed in real time, often due to a lack of absorbance/fluorescence spectra change between the substrate and product. In this case, an assay is stopped after a period of time and analyzed in a way that allows for measurement of the product (269).

When measuring enzyme activity, measurements can be performed either directly or indirectly (269). When measuring activity directly, a researcher measures either the increase in concentration of the product, or the decrease in concentration of the substrate. When measuring indirectly, neither the product nor substrate can be detected by available methods. However, it is often possible to couple a second enzyme reaction to the first. This reaction must be several orders of magnitude more rapid than the one being measured, and the product of that reaction must be detectable in some manner. In this way, it can be assumed that when a substrate is converted to product by the enzyme being studied, the product is immediately used as a substrate by the coupled enzyme. Thus, measuring the product formation of the coupled enzyme will give information as to the rate of product formation of the studied enzyme (269). One way of measuring product formation specific to CAZymes is to measure the presence of reducing ends of oligo- and polysaccharides – an increase in reducing ends indicates that the enzyme has made cuts within the chain (270). This can be done using the 3,5-Dinitrosalicylic acid assay, as was done in **Paper I** and **Paper III** (270).

Often, it is not possible when conducting enzyme activity assays to use the natural substrate, either because it is too difficult to synthesize or isolate, it is too unstable under the tested conditions, or there is no way to detect the change in substrate/product concentration. A model substrate should be able to reach a high enough concentration to saturate the enzyme in solution completely (269). Model substrates can also be designed with properties that make either them or their products easier to detect, to allow for better measurement of the rate of catalysis by the enzyme (271). Enzyme activities were tested in **Paper I** and **Paper III** using model substrates for the CE15 domains, and more natural substrates for the GH8 and GH10 domains, as well as for synergy studies between the domains.

4.4.2 Measure Enzyme Properties

As mentioned earlier, enzymes have many properties that are crucial in determining their suitability for industrial use. Many analysis techniques have been developed to investigate these features of enzymes, some of which will be summarized in this section.

4.4.2.1 Temperature Dependence

Optimal enzyme operating temperature is a highly important property in an industrial context. It is important to have an enzyme that works most effectively around the temperature of the reaction being run (or conversely, to set the temperature of the reaction to the optimal temperature of the enzyme being used) – if the temperature is too low, the enzyme effectiveness decreases, and if it is too high, the enzyme may become permanently damaged, rendering it useless (272). Enzyme optimum temperature can be estimated using the optimal growth temperature of its producing organism, as the optimal temperature of an enzyme often (but not always) correlates strongly with the growth temperature (273). Several computational methods have also been developed to attempt to calculate optimum temperature based on amino acid sequence (273,274).

While predictions provide a good starting point for determining enzyme temperature optimum, current methods do not take into account the impact of reaction conditions on the optimum temperature (275). Rigorous testing is required, measuring the enzyme activity at various temperatures and time lengths, in order to determine the ideal operating temperature (275). Enzyme temperature optima are determined in **Paper I**, although indirectly for one of the domains. For the CE15 domain in *Ck*Xyn10C-GE15A, there is no thermostable model substrate, so the temperature optimum had to be estimated from melting temperature of the enzyme. Such estimations do not always prove accurate, and its accuracy was unable to be confirmed. If the estimation is indeed accurate (or at least close), then the *Ck*GE15A domain is the most thermostable CE15 domain published to date.

4.4.2.2 pH

Enzyme activity is generally highly pH dependent, with enzymes often only effective over a narrow pH range (although there are exceptions) (276). The optimal pH for an enzyme can be difficult to predict, as it is not necessarily the same as the pH of its natural environment – in addition, there are also often microenvironments which have a different pH than the overall environment the enzyme is found in, further complicating matters (277). Ultimately, the pH dependence of the enzyme is determined by its overall fold and its amino acid side chains (277). Different amino acids have side chain pK_a values of between 4 and 12.5, and the frequency and position of these amino acids can have a large impact on the enzyme's pH stability

(278). These pK_a values are for the amino acids in solution by themselves – within a protein, nearby amino acids can impact the pK_a values (279). Therefore, while predictions of optimal pH based on the expected environment can be useful, ultimately, it is necessary to test each enzyme to find its optimal pH.

Enzyme pH optima can be tested for in a similar way to the temperature optima discussed earlier (280). Activity measurements can be carried out at different pH values and the highest activity determined from these measurements (280). It is important to note that the buffering agent can often have an impact on enzyme activity, so ideally one should test multiple buffering agents at each pH (281). **Paper I** and **Paper III** both involve testing the pH optima of different enzyme domains. **Paper I** especially showed a somewhat interesting pH profile for *Ck*Xyn10C, which appeared to have a dual pH optimum.

4.4.2.3 Inhibition

Enzymes are often subject to inhibition, either through the presence of small molecules that interfere with the enzymes ability to catalyze a reaction, or from a buildup of products from the enzyme-catalyzed reaction (282). Both of these methods of inhibition can occur for CAZymes, but the more common type of inhibition faced by these enzymes is product inhibition (282,283). Inhibition can be determined and measured by adding potential inhibitors to an enzyme-catalyzed reaction and analyzing whether the rate decreases (283). By varying inhibitor concentrations in these measurements, one can determine the effectiveness of the inhibitor in preventing enzyme activity (283).

4.4.3 Protein Structure Determination

Information that comes from protein structure can be incredibly useful in understanding enzyme functions, substrate binding sites, catalytic residues, and more (77). A protein structure can even be determined for a protein or enzyme interacting with its ligand or substrate (although gaining structural information without allowing catalysis can be challenging), giving information about how this interaction occurs, and what residues mediate it. Such information is often key in helping a researcher understand key amino acid residues that can be substituted to improve or direct function in a desired manner or how a protein compares to related proteins (77). The following sections will discuss several methods of obtaining protein structure, their advantages and drawbacks, and when each might be preferable to use. It is important to note that none of the techniques is “better” than the others, just that one may be more relevant to obtain the information a researcher is attempting to discover. Used properly, the techniques can complement each other, giving more information together than any could separately.

4.4.3.1 X-Ray Crystallography

The capacity of proteins to form crystals has been known for over 150 years (284). Only many decades later, however, did the use of x-ray crystallography to determine the three-dimensional structure of proteins come into use (and even then, decades more were required for it to become a more commonplace technique) (285). Some of the earliest protein structures would take researchers decades of working on the same problem to produce (286,287). Today, there are nearly 160 000 x-ray crystallography structures in the PDB, and numbers grow every year at an exponential rate (the work contained in this thesis contains five of the new structures added this year, in **Paper II**) (201).

The process of x-ray crystallography first requires the production and isolation of a significant quantity of protein, the methods for which will not be discussed here (285). It is interesting to note that while protein crystallization was initially used as a technique to purify proteins, modern approaches require the starting protein to be pure, in order to limit the chances of crystallizing the wrong protein (285). Although the methods of producing crystals from purified proteins vary slightly, they involve slowly mixing the protein solution with that of a known precipitant solution (287). However, it is extremely difficult to predict whether a specific protein will produce a crystal when mixed with a given precipitant solution, so researchers often set up large screens involving dozens or hundreds of different conditions (287).

This initial stage can take a significant amount of time to produce a protein crystal (Figure 4.6), if one can be produced at all, and numerous methods of protein modification can be performed to try to compensate for a poorly-crystallizing protein (287). However, assuming one is able to obtain a crystal, one is still a long way from solving a protein structure. The crystal itself must be frozen using liquid

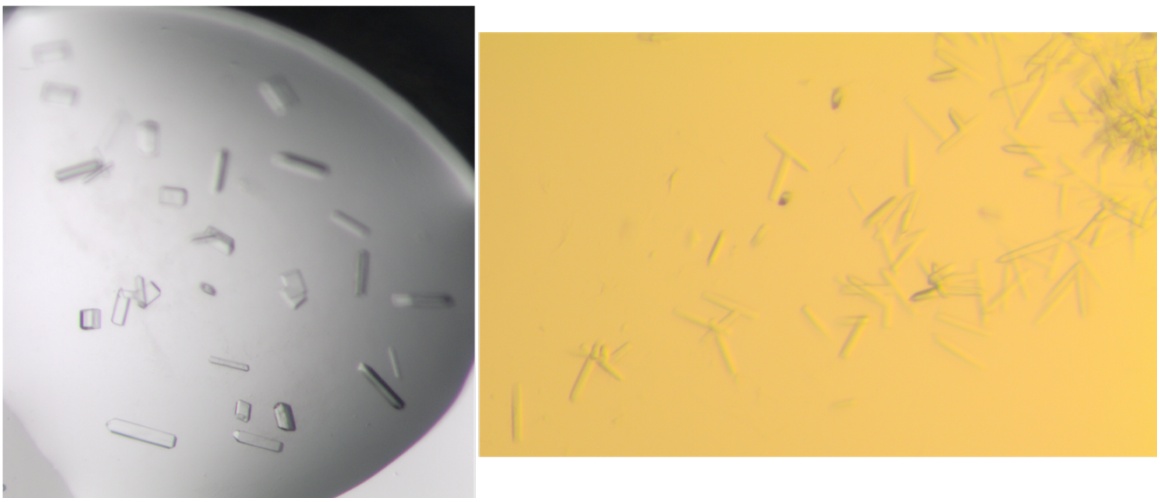

Figure 4.6: Protein crystals produced in experiments done to obtain the three-dimensional structure of *Ck*GE15A (**Paper II**). Several crystals from the samples seen here were diffracted at the ESRF, and the data collected were used as the basis for structural modeling.

nitrogen, usually in some sort of cryoprotectant solution to prevent damage, then shot with an x-ray beam and if the crystal diffracts the x-rays, the diffraction pattern can be collected (288). The diffraction pattern alone unfortunately does not provide researchers with all the information needed to obtain a protein structure (288). While the physical reasons are beyond the scope of this thesis (for an excellent source of detailed information, see the textbook Biomolecular Crystallography by Bernhard Rupp), the diffracted x-rays give information about their amplitudes when collected, but not their phases, and both are required to reconstruct a protein structure from the collected diffraction pattern (Figure 4.7, example diffraction pattern). The phase information can be obtained either through incorporating heavy atoms into the crystal, either by soaking or by direct incorporation into the protein prior to crystallization, or through using the phase information from a known, previously discovered protein structure. With this information, it is "simply" (although the process seems anything but simple while undergoing it) a matter of effort and computational power to produce the end protein structure (288).

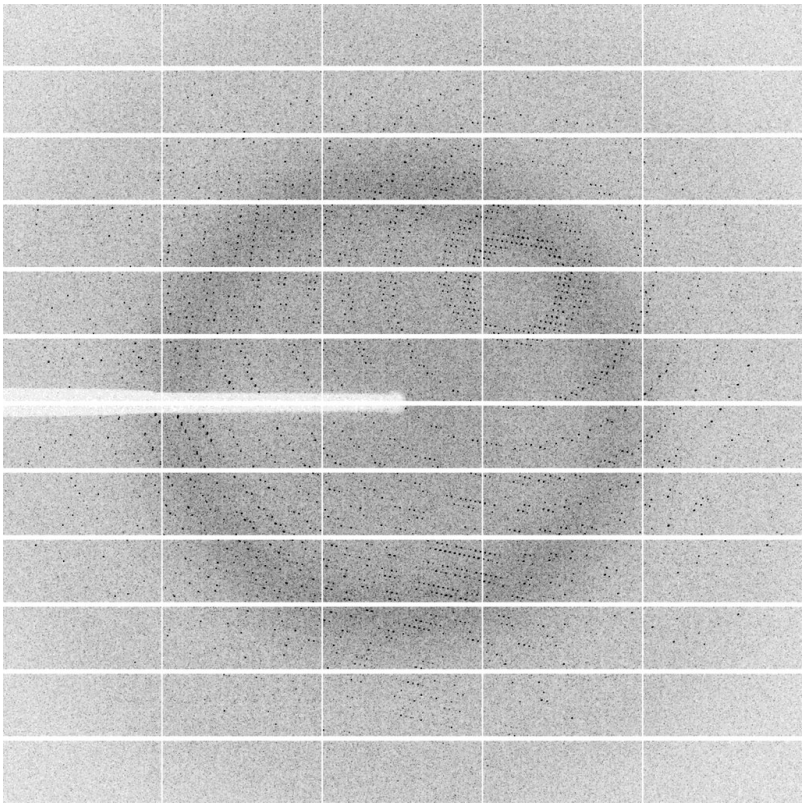

Figure 4.7: Example x-ray scattering image. This image was taken from experiments performed at the ESRF in November of 2017. The diffraction pattern is one image from a set collected for *Ck*GE15A. Several hundred of these images need to be collected and the data analyzed in combination to have diffraction data from all possible angles in order to begin to solve the protein structure.

Although it can be challenging, x-ray crystallography presents a number of advantages to researchers interested in determining protein structure. For instance,

the vast majority of the highest resolution protein structures available have been determined by x-ray crystallography, and it is arguably the most mature of the protein structural determination methods (201,289). It also requires low start up investment, as most protein researchers will be trying to obtain large amounts of their protein of interest anyways (288). Since proteins are in crystal form when obtaining structural information, they can reasonably be expected to be somewhat homogeneous in their conformation, greatly simplifying data processing. X-ray crystallography is also not limited in the size or type of protein it can analyze (with perhaps the exception of intrinsically disorders proteins) (288).

X-ray crystallography does present drawbacks that need to be taken into consideration. The process (very briefly) described above to obtain data can prove extremely challenging, and problems can occur at many of the steps involved (285,287,288). As the protein is trapped into a single state in the crystal, dynamic studies are not possible, and the structure obtained is merely a "snapshot" of the native protein (288). As well, the crystallization itself may introduce artefacts into the structure that are not naturally present, and can be overinterpreted if the researcher is not careful (288).

In this work, x-ray crystallography was used to obtain several structures in **Paper II**. Structures of a CE15 domain as well as a CBM9 domain were obtained through crystallography. With the CBM9 structure, it was also possible to obtain the structure with various ligands in the binding site. The CBM9 structure was the second ever solved, and the CE15 domain showed a lack of several inserted regions that have previously been identified in all solved bacterial CE15 enzyme structures.

4.4.3.2 Small Angle X-Ray Scattering

Small angle x-ray scattering (SAXS) is a technique used to obtain the overall shape of a protein, rather than its detailed structure in high resolution (290). Compared to x-ray crystallography, sample preparation is relatively straightforward. The protein of interest must be as pure as possible, and ideally, prepared in several different concentrations (291). The concentration required is inversely related to the size of the protein – smaller proteins generally require a higher concentration to obtain a good signal (291). An x-ray beam is directed towards the protein sample, and the scattering of the x-rays by the sample is recorded (290). Through mathematical modeling of the obtained data, data points resulting from aggregation or protein interaction can be excluded, and a model of the protein shape determined (291).

While sample preparation for SAXS is vastly simpler than for other methods of structural determination, the major drawback is that the resolution is limited to 10 Å (compared to the possible sub-ångström resolution for x-ray crystallography) (285,291). SAXS data therefore only provides the overall shape of the protein, and no information on individual amino acid positions or interactions can be obtained

(291). The measurements are also sensitive to protein aggregation (although this can be somewhat compensated for), even when that aggregation is not noticeable beforehand. However, SAXS does have the advantages that there is no size limitation on the proteins being studied, and it provides insight into protein structures and behaviors in solution. This can be used to explore changes in protein conformation, as well as interactions between proteins (291).

SAXS was used to explore the structure of the first three domains of *Ck*Xyn10C-GE15A in **Paper II**. While the interaction of the three domains could be determined using this technique, and they were seen to be compact rather than elongated, the data ultimately did not answer all questions about the domains. SAXS, while not theoretically limited in application temperature, is limited by the equipment used. Most, if not all, SAXS capillaries (used for storing the protein sample when in the beamline) are incapable of handling temperatures above 50°C. Thus, *Ck*Xyn10C-GE15A was unable to be tested at its natural environmental temperature of 80°C, which may have produced a different conformational result.

4.4.3.3 Nuclear Magnetic Resonance

Another method of obtaining protein structure is using nuclear magnetic resonance (NMR) spectroscopy (201). NMR is currently the second most common method of structural determination for structures deposited in the PDB (although there is a large gap between it and x-ray crystallography), with nearly 13 500 structures available to date (201). This technique works by exploiting the magnetic spin properties of atomic nuclei by placing the sample under the influence of an extremely strong magnet, and then probing it with radio waves and measuring their absorption (292). This gives information on the environment of each atomic nucleus, as well as which atomic nuclei are nearby, and from this (along with the protein sequence), researchers can build a three-dimensional model of the protein (293).

NMR does have some significant drawbacks, including the expense and difficulty of sample preparation – in order to obtain NMR signals, radioisotopes must be used, commonly two of ^{1}H, ^{13}C, and ^{15}N (although ^{31}P is also possible) (293,294). Additionally, the data obtained from NMR techniques, especially on larger proteins, can be very difficult to interpret correctly (293). Finally, NMR has a significant size limitation – very few structures over 50 kDa in size have been solved by solution state NMR (the most common type) (295). While newer techniques such as solid state NMR promise – at least in theory – to overcome this limitation, more work needs to be done in the area for it to become a generally viable technique at larger protein sizes (295). However, despite these disadvantages, NMR offers several appealing qualities that make it favorable for some research applications. For instance, NMR can offer a high quality structure of a protein in solution (assuming solution state NMR and not solid state NMR is performed), and can also provide information on protein dynamics, conformational changes, and ligand binding (296).

4.4.3.4 Cryo-Electron Microscopy

The last experimental method to be discussed in this thesis is cryogenic-electron microscopy (cryo-EM). The first structure released using electron microscopy only became available in 1991, and it took five years from that time for a second to be released (201). To date, fewer than 8 500 structures have been determined using cryo-EM, with over 75% of them being released in the past five years (201). Technological capability and availability for researchers to perform cryo-EM to obtain detailed protein structures has only begun to become more widely available in the last decade (297). To conduct cryo-EM experiments, a researcher will generally freeze their protein sample in a support mesh, and then using an electron microscope, collect between several hundred and 50 000 images of individual particles (298). These images are then constructed into a three dimensional model using computer software (298).

The largest advantage for researchers choosing to pursue cryo-EM is that sample preparation is theoretically extremely simple – the researcher only needs to apply the protein sample to a support grid in a thin layer and freeze it (298). Individual proteins may prove to be more difficult, however, and the sample preparation step is incredibly important to have correct before moving on in the cryo-EM experiment (298). Cryo-EM also does not become more difficult with larger proteins (299). This, however, leads in to one of the major disadvantages of cryo-EM: small proteins are extremely difficult to work with, and there is a lower size limit of approximately 50 kDa on the proteins that can be analyzed (299). While there are some techniques to help get around this size limitation, very few structures have been solved below 50 kDa in size (and even structures at that size are rare) (300). Additionally, historically, resolution obtained with cryo-EM has been considered poor, however, this has improved greatly in recent years (201,300).

4.4.3.5 Computer Modeling

For a long time, the field of protein structure has had a significant “protein folding problem” – knowledge of a proteins primary structure was easy to come by, but understanding how the protein managed to fold into its three dimensional conformation is incredibly difficult (301). However, with the huge growth of known protein structures in recent years, the problem has come significantly closer to being solved (201,301). Using existing information, a number of different protein modelling solutions have become available to researchers to allow them to obtain an approximation of their protein structure of interest, taking only time on a computer rather than work in a lab. Several of these methods will be discussed briefly below, however, this is far from an exhaustive list of the programs available.

SWISS-MODEL was the first fully automated protein homology server, where a researcher needs to only input the amino acid sequence of the protein of interest and

then wait for results (302). The software can perform sequence alignments to proteins of known structure and attempt to generate a structure for the inputted protein sequence based on close homologs. Energy minimization is performed by the software to resolve any clashes, and the model is output along with statistics estimating its accuracy (302).

Another protein homology server that sees a significant amount of use by researchers is Phyre2, the successor to the original Phyre (Protein Homology/analogY Recognition Engine) server (303). Like SWISS-MODEL, Phyre2 is a homology modelling server in which a user inputs an amino acid sequence, and the server searches for homologs and attempts to build a model. In this case, a series of models are generated, and different templates can be used for different regions of the protein. *Ab initio* modelling is performed on regions without clear templates, and the models are output along with statistics to estimate their accuracy (on a per-residue basis) (303). Phyre2 was used to generate the models of *Be*CE15A-Rex8A in **Paper III**, which enabled several hypotheses to be generated and tested through mutational analysis. The non-crystallized CBM9 modules from *Ck*Xyn10C-GE15A were also modeled using Phyre2 in **Paper II**, revealing potential reasons for the differences in binding between them, and suggesting a significantly different structure for the first CBM9 domain.

Rosetta, unlike the previous two examples, is a software package to perform *de novo* structure prediction on a target protein sequence (304). Rosetta is one of the most successful *de novo* modelling software packages and works by assembling the protein model as small blocks of short, modelled peptides. Multiple models can be generated at this stage (depending on user input), which then undergo significant refinement and energy minimization steps. The resulting models are output to the user, and can in many cases be more accurate than the homology modeling servers discussed above (304).

The newest and perhaps most exciting development in the protein modelling field is the introduction of AlphaFold. AlphaFold uses trained AI neural networks to predict protein structures with incredible accuracy (305). In the past two Critical Assessment of protein Structure Prediction (CASP) competitions, CASP13 and CASP14, AlphaFold has shown to be a significant improvement over its competitors (CASP14 data to be formally published late 2021) (306). AlphaFold has since been used to generate a database of over 350 000 protein structure predictions which are freely available to researchers (307). While AlphaFold does not represent a complete solution to the protein folding problem, and a significant portion of its predictions are not perfect, it does represent a significant leap forward in protein prediction, and the improvements in version 2 (the version entered into CASP14) are exciting (305). Perhaps the biggest hurdles preventing AlphaFold from being more widely used are its lack of a web server for submissions (available for the

previous three methods), its rather specific computer hardware requirements, and that it is an extremely computationally intensive method (which should become less of an issue in the future, as computers continue to improve and prices lower).

