FUNCTIONAL CHARACTERISATION OF A THERMOPHILIC CELLULASE FROM A MALAWIAN METAGENOMIC LIBRARY

Timna January

A thesis submitted in fulfillment of the requirements for the degree of
MAGISTER SCIENTIAE (M. Sc)

Institute for Microbial Biotechnology and Metagenomics
Department of Biotechnology, University of the Western Cape,
Bellville, Cape Town
South Africa

Supervisor: Assoc. Prof. Marla Tuffin
Co-supervisors: Dr. Robert Huddy and Dr. Mariette Smart

DECLARATION

I, Timna January, hereby declare that "**Functional characterisation of a thermophilic cellulase from a Malawian metagenomic library**" is my own original work and that I have accurately reported and acknowledged all sources, and that this document has not previously, in its entirety or in part been submitted at any university for the purpose of obtaining an academic qualification.

…… ……

(Signature) (Date)

ABSTRACT

Biofuels are currently recognised as the most viable source of energy to replace depleting fossil fuel reserves, with bioethanol the most popular alternative alcohol fuel. Producing bioethanol from agricultural waste residues is a feasible socio-economic industrial process. Lignocellulose, from which plant material is composed, is highly recalcitrant to enzymatic degradation and therefore requires a suite of enzymes for complete hydrolysis of the biomass. Metagenomes, particularly from extreme environments, represent an unlimited resource for the discovery of novel biocatalysts for inclusion in industrial processes. Here we report on the cloning and functional characterisation of a novel thermophilic cellulase identified by the functional screening of a Malawian, hotspring sediment metagenomic library. The gene encoding the cellulase, *celMHS*, composed of 2,705 nucleotides and encoded a polypeptide of 905 amino acids with a predicted molecular mass of about 98 kDa. The *in silico* translated protein, CelMHS, contained a putative transmembrane domain, a family 4 carbohydrate binding motive (CBM 4), a truncated glycoside hydrolase family 42 (GH42) domain and a N-terminal region that does not have sequence similarity to any previously described domains. Functional characterisation of the recombinant CelMHS demonstrated that the protein displayed an optimal pH of 6.0 and temperature of 100°C. CelMHS had high specific activity toward substrates comprising of β-1,4 linked glucose subunits such as carboxymethyl cellulose, β-D-glucan from barley and lichenan, however, some activity was also observed against avicel, a crystalline cellulose substrate. HPLC analysis of the hydrolysis products produced by CelMHS indicates that this particular enzyme prefers longer chain oligosaccharides. This is, to the best of our knowledge, the first investigation describing the cloning and characterization of a carbohydrate hydrolysing enzyme comprised of the unique sequence architecture: a partial GH42 catalytic domain, a CBM 4 and a unique N-domain sequence.

Key words: cellulose, cellulases, lignocellulosic biomass, bioethanol, saccharification, hydrolysis, metagenomic library, thermophilic

ACKNOWLEDGEMENTS

I would like to express my sincere appreciation to the following people for their contributions to my work:

- To God, in whom I trust and believe, for carrying me through these years.
- Associate Professor Marla Tuffin, thank you for giving me the opportunity and funding to work in a world class laboratory. Thank you for all your patience and understanding.
- NRF for financial support and funding of this project.
- Dr.Heide Goodman and Dr.Rolene Bauer for encouragement and administrative help throughout my years at IMBM.
- Dr. Robert Huddy and Dr. Mariette Smart, as co-supervisors involved in this project, words cannot describe how thankful I am to the both of you. Together, you have played an amazing role in helping me to complete this thesis and molding me as a scientist. I am appreciative of all the hours you have spent guiding and advising me throughout my MSc.
- I am also deeply indebted to Dr. Inonge Mulako, Dr. Colin Anthony and Lesley-Ann Jeftha for reading and editing drafts of this thesis.
- I am appreciative to individuals who took the time to train me over the course of my MSc, especially Dr Inonge Mulako, Dr.Thulani Makhalanyane and William Mavengere.
- To everyone that contributed to the day to day running of the laboratory, including Aunty Anna, Carmen and Ruth. I am very thankful for everything that you did to assist me and keep the lab running.
- I am grateful to a number of people at IMBM who have been kind to me over the course of my MSc. Particularily, 'Bench Inc.™' (Dean, Lesley, Megan, Layla and Randal), Thulani, Rhulani, Munaka, Rudzani, Gao, Natasha, Tanya, Jarishma, Grant, Oratilwe and Xiao Ping.

- Xiao Ping Hu and Ilana Ackermann for their practical contributions towards this study. Xiao Ping for construction and initial screening of the library and Ilana for providing clone, *celMHSΔTMR*
- I would like to thank all the academic staff in the Biotechnology department, particularly Charles Gelderbloem, Clifford Jacobs and Frances Starkey.
- Thank you to Ron McEwan, Anika De Beer, Martmari Botha and the rest of my colleagues at Kapa Biosystems, South Africa. Your support and understanding during the period I was crafting my thesis will forever be remembered.
- I would like to thank my family for their support over the past few years. Especially Nilton, my cousins (Earl, Bianca, Tarin, Miguel, Dale, Marcello, Sharna, Mishay, Ashley, Lucille) my Uncles Johnny, Stephen, David, my Aunts Yolanda, Una, Amana, Sandra, Myrna and my parents in-law.
- Finally, my husband and rock, Justin Daniels thank you! Throughout this entire process, you have been there supporting me regardless. Thank you for understanding the numerous evenings and weekends that I spent working on this thesis. I love you so very much, and promise to make up for all the time we have lost.

DEDICATION

To my mother and father, your constant support, guidance and encouragement has enabled me to complete this thesis. Thank you for always being there and praying me through. I love you always.

ABBREVIATIONS

Amp	Ampicillin
APS	Ammonium persulphate
bp	Base pairs
BSA	Bovine serum albumin
BLAST	Basic Local Alignment Search Tool
Cam	Chloramphenicol
celMHS	cellulase from a Malawian Hot Spring\
CMC	Carboxymethyl cellulose
dH2O	Demineralised water
DNA	Deoxyribonucleic acid
DNS	3,5-dinitrosalicylic acid
dNTP	Deoxyribonucleotides
et al	*et alia* (and others)
EDTA	Ethylenediamine tetra-acetic acid
g	Gram
HCl	Hydrochloric acid
HPLC	High pressure liquid chromotography
hrs	Hours
IPTG	Isopropyl-b-D-thiogalactopyranoside
Kan	Kanamycin
kDa	Kilo Dalton
V	Volts
L	Litre
LB	Luria Bertani

LA	Luria Bertani agar
M	Molar
MES	2-(N-morpholino)ethanesulfonic acid
min	minute(s)
µ	micro
ml	Millilitre
µl	Microlitre
MW	Molecular weight
NCBI	National Center for Biotechnology Information
ng	Nanogram
ORF	Open reading frame
PAGE	Polyacrylamide gel electrophoresis
PCR	Polymerase chain reaction
rpm	revolutions per minute
s	Seconds
SCB	Sugarcane bagasse
SOB	Super Optimal Broth
SOC	Super Optimal broth with Catabolic repressor
SDS	Sodium Dodecyl Sulphate
TAE	Tris Acetate EDTA
TEMED	N,N,N',N'-Tetramethylethylenediamine
TMR	Transmembrane region
Tris	Tris-hydroxymethyl-aminomethane
UV	Ultraviolet
v/v	Volume per volume
w/v	Weight per volume
xg	Centrifugal force

Δ	delta
°C	Degrees Celsius

Content

Chapter 1

Thermophilic enzymatic saccharification of lignocellulosic biomass for bioethanol production

1.1 Introduction

Modern society is dependent on fossil fuel to sustain electric power plants, transportation systems as well as gas fired furnaces (Huber, 2009). Fossil fuels dominate the world's energy market, which is estimated to be worth approximately 1.5 trillion US$ annually (Goldemberg, 2006). These fossil fuels are non-renewable energy sources and there is immense pressure on the sustainability of the remaining fossil fuel reserves which are estimated to be depleted by 2030 (Shafiee and Topal, 2009). Consciousness of the finite nature of the world's remaining fossil fuel quantities, the high costs associated with extracting fossil fuels and global awareness of the adverse effect fossil fuel combustion has on the environment has driven a demand for alternative renewable energy sources (Demibras, 2008).

1.2 Alternative fuel sources

Alternative energy sources are primarily based on sustainable, renewable and ecologically friendly processes (Kumar *et al*., 2008), and especially biofuels are currently recognised as one of the most viable alternative energy sources to replace the use of fossil fuels particularly within the transportation sector (Demibras, 2008). Biofuels are considered to be 'environmentally friendly' and benefit the economy and consumer in that they are biodegradable, reduce the net emission of CO_2 and contribute to sustainability (Margoet *et al.*, 2009). As they are primarily produced from biological material derived from living, or recently living organisms (Kumar *et al.*, 2008), biofuels are described as infinite fuels (Aleklett and Campbell, 2003; Demibras, 2007; Margoet *et al.*, 2009). Examples of biofuels include bioethanol (Margoet *et al.*, 2009), biomethanol (Nakagawa *et al.*, 2007), biodiesel (Ma and Hanna, 1999), biogas (Taherzadeh and Karimni, 2008), biosyngas (Srinivas *et al*., 2007), bio-oil (Mohan *et al*., 2006), Fischer-Tropsch liquids (Tijmensen *et al*., 2002) and biohydrogen (Lo *et al.,* 2009). Of these, bioethanol and biodiesel have successfully been used within the transportation sectors of Brazil and the United States of America (USA) by blending these fuels with petrol to produce a mixture known as E10 [10% (v/v) ethanol and 90%

(v/v) petrol] to substitute conventional motor fuel without affecting the car's performance (Balat and Balat, 2009; Demibras, 2009).

1.2.1 South African perspective on biofuels

South Africa has substantial coal reserves, and as a result has developed an extensive power production system that relies heavily on the combustion of coal to meet the country's energy demands. The South African government has recognised the negative impact that greenhouse gas emission from fossil fuel combustion, such as coal, has on the environment and in response has published a White Paper on Renewable Energy (November 2003; http://unfccc.int/files/meetings/seminar/application/pdf/sem_sup1_south_africa.pdf) and a Biofuels Industrial Strategy (December 2007; http://www.info.gov.za/view/DownloadFileAction?id=77830) to map the country's progress and implementation of renewable energy. The South African Constitution (Act No. 108 of 1996) requires that: "Government establish a national energy policy to ensure that national energy resources are adequately tapped and delivered to cater for the needs of the nation; further, the production and distribution of energy should be sustainable and lead to an improvement in the standard of living of citizens." In addition, South Africa agreed to join the Kyoto Protocol in March 2002 and in so doing committed itself to reducing greenhouse gas emissions levels from those observed in 1990, by 2012. The White Paper on Renewable Energy states that South Africa has substantial renewable energy resources, although these resources are largely underutilised. Transportation fuels comprise approximately 30% of fossil fuel energy used and 70% of the value of total energy used in South Africa. Therefore, the development of a locally-based renewable energy production sector and the technologies associated with these processes, particularly for transportation fuels, would be a key driver in South Africa's economy and long-term sustainability.

1.2.2 Bioethanol as an alternative fuel

Bioethanol is ethanol produced from a biological source, most often plant material (Sun and Cheng, 2002), and is termed renewable as the feedstock can be re-grown or is a by-product produced during the course of normal agricultural practises (Taylor *et al*., 2009). Bioethanol is currently the most promising alternative transportation energy source, and this is supported by the exponential growth in bioethanol technologies and production over a recent seven year period (Fig. 1.1) (Balat and Balat, 2009). The Renewable Fuels Association state that roughly 49.2 billion litres of ethanol was produced globally in 2007 and approximately 85% of this was produced in the USA and Brazil (2007 World Fuel Ethanol Production; http://www.ethanolrfa.org/industry/statistics /). Scientists propose that bioethanol is likely to develop as the principle petrol substitute for road transportation vehicles (Balat and Balat, 2009).

Bioethanol (or ethyl alcohol; C_2H_5OH) is a clear colourless liquid which is biodegradable, has a low level of toxicity and causes minimal environmental pollution (Demibras and Karslioglu, 2007; Balat and Balat, 2009). Bioethanol has a higher octane rating than petrol (Balat and Balat, 2009), which allows for a higher compression ratio and shorter burn time within internal combustion engines. This provides enhanced engine performance in comparison to vehicles which run solely on petrol (Balat, 2007). In addition, blending bioethanol with petrol ensures that the fuel mixture is oxygenated and burns to completion, which in turn helps to reduce polluting emissions (Yoosain and Sorapipatana, 2007). These bioethanol fuel blends are popular in the countries in which bioethanol is produced on a large scale, such as USA and Brazil, as this reduces these countries' dependence on fossil fuel imports (Demibras, 2007). The production of bioethanol is an extensive process which may occur via a sugar fermentation process (Harun and Danquah, 2011), or through the chemical reaction of ethylene with steam (Gray *et al*., 2006; Christensen *et al*., 2008; Li *et al.,* 2009). Currently, the principle sources of fermentable sugars required for ethanol production are derived from plant material, including maize and/or sugar cane (Lin and Tanaka, 2006).

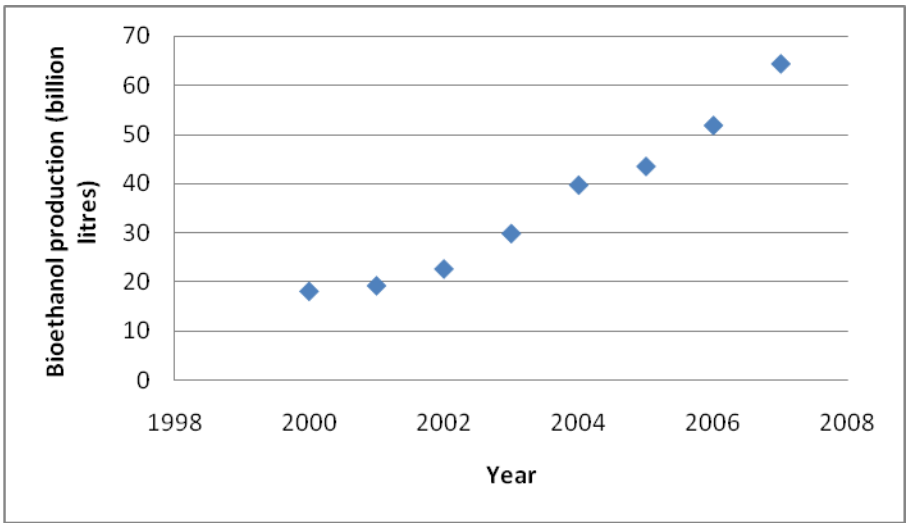


Figure 1.1 The amount of bioethanol produced globally from the year 2000 to 2007 (Balat and Balat, 2009).

1.2.3 Production of Bioethanol

Globally, there is significant interest in the discovery of new and improved plant-based carbohydrate sources for the production of bioethanol (Balat and Balat, 2009). This is a consequence of governments world-wide coming to the realisation that domestic production and use of bioethanol for fuel can decrease their country's dependence on foreign oil imports, reduce trade deficits, create jobs in rural areas, reduce air pollution and reduce carbon dioxide build up and as a result combat climate change (Margeot *et al*., 2009a).

1.2.3.1 First generation bioethanol

First generation bioethanol production relies on the use of plant biomass containing a high concentration of simple sugars (Fig. 1.2) (Gray *et al*., 2006; Balat and Balat, 2009) as a fermentation feedstock for ethanol production by variants of *Saccharomyces cerevisiae* (Taylor *et al.,* 2009). In the USA ethanol production is based on maize as a feedstock, while in Brazil the primary feedstock is sugar cane (Taylor *et al.,* 2009).

Therefore, first generation bioethanol production relies upon crops cultivated for human and animal consumption as the fermentation feedstock (Lin and Tinaka, 2006). The use of food crops for bioethanol production has been severely criticised, as it potentially diverts agricultural production away from the human food chain, which may in turn lead to food shortages and price increases in commodity food stuffs (Lin and Tinaka, 2006). Within developing countries, these food groups form the staple diets of most of the population. Alternative sustainable feedstocks for bioethanol production therefore needs to be investigated and developed (Gever *et al.,* 1985; Margeot *et al.,* 2009b).

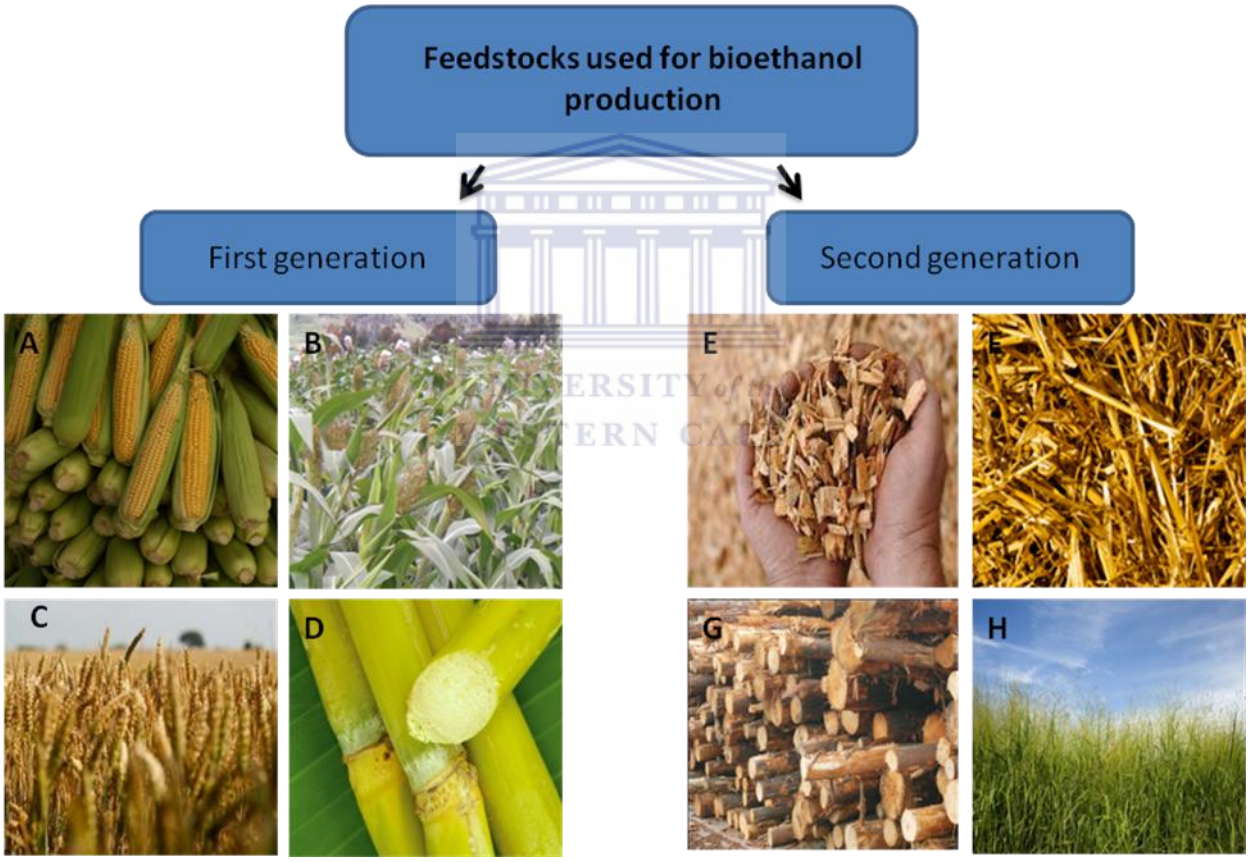


Figure1.2 First and second generation feedstocks utilised for the production of bioethanol. First generation feedstocks may include maize (*Zea mays*) (A), sorghum (*Sorghum bicolor*) (B), wheat (*Triticum aestivum*) (C), sugarcane (*Saccharum officinarum*) (D); and second generation feedstocks may include wood chips (E), straw (F), deciduous trees (G) and switch grass (*Panicum virgatum*) (H)

1.2.3.2 Second generation bioethanol

There is a significant resource of fermentable sugars within lignocellulosic biomass residues which may be utilised as a second generation feedstock (Taylor *et al.,* 2009). Lignocellulosic biomass is essentially surplus plant biomass, mostly generated as waste products during agricultural activities (Naik *et al*., 2010). Second generation bioethanol production processes are thought to be more energy efficient than first generation production, mainly because of the flexibility regarding the type of biomass feedstock that is utilised and the quantity of waste biomass produced annually on a global scale (Naik *et al.*, 2010). Taylor *et al.* (2009) suggest that despite first generation bioethanol continuing to dominate in the short to medium term, on-going development and commercialization of second generation processes will begin to supplement first generation production within the next 20-30 years, before possibly dominating ethanol production.

Feedstocks for second generation ethanol production can be derived from waste material generated from forestry and agriculture, and may include the residual non-food parts of plants such as stems, leaves and husks (Hendriks and Zeeman, 2008; Kumar *et al.*, 2008). Furthermore, various non-food crops such as *Panicum virgatum* (switch grass), *Miscanthus giganteus* (perennial grass), the deciduous tree *Populus nigra* and cereal crops (Lin and Tinaka, 2006; Kumar *et al*., 2008; Balat and Balat, 2009) as well as industrial organic waste such as wood chips and pulp from fruit pressing can serve as feedstocks (Fig. 1.2) (Kumar *et al*., 2008).

To the best of our knowledge there is no locally-based industrial-scale bioethanol industry. However, the conversion of lignocellulosic biomass to bioethanol has the potential to positively impact the South African economy (as outlined in section 1.2.2). The success of such an industry within SA is largely dependent on the development of low cost conversion technologies. Higher education facilities within SA are currently conducting a number of research programs aimed at the development of a bioethanol production platform in SA. Some of these include research at the University of

Stellenbosch into the potential use of *Saccromyces cerevisiae* as a candidate for the consolidated bioprocessing of cellulose to ethanol (den Haan *et al*., 2006; Van Zyl *et al*., 2010; Ilmén *et al*., 2011). Furthermore, researchers at Rhodes University are focusing on the characterisation of a cellulosome for enzymatic saccharification of lignocellulose produced by a strain of *Bacillus licheniformis* (Van Dyk *et al*., 2009; Van Dyk *et al*., 2010) as well as optimising and assessing the effects of various pre-treatment strategies on the biomass conversion of locally available lignocellulosic substrates (Beukes *et al*., 2008; Beukes *et al*., 2010; Beukes *et al*., 2011; Dredge *et al*., 2011). A large collaborative Technology Innovation Agency funded program, between academic and industrial researchers from the Biosciences division of the Council of Scientific and Industrial Research (CSIR) and the Universities of Stellenbosch (US) and the Western Cape (UWC), aimed at isolating and characterising novel lignocellulosic enzymes that can be utilised for enzymatic saccharification of biomass for bioethanol production is also ongoing. This project aimed to employ a combination of metagenomic (CSIR and UWC) and metatranscriptomic (US) technologies to identify the enzymes of interest (Rashamuse *et al.,* 2009a; Rashamuse *et al*., 2009b; Rashamuse *et al*., 2012a; Rashamuse *et al.*, 2012b; Rashamuse *et al.*, 2012c). The discovery of novel enzymes with improved process characteristics, such as turnover and thermostability, for possible inclusionin thermogenic bioethanol production processes may help to improve enzymatic saccharification of lignocellulosic biomass for locally-based bioethanol production. The proposed benefits of a thermogenic bioethanol production process will be discussed in section 1.4.21.

1.3 Lignocellulosic biomass

Lignocellulose is defined as the three dimensional polymeric composite formed by plants as their structural material (Fig. 1.3) and consists of variable amounts of cellulose, hemicellulose and lignin (Table 1.1) (Fuduka *et al.,* 2008; Hendriks and Zeeman, 2008; Kumar *et al.,* 2008).

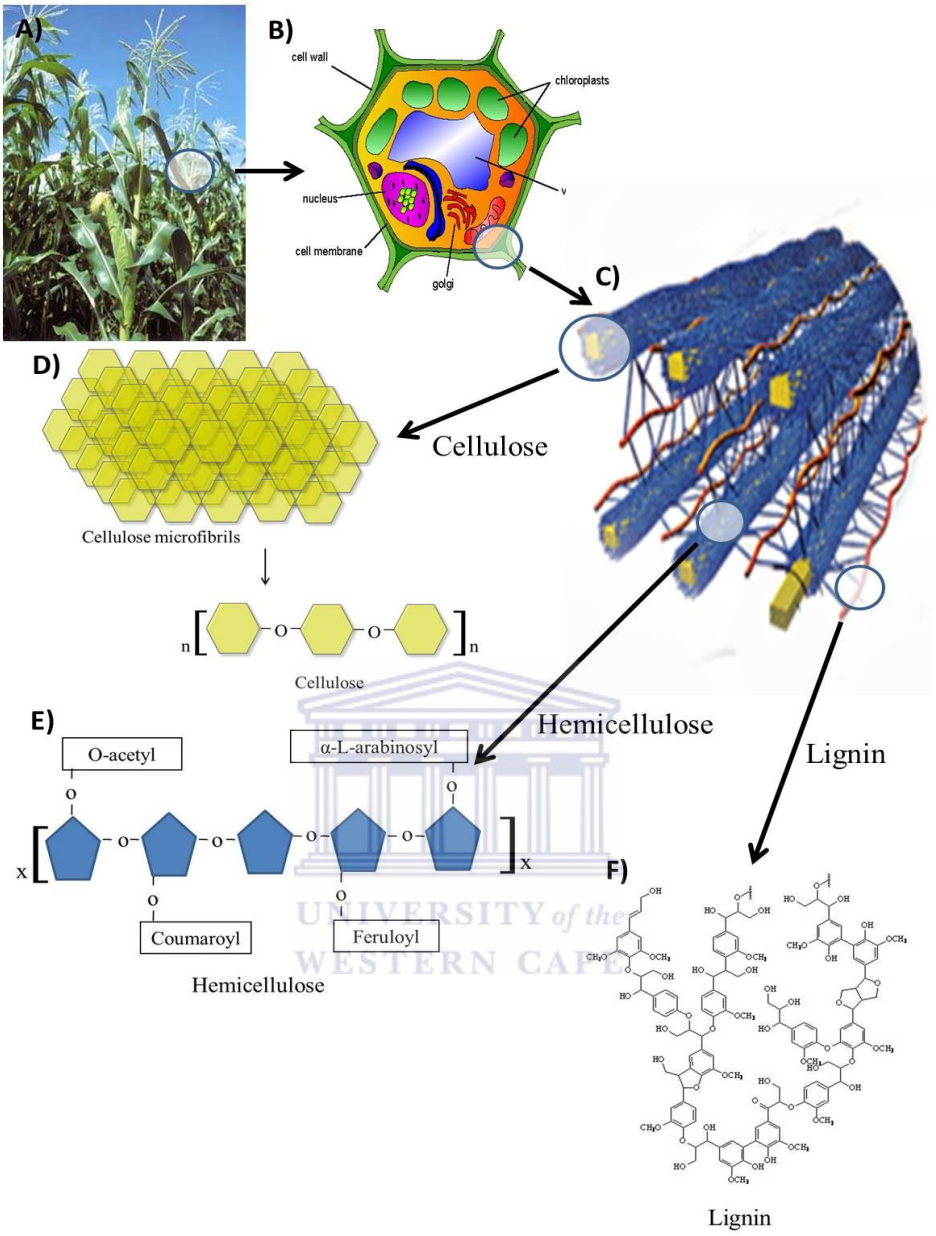


Figure 1.3 Graphical representation of the composition of lignocellulosic materials. Depicting maize as an example of plant biomass (A), the location of lignocellulose within the plant cell wall (B), the lignin-carbohydrate complex that comprises lignocellulose (C) and the three main components of lignocellulose, namely cellulose microfibrils made up of cellulose polymers (D), hemicellulose polymers including some of the predominant accessory side-chain groups (E) and lignin (F) (Figure courtesy of Rob Huddy).

Table 1.1 The relative composition of lignocellulose, within different plant sources (Jørgensen *et al*., 2007).

Lignocellulosic material	Cellulose (%)	Hemicellulose (%)*	Lignin (%)*
Hardwood stems	40-55	24-40	18-25
Softwood stems	45-50	25-35	25-35
Nut shells	25-30	25-30	30-40
Maize cob	45	35	15
Paper	85-99	0	0-15
Wheat straw	30	50	15
Rice straw	32	24	18
Sorted refuse	60	20	20
Leaves	15-20	80-85	0
Cotton seed hairs	80-95	5-20	0
Newspaper	40-55	25-40	18-30
Waste paper from chemical pulps	60-70	10-20	24-29
Fresh bagasse	33	30	19
Solid cattle manure	2-5	1-3	3-6

1.3.1 Cellulose

Cellulose is the most abundant bio-polymer found within terrestrial ecosystems on earth (Bielecki *et al.,* 2005). It consists of D-glucose subunits linked by β-1, 4 glycosidic bonds (Fig. 1.4) (Bayer *et al.*, 1998). Linear cellulose polymers have been estimated to contain approximately 5,000 glucose units (Jullander, 1947; Stamm, 1977; Van der Hart and Atalla, 1984). In plants, cellulose exists in two states, either in a highly organised crystalline form or a highly disorganised amorphous state (French *et al*., 2002; Walker, 2006). Cellulose strands are grouped together to form densely packed cellulose microfibrils, which are also referred to as cellulose bundles (Hohl *et al*., 1968; Bhatnagar and Sain, 2005). These cellulose fibrils are weakly bound together by hydrogen bonds and, within lignocellulose, are in-cased in a network of hemicellulose polymers (Fig. 1.3) (Laureano-Perez *et al*., 2005; Hendriks and Zeeman, 2009).

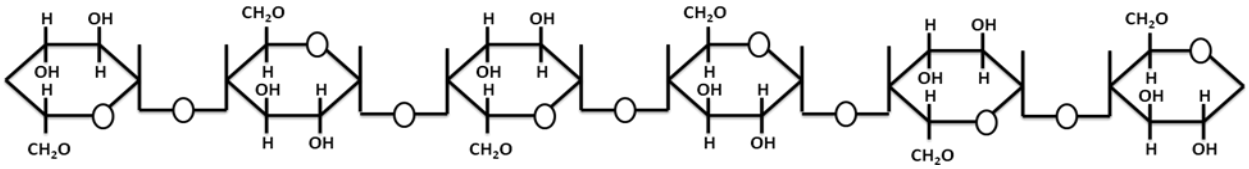


Figure 1.4 Showing a fraction of the structure of the repeat unit of cellulose constituting a continuous polymer (Reilly, 1991).

1.3.2 Hemicellulose

Hemicellulose is highly variable in structure and differs from cellulose as it may consist of a mixture of polymers, often comprising pentose (D-xylose and L-arabinose) and hexose (D-galactose, L-galactose, D-mannose, L-fructose and L-rhamnose) sugars, and sugar acids (ferulic acid, D-glucuronic acid, 4-O-methyl-D-glucoronic acid and D-galacturonic acid) (Fig. 1.5) (Peng *et al*., 2009; Scheller and Ulvskov, 2010). The dominant component of hemicellulose from hardwood sources and agricultural plants is xylan, while hemicellulose from softwood contains mainly glucomannan (Fukuda *et al*., 2009). Hemicellulose has a lower molecular weight than cellulose and is generally highly branched with short side chains comprising different sugar residues, such as pentose and hexose sugars and sugar acids (Fig. 1.5) (Kumar *et al*., 2009; Agbor *et al*., 2011). Hemicellulose contributes towards the rigidity of the lignin-carbohydrate structure, as it provides a connection between the lignin (section 1.3.3) and the cellulose microfibrils which is facilitated by the presence of side-chain groups on the xylan backbone of hemicellulose (Hendriks and Zeeman, 2009; Scheller and Ulvskov, 2010).

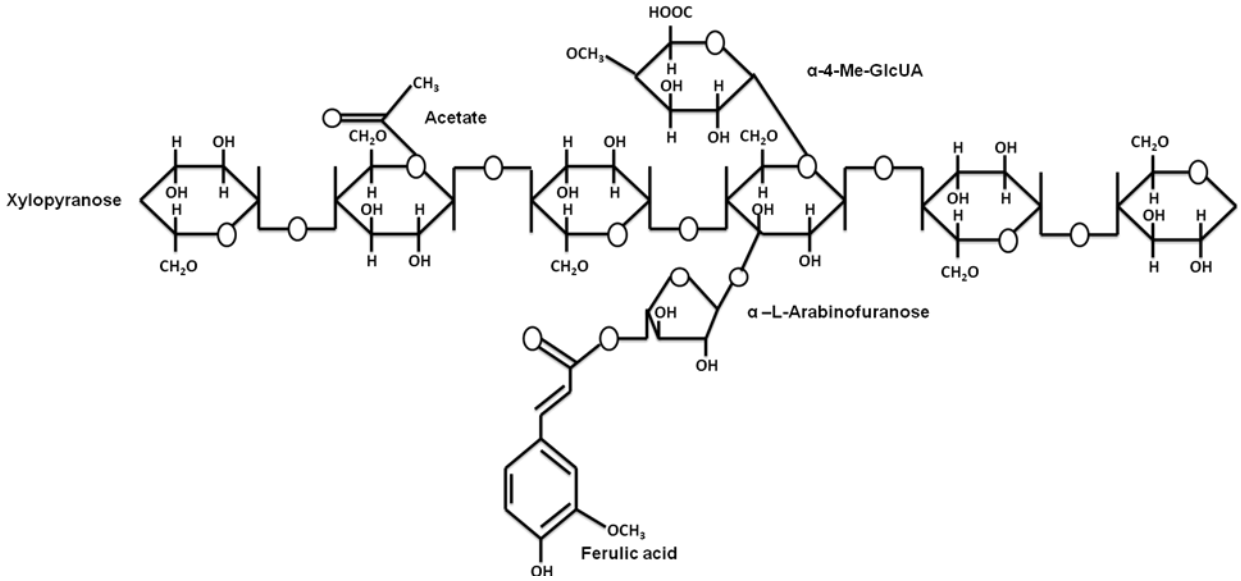


Figure 1.5 A schematic diagram showing the general structure of hemicellulose and the various side-chain linkages found within the branching polymer (adapted from Grohmann *et al*., 1985).