4.5 Enzyme Discovery

The discovery of new enzymes with more beneficial properties for biorefineries and other industrial uses is a major driving force behind research into lignocellulose-degrading organisms. Discovery of rapid and cost effective DNA sequencing has caused an explosion in the number of known or computationally predicted protein sequences (308). New sequences are being discovered at a torrid pace; the TrEMBL database currently houses over 200 million predicted sequences, with nearly fifteen million being added between April and June 2021 alone (the database itself was established in 1996 and has grown exponentially almost ever since) (309,310). With this information, it has become easier than ever to discover new enzymes with desirable characteristics, although researchers may still source organisms from field research (311).

Using databases such as UniProt (310) and CAZy (170), one can find an extensive list of proteins and enzymes with computationally predicted, but untested, characteristics. These databases are typically assembled computationally, analyzing new genome sequences as they are published for predicted genes, translating protein-encoding regions, and comparing the sequences to characterized proteins to predict function (170,310). Unfortunately, this approach does not always produce the correct results – in some enzyme classes, up to 78% of the database sequences have been shown to be misannotated (although the error rate is generally much lower) (312). Additionally, over 20% of predicted protein domains are described as being of “unknown function”, suggesting a significant shortage of experimental information to complement these computer-generated databases (313).

Despite these drawbacks, the first stop of researchers searching for new enzymes is often these databases (although some researchers prefer using functional screening to find and isolate enzymes) (314,315). A database search, combined with analysis of the genomic context of the protein-encoding gene and the environmental context in which the encoding organism lives can help identify potential enzymes of interest. Continuing further, a sequence alignment with characterized enzymes of the same type (to confirm essential features are intact), model structure building using various software, and other bioinformatic analysis can significantly aid a researcher in identifying a new enzyme of interest (315). Of course, the identified enzyme or enzymes must still be characterized in a laboratory in order to confirm they have the desired function and properties (315).

If a researcher is looking for a novel enzyme function, discovery can be significantly more complex. This generally requires sampling of an environment in which the desired function is thought to exist within the microbial community (if it is already known that an organism with a sequenced genome can perform the function of interest, environmental sampling can be bypassed) (316,317). It is possible to enrich samples by adding a substrate of interest in abundance (for example, if a researcher were searching for a xylanase, adding xylan to the environmental sample), but this can have the drawback of encouraging growth of organisms that metabolize the end product of the desired enzyme, rather than the organism which produces said enzyme (316). After the desired activity has been verified in the sample, some sort of sequencing must be performed, either metagenomic, metatranscriptomic, or metaproteomic. From there, the proteins of unknown function can be analyzed in the laboratory to determine which of them has the desired function (316).

4.6 Modern Enzyme Producers

As mentioned in chapter three, enzymes are key tools utilized to degrade plant biomass. These enzymes are sourced from organisms that can be found in the natural environment. A plethora of organisms have evolved to survive off of and degrade lignocellulose, and are potential sources for these enzymes (130,318). While still highly debated in the scientific literature, recent data suggest that the ability to degrade lignocellulose has existed as long as there has been lignocellulose to degrade (130,319-321). These lignocellulose-degrading organisms (a group containing representatives from the bacterial and archaeal domains and the fungal kingdom) exist in all environments on earth that one can reasonably expect to find lignocellulose available, and have evolved a diverse set of characteristics to enable them to survive in these wildly different environments (322). Historically, fungi have been the best-studied of these organisms, however, recently focus on bacteria which can degrade lignocellulose has increased dramatically (130). In order for any of these organisms to degrade lignocellulose, they invariably need to use a diverse system of enzymes (323). An astonishing number of enzymes have evolved in different microorganisms to allow for the breakdown of every known bond within lignocellulosic materials (323). It is the lignocellulose-degrading enzymes from these organisms which are often exploited in industrial environments, and it is the search for these enzymes which largely fuels research into these organisms (324).

4.6.1 *Trichoderma reesei*

Currently, the undisputedly most important organism for enzyme production in a biorefinery context is *Trichoderma reesei* (325). *T. reesei* is a mesophilic filamentous fungus first identified during the Second World War, when it would destroy the cotton tents set up by the US army in the Solomon Islands (326,327). As early as the 1960's, this organism was used for industrial-scale production of cellulases (328). Although other fungi were used for similar reasons at the time, it was soon

discovered that *T. reesei* was a more efficient cellulase-producer than any other known organism, and work focused largely on improving yields from it (328,329). Starting in the early 1970's, experiments to increase the yield of cellulase production from *T. reesei* were undertaken (327). Although researchers of the time lacked the sophisticated molecular biology tools available today, by the end of the decade the strain RUT-C30 was developed, which would serve as the basis for almost all cellulase production from *T. reesei* organisms going forward (327,330,331). Since having its genome sequenced in 2008, the potential for *T. reesei* genetic manipulation and directed improvements has exploded (332). Today, *T. reesei* remains a highly relevant organism for industrial cellulase production, and modern strains can produce over 100 g of enzyme per liter of culture (333,334).

4.6.2 *Caldicellulosiruptor*

Recently, bacteria from the genus *Caldicellulosiruptor* have been of great interest to researchers for their potentially desirable qualities for use in biorefineries. First discovered in 1987 in a New Zealand hot spring, *Caldicellulosiruptor* organisms displayed an ability to grow on cellulose, hemicelluloses, and other complex glycans (335). These Gram-positive, anaerobic bacteria, which have since been isolated from locations globally, naturally exist at temperatures in the 70-80°C range, making them an idea target to search for new thermostable lignocellulose-degrading enzymes (336). Since the initial discovery of the genus, at least 14 separate species have been identified worldwide, with genome sequences available for all of them (337). In addition, tools to manipulate the genome are available for two species of the genus (338-340). This is somewhat of a rarity, as few anaerobic organisms have had genome-editing tools developed for them (341). This has allowed *Caldicellulosiruptor* strains to be engineered to produce ethanol from cellulose, however, the yields appear to decrease drastically at higher temperatures, and the maximum obtained yield to date (with pure cellulose as a carbon source) is still an order of magnitude lower than what *Saccharomyces cerevisiae* is regularly able to achieve (with Kraft pulp as a carbon source) (342-345).

Experimental evidence has shown these organisms to have many desirable qualities for biorefinery usage. For example, previous studies have shown that in their native environments, the microbial community is dominated by *Caldicellulosiruptor* at higher temperatures when grown on cellulose (337,346). This dominance suggests that it is very difficult for other organisms to grow at these temperatures, and that *Caldicellulosiruptor* are extremely efficient and effective at utilizing cellulose. Perhaps because of their incredibly challenging environment, organisms within the *Caldicellulosiruptor* genus have developed a truly impressive array of enzymes for degrading lignocellulose (347). The most studied of these is undoubtably the CelA cellulase from *Caldicellulosiruptor bescii* (226,348,349). This enzyme has been shown to be up to 6-fold more efficient than the most prominent *T. reesei* cellulase, Cel7A (226). It also remains more efficient even when both CelA and Cel7A are used under

the optimal reaction conditions for Cel7A (349). One of the reasons for this efficiency is a somewhat unique enzyme architecture, where multiple catalytic domains exist within the same protein (226). One of these multicatalytic *Caldicellulosiruptor* enzymes, CkXyn10C-GE15A from *C. kristjanssonii*, is studied in detail in **Paper I** and **Paper II**. Multicatalytic enzymes will be discussed in more detail in chapter five.

4.6.3 *Bacteroides*

Although *Caldicellulosiruptor* organisms are known for their production of multicatalytic enzymes, they are not the only genus to produce them. Organisms from the genus *Bacteroides* also have a demonstrated ability to produce various multicatalytic enzymes targeting the degradation of lignocellulose (350-352). While these organisms have previously been studied for other interesting lignocellulose-degrading mechanisms, they are coming into more focus lately as producers of interesting and unique multicatalytic enzymes (200,350-352).

Bacteroides are Gram-negative, anaerobic bacteria found in the intestinal tract of a variety of animals, including humans (350,353,354). They make up a substantial portion of the intestinal microflora wherever they are found; in humans they account for approximately 25% of all intestinal microbes (355). They are incredibly important organisms for human health, with implications in roles from obesity to brain development to bone strength (356-358). *Bacteroides* generally exist in a mutualistic relationship with their host organism, utilizing material that is not digestible by the host as a source of energy, and providing nutrients such as short-chain fatty acids to the host in exchange (359-361). Not only do they have a symbiotic relationship with their host, *Bacteroides* have also been shown to provide nutrients to other gut bacteria, making them a valuable contributor to overall gut health (362-364). Surprisingly, certain *Bacteroides* species have been observed to produce and secrete enzymes to aid in the degradation of polysaccharides that they themselves cannot utilize, suggesting this is done to benefit the overall microbial community, rather than the secreting organism (364). Within the gut itself, *Bacteroides* has access to a smorgasbord of lignocellulosic biomass (depending on the host diet). This biomass is typically consumed by the host organism, who is then unable to digest it, as no animal is currently known to be able to degrade lignocellulose to sugar monomers (365). This situation has led to *Bacteroides* developing very efficient strategies for breaking down lignocellulose into simple sugars that can be utilized by bacteria and host alike (366).

A semi-unique strategy utilized by *Bacteroides* is the organization of genes related to polysaccharide utilization into polysaccharide utilization loci (PULs) (367). Although not the only genus of organism to organize genes into PULs (this is a trait that occurs throughout the phylum Bacteroidetes, and similar, although less sophisticated, genomic structures are being discovered in many other bacteria),

Bacteroides are the most prolific at it, with almost 30% of known or predicted PULs contained within their genomes, and some species devoting as much as 20% of their entire genome to PULs (368-370). This is despite them accounting for only 12% of sequenced genomes contained within PULdb, a database of all known and predicted PULs (368).

The first discovered PUL was found to be involved in starch utilization in *Bacteroides thetaiotaomicron*, a human gut symbiont (371-374). To be classified as a complete PUL, the section of genome must encode proteins to bind polysaccharides, partially degrade them extracellularly, import the partially-degraded polysaccharides, complete the polysaccharide degradation inside the cell, and proteins which can act as regulators for the entire system (367). PULs can be predicted from genomic data by searching for the presence of SusC and SusD homologs (named from the first studied PUL, the Starch Utilization System). SusC and homologs in other PULs are TonB-dependent transporters importing partially degraded oligosaccharides into the cell, and SusD and homologs are responsible for polysaccharide recognition and capture (367,375). The presence of both of genes encoding both of these proteins in the same region of a genome on the same DNA strand is a very strong indication of the presence of a PUL (367). Analysis of PULs suggests that they are spread amongst *Bacteroides* via horizontal gene transfer (375). Additionally, evolutionary pressure has allowed for the adaptation of PULs to various polysaccharide substrates. Overall, these two features have allowed *Bacteroides* to develop an adaptable polysaccharide utilization strategy, which has likely contributed significantly to their evolutionary success (375).

Although they are not found in every *Bacteroides* species, multicatalyic polysaccharide-targeting enzymes exist in many of the published *Bacteroides* genomes (170). Unlike with *Caldicellulosiruptor*, however, these have been studied relatively little by the scientific community. Only a few examples of publications highlighting multicatalytic enzymes from *Bacteroides* exist (**Paper III**)(352,376). Despite this, they remain a promising avenue for the development of novel enzyme cocktails for industrial use, especially when taken in the context of the overall efficiency of *Bacteroides* at degrading plant biomass.

4.6.4 Industrial Enzyme Production

Enzymes for industrial use are almost always microbial in origin, due to the fast growth rate, consistent production levels, and ease of enzyme production within microbes as compared to higher organisms, as well as difficulty in isolating mammalian or plant enzymes from source organisms and the limited range of temperature and pH those enzymes are generally active at (377,378). Microbes are also often simple to genetically modify, to increase production of the enzyme of interest in the natural organism, or to produce it in an entirely different organism instead (378). The ideal production host depends on a large number of factors,

including protein origin, available equipment, and end use of the enzyme being produced (377). Several (but certainly not all possible) production hosts will be discussed below. As *T. reesei* has already been discussed above, it will not be covered again here, but it is worth noting that several industrially important CAZymes are produced by it (378).

Fungi from the genus *Aspergillus* have long been the most commonly used industrial enzyme producers (378-380). *Aspergillus* are generally used to produce native *Aspergillus* enzymes, with various research efforts focusing on strain improvement to increase yields of relevant enzymes (381). Species in this genus are known to be lignocellulose-degrading, and have been used to produce CAZymes for industrial purposes (381). *Aspergillus* has also shown increased enzyme yields when co-cultured with other organisms – either multiple *Aspergillus* species, or one *Aspergillus* and another fungus such as *T. reesei* (381,382).

Another highly useful genus of microorganisms, bacterial rather than fungal, is the *Bacillus* genus. *Bacilli* are easy to manipulate genetically, and also produce a number of industrially-relevant enzymes naturally (383). They are known to natively produce several CAZymes, and are able to easily be genetically modified to produce CAZymes from other bacterial species (378,383).

While not often used for industrial production (although there are cases where it is used), *Escherichia coli* is one of the most important microorganisms for enzyme production, simply because it is quite often the first choice of researchers investigating a new enzyme (378,384). This organism has highly effective genetic manipulation methods available to researchers and is extremely quick to grow and produce protein. Its production levels rarely match those of industrial producers, but its significant advantages lead to it being utilized very often by researchers, and sometimes by industrial processes as well (384).

Chapter 4: Summary

- A large variety of CAZymes is responsible for the degradation of plant biomass
- GH enzymes include cellulases, xylanases, and mannanases, and are responsible for the degradation of glycosidic bonds
- CE enzymes break ester bonds within lignocellulosic biomass
- CBMs have a variety of roles, not solely in carbohydrate binding
- Enzymes can be characterized in a variety of ways, both biochemically and structurally
- A variety of different CAZyme producers exist and can be exploited

Chapter 5: Multicatalytic Enzymes

As briefly discussed in chapter four, multicatalytic enzymes are those in which multiple independently folded catalytic domains are present within the same polypeptide chain (385). These enzymes can present several distinct advantages over other enzyme architectures (discussed below) which make them extremely interesting targets for study. While they are not without their challenges, the potential of these enzymes to aid in lignocellulosic biorefineries is extremely exciting. This chapter will discuss different enzyme architectures, including multicatalytic enzymes, producers of multicatalytic CAZymes, and methods of producing these enzymes more effectively in the future.

5.1 Enzyme Architectures

With the overall size and diversity of CAZymes, it should come as no surprise that they present a variety of enzyme architectures which allow them to function successfully. Evolution has allowed organisms to develop three main enzymatic architectures to degrade lignocellulose, each with its own advantages and disadvantages (386,387).

5.1.1 Free Enzymes

Perhaps the most basic, and by far the most common, strategy used for lignocellulose degradation is the secretion of free enzymes into the extracellular environment (Figure 5.1 A) (387). This system is adopted with a high degree of success by aerobic cellulolytic fungi, some of the first cellulolytic organisms to be studied (388). In this system, the organism produces a large quantity of CAZymes suitable for the environment, and secretes them extracellularly (334). The enzymes must be secreted, as the overall polysaccharide to be degraded is far too large to bring into the cell. Typically, these consist of a single catalytic domain, at times appended with one or more CBMs. Free, secreted enzymes have the distinct advantage of being relatively simple to produce and secrete for the microorganism (389). Unlike with other architectures, they do not require extra domains for assembly, and they are fairly limited in size and complexity (386,387,389). Unfortunately for the microbes, this ease of production comes with the downside that there is nothing to force the free enzymes to exist together in one place on the polysaccharide. This means that enzymes which may work synergistically are not necessarily distributed on the polysaccharide for optimal degradation, and must rely on chance for the opportunity to work together (387).

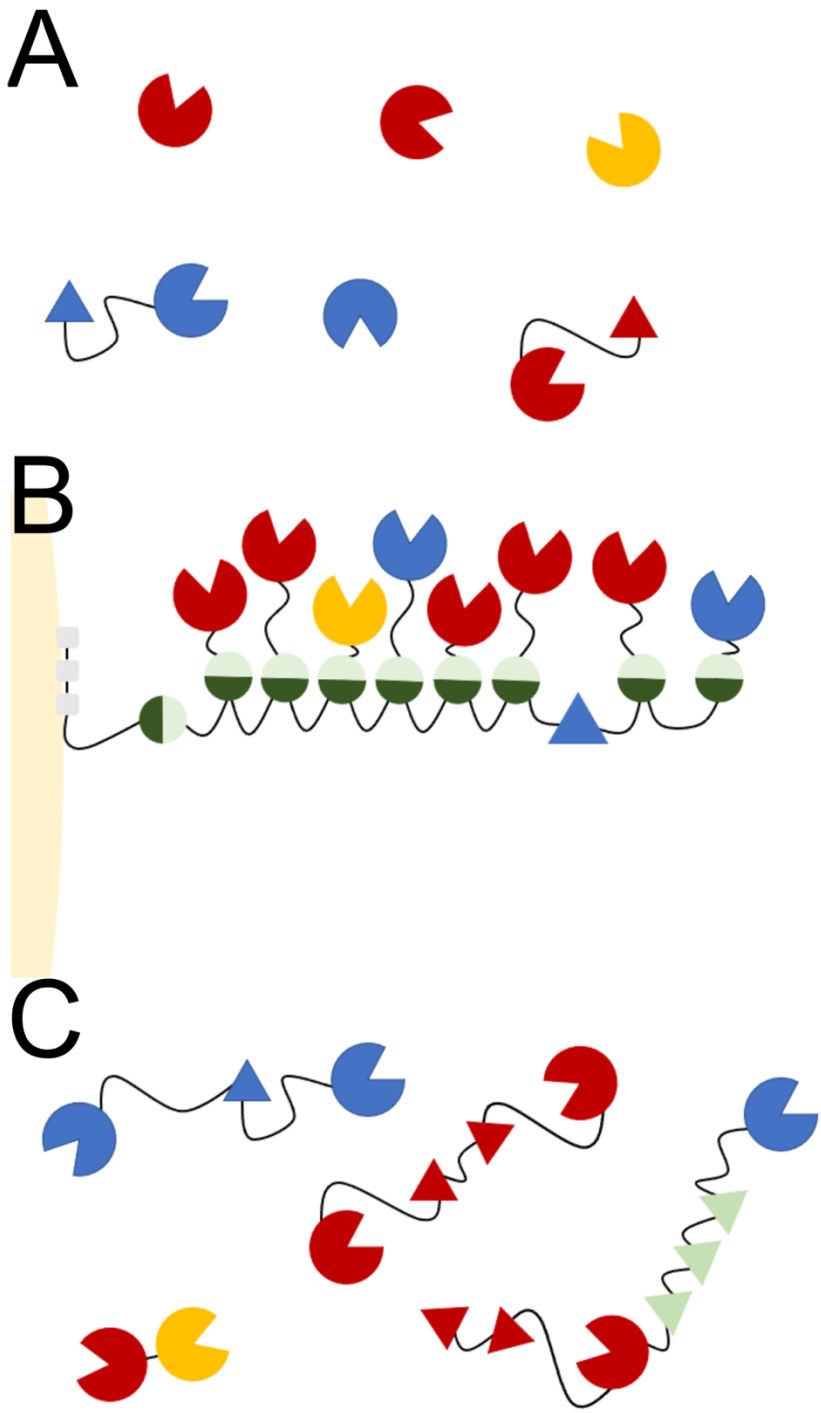

Figure 5.1: Schematic diagram of the differences between free enzymes (**A**), cellulosomes (**B**) and multicatalytic enzymes (**C**). Enzymes are represented by circles with a missing wedge (xylanases in red, cellulases in blue, and CE domains in yellow), CBM domains are represented by triangles (xylan-interacting domains in red, cellulose interacting domains in blue, and mannan interacting domains in green), SLH domains represented by squares, and cohesin (dark green) and dockerin (light green) domains are represented by half circles. Linker regions are represented as black lines, and the bacterial cell as a light yellow curved and cut circle in (**B**).

5.1.2 Cellulosomes

Significantly more complex than just free enzyme secretion, the cellulosome is a massive protein structure produced by some anaerobic fungi and bacteria (Figure 5.1 B) (387,390). This structure consists of multiple enzyme domains, from as few as six to more than 100 in some cases, linked together through interactions with a large scaffold domain known as a scaffoldin (391,392). The scaffoldin contains multiple cohesin modules, which bind with dockerin modules integrated into the sequence of the enzyme domain containing polypeptides (388). Although the precise mechanisms behind cellulosomal arrangement are poorly understood (it is known, however, that assembly is not random), different classes of cohesin domains interact with different dockerins, allowing some selectivity in arrangement (393,394). As well, interactions between cohesin domains within the scaffoldin protein can impact which enzyme domains bind in which positions (395). It is important to note that cellulosomes are almost exclusively anchored to the outer membrane of the cell and are not freely diffused into the extracellular environment (226,388,389,393,395).

Cellulosomes present several major advantages for the producing organism. Within the cellulosome, enzyme domains that perform related functions are often located in close proximity in the overall structure, allowing for reaction synergy, and greatly increasing the rate of polysaccharide degradation (388,396). Secondly, by keeping the cellulosome cell-anchored, the microorganism ensures that any sugars released by the cellulosome are in close proximity, and easily available for uptake into the cell (389). This proximity of the cell to the polysaccharide being degraded also helps prevent any product inhibition of the enzymes within the cellulosome, as the products are taken up by the cell as they are produced (389).