1.3.3 Lignin

Together with cellulose and hemicellulose, lignin is one of the most abundant biopolymers in nature (Fukuda *et al*., 2009; Hendriks and Zeeman, 2009). It is an amorphous heteropolymer consisting of three different phenyl propane units (Fig. 1.6), namely p-coumaryl, coniferyl and sinapyl alcohol, held together by carbon or ether linkages (Hendriks and Zeeman, 2009; Agbor *et al*., 2011). Lignin plays a number of important roles in the plant cell wall as it provides structural support, resistance against microbial attack and protection against oxidative stress (Fengel and Wegener, 1984; Hendriks and Zeeman, 2009). Lignin is the most recalcitrant component of the plant cell wall (Sánchez, 2009; Sannigrahi *et al*., 2010), and its presence decreases the availability of the sugar polymers contained within the cellulose and hemicellulose during enzymatic hydrolysis of lignocellulosic biomass. Lignin increases the recalcitrance of biomass feedstocks, by reducing the surface area available to enzymatic penetration and, therefore, enzymatic activity upon the carbohydrate polymer (Haug, 1993). Due to the recalcitrance and complexity of lignocellulose, a pre-treatment

step is required to physically and/or chemically disrupt the complex structure of lignocellulose (pre-treatment is discussed in section 1.4.1). This facilitates increased access to the sugar polymers contained within the cellulose and hemicellulose portions, before enzymatic saccharification of the complex sugars to monosaccharides which serve as the fermentable sugar substrates for the microbial production of bioethanol.

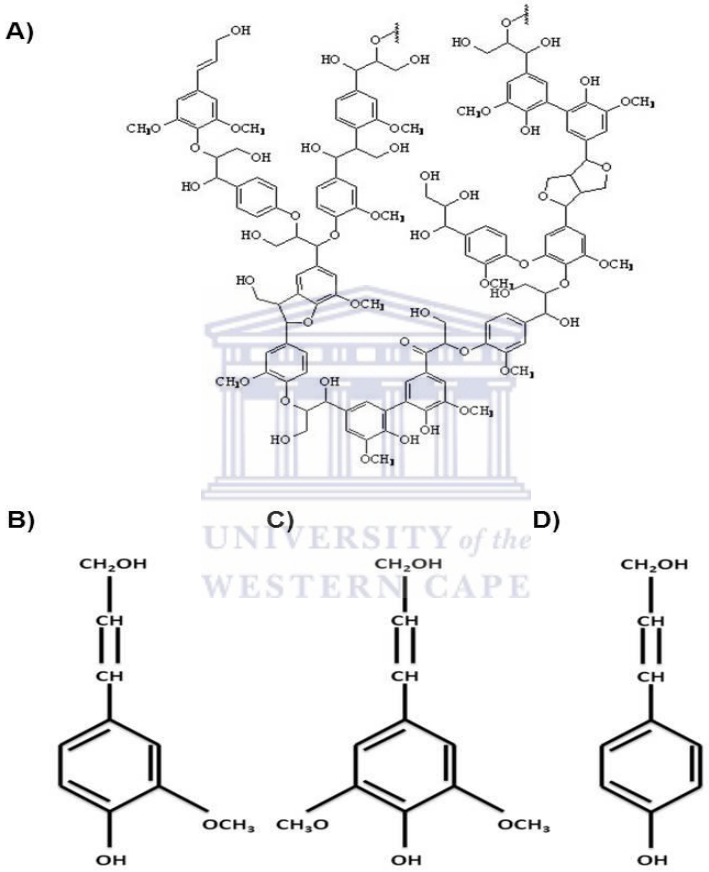


Figure 1.6 Diagrammatic representation of the complex structure of lignin (A) and the three different phenyl propane units within lignin, namely coniferyl alcohol (B), sinapyl alcohol (C) and p-coumaryl alcohol (D) (Adapted from Chakar and Ragauskas, 2004).

1.4 Lignocellulosic ethanol production

The production of second generation bioethanol from lignocellulosic material, can be divided into five key processes, namely pre-treatment, enzymatic saccharification, microbial fermentation, product separation and post-treatment of the liquid fraction (Fig. 1.7) (Margoet *et al*., 2009).

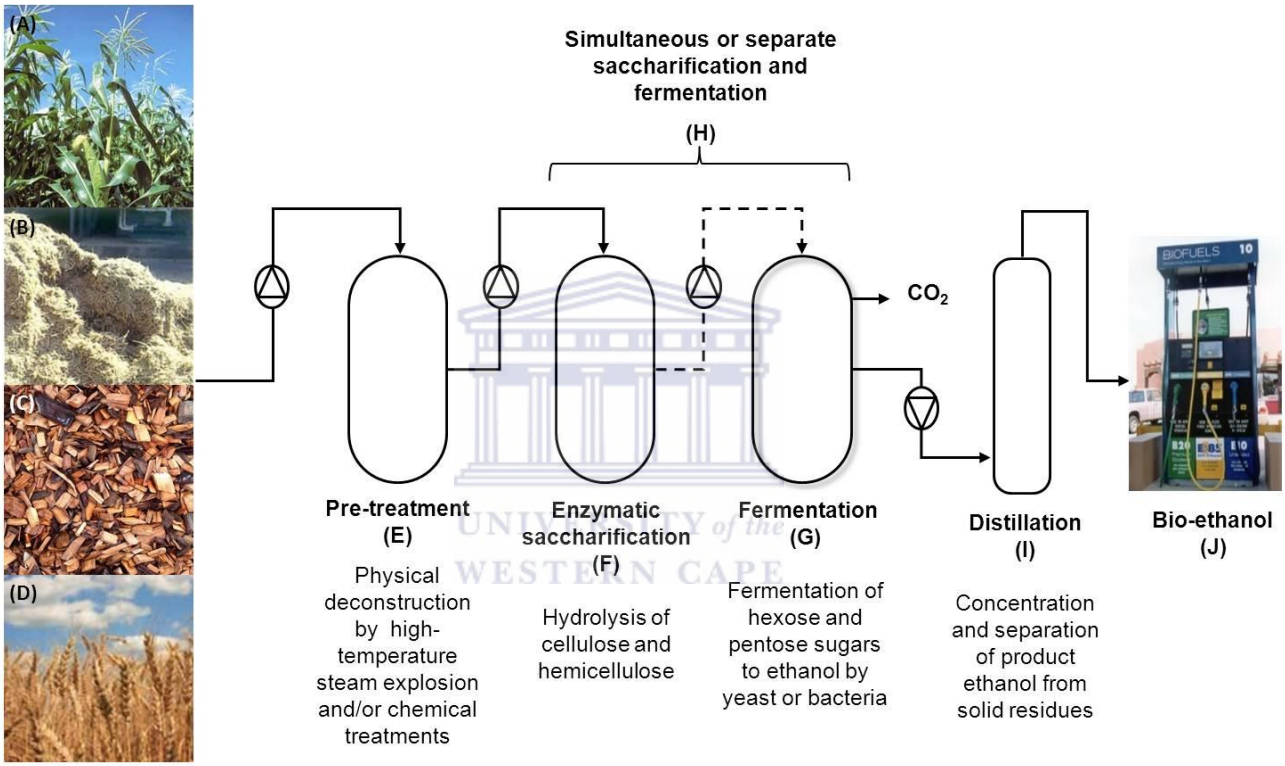


Figure 1.7 Graphical representation of the major steps involved in the bioethanol production cycle. Lignocellulosic feedstocks, such as maize husks (A), sugar cane bagasse (B), wood chips (C) or wheat husks (D) are pre-treated (E). Thereafter enzymatic saccharification (F) liberates fermentable sugars which are subsequently fermented to ethanol by microorganisms within biorefineries (G). The simultaneous saccharification and fermentation, by an appropriate microorganism, would be the ideal bioethanol production process (H). Thereafter, distillation of the ethanol is conducted to separate ethanol from other residues (I) before the final product, ethanol, is ready to be blended with petrol and provided to fuel stations (J) (Adapted Viikari *et al.,* 2007 and reproduced with permission from R. Huddy).

1.4.1 Pre-treatment

A major barrier to the second-generation production of ethanol is the recalcitrance of lignocellulosic biomass to enzymatic hydrolysis (see section 1.3.3) (Balat, 2011). A technical approach to overcome or reduce the recalcitrance associated with lignocellulose involves an initial pre-treatment step in which biomass is mechanically and/or chemically disrupted to decrease the lignin content and increase the accessibility of enzymes to the carbohydrate polymers for subsequent degradation (Fig. 1.7) (Demibras, 2003). Pre-treatment, therefore, improves the rate at which fermentable sugars are produced and the total sugar yield during the enzymatic hydrolysis step, while removing potential inhibitors, like lignin and phenolic compounds, which may impede microbial fermentation and ethanol production (Hendriks and Zeeman, 2008). Comparative studies have shown that pre-treatment of lignocellulose increases the sugar yield after enzymatic hydrolysis by more than 70% compared to hydrolysis of lignocellulose without initial pre-treatment (Hamelinck *et al*., 2006). Pre-treatment may be divided into different types, of which the major categories are chemical, mechanical, heat and pressure treatments, and/or a combination of these methods (Kumar *et al.,* 2009).

Mechanical pre-treatment is performed to reduce the particle size of the biomass by milling, chipping or grinding of the feedstock (Kumar *et al.,* 2009) and requires the use of expensive machinery and skilled operators. Chemical pre-treatments may include the treatment of lignocellulose with an acid or alkaline solution. Acid pre-treatment aims for high yields of sugars from lignocellulosic biomass and includes the use of either sulphuric acid (Parajo *et al*., 1993) hydrochloric acid (Kurakake *et al*., 2005), peracetic acid (Teixeira *et al*., 1999), nitric acid (Brink, 1993) or phosphoric acid (Hussein *et al*., 2001). Alkaline pre-treatment has the advantage of being carried out at ambient temperature with low pressure, however; it requires prolonged processing at these conditions thereby decreasing the rate of production (Silvertein, 2004).

Certain pre-treatment methods require biomass to be heated to extreme temperatures and pressures to reduce the complexity of the substrate (Kumar *et al.,* 2009). One such method is steam explosion, during which the biomass is treated with saturated steam at high pressure (Kumar *et al.,* 2009). Steam explosion results in hemicellulose degradation and lignin transformation and, therefore, increases the potential for enzymatic cellulose hydrolysis (Grous *et al*., 1986; Ballesteros *et al*., 2000; Cara *et al*., 2008; Kumar *et al.,* 2009).

1.4.2 Enzymatic saccharification

As outlined above, during the bioethanol production process the biomass undergoes enzymatic hydrolysis after the pre-treatment step. Carbohydrate polymers are converted to their simple monomeric sugars, such as hexoses and pentoses (Howard *et al.,* 2003), by a cocktail of lignocellulosic enzymes (Fig. 1.7) (Kumar *et al.,* 2009). The cost associated with acquiring lignocellulosic enzymes for inclusion during this process remains a crucial barrier to the commercialisation of second generation ethanol production (Merino and Cherry, 2007). Worldwide, there is a growing interest in research towards the identification of new and improved enzymes that can be employed in this step of the process (Banerjee *et al.,* 2010), or, the development of microorganisms capable of simultaneous saccharification and fermentation (SSF) (Fig.1.7) (Taylor *et al*., 2009). Industrial production of ethanol, using second generation processes, are currently making use of commercial enzyme preparations such as the liquid cellulase preparation of *T. reesei*, known as Celluclast® and produced by Novazyme (Sannigrahi and Ragauskaus, 2010). This enzyme preparation is used in numerous industrial scale mesophilic processes. However, the use of robust thermostable lignocellulolosic enzymes, able to perform hydrolysis at temperatures above 60°C, for inclusion in second generation lignocellulose hydrolysis is currently under investigation. Using thermostable enzymes and performing enzymatic treatments at elevated temperatures may reduce some of the energy inputs, and therefore production costs, required to reduce the temperature of steam exploded feedstocks following pre-treatment, before hydrolysis can be performed as well as the costs

associated with maintaining large reactors at mesophilic temperatures (Liang *et al*., 2009; Maki *et al*., 2009; Brumm *et al*., 2011; Geng *et al*., 2011; Ilmberger *et al*., 2011).

1.4.2.1 Mesophilic versus thermophilic saccharification processes

The enzymatic hydrolysis of lignocellulose at elevated temperatures has a number of advantages over processes operated under mesophilic conditions (Turner *et al.,* 2007). The most important of these are that under thermophilic conditions the solubility of polymeric substrates is significantly enhanced, which increases the enzyme(s) penetration and substrate accessibility. Thermostable enzymes also have higher specific activities compared to their mesophilic counterparts, which results in the use of less enzyme during hydrolysis (Viikari *et al*., 2007). These enzymes also display enhanced stability and flexibility which may improve enzyme performance within the hydrolysishave enhanced tolerance to organic solvents, reduced risk of microbial and phage contamination, and minimal loss of activity during processing (Turner *et al.,* 2007).

Extremophiles are defined as microorganisms that can survive under conditions that from a human viewpoint are clearly extreme (Gerday and Glansdorf, 2007). Amazingly, these microorganisms do not merely tolerate these severe habitats, but thrive and in many cases require one or more extreme physiological conditions in order to survive and/or remain viable (Gerday and Glansdorf, 2007). The isolation and characterisation of microorganisms from extreme environments is of physiological, ecological and biotechnological interest as the enzymes that these microorganisms produce and enable these extremophilic organisms to function in these harsh ecological niches may be able to catalyse reactions under harsh industrial conditions (Zhang *et al*., 2006). The optimal growth temperature of thermophiles ranges between 55 and 65°C, with a minimum temperature requirement of approximately 45°C (Blumer-Schuette *et al*., 2008), while hyperthermophiles have a temperature optima above 65°C. Thermostable enzymes have many applications in various industries, including the textile, detergent,

paper and food processing industries (Gerday and Glansdorf, 2007). To meet the requirements of thermophilic industrial process, novel robust biocatalysts need to be discovered and their suitability accessed (Liang *et al*., 2009; Maki *et al*., 2009; Brumm *et al*., 2011; Geng *et al*., 2011; Ilmberger *et al*., 2011).

1.4.3 Ethanol production by microbial fermentation

Monomeric hexose and pentose sugars, generated during the enzymatic hydrolysis of lignocellulose, can be subsequently fermented to ethanol by certain strains of bacteria and yeasts (Dien *et al.*, 2003). Thus far, *Saccharomyces cerevisiae* has been the preferred organism for fermentative ethanol production. However, it lacks the ability to utilise pentose sugars, such as xylose and arabinose (Margeot *et al*., 2009a). During the 1970's and 1980's the gram-negative bacterium *Zymomonas mobilis* was identified as a possibly superior ethanol production strain in comparison to *S. cerevisiae*, however, it was later established that it only ferments glucose, fructose and sucrose (Dien *et al*., 2003). One of the limitations most organisms exhibit is that the ethanol produced during the fermentation process acts as a growth inhibitor for most microorganisms, lowering the yield of bioethanol produced (Dien *et al.*, 2003; Fuduka *et al.*, 2008). Engineering an *E. coli* strain capable of ethanol production was showing to be largely unsuccessful, as *E. coli* cultures are prone to shear sensitivity during the industrial processes, as well as having a narrow pH range for growth (Dien *et al*., 2003). The lack of industrially suitable microorganisms for the conversion of biomass to bioethanol is currently viewed as a major technical bottleneck in the second generation bioethanol production process.

1.4.4 Simultaneous saccharification and fermentation

As the conditions during enzymatic hydrolysis and ethanol fermentation are relatively similar, it would be theoretically feasible to perform these steps simultaneously in one vessel, provided that one could obtain a microorganism cable of producing the required lignocellulosic enzymes and fermenting the resulting sugars to ethanol (Kádár *et al.,*

2004). The process of simultaneous saccrification and fermentation (SSF) offers the advantages of enhanced production rates, high ethanol yields and lowered cost input compared to other processes (Wyman *et al*., 1992). However, SSF has the inherent problem of possible mismatches between the optimum temperature for hydrolysis and that required for microbial growth and fermentation. Research on the identification and improvement of thermophilic yeasts and bacteria that are able to produce ethanol at higher temperatures is therefore being conducted (Wu and Lee, 1998). The use of thermostable, metabolically engineered production strains such as *Geobacillus thermoglucosidasius*, *Thermoanerobacterium saccrolyticum* and *Thermoanerobacter mathranii* with enhanced versatility and higher degrees of ethanol tolerance are currently being investigated as possible host systems for SSF (Taylor 2009).

1.4.5 Product purification

As ethanol is more volatile than water, the ethanol produced during microbial fermentation is recovered from the fermentation broth by distillation (Fig. 1.7) (Kumar *et al.,* 2008; Balat, 2011). Distillation is a concentrated counter current vapour and liquid mass transfer system in which liquid descends by gravity through directed channels (Madson and Lococo, 2000). Distillation is readily up-scalable and, therefore, ideally suited for use in industrial-scale production of bioethanol. Furthermore, conducting bioethanogenesis under thermophilic conditions would significantly reduce the energy input required for the distillation of ethanol, compared to a mesophilic production process, as the fermentation broth is already at a temperature of approximately 60°C leading to greatly simplified purification (Taylor *et al.,* 2009).

1.5 Lignocellulolytic enzymes

Due to the complex composition of lignocellulosic biomass, the action of three enzyme classes, namely cellulases (EC 3.2.1.4), hemicellulases (EC 3.2.1.8) and lignin-degrading enzymes (EC 1.11.1.14), are required for its enzymatic hydrolysis (Howard *et al.,* 2003). These carbohydrate active enzymes, also known as glycoside hydrolases

(GHs) (EC 3.2.1.x), catalyse the hydrolysis of glycosidic bonds between two or more carbohydrate moieties and/or between a carbohydrate and a non-carbohydrate moiety (Henrissat and Davies, 2000). The classification of carbohydrate active enzymes is based upon the recommendations specified by the International Union of Biochemistry and Molecular Biology (IUBMB) which assigns an EC number for a given enzyme. Furthermore, GHs are grouped and classified into families based on their amino acid sequences (Henrissat, 1991;Henrissat and Davies, 2000). There are currently 130 GH families as described on the carbohydrate active enzymes, CaZY, database (www.cazy.org). These GH families are grouped according to protein sequence and folding similarities, therefore, each GH family contains proteins that are related by sequence and by putative structure. This allows a number of useful predictions to be made, based on a GH protein's sequence, since the catalytic machinery and molecular mechanism of the GH families (Gebler *et al.,* 1992) as well as the geometry around the glycosidic bond (Davies and Henrissat, 1995) are well conserved. Generally, the catalytic mechanism of a GH enzyme involves hydrolysis of the glycosidic bond which is catalyzed by two amino acid residues of the enzyme, namely a general acid (proton donor) and a nucleophile or base (Davies and Henrissat, 1995). Depending on the spatial position of these catalytic residues, hydrolysis takes place through either the overall retention or the inversion of the anomeric configuration (Wilson, 2009). This catalytic mechanism is generally conserved within a particular GH family. Some of the key enzymes implicated in the carbohydrate metabolism of microorganisms are GHs. Therefore, microbes represent an excellent source of GH enzymes, and these enzymes are often exploited for use in industrial processes (Henrissat, 1991). The major component of lignocellulosic biomass is cellulose (Table 1.1), therefore, cellulolytic enzymes are considered to be the core enzyme class required for efficient lignocellulose decomposition.

1.5.1 Cellulases

Cellulases (EC 3.2.1.4) are a class of enzymes which hydrolyse the β-1, 4 glycosidic linkages of cellulose (Fuduka *et al.*, 2008). These enzymes are represented in 13 of the 130 GH families identified to date and are able to hydrolyse oligosaccharides and/or polysaccharides (Bayer *et al.,* 1998). Cellulases are organised into clans based on common protein folds and conservation of catalytic residues which may represent global function (Duan and Feng, 2010). The structural complexity and rigidity of cellulosic materials has driven the natural evolution of a significant number of cellulases with various catalytic and structural properties (Bayer *et al.,* 1998). In general, cellulases are modular proteins with at least two distinct modules, namely a catalytic domain and a carbohydrate binding module (CBM) (Linder and Teeri, 1997). In some cases, cellulolytic enzymes may contain multiple catalytic domains and CBMs (see Fig. 1.8 for examples of the variability of cellulase domain architectures). Examples of enzymes comprising multiple catalytic domains and CBMs include the cellobiohydrolase *cbh2* from *Tricoderma reesei* (Teeri *et al*., 1987), the endoglucanase *CenB* from *Cellulomonas fimi* (Meinke *et al*., 1991) and the β-glucosidase 3B produced by *Thermotoga neapolitana* (Pozzo *et al*., 2010).

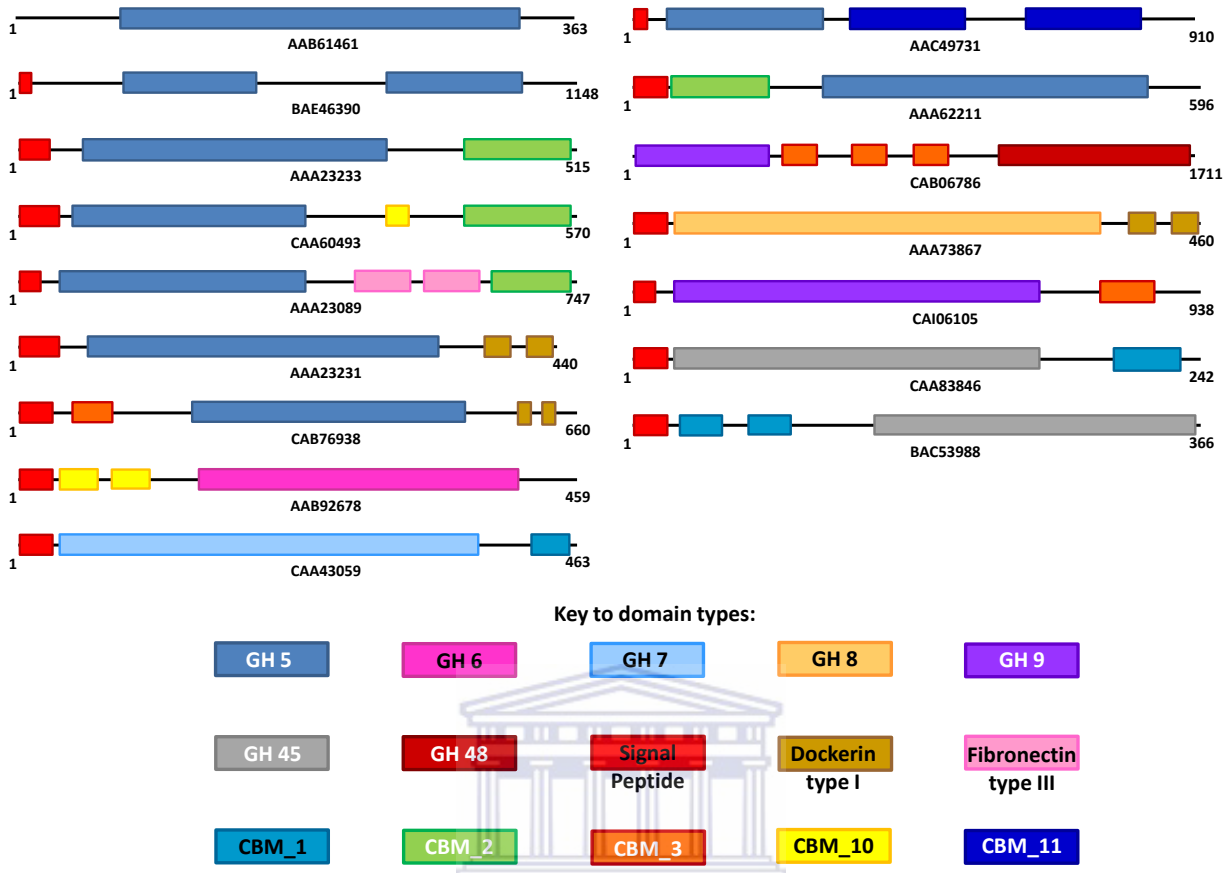


Figure 1.8 Cellulase conservation and diversity of domain architectures (adapted from Sukharnikov *et al*., 2011). Conserved domains within sequences are shown as coloured boxes corresponding to the key provided. Domain sizes are drawn to the scale of the complete protein sequence. NCBI accession numbers of proteins are given below their graphical representation.

1.5.1.1 Carbohydrate Binding Modules

A CBM is broadly defined as a domain within a carbohydrate-active enzyme that has a distinctive structural fold and carbohydrate-binding capability (Boraston *et al.*, 2001; Boraston *et al*., 2004; Shoseyov *et al*., 2006). CBMs have been identified within protein sequences derived from a wide variety of sources, including bacteria, archaea, eukaryotes and even viruses (Hildén and Johansson, 2004). CBMs vary in size, structure and the ligand they bind, and are connected to the catalytic module via highly glycosylated flexible polypeptide linkers (Fig. 1.9 and 1.10) (Boraston *et al*., 2004;

Hildén and Johansson, 2004; Shoseyov *et al.*, 2006). CBMs may be located at either the N- or C-terminus within the paternal protein and occasionally occur centrally (Fig. 1.8) (Shoseyov *et al*., 2006). CBMs are broadly classified into three types on the basis of their substrate binding (Table 1.3). These types are either surface binding (Type A), glucan chain binding (Type B) or small sugar binding (Type C) (Boraston *et al.,* 2004; Shoseyov *et al.,* 2006). CBMs are further classified into 7 fold families (Table 1.3), of which the β-sandwich (Fold family 1) is the most dominant fold-type characterised to date (Boraston *et al*., 2004; Shoseyov *et al.*, 2006). CBMs from fold family 1 have characteristic folds made up of two β-sheets, each of which consists of 3 to 6 anti-parallel β-strands, and in most cases have at least one bound metal ion. CBMs are further divided into 39 families on the basis of amino acid sequence similarity (Table 1.4). Within each family, the ligand bound is very specific; however, between families the ligand specificity is very diverse.

CBM protein domains have been identified in both hydrolytic enzymes, such as cellulases and xylanases, and non-hydrolytic proteins, such as cellulosome scaffolding sub-units and as they exist as modules within larger enzymes this differentiates them from other non-catalytic sugar binding proteins (Shoseyov *et al*., 2006). CBMs are proposed to fulfil three primary roles which assist with the activity of their respective catalytic domain(s). Firstly, CBMs maintain the proximity of the enzyme to the substrate, resulting in increased concentration of the enzyme complex on the surface of the substrate and enhanced substrate degradation (Fig. 1.9 and 1.10). Secondly, CBMs have been shown to improve enzyme action by lengthening the period of contact between the catalytic domain and substrate and target the enzyme complex to specific substrates and even distinct regions of the substrate (Fig. 1.9 and 1.10) (Bolam *et al*., 1998; Carrard *et al*., 2000; Hervé *et al*., 2010). Finally, CBMs are also thought to contribute to the activity of the enzyme by destabilizing the cellulose substrate, and in so doing help to make the substrate more accessible to the catalytic portion of the enzyme (Levy and Shoseyov, 2002; Boraston *et al.,* 2004). Studies where the CMB component of individual cellulases were removed, showed a reduction in the hydrolytic activity of the catalytic module on insoluble substrates, however, activity on soluble

amorphous substrates was unaffected (Irwin *et al*., 1993; Mansfield *et al*., 1999; Raghothama *et al*., 2000).

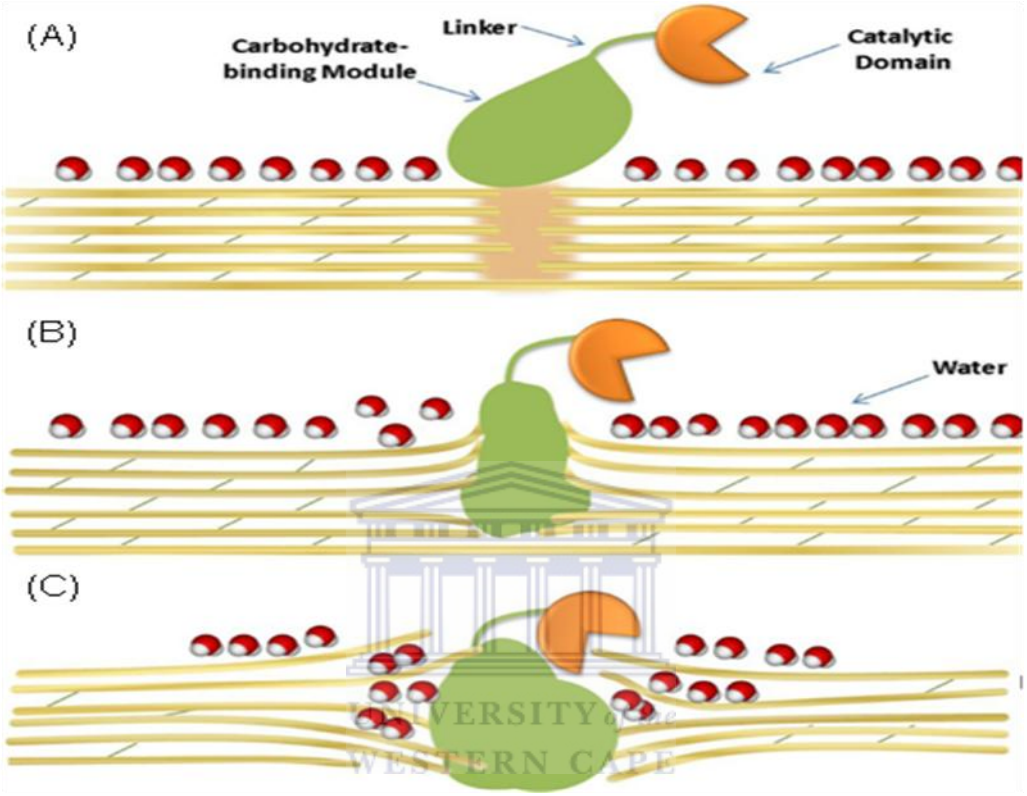


Figure 1.9 Schematic representation of amorphogenesis of cellulose fibres mediated by the carbohydrate-binding module (CBM) of cellobiohydrolase I (CBHI). It is generally accepted that cellulose hydrolysis occurs on the surface of cellulose aggregations. Firstly, the CBM is absorbed into micro-cracks found within the cellulose molecule (A). Secondly, the CBM penetrates the interfibrillar space resulting in dispersion of the tightly packed cellulose fibres. Thirdly, water penetrates the interfibrillar space and cleavage of hydrogen bonds occurs which releases free chain ends on which absorbed enzymes act (Arantes and Saddler, 2010).

Table 1.3 Classification of CBM families into CBM Types (Boraston *et al.,* 2004).

Type	Fold family	CBM families	Ligand specificity
A	1,3,4,5	1, 2a, 3, 5, 10	Crystalline cellulose; chitin (McLean *et al.,* 2000)
B	1	2b, 4,6, 15,17, 20, 22, 27, 28, 29, 34, 36	Large oligosaccharide chains (Pell *et al.,* 2003)
C	1,2,6,7	9,13, 14, 18, 32	Mono-, di- and tri-saccharides (Boraston *et al.,* 2003)

Kraulis *et al.* (1989) was the first author to describe the structure of a CBM, namely CBHI-CBD from the *Trichoderma reesei* cellobiohydrolase I. Since then, a significant number of CBM structures have been resolved across the various CBM types and families. In many cases, the CBM was crystalised in complex with their respective ligands (Boraston *et al*., 2002; Flint *et al*., 2004; Bae *et al*., 2008).

Table 1.4 Classification of CBM families into fold families (Boraston *et al.*, 2004).

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

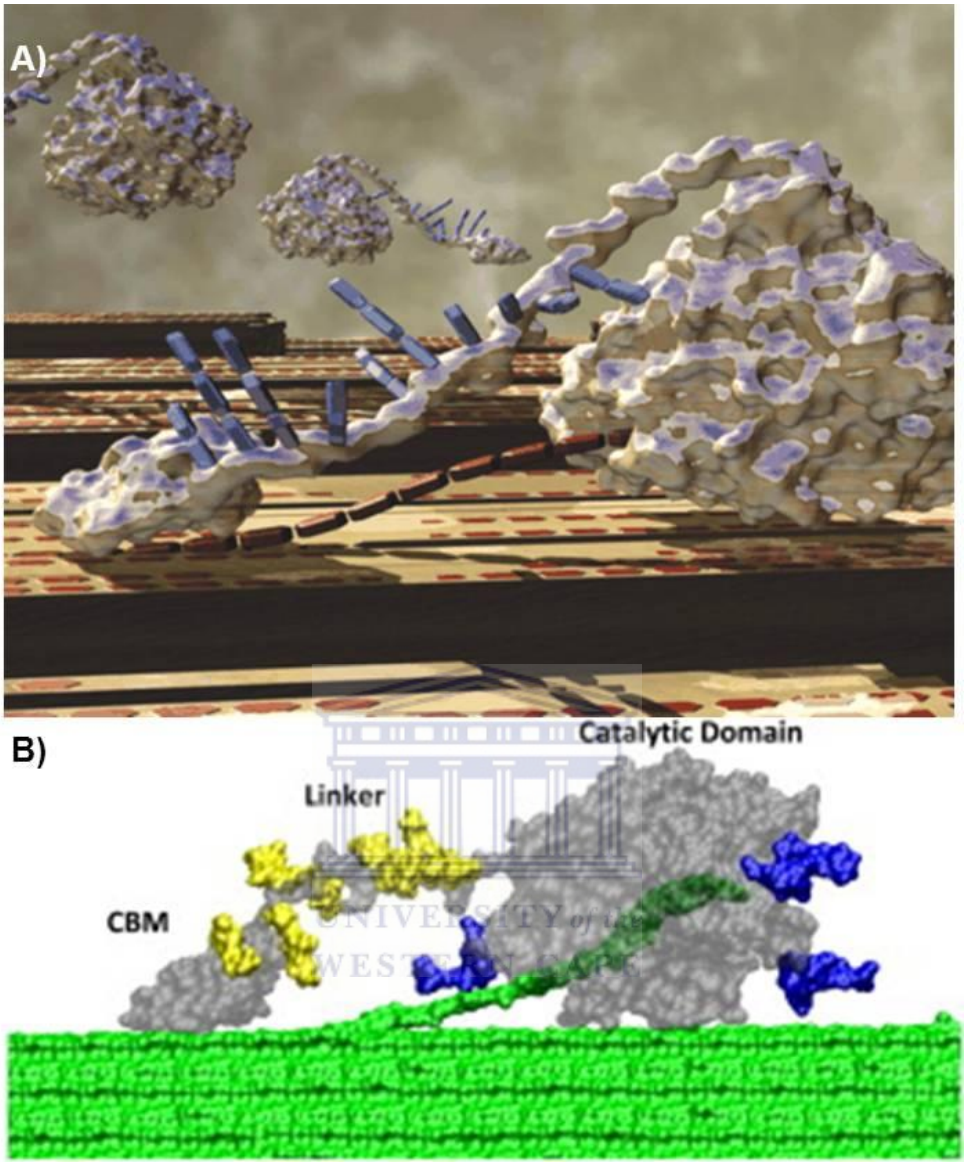


Figure 1.10 Artistic impression of a *T. reesei* exoglucanase (cellobiohydrolase I) hydrolysing a crystalline cellulose substrate (A and B). The CBM is depicted recognizing and binding to the cellulose surface. A single polymer of cellulose is "decrystallized" and directed into the active-site tunnel of the catalytic domain. The CBM and catalytic domain enzyme complex, joined by a polypeptide linker, proceeds along a cellulose chain cleaving one cellobiose unit per catalytic event until the chain ends or the enzyme becomes inactivated (Rouvinen *et al.,* 1990; Vršanská and Biely, 1992; Zhong *et al*., 2009).

. 1.5.1.2 Industrial applications of cellulases

Globally, cellulases are the third largest group of commodity industrial enzymes produced. They have a broad range of applications in various industries, including the textile industry, in which they are used for cotton softening and denim finishing, as well as in the detergent industry where they are used for colour care, cleaning and anti-deposition. In the food industry cellulases assist in the processing of fruit juices and flour, and in the paper and pulp industries they are used for de-inking, drainage improvement and fibre modification (Ibrahim and El-diwany, 2007). The market for commodity cellulase enzyme preparations is expected to expand dramatically over the next decade as these enzymes are likely to be used extensively within the biofuels and fine chemical production industries for enzymatic saccharification of pre-treated lignocellulosic biomass (Zhang *et al*., 2006; Duan and Feng, 2010; Shabed *et al*., 2010).

Cellulases are relatively costly enzymes to produce, and a significant reduction in cost will be important to increase their feasibility for commercial use in bio-refineries (Zhang *et al*., 2006; Shabed *et al.*, 2010; Sun *et al.,* 2010; McMillan *et al*., 2011). The large market potential and the important role that cellulases play in the emerging bio-energy and bio-based product industries provide a great motivation to develop superior cellulase preparations for biomass hydrolysis (Zhang *et al*., 2006). Thermally active and stable cellulases are highly sought after for lignocellulose deconstruction, as the hydrolysis of lignocellulosic biomass at high temperatures in thermogenic processes holds a number of key process advantages over comparable mesophilic processes (as summarised in section 1.4.2.1). Some of these include: easy mixing; better substrate solubility and, a reduced risk of microbial and/or phage contamination (Turner *et al*., 2007). The conditions utilised for many pre-treatment processes, in an effort to reduce the recalcitrant nature of the biomass, are extreme and may include employing high temperatures and low pH levels. Thus, the identification and engineering of thermophilic cellulases have been the aim of many research groups, as these durable enzymes can withstand harsh processing conditions which will minimise the cost of adjusting process conditions between pre-treatment and hydrolysis. Cellulases suitable for inclusion in

these processes should have characteristics such as higher catalytic efficiencies on insoluble cellulosic substrates, be able to completely hydrolyse biomass, increased stability at elevated temperatures, be active at the relevant pH and a higher tolerance to end product inhibition (Zhang *et al*., 2006).

1.5.1.3 Cellulases of microbial origin

In nature, cellulolytic microorganisms produce a wide variety of cellulases which may act synergistically to degrade plant biomass. Microbial enzymes are often more useful in industrial applications than those enzymes derived from plants or animals (Wiseman, 1995), as enzymes from a microbial source often have a broader range of catalytic activities (Konietzny and Griener, 2002). These enzymes are also mostly produced at high yields (Kirk *et al*., 2002), can easily be subjected to genetic manipulation (Iftikhar *et al*., 2010) and are inexpensive to produce as a consequence of the rapid growth of microorganisms on inexpensive media and the absence of seasonal fluctuations (Illanes, 1999; Kirk *et al*., 2002). Microbial enzymes are also generally more stable than equivalent enzymes from plants and animals (Wiseman, 1995). Several bacterial species of *Bacillus* (Ito, 1997), *Bacteriodes* (Forsberg *et al.,* 1981), *Clostridium* (Sleat *et al*., 1984), *Cellulomonas* (Stewart and Leatherwood, 1976), *Erwina* (Boyer *et al*., 1984), *Microbispora* (Wilson, 1992), *Streptomyces* (Jang and Chen, 2003) and *Thermomonospora* (Spiridonov and Wilson, 1998) and fungal species of *Trichoderma* (Beldman *et al.,* 2005), *Penicillium* (Wood and McCrae, 1986), *Fusarium* (Wood, 1971), *Phanerochaete* (Bao and Renganathan, 1992) and *Schizophyllum* (Clarke, 1987) are known to produce cellulases.

Bacterial cellulases may exist as either singular secreted enzymes or discrete multi-enzyme complexes, called cellulosomes (Bayer *et al.,* 2007). Cellulosomes consist of multiple subunits including a scaffoldin subunit, dockerin domain and the carbohydrate binding domain. These domains interact with each other synergistically, resulting in the degradation of cellulosic substrates (Bayer *et al*., 1998). The cellulosome is thought to

enable concerted enzyme activity in close proximity to the bacterial cell, enabling optimum synergism between the various cellulolytic enzymes presented on the cellulosome (Kumar *et al*., 2008). In addition, the presence of a cellulosome also minimises the distance over which cellulose hydrolysis products diffuse, thus enabling efficient uptake of oligosaccharides by the host cell (Kumar *et al*., 2008). Many of the cellulosomes identified to date may also include different types and varied compositions of hemicellulases such as: xylanases; mannanases; arabinofuranosidases; lichenases and pectin lyases, in addition to cellulases (Bayer *et al.*, 2004).

1.5.1.4 Synergistic action between cellulases and oxidative enzymes

The widely accepted mechanism for the enzymatic hydrolysis of cellulose involves the synergistic action of at least three enzymes, namely endoglucanases (EC 3.2.1.2), exoglucanases (EC 3.2.1.4) and β-glucosidases (EC 3.2.1.21) (Fig. 1.10) (Howard *et al.,* 2003). Firstly, endoglucanases randomly hydrolyse accessible intramolecular β-1-4 glycosidic bonds within the cellulose polymer to generate new chain ends (Howard *et al.,* 2003). Exoglucanases then progressively cleave cellulose chains to release soluble cellobiose units consisting of two β-1,4-linked glucose monomers. Finally, β-glucosidases hydrolyse cellobiose to release two molecules of glucose, thus eliminating any possible inhibition a buildup of cellobiose may have on further exoglucanase activity (Howard *et al.,* 2003). A recent review by Horn *et al*. (2012) highlights the benefits associated with the inclusion of copper dependant lytic polysaccharide monooxygenases during lignocellulose hydrolysis by carbohydrate degrading enzymes. These enzymes have a general oxidative action and are thought to disrupt polymer packing in complex substrates, thereby increasing the accessibility to the substrate for hydrolytic attack (Quinlan *et al*., 2011; Beeson *et al*., 2012). These enzymes were formerly classified as GH61 and CBM33 family members, however, have been reclassified as auxillary activity 9 family members on the CaZy database (www.cazy.org). Companies such as Novazymes® are incorporating GH61 in cellulolytic cocktails to enhance the overall performance of the enzyme in saccrification of complex lignocellulosic biomass (Teter, 2012).

1.5.2 Hemicellulases

Hemicellulases represent a diverse group of enzymes that synergistically hydrolyse hemicellulose substrates. Due to the highly variable structure of hemicellulose, a large number of enzymes acting synergistically are required for its complete hydrolysis (Table 1.5). The catalytic mechanisms of hemicellulases employ either GH machinery in which glycosidic bonds are hydrolysed or carbohydrates esterase (CE) machinery in which ester linkages and ferulic acid side chains are hydrolysed (Shallom and Shoham, 2003). Complete degradation of hemicellulose initially requires endo-xylanase(s) to cleave the xylan backbone, whilst β-1,4 xylosidases hydrolyse xylooligosaccarides to xylose units. Thereafter, accessory enzymes such as arabinofuranosidase cleave arabinofuranosyl residues and glucoronidases cleave 4-*O*-methyl glucuronic acid side chain groups from the xylan polysaccharide backbone, before the esterification of acetic acid, ferulic acid and *p*-coumaric acid side chains by acetylxylan and ferulic acid and *p*-coumaric acid esterases.

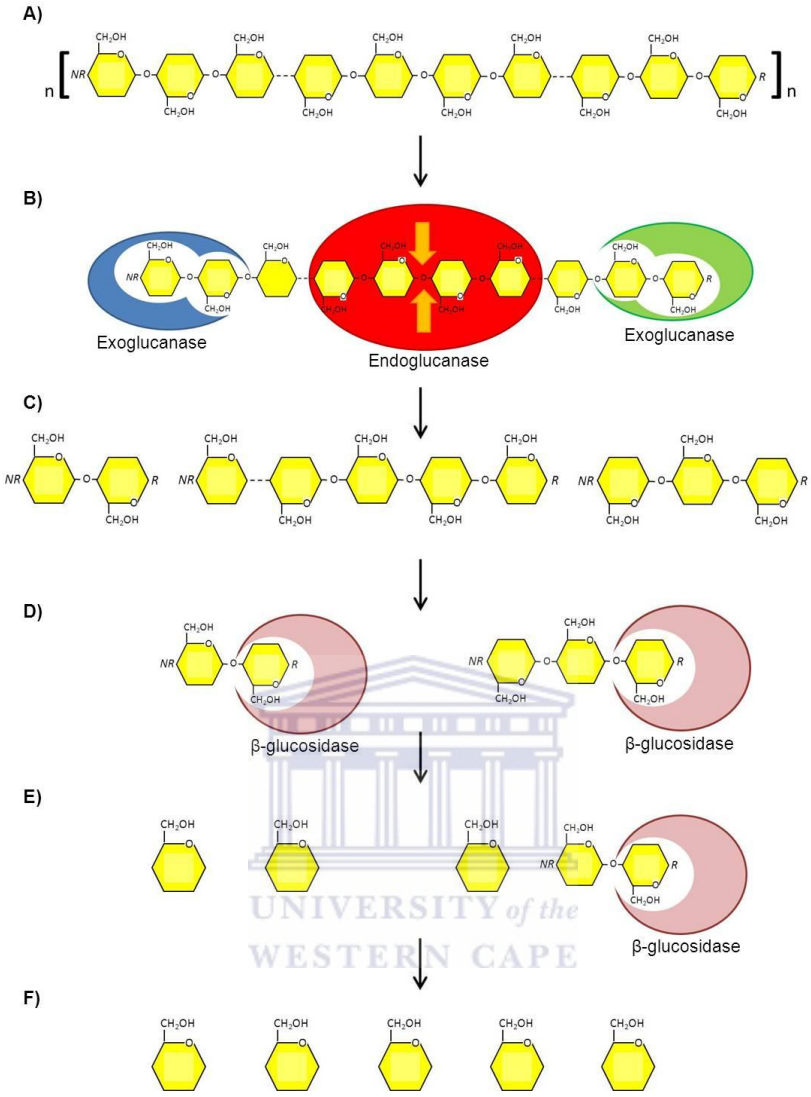


Figure 1.11 Diagram showing the synergistic relationship between the three major classes of cellulase enzymes, endoglucanases, exoglucanases and β-glucosidases. These enzymes act synergistically to sequentially hydrolyse a cellulose polymer (A). Firstly, as indicated in red, an endoglucanase cleaves the β-1,4 linkages within the cellulose polymer (B). New chains formed from the action of the endoglucanase are further degraded by an exoglucanase, by both reducing (indicated in green) and non-reducing (indicated in blue) action, which releases cellobiose units (B,C). Cellobiose (D) is hydrolysed (E) to monomeric glucose residues by β-glucosidase (F) (Adapted from Watanabe *et al*., 2010).

Table 1.5 Classification of hemicellulolytic enzymes (Adapted from Shallom and Shoham, 2003).

EC number	Enzyme	Substrate	Family
3.2.1.8	Endo-β 1,4-xylanase	β 1,4-xylan	GH 5, 8, 10 and 43
3.2.1.37	Exo-β 1,4-xylosidase	β 1,4-xylooligomers, xylobiose	GH 3, 39, 43, 52 and 54
3.1.2.55	α-L-arabinofuranosidase	α-L-arabinofuranosyl, xylooligomers, α-L-arabinan	GH 3, 43, 51, 54 and 62
3.2.99	Endo α-1,5-arabinase	1,5-arabin	GH 43
3.2.1.139	α-glucuronidase	4-*0*-methyl-α-glucuronic acid, xylooligomers	GH 67
3.2.1.78	Endo-β 1,4-mannase	β 1,4-mannan	GH 5 and 26
3.2.1.25	Exo-β 1,4-mannosidase	β 1,4-manooligomers, mannobiose	GH 1, 2 and 5
3.2..1.22	α-galactosidase	α-galactopyranoside, mannooligomers	GH 4, 27, 36 and 57
3.2.1.21	β-glucosidase	β-glucopyranose, mannopyranose	GH 1 and 3
3.2.1.89	Endo-galactanase	β 1,4 galactan	GH 53
3.1.1.72	Acetyl xylan esterase	2- or 3- *0*- acetyl xylan	CE 1, 2, 3, 4, 5, 6 and 7
3.1.1.6	Acetyl mannan esterase	2- or 3- *0*- acetyl mannan	CE 1
3.1.1.73	Ferulic and *p*-cumaric acid esterase	Ferulic acid, *p*-cumaric acid	CE 1

1.5.3 Lignin degrading enzymes

The three enzyme families implicated in the biological degradation of lignin include lignin peroxidases (EC 1.11.1.14), manganese peroxidases (EC 1.11.1.13) and laccases (EC 1.10.3.2) (Maciel *et al*., 2010). Lignin peroxidases and manganese peroxidases belong to the oxidoreductase family, while laccases are members of the multicopper oxidase family (Martinez *et al*., 2005; Alcalde 2007). The classification of lignin degrading enzymes is based upon their ability to cleave different sites within the lignin molecule. Lignin peroxidases and manganese peroxidases cleave the lignin polymer in a process known as lignin mineralisation, while the phenolic compounds are oxidised by laccases. A large number of microorganisms have been shown to produce ligninases (Tien *et al.*, 1987; Pandey *et al*., 1999; Gomes *et al*., 2009), with the main

producers of lignin degrading enzymes being white rot fungi, such as *Pleurotus* species (Caramelo, 1999) and *Phanerochaete chrysosporium* (Pandey *et al*., 1999), and bacteria, such as *Streptomyces viridosporous* and *Clostridium stercoranium* (Ren *et al*., 2009).

1.6 Metagenomic Gene Discovery

With an estimated 4 to 6 x 10^{30} prokaryotic cells on earth (Whitman *et al.,* 1998), predicted to comprise close to 9 million different species (Mora *et al*., 2011), microorganisms represent a significant reservoir of genetic material for the identification of novel biocatalysts with novel metabolic abilities (Simon and Daniel, 2011). However, it is thought that less than 1% of microorganisms within terrestrial and 0.01-0.001% in aquatic environments can be cultured using current technologies and techniques (Amann *et al.,* 1995). Metagenomic technologies represent a culture-independent approach to gene discovery which allows access to the enzymes contained within the genomes of unidentified and potentially unculturable microorganisms (Lorenz *et al.*, 2002; Reisenfeld *et al.*, 2004). Therefore, metagenomics makes it possible to harness the taxonomic diversity and metabolic complexity of the uncultivated majority of microorganisms. As an example, various studies have demonstrated that the unculturable proportion of the microorganisms found within terrestrial soils may contain 1,000-10,000 different species (Torsvik *et al*., 1990; Amann *et al.,* 1995; Torsvik *et al*., 2002; Handelsman *et al*., 2004). Furthermore, the soil metagenome has proved to be an important resource for the isolation of useful biocatalysts and bioactive molecules (Handelsman *et al*., 1998; Schmeisser *et al.,* 2007).

The metagenomic gene discovery process involves the isolation of total genomic DNA, referred to as the metagenome, from an appropriate environmental source, followed by cloning of the metagenome into a heterologous host/vector system to generate a metagenomic library (Fig. 1.12) (Handelsman *et al.*, 1998; Cowan *et al*., 2005). The choice of environmental source from which the metagenomic library is constructed is critical to the successful identification of target genes encoding relevant biocatalysts, as

this can be viewed as a natural enrichment for the genes of interest. Examples of environmental samples thought to contain a large number of lignocellulosic enzymes include i) the rumen of cows (Wang *et al.,* 2009) and buffalo (Duan *et al.,* 2008; Nguyen *et al*., 2012), as the diet of ruminants is comprised of plant material which is then broken down within the rumen by host and microflora encoded hydrolytic enzymes;, ii) digestive tract of termites (Warnecke *et al.,* 2007), as termites are known to have an obligate synergistic relationship with wood-degrading enteric microorganisms; and iii) the gut microflora of abalone which have been shown to produce glucan-degrading enzymes (Kim *et al*., 2011). Metagenomic libraries, constructed from an appropriate environmental source, can be screened for desired target gene(s) or gene product(s) using one of two approaches, namely sequence or function-based screening.

Sequence based screening employs linkages between phylogeny and function (Steele *et al*., 2009). This strategy has provided researchers with reconstructed genomes for yet un-cultured organisms, however, only allows for the identification of enzymes with sequence homology to known functional enzyme classes (Simon and Daniel, 2009). Sequence based screening involves designing primers which are derived from conserved regions of already-known genes or protein families, investigating microbial families by sequencing a marker gene of choice or sampling fragments of whole genomes and/or transcriptomes without the need for culturing (Simon and Daniel, 2009). On the other hand, functional screening allows for the identification of genes encoding novel biomolecules based on the metabolic activities of the metagenomic library clones (Simon and Daniel, 2009). This strategy may potentially lead to the identification of entirely novel gene classes, encoding known functions. Function based screening may employ either phenotypical detection of the desired activity, heterologous complementation of host strains or mutants and induced gene expression (Simon and Daniel, 2009).

With the advent of whole genome sequencing and bioinformatic annotation of genome and metagenome sequences, many of the bottlenecks that previously hampered

metagenomic investigations, such as the cloning and heterologous expression of genomic fragments, are being overcome. To date, despite the numerous challenges associated with metagenomics, it has successfully been used to identify a wide range of biocatalysts, including lipases (Hardeman and Sjoling, 2007), esterases (Jeon *et al*., 2009), proteases (Waschkowitz *et al*., 2009) and DNA polymerases (Simon *et al*., 2009) from a variety of environments and potentially non-culturable microorganisms (Schloss and Handelsman, 2003; Streit *et al*., 2004; Steele and Streit, 2005).

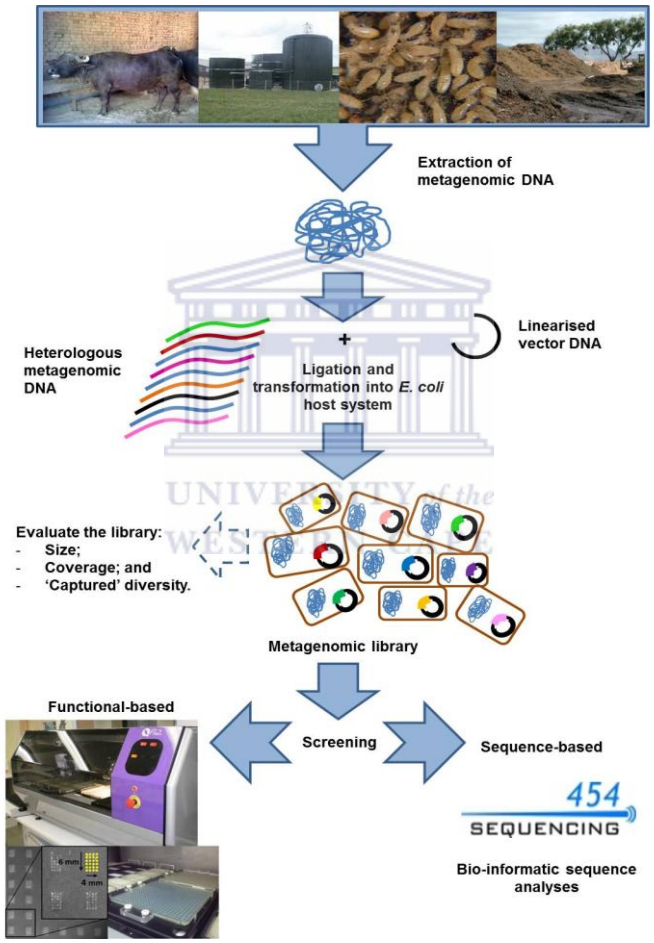


Figure 1.12 Schematic overview of metagenomic library construction and the metagenomic approach to gene discovery which involves extraction of metagenomic DNA from an environmental source, metagenomic library construction and novel gene discovery using function based screening and sequence analysis. (Handelsman, 2004).

1.6.1 Isolation of cellulases using metagenomics

The isolation of novel cellulases from uncultured microorganisms may potentially result in the identification of cellulolytic enzymes with presently unknown biochemical characteristics and properties that may define them as ideal candidates for use as industrial biocatalysts (Healy *et al.,* 1995; Rees *et al*., 2003). Mining cellulases, using metagenomic technologies, from environmental sources allows for the identification of novel cellulases from specific environmental niches (Voget *et al.,* 2006; Duan and Feng, 2010).

Metagenomic approaches have been widely employed to isolate cellulases from samples where plant materials are degraded, including soil (Kim *et al*., 2008, Jiang *et al*. 2009), the rumen of ruminants (Duan *et al.,* 2008; Wang *et al.,* 2009; Bao *et al.,* 2011; Nguyen *et al.*, 2012), the hindgut of termites (Warnecke *et al*. 2007), compost (Pang *et al*. 2009), rabbit cecum contents (Feng *et al*. 2007), sludge from a biogas reactor (Jiang *et al*. 2010) and enrichment cultures (Rees *et al*. 2003; Voget *et al*. 2003; Grant *et al*. 2004) (Summarised in Table 1.5). Interestingly, to the best of our knowledge, there have been no published examples of thermostable cellulases identified using metagenomics from hot spring water or sediments. Metagenomic libraries constructed using genomic DNA from several environmental sources such as psycrophillic Antarctic dry valley soil (Berlemont *et al.,* 2009), mesophillic compost heaps (Pang *et al*., 2009) and thermophillic environments (Duan and Feng, 2010) have also resulted in the identification of a number of cellulolytic genes with thermal properties that correspond to the environmental conditions of the sample from which the library was constructed. Despite these investigations, the largely untapped bacterial diversity associated with terrestrial and aquatic environments suggests that the biotechnological potential for the identification of novel cellulases from uncultured communities has not been fully explored.

Table 1.5 Summary of some of the research conducted to date towards the isolation of cellulases from various environmental sources using metagenomic technologies.

Sample source	Host (vector) system	Library size (Coverage)	Hit rate (%)	Library screening		Cellulase (Characteristics)	Reference
				Sequence-based	Functional-based		
Yak (*Bos grunniens*) rumen	*E. coli* (pWEB Cosmid)	4,000 (nd)	0.025	N/A	Solid-phase screening with CMC and Congo red staining.	A novel cellulase, Rucel5B (with both exo- and endoglucanase activities).	Bao *et al.* (2011)
Antarctic soils	*E. coli* (CopyControl pCC1BAC)	113,742 (650 Mb)	0.12	N/A	Solid-phase screening with CMC and Trypan blue.	11 Cellulases isolated (apparent maximum activities between 35-55°C).	Berlemont *et al.* (2011)
Rabbit cecum	*E. coli* (pWEB Cosmid)	32,500 (114 Mb)	0.033	N/A	Solid phase screening with CMC, 4-methylumbelliferyl-β-D-cellobioside and, esculin hydrate and ferric ammonium citrate.	Four endo-β-1,4-glucanases and seven β-glucosidases (Optimal pH 5.5-8.0 and optimal temperature 40-55°C).	Feng *et al.* (2007)
Buffalo rumen	*E. coli* (pWEB Cosmid)	15,000 (525 Mb)	0.41	N/A	As conducted by Feng *et al.* (2007)	14 Cellulases; One novel cellulase, C67-1, showed optimal activity at pH4.5.	Duan *et al.* (2008)
Hindgut paunch of *Nasutitermes* species termites	*E. coli* [Small (2-4 kb) insert library in pBK-CMV and large (32-kb) insert library in pCC1Fos]	68.47 and 2.91 Mb from the small- and large insert libraries, respectively	nd	Sanger and 454® next-generation sequencing of clones as well as direct sequencing of the metagenome	N/A	>100 putative genes implicated in cellulose hydrolysis identified.	Warnecke *et al.* (2007)
Agarolytic-enriched soil sample	*E. coli* (pWE15 Cosmid)	>2,000 (nd)	0.522	N/A	Solid-phase screening with CMC and Congo red staining.	8 Cellulases; one of which is the halotolerant cellulase, Cel5A.	Voget *et al.* (2003 and 2006)

Table 1.5 (*Continued*)

Sample source	Host (vector) system	Library size (Coverage)	Hit rate (%)	Library screening		Cellulase (Characteristics)	Reference
				Sequence-based	Functional-based		
Cow rumen	*E. coli* (λ ZAP Express vector)	35,700 (0.2 Gb)	0.4	N/A	Solid-phase screening with CMC (Cellulase) and carboxymethyl cellulose or Esculin hydrate and ferric ammonium citrate (β-glucosidase).	Two GH family 5 cellulases and GH family 3 β-glucosidases.	Wang *et al.* (2009)
Elephant dung	*E. coli* (λ ZAP Express vector)	28,900 (0.18 Gb)	0.2	N/A	Solid-phase screening with CMC (Cellulase) and carboxymethyl cellulose or Esculin hydrate and ferric ammonium citrate (β-glucosidase).	Two cellulases isolated (GH families 5 and 9).	Wang *et al.* (2009)
Buffalo (*Bubalus bubalis*) rumen	*E. coli* (CopyControl™ pCC1FOS™ system)	10,000 (0.3 Gb)	0.93	N/A	Solid-phase screening with AZCL-HE-CMC overlay.	Characterization of the clone with the highest activity revealed a cellulase with apparent optimal activity at pH7 and 37°C.	Nguyen *et al.* (2012)
Abalone digestive tract sample	*E. coli* (SuperCos1)	90,000 (3.24 Gb)	0.026	N/A	Solid-phase screening with CMC and Congo red staining.	Cellulase, CelAM11, has apparent optimal activity at pH 7 and 37°C.	Kim *et al.* (2011)
Grassland soil	*E. coli* (pCC1FOS™)	147,888 (3.44 Gb)	0.021	N/A	Hydroxyethyl cellulose (HEC_{red})	Novel cellulase, Cel01, (GH family 9) contains a family 9 CBM; Apparent optimal activity at pH6-7 and 45-50°C.	Nacke *et al.* (2012)
Hot spring sediment	*E. coli* (pCC1FOS™)	10,000 (0.3 Gb)	0.15	N/A	Solid-phase screening with CMC and Congo red staining.	One of the cellulases has an apparent optimal activity at pH4 and 50°C.	X.P. Hu, unpublished (2010)

nd: not determined; N/A: not applicable

1.7 Concluding remarks

Ethanol produced from lignocellulosic biomass utilising a second generation process is recognised as a viable alternative to fossil fuel based transportation fuels (Balat and Balat, 2008; Balat and Balat, 2009; Margoet *et al*., 2009). An important stage in a second generation process is the enzymatic saccharification of the plant derived biomass to fermentable sugars for ethanol production (Ren *et al*., 2009; Wen *et al*., 2009). Due to the complex nature of plant material, a number of lignocellulosic enzymes including cellulases are required to efficiently hydrolyse the substrate to fermentable sugars (Wilson, 2009). Currently, the majority of industrial bioethanol production processes are mesophilic, although, thermophilic bioethanogenesis is reported to have a number of advantages (Taylor *et al.,* 2009). Thermophilic bioethanol production processes require the inclusion of thermostable lignocellulosic enzymes during the enzymatic hydrolysis step and thermophilic microorganisms, capable of growth at temperatures greater than 65°C, represent an excellent source of potentially novel thermostable biocatalysts. The use of metagenomic methods enables researchers to access the genetic potential of mostly unculturable and currently unknown microorganisms within a given environmental sample. The isolation of novel genes may aid the development of enzymes with significant biotechnological potential for use in the biofuels industry.

1.8 Aims and objectives of this study

The present investigation forms part of a larger study, in which researchers based at the Institute for Microbial Biotechnology and Metagenomics (UWC), together with collaborators at the University of Stellenbosch and CSIR Biosciences, are employing metagenomic technologies to identify novel lignocellulosic enzymes. The hypothesis of this study was that enzymes identified from a Malawian hot spring metagenomic would display similar thermophilic characteristics to that of the environmental source. The temperature and pH measured at the sample collection site was 72-78°C and pH 6.2 respectively. As these characteristics are within the scope of conditions being evaluated for use during the thermogenic sacccharification of lignocellulosic biomass,

carbohydrate degrading enzymes identified from this library may be suitable for the inclusion in this process. The primary aim of this study was therefore to identify and characterise a novel cellulase, isolated from this thermophilic metagenomic library, and determine its suitability for use as a biocatalyst in the enzymatic saccharification of lignocellulosic biomass.

A metagenomic library, constructed from sediment material collected from the Mphizi hot spring (Chiweta, Malawi), was functionally screened for clones displaying cellulolytic activity by XP Hu (MSc thesis, 2010). A number of *E. coli* fosmid clones displaying cellulase activity were identified. The specific objective of the current study was to further investigate the cellulolytic activity displayed by one of these fosmid clones, pFos_032c10. Following 454 sequencing of the fosmid, *in silico* gene mining was used to identify putative cellulolytic gene candidates contained on the fosmid. A gene sequence containing a carbohydrate binding motive was cloned and the recombinant protein, expressed in *E. coli*, was fully characterised. The results and discussions in the chapters to follow, demonstrate the value of employing metagenomic techniques to isolate novel enzymes for use in industrial processes.

Chapter 2

Materials and Methods

2.1 General techniques used in this study

This chapter gives a detailed account of the technical procedures that were followed during the course of this investigation.

2.2.1 Material suppliers

All reagents used were supplied by either Merck Chemicals and Laboratory Supplies, Sigma-Aldrich Chemical Company or Kimix Chemical and Laboratory Supplies. Culture media used were supplied by Oxoid Ltd and Biolabs. Protein molecular weight markers and restriction enzymes were supplied by Fermentas.

2.1.2 Media

The following media was used for culturing of microorganisms:

i) Luria-Bertani broth (LB; Sambrook and Russell, 2001) 1% (w/v) tryptone, 0.5% (w/v) yeast extract and 1% (w/v) NaCl.
ii) Luria-Bertani agar (LA; Sambrook and Russell, 2001), solid media, was prepared from LB with the addition of 1.5% (w/v) bacteriological agar.
iii) Super Optimal Broth (SOB; Sambrook and Russell, 2001) 2% (w/v) tryptone, 0.5% (w/v) yeast extract, 0.05% (w/v) NaCl and 0.02% (w/v) KCl.
iv) Super Optimal broth with Catabolite repression (SOC; Sambrook and Russell, 2001) was prepared from SOB with the addition of filter sterilized 1mM $MgCl_2$ and 20mM D-glucose.
v) Ψ-broth (Dower *et al.*, 1988) 2% (w/v) tryptone, 0.5% (w/v) yeast extract, 0.4% (w/v) $MgSO_4$ and 0.075% (w/v) KCl.