Cellulosomes, although they provide several advantages over free enzymes, are not without their drawbacks. The many non-enzymatic components involved in the cellulosome inevitably require more resource investment from the cell than simply secreting free enzymes, and a much higher energy cost to the cell before it starts to benefit. The proximity of the cellulosome to both the cell and the polysaccharide also precludes any community degradation of the substrate as discussed earlier, requiring the cell to produce most (or all) enzymatic components necessary on its own. Cellulosomes also have several disadvantages in the context of their use in lignocellulosic biorefineries. First, cellulosomes are generally limited to the surface of the producing cell, and are not freely diffused into the environment (397). Because of this, the overall number of cellulosomes produced is limited (397). Cellulosomes have also been seen to be incapable of two dimensional diffusion across the cell surface, further limiting their spatial distribution (398). Adapting natural cellulosomes for use also has the problem of cellulosome assembly – cohesin and dockerin domains are often species-specific, limiting their utilization and customization (399). Cellulosomes are also difficult to produce at an industrially-relevant scale, which is a significant hurdle to their adaptation in biorefineries (400).

5.1.3 Multicatalytic Enzymes

A third approach to CAZyme production is a somewhat hybrid strategy of the previous two. In this approach, the organism produces multicatalytic enzymes, where multiple catalytic domains exist within the same polypeptide chain (Figure 5.1 C) (386). Indeed, enzymes with this character are some of the most efficient CAZymes discovered so far, indicating that this strategy is quite efficient (226). It is important at this point to note the distinction between multicatalytic and multi-domain proteins: while multicatalytic enzymes have multiple active enzyme domains within one polypeptide chain, multi-domain proteins only need to have multiple distinct protein domains within the same polypeptide chain. These domains do not have to be enzymatically active. Thus, all multicatalytic enzymes are multi-domain proteins, but not all multi-domain proteins are multicatalytic enzymes.

Utilizing this multicatalytic enzyme strategy, microorganisms are able to exploit many of the advantages of cellulosomes with the ease of production and mobility of free enzymes. For example, many multicatalytic enzymes show increased synergy between the catalytic domains when added to substrates as a full-length protein, as compared to the two catalytic domains added separately (349,352,401,402). Somewhat surprisingly, synergistic action of the included enzyme domains does not seem to be a general feature of multicatalytic enzymes, as there are many examples where no synergy has been observed, and in these cases, it is somewhat unclear what the advantage to the multicatalytic architecture is (261,403). Multicatalytic enzymes are able to be both freely secreted into the extracellular medium similar to free enzyme domains, as well as anchored to the cell surface like a cellulosome (this is seen in *Ck*Xyn10C-GE15A, studied in **Paper I** and **Paper II** (the cell anchoring domains themselves were not studied, however) (51,226,403). This allows flexibility to the producing organism in that it allows the enzymes to function in the most effective way possible, whether that is anchoring the organism to the polysaccharide, or burrowing into crystalline cellulose (226).

The study of multicatalytic enzymes can present particular challenges to the researcher. Due to their typically large size, they can be extremely difficult to produce recombinantly (see **Paper I**, **Paper II**, **Paper IV**, and the section discussing challenges producing multicatalytic enzymes later in this thesis). It is therefore often necessary to divide the overall protein into its individual domains, and express and investigate them individually. This adds difficulty in the investigation of the synergy between enzyme domains. Studying multicatalytic enzymes should also involve studying the synergy between enzyme domains. As mentioned previously, domains linked in a multicatalytic fashion are expected to show synergy in their activities. In testing for this, the use of model substrates is often limited or impossible, as they may only interact with one of the enzyme domains (**Paper I**, **Paper III**). Therefore, an assay utilizing a natural substrate, or one as close to the natural substrate as

possible, must be devised. As previously mentioned, synergy studies are also better performed on the intact enzyme, but in some cases this is impossible (**Paper I**). Both enzyme domains can be added to the substrate together and observed for synergy, but this does not always produce synergistic results (**Paper I**, **Paper III**) (352).

The structural study of multicatalytic enzymes can also be a particular challenge. Although x-ray crystallography is generally the most common way to obtain protein structures, multicatalytic proteins present difficulties with this technique, as the flexible linkers often render proper crystallization impossible (**Paper I**, **Paper II**). In this case, the use of other techniques, such as SAXS or cryo-EM might be used, however, both have their drawbacks (and may ultimately prove unsuccessful in this case) compared to x-ray crystallography, as discussed earlier. It can often be necessary to combine multiple structural techniques in order to obtain a full picture of the structure of the enzyme; for example, x-ray crystallography of the individual domains, and whole protein SAXS, similar to what was attempted in **Paper II**.

Despite the challenges associated with their study, multicatalytic enzymes present an exciting prospect for industrial use. They often show superior stability and specific activity, without the need for non-covalent cohesin-dockerin interactions (386). They can also degrade lignocellulosic biomass more effectively than free enzymes (386). With their somewhat simplistic modular nature, they are also more straightforward to engineer, although small differences can have a large impact on enzyme synergy (402,404).

5.2 Organisms That Produce Multicatalytic Enzymes

Thus far, multicatalytic enzymes are not extremely common in the scientific literature, although reports on these enzymes appear to be increasing in frequency recently (51,352,385,405). However, in browsing through CAZy, one can find that many genomes appear to have at least one annotated multicatalytic enzyme (170). This suggests that this enzyme architecture is beneficial to many microorganisms, at least in some lignocellulose degradation processes. The presence of multicatalytic enzymes within several genera (for a more detailed discussion, see chapter four) which appear to have an elevated amount of these enzymes is discussed here.

5.2.1 Multicatalytic Enzymes in *Caldicellulosiruptor*

As of the time of writing, there are ten *Caldicellulosiruptor* genomes annotated within CAZy, containing a total of 40 multicatalytic enzymes (oddly, *Caldicellulosiruptor hydrothermalis* does not appear to encode any multicatalytic enzymes) (170). These multicatalytic enzymes appear to mostly contain GH enzymes, although there are also occasional CE and glycosyltransferase (GT) enzymes involved (170). A number of these enzymes have been studied, and most

display a synergy effect where the linked domains are more efficient linked together rather than separated (226,349,406-408).

Although a few have shown the ability to degrade xylan, the majority of studied multicatalytic enzymes from *Caldicellulosiruptor* appear to be cellulose degrading (407). The most well-studied of these, CelA from *C. bescii*, is not only more efficient than leading industrial cellulases, but displays a novel mechanism of action that takes advantage of its multicatalytic architecture (226,349). In this mechanism, CelA was shown to dig cavities into polymeric substrates, degrading not only the top layer of polysaccharide, but also the layers underneath, enabling it to work much faster than traditionally used cellulases (226,349). While such in-depth studies have not been performed for other *Caldicellulosiruptor* multicatalytic enzymes, the synergies displayed by these enzymes is a promising indication of their biorefinery potential.

5.2.2 Multicatalytic Enzymes in *Bacteroides*

Significantly more *Bacteroides* genomes than *Caldicellulosiruptor* genomes are currently annotated in CAZy – 66 in total (170). The genus contains nearly 450 multicatalytic enzymes, with each individual species having at least one, and some having up to 25 (170). Despite the large number of multicatalytic enzymes produced by this genus, they remain severely understudied, with far fewer published examples of *Bacteroides* multicatalytic enzymes than those from *Caldicellulosiruptor*. In fact, there are only two publications on the characterization of multicatalytic enzymes from *Bacteroides* species, a CE6-CE1 enzyme from *Bacteroides ovatus* (*Bo*CE6-CE1), and a GH8-CE15 enzyme from *Bacteroides eggerthii* (*Be*CE15A-Rex8A, the enzyme studied in **Paper III**) (200,352,385).

5.2.3 Examples of Multicatalytic Enzymes in Other Organisms

As previously mentioned, it is relatively common for the genomes of polysaccharide-degrading organisms to encode at least one multicatalytic CAZyme. Although few characterized examples exist in the literature, there are several notable enzymes that will be discussed briefly here.

ChiA from *Flavobacterium johnsoniae* is a multicatalytic chitinase which shows synergy between its two glycoside hydrolase family 18 catalytic domains (409). The synergy between these is remarkable, showing an approximately 20-fold increase in activity when the individual enzyme domains are added into solution together (as opposed to individually), and a further 6-fold increase in activity was shown for the full polypeptide (409). *Fj*CE6-CE1 is another interesting enzyme produced by *F. johnsoniae* (352). It is one of only three multicatalytic CAZymes in which all enzymatic domains are CE enzymes that have been characterized to date (the others being *Bo*CE6-CE1 from *B. ovatis* and *Dm*CE1-CE1 from *Dysgonomonas mossii*) (351,352). Oddly, this enzyme did not show synergy when tested, although it still

proved more effective than *Bo*CE6-CE1 at increasing xylanase activity when combined with a GH11 xylanase (352). It is possible that the appropriate biomass was not selected for testing for this enzyme, but if this is the case, it highlights the complexity and variety of biomass, and the plethora of strategies that have evolved to degrade it (352).

5.3 Production of Multicatalytic Enzymes

Multicatalytic enzymes can of course be studied after homologous production by the encoding organisms, and has been demonstrated (226). Often, however, this is not a feasible approach due to the difficult conditions required to cultivate many of these organisms and the lack of expression control (226). The use of recombinant production methods is more common, but it too is not without difficulty (51,200,261,352). Often recombinant methods are only capable of producing individual domains, and not the full-length protein, suggesting that there is no one correct approach for the production of these enzymes (51,226,261).

4.3.1 Challenges in Producing Multicatalytic Enzymes

Production of multicatalytic enzymes can be extremely challenging, and is often unsuccessful (this was unceremoniously demonstrated in **Paper I** and **Paper II**, in which the full length enzyme was never successfully produced and purified) (51,261). As an example, in the case of CelA from *C. bescii*, almost all production of the protein discussed in the scientific literature involves using homologous expression of the enzyme or recombinant production within the native organism, rather than production in more common expression hosts (226,348,349). Even this production can prove to be highly challenging – Brunecky *et al.* required the use of a 600 L fermenter growing *C. bescii* (at 75°C and under anaerobic conditions, which are conditions that may prove difficult for many laboratories) in order to obtain enough CelA for their experiments (226). Attempts at production of full-length CelA in *E. coli* have repeatedly failed, although fragments have been successfully expressed (226,410). The only successful expression of recombinant full-length CelA so far has been in the uncommon expression host *Bacillus megaterium* (410).

A major problem with recombinant production of multicatalytic enzymes tends to be the size of the enzyme itself. Using *E. coli*, the most common laboratory recombinant protein production system, the larger the protein being produced, the less likely it is to be produced successfully (411). Even with other systems, there is no simple answer to the problem of large protein production (412-414).

5.4 Multicatalyic Enzymes Studied in This Thesis

In this work, two multicatalytic enzymes have been studied in detail. In **Paper I** and **Paper II**, the enzyme *Ck*Xyn10C-GE15A from *C. kristjanssonii* was studied. This enzyme consists of two CBM22 domains, a GH10 domain, three CBM9 domains,

and a CE15 domain (Figure 5.2). An unstudied C-terminal portion also includes a cadherin domain and two SLH domains. This enzyme was studied, in part, because the CE15 domains in *Caldicellulosiruptor* organisms are rare (while the presence of GH10 members is extremely common) (170). Analysis of the enzyme domains revealed a different pH optimum for each, which was surprising, since they are physically linked together. Further activity analysis was unable to detect any synergy between the two enzyme domains. However, the rare enzyme architecture could indicate that there is something unique about the lignocellulosic biomass in the environment in which *C. kristjanssonii* is found that was not adequately replicated in the synergy experiments.

Figure 5.2: Diagram of the domain layout for *Ck*Xyn10C-GE15A. CBM22 domains are in red, the GH10 domain in blue, CBM9 domains in cyan, and the CE15 domain in orange. Linker regions are in green. All domains are models generated with Phyre2 (303), with the exception of the GH10 and third CBM9 domain, which were crystalized and had structures solved in **Paper II**.

Paper II was focused on the same enzyme, but with more emphasis on structural characterization. The structure of the CE15 domain, along with that of the third CBM9 domain were obtained, and the overall shape of the first three domains were obtained using SAXS. Combined with computer modeling, this was able to provide a picture of the whole enzyme (Figure 5.2). The extended conformation of the protein suggests that the enzyme domains can act at a significant distance from the cell surface, especially when the linker between the CE15 domain and the cadherin and SLH domains is taken into account.

*Be*CE15A-Rex8A (Figure 5.3) studied in **Paper III** consists of a CE15 domain closely linked to a GH8 Rex domain. For this enzyme, the presence of a CE15

domain fused to a Rex domain was a surprise, as those enzymes are typically expected to function in different physical locations (the periplasm for a Rex domain and the extracellular environment for the CE15A domain) (194,195,239). The CE15 domain also displayed an unusual active site, in which a normally-conserved arginine was substituted with a phenylalanine. Because of its poor activity, the phenylalanine was replaced with the arginine, but no activity was not recovered – in fact, no activity was seen at all with the arginine-containing variant. No synergy was seen between the two domains (wild-type Rex8A and CE15A) when tested on complex biomass substrates. It therefore seems possible that, in combination with the atypical active site residues present within the CE15 domain, the true target of this domain is not the typical CE15 target at all.

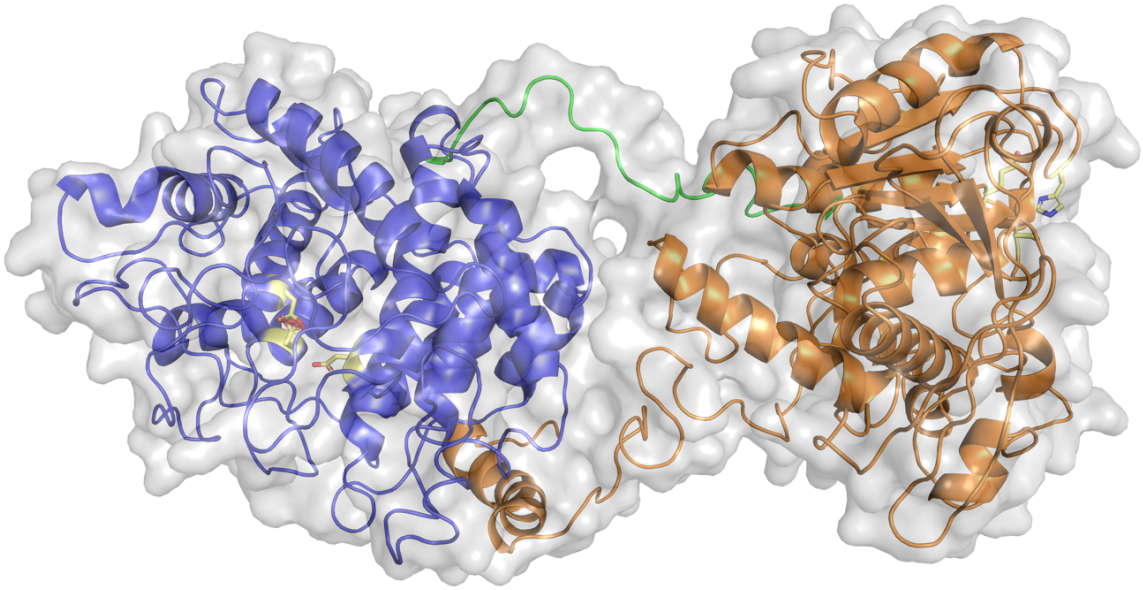

Figure 5.3: Diagram of the domain layout for *Be*CE15A-Rex8A, generated using Phyre2 (303). The GH8 domain is seen in blue, the CE15 domain in orange, and the potential linker region in green. Catalytic residues are highlighted in yellow.

Chapter 5: Summary

- CAZymes can be found existing as single domain enzymes, multidomain enzymes (with one catalytic domain), multicatalytic enzymes, or cellulosomes
- Multicatalytic enzymes provide an interesting new opportunity for more efficient biorefinery enzymes
- A large number of organisms produce multicatalytic enzymes, but some produce more than others
- Multicatalytic enzymes can present challenges in the production process

Chapter 6: Designer Enzymes

Enzymes in nature have evolved to help the producing organisms thrive in a large variety of situations (415). However, enzymes that are produced in nature to meet the needs of various organisms may not be appropriate to meet the needs of industrial processes (415). Additionally, as seen in the previous chapter, natural enzymes can be difficult to produce artificially, or in large quantities. While the field of protein engineering is still in its infancy, it has seen significant advances in recent years, and the production of unnatural enzymes for industrial purposes is closer than ever to being a reality (416).

6.1 A Brief History

Since the early 1980's, enzyme modifications have been a tool widely used by researchers to study the role of individual amino acids in enzyme function, via substitution with other amino acid residues (416). Shortly after that, the first enzyme engineered to improve specific properties (in this case, resistance to oxidation by hydrogen peroxide) was reported (417). While this modification only required a single amino acid substitution (and resulted in a corresponding decrease in enzyme activity), it opened the door to the field of rational enzyme design (417). An early focus of enzyme design was based on evolution and mutation – large libraries of mutant genes were established, expressed, exposed to evolutionary pressures and tested to obtain functional variants of the enzyme of interest that displayed modified properties (308,418). This Nobel Prize-winning work led to the production of enzymes capable of withstanding organic solvents, enzymes with increased thermostability, activity towards novel substrates, and more desirable industrial properties (308). A key takeaway from this work was that often, beneficial mutations were found in unexpected areas of the gene of interest, suggesting an overall incomplete picture of how amino acid sequence and protein folding impart different properties on enzymes (308,419).

While this directed evolutionary approach produced many important results, the method has a significant downside in that an overwhelming number of enzyme variants can be produced, and screening can be a massive time investment (308). Advances in knowledge surrounding enzyme function, as well as bioinformatic tools which can allow for the modeling of substrate interactions, as well as modeling of the enzyme behavior in solution, have allowed for the targeting of specific regions of genes for mutation, greatly lowering the number of variants that need to be screened (308). Even *de novo* enzyme design has proved possible (the Rosetta software, described earlier, is an excellent tool for this), although these

computationally designed enzymes are still often slow acting (308,420,421). With the knowledge gained over four decades of protein engineering, and advances in bioinformatics and protein modeling, the field appears to be heading towards rational design of enzymes for engineering favorable properties (421). By moving much of the work to *in silico* methods, enzyme engineering can be performed quicker, and with much less laboratory screening to obtain the same results (421).

6.2 Potential for Engineering of CAZymes

Despite their industrial importance, literature on CAZyme engineering is scarce, perhaps because of the abundance of CAZymes from natural sources. The few examples that exist in the literature are mostly done through rational design rather than directed evolution and show both successes and failures regarding increased enzyme activity, thermostability, and environmental tolerance (422-428). It is important to remember that despite the relative lack of enzyme engineering in the CAZyme field, the enzymes evolved to the needs of the producing organism likely do not have all desired properties sought for industrial processes, so engineering may become a necessity in the near future.

6.3 Split Inteins for Producing and Designing Enzymes

One of the earlier base hypotheses of molecular biology, the one gene one enzyme (later one gene one polypeptide) hypothesis states that one gene is transcribed into RNA, translated by ribosomes, and becomes one complete protein (minus any posttranslational modifications such as glycosylation) (429). While certainly true in many cases, there are a large number of exceptions, such as instances of alternative splicing, where this does not hold true (430).

A more recently discovered and less investigated exception to the one gene, one polypeptide hypothesis is the presence of “split inteins” within the genome of an organism (431-433). Regular inteins are polypeptide sequences contained within a protein which have the ability to excise themselves from the protein they are contained in, and then rejoin the two remaining parts of the protein through a normal peptide bond (Figure 6.1) (434,435). While the existence of such motifs is fascinating, split inteins perform much the same function, but are split across two separate genes (431). When these genes are transcribed and translated, the two split intein segments are capable of binding each other in solution, excising themselves from their polypeptide chains, and simultaneously joining those chains together through a new peptide bond (431).