Media was prepared by mixing media components together, and where appropriate the pH was adjusted to 7.0, before autoclaving. Media was allowed to cool to approximately 50°C and where appropriate filter sterilized antibiotics were aseptically added. Unless

otherwise stated the final concentrations of the antibiotics used were 100 µg/ml ampicillin (Amp), 50 µg/ml kanamycin (Kan) and 12.5 µg/ml chloramphenicol (Cam).

2.1.3 Bacterial strains and plasmids

The bacterial strains and plasmids used and generated during the course of this study are listed in Table 2.1 and 2.2, respectively.

Table 2.1 Bacterial strains used in this study.

Strain	Genotype/Relevant features[a]	Reference
E. coli Genehog	F- *mcr*A Δ(*mrr-hsd*RMS-*mcr*BC) φ80*lac*ZΔM15 Δ*lac*X74 *rec*A1 *ara*D139 Δ(*ara-leu*)7697 *gal*U*gal*K*rps*L (StrR) *end*A1 *nup*G*fhu*AIS2	Invitrogen, USA
E. coli EPI300™	F- mcrA D(*mrr-hsd*RMS-*mcr*BC) f80d*lac*ZDM15 D*lac*X74*rec*A1 *end*A1 *ara*D139 (*ara*, *leu*) 7697 *gal*K1- *rsp*LnupG*trf*A*ton*A*dhfr*	Epicentre Biotechnology USA
E. coli Rosetta2	F-*ompThsdS*$_B$($r_B^-m_B^-$) *gal dcm*pRARE2[3],Cam^R	Novagen, USA
E. coli BL21(DE3)	F- *omp*T*hsd*SB(r_B-m_B-)*gal dcm gal* λ(DE3)	Invitrogen, USA

[a]The bacterial strain's antibiotic resistance is indicated as Cam^R, Chloramphenicol resistance or Kan^R, Kanamycin resistance.

2.1.4 Growth of Escherichia coli strains

Escherichia coli (*E. coli)* strains were cultured on solid or liquid growth media supplemented with appropriate antibiotics (section 2.1.2; Table 2.1). When *E. coli* strains were cultivated in liquid media, incubation was accompanied by constant agitation on a rotary shaker at 150 to 225 rpm. Unless otherwise stated, cultures were incubated at 37°C.

Table 2.2 Plasmids used and generated during this study.

Plasmid	Genotype/Relevant features[a]	Reference
pJET1.2/blunt	Suicide vector (*eco47IR*),blunt DNA ends for ligation with insert, T7 promoter, Amp^R	Fermentas, USA
pET21a(+)	C- terminal His-tag, Amp^R	Novagen, USA
pET28a(+)	N- and C-terminal His-tag T7 *lac* T7 Tag, Kan^R	Novagen, USA
pFos_032c10	Copy Control pCC1FOS vector containing approximately 30 kb of metagenomic DNA insert with cellulase activity, Cam^R	Xiao Ping Hu, IMBM, UWC, SA
pJET_*celMHS*	pJET1.2/blunt containing *celMHS* ,Amp^R	This study
pET21_*celMHS*	pET21a(+) containing *celMHS*, Amp^R	This study
pET28_*celMHS*	pET28a(+) containing *celMHS*, Cam^R and Kan^R	This study
pET28_ celMHSΔTMR	pET28a(+) containing *celMHS*, Cam^R and Kan^R	Ilana Ackermann, IMBM, UWC, SA

[a]Plasmid encoded antibiotic resistance is indicated as Amp^R, ampicillin resistance; Cam^R, chloramphenicol resistance; and Kan^R, kanamycin resistance.

2.2 Molecular biology techniques

2.2.1 Fosmid DNA extraction

An *E. coli* Epi300™ (Table 2.1) clone harbouring the recombinant fosmid construct pFos_032c10 (Table 2.2) was streaked from a frozen glycerol stock onto the surface of an LA plate supplemented with 12.5 µg/ml Cam and incubated at 37°C for 16 hours (hrs). A single colony was inoculated into 5ml LB containing 12.5 µg/ml Cam and incubated overnight as described in section 2.1.3. One ml of this overnight culture was used as starter culture to inoculate 100 ml fresh LB containing 0.01% (w/v) L-arabinose and 12.5 µg/µl Cam, and incubated for 5 hrs at 37°C with agitation. The cells were

harvested by centrifugation (4,000x*g* for 30 min at 4°C) and fosmid DNA was extracted from the cell pellet using the Plasmid Midi Prep Kit (Qiagen, Germany), according to the manufacturer's instructions. Extracted fosmid DNA was quantitated using the NanoDrop spectrophotometer (ND-1000; NanoDrop Technologies, USA).

2.2.2 Agarose gel electrophoresis

Nucleic acid samples were separated according to the method outlined by Sambrook and Russell (2001). Briefly, agarose gels [0.8-1% (w/v) agarose] were prepared in Tris-Acetate EDTA (TAE) buffer [0.2% (w/v) Tris-base, 0.5% (v/v) glacial acetic acid and 1% (v/v) 5 M ETDA, pH 8.0] (Sambrook and Russell, 2001). Ethidium bromide (0.5 µg/ml), was added to the molten agarose during preparation to allow post-electrophoresis visualization of nucleic acids. Sample preparation involved mixing 6x concentrated loading buffer [20% (v/v) glycerol, 60 mM EDTA and 5 mg/ml bromophenol blue] with the specific sample, before loading onto the agarose gel. Electrophoresis was performed in 1x TAE buffer at 10-12V/cm. Gels were visualized using ultraviolet (UV; 302 nm) light illumination and photographed with a digital imaging system (Alphalmager 2000, Alpha Innotech, USA). Λ DNA was restricted to completion with *Pst*I and used as a DNA size marker.

2.2.3 DNA extraction from agarose gels

DNA extracted from agarose gels was purified using the NucleoSpin PCR/DNA and gel band purification kit (Clonetech Laboratories, USA) according to the manufacturer's specifications.

2.2.4 Restriction enzyme digestion

Restriction enzyme digestions were performed in reaction volumes of 20-50 µl, according to the manufacturer's instructions (Fermentas). The reactions contained 1U of

restriction endonuclease per µg of DNA and the appropriate buffer as per the manufacturer's specifications. Reactions were incubated for approximately 16 hrs at 37°C, unless otherwise specified. Thereafter, restriction endonuclease reactions were stopped by heat inactivation and addition of loading buffer (section 2.2.2). Restriction enzyme digestion products were visualised following separation of DNA fragments by agarose gel electrophoresis (section 2.2.2).

2.2.5 Preparation and transformation of competent E. coli cells

Electrocompetent *E. coli* cells were prepared according to the method outlined by Miller and Nickioloff (1995). In addition, chemically competent *E. coli* cells were prepared according to the method described by Sambrook and Russell (2001). Once prepared all *E. coli* competent cells were stored at -80°C. The following transformation protocols were used to generate recombinant clones.

2.2.5.1 Electroporation

Electrocompetent *E. coli* cells (section 2.2.5) were transformed, according to Dower *et al.* (1988), by addition of 50 ng plasmid DNA to 50 µl freshly thawed electrocompotent *E. coli* cells in a sterile pre-chilled electroporation cuvette (0.1 mm gap junction; Bio-Rad Laboratories, USA). The cells were incubated on ice for 10 min prior to electroporation at 1.8 kV, 25 µF and 200 Ω (Gene Pulser, Bio-Rad, USA). After the addition of 950 µl ice-cold SOC (section 2.1.2) the transformation mixture was incubated in a sterile 1.5ml microfuge tube for1 hour (hr) at 37°C with constant agitation (225 rpm), before being plated onto LA supplemented with an appropriate antibiotic.

2.2.5.2 Heat shock transformation

Chemically competent *E. coli* cells were transformed following the addition of 10ng plasmid DNA to freshly thawed competent cells on ice following the protocol by Van Die *et al.,* 1983. Briefly, the cells were kept on ice for 2 min, heat shocked by incubation at 37°C for 60seconds and again placed back on ice for 2 min. After the addition of 800 µl ψ-broth (section 2.1.2) the transformation mixtures were incubated for 1 hr at 37°C with constant agitation at 225 rpm, before being plated onto LA supplemented with an appropriate antibiotic.

2.2.6 Cell lysis by sonication

E. coli cells harvested by centrifugation (6,000x*g* for 20 min at 4°C) were resuspended in 0.1 M sodium phosphate buffer (pH 6) prepared by adding 12 ml 1M NaH_2PO_4 and 88 ml 1M Na_2HPO_4 to 900 ml distilled water and the pH adjusted to 6.0. The resuspended cells were subjected to hydrodynamic shearing by exposure to 6 pulses of sonication (Bandelin Sonoplus Ultrasonic Homogenizer, Germany) for 20 seconds followed by 30 seconds of no sonication. Throughout the sonication procedure the cell suspension was maintained on ice to prevent heat build-up within the sample. Following sonication the insoluble cell debris was removed by centrifugation (6,000x*g* for 20 min at 4°C). Both the supernatant, containing the soluble proteins, and the pellet, containing insoluble proteins and cellular debris, were retained for subsequent analyses.

2.3 Cloning of celMHS

2.3.1 Metagenomic Fosmid Library Construction and Screening

A Malawian hot spring metagenomic library was constructed, validated and screened by Xiao Ping Hu during the course of her MSc research project (Hu, MSc thesis, 2010). Briefly, a sample of sediment was collected from Mphizi hot springs, Chiweta, Malawi (10.68°22'07"S 34.18°57'56"E) by Prof. D Cowan in March 2009. The temperature range and pH measured at the source were 72-78°C and 6.2, respectively. High molecular weight metagenomic DNA was extracted from this sample according to the method outlined by Zhou *et al.* (1996). This DNA was used to construct a metagenomic library using the Copy Control™ Fosmid Library production kit (Park *et al*., 2008) comprising the pCC1FOS™ plasmid vector (Fig. 2.1; EpiCentre Biotechnologies) in *E. coli* EPI300™ (Table 2.1) cells. The library comprised approximately 10,000 *E. coli* clones with an average insert size of 30 kb, which represented approximately 100 bacterial genomes (Hu, MSc thesis, 2010).

E. coli EPI300™fosmid metagenomic clones were functionally screened for cellulase activity using Congo red (Teather and Wood, 1982) and Remazol Brilliant Blue coupled CMC (RBB-CMC; Biely *et al*., 1985) screening methods. Briefly, the Congo Red colorimetric assay involved plating *E. coli* EPI300™recombinant fosmid metagenomic clones on the surface of LA supplemented with 1% carboxymethyl cellulose (CMC; Sigma), 0.01% (w/v) L-arabinose and 12.5 µg/ml Cam, and incubating for five days at 37°C. Subsequently, cellulase activity was visualised by staining the growth medium with 1% (w/v) Congo red (Merck, USA), after removing bacterial colonies from the surface of the plate by scraping. Plates with Congo red solution were incubated at 22°C for 20 min thereafter the stain was discarded and the plates destained by the addition of 1 M NaCl and incubated at room temperature for 20 min. Cellulose hydrolysing clones were identified visually by the presence of a yellow halo surrounding the colony against a red background. Putative positive cellulase clones were also screened using RBB-

CMC, prepared according to the method outlined by (Biely *et al*., 1985). Briefly, *E. coli* EPI300™ fosmid clones were plated onto LA containing 0.7% (w/v) RBB-CMC, 0.3% (w/v) CMC, 0.01% (w/v) L-arabinose and 12.5 µg/ml Cam, and incubated at 37°C. An extracellular cellulose phenotype was identified by visual detection of zones of clearing surrounding bacterial colonies.

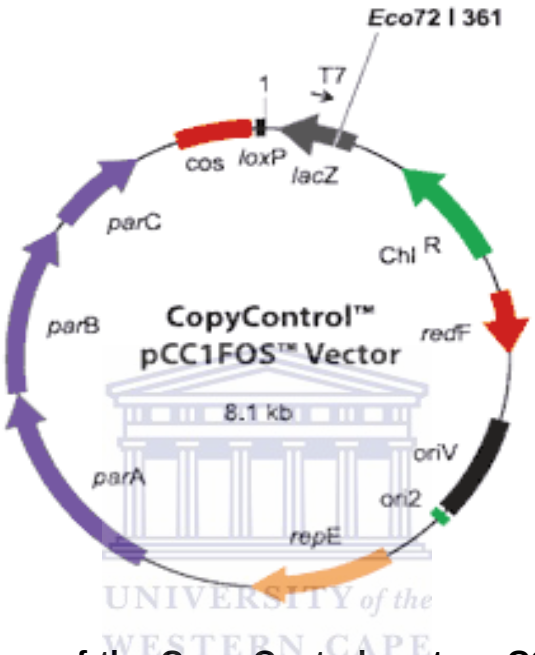


Figure 2.1 **A vector map of the Copy Control vector pCC1FOS**. The positions of the chloramphenicol antibiotic resistance gene (Chl^R) the T7 *LacZ*gene, and the blunt-ended cloning site (*Eco*72I361) are indicated (Epicentre Biotechnologies).

During the course of X.P. Hu's MSc research project several putative cellulase clones were identified, as described above (Hu, MSc thesis, 2010). One of these fosmid clones, designated pFos_032c10 (Table 2.2), was selected for further investigation and formed the basis for this investigation. Verification of the cellulase activity conferred by pFos_032c10 was performed as follows. Fosmid DNA was extracted as described in section 2.2.1 and re-transformed into electrocompetent *E.coli* EPI300™cells (section 2.2.5.1). The cellulase activity of *E. coli* EPI300™ harbouring pFos_032c10 and a negative control *E. coli* EPI300™ was assessed by cultivation on LA supplemented with CMC and subsequent Congo red staining, as described previously.

2.4 Sequencing and assembly of pFos_032c10

2.4.1 Sequencing of pFos_032c10

The recombinant fosmid clone, pFos_032c10, was end-sequencedusing the T7 Forward and pCC1Fos Reverse primers (Table 2.3) (Macrogen, Korea).Sequence data was edited using CHROMAS v2.1 software (Technelysium Pty Ltd., Australia) and DNAMAN for windows v4.13 software (Lynnon Biosoft, Canada).

Table 2.3 Oligonucleotide primers used in this study

Primer	Primer sequence (5'-3')	Reference
T7 Forward	GGATGTGCTGCAAGGCGATTAAGTTGG	Novagen, USA
pCC1FOS Reverse	TACGCCAAGCTATTTAGGTGGTGAGA	EpiCentre, USA
pJET1.2 blunt Forward	CGACTCACTATAGGGAGAGCGGC	Fermentas, USA
pJET1.2 blunt Reverse	AAGAACATCGATTTTCCATGGCAG	Fermentas, USA
T7 Promoter	CCATATGAATACAGATAAACGCAG	Novagen, USA
T7 Terminator	GCTCGAGTTTAATAAATTTTTCGCTC	Novagen, USA
celMHS3_Forward	AT**CATATG**GCGCGGGTGGTGGCTG	This study
celMHS3_Reverse	**CTCGAG**TCACTACAGCGACTTC	This study
celMHSΔTMR _Fwd	GCGTTAT**CCATGG**GCATGCGA	Ilana Ackermann, IMBM, UWC, SA
celMHSΔTMR _Rev	GCTAT**GCGGCCGC**ACTCTACA	Illana Ackermann, IMBM, UWC, SA

Restriction endonuclease sites incorporated into the PCR primers to facilitate directional cloning into expression vectors, are indicated in coloured bold type. Red indicates *Xho*I, green indicates *Nde*I, blue indicates *Nco*I and pink indicates *Not*I.

The fosmid DNA from clone pFos_032c10 was extracted, as described in section 2.2.1, and sequenced, together with 7 other recombinant fosmids, using the 454 GS-FLX sequencing platform (Inqaba Biotech, South Africa). Fosmid DNA preparations were pooled together in an equi-molar ratio, according to the specifications of the service provider, prior to 454 GS-FLX sequencing. Assembly of contig sequences, comprising sequence reads, was performed using CLC Genomics Workbench (www.clcbio.com) and provided by Inqaba Biotech. The 454 GS-FLX contig sequences were compared to the T7 Forward and pCC1Fos Reverse end-sequences (Table 2.3) using Sequencher (www.sequencher.com) in order to identify the contig(s) that represented the fosmid clone pFos_032c10. A contig sequence (designated Contig#23) was identified from the pool of contigs as being a portion of the recombinant fosmid pFos_032c10.

Softberry FgenesH online (www.softberry.com), GLIMMER (TIGR) (Kelley *et al*., 2011) and Genemark (Borodovsky *et al*., 2002) were used to predict putative open reading frames (ORFs) contained within Contig#23. Putative gene identities and functions of the predicted ORFs were determined by comparing *in silico* translated protein sequences with those contained in databases, including the National Center for Biotechnology Information (NCBI) database using the Basic Local Alignment Search Tool (BLASTp) (Altschul *et al*., 1990), Conserved Domain Database (CDD) (Fong and Marchler-Bauer, 2008) and InterProScan (Hunter *et al*., 2011). A putative cellulase gene, *celMHS*, identified within Contig#23, and therefore pFos_032c10, was selected for further investigation and formed the focus of this study.

2.4.2 Bio-informatic analysis and annotation of celMHS

The predicted protein sequence, CelMHS, was compared to related protein sequences contained within the NCBI database using BLASTp (Altschul *et al.*, 1990). In addition, InterProscan (Zdobnov and Apweiler, 2001) and Pfam (Finn *et al.*, 2010) databases were used to predict conserved domains and protein-based families within the CelMHS. TMHMM server Version 2.0 (http://www.cbs.dtu.dk/ services/TMHMM/) was used to predict the presence and possible location of secondary structures within the deduced

protein sequence. The quantity of rare codons present within the sequence was determined using Rare codon caltor (Ergin *et al*., 2007).

Phylogenetic analysis of CelMHS CBM was performed following a multiple sequence alignment with related sequences using Clustal_X (Larkin *et al.*, 2007) and MEGA version 5.0 (Higgins *et al*., 1996, Tamura *et al.*, 2011). These conserved cellulase sequences were used to construct a phylogenetic tree in MEGA based on Bayesian statistics by means of the WAG model. Bootstrap values were based upon 1,000 re-sampled data sets (Felsenstein, 1985) and only bootstrap values of greater than 40% are shown.

2.4.3 Structural modelling of CelMHS CBM

The web tool, SWISS MODEL (Peitsch, 1995; Schwede *et al*., 2003; Arnold *et al*., 2006; Kiefer *et al*., 2009), was used to generate a homology model of the CBM within CelMHS. Briefly, SWISS MODEL performs a BLAST search against the SWISS MODEL structural database, using the CelMHS sequence as a query. The alignment to the best match protein within this database is then used for the homology modeling. The accuracy of the model was assessed using the web tool Rampage (Koteiche *et al*., 2005). This tool provides an indication of whether the modeled structure is favourable by generating a Ramachandrian plot, showing the phi and psi bond angles for the homology model. Figures were generated using PyMolv1.5 (The PyMOL Molecular Graphics System, Version 1.2r3pre, Schrödinger, LLC) (Mura *et al*., 2010).

2.5 Amplification and cloning of celMHS

To assess the biochemical characteristics of CelMHS experimentally, the coding sequence of *celMHS* was cloned into *E. coli* expression vectors to enable the expression and purification of recombinant CelMHS for characterisation.

2.5.1 Cloning celMHS into pET21(a)+

The coding region of the putative cellulase gene, *celMHS*, was amplified by Polymerase Chain Reaction (PCR). The PCR reaction included 50 ng of pFos_032c10 template DNA, 1x KAPA HiFi buffer (containing 25mM $MgCl_2$ and 0.3 mM dNTPs; Kapa Biosystems, South Africa), 0.3 µM celMHS3_Forward and celMHS3_Reverse primers (Table 2.3) and 0.5 U of KAPA HiFi DNA polymerase (Kapa Biosystems, South Africa). The oligonucleotide primers, celMHS3_Forwardand celMHS3_Reverse, were designed to amplify the gene directly from the recombinant fosmid, pFos_032c10, and incorporated the restriction endonuclease recognition sequences for *Nde*I and *Xho*I, respectively (Table 2.3), to aid subsequently cloning into the pET21a(+) expression vector (Fig. 2.2). The PCR reactions were performed in a GeneAmp® 2700 thermocycler (Applied Biosystems) using the following reaction conditions: initial denaturation temperature of 95°C for 5min; followed by 98°C for 20 sec, annealing at 65°C for 15 sec and extension at 72°C for 90 sec repeated for 25 cycles prior to a final extension step at 72°C for 5 min. PCR amplicons were subjected to electrophoresis on a 1% (w/v) TAE agarose gel (section 2.2.4) prior to being visualized using UV light illumination (Alphalmager 2000, Alpha Innotech, USA) to determine whether the gene of interest was successfully amplified.

The amplified PCR product was excised from a 1% (w/v) TAE agarose gel (section 2.2.3), under long wavelength UV light (365 nm), and purified as described in section 2.2.3. Subsequently, the purified PCR product was ligated into the pJET1.2/blunt vector system (Table 2.1; Fermentas, Lithuania) according to manufacturer's instructions. The resulting recombinant constructs were transformed into chemically competent *E. coli* Genehog (section 2.2.5.2), plated onto LA containing 100µg/ml Amp and incubated for 16 hrsat 37°C. Eight *E. coli* Genehog transformants, capable of growth on Amp, were inoculated into 5 ml LB containing 100 µg/ml Amp and incubated overnight at 37°C with shaking (225 rpm). Plasmid DNA was extracted from the overnight cultures using the Qiagen mini-prep kit (Qiagen, Germany) and restricted with *Nde*I and *Xho*I (Fermentas), according to the manufacturer's instructions, to confirm the presence of the desired

insert. Two of these pJET1.2_*celMHS* clones were selected for sequencing, using pJET1.2 blunt Forward and pJET1.2 blunt Reverse (Table 2.3), to ensure the absence of PCR sequence errors within the amplified and cloned *celMHS* sequence. One of these clones, with 100% identity to the *celMHS* gene annotated from Contig#23 was chosen for further study. This pJET1.2_*celMHS* clone and the expression vector pET21a(+) (Table 2.1 and Fig. 2.2; Novagen, USA) were restricted with *Nde*I and *Xho*I to allow directional cloning of *celMHS*. The *NdeI* and *XhoI* restricted insert DNA fragments, comprising the *celMHS* gene and linearised pET21a(+) were ligated together at a 1:1 molar ratio using T4 DNA ligase (Fermentas, Lithuania) according to the manufacturer's instructions to generate pET21_*celMHS*. Subsequently the ligation reaction was transformed into chemically competent *E. coli* Genehog cells (section 2.2.5.2), before being plated onto LA containing 100 µg/ml Amp and incubated for 16 hrs at 37°C. *E. coli* Genehog cells capable of growth on Amp were inoculated into 5ml LB supplemented with 100 µg/ml Amp and incubated overnight at 37°C with shaking (225 rpm). Plasmid DNA was isolated from the overnight cultures using the Plasmid Mini Prep Kit (Qiagen, Germany), according to the manufacturer's instructions. The presence of the *celMHS* gene was confirmed by performing a diagnostic restriction enzyme digest with *NdeI* and *XhoI*. A recombinant construct containing the *celMHS* gene was identified and designated pET21_*celMHS* (Table 2.2). The successful in-frame cloning of *celMHS* into pET21a(+) was confirmed by sequencing using the T7 promoter and terminator primers (Table 2.3; Macrogen, Korea).The recombinant construct, pET21_*celMHS* was extracted and transformed into chemically competent *E. coli* Rosetta 2 and *E.coli* BL21(Table 2.1) as described in section 2.2.5.2.

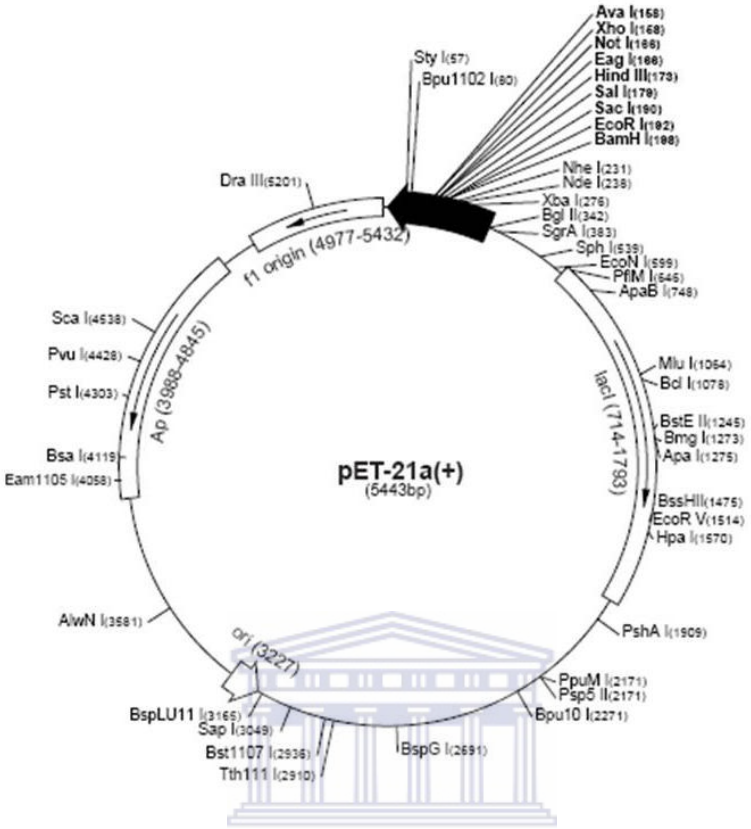


Figure 2.2 **A vector map of the pET21a(+) expression vector used in the sub-cloning and expression of the *celMHS* gene**. The positions of the restriction enzyme recognition sequences, in base pairs (bp), are indicated in brackets. The position of the Amp antibiotic resistance gene is indicated (Novagen).

2.5.2 Cloning *celMHS* into pET28a(+)

The plasmid,pET21_*celMHS*, was extracted from *E. coli* Genehog cells harboring pET21_*CelMHS* and restriction endonuclease digested using *Nde*I (Roche) and *Xho*I (Fermentas), according to the manufacturer's instructions, prior to directional cloning into the pET28a(+) expression vector (Table 2.1 and Fig. 2.3; Novagen, USA) restricted with the same enzymes. The *NdeI* and *XhoI* digested *pET21_celMHS* plasmid DNA, comprising the *celMHS* gene and linearised pET28a(+) were ligated together at a 1:1 molar ratio using T4 DNA ligase (Fermentas, Lithuania) according to the manufacturer's instructions. Subsequently the ligation reaction was transformed into chemically

competent *E. coli* Genehog (section 2.2.2.5.2), plated onto LA containing 50 µg/ml Kan and incubated for 16 hrs at 37°C.The presence of the *celMHS* gene was confirmed using a diagnostic restriction enzyme digest with *Nde*I and *Xho*I, according to the manufacturer's instructions. A recombinant construct containing the *celMH*S gene was identified and designated pET28_*celMHS* (Table 2.2). Furthermore, the successful in-frame cloning of *celMHS* into pET28a(+) was confirmed by sequencing using the T7 promoter and terminator primers (Table 2.3; Macrogen, Korea). Thereafter, pET28_*celMHS* plasmid DNA was extracted from *E. coli* Genehog and transformed into chemically competent *E.coli* Rosetta 2 and *E.coli* BL21 cells (Table 2.1), as described in section 2.2.5.2.

2.5.3 Cloning celMHSΔTMR into pET28(a)+

The coding region of the putative cellulase gene without the nucleotides coding for the signal peptide, ***celMHSΔTMR***, was PCR amplified and cloned into pET28a(+) by Ilana Ackermann (IMBM, UWC, SA)during the course of her PhD study. The oligonucleotide primers, ***celMHSΔTMR*** _Fwd and ***celMHSΔTMR*** _Rev (Table 2.3), were designed to amplify the mature *celMHS* gene, without the putative signal peptide sequence. These primers included the restriction endonuclease recognition sequences for *NcoI* and *Not*I, respectively (Table 2.3), which would subsequently aid in cloning of ***celMHSΔTMR*** into the pET28a(+) expression vector. The amplified PCR product was digested with *Nco*I and *Not*I (Fermentas), according to the manufacturer's instructions, prior to cloning into the expression vector pET28a(+) (Novagen, USA) that had been linearised by digestion with *Nco*I and *Not*I.The digested insert DNA fragment, comprising the ***celMHSΔTMR*** gene and linearised pET28a(+) were ligated together at a 1:1 molar ratio using T4 DNA ligase (Fermentas, Lithuania) according to the manufacturer's instructions. Subsequently, the ligation reaction was transformed into chemically competent *E. coli*Genehog cells (section 2.2.1) as described in section 2.2.5.2, plated onto LA containing 50 µg/ml Kan and incubated for 16 hrs at 37°C. *E. coli* Genehog cells capable of growth on Kan were inoculated into 5ml LB supplemented with 50 µg/ml Kan and incubated overnight at 37°C with shaking (225 rpm). Plasmid DNA was isolated

from the overnight cultures using the Plasmid Mini Prep Kit (Qiagen, Germany), according to the manufacturer's instructions. The presence of the *celMHSΔTMR* gene was confirmed using restriction endonuclease analysis with *Nco*I and *Not*I. A recombinant construct containing the *celMHSΔTMR* gene was identified and designated pET28_ *celMHSΔTMR* (Table 2.2). The successful in-frame cloning of *celMHSΔTMR* into pET28a(+) was confirmed by sequencing using the T7 promoter and terminator primers (Table 2.3; Macrogen, Korea).Recombinant pET28_*celMHSΔTMR* plasmid DNA was isolated from an overnight culture using the Qiagen mini-prep kit, as per manufacturer's description and transformed into chemically competent *E.coli* Rosetta 2 and *E.coli* BL21 cells (Table 2.1) as described in section 2.2.5.2.

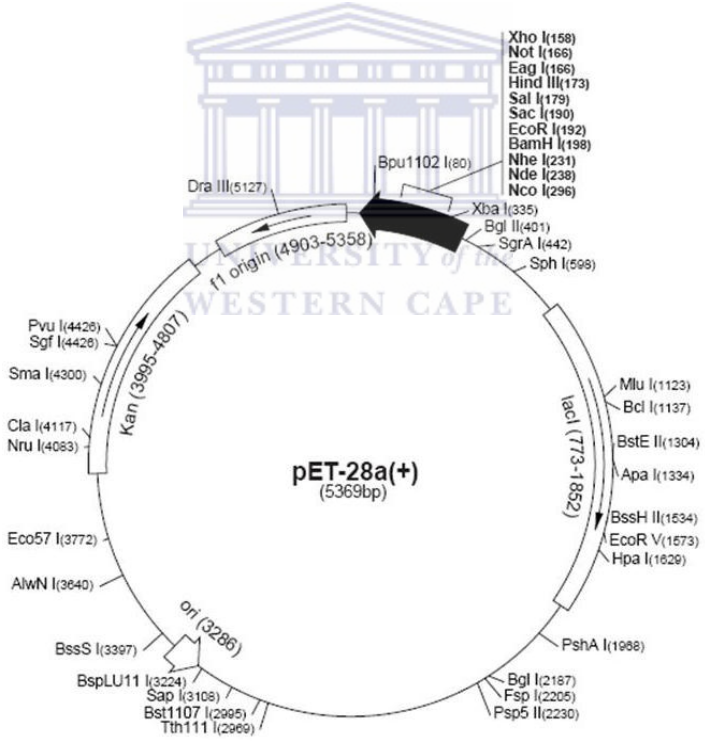


Figure 2.3 A vector map of the pET28a(+) expression vector used to sub-clone and express the *celMHS* gene. The positions of the restriction enzyme recognition sequences, in bp, are indicated in brackets. The position of the Kan antibiotic resistance gene is indicated (Novagen).

2.5.4 Comparative expression trial

E. coli Rosetta 2 and *E. coli* BL21 (Table 2.1) transformants harbouring pET28_*celMHS* and pET28_*celMHSΔTMR*, respectively, were selected for a comparative expression trial to determine which strain had enhanced expression of the recombinant protein. Briefly, a single colony of each of the above strain harbouring the respective recombinant plasmid was inoculated into 5 ml LB containing 12.5 µg/ml Cam and 50 µg/ml Kan for the *E. coli* Rosetta 2 clones and 50 µg/ml Kan for the *E. coli* BL21 clones, and incubated for 16 hrs at 37°C with agitation (225rpm). The entire volume of the overnight cultures was inoculated into 100 ml LB supplemented with the appropriate antibiotics, and incubated at 37°C with agitation until an OD_{600}, determined spectrophotometrically on a Biomate3 spectrophotometer (Thermo Scientific, USA) of approximately 0.6 was obtained. Subsequently, expression of the protein of interest induced with 0.5 mM Isopropyl β-D-1-thiogalactopyranoside (IPTG) at 37°C for 4 hrs following induction. Thereafter, the cells were harvested by centrifugation at 6,000x*g* for 10 min at 4°C and the pellets stored at -20°C overnight.

The frozen cell pellets were defrosted on ice, before being gently resuspended in pre-chilled 0.1 M phosphate buffer, pH 6.0. The resuspended cells were lysed by sonication, as described in section 2.2.6, and analysed using SDS PAGE electrophoresis. Briefly, the soluble and insoluble proteins were analysed by SDS-PAGE according to (Laemmli, 1970). SDS-PAGE gels were cast using the Hoefer® Mighty Small System (Hoefer incorporated, USA), with a 10% (w/v) polyacrylamide separating gel [0.375 M Tris-HCl (pH 8.8), 0.1% (w/v) SDS, 9.9% (w/v) acrylamide (Sigma), 0.264% (w/v) bis-acrylamide, 0.05% (w/v) ammonium persulphate and0.1% (v/v) tetramethylethylenediamine (TEMED)] and 4% (w/v) polyacrylamide stacking gel [0.125 M Tris-HCl (pH 6.8), 0.1% (w/v) SDS, 4% (w/v) acrylamide, 0.1% (w/v) bisacrylamide, 0.03% (w/v) ammonium persulphate and 0.1% (v/v) TEMED]. Sample preparation consisted of mixing equal volumes of 2x loading dye [80 mMTris-HCl (pH 6.8), 10% (v/v) glycerol, 2% (v/v) SDS, 1% (v/v) β-mercaptoethanol and 0.02% (w/v) bromophenol blue] and the protein sample. Thereafter, the protein samples were heated to 95°C for 5 min prior to loading

onto the polymerized polyacrylamide gel. The protein samples were separated under denaturing conditions at a constant voltage of 100 and 150 volts through the stacking and separating gels, respectively, until the dye front had migrated off the polyacrylamide gel. Following electrophoresis, the polyacrylamide gel was visualised by staining with Coomassie Brilliant Blue solution [40% (v/v) methanol, 10% (v/v) glacial acetic acid and 0.025% (w/v) Coomassie Brilliant Blue R-250] for 2 hrs and then destained with destaining solution [20% (v/v) methanol and 10% (v/v) glacial acetic acid] for 16 hrs(Sambrook and Russell, 2001). The molecular weight was determined by comparison to the PageRuler (Fermentas, Germany) molecular weight ladder separated on the same polyacrylamide gel.

Zymograms were prepared for the determination of cellulase activity, according to the method of Watson *et al.* (2009) with minor modifications. The soluble fraction of *E. coli* Rosetta 2 (pET28_*celMHS*) and *E. coli* Rosetta 2 (pET28_ ***celMHSΔTMR***) as well as *E. coli* BL21 (pET28_*celMHS*) and *E. coli* BL21 (pET28_ ***celMHSΔTMR***) were prepared and run in duplicate, as described in section 2.5.4,on SDS-PAGE gels containing 1% (w/v) CMC. One SDS-PAGE gel was stained and destained as described in section 2.5.4. The other SDS-PAGE gel was washed with 2.5% (v/v) Triton X-100 for 4hrs at room temperature. The SDS-PAGE gels were washed twice in distilled water and further incubated overnight in 0.2 M MES pH 6.0. Subsequently the gel was washed in 0.1% (w/v) Congo red solution for 2 hrs and cellulose activity visualized following destaining the gel in 1 M NaCl solution for 30 min.

The soluble fraction of *E. coli* Rosetta 2(pET28_*celMHS*) and *E. coli* Rosetta 2 (pET28_ ***celMHSΔTMR***) as well as *E. coli* BL21 (pET28_*celMHS*) and *E. coli* BL21 (pET28_ ***celMHSΔTMR***) were used in a comparative liquid assay to assess which expression strain resulted in the optimal heterologous expression of the recombinant protein, CelMHS. The protein concentration of the recombinant constructs was determined using a modified Bradford assay (Bradford, 1976). Briefly, the Bradford assay was performed by combining 5 µl of the protein sample and 200 µl of Bradford's reagent (Sigma) in a well of a flat-bottom 96-well microtitre plate. The samples were mixed by

gently pipetting up and down several times and the reaction incubated at room temperature for 20 min, before the absorbance was determined at 595nm using a microplate reader (Spectrostarnano, BMG Labtech, Germany). A standard curve was generated using a set concentration range (0.2-1.6 mg/ml) of Bovine Serine Albumin (BSA) prepared in buffer, according the manufacturer's instructions. The concentration of the BSA standards was confirmed at 595 nm (OD_{595}) using the microtitre plate (MTP) reader. The protein concentration of the unknown samples was determined by comparison to the standard curve of the BSA standards.

Cellulase activity was determined by measuring the amount of reducing sugar released following cellulase-catalyzed hydrolysis of the substrate CMC (Sigma) using the dinitrosalicylic acid (DNS) reducing sugar assay (Miller, 1959).Briefly, the reaction mixture consisted of 35 µl of the soluble cell extract prepared as described above, was assayed with 0.3% (w/v) CMC in 100 mM sodium phosphate buffer (pH 6). The enzyme reaction was performed for 20 min at an incubation temperature of 80°C, before the reaction was terminated by the addition of 150 µl of DNS reagent [21.6% (w/v) NaK-tartrate (Merck), 1.4% (w/v) NaOH, 0.7% (w/v) 3,5-Dinitrosalycyclic Acid (DNS) (Sigma), 0.6% (w/v) Na-Metabisulphate, 0.5% (w/v) phenol (Sigma)] and boiled for exactly 5 minutes, before rapidly cooling on ice. Following the addition of 800 µl distilled water, the absorbance at 510 nm (OD_{510}) was determined spectrophotometrically (Biomate 3, Thermo Fischer Scientific, USA). For the assay blank the enzymatic reaction was prepared, as described above, excluding the soluble protein extract but with the addition of 35 µl 0.1 M phosphate buffer, pH 6.0 and assayed as described above. The amount of reducing sugar release was determined using a standard curve generated with D-glucose (0-1.6 µg/ml) (Sigma) assayed using the DNS method, as described above. Subsequently the quantity of reducing sugar liberated from the substrate was determined by comparison of the absorbance readings to the D-glucose standard curve. One unit of cellulose activity (U) is the amount of enzyme required to release 1 micromole of reducing sugar per min per milligram protein (µmol/min/mg). All experiments were carried out in triplicate.

2.6 Protein expression and purification

2.6.1 Small scale expression of CelMHSΔTMR

E. coli Rosetta 2 harbouring pET28_*celMHSΔTMR* was investigated in small-scale expression trials to determine the optimal IPTG concentration for induction of *celMHS* expression. Briefly, single colonies of *E. coli* Rosetta 2 clones harbouring pET28_ *celMHSΔTMR* or the pET28a(+) control vector were inoculated into 5 ml LB supplemented with 12.5 ug/ml Cam and 50 ug/ml Kan, and incubated for 16 hrs at 37°C with agitation (225rpm). Three ml of the overnight culture was used as starter culture to inoculate 250 ml flasks containing50 ml LB supplemented with 12.5 µg/ml Cam and 50 µg/ml Kan. The cultures were incubated at 37°C with agitation until an OD_{600} (Biomate3, Thermo Scientific, USA) of approximately 0.6 was obtained. Protein expression was induced by the addition of 0.05, 0.1 or 0.5 mM IPTG. For the pET28a(+) control only 0.5 mM IPTG was added. Cultures were further incubated at 37°C for four hrs and the cells harvested by centrifugation at 6,000x*g* for 10 min at 4°C. Cell pellets were stored at -20°C overnight. Soluble and insoluble cell fractions were prepared and analysed by SDS-PAGE and liquid activity assays as described previously (section 2.5.4).

2.6.2 Large scale expression and purification of celMHSΔTMR

A single colony of *E. coli* Rosetta 2 (Table 2.1) harbouring pET28_*celMHSΔTMR* (Table 2.2) was inoculated into 5 ml LB containing 12.5 ug/ml Cam and 50 ug/ml Kan, and incubated for 16 hrs at 37°C with constant agitation at 225rpm. The entire volume of this overnight culture was used as starter culture to inoculate 200 ml LB containing 12.5 ug/ml Cam and 50 ug/ml Kan, and incubated at 37°C with agitation until an OD_{600} of 0.6 was obtained. A final concentration of 0.5 mM IPTG was added to the culture and incubated at 37°C for an additional 4 hrs. Thereafter, the cells were harvested by centrifugation (6,000x*g* for 10 min at 4°C) and the pellet stored at -20°C overnight. The frozen cell pellets were defrosted on ice, before being resuspended in binding buffer [0.5 M NaCl, 5 mM imidazole and 20 mM Tris-HCl (pH 7.9)] and lysed by sonication, as

described in section 2.2.6.The insoluble cell debris was removed by centrifugation (6,000x*g* for 30 min at 4°C) and the supernatant containing the soluble proteins retained for subsequent nickel-affinity chromatographic purification of *celMHSΔTMR*.

The recombinantly expressed His-tag fused protein CelMHSΔTMR was purified by Ni-chelation chromatography with the His-Bind® Resin and Buffer kit (Novagen, USA), according to the manufacturer's instructions. Briefly, the column was prepared and equilibrated with 3 column volumes of distilled water, 5 column volumes of charge buffer (50mM $NiSO_4$) and 3 column volumes of binding buffer. The cell-free extract (section 2.8.3) containing CelMHSΔTMR was filtered through a 0.45 µm syringe filter (Millipore, USA), before being loaded onto the charged and equilibrated column using a peristaltic pump (Perista pump SJ-1211, Atto Corporation, Japan) at a flow rate of 1 ml/min. Thereafter, the column was washed with 10 column volumes of binding buffer and 6 column volumes of modified wash buffer [0.5 M NaCl, 30mM imidazole and 20 mMTris-HCl(pH 7.9)]. CelMHSΔTMR was eluted from the column with 6 column volumes of modified elution buffer [0.5 M NaCl, 250 mM imidazole and 20 mM Tris-HCl (pH 7.9)]. Fractions were collected at each step of the Ni-chelation chromatography process for analyses by SDS-PAGE (Section 2.8.1.). Column purified CelMHSΔTMR was subsequently dialysed in dialysis tubing with a 10,000 kDa molecular weight cut off(Thermo Fisher Scientific, USA) for 16 hrs against approximately 200 volumes of 0.2 mM2-*N*-morpholinoethanesulfonic acid(MES) buffer (pH 6.0) at 22°C. Purified CelMHSΔTMR was stored at 4°C following dialysis.

2.7 Cellulase enzyme assays

2.7.1 Determining cellulase activity

Cellulase activity was determined using and 1 µg purified CelMHSΔTMR was used in a 50 µl assay performed as described previously in section 2.5. A Bradford assay was performed, as previously described in section 2.5.

2.8 Characterisation of CelMHSΔTMR

2.8.1 Determining the linear rate of CelMHSΔTMR activity

In order to determine the linearity of CelMHSΔTMR enzyme activity, the purified enzyme, was assayed at a constant CMC concentration over 2 to 60 min. CelMHS activity was assessed at 0.3% (w/v) CMC and the reducing sugars liberated were measured by the DNS reducing sugar assay, as described in section 2.5.4.

2.8.2 Temperature and pH optima of CelMHSΔTMR

In order to assess the effect of pH on CelMHSΔTMR, purified CelMHSΔTMR was assayed, as described in section 2.5.4 across a pH range of 3.5-9.0. The buffers used were 50 mM citrate buffer (pH 3.5-5.5), 50 mM MES buffer (pH 5.5-7.0) and 50 mMTris-HCl buffer (pH 7.0–9.0). Similarly, purified CelMHSΔTMR was assayed in 50 mM MES buffer (pH 6.0) at various incubation temperatures (40, 50, 60, 70, 80, 90, 95 and 100°C) to assess the effect of incubation temperature on CelMHSΔTMR activity. A negative control was included for each of these temperatures to test the non enzymatic hydrolysis of CMC at these high temperatures.

2.8.3 CelMHSΔTMR Thermostability

To determine the ability of CelMHSΔTMR to resist irreversible changes after incubation at high temperatures a thermostability assay was performed. CelMHSΔTMR was incubated at temperatures on of 60, 70, 80, 90 and 100°C and enzyme aliquots were removed at time intervals of 15, 30, 45 and 60 min and immediately placed on ice for the remainder of the experiment. In addition, an aliquot of CelMHSΔTMR was incubated on ice for the duration of the experiment as an untreated control. Thereafter, the effect of thermal pre-incubation was determined by comparing the residual cellulase activity, determined by assaying the enzyme (as described in section 2.5.4) in 50 mM MES (pH 6.0), of the heated samples to the untreated control.

2.8.4 Substrate specificity

The substrate specificity of purified CelMHSΔTMR was assessed against various polymeric substrates, including: 0.2% (w/v) rye arabinoxylan (Megazyme,Ireland); 0.2% (w/v) CMC (Sigma); 0.2% (w/v) avicel (Sigma); 0.2% (w/v) β-Mannan (Megazyme); 0.2% (w/v) birchwoodxylan (Sigma); 0.2% (w/v) beechwoodxylan (Sigma); 0.2% (w/v) arabinan (Megazyme); 0.2% (w/v) laminarin (Megazyme); 0.2% (w/v) β–Glucan (Megazyme), and 0.2 % (w/v) lichenan (Megazyme). The activity of CelMHSΔTMR on each substrate was determined after performing a liquid assay in 50 mM MES (pH 6.0) using the DNS reducing sugar assay and comparing the data to a glucose standard curve (section 2.5.4). A no enzyme control, or blank reaction lacking purified CelMHSΔTMR, was performed for each substrate.

2.8.5 Steady state kinetics of CelMHSΔTMR

For the determination of CelMHSΔTMR steady state kinetics, the initial rate of the enzyme reaction was measured over a CMC substrate concentration range between 1 and 20 mg/ml, by performing a liquid assay in 50 mM MES (pH 6.0) and determining the amount of reducing sugars released by DNS assay, as described in section 2.5.4.

Kinetic parameters, K_M and V_{MAX}, were determined by non-linear fitting of data to hyperbolic curves according to Michaelis-Menten (GraphPad Prism v.4.00, GraphPad Software, Inc., San Diego, CA, USA).

2.9 CelMHSΔTMR hydrolysis of cello-oligosaccharides

The capacity of CelMHSΔTMR to hydrolyse cello-oligosaccharides was assessed in small scale trials and analysed by HPLC. Reaction mixtures were prepared for each cello-oligosaccharide, namely 100 mM glucose (Sigma), 100 mM cellobiose (Sigma), 100 mM cellotriose (Megazyme, Ireland), 100 mM cellotetrose (Megazyme), and100 mM cellopentose (Megazyme), in 50 mM MES buffer pH 6.0. Reactions were initiated by the addition of 1 µg of CelMHSΔTMR and allowed to proceed for 16 hrs at 60°C with constant agitation (225 rpm). In addition, undigested cello-oligosaccharides standardswere prepared in 50 mM MES buffer (pH 6.0). The CelMHSΔTMR oligosaccharide digestion products and undigested standards were analysed on a Dionex Ultimate 3000 HPLC (Thermo Scientific, USA). The samples were injected onto a Resex™ RSO oligosaccharide Ag+4% column (200 x 10.0 mm; Phenomenex, USA) at a flow rate of 0.3 ml/min using 30% (v/v) methanol, 1% (v/v) acetic acid as the mobile phase and maintained at 48°C throughout. Oligosaccharides were detected using the refractive index (RI) and the results analysed using Chromeleon Dionex version 6.80 software (Dionex Corporation, USA).

2.10 CelMHSΔTMR hydrolysis of CMC and β-glucan

The capacity of CelMHSΔTMR to hydrolyse CMC and β-glucan was assessed and analysed by HPLC. Reaction mixtures were prepared for each substrate in 50 mM MES buffer pH 6.0. Reactions were initiated by the addition of 1 µg of CelMHSΔTMR and allowed to proceed for 16 hrs at 60°C with constant agitation (225 rpm). Hydrolysis of the substrates by CelMHSΔTMR were analysed on a Dionex Ultimate 3000 HPLC (Thermo Scientific, USA). The samples were injected onto a Resex™ RHM Monosaccharide H^+ column (300 x 7.8 mm; Phenomenex, USA) at a flow rate of

0.3 ml/min using 500mM H_2S0_4 as the mobile phase and maintained at 48°C throughout. Hydrolysis products were detected using the refractive index (RI) and the results analysed using Chromeleon Dionex version 6.80 software (Dionex Corporation, USA).

2.11. Pre-treatment of sugarcane bagasse

Sugarcane bagasse (SCB) provided by Prof. J Görgens (Process Engineering, University of Stellenbosch, South Africa) was pre-treated by Dr. Garcia Aparicio and Paul McIntosh (Process Engineering, University of Stellenbosch, South Africa) by steam explosion, 195°C for 11 min, using a steam gun. The properties of the resulting slurry in percentage composition was 37.9±1.4% total solids, of which 32.6±0.8% was water insoluble (WIS) and 5.3±0.7%water soluble (WSS). Steam exploded SCB was manually pressed using a 4 ton hydraulic jack to separate the liquid and solid fractions. The SCB solid fraction was washed with 200 volumes of room temperature tap water to remove residual sugars and contaminants. The remaining water was removed from the washed SCB by low speed centrifugation, before the SCB was spread out onto large trays and dried in an oven at 30-40°C until it reached a relative moisture content of approximately 10% (w/v). The dried SCB was thoroughly mixed and milled to a particle size of less than 0.5 mm using an ultracentrifugal mill (Retch ZM200, Haan, Germany). All composition analyses of the sugarcane bagasse were performed as outlined by the National Renewable Energy Laboratory (NREL; www.nrel.gov/biomass/analytical procedures.html) by Dr M Garcia Aparicio. Briefly, the solid SCB fraction consisting of glucan, xylan, arabinan and lignin was analysed by the standard methods, as described by (Sluiter *et al*., 2008). Sugars present in the solid fraction were further analysed by HPLC. Samples were diluted to within the detectable HPLC range, before being centrifuged (16,000x*g* for 5 min at 22°C) and filtered through 0.22 µm syringe filters to remove particulate matter. Analysis of D-glucose, D-xylose and L-arabinose was performed on a Dionex Ultimate® 3000 system equipped with a Carbo Pac PA1 column operated at 25°C with a mobile phase of 30 mM sodium hydroxide and a flow rate of 1 ml/min. In addition, sugars present in the liquid fraction were assessed by performing a

1/10 dilution of the liquid fraction and resolved on a Rezex R50 oligosaccharide column (Phenomenex). Monosaccharide standards (D-fructose, D-xylose, D-glucose, D-sucrose and L-arabinose) were prepared at a concentration of 40 mM and separated under the same conditions. Sugars were analysed before, and after, hydrolysis with 4% (v/v) sulphuric acid to measure both monomeric and oligomeric sugars. Approximately 20 µl per sample was injected onto the column and a run time of 40 min in a mobile phase of double distilled water was employed, while the column was maintained at 48°C throughout.

2.12 CelMHSΔTMR hydrolysis of pre-treated sugarcane bagasse

The ability of CelMHSΔTMR to hydrolyse pre-treated SCB was assessed to determine the suitability of this enzyme for large scale biomass hydrolysis. Briefly, all hydrolysis reactions were performed in 100 mM MES buffer (pH 6.0) in 2 ml volumes. Approximately 50 µg CelMHS protein was used to hydrolyse approximately 50mg SCB [2% WIS (w/v) equating to approximately 1% glucan]. Concurrently, 50 µg CelMHS protein was used to hydrolyse approximately 0.3% (w/v) CMC as a positive control under the same hydrolytic conditions. Any possible inhibitory effects of SCB on CelMHS enzyme activity were determined by incubating the enzyme, as described above, with 0.3% (w/v) CMC in the presence of 50 mg SCB. In addition, the SCB was hydrolysed using 15 U of a commercial *Tricoderma reesi* cellulase preparation (Sigma) under the same reaction conditions. For all reactions a no enzyme background control, comprising SCB and 100mM MES buffer pH 6.0, was performed to measure the non-enzymatic release of reducing sugars over the course of the incubation period. The reaction mixtures were allowed to proceed for approximately 14 hrs at 60°C with shaking (225 rpm), after which the free reducing sugars were measured by the DNS reducing sugar assay, as described in section 2.5.4.

Chapter 3

Results and Discussion

3.1 Functional screening of a cellulase-producing fosmid clone

High molecular weight metagenomic DNA was previously isolated from Malawian hot spring sediment and used to construct a fosmid metagenomic library by Xiao Ping Hu (MSc project, unpublished data). The library contained approximately 10,000 clones with an average insert size of approximately 30 kb. The library was functionally screened for genes conferring cellulase activity. An *E. coli* Epi300 fosmid clone, pFos_032c10, with extracellular cellulase activity indicated by the zone of clearing surrounding the colony when cultivated on CMC was identified (Fig. 3.1A). This clone formed the basis of this investigation.

Functional based screening of metagenomic libraries may result in the discovery of entirely novel functional genes from cloned metagenomes (Simon and Daniels, 2009). These methods have been employed to identify numerous genes encoding cellulases from various environmental niches, including soil (Voget *et al*., 2006; Kim *et al*., 2008; Jiang *et al*., 2009; Lui *et al*., 2010; Nacke *et al*., 2011), buffalo rumen (Nguyen *et al*., 2012) and the gut microflora of marine invertebrates, such as abalone (Kim *et al*., 2011). During the current study, the cellulolytic activity of the fosmid clone, pFos_032c10, identified from a hot spring metagenomic library was confirmed by re-transforming the fosmid into *E. coli* Epi300 cells. The resulting *E. coli* transformants were observed to have extracellular cellulase activity on RBB-CMC screening plates (Fig. 3.1B). This confirmed that the cellulase activity was encoded by the genetic information contained within pFos_032c10.

3.2 Sequencing and de novo sequence assembly of pFos_032c10

The recombinant fosmid, pFos_032c10, was pooled with a number of other fosmids and sequenced. Following the *de novo* assembly of the 454 GS-FLX sequence reads, assembled contigs were aligned to the end sequences of pFos_032c10, obtained from Sanger sequencing runs using the T7 forward and pCC1Fos reverse primers to identify

the contig(s) which contained the sequence information of pFos_032c10. A single contig (Contig#23) 11,294 bp in length was identified (Supplementary Figure 1). The sequence of Contig#23 had a %G+C content of 58%. In addition, bioinformatic analysis of Contig#23 identified a total of ten putative open reading frames (ORFs) (Table 3.1 and Fig. 3.2), and all showed relatively low sequence similarity to proteins contained within the Genbank database. Approximately 50% of the predicted proteins within Contig#23 showed the greatest similarity to proteins that originate from thermophilic microorganisms. This was not surprising, as the metagenomic DNA used for library construction was from a thermophilic source, sediment from Mphizi hot springs (Chiweta, Malawi), which can reach temperatures of close to 80°C.

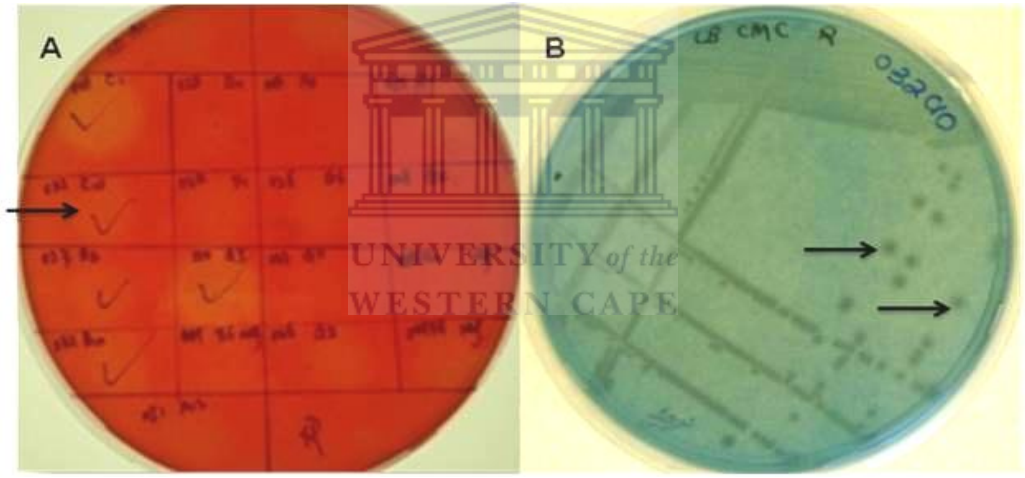


Figure 3.1 Secondary functional screening of recombinant fosmid clones from a Malawian hot spring sediment metagenomic library for cellulase activity on CMC (A; courtesy of X.P. Hu) and functional confirmation of the cellulolytic phenotype of *E. coli* Epi300 cells harbouring fosmid clone pFos_032c10 on RBB-CMC (B). The arrows indicate the zones of CMC hydrolysis surrounding *E. coli* Epi300 pFos_032c10 colonies following incubation at 37°C for 5 days on screening plates.

None of the putative ORF's on Contig#23 showed significant homology to previously characterised cellulolytic genes. However, one of the putative genes (Gene no. 3; Table 3.1 and Fig. 3.2) showed homology to a carbohydrate–binding CenC domain containing

protein from the thermophilic archaean species, *Ignisphaera aggregans* (Niederberger *et al*., 2006). As discussed in chapter 1, CBMs are a common feature of many cellulases. Therefore, we hypothesized that this putative ORF (Gene no. 3) could be responsible for the cellulolytic activity encoded by pFos_032c10. We designated this putative cellulolytic gene *celMHS* (cellulase from a Malawian Hot Spring) and investigated it further to confirm whether it encoded a protein with cellulolytic activity and whether this protein had characteristics similar to other known cellulases.

Table 3.1 Sequence identity of predicted ORF's in Contig23 to sequences in the GenBank database (accessed 22/02/2013).

Gene no.	Gene size (bp)	Putative gene function/Characteristic features	Accession no.	% Identity	E-value	Organism
1	767	Hypothetical protein; putative membrane protein	YP_003205282.1	44	9e-63	*Methylomirabilis oxyfera*
2	305	30S ribosomal protein S20	YP_004603483.1	45	1e-13	*Flexistipes sinusarabici* DSM 4947
3	2,715	Carbohydrate-binding CenC domain-containing protein	YP_003860538.1	44	0.0	*Ignisphaera aggregans* DSM 17230
4	861	Xylose isomerase; sugar phosphate isomerase	ZP_11048173.1	32	4e-14	*Acinetobacter ursingii* DSM 16037
5	1,089	Oxidoreductase domain-containing protein	ZP_21557377.1	34	3e-41	*Natrialba aegyptia* DSM 13077
6	918	Periplasmic binding protein; ABC substrate-binding transporter protein	YP_003825802.1	39	1e-59	*Thermosediminibacter oceani* DSM 16646
7	849	Endonuclease V	ZP_08429101.1	50	2e-61	*Moorea producens* 3L
8	807	Hypothetical protein	ZP_03725128.1	31	5e-17	*Diplosphaera colitermitum* TAV2
9	984	Hypothetical protein	ZP_11092823.1	47	1e-81	*Schlesneria paludicola* DSM 18645

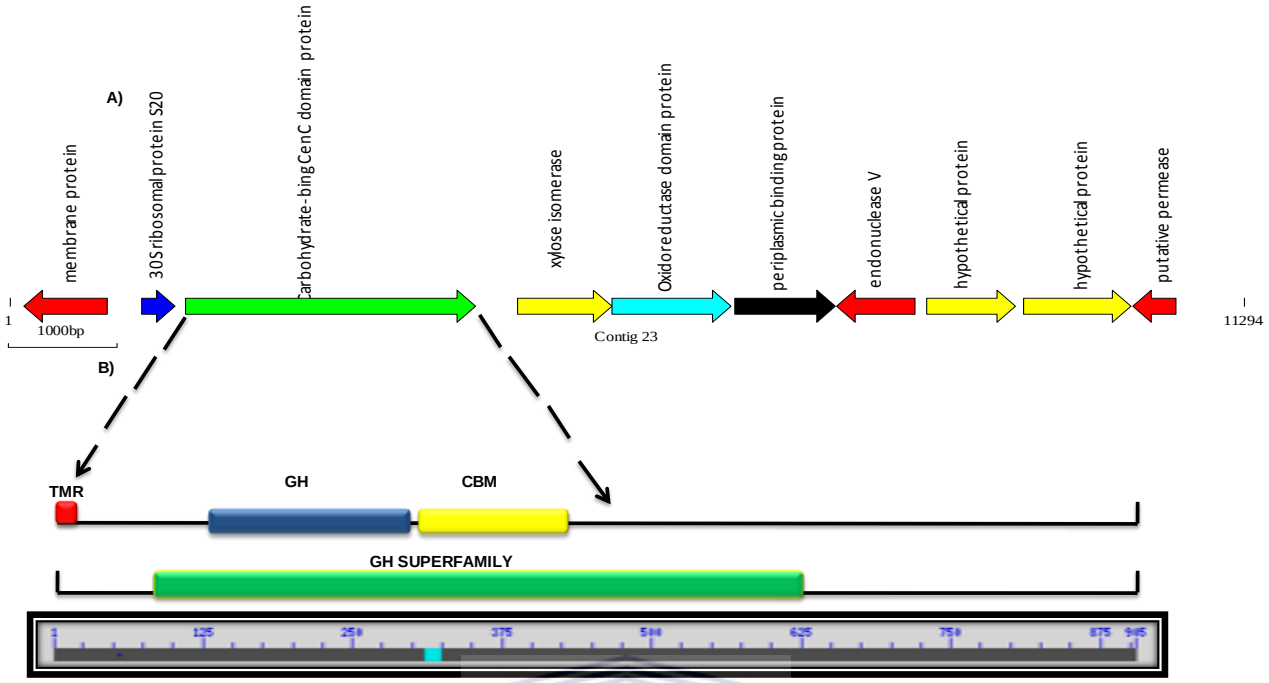


Figure 3.2 Diagrammatic representation of the ORFs predicted within Contig#23 (A) and conserved domains architecture within the deduced CelMHS protein sequence (B). The coloured block arrows in A indicate the position, orientation and relative size of the predicted ORFs (www.ncbi.nlm.noh.gov). TMR, transmembrane region; GH, glycoside hydrolase and; CBM, Carbohydrate binding domain.