Although the identification of split intein pairs within a genome can be challenging, multiple such pairs have been found in nature (432,436,437). Interestingly, the first split intein mentioned in the scientific literature is not naturally occurring, but was artificially split by the researchers who reported it (438). Shortly thereafter, the first

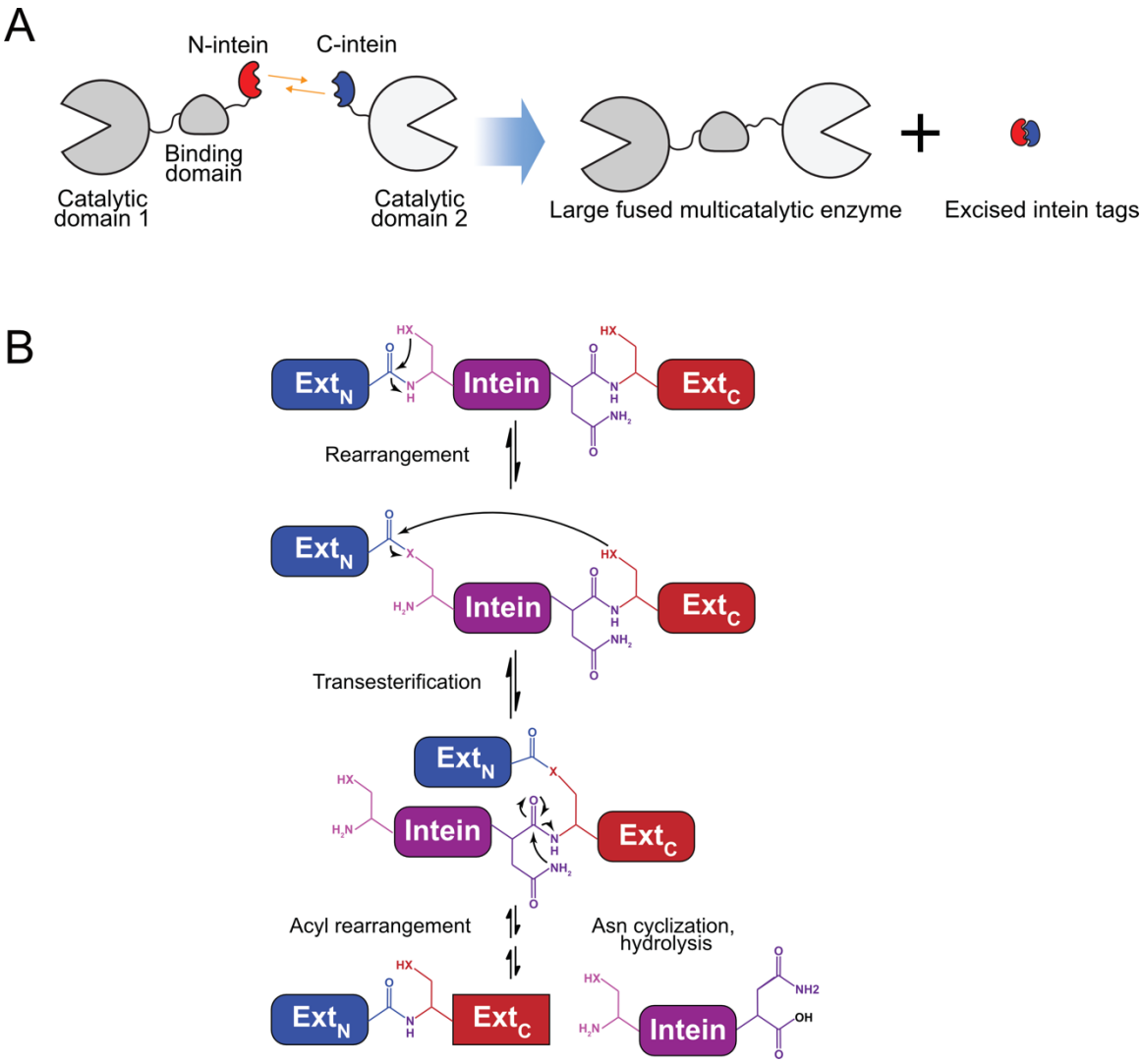

Figure 6.1 : Diagram of split intein mechanism. Panel **A** shows an overall diagram of the use of split inteins, in which an intein pair comes together and creates a fused protein and the excised tags. Panel **B** shows the detailed mechanism of the fusion of the two polypeptides and the intein excision.

naturally split intein was discovered (439). While the rate of reaction for these initial split inteins was quite slow (with a $t_{1/2}$ of between 6 minutes and 722 minutes), split inteins have since been discovered with $t_{1/2}$ values as little as 16 seconds (436,440). While many of the initially-discovered split inteins are cross reactive with one another, more recent work has identified many fast-acting split inteins that are specific to their partners, and not cross reactive with other sets of split inteins (436).

There are many examples in the scientific literature of split inteins being used for various different purposes. For example, split inteins have been used to label proteins with isotopes for NMR, in some cases ensuring only one domain of a multidomain protein was labelled (441,442). They have been used for allowing double plasmid selection from one antibiotic marker, the tagless purification of recombinantly-expressed proteins, and the selective killing of antibiotic-resistant

bacteria within mixed bacterial populations, among many other applications (443-445). Methods have even been developed to allow for selective triggering of the split intein reaction, so that combination of protein fragments does not occur spontaneously, but only on cue (446).

6.3.1 Applications of Split Inteins within the CAZyme Field

When it comes to multicatalytic CAZymes, designing desired functions can be done in a rational manner using split inteins. To date, however, there is no information in the literature about attempts to use split inteins with CAZymes. In **Paper IV**, split inteins were used to construct multicatalytic CAZymes, including combinations that do not exist in nature. This has allowed for the difficulty in recombinant production of large multicatalytic enzymes discussed earlier in chapter six to be bypassed, by producing each individual domain separately. The combination of different CAZyme domains proved possible using multiple different split intein pairings, as well as multiple CAZyme domain families, and tested domains remained active after combining.

The work described in **Paper IV** represents a starting point in the construction of customized multicatalytic CAZymes. Although not all split intein pairs tested in **Paper IV** were successful in creating combinations, the majority did appear to work. Several different proteins were created using split inteins, and multiple pairs were tested successfully. There still remain several pairs to test, and more work can be done to find conditions in which the seemingly non-functional (in these experiments) pairs will work correctly. It also remains to be seen how large the final proteins can be, but theoretically there is no limit to the final size. More investigation is necessary to determine the efficiency of conducting multiple split intein reactions at once, and to determine the effect of linker regions on the final constructed proteins.

Previously, if a researcher were interested in making a designed multicatalytic CAZyme, it would require manipulation at a genetic level, using molecular biology techniques to combine the domains into a single reading frame, and then expressing the chimeric protein. As mentioned, this would also mean a very likely risk of no expression of the target protein due to the protein size limitations discussed previously. By utilizing split inteins, this process is extremely simplified – as an example, producing every possible linear combination of six different domains would require producing, and then expressing, 720 different genes. To do so using different split intein pairs would only require 36. In this way, split inteins can greatly simplify the process of creating and testing novel multicatalytic enzymes for industrial processes, and potentially uncovering previously unknown intramolecular synergy between catalytic domains.

Chapter 6: Summary

- Enzyme modifications have been pursued since at least the 1980's
- Engineering of CAZymes presents an interesting possibility to improve many of their properties
- Split inteins are protein segments that are capable of joining two polypeptides together
- Split inteins have proven useful for the assembly of multicatalytic CAZymes

Chapter 7: Conclusions

The work presented in this thesis focused primarily on the investigation and characterization of multicatalytic enzymes containing a CE15 catalytic domain. To this end, a thorough investigation of the enzyme *Ck*Xyn10C-GE15A from *C. kristjanssonii* was conducted. The xylanase and glucuronoyl esterase domain were both biochemically characterized, with the xylanase showing by far its highest activity on wheat arabinoxylan, and the glucuronoyl esterase domain showing limited activity on tested model substrates (**Paper I**). Oddly, investigation into potential synergy between the two enzymatic domains did not reveal any. This result was unexpected, as it was expected that there would be a measurable advantage to the combination of the two domains, where the proximity of these domains would provide synergistic action on the targeted biomass. It is still possible, however, that synergy would be seen if tested on a biomass that *C. kristjanssonii* could be realistically expected to encounter in its natural environment. As the CE15 family of enzymes has only recently been of more focus in the scientific literature, it is also possible that this enzyme has unexpected substrate specificities that have yet to be determined.

As the CE15 domain of *Ck*Xyn10C-GE15A was the first thermostable CE15 domain studied (**Paper I**), a structural investigation of the domain was conducted (**Paper II**). While the study did not reveal any direct explanations of the thermostability of the enzyme, it did reveal that the enzyme does not completely align with previously characterized bacterial or fungal CE15 domains, and appears to be a hybrid of both, showing only one of the three insert regions that are otherwise present in bacterial CE15 domains when contrasted with their fungal counterparts. A structural investigation of this insert suggests it may be involved in binding to or coordinating the lignin component of the LCC. Additionally, substitution of a previously identified substrate stabilizing tryptophan with a glycine residue was observed in the active site of *Ck*CE15A, suggesting a preference for larger substrates when compared with other CE15 domains.

A second multicatalytic enzyme containing a CE15 domain, *Be*CE15A-Rex8A, was also biochemically characterized (**Paper III**). Again, the observed lack of synergy between the two catalytic domains was unexpected. In this case, while the CE15 domain did have activity on model substrates, it was significantly less active than what is typically seen for enzymes in this family. A normally conserved arginine residue was observed to be substituted with a phenylalanine, which was thought to be a contributing factor to the low activity levels. However, replacement of this phenylalanine with an arginine did not restore activity as was predicted, and instead

removed it entirely. This again suggests that there is more about the function of these domains that is unknown, and hints that there may be a different activity present within the family that has yet to be fully uncovered.

Two more main goals of this thesis work were the attempt to produce novel multicatalytic enzymes not found in nature, and to successfully produce large multicatalytic enzymes recombinantly. These goals were addressed in **Paper IV**. Although production of novel multicatalytic enzymes was initially planned to be done with a DNA library, the project was adapted to utilize split inteins instead, in order to address the challenges posed by recombinant expression of multicatalytic enzymes. This production did show some difficulties, as not all split intein combinations proved to work successfully, and not all CAZyme domains were able to be expressed with split intein tags attached. Importantly, the work presented in **Paper IV** represents a proof of concept for this approach, and a solid starting point for the work to be expanded upon in the future.

Overall, the majority of goals set out at the beginning of this work were completed. Large, multicatalytic enzymes were investigated, and the roles of CE15 enzyme domains within multicatalytic enzymes were probed, providing insights into this fascinating group of multicatalytic enzymes. While a custom multicatalytic enzyme library was not set up, the techniques and tools to do so were investigated and prepared, and several domains with split intein tags were prepared as a start to the library.

Chapter 8: Future Perspectives

Despite the advances presented in this thesis, there is still significant work to be done in the investigation of multicatalytic CAZymes. The overall lack of synergy seen in any of the enzymes investigated (**Paper I**, **Paper II**, **Paper III**) is a very curious result, and certainly warrants further study. From an evolutionary perspective, it is difficult to conceive of a purpose for multicatalytic architecture that does not provide some benefit to the producing organism. If the enzyme domains are not working to benefit each other in some way, the cell is expending considerable effort in producing the whole polypeptide for no apparent benefit. It also loses the ability to regulate the expression of the different enzyme domains separately, so it may even be detrimental. It therefore stands to reason that there is, in fact, some purpose behind this multicatalytic architecture, and it simply hasn't been discovered yet.

Perhaps an explanation for the observed lack of synergy, and also an avenue for further investigation, is the presence of CE15 domains in both studied enzymes. Compared to many other CAZyme families, CE15 enzymes have been studied relatively little. While model substrates for the enzymes exist and can be used to characterize them to some extent, natural substrates are incredibly difficult to isolate and even more so characterize reactions on. Furthermore, the model substrates do not cover the substrate diversity that is assumed to exist in nature. It is assumed that the enzymes are significantly more efficient on bulkier components within LCCs (which is supported by their universally increased activity on the larger model substrates), but what those substrates are exactly, how the activity is impacted by different lignin components, and how efficient the CE15 enzymes are in their natural environment all remain to be elucidated.

Some of the work presented here on CE15 domains would seem to suggest that the model substrates do not always accurately reflect the role of the domains *in vivo*. Besides the lack of synergy, looking at *Be*CE15A-Rex8A from **Paper III**, the enzyme domain combination does not make sense if the CE15 domain is acting on bonds within lignin. Previously characterized Rex domains have been shown to not be secreted by the cell, and act on xylooligosaccharides imported into the cell. The CE15 domains, on the other hand, are all secreted extracellularly by their producing organism, as lignin is too large to be brought into the cell. This makes the domain combination seen in *Be*CE15A-Rex8A somewhat of a mystery, as these domains would not be expected to function in the same physical location. This would suggest that further study on the CE15, the Rex, or both is needed to elucidate the natural substrate of these enzymes, as well as what environment they function in.

A large part of the work in this thesis was performed in order to build towards the construction of a library of both natural and artificially designed multicatalytic enzymes. While no such library was completed, the techniques to build one were established, and many domains of interest for inclusion in the library were identified. The use of split inteins represents an exciting prospect for the construction of this library, and the rapid construction of novel multicatalytic enzymes. Ideally this method could be used to produce difficult enzymes, such as *Ck*Xyn10C-GE15A from **Paper I**, which are unable to be produced as full length enzymes through other means. In this specific case, it would allow a more complete study of the potential synergy between the two catalytic domains.

While the initial functionality of the split intein method for multicatalytic enzyme construction was demonstrated in **Paper IV**, much work is left to be done in order to perfect the technique and construct desired libraries. While many split intein pairs were used, not all proved to be successful, and there are still more available to be tested. The pairs were also not tested for cross-reactivity. A lack of cross-reactivity has previously been reported in literature, but this has not been tested in a multicatalytic enzyme context, and would still need to be verified before library construction begins in earnest (436,437).

The chaining together of split intein reactions is another untested component of the library construction process. Currently, it is assumed that the presence of different split intein tags on the same protein will not interfere with the fusion process. There is, however, little experimental verification of this assumption. While there are several examples in the literature of the use of dual split intein pairs in protein construction (and in **Paper IV**, components with multiple intein tags were successfully used in protein construction), but data on the use of three or more intein pairs remains elusive.

A factor that was not explored in **Paper IV**, but remains extremely important for construction of custom multicatalytic enzymes, is the analysis of linker regions between the enzyme domains. These linker regions can have a significant impact on the distance between catalytic domains, their orientation, and potentially how well they work together. Linker regions for one multicatalytic enzyme were briefly discussed in **Paper II**, and even in that one enzyme it was clear that the linkers can have radically different lengths and compositions. The effects of these linkers on their associated catalytic domains are little studied, and need to be examined in detail for the multicatalytic enzyme library to be effective. Ideally, different linkers will be able to be produced with split intein tags, and then can be swapped into customized enzymes as needed. This split intein method could also be used to investigate the presence of multiple CBM domains within CAZymes. This is a commonly occurring architecture, where multiple CBM domains of the same family

are present attached to a single CAZyme domain (seen in **Paper I** and **Paper II**). The split intein construction method would allow for the simple inclusion or exclusion of one or more CBM domains, and even the inclusion of CBM domains not natively found in the enzyme studied, in order to better elucidate their functions.

Perhaps the most intriguing possibility for this split intein technique would be its application to the construction of unnatural multicatalytic enzymes. A clear example of this would be the inclusion of lytic polysaccharide monooxygenases (LPMOs) within a multicatalytic architecture. These relatively newly discovered enzymes have shown to be incredibly efficient at preparing crystalline cellulose for degradation, and have significant industrial relevance, however, their inclusion in natural multicatalytic enzymes is rare (170). While there is still much debate pertaining to LPMOs in the literature, their inclusion in a multicatalytic architecture would allow for the possibility to construct highly efficient multicatalytic cellulases, and possibly give more insight into the mechanism and function of LPMOs as they relate to other catalytic domains.

Utilizing split inteins with the overall goal of producing unnatural multicatalytic enzymes for industrial use should allow for significantly increased efficiency of industrial processes. Enzymes currently used are those that have evolved to benefit their producing organisms first and foremost, and any benefit to industrial processes is merely a side effect of their evolutionary purpose. The ability to design enzymes specifically tailored for industrial use, and the ability to combine and (hopefully) instill synergy between enzymes from unrelated species is expected to be of great benefit to any lignocellulosic biorefinery.

Acknowledgements

Completing a PhD degree can be difficult at the best of times, and can often seem like an insurmountable challenge. None of us who go through it would be able to reach the end without significant help from those around us, and I would be remiss if I did not take the opportunity here to thank (in no particular order) some of the many people who have supported me throughout this process.

I certainly wouldn't be finishing my PhD now if, nearly five years ago, Johan and Lisbeth hadn't decided to take a chance on accepting me into the lab, and bringing me across an ocean to start work on this project. I can definitely appreciate that hiring someone that they had never met in person, and who had never before set foot in Sweden, could not have been an easy choice to make. I am extremely grateful for the choice that they made then, as it has afforded me the opportunity to experience so many new things, to gain more scientific knowledge and experience, and to develop more as a person. I should also take this opportunity to thank the both of them for their help in editing this thesis, the time they spent on it and their suggestions to improve it were always appreciated.

Throughout my PhD, I had the opportunity to visit Copenhagen University several times. I always felt extremely welcomed by everyone in the lab there, and greatly appreciate the support from Leila, Rasmus, Jens-Christian, Tobias, and Kim that I received while I was there (and while on synchrotron trips with them as well). Special thanks to Leila for hosting me in her lab, and for the many extremely helpful conversations and teaching sessions that helped guide me through much of the structural work in this thesis.

Thank you to my co-supervisor Lisbeth and to my examiner Pernilla for fulfilling the duties those roles entail impeccably, and with enthusiasm for the development of us early-career researchers. I appreciate the both of you pushing me to produce my best possible work – this thesis, my previous publications, and everything I've created during my PhD are definitely better for it.

I would also like to thank the members of my defence committee, Paul, Mirjam, Kristian, Yvonne, and my opponent Maija, for taking the time to read and (by the time anyone reads this) approve this thesis, and for what I'm sure will be excellent and challenging questions at the defence itself.

A phenomenal amount of work happens behind the scenes at any university to make the research possible, and Chalmers is no exception to this. A huge thank you to

Anna, Anne-Lise, Gunilla, Erica, Barbro, Jenny for their work on the administration side of the department, and to Julia, Julia, Thenia, Vijay, and Pun for making sure the lab ran smoothly (or as smoothly as possible), orders were filled, and that we faced as little hinderance as possible while conducting research.

Johan. The supervisor of the year. This project wouldn't have happened without you, and I can't imagine any other supervisor doing a better job than you have. I appreciate you always being there to support me and help me deal with the stresses of the PhD, even though I often took far too long to give indications that the stress may be getting to me. You've always been more than willing to listen to my ideas for the project and have supported me in exploring avenues that diverge from the initial plans. Under your supervision I've had every opportunity to grow as a researcher, to manage my own project, and to explore and expand my horizons. Thank you.

Scott, when I first applied to work in IndBio, I had no idea you were already here representing MCB. I'm sure you were equally surprised when Johan told you I was one of the applicants for the position – you had probably assumed you'd never run into that undergraduate project student from Stephen's lab halfway across the world. Thank you for helping me transition from Canadian life to Swedish life, for helping me get started in the lab here (and back in Canada as well), for being available for questions and advice whenever you were needed, and for the work you've contributed co-authoring two of the papers in this thesis.

Of course, Scott was not the only co-author on my publications, and I appreciate the contributions made by all of the excellent scientists who have contributed to the work presented here: Johan, Scott, Haley, Yusuf, Jens-Christian, Adeline, Nicole, Leila, and Cathi. I thoroughly enjoyed working with you all, and only regret that I was only able to actually meet some of you in person.

I would also like to thank the remainder of the Maple team for all of their assistance both in and out of the lab during my time in the group, and to also thank all the other IndBio members (far too numerous to name here individually) who have been incredibly helpful these past four and a half years.

Thus far, I've only thanked people for their incredible contributions to my professional life - work in the lab, work at Chalmers, collaborations, etc. But a PhD wouldn't be achievable without some sort of support away from all of the hard work that goes into completing a project. To make the transition from work-life thanks to personal-life thanks, I'd like to start by thanking my office mates during my time at Chalmers. From my first office, thanks to Emma, Vero, Elena, and Cecilia, for making the office a fun place to get away from the lab for a few quick breaks during the day. From my second office, thanks to Jocke for our many hockey and music

related conversations, and thanks to Tom for, well, putting up with me (I have no doubt I was not the easiest office mate to get assigned to upon your arrival).

Numerous people from have helped arrange and participate in different sports to help blow off steam during wellness hours or on the weekends. Thanks to Jocke, David, Emma, Rike, Andrea, Tyler, Jeroen, Fabio, and others for being the more regular participants in floorball/basketball/squash.

Evening movie nights, barbeques, trips to Liseberg, restaurants, vacations to the frozen north, and all the ridiculousness and insanity they brought would not have been possible without the Hey Google/Alphabet group. Thanks to Andrea, Elena, Gerard, Jae, Javi, Otavio, Rike, Thenia, Vero, Daniel, and Jeroen (yes, even Jeroen) for the levity and fun they brought every time we met.

Adventures into the Swedish wilderness, coast, cooking over campfires, canoeing, kayaking (sometimes upside down), escaping from locked rooms, and eating tons of sliders and tacos provided great fun, relaxation, and distraction from the stresses of work during the beautiful Swedish summers, none of which would have been possible without Amanda, Cathi, Jonas, and Rike – thank you all for the amazing times we've spent together, and hopefully they won't end any time soon. After all, we still have our American road trip to go on at some point!

At this point, I think it's fair to say that I would probably not be finishing this PhD without the support I've gotten from Rike. She's always been available to help read over whatever I've written, whether it be work for courses, paper drafts, or this thesis. She has supported me unconditionally as I dealt with the stress of meeting all required deadlines, of spending late nights in the lab on failing experiments, and often feeling that I would never manage to get through all the work necessary to get to the point I'm at now. I am sincerely extremely grateful for all that she has done (and, I'm sure, will continue to do) to assist and support me in everything.

I would also like to thank my family in Canada for everything they've done to support me in this endeavour. A PhD is a large undertaking, but to even get to the point of starting one encompasses years of work which I would not have been able to do without them. Leaving them an ocean away has not been easy, for either me or them, and has only been made more difficult with travel restrictions these past two years. I've missed many birthdays, Thanksgivings, and Christmases in being here, which I know was as difficult for them as it has been for me. But I know I still have their unconditional support and their unwavering belief that I can succeed here, both in this PhD and in whatever comes next, and for that I am eternally grateful.

And finally, thank you Cathi.

References

1. Amoah, J., Kahar, P., et al. (2019) Bioenergy and biorefinery: feedstock, biotechnological conversion, and products. *Biotechnol. J.* **14**, 1800494
2. Duwe, A., Tippkötter, N., et al. (2017) Lignocellulose-biorefinery: ethanol-focused. in *Biorefineries*, Springer. pp 177-215
3. Mbow, H.-O. P., Reisinger, A., et al. (2017) Special Report on climate change, desertification, land degradation, sustainable land management, food security, and greenhouse gas fluxes in terrestrial ecosystems (SR2). *Ginevra, IPCC*
4. de Coninck, H., Revi, A., et al. (2018) Strengthening and implementing the global response. in *Global warming of 1.5° C: Summary for policy makers*, IPCC-The Intergovernmental Panel on Climate Change. pp 313-443
5. Shen, M., Huang, W., et al. (2020) (Micro) plastic crisis: Un-ignorable contribution to global greenhouse gas emissions and climate change. *Journal of Cleaner Production* **254**, 120138
6. Geyer, R., Jambeck, J. R., et al. (2017) Production, use, and fate of all plastics ever made. *Science advances* **3**, e1700782
7. Hankermeyer, C. R., and Tjeerdema, R. S. (1999) Polyhydroxybutyrate: plastic made and degraded by microorganisms. *Reviews of environmental contamination and toxicology*, 1-24
8. Robinson, P. K. (2015) Enzymes: principles and biotechnological applications. *Essays Biochem.* **59**, 1
9. de Pina Mariz, B., Carvalho, S., et al. (2021) Artificial enzymes bringing together computational design and directed evolution. *Org. Biomol. Chem.* **19**, 1915-1925
10. Schomburg, I., Chang, A., et al. (2012) BRENDA in 2013: integrated reactions, kinetic data, enzyme function data, improved disease classification: new options and contents in BRENDA. *Nucleic Acids Res.* **41**, D764-D772
11. Commission, E. Innovating for Sustainable Growth: A Bioeconomy for Europe: Communication from the Commission to the European Parliament, the Council and the European Economic and Social Committee and the Committee of the Regions. Accessed: July 19, 2021. https://ec.europa.eu/research/bioeconomy/pdf/bioeconomycommunicationstrategy_b5_brochure_web.pdf
12. Stegmann, P., Londo, M., et al. (2020) The circular bioeconomy: Its elements and role in European bioeconomy clusters. *Resources, Conservation & Recycling: X* **6**, 100029
13. Strategy, U. B. (2018) A sustainable bioeconomy for Europe: strengthening the connection between economy, society and the environment. *https://ec.europa.eu/research/bioeconomy/pdf/ec_bioeconomy_strategy_2018.pdf*
14. El-Chichakli, B., von Braun, J., et al. (2016) Policy: Five cornerstones of a global bioeconomy. *Nature News* **535**, 221
15. Raney, T. (2009) The state of food and agriculture: livestock in the balance. *Food and Agriculture Organization of the United Nations, Rome, Italy*
16. Philp, J. (2018) The bioeconomy, the challenge of the century for policy makers. *N. Biotechnol.* **40**, 11-19
17. Zetterholm, J., Bryngemark, E., et al. (2020) Economic evaluation of large-scale biorefinery deployment: A framework integrating dynamic biomass market and techno-economic models. *Sustainability* **12**, 7126
18. Sedjo, R. A. (1997) The economics of forest-based biomass supply. *Energy Policy* **25**, 559-566
19. Berntsson, T., Sandén, B. A., et al. (2012) What is a biorefinery? , 16-25
20. Demirbas, A., and Demirbas, M. F. (2010) Biorefineries. in *Algae energy*, Springer. pp 159-181
21. Cherubini, F. (2010) The biorefinery concept: using biomass instead of oil for producing energy and chemicals. *Energy Convers. Manage.* **51**, 1412-1421
22. Takkellapati, S., Li, T., et al. (2018) An overview of biorefinery-derived platform chemicals from a cellulose and hemicellulose biorefinery. *Clean technologies and environmental policy* **20**, 1615-1630

23. Gavrilescu, M. (2014) Biorefinery systems: an overview. *Bioenergy research: advances and applications*, 219-241
24. Ingelheim, B. 1885 – 1948: Innovative beginnings. Accessed: October 7, 2021. https://www.boehringer-ingelheim.com/history/history-milestone/1885-1948
25. Chalkley, A. P. (1917) *Diesel engines for land and marine work*, Constable, Limited
26. Bergius, F. (1937) Conversion of wood to carbohydrates. *Ind. Eng. Chem.* **29**, 247-253
27. Devaney, L. A., and Henchion, M. (2017) If opportunity doesn't knock, build a door: Reflecting on a bioeconomy policy agenda for Ireland. *The Economic and Social Review* **48**, 207-229
28. Fielding, M., and Aung, M. T. (2018) *Bioeconomy in Thailand: a case study*, Stockholm Environment Institute.
29. BioBase4SME. Bioeconomy Factsheet United Kingdom. Accessed: July 20, 2021. https://www.nweurope.eu/media/4664/180369_biobase4sme_2luik_uk_v4_lr.pdf
30. Wenger, J., and Stern, T. (2019) Reflection on the research on and implementation of biorefinery systems–a systematic literature review with a focus on feedstock. *Biofuels, Bioproducts and Biorefining* **13**, 1347-1364
31. National Academies of Sciences, E., and Medicine. (2020) *Safeguarding the Bioeconomy*, National Academies Press
32. Lv, Y., Jiang, Y., et al. (2021) Genetic manipulation of non-solvent-producing microbial species for effective butanol production. *Biofuels, Bioproducts and Biorefining* **15**, 119-130
33. Fayyaz, M., Chew, K. W., et al. (2020) Genetic engineering of microalgae for enhanced biorefinery capabilities. *Biotechnol. Adv.* **43**, 107554
34. Sarrouh, B., Santos, T. M., et al. (2012) Up-to-date insight on industrial enzymes applications and global market. *J Bioprocess Biotech* **4**, 002
35. Filiatrault-Chastel, C., Heiss-Blanquet, S., et al. (2021) From fungal secretomes to enzymes cocktails: The path forward to bioeconomy. *Biotechnol. Adv.*, 107833
36. Chapman, J., Ismail, A. E., et al. (2018) Industrial applications of enzymes: Recent advances, techniques, and outlooks. *Catalysts* **8**, 238
37. Lv, L., Dai, L., et al. (2021) Progress in enzymatic biodiesel production and commercialization. *Processes* **9**, 355
38. Adsul, M., Sandhu, S. K., et al. (2020) Designing a cellulolytic enzyme cocktail for the efficient and economical conversion of lignocellulosic biomass to biofuels. *Enzyme and microbial technology* **133**, 109442
39. Dumorné, K., Córdova, D. C., et al. (2017) Extremozymes: a potential source for industrial applications.
40. Raddadi, N., Cherif, A., et al. (2015) Biotechnological applications of extremophiles, extremozymes and extremolytes. *Applied microbiology and biotechnology* **99**, 7907-7913
41. Toplak, A., Wu, B., et al. (2013) Proteolysin, a novel highly thermostable and cosolvent-compatible protease from the thermophilic bacterium *Coprothermobacter proteolyticus*. *Applied and environmental microbiology* **79**, 5625-5632
42. Egorova, K., and Antranikian, G. (2005) Industrial relevance of thermophilic Archaea. *Curr. Opin. Microbiol.* **8**, 649-655
43. Van Den Burg, B. (2003) Extremophiles as a source for novel enzymes. *Curr. Opin. Microbiol.* **6**, 213-218
44. Littlechild, J., Novak, H., et al. (2013) Mechanisms of thermal stability adopted by thermophilic proteins and their use in white biotechnology. in *Thermophilic Microbes in Environmental and Industrial Biotechnology*, Springer. pp 481-507
45. Song, L., Tsang, A., et al. (2015) Engineering a thermostable fungal GH10 xylanase, importance of N-terminal amino acids. *Biotechnology and bioengineering* **112**, 1081-1091
46. Han, H., Ling, Z., et al. (2019) Improvements of thermophilic enzymes: From genetic modifications to applications. *Bioresour. Technol.* **279**, 350-361
47. Li, W., Zhou, X., et al. (2005) Structural features of thermozymes. *Biotechnol. Adv.* **23**, 271-281
48. Li, H., Kankaanpää, A., et al. (2013) Thermostabilization of extremophilic *Dictyoglomus thermophilum* GH11 xylanase by an N-terminal disulfide bridge and the effect of ionic liquid [emim] OAc on the enzymatic performance. *Enzyme and microbial technology* **53**, 414-419
49. Liu, T., Wang, Y., et al. (2016) Enhancing protein stability with extended disulfide bonds. *Proceedings of the National Academy of Sciences* **113**, 5910-5915
50. Yin, X., Hu, D., et al. (2015) Contribution of disulfide bridges to the thermostability of a type A feruloyl esterase from *Aspergillus usamii*. *PLoS One* **10**, e0126864

51. Krska, D., Mazurkewich, S., et al. (2021) Structural and Functional Analysis of a Multimodular Hyperthermostable Xylanase-Glucuronoyl Esterase from *Caldicellulosiruptor kristjansonii*. *Biochemistry*
52. Rader, A. (2009) Thermostability in rubredoxin and its relationship to mechanical rigidity. *Phys. Biol.* **7**, 016002
53. Rathi, P. C., Jaeger, K.-E., et al. (2015) Structural rigidity and protein thermostability in variants of lipase A from *Bacillus subtilis*. *PLoS One* **10**, e0130289
54. de Souza, A. R., de Araújo, G. C., et al. (2016) Engineering increased thermostability in the GH-10 endo-1, 4-β-xylanase from *Thermoascus aurantiacus* CBMAI 756. *Int. J. Biol. Macromol.* **93**, 20-26
55. Auerbach, G., Ostendorp, R., et al. (1998) Lactate dehydrogenase from the hyperthermophilic bacterium *Thermotoga maritima*: the crystal structure at 2.1 Å resolution reveals strategies for intrinsic protein stabilization. *Structure* **6**, 769-781
56. Natesh, R., Bhanumoorthy, P., et al. (1999) Crystal structure at 1.8 Å resolution and proposed amino acid sequence of a thermostable xylanase from *Thermoascus aurantiacus*. *J. Mol. Biol.* **288**, 999-1012
57. Elcock, A. H. (1998) The stability of salt bridges at high temperatures: implications for hyperthermophilic proteins. *J. Mol. Biol.* **284**, 489-502
58. Cowan, D. (1997) Thermophilic proteins: stability and function in aqueous and organic solvents. *Comparative Biochemistry and Physiology Part A: Physiology* **118**, 429-438
59. Daniel, R. M., Cowan, D. A., et al. (1982) A correlation between protein thermostability and resistance to proteolysis. *Biochemical Journal* **207**, 641-644
60. Viikari, L., Alapuranen, M., et al. (2007) Thermostable enzymes in lignocellulose hydrolysis. *Biofuels*, 121-145
61. Blumer-Schuette, S. E., Brown, S. D., et al. (2014) Thermophilic lignocellulose deconstruction. *FEMS Microbiol. Rev.* **38**, 393-448
62. Rahikainen, J. L., Moilanen, U., et al. (2013) Effect of temperature on lignin-derived inhibition studied with three structurally different cellobiohydrolases. *Bioresour. Technol.* **146**, 118-125
63. Jin, M., Gai, Y., et al. (2019) Properties and applications of extremozymes from deep-sea extremophilic microorganisms: A mini review. *Mar. Drugs* **17**, 656
64. Sharma, A., Kawarabayasi, Y., et al. (2012) Acidophilic bacteria and archaea: acid stable biocatalysts and their potential applications. *Extremophiles* **16**, 1-19
65. Jönsson, L. J., and Martín, C. (2016) Pretreatment of lignocellulose: formation of inhibitory by-products and strategies for minimizing their effects. *Bioresour. Technol.* **199**, 103-112
66. Maurelli, L., Giovane, A., et al. (2008) Evidence that the xylanase activity from *Sulfolobus solfataricus* Oα is encoded by the endoglucanase precursor gene (sso1354) and characterization of the associated cellulase activity. *Extremophiles* **12**, 689-700
67. Eckert, K., and Schneider, E. (2003) A thermoacidophilic endoglucanase (CelB) from *Alicyclobacillus acidocaldarius* displays high sequence similarity to arabinofuranosidases belonging to family 51 of glycoside hydrolases. *Eur. J. Biochem.* **270**, 3593-3602
68. Fujinami, S., and Fujisawa, M. (2010) Industrial applications of alkaliphiles and their enzymes–past, present and future. *Environ. Technol.* **31**, 845-856
69. Ben Hmad, I., and Gargouri, A. (2017) Neutral and alkaline cellulases: Production, engineering, and applications. *J. Basic Microbiol.* **57**, 653-658
70. Madern, D., Pfister, C., et al. (1995) Mutation at a single acidic amino acid enhances the halophilic behaviour of malate dehydrogenase from *Haloarcula marismortui* in physiological salts. *Eur. J. Biochem.* **230**, 1088-1095
71. Bhalla, A., Bansal, N., et al. (2013) Improved lignocellulose conversion to biofuels with thermophilic bacteria and thermostable enzymes. *Bioresour. Technol.* **128**, 751-759
72. Raddadi, N., Cherif, A., et al. (2013) Halo-alkalitolerant and thermostable cellulases with improved tolerance to ionic liquids and organic solvents from *Paenibacillus tarimensis* isolated from the Chott El Fejej, Sahara desert, Tunisia. *Bioresour. Technol.* **150**, 121-128
73. Ortega, G., Laín, A., et al. (2011) Halophilic enzyme activation induced by salts. *Sci. Rep.* **1**, 1-6
74. Suzuki, T., Nakayama, T., et al. (2001) Cold-active lipolytic activity of psychrotrophic *Acinetobacter* sp. strain no. 6. *Journal of bioscience and bioengineering* **92**, 144-148
75. Serour, E., and Antranikian, G. (2002) Novel thermoactive glucoamylases from the thermoacidophilic Archaea *Thermoplasma acidophilum*, *Picrophilus torridus* and *Picrophilus oshimae*. *Antonie Van Leeuwenhoek* **81**, 73-83

76. Robin, T., Reuveni, S., et al. (2018) Single-molecule theory of enzymatic inhibition. *Nature communications* **9**, 1-9
77. Berg, J. M., Tymoczko, J. L., et al. (2002) *Biochemistry: International Version*, Granite Hill Publishers
78. Montgomery, A. P., Xiao, K., et al. (2017) Computational glycobiology: Mechanistic studies of carbohydrate-active enzymes and implication for inhibitor design. *Adv. Protein Chem. Struct. Biol.* **109**, 25-76
79. Tramontina, R., Brenelli, L. B., et al. (2020) Enzymatic removal of inhibitory compounds from lignocellulosic hydrolysates for biomass to bioproducts applications. *World J. Microbiol. Biotechnol.* **36**, 1-11
80. Williams, C. L., Emerson, R. M., et al. (2017) Biomass compositional analysis for conversion to renewable fuels and chemicals. *Biomass volume estimation and valorization for energy*, 251-270
81. Alberts, B., Johnson, A., et al. (2002) The plant cell wall. in *Molecular Biology of the Cell. 4th edition*, Garland Science. pp
82. Zhang, B., Gao, Y., et al. (2021) The plant cell wall: biosynthesis, construction, and functions. *J. Integr. Plant Biol.* **63**, 251-272
83. Houston, K., Tucker, M. R., et al. (2016) The plant cell wall: a complex and dynamic structure as revealed by the responses of genes under stress conditions. *Frontiers in plant science* **7**, 984
84. Falcioni, R., Moriwaki, T., et al. (2020) Cell wall structure and composition is affected by light quality in tomato seedlings. *J. Photochem. Photobiol. B: Biol.* **203**, 111745
85. Fry, S. C. (2004) Primary cell wall metabolism: tracking the careers of wall polymers in living plant cells. *New Phytol.* **161**, 641-675
86. Carpita, N. C., and McCann, M. C. (2020) Redesigning plant cell walls for the biomass-based bioeconomy. *J. Biol. Chem.* **295**, 15144-15157
87. Iiyama, K., Lam, T. B.-T., et al. (1994) Covalent cross-links in the cell wall. *Plant Physiol.* **104**, 315
88. Cosgrove, D., and Jarvis, M. C. (2012) Comparative structure and biomechanics of plant primary and secondary cell walls. *Frontiers in plant science* **3**, 204
89. Yakubov, G. E., Bonilla, M. R., et al. (2016) Mapping nano-scale mechanical heterogeneity of primary plant cell walls. *J. Exp. Bot.* **67**, 2799-2816
90. Canilha, L., Chandel, A. K., et al. (2012) Bioconversion of sugarcane biomass into ethanol: an overview about composition, pretreatment methods, detoxification of hydrolysates, enzymatic saccharification, and ethanol fermentation. *J. Biomed. Biotechnol.* **2012**
91. Al-Ahmad, H. (2018) Biotechnology for bioenergy dedicated trees: meeting future energy demands. *Zeitschrift für Naturforschung C* **73**, 15-32
92. Keshwani, D. R., and Cheng, J. J. (2009) Switchgrass for bioethanol and other value-added applications: a review. *Bioresour. Technol.* **100**, 1515-1523
93. Wang, F., Zhang, D., et al. (2019) Characteristics of corn stover components pyrolysis at low temperature based on detergent fibers. *Frontiers in bioengineering and biotechnology* **7**, 188
94. Rocha-Meneses, L., Raud, M., et al. (2017) Second-generation bioethanol production: A review of strategies for waste valorisation. *Agron. Res.* **15**, 830-847
95. Cummings, J., and Stephen, A. (2007) Carbohydrate terminology and classification. *Eur. J. Clin. Nutr.* **61**, S5-S18
96. Miljkovic, M. (2009) *Carbohydrates: synthesis, mechanisms, and stereoelectronic effects*, Springer Science & Business Media
97. Guo, M. Q., Hu, X., et al. (2017) Polysaccharides: structure and solubility. *Solubility of polysaccharides*, 7-21
98. Dixon, H. (1982) Polysaccharide nomenclature. *Pure and Applied Chemistry* **8**, 1523-1526
99. Nelson, D. L., Lehninger, A. L., et al. (2008) *Lehninger principles of biochemistry*, Macmillan
100. Prestegard, J. H., Liu, J., et al. (2017) Oligosaccharides and polysaccharides.
101. Lapébie, P., Lombard, V., et al. (2019) Bacteroidetes use thousands of enzyme combinations to break down glycans. *Nature communications* **10**, 1-7
102. George, J., and Sabapathi, S. (2015) Cellulose nanocrystals: synthesis, functional properties, and applications. *Nanotechnology, science and applications* **8**, 45
103. Hsieh, Y. (2007) Chemical structure and properties of cotton. *Cotton: Science and technology*, 3-34
104. Scheller, H. V., and Ulvskov, P. (2010) Hemicelluloses. *Annu. Rev. Plant Biol.* **61**, 263-289
105. Gírio, F. M., Fonseca, C., et al. (2010) Hemicelluloses for fuel ethanol: a review. *Bioresour. Technol.* **101**, 4775-4800

106. Li, M., Pu, Y., et al. (2017) Study of traits and recalcitrance reduction of field-grown COMT down-regulated switchgrass. *Biotechnology for biofuels* **10**, 1-12
107. Normark, M., Winestrand, S., et al. (2014) Analysis, pretreatment and enzymatic saccharification of different fractions of Scots pine. *BMC Biotechnol.* **14**, 1-12
108. Brumm, P. J., De Maayer, P., et al. (2015) Genomic analysis of six new *Geobacillus* strains reveals highly conserved carbohydrate degradation architectures and strategies. *Front. Microbiol.* **6**, 430
109. Collins, T., Gerday, C., et al. (2005) Xylanases, xylanase families and extremophilic xylanases. *FEMS Microbiol. Rev.* **29**, 3-23
110. Beg, Q., Kapoor, M., et al. (2001) Microbial xylanases and their industrial applications: a review. *Applied microbiology and biotechnology* **56**, 326-338
111. Hatfield, R. D., Rancour, D. M., et al. (2017) Grass cell walls: a story of cross-linking. *Frontiers in plant science* **7**, 2056
112. Busse-Wicher, M., Gomes, T. C., et al. (2014) The pattern of xylan acetylation suggests xylan may interact with cellulose microfibrils as a twofold helical screw in the secondary plant cell wall of *Arabidopsis thaliana*. *The Plant Journal* **79**, 492-506
113. Busse-Wicher, M., Li, A., et al. (2016) Evolution of xylan substitution patterns in gymnosperms and angiosperms: implications for xylan interaction with cellulose. *Plant Physiol.* **171**, 2418-2431
114. Kiyohara, M., Sakaguchi, K., et al. (2005) Molecular cloning and characterization of a novel β-1, 3-xylanase possessing two putative carbohydrate-binding modules from a marine bacterium *Vibrio* sp. strain AX-4. *Biochemical Journal* **388**, 949-957
115. Dodd, D., and Cann, I. K. (2009) Enzymatic deconstruction of xylan for biofuel production. *Gcb Bioenergy* **1**, 2-17
116. Sharma, K., Khaire, K. C., et al. (2020) Acacia xylan as a substitute for commercially available xylan and its application in the production of xylooligosaccharides. *ACS omega* **5**, 13729-13738
117. Fu, G.-Q., Hu, Y.-J., et al. (2019) Isolation, purification, and potential applications of xylan. in *Production of Materials from Sustainable Biomass Resources*, Springer. pp 3-35
118. Pereira, C. S., Silveira, R. L., et al. (2017) Effects of xylan side-chain substitutions on xylan–cellulose interactions and implications for thermal pretreatment of cellulosic biomass. *Biomacromolecules* **18**, 1311-1321
119. Lyczakowski, J. J., Wicher, K. B., et al. (2017) Removal of glucuronic acid from xylan is a strategy to improve the conversion of plant biomass to sugars for bioenergy. *Biotechnology for biofuels* **10**, 1-11
120. Moreira, L. (2008) An overview of mannan structure and mannan-degrading enzyme systems. *Applied microbiology and biotechnology* **79**, 165-178
121. Yu, L., Lyczakowski, J. J., et al. (2018) The patterned structure of galactoglucomannan suggests it may bind to cellulose in seed mucilage. *Plant Physiol.* **178**, 1011-1026
122. Donev, E., Gandla, M. L., et al. (2018) Engineering non-cellulosic polysaccharides of wood for the biorefinery. *Frontiers in plant science* **9**, 1537
123. Yamabhai, M., Sak-Ubol, S., et al. (2016) Mannan biotechnology: from biofuels to health. *Crit. Rev. Biotechnol.* **36**, 32-42
124. Park, Y. B., and Cosgrove, D. J. (2015) Xyloglucan and its interactions with other components of the growing cell wall. *Plant and Cell Physiology* **56**, 180-194
125. Nishinari, K., Takemasa, M., et al. (2007) Storage plant polysaccharides: Xyloglucans, galactomannans, glucomannans. *Comprehensive glycoscience*, 613-652
126. Piqué, N., Gómez-Guillén, M. D. C., et al. (2018) Xyloglucan, a plant polymer with barrier protective properties over the mucous membranes: an overview. *Int. J. Mol. Sci.* **19**, 673
127. Du, B., Meenu, M., et al. (2019) A concise review on the molecular structure and function relationship of β-glucan. *Int. J. Mol. Sci.* **20**, 4032
128. Harholt, J., Suttangkakul, A., et al. (2010) Biosynthesis of pectin. *Plant Physiol.* **153**, 384-395
129. Mohnen, D. (2008) Pectin structure and biosynthesis. *Curr. Opin. Plant Biol.* **11**, 266-277
130. Janusz, G., Pawlik, A., et al. (2017) Lignin degradation: microorganisms, enzymes involved, genomes analysis and evolution. *FEMS Microbiol. Rev.* **41**, 941-962
131. Pollegioni, L., Tonin, F., et al. (2015) Lignin-degrading enzymes. *The FEBS journal* **282**, 1190-1213
132. Ragauskas, A. J., Beckham, G. T., et al. (2014) Lignin valorization: improving lignin processing in the biorefinery. *Science* **344**