3.3 Sequence analysis and classification of celMHS

The predicted gene sequence of *celMHS* comprised of 2,715 nucleotides and analysis of the upstream sequence predicted the presence of a -10 and -35 box at positions -15 and -36, upstream of the predicted transcription start site (Fig. 3.3) (BPROM; www.softberry.com). A transcription factor binding site, rpoD15, was also identified at position -12. The nucleotide sequence of *celMHS* encodes a putative protein of 905 amino acids (Fig 3.3) with a predicted molecular mass of approximately 98 kDa. Sequence analysis of CelMHS revealed relatively low sequence identity to known proteins within the NCBI database (Table 3.2). However, regions of homology to conserved GH family 42 (β-galactosidases) and CBM domains were evident within the deduced CelMHS amino acid sequence. A transmembrane helix was predicted when

the secondary structure of CelMHS was analysed on the TMHMM server Version 2.0 (http://www.cbs.dtu.dk/ services/TMHMM/) (Fig. 3.3). The transmembrane helix is predicted to span amino acids number five to 22, while the first four amino acids are predicted to be intracellular and the amino acids 23 to 905, which include the catalytic and carbohydrate binding domains, protruding to the exterior of the membrane. This hydrophobic transmembrane region may anchor the protein within the lipid bilayer of the native host organism's cell membrane, thus allowing close proximity of the hydrolytic products produced from enzymatic action outside the cell.

```
-177      CATTGACAAGACGGCGTCTGCGTGCCGCATCATTGACAAGACGGCGTCTAAAGGGGTCAT

-118      CAAAAAAGGTGCAGCTGCCCGCTATAAGTCCCGCTTAATGCGCCGACTGAACGAGCTCAT

-59        CCATCAAAAGGCTGTCGTTGAAGCGGCGTAAGGGCTGTAGCAGTGTTGGGGTGATGAAAG

1          ATGCGCGGGTGGTGGCTGTTGGTGGCTTGTGCCATGTTGGCTGGTATGGGAGGATGGTGG
1           M  R  G  W  W  L  L  V  A  C  A  M  L  A  G  M  G  G  W  W
61         ACAATGCGACAGGGAGCATCCCAAGCAACGCCGCAAACTGCTTCCCCACTGTTCCCTTTC
21          T  M  R  Q  G  A  S  Q  A  T  P  Q  T  A  S  P  L  F  P  F
121        GTGCTCCCGTGGGATGACAACGCACCGTCGGTGACCAACCTCAACGGTTGGCTCCACCAA
41          V  L  P  W  D  D  N  A  P  S  V  T  N  L  N  G  W  L  H  Q
181        CCAGCGGGAAGGTTCGGGCATGTGCGCGTGGGTGCGGACGGGCACCTCTATGTCGGTCAG
61          P  A  G  R  F  G  H  V  R  V  G  A  D  G  H  L  Y  V  G  Q
241        CGGCGCATCCGCTTTTTCGGAGTCAACCTGTGCTTCGGCGCTTGCTTCCCCCGCAAAGAA
81          R  R  I  R  F  F  G  V  N  L  C  F  G  A  C  F  P  R  K  E
301        GATGCCGAAAAGATTGCTGCCCGCATGGCGAAGTTCGGCATCAACATCGTCCGCTTCCAC
101         D  A  E  K  I  A  A  R  M  A  K  F  G  I  N  I  V  R  F  H
361        CACATGGACATGAACCCCTTCCCAAACGGGATCCGTGCCCGCAATGTCCCCCACACCCGC
121         H  M  D  M  N  P  F  P  N  G  I  R  A  R  N  V  P  H  T  R
421        GACCTTGACCCTGAAGCTTTAGATCGCTTGGACTACCTGATTGCCCAATTGAAACGCCAC
141         D  L  D  P  E  A  L  D  R  L  D  Y  L  I  A  Q  L  K  R  H
481        GGCATCTATGTCAACTTGAACTTGTTGGTCTCTCGCCCCTTCAATGCTGCCGACGGGTTG
161         G  I  Y  V  N  L  N  L  L  V  S  R  P  F  N  A  A  D  G  L
541        CCGAAAGAGATTGAGCAACTCGGCTGGAAAGAACGGCACATTGTCGGCTTTTTCTATGAG
181         P  K  E  I  E  Q  L  G  W  K  E  R  H  I  V  G  F  F  Y  E
601        CCATGTTTGGAGCTGCAGAAGGAGTATGCCCGCAAGCTGCTGACCCACCGCAACCCCTAC
201         P  C  L  E  L  Q  K  E  Y  A  R  K  L  L  T  H  R  N  P  Y
661        ACAGGATTGACTTATGCGGAAGCCCCCGTAGTCGCTTTCGTGGAAATCAACAATGAAAAT
221         T  G  L  T  Y  A  E  A  P  V  V  A  F  V  E  I  N  N  E  N
721        GGCTTGTTGCACGCTTGGTTGGGAGGCGATGTGGACCAATTGCCGCCAGTGTTTCTGCAA
241         G  L  L  H  A  W  L  G  G  D  V  D  Q  L  P  P  V  F  L  Q
781        GAATTGCAGCGGCAGTGGAATGGGTGGCTCAAAGCCCGCTACGGCAGCACGGCAAAGCTG
261         E  L  Q  R  Q  W  N  G  W  L  K  A  R  Y  G  S  T  A  K  L
841        CGGCAAGCATGGGGCGTTCAGGAAGAACCGTTAGGCACCGAACTGCTGCAAAACACCACC
281         R  Q  A  W  G  V  Q  E  E  P  L  G  T  E  L  L  Q  N  T  T
901        TTCGCGACAGGGTTGCAGGGATGGGTGTTGGAGCAACACGCCAACGCCCAAGCGGCAGCG
301         F  A  T  G  L  Q  G  W  V  L  E  Q  H  A  N  A  Q  A  A  A
```

```
961         GAAGTGGTAGCCGACCCCATTCCCGAGTTGAAAGGGATGCGCTTTGTGCGCATCACCGTC
321          E  V  V  A  D  P  I  P  E  L  K  G  M  R  F  V  R  I  T  V
1021        ACCCGCCCAGGGACAGCAGGGTGGCATGTGCAATTTCACCAACCAGGGCTCAAAGTTGCC
341          T  R  P  G  T  A  G  W  H  V  Q  F  H  Q  P  G  L  K  V  A
1081        CCCGATCGCCCTTACACCCTGTCCTTCTGGGCGAGAGCGGAGCGCCCTTGCCGCATCACC
361          P  D  R  P  Y  T  L  S  F  W  A  R  A  E  R  P  C  R  I  T
1141        GTCGGTCTCAGTCAAGCCCACGAACCGTGGCAAAACTTGGGCTTCTCCGCCAATGTGGAC
381          V  G  L  S  Q  A  H  E  P  W  Q  N  L  G  F  S  A  N  V  D
1201        CTGACGCAAGAGTGGCGGGAGTATCGGTTCACCTTCGCCCTCGCCCGCGGGGACGACAAT
401          L  T  Q  E  W  R  E  Y  R  F  T  F  A  L  A  R  G  D  D  N
1261        GCGCGGGTTATCTTCAGCAATCTGGGTGCCCAGACAACCACCTATTGGTTTGCCGCGCCT
421          A  R  V  I  F  S  N  L  G  A  Q  T  T  T  Y  W  F  A  A  P
1321        TCCTTGCGCCCTGGAGGGATTGTGGGCTTGAAGGAAGACGAACGGATAGAAGACACCACC
441          S  L  R  P  G  G  I  V  G  L  K  E  D  E  R  I  E  D  T  T
1381        GTGCCTCCCTTCCGACGGGACCGATTGGGCGAACGAACACCGGAAGCGCAACGGGATTGG
461          V  P  P  F  R  R  D  R  L  G  E  R  T  P  E  A  Q  R  D  W
1441        ATGCGGTTCCTGTGGGAAACGGAAGATCGCTACTGGCAGACCCTCTACCGCTACTTGAAA
481          M  R  F  L  W  E  T  E  D  R  Y  W  Q  T  L  Y  R  Y  L  K
1501        GACGAGTTGAAAGTTAAGGCGTTGGTCATGGGCACCATCGTTGGTTGCAGCACCCCCAAC
501          D  E  L  K  V  K  A  L  V  M  G  T  I  V  G  C  S  T  P  N
1561        TTAATGGCGAAACTGGACGGGGTGGACACCCATGCCTATTGGCAGCACCCCCTGTTTCCC
521          L  M  A  K  L  D  G  V  D  T  H  A  Y  W  Q  H  P  L  F  P
1621        ACCCGCCCGTGGGACCCTGAAGACTGGATTGTGCCGAACCGCACAATGGTGAATGAGCGC
541          T  R  P  W  D  P  E  D  W  I  V  P  N  R  T  M  V  N  E  R
1681        GGGGGCACCTTGCCTGGGCTCGCTTTGCGGCGCGTGTTAGGCAAACCCCATTCGTGCACC
561          G  G  T  L  P  G  L  A  L  R  R  V  L  G  K  P  H  S  C  T
1741        GAATACAACCATCCCGCTCCCAATACTTACAGCAGTGAAGGCTTCTTGCTGTTAGCTGCT
581          E  Y  N  H  P  A  P  N  T  Y  S  S  E  G  F  L  L  L  A  A
1801        TATGCTGCCCTGCAGGATTGGGATGCTATCTACGCTTTCTCCTACTCCCACCGCCGTGAC
601          Y  A  A  L  Q  D  W  D  A  I  Y  A  F  S  Y  S  H  R  R  D
1861        GAATGGGATTTACGCCGCATCCCCAACTTCTTTGACATTGACCAGCACCCCACGAAGATG
621          E  W  D  L  R  R  I  P  N  F  F  D  I  D  Q  H  P  T  K  M
1921        GTGACGCTGGTCCCCGCCGCAGCACTGTTCTTGCGCGGGGATGTGAAACCGGCGAAGCAG
641          V  T  L  V  P  A  A  A  L  F  L  R  G  D  V  K  P  A  K  Q
1981        CAGGTGGTCGTGCGGCTGACAAAGGAGCAAGAAGTGGACTTGCTGCGCCGCAGTTGGGCG
661          Q  V  V  V  R  L  T  K  E  Q  E  V  D  L  L  R  R  S  W  A
2041        TGGGTCCTCGTGCATGCAGGGCATGTCGGGGTGCCGAACGAGGTCGCGCTTGTCCATCGC
681          W  V  L  V  H  A  G  H  V  G  V  P  N  E  V  A  L  V  H  R
2101        GTCGCCATCGCAACGGAGGGCAAACGGGTGCCGCCCACAGCGTTGAAACCTGAACAAGTG
701          V  A  I  A  T  E  G  K  R  V  P  P  T  A  L  K  P  E  Q  V
2161        AAAATCGCAGGCGAGAAGTTTGTTTCAGACACAGGCGAGCTCGTGTGGGACTTGACCGAG
721          K  I  A  G  E  K  F  V  S  D  T  G  E  L  V  W  D  L  T  E
2221        AAGGGGCGGGGTGTCGTGACGATCAACACGCCCAACAGCAAAGCCGTCATCGGTTACGGG
741          K  G  R  G  V  V  T  I  N  T  P  N  S  K  A  V  I  G  Y  G
2281        GGCGGCAAAGTGTTCCCTTTGGGCAATGTGGTCATTGCGCCTGGTCAGACCCTGCAAGAA
761          G  G  K  V  F  P  L  G  N  V  V  I  A  P  G  Q  T  L  Q  E
2341        GGGTGGTGCGCCATCACCCTGACGGTTATGCAAGGCAAATTGCCTGCGAAGATGACCTCC
781          G  W  C  A  I  T  L  T  V  M  Q  G  K  L  P  A  K  M  T  S
2401        ACCGTGCCCCGCAATTTGCGGTTGCTGATCACCGCCACAGGCTATGCCGAGAACACCAAC
801          T  V  P  R  N  L  R  L  L  I  T  A  T  G  Y  A  E  N  T  N
2461        ATCGGTTGGAAAGAAGTGCCAGGGTATCCGCCCAAATCCAGCTGCGGTCGCAATTGGGGT
821          I  G  W  K  E  V  P  G  Y  P  P  K  S  S  C  G  R  N  W  G
2521        TCCCCTCCGTCGCTGGTGGAAGGCATTCCCGCCCGCATCACTTTGCCCCTGCCTGCCAAG
841          S  P  P  S  L  V  E  G  I  P  A  R  I  T  L  P  L  P  A  K
2581        CGCGTGAAAGCCTGGGCGTTGGACGAACGCGGGCACCGCAAAACCGCCTTGCCTGTCAGC
861          R  V  K  A  W  A  L  D  E  R  G  H  R  K  T  A  L  P  V  S
```

```
2641      GCCGATGCAAAGGGCAGCGCCGTGGTGGAAATCGGACCGCCGTATCAAACGCTGTGGTAT
881        A  D  A  K  G  S  A  V  V  E  I  G  P  P  Y  Q  T  L  W  Y
2701      GAAGTCGCTGTAGAGTGA
901        E  V  A  V  E  *
```

Figure 3.3 The nucleotide sequence of *celMHS* and the deduced amino acid sequence of CelMHS. The ATG start codon is highlighted in green, while the asterisk (*) and red highlighting indicates the TGA stop codon. The predicted transmembrane region (TMHMM Server Version 2.0) is highlighted in blue. The GH superfamily domain A^{62}-F^{631} is underlined (http://supfam.org/SUPERFAMILY/). The glycoside hydrolase family 42 (E^{103}-V^{286}) is highlighted in pink and the carbohydrate binding like protein domain (E^{294}-F^{425}) is highlighted in turquoise (http://pfam.sanger.ac.uk/). The 5' untranslated region, indicated in bold, was predicted using BPROM (www.softberry.com) and the -10 and -35 boxes are double underlined. A putative transcription factor binding site, rpoD15, is indicated in bold italics. The proposed transcription start site is indicated with an arrow.

The *in silico* deduced CelMHS sequence was analysed by BLASTp analysis against the NCBI protein database. CelMHS has low sequence identity to other carbohydrate binding domain proteins as the percentage identities are all below 50% (Table 3.2). According to Hoffman (2000) sequence identity is inferred when the alignment generated between a sequence of interest and in a queried database exceeds a specific alignment score, and the score that governs this is the e-value. Therefore, a low e-value indicates a high similarity between sequences. Across the CelMHS full length protein sequence, it showed the highest similarity to proteins derived from *Ignisphaera aggregans* (YP_003860538.1), *Chthoniobacter flavus* (ZP_03130152.1), *Verrucomicrobiae bacterium* (ZP_05055536.1), *Opitutus terrae* (YP_001820698.1) and *Teredinibacter turnerae* (YP_003074810.1; Table 3.2). Phylogenetic analysis of CelMHS was conducted using closely related protein sequences obtained from the BLASTp analysis (Table 3.2). A neighbour joining tree confirmed that CelMHS is most closely related to the Carbohydrate-binding CenC domain-containing protein of *Chthoniobacter flavus* (Fig. 3.4). To the best of our knowledge none of these closely related proteins have been characterised, and CelMHS is the only one that was

identified from a functionally-screened metagenomic library while the others were identified and annotated from microbial genome sequences.

Table 3.2 Identities of the closest related protein sequences, based on E-values, obtained from a BLASTp search of the GenBank database (accessed 28/02/2013), to the deduced amino acid sequence of CelMHS.

Organism	Accession no.	Description	% Coverage	% Identity	E-Value
Ignisphaeraaggregans DSM 17230	YP_003860538.1	Carbohydrate-binding CenC domain-containing protein	97	44	0.0
Chthoniobacterflavus Ellin428	ZP_03130152.1	Carbohydrate-binding CenC domain-containing protein	96	46	0.0
Verrucomicrobiae bacterium DG1235	ZP_05055536.1	Carbohydrate binding domain protein	96	41	0.0
Opitutus terrae PB90-1	YP_001820698.1	Carbohydrate-binding CenC domain-containing protein	96	42	0.0
Teredinibacterturnerae T7901	YP_003074810.1	Carbohydrate binding module family 4 domain-containing protein	95	38	0.0
Opitutaceae bacterium TAV5	ZP_09593983.1	Carbohydrate-binding CenC domain-containing protein	96	37	2e-157
Opitutaceae bacterium TAV1	ZP_10212840.1	Carbohydrate binding module family 4 domain-containing protein	91	33	6e-122
Verrucomicrobiumspinosum DSM 4136	ZP_02928539.1	Hypothetical protein VspiD_17855	98	33	2e-121
MelioribacterroseusP3M	YP_006527007.1	Carbohydrate-binding CenC domain protein	95	32	2e-120
Opitutaceae bacterium TAV5	ZP_09593046	Hypothetical protein Opit5DRAFT_1100	96	33	2e-115
Opitutus terrae PB90-1	YP_001819026.1	Hypothetical protein Oter_2143	95	33	9e-107
Diplosphaeracolitermitum TAV2	ZP_03728058.1	Carbohydrate-binding CenC domain-containing protein	48	41	6e-81

A sequence alignment of CelMHS with its closest sequences revealed that the putative catalytic residues such as asparagine (D^{140} and D^{174}), glutamate (E^{258} and E^{264}) and histidine (H^{137}, H^{138} and H^{240}) are well conserved in the region containing the catalytic domain which is represented by residues E^{103} to L^{286} (Fig. 3.5). Substrate binding residues such as tyrosine (Y), tryptophan (W) and histidine (H) are all conserved in the region of the CBM which consists of residues T^{296} to R^{444} (Tomme *et al*., 1996; Boraston *et al*., 2004). Interestingly, this region contains multiple tryptophan residues (W^{290}, W^{308}, W^{348} and W^{430}) suggesting that this region of the protein may be involved in substrate binding (Tomme *et al*., 1996; Boraston *et al*., 2004). A large region of the CelMHS protein sequence [Fig. 3.3 and 3.5 (indicated by a pink arrow)] shows no sequence similarity to previously characterized protein domains, however, this region is conserved between CelMHS and its closest related protein sequences. We may therefore speculate that this region contains an unclassified domain.

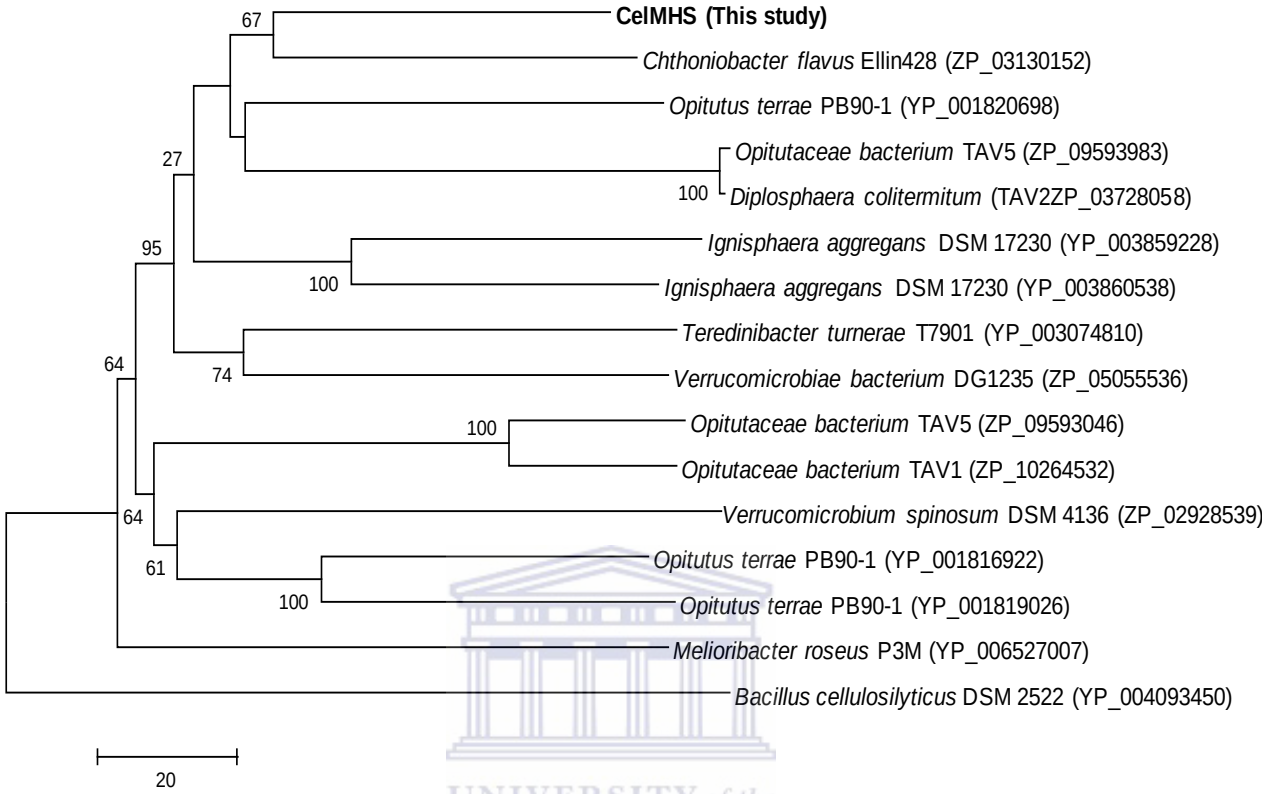
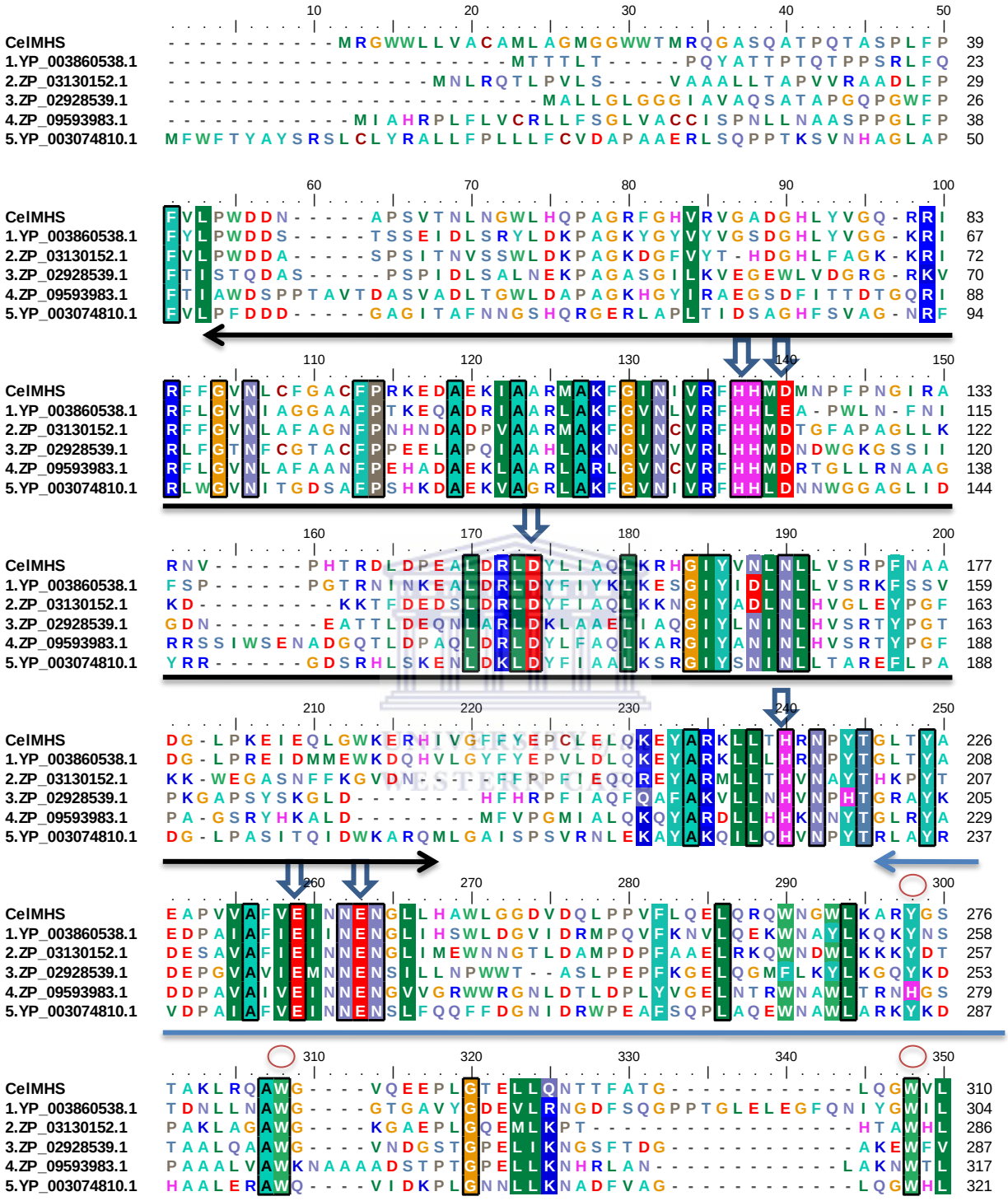
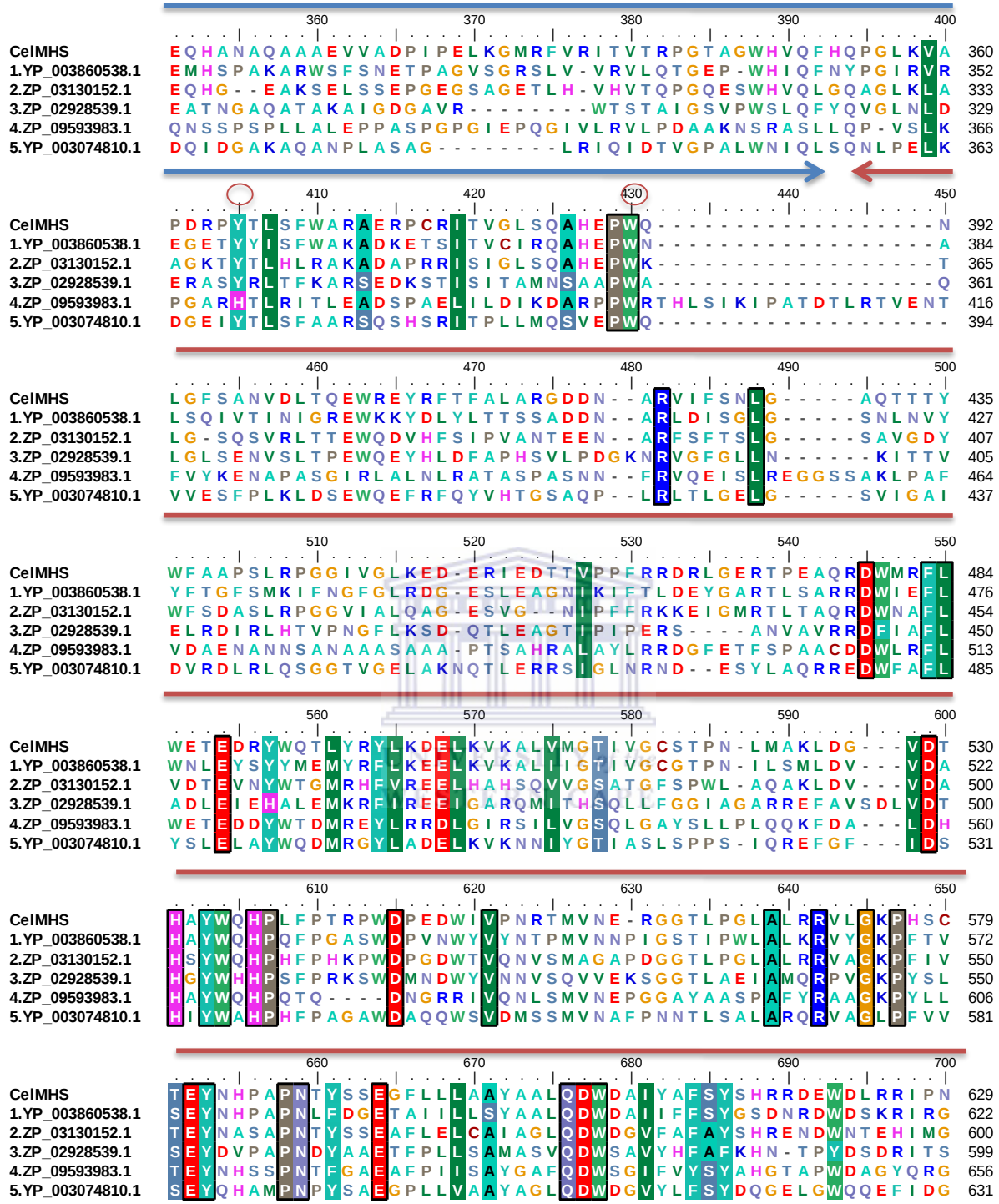


Figure 3.4 Unrooted phylogenetic tree of CelMHS (highlighted in bold) and closely related protein sequences based upon the sequence alignment of the full-length amino acid sequences. The tree was constructed using the neighbour joining method. Bootstrap values are based upon 1,000 resampled data sets and only values of greater than 40% are indicated. A *Bacillus cellulosilyticus* DSM 2522 GH family 5 cellulase (YP_004093450) was included as an outlier. Accession numbers are shown in brackets. The scale bar represents 20 substitutions per amino acid position, conducted in MEGA version 5.0.


CelMHS
1.YP_003860538.1
2.ZP_03130152.1
3.ZP_02928539.1
4.ZP_09593983.1
5.YP_003074810.1


CelMHS
1.YP_003860538.1
2.ZP_03130152.1
3.ZP_02928539.1
4.ZP_09593983.1
5.YP_003074810.1

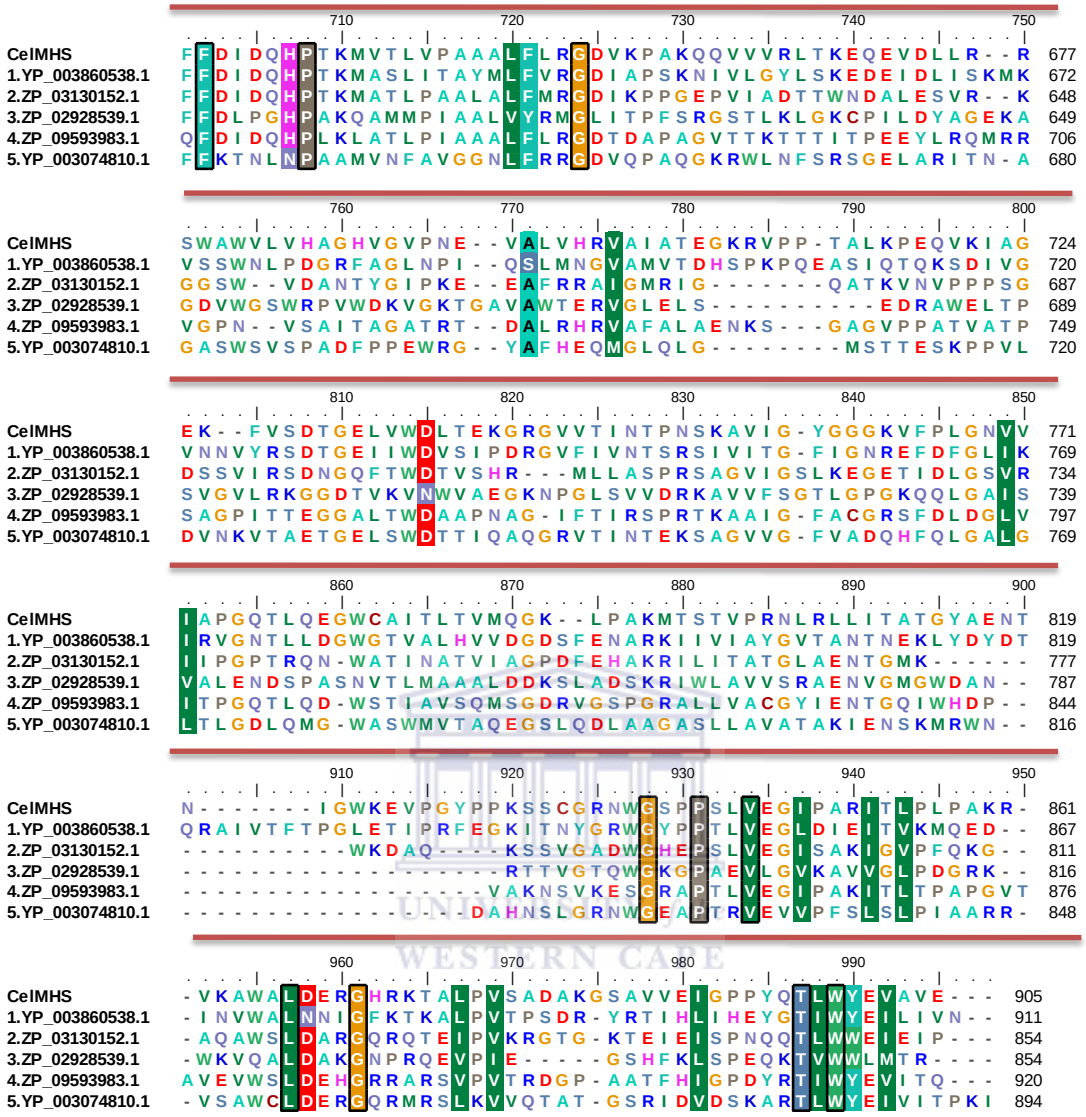


Figure 3.5 Multiple sequence alignment of the deduced CelMHS amino acid sequence with closely related protein sequences. Accession numbers (where available) are indicated. Boxed residues indicate sequence similarity with a threshold value of 100%. The colours of the respective boxes indicate amino acid and side chain properties, namely hydrophobic residues (green), positively (blue) and negatively (red) charged residues, polar uncharged residues (purple) and special case residues (grey). The black arrow designates the partial GH family 42 domain, while the blue arrow indicates CelMHS_CBM, and the pink arrow shows the region of CelMHS without any homology to known proteins. Catalytic residues are indicated by blue arrows and substrate binding residues are indicated by red circles.

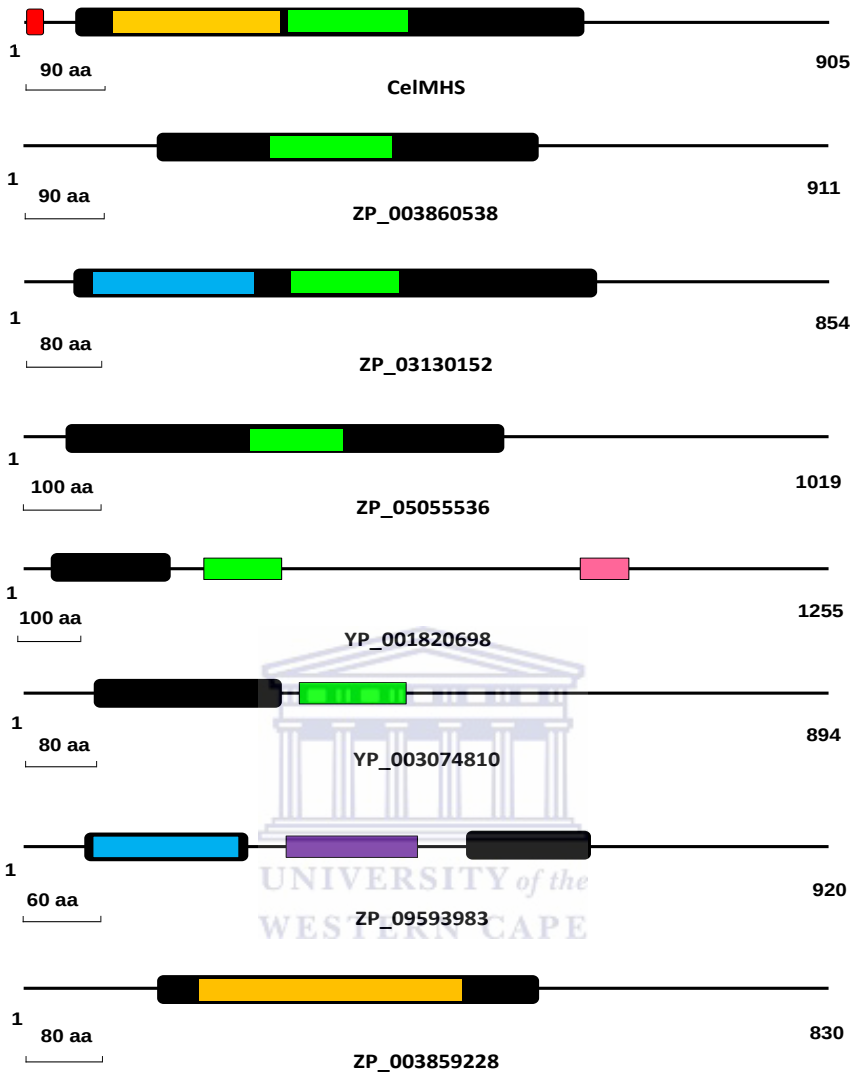


Figure 3.6 Domain architectures of CelMHS and related proteins, based upon InterProScan predicted domains. A red block indicates a transmembrane helix, a black block indicates a GH superfamily domain, a blue block designates a GH family 5 catalytic domain, an orange block indicates a GH 42 N-terminal domain, green blocks indicate CBM family 4 domains, a pink wedge indicates an immunoglobulin-like domain, and a purple block designates a galactose-binding domain. This figure has been drawn to scale.