133. Vanholme, R., De Meester, B., et al. (2019) Lignin biosynthesis and its integration into metabolism. *Curr. Opin. Biotechnol.* **56**, 230-239
134. Liu, Q., Luo, L., et al. (2018) Lignins: biosynthesis and biological functions in plants. *Int. J. Mol. Sci.* **19**, 335
135. Vanholme, R., Demedts, B., et al. (2010) Lignin biosynthesis and structure. *Plant Physiol.* **153**, 895-905
136. Xie, T., Liu, Z., et al. (2020) Structural basis for monolignol oxidation by a maize laccase. *Nature plants* **6**, 231-237
137. Barakat, A., Winter, H., et al. (2007) Studies of xylan interactions and cross-linking to synthetic lignins formed by bulk and end-wise polymerization: a model study of lignin carbohydrate complex formation. *Planta* **226**, 267-281
138. Sun, Z., Cheng, J., et al. (2020) Downstream Processing Strategies for Lignin-First Biorefinery. *ChemSusChem* **13**, 5134-5134
139. Li, C., Chen, C., et al. (2019) Recent advancement in lignin biorefinery: With special focus on enzymatic degradation and valorization. *Bioresour. Technol.* **291**, 121898
140. Huang, Y., Duan, Y., et al. (2018) Lignin-first biorefinery: a reusable catalyst for lignin depolymerization and application of lignin oil to jet fuel aromatics and polyurethane feedstock. *Sustainable Energy & Fuels* **2**, 637-647
141. Si, M., Yan, X., et al. (2018) In situ lignin bioconversion promotes complete carbohydrate conversion of rice straw by *Cupriavidus basilensis* B-8. *ACS Sustainable Chemistry & Engineering* **6**, 7969-7978
142. Shi, Y., Yan, X., et al. (2017) Directed bioconversion of Kraft lignin to polyhydroxyalkanoate by *Cupriavidus basilensis* B-8 without any pretreatment. *Process Biochem.* **52**, 238-242
143. Sawant, S. S., Salunke, B. K., et al. (2015) Degradation of corn stover by fungal cellulase cocktail for production of polyhydroxyalkanoates by moderate halophile Paracoccus sp. LL1. *Bioresour. Technol.* **194**, 247-255
144. Tarasov, D., Leitch, M., et al. (2018) Lignin–carbohydrate complexes: properties, applications, analyses, and methods of extraction: a review. *Biotechnology for biofuels* **11**, 1-28
145. Nishimura, H., Kamiya, A., et al. (2018) Direct evidence for α ether linkage between lignin and carbohydrates in wood cell walls. *Sci. Rep.* **8**, 1-11
146. Henriksson, G., Lawoko, M., et al. (2007) Lignin-carbohydrate network in wood and pulps: a determinant for reactivity.
147. Lawoko, M., Henriksson, G., et al. (2005) Structural differences between the lignin–carbohydrate complexes present in wood and in chemical pulps. *Biomacromolecules* **6**, 3467-3473
148. Balan, V., Sousa, L. d. C., et al. (2009) Enzymatic digestibility and pretreatment degradation products of AFEX-treated hardwoods (*Populus nigra*). *Biotechnology progress* **25**, 365-375
149. Marcia, M. d. O. (2009) Feruloylation in grasses: current and future perspectives. *Molecular plant* **2**, 861-872
150. Sapouna, I., and Lawoko, M. (2021) Deciphering lignin heterogeneity in ball milled softwood: unravelling the synergy between the supramolecular cell wall structure and molecular events. *Green Chem.* **23**, 3348-3364
151. Das, K., Singh, K., et al. (2011) Pyrolysis characteristics of forest residues obtained from different harvesting methods. *Appl. Eng. Agric.* **27**, 107-113
152. Tao, G., Lestander, T. A., et al. (2012) Biomass properties in association with plant species and assortments I: A synthesis based on literature data of energy properties. *Renewable and Sustainable Energy Reviews* **16**, 3481-3506
153. Hörhammer, H., Dou, C., et al. (2018) Removal of non-structural components from poplar whole-tree chips to enhance hydrolysis and fermentation performance. *Biotechnology for biofuels* **11**, 1-12
154. Sluiter, A., Ruiz, R., et al. (2005) Determination of extractives in biomass. *Laboratory analytical procedure (LAP)* **1617**, 1-9
155. Sluiter, J. B., Ruiz, R. O., et al. (2010) Compositional analysis of lignocellulosic feedstocks. 1. Review and description of methods. *Journal of agricultural and food chemistry* **58**, 9043-9053
156. Valette, N., Perrot, T., et al. (2017) Antifungal activities of wood extractives. *Fungal Biol. Rev.* **31**, 113-123
157. Devappa, R. K., Rakshit, S. K., et al. (2015) Forest biorefinery: Potential of poplar phytochemicals as value-added co-products. *Biotechnol. Adv.* **33**, 681-716

158. Kenney, K. L., Smith, W. A., et al. (2013) Understanding biomass feedstock variability. *Biofuels* **4**, 111-127
159. Reza, M. T., Emerson, R., et al. (2015) Ash reduction of corn stover by mild hydrothermal preprocessing. *Biomass Conversion and Biorefinery* **5**, 21-31
160. Stevens, C., and Brown, R. Thermochemical processing of biomass: conversion into fuels, chemicals and power. 2011. John Wiley & Sons
161. Jenkins, B., Baxter, L., et al. (1998) Combustion properties of biomass. *Fuel Process. Technol.* **54**, 17-46
162. He, Y., Fang, Z., et al. (2014) De-ashing treatment of corn stover improves the efficiencies of enzymatic hydrolysis and consequent ethanol fermentation. *Bioresour. Technol.* **169**, 552-558
163. Kemppainen, K., Inkinen, J., et al. (2012) Hot water extraction and steam explosion as pretreatments for ethanol production from spruce bark. *Bioresour. Technol.* **117**, 131-139
164. Frankó, B., Carlqvist, K., et al. (2018) Removal of water-soluble extractives improves the enzymatic digestibility of steam-pretreated softwood barks. *Applied biochemistry and biotechnology* **184**, 599-615
165. van Dijk, M., Mierke, F., et al. (2020) Nutrient-supplemented propagation of *Saccharomyces cerevisiae* improves its lignocellulose fermentation ability. *AMB Express* **10**, 1-10
166. Chen, X.-Y., and Kim, J.-Y. (2009) Callose synthesis in higher plants. *Plant signaling & behavior* **4**, 489-492
167. Torres, F. G., Troncoso, O. P., et al. (2019) Natural polysaccharide nanomaterials: an overview of their immunological properties. *Int. J. Mol. Sci.* **20**, 5092
168. Nasrollahzadeh, M., Sajjadi, M., et al. (2020) Starch, cellulose, pectin, gum, alginate, chitin and chitosan derived (nano) materials for sustainable water treatment: A review. *Carbohydr. Polym.*, 116986
169. Garron, M.-L., and Henrissat, B. (2019) The continuing expansion of CAZymes and their families. *Curr. Opin. Chem. Biol.* **53**, 82-87
170. Lombard, V., Golaconda Ramulu, H., et al. (2014) The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res.* **42**, D490-D495
171. Chettri, D., Verma, A. K., et al. (2020) Innovations in CAZyme gene diversity and its modification for biorefinery applications. *Biotechnology Reports*, e00525
172. Kuhad, R. C., Gupta, R., et al. (2011) Microbial cellulases and their industrial applications. *Enzyme research* **2011**
173. Polizeli, M., Rizzatti, A., et al. (2005) Xylanases from fungi: properties and industrial applications. *Applied microbiology and biotechnology* **67**, 577-591
174. Bourne, Y., and Henrissat, B. (2001) Glycoside hydrolases and glycosyltransferases: families and functional modules. *Curr. Opin. Struct. Biol.* **11**, 593-600
175. Michaelis, L., and Menten, M. L. (1913) Die kinetik der invertinwirkung. *Biochem. z* **49**, 352
176. McKee, L. S., Peña, M. J., et al. (2012) Introducing endo-xylanase activity into an exo-acting arabinofuranosidase that targets side chains. *Proceedings of the National Academy of Sciences* **109**, 6537-6542
177. Nguyen, S. T., Freund, H. L., et al. (2018) Function, distribution, and annotation of characterized cellulases, xylanases, and chitinases from CAZy. *Applied microbiology and biotechnology* **102**, 1629-1637
178. Juturu, V., and Wu, J. C. (2014) Microbial exo-xylanases: a mini review. *Applied biochemistry and biotechnology* **174**, 81-92
179. Kubata, B. K., Suzuki, T., et al. (1994) Purification and characterization of *Aeromonas caviae* ME-1 xylanase V, which produces exclusively xylobiose from xylan. *Applied and Environmental Microbiology* **60**, 531-535
180. Rakitin, A. L., Ermakova, A. Y., et al. (2015) Novel endoxylanases of the moderately thermophilic polysaccharide-degrading bacterium *Melioribacter roseus*. *Journal of microbiology and biotechnology* **25**, 1476-1484
181. Chakdar, H., Kumar, M., et al. (2016) Bacterial xylanases: biology to biotechnology. *3 Biotech* **6**, 1-15
182. Gupta, V., Garg, S., et al. (2015) Production of thermo-alkali-stable laccase and xylanase by co-culturing of *Bacillus* sp. and *B. halodurans* for biobleaching of kraft pulp and deinking of waste paper. *Bioprocess and biosystems engineering* **38**, 947-956
183. Jeffries, T. W., Patel, R. N., et al. (1992) Enzymatic solutions to enhance bonding, bleaching and contaminant removal. *MRS Online Proceedings Library (OPL)* **266**

184. Harris, A. D., and Ramalingam, C. (2010) Xylanases and its application in food industry: a review. *Journal of Experimental Sciences* **1**
185. Butt, M. S., Tahir-Nadeem, M., et al. (2008) Xylanases and their applications in baking industry. *Food Technology and Biotechnology* **46**, 22-31
186. Karlsson, E. N., Schmitz, E., et al. (2018) Endo-xylanases as tools for production of substituted xylooligosaccharides with prebiotic properties. *Applied microbiology and biotechnology* **102**, 9081-9088
187. Manisseri, C., and Gudipati, M. (2012) Prebiotic activity of purified xylobiose obtained from Ragi (*Eleusine coracana*, Indaf-15) Bran. *Indian J. Microbiol.* **52**, 251-257
188. Bajaj, P., and Mahajan, R. (2019) Cellulase and xylanase synergism in industrial biotechnology. *Applied microbiology and biotechnology* **103**, 8711-8724
189. Nakamichi, Y., Fouquet, T., et al. (2019) Mode of action of GH30-7 reducing-end xylose-releasing exoxylanase A (Xyn30A) from the filamentous fungus *Talaromyces cellulolyticus*. *Applied and environmental microbiology* **85**, e00552-00519
190. Sunagawa, N., Tajima, K., et al. (2012) Cellulose production by *Enterobacter* sp. CJF-002 and identification of genes for cellulose biosynthesis. *Cellulose* **19**, 1989-2001
191. Ellilä, S., Bromann, P., et al. (2019) Cloning of novel bacterial xylanases from lignocellulose-enriched compost metagenomic libraries. *AMB Express* **9**, 1-12
192. Chen, X.-L., Zhao, F., et al. (2018) A new group of modular xylanases in glycoside hydrolase family 8 from marine bacteria. *Applied and environmental microbiology* **84**, e01785-01718
193. Brennan, Y., Callen, W. N., et al. (2004) Unusual microbial xylanases from insect guts. *Applied and Environmental Microbiology* **70**, 3609-3617
194. Valenzuela, S. V., Lopez, S., et al. (2016) The glycoside hydrolase family 8 reducing-end xylose-releasing exo-oligoxylanase Rex8A from *Paenibacillus barcinonensis* BP-23 is active on branched xylooligosaccharides. *Applied and environmental microbiology* **82**, 5116-5124
195. Jiménez-Ortega, E., Valenzuela, S., et al. (2020) Structural analysis of the reducing-end xylose-releasing exo-oligoxylanase Rex8A from *Paenibacillus barcinonensis* BP-23 deciphers its molecular specificity. *The FEBS journal* **287**, 5362-5374
196. Grunwald, P. (2011) *Carbohydrate-modifying biocatalysts*, CRC Press
197. Tenkanen, M., Vršanská, M., et al. (2013) Xylanase XYN IV from *Trichoderma reesei* showing exo-and endo-xylanase activity. *The FEBS journal* **280**, 285-301
198. Hong, P.-Y., Iakiviak, M., et al. (2014) Two new xylanases with different substrate specificities from the human gut bacterium *Bacteroides intestinalis* DSM 17393. *Applied and environmental microbiology* **80**, 2084-2093
199. Lagaert, S., Van Campenhout, S., et al. (2007) Recombinant expression and characterization of a reducing-end xylose-releasing exo-oligoxylanase from *Bifidobacterium adolescentis*. *Applied and environmental microbiology* **73**, 5374-5377
200. Kmezik, C., Krska, D., et al. (2021) Characterization of a novel multidomain CE15-GH8 enzyme encoded by a polysaccharide utilization locus in the human gut bacterium *Bacteroides eggerthii*. *Sci. Rep.* **11**, 17662
201. Berman, H. M., Westbrook, J., et al. (2000) The protein data bank. *Nucleic Acids Res.* **28**, 235-242
202. Pollet, A., Delcour, J. A., et al. (2010) Structural determinants of the substrate specificities of xylanases from different glycoside hydrolase families. *Crit. Rev. Biotechnol.* **30**, 176-191
203. Adachi, W., Sakihama, Y., et al. (2004) Crystal structure of family GH-8 chitosanase with subclass II specificity from *Bacillus* sp. K17. *J. Mol. Biol.* **343**, 785-795
204. Chadha, B. S., Kaur, B., et al. (2019) Thermostable xylanases from thermophilic fungi and bacteria: current perspective. *Bioresour. Technol.* **277**, 195-203
205. Naumoff, D. (2016) GH10 family of glycoside hydrolases: structure and evolutionary connections. *Mol. Biol.* **50**, 132-140
206. Pell, G., Taylor, E. J., et al. (2004) The mechanisms by which family 10 glycoside hydrolases bind decorated substrates. *J. Biol. Chem.* **279**, 9597-9605
207. Chu, Y., Tu, T., et al. (2017) Insights into the roles of non-catalytic residues in the active site of a GH10 xylanase with activity on cellulose. *J. Biol. Chem.* **292**, 19315-19327
208. Andrews, S. R., Taylor, E. J., et al. (2004) The use of forced protein evolution to investigate and improve stability of family 10 xylanases: the production of Ca2+-independent stable xylanases. *J. Biol. Chem.* **279**, 54369-54379

209. Wakarchuk, W. W., Campbell, R. L., et al. (1994) Mutational and crystallographic analyses of the active site residues of the *Bacillus circulans* xylanase. *Protein Sci.* **3**, 467-475
210. Martínez, J. P., Falomir, M. P., et al. (2009) Chitin: a structural biopolysaccharide. *eLS*
211. Soni, S. K., Sharma, A., et al. (2018) Cellulases: role in lignocellulosic biomass utilization. *Cellulases*, 3-23
212. Thoresen, M., Malgas, S., et al. (2021) Revisiting the Phenomenon of Cellulase Action: Not All Endo-and Exo-Cellulase Interactions Are Synergistic. *Catalysts* **11**, 170
213. Den Haan, R., Van Zyl, J. M., et al. (2013) Modeling the minimum enzymatic requirements for optimal cellulose conversion. *Environmental Research Letters* **8**, 025013
214. Ganner, T., Bubner, P., et al. (2012) Dissecting and reconstructing synergism: in situ visualization of cooperativity among cellulases. *J. Biol. Chem.* **287**, 43215-43222
215. Baker, J. O., McCarley, J. R., et al. (2005) Catalytically enhanced endocellulase Cel5A from *Acidothermus cellulolyticus*. *Applied biochemistry and biotechnology* **121**, 129-148
216. Ravachol, J., Borne, R., et al. (2014) Characterization of all family-9 glycoside hydrolases synthesized by the cellulosome-producing bacterium *Clostridium cellulolyticum*. *J. Biol. Chem.* **289**, 7335-7348
217. Watanabe, H., and Tokuda, G. (2001) Animal cellulases. *Cellular and Molecular Life Sciences CMLS* **58**, 1167-1178
218. Watanabe, H., Noda, H., et al. (1998) A cellulase gene of termite origin. *Nature* **394**, 330-331
219. Suzuki, K. i., Ojima, T., et al. (2003) Purification and cDNA cloning of a cellulase from abalone *Haliotis discus* hannai. *Eur. J. Biochem.* **270**, 771-778
220. Dehal, P., Satou, Y., et al. (2002) The draft genome of *Ciona intestinalis*: insights into chordate and vertebrate origins. *Science* **298**, 2157-2167
221. Sathya, T., and Khan, M. (2014) Diversity of glycosyl hydrolase enzymes from metagenome and their application in food industry. *J. Food Sci.* **79**, R2149-R2156
222. Schaechter, M. (2009) *Encyclopedia of microbiology*, Academic Press
223. Duan, C.-J., Huang, M.-Y., et al. (2017) Characterization of a novel theme C glycoside hydrolase family 9 cellulase and its CBM-chimeric enzymes. *Applied microbiology and biotechnology* **101**, 5723-5737
224. Wu, L., and Davies, G. J. (2018) Structure of the GH9 glucosidase/glucosaminidase from *Vibrio cholerae*. *Acta Crystallographica Section F: Structural Biology Communications* **74**, 512-523
225. Zhou, W., Irwin, D. C., et al. (2004) Kinetic studies of *Thermobifida fusca* Cel9A active site mutant enzymes. *Biochemistry* **43**, 9655-9663
226. Brunecky, R., Alahuhta, M., et al. (2013) Revealing nature's cellulase diversity: the digestion mechanism of *Caldicellulosiruptor bescii* CelA. *Science* **342**, 1513-1516
227. Liu, Y.-J., Liu, S., et al. (2018) Determination of the native features of the exoglucanase Cel48S from *Clostridium thermocellum*. *Biotechnology for biofuels* **11**, 1-13
228. Kostylev, M., and Wilson, D. B. (2011) Determination of the catalytic base in family 48 glycosyl hydrolases. *Applied and environmental microbiology* **77**, 6274-6276
229. Brunecky, R., Alahuhta, M., et al. (2017) Natural diversity of glycoside hydrolase family 48 exoglucanases: insights from structure. *Biotechnology for biofuels* **10**, 1-9
230. Kostylev, M., Alahuhta, M., et al. (2014) Cel48A from *Thermobifida fusca*: structure and site directed mutagenesis of key residues. *Biotechnology and bioengineering* **111**, 664-673
231. Berger, E., Zhang, D., et al. (2007) Two noncellulosomal cellulases of *Clostridium thermocellum*, Cel9I and Cel48Y, hydrolyse crystalline cellulose synergistically. *FEMS Microbiol. Lett.* **268**, 194-201
232. Wilson, D. B. (2010) Demonstration of the importance for cellulose hydrolysis of CelS, the most abundant cellulosomal cellulase in *Clostridium thermocellum*. *Proceedings of the National Academy of Sciences* **107**, 17855-17856
233. Fujita, K., Shimomura, K., et al. (2006) A chitinase structurally related to the glycoside hydrolase family 48 is indispensable for the hormonally induced diapause termination in a beetle. *Biochemical and biophysical research communications* **345**, 502-507
234. Sukharnikov, L. O., Alahuhta, M., et al. (2012) Sequence, structure, and evolution of cellulases in glycoside hydrolase family 48. *J. Biol. Chem.* **287**, 41068-41077
235. Armendáriz-Ruiz, M., Rodríguez-González, J. A., et al. (2018) Carbohydrate esterases: An overview. *Lipases and Phospholipases*, 39-68