The presence of related motifs and significant profile–profile comparison scores are our primary indicators that uncharacterised proteins, such as CelMHS, may have similar functions to related GH family members that have previously been characterised (Henrissat, 1991; Marchler-Bauer *et al.,* 2010). As previously mentioned, a conserved GHF42 (EC 3.2.1.23) (E-value =1.58e-0.4) domain and a CBM (E-value = 4.45e-10) were identified within CelMHS (Fig. 3.5).The GH family 42 enzymes characterized to date have been shown to exhibit β-galactosidase activity, which is also prevalent among GH families 1, 2 and 35 (Di Lauro *et al*., 2008). The majority of β-galactosidases that belong to GH family 42 are either thermophilic, psycrophilic or halophilic (Lee *et al*., 2011, Hidaka *et al*., 2002, Sheridan and Brenchley, 2000). GH family 42 β-galactosidases, such as BgaH characterized from the Archaeal isolate *Haloferax lucentense* SB1 (Genbank accession no.: AAB40123), are generally comprised of three separate domains, namely an N-terminal domain, a central trimerisation domain and a small C-terminal domain (Holmes and Dyall-Smith, 2002). Although the GH family 42 domain identified within CelMHS has homology to the N-terminal domain of β-galactosidases, these N-terminal domains are generally 370-390 amino acids in length, while the GH family 42 domain identified within CelMHS comprises of only 183 amino acids (Fig. 3.4 and 3.5). This suggests that it is a not an entire GH family 42 N-terminal domain, but rather a truncated portion of a N-terminal GH family 42 domain (Fig. 3.7).

Figure 3.7 Conserved domain architecture within the deduced CelMHS protein sequence. Deduced amino acids E^{98} to V^{286} have homology to GH family 42 (indicated by the blue block) and amino acids E^{294} to F^{425} (indicated by the red block) have homology to CBM family 4 domains within the CBM superfamily. Figure drawn to scale (www.ncbi.nlm.noh.gov).

Based upon a multiple sequence alignment of the CelMHS GH42 domain and GH42 N-terminal domains from previously annotated β-galactosidases (Fig. 3.8) identified from *Deinococcus geothermalis* (DSM 11300; ABF44258), *Thermotoga maritima* (MSB8; AE001776_6), *Haloferax lucentense* (SB1; AAB40123), *Bacillus licheniformis* (DSM12; ATCC 14580; AAU43090), *Alicyclobacillus acidocaldarius* (DSM446; ACV59895) and

an uncultured bacterium (BgaP412; AFD29748) it is clear that the GH42 domain of CelMHS lacks a significant number of amino acid residues that are highly conserved between β-galactosidases. Furthermore, CelMHS does not contain the conserved glutamate residue, indicated by a black arrow, implicated as an acid/base catalyst in the activity of β-galactosidases (Fig. 3.6) (Shaikh *et al.,* 2007).

Phylogenetic analysis of CelMHS and closely related proteins indicated that CelMHS is most similar to a carbohydrate-binding CenC domain-containing protein from *Chthoniobacter flavus* (ZP_03130152) (Fig. 3.4). All of the putative cellulase proteins to which CelMHS showed high levels of homology contain GH superfamily domains (Fig. 3.6). However, despite being closely related to these proteins CelMHS is the only protein which contains a truncated GH family 42 and a family 4 CBM within the GH superfamily domain (Fig. 3.6). It has been reported that family 4 CBMs are distinct in their ability to bind amorphous and not crystalline cellulose (Brun *et al*., 2000). However, the carbohydrate binding module of CelMHS also displayed homology to the CenC gene of *Cellulomonas fimi* which has been characterised to have both endo- and exoglucanase activity (Tomme *et al.,* 1996). Phylogenetic analysis was conducted using the sequence of the CBM to determine its evolutionary relationship to other characterised carbohydrate binding modules (Fig. 3.9). The CBM of CelMHS grouped with other family 4 CBM sequences. This family is often described as having binding grooves or clefts that recognize one polysaccharide strand (Boraston *et al.,* 2004). Biochemical studies of family 4 CBMs have revealed increased affinity for long chain hexoses and negligible interaction with oligosaccharides with a degree of polymerization of three or less and therefore have been described as “chain binders” (Pell *et al*., 2003).

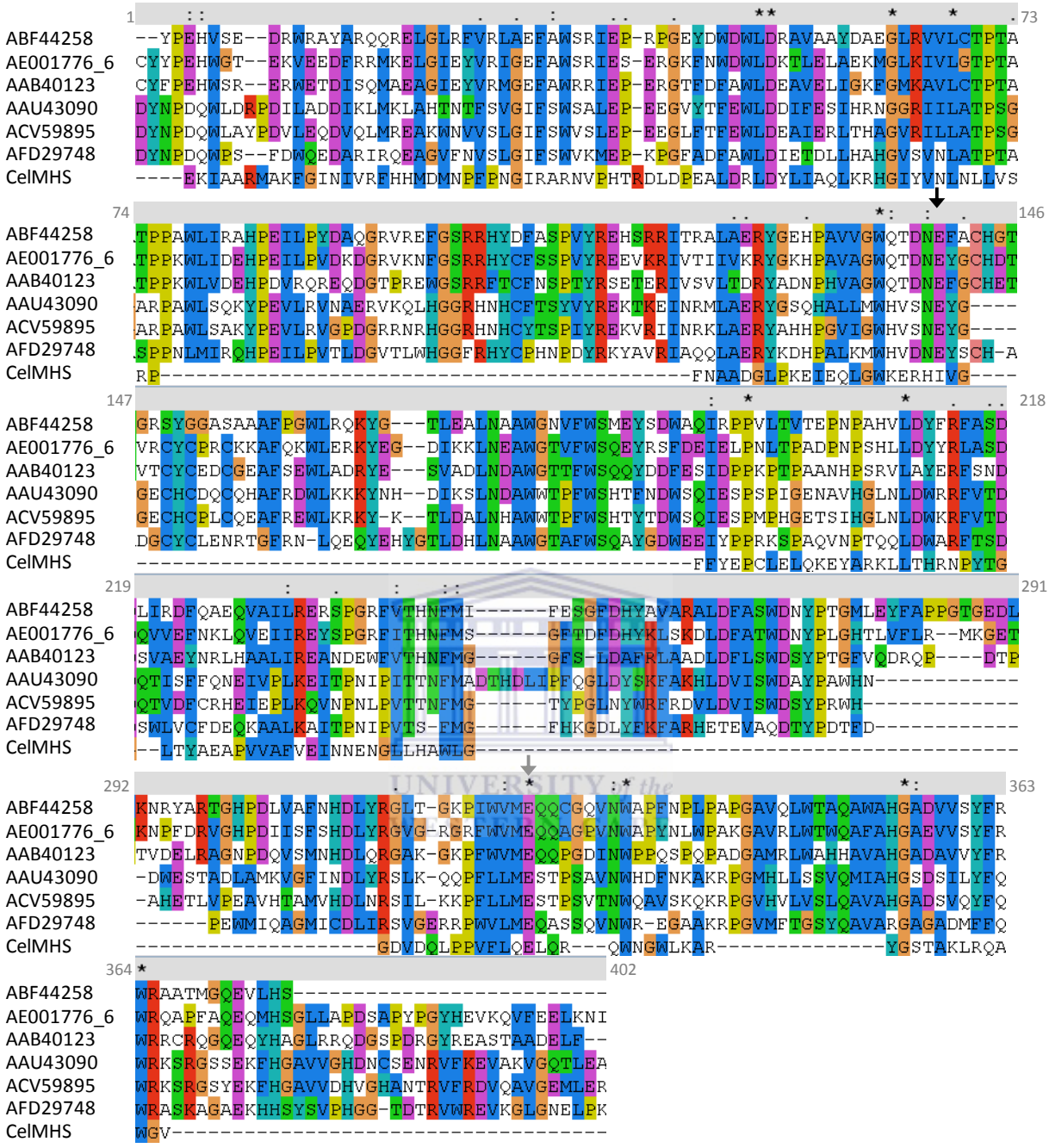


Figure 3.8 Multiple sequence alignment of GH family 42 N-terminal domains of six β-galactosidases and the partial GH family 42 N-terminal domain of CelMHS. The conserved acid/base catalyst (black arrow) and nucleophile (grey arrow) glutamate residues are indicated. Accession numbers are shown and the bacterial species from which the sequences originate are given in text.

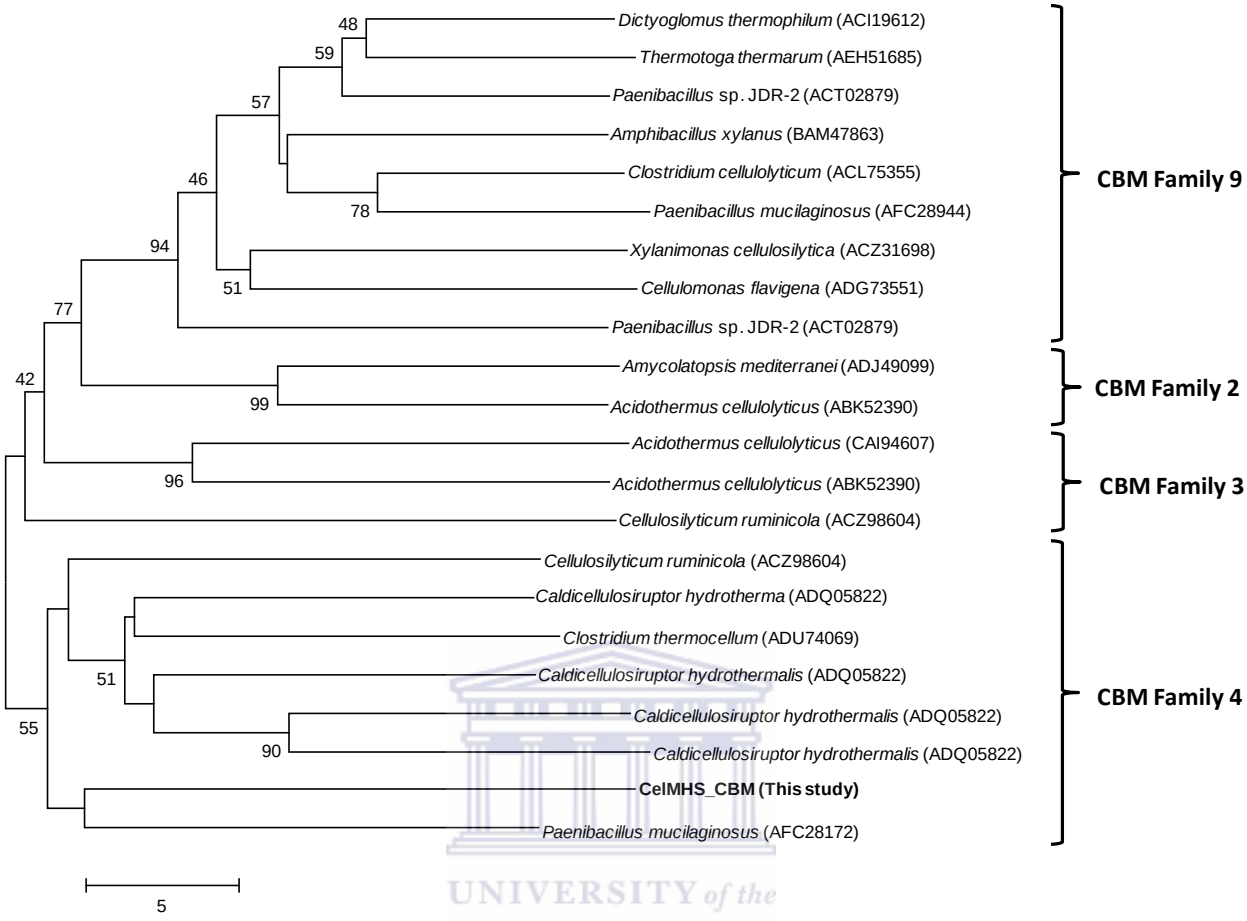


Figure 3.9 Unrooted phylogenetic tree of carbohydrate binding domain (CBM) families 2, 3, 4 and 9, including the CelMHS CBM (highlighted in bold) based on the sequence alignment of a common length portion. The tree was constructed using the neighbour joining method. Bootstrap values are based upon 1,000 resampled data sets and only values of greater than 40% are indicated. CBM families 2, 3 and 9 were included as out-groups. Accession numbers (where available) are shown in brackets. The scale bar represents 5 amino acid substitutions per amino acid position, conducted in MEGA version 5.0.

A three dimensional (3-D) model of the putative CBM of CelMHS was modeled against the β-1,3 glucan specific family 4 CBM from the marine hyperthermophile *Thermotoga maritima* (*Tm*CBM4-2) [Protein Data Bank (PDB) identifier 1guiA] (Fig. 3.10). Quality assessment of the model revealed a QMEAN4 score of 0.56 and a QMEAN Z score of -2.575 with only 27.2% sequence identity to *Tm*CBM4-2. The QMEAN4 score obtained

indicated that the model was unreliable. A low QMEAN Z score, such as this, is indicative of poor agreement with the predicted protein features as well as an increase of force potential energy. To further assess the quality of the predicted model a ramachandran plot analysis was generated (Supplementary Fig. 3). The total number of residues within the CBM of CelMHS is 147, the number of amino acids which fall within the favoured, allowed and outlier regions were 127 (86.4%), 17 (11.6%) and 3 (2.0%) respectively. Homology modeling is designed to infer an unknown sequence structure based on the solved similarly folded protein structure available (Zhang and Skolnick, 2008). The CBM of CelMHS was modeled to gain understanding of the function of this potentially novel protein at a molecular level. This information may be used to establish a basis for future experiments to reveal the biological function and interaction properties of sequences. Boraston *et al.,* 2001 reported that *Tm*CBM4-2, the template used to model the CBM of CelMHS, is specific for binding β-1,3 glucans, namely laminari-oligosaccharides and to a lesser extent cello-oligosaccarides. Therefore, it would be interesting to investigate whether CelMHS shares similar binding specificities. The accuracy of the model remains relatively low; however, this is not unexpected as the CBM domain showed a low sequence similarity to the template sequence and structure. The inconsistency between the models is based upon the fact that the CBM of CelMHS is at the limit of what can be modeled. In addition; the PDB database contains a limited number of resolved structures suitable for use as modeling templates. This model may, therefore, only serve as a general guide as to the proposed 3-D structure of the CBM of CelMHS, and the crystal structure of the entire CelMHS cellulase will need to be determined experimentally to obtain accurate protein structure information.

*Tm*CBM4-2, the CBM to which that of CelMHS was modeled, consists of residues S^{488} to E^{642} of the *T. maritima* laminarinase protein and displays a β-jelly roll formation, whereas the CBM of CelMHS comprises residues L^{296} to V^{444} and is predicted to fold into a single domain of twelve consecutive β-strands that in turn form a β-sandwich (Boraston *et al.,* 2004) (Fig. 3.10). Many of the characterised CBMs have structures based on different organisations of β-sheets (Linder and Teeri, 1997; Bayer *et al*., 1998). As reported by Fiser *et al*. (2000) functional differences between members of the

same protein family are usually a consequence of structural differences on the protein surface. In a given protein family these differences are generally produced by insertions and deletions of residues between members of the family. Boraston *et al.* (2004) suggests that CBMs have a key function in substrate recognition and binding and in the activity of glycolytic lignocellulosic enzymes. Thus, a greater understanding of the specific role of CBMs and the interaction between CBMs and their respective catalytic domains, including how CBMs bind to their target substrate, can provide critical information which can be used within biotechnological applications to enhance carbohydrate-recognition and cleavage.

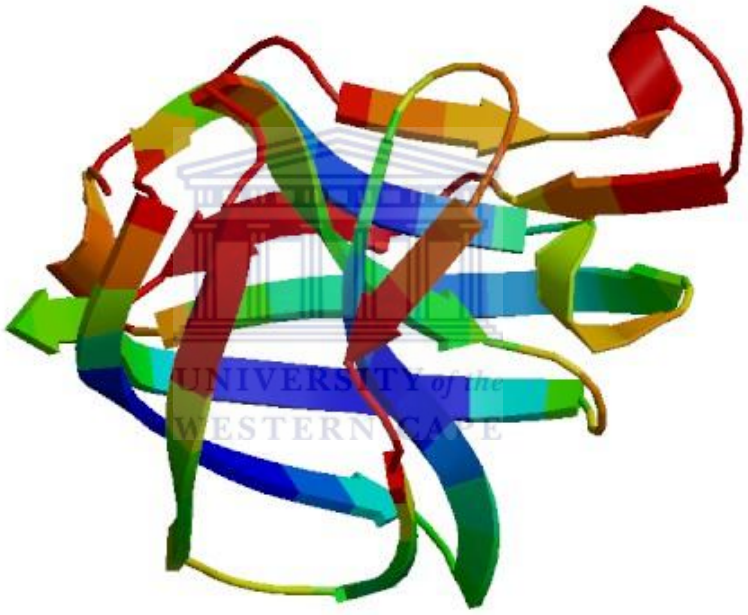


Figure 3.10 Cartoon representation of the homology model of CelMHS CBM. The model was created using a β-1,3-glucan specific family 4 CBM from the marine hyperthermophile *Thermotoga maritima* [Protein Data Bank (PDB) identifier 1guiA]. The estimated residue error is visualised using a colour gradient in which blue regions are more reliable and red regions potentially unreliable.

CBM family 4 domains are present in proteins which bind to an array of different polysaccharides. Characterised members of this family have been documented to bind β-1,3-, β-1,4-, β-1,6-linked xylan and mixed β-1,3- and β-1,4-linked β-glucan substrates (Tomme, 1996; Hachem *et al*., 2000; Boraston *et al*., 2001).

The top five proteins to which CelMHS showed sequence identity were obtained from genome annotations of organisms originating from a broad range of environments. The isolate, *I. aggregans*, is a hyperthermophilic archaean isolated from a hot spring in New Zealand (Niederberger *et al*., 2006), *C. flavus* is characterised as a mesophilic soil bacterium (Sangwan *et al*., 2004), *V. bacterium* DG1235 (Hart, 2007) and *T. turnerae* T7901 (Yang *et al*., 2009) were isolated from marine environments, whereas *O. terrae* TB90-1 was isolated from anoxic rice paddy soil (Chin *et al*., 1999). CelMHS was identified from a metagenomic library constructed from DNA isolated from sediment obtained at a Malawian hot spring, which is a similar environment to that from which *I. aggregans* was isolated. Recently, a metagenomic approach was employed to identify a multi-domain hyperthermophilic cellulase from a geothermal consortium of which *I. aggregans* represented the dominant organism, and is reported to be the most thermostable cellulase identified to date (Graham *et al*., 2011). Therefore, given that CelMHS was identified from a similar thermophilic environmental source, and that one of the closest related sequences is a cellulase from the hyperthermophile, *I. aggregans*, it is plausible that CelMHS could potentially display a high level of thermostability. This may make it highly feasible for inclusion in thermogenic industrial processes such as those proposed for bioethanol production from lignocellulosic feedstocks.

3.4 Cloning celMHS

The *celMHS* coding region, of approximately 2,717 bp, was successfully amplified by PCR (Fig. 3.11) and cloned into the pET28a(+) vector for recombinant protein expression in *E. coli*. Sequencing of the resulting recombinant construct, pET28_*celMHS*, confirmed that the expected fragment comprising the *celMHS* gene had been cloned without any PCR errors and was in-frame with the vector encoded histidine tag. Subsequently, *E. coli* BL21 and Rosetta 2 strains harbouring pET28_*celMHS* were functionally screened using CMC as the substrate and Congo red stained to confirm recombinant cellulase activity (data not shown). Protein expression from these recombinant clones was compared to expression of a truncated *celMHS* gene, *celMHSΔTMR* lacking the nucleotides encoding the predicted transmembrane

region of CelMHS (W^{5}-G^{25}). This truncated gene was cloned into the pET28a(+) vector to generate the pET28_*celMHSΔTMR* construct by I. Ackerman (PhD student, IMBM, UWC).

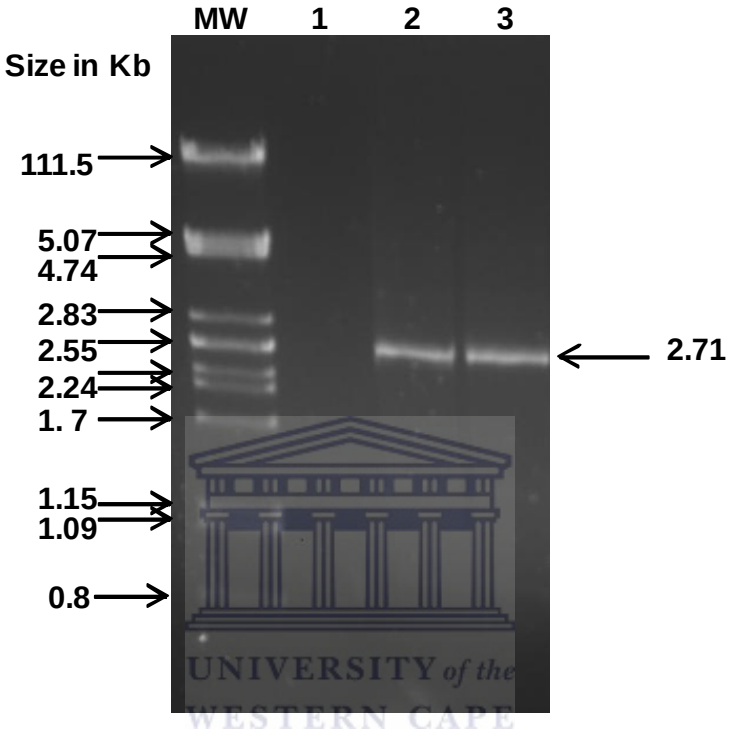


Figure 3.11 PCR amplification of *celMHS* using gene specific primers. Lane MW, λ*Pst*I DNA molecular marker; Lane 1, negative control and; Lane 2 and 3, *celMHS* PCR amplified. The approximate sizes of the molecular weight marker and the size of the amplified *celMHS* fragment are indicated.

3.5 Over expression and purification of CelMHS

One of the main factors that affect gene expression is differences in codon usage between the native organism from which the protein was derived, and the heterologous expression host (Heitzer, 2007). Furthermore, the heterologously expressed gene may contain regulatory elements in its coding sequence which limits expression (Heitzer, 2007). CelMHS was predicted to have a 10.2% frequency of rare codon occurrence which may affect its downstream expression in *E. coli*. The likelihood of high protein expression is determined by the Codon Adaption Index (CAI) value of a protein. A CAI

value of 1.0 is considered to be ideal; however, CelMHS was determined to have a CAI of 0.68, thereby indicating that there is a chance that the expression of *celMHS* may be negatively influenced by rare codon content. Therefore, a comparative expression trial was conducted in order to determine the optimal *E. coli* expression host for *celMHS* expression. The crude, soluble cellular extracts of both *E. coli* BL21 (DE3) and *E. coli* Rosetta 2 cells harbouring pET28a(+), pET28_*celMHS* or pET28_***celMHSΔTMR*** were analyzed to investigate the levels of CelMHS obtained from the two *E. coli* host strains (Fig. 3.12A).

An increase in CelMHS band intensity, as observed visually for the protein species at approximately 105 kDa on Coomassie stained SDS-PAGE gels, is indicative of enhanced protein expression levels. Furthermore, substrate SDS-PAGE analysis revealed that recombinant CelMHS, corresponding to approximately 98 kDa, had significant endoglucanase activity against CMC (Fig. 3.12B). Cellulase activity was observed for *E. coli* BL21 (DE3) and *E. coli* Rosetta 2 harbouring pET28_***celMHSΔTMR***, respectively, and not for the host strains harbouring pET28a(+). This result confirmed that the cellulase activity was a result of the cloned insert encoding *celMHS*. Higher cellulase activity was also observed from the pET28_***celMHSΔTMR*** construct compared to pET28_*celMHS* when the soluble protein was assayed from a crude cell free fraction (Fig. 3.13).

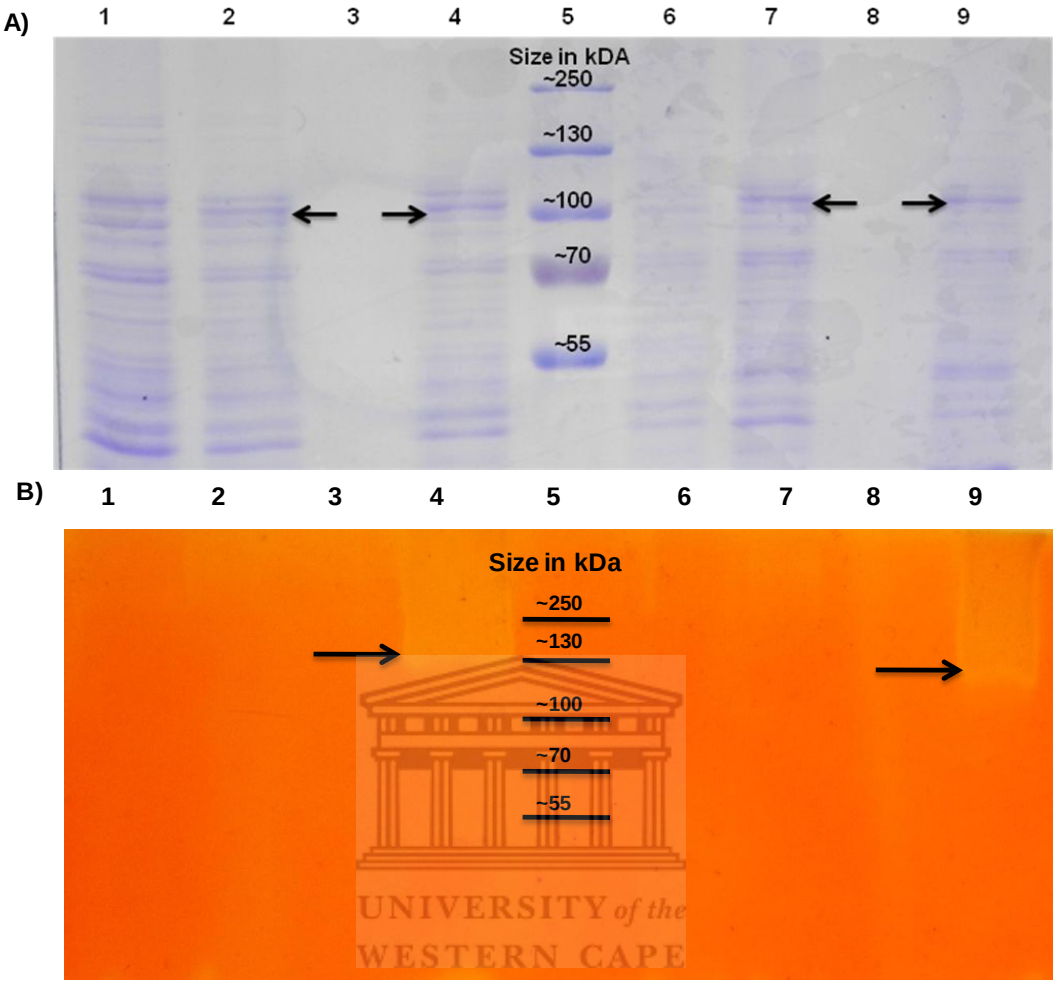


Figure 3.12 Comparative analysis of recombinant protein expression of CelMHS within *E. coli* BL21 (DE3) pLysS and *E. coli* Rosetta 2 expression hosts. Coomassie stained SDS PAGE analysis (A) and a zymogram indicating in-gel hydrolysis of CMC (B). Lane 1, *E. coli* BL21 (DE3) pET28a soluble fraction; Lane 2, *E. coli* BL21 (DE3) pET28_*celMHS* soluble fraction; Lane 4, *E. coli* BL21 (DE3) pET28_*celMHSΔTMR* soluble fraction; Lane 5, Protein molecular weight marker (Fermentas); Lane 6, *E. coli* Rosetta 2 pET28a(+) soluble fraction; Lane 7, *E. coli* Rosetta 2 pET28_*celMHS* soluble fraction; and Lane 9, *E. coli* Rosetta 2 pET28_*celMHSΔTMR* soluble fraction. The approximate sizes of the molecular weight marker are indicated and the approximate position of CelMHS is indicated by the black arrows.

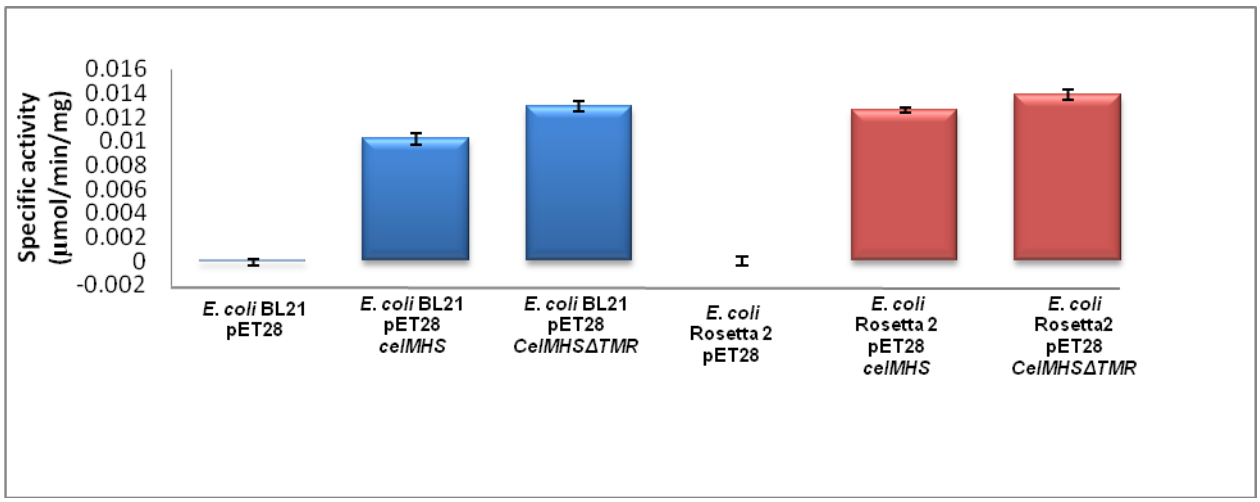


Figure 3.13 Specific activity of CelMHS and CelMHSΔTMR within either *E. coli* BL21 (DE3) or *E. coli* Rosetta 2 as determined using the DNS reducing sugar assay with 0.3% CMC as the substrate for 10 min at 80°C (100 mM MES buffer pH 6.0). Data represents the mean (n=3) ± standard error.

Cellulase activity assays revealed that both *E. coli* BL21 (DE3) and Rosetta 2 harbouring pET28_*celMHS* and pET28_*celMHSΔTMR*, respectively, produced a substantial quantity of the recombinant protein (Fig. 3.13). Furthermore, higher levels of the recombinant protein were produced by *E. coli* Rosetta 2 harbouring the pET28_*celMHSΔTMR* construct. *E. coli* Rosetta 2 enhances protein expression from target genes containing a significant proportion of rare codons (Brinkmann *et al*., 1989; Seidel *et al*., 1992; Rosenberg *et al.,* 1993; Del Tito *et al*., 1995), which are thought to hamper translation in heterologous host systems (Wald *et al.,* 2012). When the mRNA encoded by the heterologous gene is transcribed in *E. coli*, differences in codon preferences between the gene and the host may interfere with the translation process, as one or more tRNAs required for the rare codons may be rare or lacking in the mRNA population of the host bacterial strain. The presence of numerous rare codons within the mRNA of a cloned gene may lead to a reduction in either the quantity or quality of the translated protein species. The use of the *E. coli* Rosetta 2 host strain therefore circumvents the host's synthesis of codon optimized genes as it contains the pRARE plasmid which encodes tRNA genes for the infrequently used amino acid codons

(Brinkmann *et al*., 1989; Seidel *et al*., 1992; Rosenberg *et al.,* 1993; Del Tito *et al*., 1995). A comparative analysis of *celMHS* constructs expressed in *E. coli* Rosetta 2 and *E. coli* BL21 (DE3) strains suggest that the use of Rosetta 2 may result in enhanced expression of CelMHS (Fig. 3.12 and 3.13), which correlates with the presence of numerous rare codons within *celMHS*. Therefore, *E. coli* Rosetta 2 was used for the rest of this investigation for *celMHS* expression and CelMHS purification. In addition, *E. coli* Rosetta 2 harbouring pET28_*celMHSΔTMR* displayed higher cellulase activity than the other *celMHS* constructs assessed during the course of this investigation (Fig. 3.13), and therefore, pET28_*celMHSΔTMR* was used for further expression and purification of CelMHS during the course of this study.

Optimization of the over-expression of *celMHS* revealed that Rosetta 2 harbouring pET28_*celMHSΔTMR* was optimally induced for the overexpression of CelMHS with 1 mM IPTG at 37°C for 4 hours (Fig. 3.14). These conditions were used to express CelMHS for purification by His affinity chromotography. The eluted fraction was analyzed on a SDS-PAGE, which showed a single protein band of approximately 100 kDa (Fig. 3.15). The molecular weight of the resulting protein is consistent with the theoretical value of approximately 98 kDa calculated from the sequence of the recombinant protein. However, the apparent molecular mass of the protein should be further investigated using protein molecular weight standards and size exclusion chromatography.