236. Mazurkewich, S., Poulsen, J.-C. N., et al. (2019) Structural and biochemical studies of the glucuronoyl esterase *Ot*CE15A illuminate its interaction with lignocellulosic components. *J. Biol. Chem.* **294**, 19978-19987
237. De Santi, C., Gani, O. A., et al. (2017) Structural insight into a CE15 esterase from the marine bacterial metagenome. *Sci. Rep.* **7**, 1-10
238. Mosbech, C., Holck, J., et al. (2018) The natural catalytic function of *Cu*GE glucuronoyl esterase in hydrolysis of genuine lignin–carbohydrate complexes from birch. *Biotechnology for biofuels* **11**, 1-9
239. Bååth, J. A., Mazurkewich, S., et al. (2018) Biochemical and structural features of diverse bacterial glucuronoyl esterases facilitating recalcitrant biomass conversion. *Biotechnology for biofuels* **11**, 1-14
240. Guillén, D., Sánchez, S., et al. (2010) Carbohydrate-binding domains: multiplicity of biological roles. *Applied microbiology and biotechnology* **85**, 1241-1249
241. Armenta, S., Moreno-Mendieta, S., et al. (2017) Advances in molecular engineering of carbohydrate-binding modules. *Proteins: Structure, Function, and Bioinformatics* **85**, 1602-1617
242. Georgelis, N., Yennawar, N. H., et al. (2012) Structural basis for entropy-driven cellulose binding by a type-A cellulose-binding module (CBM) and bacterial expansin. *Proceedings of the National Academy of Sciences* **109**, 14830-14835
243. Boraston, A. B., Bolam, D. N., et al. (2004) Carbohydrate-binding modules: fine-tuning polysaccharide recognition. *Biochemical journal* **382**, 769-781
244. Montanier, C., Van Bueren, A. L., et al. (2009) Evidence that family 35 carbohydrate binding modules display conserved specificity but divergent function. *Proceedings of the National Academy of Sciences* **106**, 3065-3070
245. Singh, A. K., Pluvinage, B., et al. (2014) Unravelling the multiple functions of the architecturally intricate *Streptococcus pneumoniae* β-galactosidase, BgaA. *PLoS pathogens* **10**, e1004364
246. van Bueren, A. L., Ficko-Blean, E., et al. (2011) The conformation and function of a multimodular glycogen-degrading pneumococcal virulence factor. *Structure* **19**, 640-651
247. Meekins, D. A., Raththagala, M., et al. (2014) Phosphoglucan-bound structure of starch phosphatase Starch Excess4 reveals the mechanism for C6 specificity. *Proceedings of the National Academy of Sciences* **111**, 7272-7277
248. Pires, V. M., Henshaw, J. L., et al. (2004) The crystal structure of the family 6 carbohydrate binding module from *Cellvibrio mixtus* endoglucanase 5a in complex with oligosaccharides reveals two distinct binding sites with different ligand specificities. *J. Biol. Chem.* **279**, 21560-21568
249. Yaniv, O., Frolow, F., et al. (2012) Interactions between family 3 carbohydrate binding modules (CBMs) and cellulosomal linker peptides. *Methods Enzymol.* **510**, 247-259
250. Hershko Rimon, A., Livnah, O., et al. (2021) Novel clostridial cell-surface hemicellulose-binding CBM3 proteins. *Acta Crystallographica Section F: Structural Biology Communications* **77**
251. Gilad, R., Rabinovich, L., et al. (2003) CelI, a noncellulosomal family 9 enzyme from *Clostridium thermocellum*, is a processive endoglucanase that degrades crystalline cellulose. *J. Bacteriol.* **185**, 391-398
252. Irwin, D., Shin, D.-H., et al. (1998) Roles of the catalytic domain and two cellulose binding domains of *Thermomonospora fusca* E4 in cellulose hydrolysis. *J. Bacteriol.* **180**, 1709-1714
253. Shimon, L. J., Belaich, A., et al. (2000) Structure of a family IIIa scaffoldin CBD from the cellulosome of *Clostridium cellulolyticum* at 2.2 Å resolution. *Acta Crystallogr. Sect. D. Biol. Crystallogr.* **56**, 1560-1568
254. Yaniv, O., Shimon, L. J., et al. (2011) Scaffoldin-borne family 3b carbohydrate-binding module from the cellulosome of *Bacteroides cellulosolvens*: structural diversity and significance of calcium for carbohydrate binding. *Acta Crystallogr. Sect. D. Biol. Crystallogr.* **67**, 506-515
255. Yaniv, O., Halfon, Y., et al. (2012) Structure of CBM3b of the major cellulosomal scaffoldin subunit ScaA from *Acetivibrio cellulolyticus*. *Acta Crystallographica Section F: Structural Biology and Crystallization Communications* **68**, 8-13
256. Yaniv, O., Fichman, G., et al. (2014) Fine-structural variance of family 3 carbohydrate-binding modules as extracellular biomass-sensing components of *Clostridium thermocellum* anti-σI factors. *Acta Crystallogr. Sect. D. Biol. Crystallogr.* **70**, 522-534
257. Tormo, J., Lamed, R., et al. (1996) Crystal structure of a bacterial family-III cellulose-binding domain: a general mechanism for attachment to cellulose. *The EMBO journal* **15**, 5739-5751

258. Notenboom, V., Boraston, A. B., et al. (2001) Crystal structures of the family 9 carbohydrate-binding module from *Thermotoga maritima* xylanase 10A in native and ligand-bound forms. *Biochemistry* **40**, 6248-6256
259. Sainz-Polo, M. A., González, B., et al. (2015) Exploring multimodularity in plant cell wall deconstruction: structural and functional analysis of Xyn10C containing the CBM22-1–CBM22-2 tandem. *J. Biol. Chem.* **290**, 17116-17130
260. Xie, H., Gilbert, H. J., et al. (2001) *Clostridium thermocellum* Xyn10B carbohydrate-binding module 22-2: the role of conserved amino acids in ligand binding. *Biochemistry* **40**, 9167-9176
261. Krska, D., and Larsbrink, J. (2020) Investigation of a thermostable multi-domain xylanase-glucuronoyl esterase enzyme from *Caldicellulosiruptor kristjanssonii* incorporating multiple carbohydrate-binding modules. *Biotechnology for biofuels* **13**, 1-13
262. Fontes, C., Hazlewood, G., et al. (1995) Evidence for a general role for non-catalytic thermostabilizing domains in xylanases from thermophilic bacteria. *Biochemical Journal* **307**, 151-158
263. Lee, Y.-E., Lowe, S., et al. (1993) Characterization of the active site and thermostability regions of endoxylanase from *Thermoanaerobacterium saccharolyticum* B6A-RI. *J. Bacteriol.* **175**, 5890-5898
264. Meng, D.-D., Ying, Y., et al. (2015) Distinct roles for carbohydrate-binding modules of glycoside hydrolase 10 (GH10) and GH11 xylanases from *Caldicellulosiruptor* sp. strain F32 in thermostability and catalytic efficiency. *Applied and environmental microbiology* **81**, 2006-2014
265. Dias, F. M., Goyal, A., et al. (2004) The N-terminal family 22 carbohydrate-binding module of xylanase 10B of *Clostridium themocellum* is not a thermostabilizing domain. *FEMS Microbiol. Lett.* **238**, 71-78
266. Ali, E., Zhao, G., et al. (2005) Functions of family-22 carbohydrate-binding module in *Clostridium thermocellum* Xyn10C. *Bioscience, biotechnology, and biochemistry* **69**, 160-165
267. Najmudin, S., Pinheiro, B. A., et al. (2010) Putting an N-terminal end to the *Clostridium thermocellum* xylanase Xyn10B story: Crystal structure of the CBM22-1–GH10 modules complexed with xylohexaose. *J. Struct. Biol.* **172**, 353-362
268. Zhao, G., Ali, E., et al. (2005) Function of the family-9 and family-22 carbohydrate-binding modules in a modular β-1, 3-1, 4-glucanase/xylanase derived from *Clostridium stercorarium* Xyn10B. *Bioscience, biotechnology, and biochemistry* **69**, 1562-1567
269. Roskoski, R. (2007) Enzyme Assays. in *xPharm: The Comprehensive Pharmacology Reference* (Enna, S. J., and Bylund, D. B. eds.), Elsevier, New York. pp 1-7
270. McKee, L. S. (2017) Measuring enzyme kinetics of glycoside hydrolases using the 3, 5-dinitrosalicylic acid assay. in *Protein-Carbohydrate Interactions*, Springer. pp 27-36
271. Mao, G., Chen, T.-H., et al. (2016) Cleavage of model substrates by *Arabidopsis thaliana* PRORP1 reveals new insights into its substrate requirements. *PLoS One* **11**, e0160246
272. Wojcik, M., and Miłek, J. (2016) A new method to determine optimum temperature and activation energies for enzymatic reactions. *Bioprocess and biosystems engineering* **39**, 1319-1323
273. Li, G., Rabe, K. S., et al. (2019) Machine learning applied to predicting microorganism growth temperatures and enzyme catalytic optima. *ACS synthetic biology* **8**, 1411-1420
274. Gorania, M., Seker, H., et al. (2010) Predicting a protein's melting temperature from its amino acid sequence. in *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*, IEEE
275. Almeida, V. M., and Marana, S. R. (2019) Optimum temperature may be a misleading parameter in enzyme characterization and application. *PLoS One* **14**, e0212977
276. Blanco, A., and Blanco, G. (2017) Chapter 8 - Enzymes. in *Med. Biochem.* (Blanco, A., and Blanco, G. eds.), Academic Press. pp 153-175
277. Garcia-Moreno, B. (2009) Adaptations of proteins to cellular and subcellular pH. *J. Biol.* **8**, 1-4
278. Ouellette, R. J., and Rawn, J. D. (2015) 14 - Amino Acids, Peptides, and Proteins. in *Principles of Organic Chemistry* (Ouellette, R. J., and Rawn, J. D. eds.), Elsevier, Boston. pp 371-396
279. Isom, D. G., Castañeda, C. A., et al. (2011) Large shifts in pKa values of lysine residues buried inside a protein. *Proceedings of the National Academy of Sciences* **108**, 5260-5265
280. Herlet, J., Kornberger, P., et al. (2017) A new method to evaluate temperature vs. pH activity profiles for biotechnological relevant enzymes. *Biotechnology for biofuels* **10**, 1-12
281. Schomburg, K. T., Ardao, I., et al. (2012) Computational biotechnology: prediction of competitive substrate inhibition of enzymes by buffer compounds with protein–ligand docking. *J. Biotechnol.* **161**, 391-401

282. Wang, L., Zhang, G., et al. (2019) Metagenomic analyses of microbial and carbohydrate-active enzymes in the rumen of holstein cows fed different forage-to-concentrate ratios. *Front. Microbiol.* **10**, 649
283. McGregor, N. G., Artola, M., et al. (2020) Rational design of mechanism-based inhibitors and activity-based probes for the identification of retaining α-l-arabinofuranosidases. *Journal of the American Chemical Society* **142**, 4648-4662
284. Funke, O. (1848) *Über das Milzvenenblut*,
285. McPherson, A., and Gavira, J. A. (2014) Introduction to protein crystallization. *Acta Crystallographica Section F: Structural Biology Communications* **70**, 2-20
286. Muirhead, H., and Perutz, M. (1963) Structure of hæmoglobin: A three-dimensional fourier synthesis of reduced human haemoglobin at 5.5 Å resolution. *Nature* **199**, 633-638
287. Wlodawer, A., Minor, W., et al. (2013) Protein crystallography for aspiring crystallographers or how to avoid pitfalls and traps in macromolecular structure determination. *The FEBS journal* **280**, 5705-5736
288. Rupp, B. (2009) *Biomolecular crystallography: principles, practice, and application to structural biology*, Garland Science
289. Wang, H. W., and Wang, J. W. (2017) How cryo-electron microscopy and X-ray crystallography complement each other. *Protein Sci.* **26**, 32-39
290. Putnam, D. K., Lowe Jr, E. W., et al. (2013) Reconstruction of SAXS profiles from protein structures. *Computational and structural biotechnology journal* **8**, e201308006
291. Skou, S., Gillilan, R. E., et al. (2014) Synchrotron-based small-angle X-ray scattering of proteins in solution. *Nat. Protoc.* **9**, 1727-1739
292. Campbell, I. D. (2013) The evolution of protein NMR. *Biomedical Spectroscopy and Imaging* **2**, 245-264
293. Howard, M. J. (1998) Protein NMR spectroscopy. *Curr. Biol.* **8**, R331-R333
294. Acton, T. B., Xiao, R., et al. (2011) Preparation of protein samples for NMR structure, function, and small-molecule screening studies. *Methods Enzymol.* **493**, 21-60
295. Gauto, D. F., Estrozi, L. F., et al. (2019) Integrated NMR and cryo-EM atomic-resolution structure determination of a half-megadalton enzyme complex. *Nature communications* **10**, 1-12
296. Liang, B., and Tamm, L. K. (2016) NMR as a tool to investigate the structure, dynamics and function of membrane proteins. *Nat. Struct. Mol. Biol.* **23**, 468-474
297. Callaway, E. (2020) Revolutionary cryo-EM is taking over structural biology. *Nature* **578**, 201-202
298. Passmore, L. A., and Russo, C. J. (2016) Specimen preparation for high-resolution cryo-EM. *Methods Enzymol.* **579**, 51-86
299. Henderson, R. (1995) The potential and limitations of neutrons, electrons and X-rays for atomic resolution microscopy of unstained biological molecules. *Q. Rev. Biophys.* **28**, 171-193
300. Liu, Y., Huynh, D. T., et al. (2019) A 3.8 Å resolution cryo-EM structure of a small protein bound to an imaging scaffold. *Nature communications* **10**, 1-7
301. Dill, K. A., Ozkan, S. B., et al. (2008) The protein folding problem. *Annu. Rev. Biophys.* **37**, 289-316
302. Waterhouse, A., Bertoni, M., et al. (2018) SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res.* **46**, W296-W303
303. Kelley, L. A., Mezulis, S., et al. (2015) The Phyre2 web portal for protein modeling, prediction and analysis. *Nat. Protoc.* **10**, 845-858
304. Rohl, C. A., Strauss, C. E., et al. (2004) Protein structure prediction using Rosetta. *Methods Enzymol.* **383**, 66-93
305. Jumper, J., Evans, R., et al. (2021) Highly accurate protein structure prediction with AlphaFold. *Nature*, 1-11
306. AlQuraishi, M. (2019) AlphaFold at CASP13. *Bioinformatics* **35**, 4862-4865
307. Hatch, V. DeepMind and EMBL release the most complete database of predicted 3D structures of human proteins. Accessed: 31/08/2021. https://www.embl.org/news/science/alphafold-database-launch/
308. Heckmann, C. M., and Paradisi, F. (2020) Looking back: A short history of the discovery of enzymes and how they became powerful chemical tools. *ChemCatChem* **12**, 6082
309. Boeckmann, B., Bairoch, A., et al. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* **31**, 365-370
310. (2021) UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.* **49**, D480-D489

311. Ravn, J. L., Engqvist, M. K., et al. (2021) CAZyme Prediction in Ascomycetous Yeast Genomes Guides Discovery of Novel Xylanolytic Species with Diverse Capacities for Hemicellulose Hydrolysis.
312. Rembeza, E., and Engqvist, M. K. M. (2021) Experimental and computational investigation of enzyme functional annotations uncovers misannotation in the EC 1.1.3.15 enzyme class. *PLOS Computational Biology* **17**, e1009446
313. Goodacre, N. F., Gerloff, D. L., et al. (2013) Protein domains of unknown function are essential in bacteria. *MBio* **5**, e00744-00713
314. Dilokpimol, A., Mäkelä, M. R., et al. (2018) Fungal glucuronoyl esterases: genome mining based enzyme discovery and biochemical characterization. *N. Biotechnol.* **40**, 282-287
315. Zaparucha, A., de Berardinis, V., et al. (2018) Genome mining for enzyme discovery. 1-27
316. Ufarté, L., Potocki-Veronese, G., et al. (2015) Discovery of new protein families and functions: new challenges in functional metagenomics for biotechnologies and microbial ecology. *Front. Microbiol.* **6**, 563
317. Shumilin, I. A., Cymborowski, M., et al. (2012) Identification of unknown protein function using metabolite cocktail screening. *Structure* **20**, 1715-1725
318. de Souza, W. R. (2013) Microbial degradation of lignocellulosic biomass. *Sustainable degradation of lignocellulosic biomass-techniques, applications and commercialization*, 207-247
319. Floudas, D., Binder, M., et al. (2012) The Paleozoic origin of enzymatic lignin decomposition reconstructed from 31 fungal genomes. *Science* **336**, 1715-1719
320. Nelsen, M. P., DiMichele, W. A., et al. (2016) Delayed fungal evolution did not cause the Paleozoic peak in coal production. *Proceedings of the National Academy of Sciences* **113**, 2442-2447
321. Corner, E. J. H. (2002) *The life of plants*, University of Chicago Press
322. Brink, D. P., Ravi, K., et al. (2019) Mapping the diversity of microbial lignin catabolism: experiences from the eLignin database. *Applied microbiology and biotechnology* **103**, 3979-4002
323. Cragg, S. M., Beckham, G. T., et al. (2015) Lignocellulose degradation mechanisms across the Tree of Life. *Curr. Opin. Chem. Biol.* **29**, 108-119
324. Andlar, M., Rezić, T., et al. (2018) Lignocellulose degradation: an overview of fungi and fungal enzymes involved in lignocellulose degradation. *Eng. Life Sci.* **18**, 768-778
325. Gusakov, A. V. (2011) Alternatives to *Trichoderma reesei* in biofuel production. *Trends Biotechnol.* **29**, 419-425
326. Bernardes, M. A. D. S. (2011) Biofuel Production: Recent Developments and Prospects.
327. Peterson, R., and Nevalainen, H. (2012) *Trichoderma reesei* RUT-C30–thirty years of strain improvement. *Microbiology* **158**, 58-68
328. Nevalainen, H., Suominen, P., et al. (1994) On the safety of *Trichoderma reesei*. *J. Biotechnol.* **37**, 193-200
329. Mandels, M., and Sternberg, D. (1976) Recent advances in cellulase technology. *Hakko Kogaku Zasshi;(Japan)* **54**
330. Seiboth, B., Ivanova, C., et al. (2011) *Trichoderma reesei*: a fungal enzyme producer for cellulosic biofuels. *Biofuel production-recent developments and prospects*, 309-340
331. Olsson, L., Christensen, T. M., et al. (2003) Influence of the carbon source on production of cellulases, hemicellulases and pectinases by *Trichoderma reesei* Rut C-30. *Enzyme and Microbial Technology* **33**, 612-619
332. Martinez, D., Berka, R. M., et al. (2008) Genome sequencing and analysis of the biomass-degrading fungus *Trichoderma reesei* (syn. *Hypocrea jecorina*). *Nat. Biotechnol.* **26**, 553-560
333. Druzhinina, I., and Kubicek, C. (2016) Familiar stranger: ecological genomics of the model saprotroph and industrial enzyme producer *Trichoderma reesei* breaks the stereotypes. *Adv. Appl. Microbiol.* **95**, 69-147
334. Gupta, V. K., Kubicek, C. P., et al. (2016) Fungal enzymes for bio-products from sustainable and waste biomass. *Trends in biochemical sciences* **41**, 633-645
335. Sissons, C. H., Sharrock, K. R., et al. (1987) Isolation of cellulolytic anaerobic extreme thermophiles from New Zealand thermal sites. *Applied and environmental microbiology* **53**, 832-838
336. Lee, L. L., Crosby, J. R., et al. (2020) The biology and biotechnology of the genus *Caldicellulosiruptor*: recent developments in 'Caldi World'. *Extremophiles* **24**, 1-15
337. Lee, L. L., Blumer-Schuette, S. E., et al. (2018) Genus-wide assessment of lignocellulose utilization in the extremely thermophilic genus *Caldicellulosiruptor* by genomic, pangenomic, and metagenomic analyses. *Applied and environmental microbiology* **84**, e02694-02617

338. Cha, M., Chung, D., et al. (2013) Metabolic engineering of *Caldicellulosiruptor bescii* yields increased hydrogen production from lignocellulosic biomass. *Biotechnology for biofuels* **6**, 1-8
339. Groom, J., Chung, D., et al. (2014) Heterologous complementation of a pyrF deletion in *Caldicellulosiruptor hydrothermalis* generates a new host for the analysis of biomass deconstruction. *Biotechnology for biofuels* **7**, 1-10
340. Lipscomb, G. L., Conway, J. M., et al. (2016) A highly thermostable kanamycin resistance marker expands the tool kit for genetic manipulation of *Caldicellulosiruptor bescii*. *Applied and environmental microbiology* **82**, 4421-4428
341. Chung, D.-H., Huddleston, J. R., et al. (2011) Identification and characterization of CbeI, a novel thermostable restriction enzyme from *Caldicellulosiruptor bescii* DSM 6725 and a member of a new subfamily of HaeIII-like enzymes. *Journal of Industrial Microbiology and Biotechnology* **38**, 1867
342. Branco, R. H., Serafim, L. S., et al. (2019) Second generation bioethanol production: on the use of pulp and paper industry wastes as feedstock. *Fermentation* **5**, 4
343. Chung, D., Cha, M., et al. (2014) Direct conversion of plant biomass to ethanol by engineered *Caldicellulosiruptor bescii*. *Proceedings of the National Academy of Sciences* **111**, 8931-8936
344. Williams-Rhaesa, A. M., Rubinstein, G. M., et al. (2018) Engineering redox-balanced ethanol production in the cellulolytic and extremely thermophilic bacterium, *Caldicellulosiruptor bescii*. *Metabolic engineering communications* **7**, e00073
345. Chung, D., Cha, M., et al. (2015) Cellulosic ethanol production via consolidated bioprocessing at 75 C by engineered *Caldicellulosiruptor bescii*. *Biotechnology for biofuels* **8**, 163
346. Vishnivetskaya, T. A., Hamilton-Brehm, S. D., et al. (2015) Community analysis of plant biomass-degrading microorganisms from Obsidian Pool, Yellowstone National Park. *Microb. Ecol.* **69**, 333-345
347. Zurawski, J. V., Blumer-Schuette, S. E., et al. (2014) The extremely thermophilic genus *Caldicellulosiruptor*: physiological and genomic characteristics for complex carbohydrate conversion to molecular hydrogen. in *Microbial bioenergy: hydrogen production*, Springer. pp 177-195
348. Zverlov, V., Mahr, S., et al. (1998) Properties and gene structure of a bifunctional cellulolytic enzyme (CelA) from the extreme thermophile '*Anaerocellum thermophilum*' with separate glycosyl hydrolase family 9 and 48 catalytic domains. *Microbiology* **144**, 457-465
349. Brunecky, R., Donohoe, B. S., et al. (2017) The multi domain *Caldicellulosiruptor bescii* CelA cellulase excels at the hydrolysis of crystalline cellulose. *Sci. Rep.* **7**, 1-17
350. Hehemann, J.-H., Correc, G., et al. (2010) Transfer of carbohydrate-active enzymes from marine bacteria to Japanese gut microbiota. *Nature* **464**, 908-912
351. Kmezik, C., Mazurkewich, S., et al. (2021) A polysaccharide utilization locus from the gut bacterium *Dysgonomonas mossii* encodes functionally distinct carbohydrate esterases. *J. Biol. Chem.* **296**
352. Kmezik, C., Bonzom, C., et al. (2020) Multimodular fused acetyl–feruloyl esterases from soil and gut Bacteroidetes improve xylanase depolymerization of recalcitrant biomass. *Biotechnology for biofuels* **13**, 1-14
353. Fogarty, L. R., and Voytek, M. A. (2005) Comparison of *Bacteroides-Prevotella* 16S rRNA genetic markers for fecal samples from different animal species. *Applied and Environmental Microbiology* **71**, 5999-6007
354. Sakamoto, M., and Ohkuma, M. (2013) *Bacteroides reticulotermitis* sp. nov., isolated from the gut of a subterranean termite (*Reticulitermes speratus*). *International journal of systematic and evolutionary microbiology* **63**, 691-695
355. Salyers, A. (1984) *Bacteroides* of the human lower intestinal tract. *Annu. Rev. Microbiol.* **38**, 293-313
356. Ozaki, D., Kubota, R., et al. (2021) Association between gut microbiota, bone metabolism, and fracture risk in postmenopausal Japanese women. *Osteoporosis International* **32**, 145-156
357. Mudd, A. T., Berding, K., et al. (2017) Serum cortisol mediates the relationship between fecal *Ruminococcus* and brain N-acetylaspartate in the young pig. *Gut Microbes* **8**, 589-600
358. Ridaura, V. K., Faith, J. J., et al. (2013) Gut microbiota from twins discordant for obesity modulate metabolism in mice. *Science* **341**
359. Bjursell, M. K., Martens, E. C., et al. (2006) Functional genomic and metabolic studies of the adaptations of a prominent adult human gut symbiont, *Bacteroides thetaiotaomicron*, to the suckling period. *J. Biol. Chem.* **281**, 36269-36279

360. Sonnenburg, J. L., Angenent, L. T., et al. (2004) Getting a grip on things: how do communities of bacterial symbionts become established in our intestine? *Nat. Immunol.* **5**, 569-573
361. Hooper, L. V., Midtvedt, T., et al. (2002) How host-microbial interactions shape the nutrient environment of the mammalian intestine. *Annu. Rev. Nutr.* **22**, 283-307
362. Sonnenburg, J. L., Xu, J., et al. (2005) Glycan foraging in vivo by an intestine-adapted bacterial symbiont. *Science* **307**, 1955-1959
363. Rakoff-Nahoum, S., Coyne, M., et al. (2014) An ecological network of polysaccharide utilization among human intestinal symbionts. *Curr. Biol.* **24**, 40-49
364. Rakoff-Nahoum, S., Foster, K. R., et al. (2016) The evolution of cooperation within the gut microbiota. *Nature* **533**, 255-259
365. El Kaoutari, A., Armougom, F., et al. (2013) The abundance and variety of carbohydrate-active enzymes in the human gut microbiota. *Nature Reviews Microbiology* **11**, 497-504
366. Wexler, H. M. (2007) *Bacteroides*: the good, the bad, and the nitty-gritty. *Clin. Microbiol. Rev.* **20**, 593-621
367. Terrapon, N., Lombard, V., et al. (2015) Automatic prediction of polysaccharide utilization loci in *Bacteroidetes* species. *Bioinformatics* **31**, 647-655
368. Terrapon, N., Lombard, V., et al. (2018) PULDB: the expanded database of Polysaccharide Utilization Loci. *Nucleic Acids Res.* **46**, D677-D683
369. Grondin, J. M., Tamura, K., et al. (2017) Polysaccharide utilization loci: fueling microbial communities. *J. Bacteriol.* **199**, e00860-00816
370. Martens, E. C., Chiang, H. C., et al. (2008) Mucosal glycan foraging enhances fitness and transmission of a saccharolytic human gut bacterial symbiont. *Cell Host Microbe* **4**, 447-457
371. Koropatkin, N. M., Martens, E. C., et al. (2008) Starch catabolism by a prominent human gut symbiont is directed by the recognition of amylose helices. *Structure* **16**, 1105-1115
372. Koropatkin, N. M., and Smith, T. J. (2010) SusG: a unique cell-membrane-associated α-amylase from a prominent human gut symbiont targets complex starch molecules. *Structure* **18**, 200-215
373. Kitamura, M., Okuyama, M., et al. (2008) Structural and functional analysis of a glycoside hydrolase family 97 enzyme from *Bacteroides thetaiotaomicron*. *J. Biol. Chem.* **283**, 36328-36337
374. Cameron, E. A., Maynard, M. A., et al. (2012) Multidomain carbohydrate-binding proteins involved in *Bacteroides thetaiotaomicron* starch metabolism. *J. Biol. Chem.* **287**, 34614-34625
375. McKee, L. S., La Rosa, S. L., et al. (2021) Polysaccharide degradation by the Bacteroidetes—mechanisms and nomenclature. *Environ. Microbiol. Rep.*
376. Zhivin, O., Dassa, B., et al. (2017) Unique organization and unprecedented diversity of the *Bacteroides* (*Pseudobacteroides*) *cellulosolvens* cellulosome system. *Biotechnology for biofuels* **10**, 1-19
377. Raveendran, S., Parameswaran, B., et al. (2018) Applications of microbial enzymes in food industry. *Food technology and biotechnology* **56**, 16-30
378. Singh, R., Kumar, M., et al. (2016) Microbial enzymes: industrial progress in 21st century. *3 Biotech* **6**, 1-15
379. Pandey, A., Selvakumar, P., et al. (1999) Solid state fermentation for the production of industrial enzymes. *Curr. Sci.*, 149-162
380. Berka, R., Dunn-Coleman, N., et al. (1992) Industrial enzymes from *Aspergillus* species. *Biotechnology (USA)*
381. Hu, H., Van den Brink, J., et al. (2011) Improved enzyme production by co-cultivation of *Aspergillus niger* and *Aspergillus oryzae* and with other fungi. *International Biodeterioration & Biodegradation* **65**, 248-252
382. Maheshwari, D., Gohade, S., et al. (1994) Paper mill sludge as a potential source for cellulase production by *Trichoderma reesei* QM 9123 and *Aspergillus niger* using mixed cultivation. *Carbohydr. Polym.* **23**, 161-163
383. Danilova, I., and Sharipova, M. (2020) The practical potential of *Bacilli* and their enzymes for industrial production. *Front. Microbiol.* **11**
384. Sayut, D. J., Kambam, P. K., et al. (2010) Enzyme Production in *Escherichia coli*. *Manual of Industrial Microbiology and Biotechnology*, 539-548
385. Kmezik, C. (2021) Exploring and exploiting plant biomass degradation by Bacteroidetes.
386. Glasgow, E., Vander Meulen, K., et al. (2021) Multifunctional cellulases are potent, versatile tools for a renewable bioeconomy. *Curr. Opin. Biotechnol.* **67**, 141-148

387. Bae, J., Morisaka, H., et al. (2013) Cellulosome complexes: natural biocatalysts as arming microcompartments of enzymes. *Journal of molecular microbiology and biotechnology* **23**, 370-378
388. Bayer, E. A., Belaich, J.-P., et al. (2004) The cellulosomes: multienzyme machines for degradation of plant cell wall polysaccharides. *Annu. Rev. Microbiol.* **58**, 521-554
389. Shoham, Y., Lamed, R., et al. (1999) The cellulosome concept as an efficient microbial strategy for the degradation of insoluble polysaccharides. *Trends Microbiol.* **7**, 275-281
390. Dashtban, M., Schraft, H., et al. (2009) Fungal bioconversion of lignocellulosic residues; opportunities & perspectives. *Int. J. Biol. Sci.* **5**, 578
391. Brás, J. L., Carvalho, A. L., et al. (2012) *Escherichia coli* expression, purification, crystallization, and structure determination of bacterial cohesin–dockerin complexes. *Methods Enzymol.* **510**, 395-415
392. Ding, S.-Y., Bayer, E. A., et al. (1999) A novel cellulosomal scaffoldin from *Acetivibrio cellulolyticus* that contains a family 9 glycosyl hydrolase. *J. Bacteriol.* **181**, 6720-6729
393. Bule, P., Pires, V. M., et al. (2018) Higher order scaffoldin assembly in *Ruminococcus flavefaciens* cellulosome is coordinated by a discrete cohesin-dockerin interaction. *Sci. Rep.* **8**, 1-14
394. Doi, R. H., and Kosugi, A. (2004) Cellulosomes: plant-cell-wall-degrading enzyme complexes. *Nature reviews microbiology* **2**, 541-551
395. Barth, A., Hendrix, J., et al. (2018) Dynamic interactions of type I cohesin modules fine-tune the structure of the cellulosome of *Clostridium thermocellum*. *Proceedings of the National Academy of Sciences* **115**, E11274-E11283
396. Fierobe, H.-P., Bayer, E. A., et al. (2002) Degradation of cellulose substrates by cellulosome chimeras: substrate targeting versus proximity of enzyme components. *J. Biol. Chem.* **277**, 49621-49630
397. Yamada, R., Hasunuma, T., et al. (2013) Endowing non-cellulolytic microorganisms with cellulolytic activity aiming for consolidated bioprocessing. *Biotechnol. Adv.* **31**, 754-763
398. Tanaka, T., and Kondo, A. (2015) Cell surface engineering of industrial microorganisms for biorefining applications. *Biotechnol. Adv.* **33**, 1403-1411
399. Arora, R., Behera, S., et al. (2015) Bioprospecting thermostable cellulosomes for efficient biofuel production from lignocellulosic biomass. *Bioresources and Bioprocessing* **2**, 1-12
400. Garvey, M., Klose, H., et al. (2013) Cellulases for biomass degradation: comparing recombinant cellulase expression platforms. *Trends Biotechnol.* **31**, 581-593
401. Li, X., Xia, J., et al. (2019) Construction and characterization of bifunctional cellulases: *Caldicellulosiruptor*-sourced endoglucanase, CBM, and exoglucanase for efficient degradation of lignocellulose. *Biochem. Eng. J.* **151**, 107363
402. Nath, P., Dhillon, A., et al. (2019) Development of bi-functional chimeric enzyme (CtGH1-L1-CtGH5-F194A) from endoglucanase (CtGH5) mutant F194A and β-1, 4-glucosidase (CtGH1) from *Clostridium thermocellum* with enhanced activity and structural integrity. *Bioresour. Technol.* **282**, 494-501
403. Conway, J. M., Pierce, W. S., et al. (2016) Multidomain, surface layer-associated glycoside hydrolases contribute to plant polysaccharide degradation by *Caldicellulosiruptor* species. *J. Biol. Chem.* **291**, 6732-6747
404. Brunecky, R., Subramanian, V., et al. (2020) Synthetic fungal multifunctional cellulases for enhanced biomass conversion. *Green Chem.* **22**, 478-489
405. Wang, R., Gong, L., et al. (2016) Identification of the C-Terminal GH5 Domain from Cb Cel9B/Man5A as the First Glycoside Hydrolase with Thermal Activation Property from a Multimodular Bifunctional Enzyme. *PLoS One* **11**, e0156802
406. Conway, J. M., Crosby, J. R., et al. (2018) Parsing in vivo and in vitro contributions to microcrystalline cellulose hydrolysis by multidomain glycoside hydrolases in the *Caldicellulosiruptor bescii* secretome. *Biotechnology and bioengineering* **115**, 2426-2440
407. Blumer-Schuette, S. E. (2020) Insights into thermophilic plant biomass hydrolysis from *Caldicellulosiruptor* systems biology. *Microorganisms* **8**, 385
408. Chu, Y., Hao, Z., et al. (2019) The GH10 and GH48 dual-functional catalytic domains from a multimodular glycoside hydrolase synergize in hydrolyzing both cellulose and xylan. *Biotechnology for biofuels* **12**, 1-10
409. Larsbrink, J., Zhu, Y., et al. (2016) A polysaccharide utilization locus from *Flavobacterium johnsoniae* enables conversion of recalcitrant chitin. *Biotechnology for biofuels* **9**, 1-16

410. Yi, Z., Su, X., et al. (2013) Molecular and biochemical analyses of CbCel9A/Cel48A, a highly secreted multi-modular cellulase by *Caldicellulosiruptor bescii* during growth on crystalline cellulose. *PLoS One* **8**, e84172
411. Rosano, G. L., and Ceccarelli, E. A. (2014) Recombinant protein expression in *Escherichia coli*: advances and challenges. *Front. Microbiol.* **5**, 172
412. Zhang, N., and An, Z. (2010) Heterologous protein expression in yeasts and filamentous fungi. *Manual of Industrial Microbiology and Biotechnology*, 145-156
413. Cai, D., Rao, Y., et al. (2019) Engineering *Bacillus* for efficient production of heterologous protein: current progress, challenge and prospect. *J. Appl. Microbiol.* **126**, 1632-1642
414. El-Baky, N. A., and Redwan, E. M. (2015) Therapeutic alpha-interferons protein: structure, production, and biosimilar. *Preparative Biochemistry and Biotechnology* **45**, 109-127
415. Robertson, M. P., and Scott, W. G. (2007) Designer enzymes. *Nature* **448**, 757-758
416. Lutz, S., and Iamurri, S. M. (2018) Protein engineering: past, present, and future. *Protein Eng.*, 1-12
417. Estell, D. A., Graycar, T. P., et al. (1985) Engineering an enzyme by site-directed mutagenesis to be resistant to chemical oxidation. *J. Biol. Chem.* **260**, 6518-6521
418. Chen, K., and Arnold, F. H. (1991) Enzyme engineering for nonaqueous solvents: random mutagenesis to enhance activity of subtilisin E in polar organic media. *Biotechnology. (N. Y.)* **9**, 1073-1077
419. Bornscheuer, U. T., Huisman, G., et al. (2012) Engineering the third wave of biocatalysis. *Nature* **485**, 185-194
420. Richter, F., Leaver-Fay, A., et al. (2011) *De novo* enzyme design using Rosetta3. *PLoS One* **6**, e19230
421. Korendovych, I. V. (2018) Rational and semirational protein design. *Protein Eng.*, 15-23
422. Escovar-Kousen, J. M., Wilson, D., et al. (2004) Integration of computer modeling and initial studies of site-directed mutagenesis to improve cellulase activity on Cel9A from *Thermobifida fusca*. *Applied biochemistry and biotechnology* **113**, 287-297
423. Kauffmann, I., and Schmidt-Dannert, C. (2001) Conversion of *Bacillus thermocatenulatus* lipase into an efficient phospholipase with increased activity towards long-chain fatty acyl substrates by directed evolution and rational design. *Protein Eng.* **14**, 919-928
424. Durao, P., Bento, I., et al. (2006) Perturbations of the T1 copper site in the CotA laccase from *Bacillus subtilis*: structural, biochemical, enzymatic and stability studies. *JBIC Journal of Biological Inorganic Chemistry* **11**, 514-526
425. Zhang, J., Shi, H., et al. (2015) Site-directed mutagenesis of a hyperthermophilic endoglucanase Cel12B from *Thermotoga maritima* based on rational design. *PLoS One* **10**, e0133824
426. Deng, Z., Yang, H., et al. (2014) Structure-based rational design and introduction of arginines on the surface of an alkaline α-amylase from *Alkalimonas amylolytica* for improved thermostability. *Applied microbiology and biotechnology* **98**, 8937-8945
427. Anbar, M., Gul, O., et al. (2012) Improved thermostability of *Clostridium thermocellum* endoglucanase Cel8A by using consensus-guided mutagenesis. *Applied and environmental microbiology* **78**, 3458-3464
428. Chen, Z., Friedland, G. D., et al. (2012) Tracing determinants of dual substrate specificity in glycoside hydrolase family 5. *J. Biol. Chem.* **287**, 25335-25343
429. Horowitz, N. H. (1995) One-gene-one-enzyme: Remembering biochemical genetics. *Protein Sci.* **4**, 1017-1019
430. Wang, Y., Liu, J., et al. (2015) Mechanism of alternative splicing and its regulation. *Biomedical reports* **3**, 152-158
431. Aranko, A. S., Wlodawer, A., et al. (2014) Nature's recipe for splitting inteins. *Protein Eng. Des. Sel.* **27**, 263-271
432. Dassa, B., London, N., et al. (2009) Fractured genes: a novel genomic arrangement involving new split inteins and a new homing endonuclease family. *Nucleic Acids Res.* **37**, 2560-2573
433. Shi, J., and Muir, T. W. (2005) Development of a tandem protein trans-splicing system based on native and engineered split inteins. *Journal of the American Chemical Society* **127**, 6198-6206
434. Otomo, T., Ito, N., et al. (1999) NMR observation of selected segments in a larger protein: central-segment isotope labeling through intein-mediated ligation. *Biochemistry* **38**, 16040-16044
435. Shah, N. H., and Muir, T. W. (2014) Inteins: nature's gift to protein chemists. *Chemical science* **5**, 446-461

436. Carvajal-Vallejos, P., Pallissé, R., et al. (2012) Unprecedented rates and efficiencies revealed for new natural split inteins from metagenomic sources. *J. Biol. Chem.* **287**, 28686-28696
437. Shah, N. H., Dann, G. P., et al. (2012) Ultrafast protein splicing is common among cyanobacterial split inteins: implications for protein engineering. *Journal of the American Chemical Society* **134**, 11338-11341
438. Mills, K. V., Lew, B. M., et al. (1998) Protein splicing in trans by purified N-and C-terminal fragments of the *Mycobacterium tuberculosis* RecA intein. *Proceedings of the National Academy of Sciences* **95**, 3543-3548
439. Wu, H., Hu, Z., et al. (1998) Protein trans-splicing by a split intein encoded in a split DnaE gene of *Synechocystis* sp. PCC6803. *Proceedings of the National Academy of Sciences* **95**, 9226-9231
440. Zettler, J., Schütz, V., et al. (2009) The naturally split Npu DnaE intein exhibits an extraordinarily high rate in the protein trans-splicing reaction. *FEBS Lett.* **583**, 909-914
441. Muona, M., Aranko, A. S., et al. (2010) Segmental isotopic labeling of multi-domain and fusion proteins by protein trans-splicing in vivo and in vitro. *Nat. Protoc.* **5**, 574-587
442. Busche, A. E., Aranko, A. S., et al. (2009) Segmental isotopic labeling of a central domain in a multidomain protein by protein trans-splicing using only one robust DnaE intein. *Angew. Chem. Int. Ed.* **48**, 6128-6131
443. Palanisamy, N., Degen, A., et al. (2019) Split intein-mediated selection of cells containing two plasmids using a single antibiotic. *Nature communications* **10**, 1-15
444. López-Igual, R., Bernal-Bayard, J., et al. (2019) Engineered toxin–intein antimicrobials can selectively target and kill antibiotic-resistant bacteria in mixed populations. *Nat. Biotechnol.* **37**, 755-760
445. Cooper, M. A., Taris, J. E., et al. (2018) A convenient split-intein tag method for the purification of tagless target proteins. *Current protocols in protein science* **91**, 5.29. 21-25.29. 23
446. Gramespacher, J. A., Stevens, A. J., et al. (2017) Intein zymogens: conditional assembly and splicing of split inteins via targeted proteolysis. *Journal of the American Chemical Society* **139**, 8074-8077