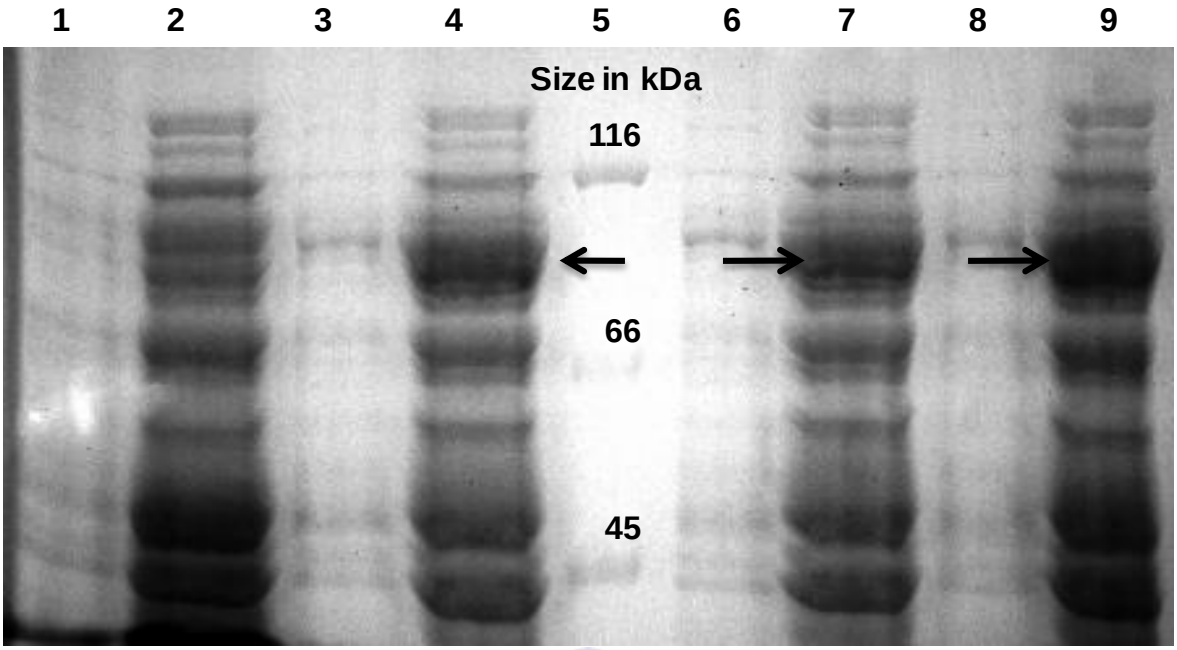


Figure 3.14 SDS-PAGE analysis of *E. coli* Rosetta 2 cell fractions induced with varying IPTG concentrations. pET28a(+) soluble (Lane 1) and insoluble fractions (Lane 2) induced with 0.5 mM IPTG; pET28_*celMHSΔTMR* soluble (Lane 3) and insoluble fractions (Lane 4) induced with 0.8 mM IPTG; unstained protein molecular weight marker (Lane 5); pET28_*celMHSΔTMR* soluble (Lane 6) and insoluble (Lane 7) fractions induced with 1 mM IPTG and; pET28_*celMHSΔTMR* soluble (Lane 8) and insoluble (Lane 9) fractions induced with 0.5 mM IPTG. The approximate sizes of the molecular weight marker and CelMHS are indicated by the numbers and arrows, respectively.

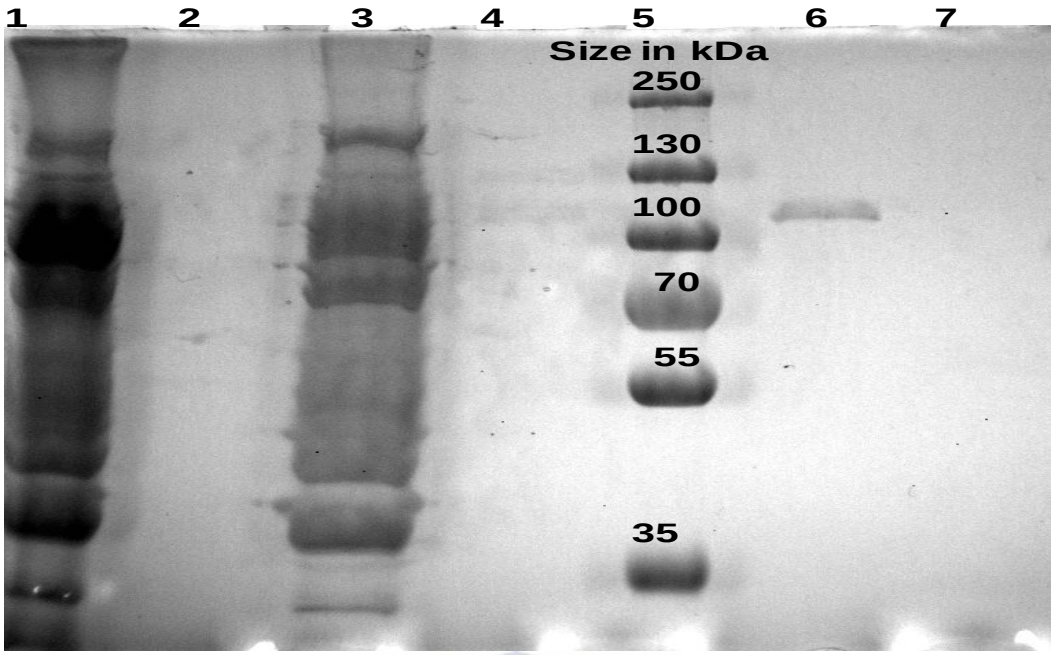


Figure 3.15 SDS-PAGE analysis of nickel-affinity column purification of recombinant CelMHS from *E. coli* Rosetta 2 harbouring pET28_*celMHSΔTMR*, following over expression of CelMHS with 0.5 mM IPTG. Lane 1, Total soluble cell-free extract loaded onto the column; Lane 2, Flow-through; Lane 3, Binding buffer wash; Lane 4, Wash buffer; Lane 5, Molecular weight marker (PageRuler protein ladder, Fermentas); Lane 6, Eluate; and Lane 7, Strip buffer. The approximate sizes of the molecular weight marker and the approximate position of CelMHS are indicated.

3.6 Functional characterisation of CelMHS

The functional characteristics of CelMHS, including optimum pH, temperature, thermostability, substrate specificity and steady state enzyme kinetics, were evaluated in order to better understand this novel cellulase.

3.6.1 Effect of pH and temperature on CelMHS activity

CelMHS was shown to be active over a broad pH range, with the highest apparent activity at pH 6.0 in 100 mM MES buffer (Fig. 3.16). This suggests that CelMHS is most active in a neutral to slightly acidic pH range, which is consistent with the physiological

conditions of the hot spring from which the metagenomic DNA encoding CelMHS was originally isolated. A number of cellulases have been identified with biochemical characteristics that match the environment from which they are isolated (Healy *et al*, 1995, Voget *et al*., 2006, Geng *et al*., 2011). Healy *et al*. (1995) demonstrated that out of 12 clones exhibiting cellulase activity the 4 chosen for further characterisation were shown to function optimally at pH 6-7 which was a similar pH as the environmental sample from which they were obtained.

Purified CelMHS displayed an optimal reaction temperature close to 100°C (Fig. 3.17). This is similar to the optimal temperature of EBI–244, a cellulase identified following archaeal enrichment by Graham *et al.* (2011). EBI-244 was observed to have maximum activity at 109°C and negligible activity at 70°C when assayed against CMC at pH 6.8 (Graham *et al.,* 2011). As expected, CelMHS had a higher temperature optimum than cellulases identified from metagenomic libraries constructed from mesophilic environments such as buffalo rumen (Duan *et al*., 2009; Lui *et al.*, 2011) and soil (Voget *et al*., 2005; Kim *et al*., 2007; Jiang *et al*., 2009; Pang *et al*., 2009; Lui *et al*., 2011; Nacke *et al*., 2011). However, CelMHS also had a higher optimum reaction temperature compared to the temperature optima of two other proteins, XPgene12 (Hu, MSc thesis, 2010) and RHgene34 (Ngobeni, MSc thesis, 2011), a cellulase and a ferulic acid esterase identified and characterised from the same Malawian hot spring soil metagenomic library. These proteins displayed temperature optima of 50°C and 45°C respectively.

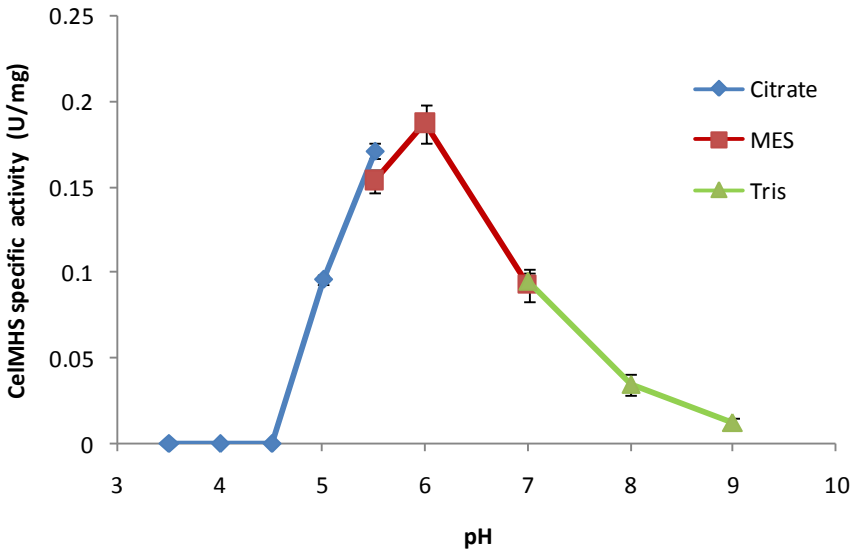


Figure 3.16 Effect of pH on purified CelMHS activity using 0.3% CMC as substrate. The activity was determined in various pH buffers at 60°C. The buffers used were citrate buffer (pH 3.5-5.5; blue line), MES buffer (pH 5.5-7; red line) and Tris buffer (pH 7–9; green line). Data represents the mean (n=3) ± standard error indicated by error bars.

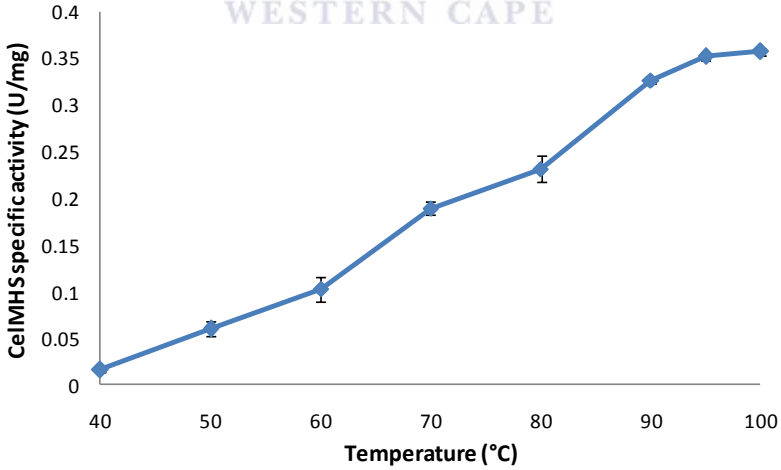


Figure 3.17 Effect of temperature on CelMHS activity using 0.3% CMC as substrate for 10 min at the indicated temperatures (pH 6.0 in 100 mM MES buffer). The amount of reducing sugars liberated was measured by DNS assay. Data represents the mean (n=3) ± standard error indicated by error bars.

3.6.2 Thermostability

CelMHS was found to be highly thermostable, retaining approximately 100% residual activity following 60 minute incubations at 60 and 70°C, respectively (Fig. 3.18). Furthermore, the enzyme retained 95 and 65% of its activity at 80 and 90°C, respectively, after 60 minutes. CelMHS was not inactivated by boiling for 1 hr; however, autoclaving resulted in the irreversible deactivation of the enzyme (data not shown). Considering that the metagenomic library from which CelMHS was isolated was constructed from a Malawian hot spring soil sample with a temperature range of 72-78°C, we had hypothesised that the enzyme may be thermally stable. To the best of our knowledge this enzyme has the highest reported thermostability when compared to other cellulases originating from metagenomic enzyme discovery studies (Voget *et al*., 2005; Feng *et al*., 2007, Kim *et al*., 2007; Jiang *et al*., 2009; Pang *et al.,* 2009; Lui *et al*., 2011; Nacke *et al*., 2011). None of these studies were, however, conducted on metagenomic DNA extracted from hot spring environments and perusal of the currently available literature did not yield any examples of studies where cellulases have been identified from hot spring environments using metagenomic approaches. The degree of thermostability of CelMHS suggests that this enzyme may be suitable for use in industries requiring a highly thermostable biocatalyst such as biorefineries.

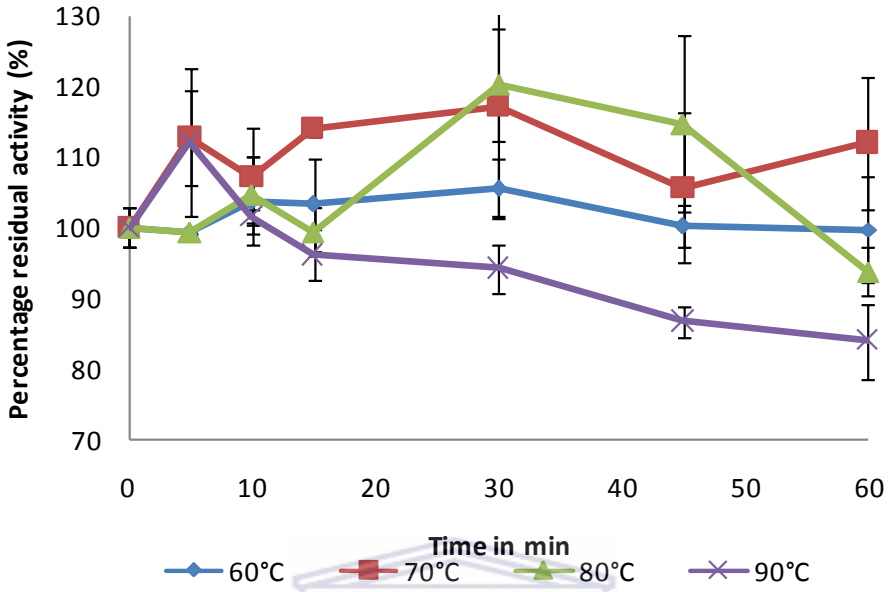


Figure 3.18 Thermostability profile of purified CelMHS at 60, 70, 80 or 90°C, as indicated, for 60 min. Residual activity was determined as a percentage of an untreated control sample incubated on ice for the duration of the experimental period. Data represents the mean (n=3) ± standard error.

3.6.2 CelMHS substrate specificity

Hydrolysis of various substrates by CelMHS was performed to investigate the substrate range of CelMHS (Table 3.5). CelMHS was initially identified following a functional screen of a metagenomic library using CMC as the substrate (see section 3.1). CMC is a specific substrate for endo-acting cellulases as the method of CMC preparation specifically decrystallizes cellulose to create amorphous sites that are ideal for endoglucanase action. Therefore, it is not surprising that the purified enzyme showed highest activity against substrates with a high number of β-1,4 linked glucose subunits, such as CMC, lichenan and β-glucan (Table 3.5). Higher CelMHS activity was detected on the linear polymeric substrates, lichenan and β-glucan, when compared to CMC a derivative of cellulose containing carboxymethyl groups. This is possibly due to CelMHS having greater accessibility to hydrolysable bonds in the more simplistic linear

substrates. The higher activity observed on β-glucan compared to lichenan may indicate that CelMHS preferentially cleaves β-1,4 linked glucose subunits, rather than β-1,3 linked subunits as the ratio of 1,4 to 1,3 linkages in β-glucan is greater than those contained within lichenan (http://secure.megazyme.com/files/BOOKLET/P-LICHN_70901_DATA.pdf). This is also supported by the fact that CelMHS had little activity on the other linear glucose polymeric substrate tested, laminarin (Table 3.5), which comprises of β-1,3 and β-1,6 linked glucose subunits. High levels of activity on β-glucan is reported to be a property of endoglucanases (Bailey *et al*., 1993) and may suggest that CelMHS acts as an endoglucanase. Interestingly, CelMHS displayed no activity against short-chain *p*-nitrophenol (*p*NP) linked substrates such as *p*NP-β-galactoside and *p*NP-β-D-glucopyranoside, however, a slight yellow colour change indicating the release of *p*NP was observed after a 30 min incubation with *p*NP-β-D-cellobioside when compared to a control reaction containing no enzyme (data not shown). CelMHS showed little activity against polysaccharides comprising of sugar residues other than glucose such as arabinan, laminarin and rye arabinoxylan. CelMHS also displayed limited activity against β-1,4-linked xylose substrates, such as birchwood and beechwood xylan (Table 3.5). Interestingly, CelMHS was shown to have activity towards a crystalline cellulose substrate, Avicel, which suggests exoglucanase activity as it has the ability to access bonds within a substrate with a high degree of crystallinity (Table 3.5).

Collectively these results suggest that CelMHS may have dual activity comprising both endo- and exoglucanase activities. A number of metagenomic studies have reported cellulases with dual endo- and exoglucanase activities (Duan *et al.,* 2008; Duan and Feng, 2009; Graham *et al.*, 2011), including the hyperthermophilic cellulase isolated from archaeal enrichment (Graham *et al.* 2011) and CenC *Cellulomonas fimi* (Tomme *et al*., 1996). Interestingly, results from the Conserved Domain Database (CDD) reveal that the closest hit to the CBM4_9 from CelMHS is the carbohydrate binding module (family 4_9) from CenC produced by *Cellulomonas fimi.*

Table 3.5 Relative activity of CelMHS against various polysaccharide substrates. The assay was performed for 20 min at 80°C, unless otherwise specified, in the presence of 0.3% (w/v) of each of the substrates. Data represents the average of three replicates ± standard error.

Substrate	Relative activity [a,b]
CMC	100.0 ± 3.1
β-Glucan	319.3 ± 0.7
Lichenan	291.4 ± 4.9
Avicel[b]	99.8 ± 4.2
Laminarin	6.5 ± 0.7
Beechwood Xylan	5.8 ± 0.9
Birchwood Xylan	2.0 ± 0.3
1,4 β-D Mannan	1.4 ± 0.3
Rye Arabinoxylan	0.71 ± 0.4
Arabinan	0.3 ± 0.2

[a]The activity of CelMHS against CMC was defined as 100%.
[b]The activity of CelMHS against the insoluble and crystalline substrate, Avicel, was measured at 60°C for 36 hours.

3.6.3 CelMHS steady state enzyme kinetics

The CelMHS steady state kinetic parameters, K_m and V_{max}, were determined from Michaelis-Menten plots of cellulase activity at 80°C using increasing CMC concentrations (Fig. 3.21). Due to the viscous nature of the CMC substrate, 20 mg_{CMC} ml^{-1} was the highest concentration which could be used to generate the Michaelis-Menten curve. The K_m, V_{max} and K_{cat} were determined to be 25.47 mg_{CMC} ml^{-1}, 1020 U mg^{-1} and 1728.53 s^{-1}, respectively. CelMHS displays a V_{max} that is higher than that of most reported endoglucanases isolated from various metagenomes. This includes BT-01 from the rumen of buffalo (294 U mg^{-1}) (Nguyen *et al*., 2012), *bgl*1C from a soil metagenome (4.75 U mg^{-1}) (Jiang *et al*., 2009) and *cel*5G from a soil metagenome (256.61 U mg^{-1}) (Lui *et al*., 2011). Interestingly, the V_{max} value of XPgene12 (1085 U mg^{-1}), which was also isolated from the same Malawian hot spring metagenomic library, was reported to be very similar (Hu, MSc thesis, 2010).

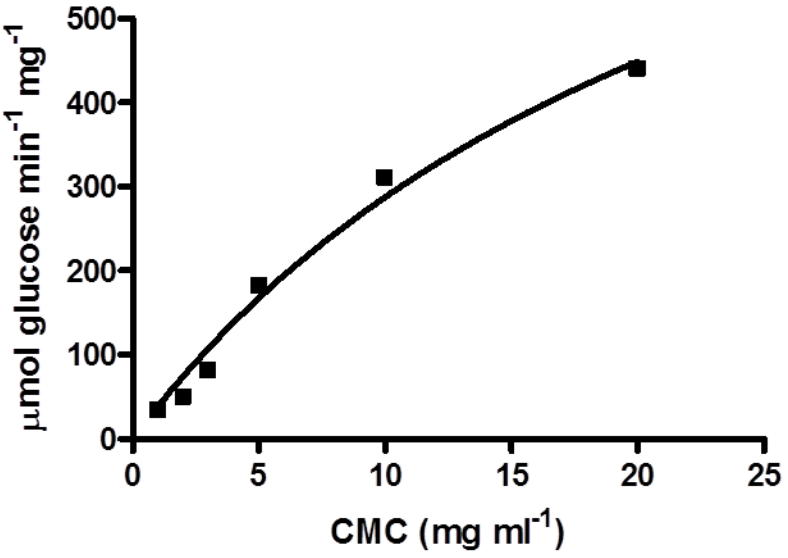


Figure 3.19 Michaelis-Menten plot of purified CelMHS obtained by measuring the rate of CMC hydrolysis at various substrate concentrations. K_m and V_{max} values were calculated using GraphPad Prism (Version 4, GraphPad Sofware, USA).

3.7 Hydrolytic capability of CelMHS

3.7.1 HPLC analysis of hydrolysis products liberated from CMC and B-glucan

Analysis of the hydrolysis products after CelMHS digestion of soluble cello-oligosaccharides with a degree of polymerization from 1 to 5 (Fig. 3.21) was performed to determine the minimum polymer length that could be digested by CelMHS. The products released following hydrolysis were analysed by HPLC and the observed refractive indexes (RI) were compared to those obtained from undigested cello-oligosaccharide standards (Fig. 3.20). CelMHS does not appear to have any significant hydrolytic activity on a cello-oligosaccharides consisting of two glucose units [cellobiose (G_2); Fig. 3.21A]. Only a small fraction of the three glucose subunit substrate [cellotriose (G_3)] was digested to release G_2 and glucose (G_1) (Fig. 3.21B), similarly a small amount of activity was observed on a similar substrate, p-nitrophenyl-cellobiose, after a 30 min incubation period (data not shown). This suggests that the minimum chain-length that CelMHS can act upon is a three sub-unit substrate. CelMHS was capable of completely hydrolyzing a four glucose subunit substrate [cellotetraose (G_4)] to yield C_3, C_2 and C_1 (Fig. 3.21C), indicating that a four sub-unit substrate is the preferred chain length. Hydrolysis of a five glucose subunit substrate [cellopentaose (G_5)] yielded both C_2 and C_3 subunits (Fig. 3.21D), which is surprising as C_3 was previously observed to be digested to C_2 and C_1 (Fig. 3.21B). The dominant product released by CelMHS was observed to be cellotriose, and to a lesser extent cellobiose and glucose.

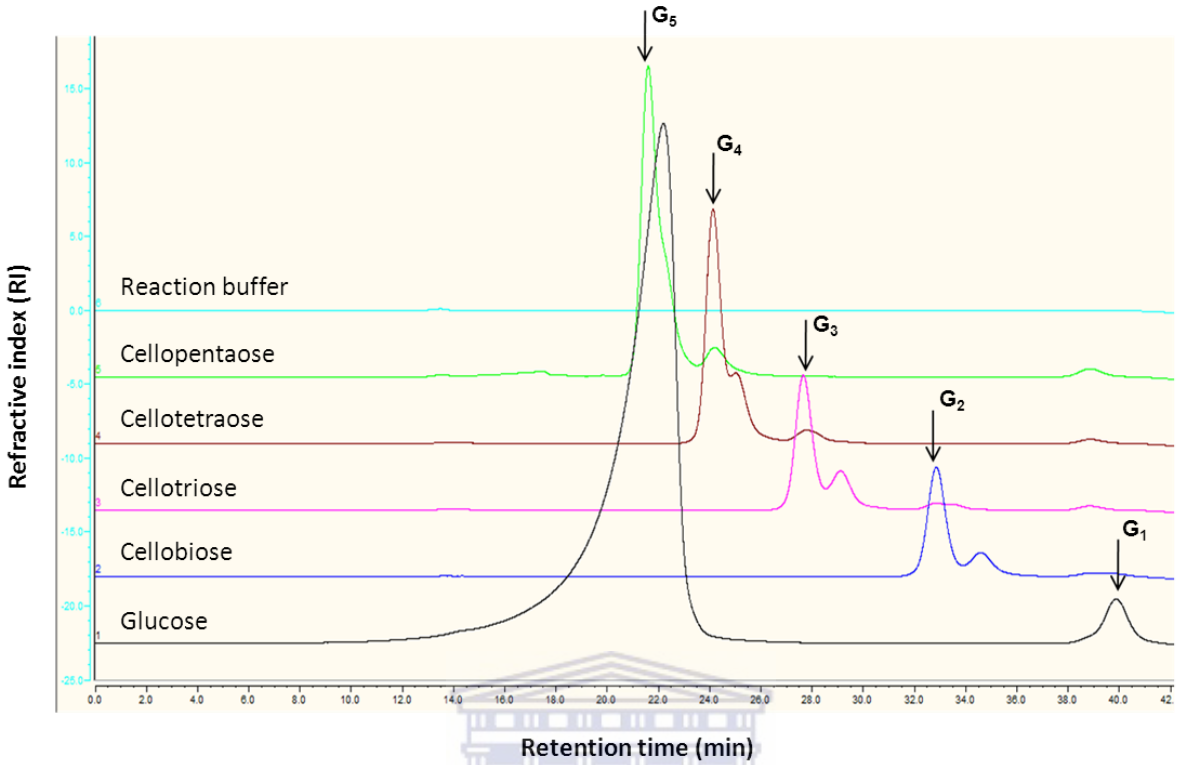


Figure 3.20 HPLC analysis of 100 mM standard oligo-saccharides, including glucose, cellobiose, cellotriose, cellotetraose and cellopentaose, and 50 mM MES (pH 6.0) reaction buffer separated by HPLC. The numbered arrows represent the oligo-saccharide peaks of: G_1, glucose; G_2, cellobiose; G_3, cellotriose; G_4, cellotetraose and; G_5, cellopentaose.

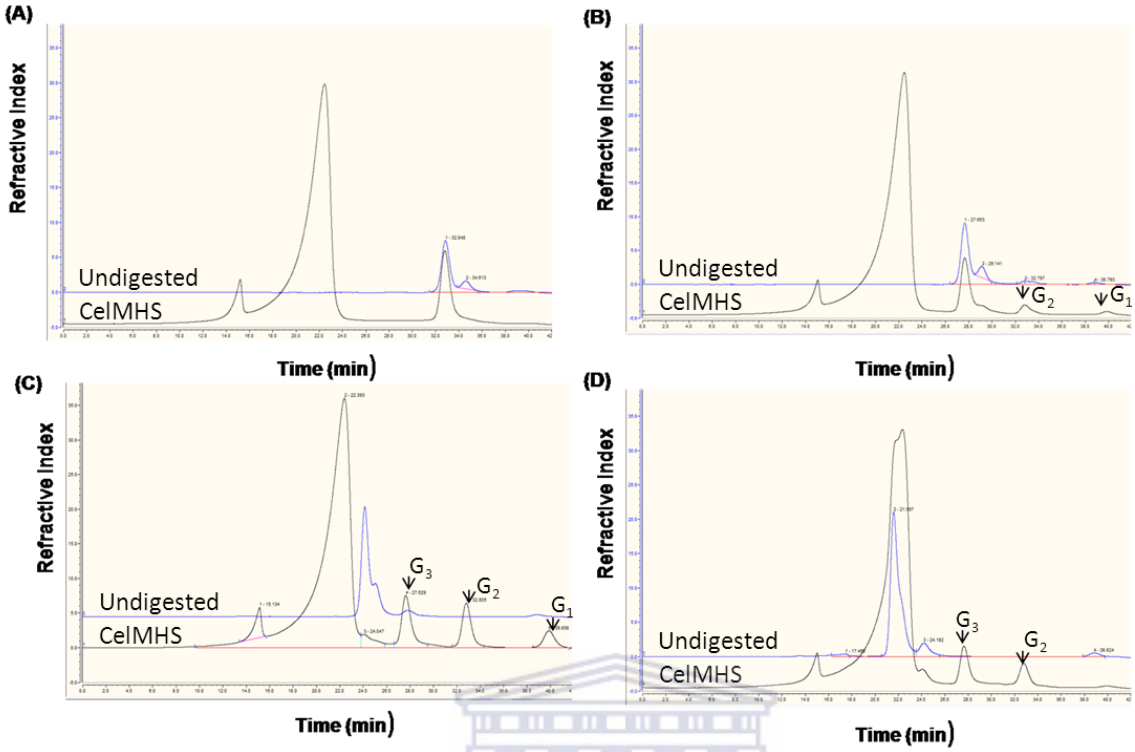


Figure 3.21 Hydrolysis of cellooligosaccarides at 60°C for 16 hours. A) Hydrolysis of cellobiose, B) Hydrolysis of cellotriose, C) Hydrolysis of cellotetraose, and D) Hydrolysis of cellopentaose. The undigested control and CelMHS digested experimental reactions are labeled, and the breakdown products are indicated with arrows and labeled: G_1, glucose; G_2, cellobiose and; G_3, cellotriose.

CMC and β-glucan were previously observed to be hydrolysed by CelMHS (section 3.6.2). Therefore, CelMHS hydrolysis of these substrates was further investigated in an attempt to better understand the effect of this cellulase on complex polymeric substrates (Fig. 3.21). CelMHS is able to act upon β-1,4-glucan, resulting in the broadening of the β-glucan substrate peak and formation of a number of smaller, but still high molecular weight products (Fig. 3.22A). Similarly, but to a far lesser extent, CelMHS does digest CMC, as the profile of the substrate peak is altered slightly, however, there is no evidence of a dominant product(s) being released. This correlates with the results reported for substrate specificity of CelMHS, in which we observed approximately 3 times more activity on β-1,4-glucan than on CMC. These results, and particularly those

of the β-glucan digestion suggest that CelMHS may have an endoglucanase mode of action. These results suggest that the catalytic domain of CelMHS is presumably influenced by the presence of the family 4 CBM as studies have confirmed that family 4 CBM employ a type B glucan chain binding mechanism which increases their affinity for longer chain oligosaccharides.

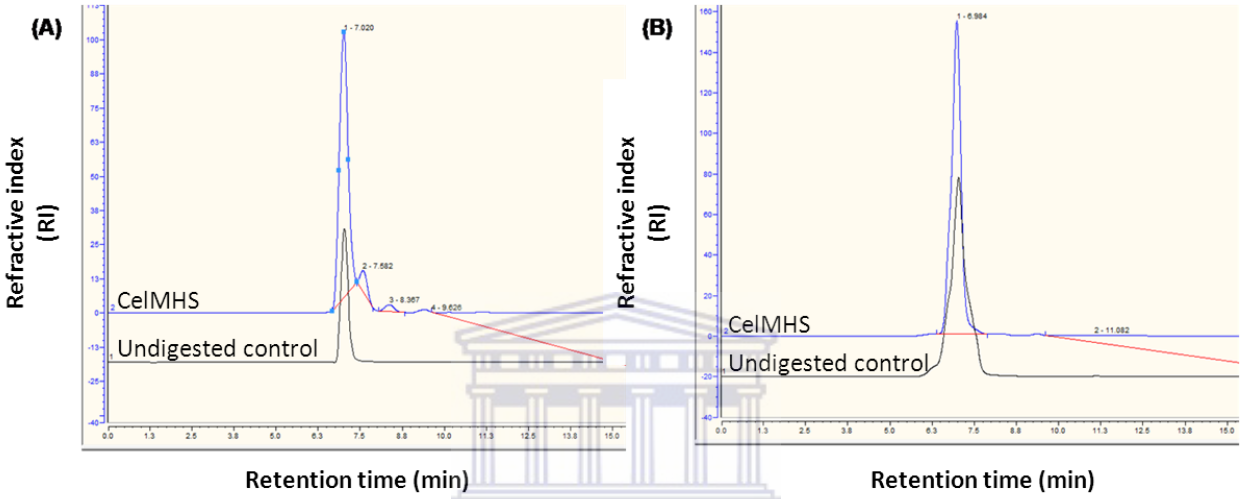


Figure 3.22 HPLC analyses of products released from β-glucan (A) and CMC (B) following hydrolysis with 5 µg of CelMHS for 16 hrs at 60°C. An undigested control that had also been incubated for 16 hrs at 60°C, of β-glucan (A) and CMC (B) was included for comparative analysis.

3.7.2 Hydrolysis of sugar cane bagasse with CelMHS

In vitro digestibility tests were performed to investigate the ability of CelMHS to digest sugar cane bagasse (SCB) (Fig. 3.23). Concurrently, the digestibility of the SCB using 15 U of a commercial *T. reesei* cellulase preparation was assessed. The activity of CelMHS under the conditions chosen for SCB hydrolysis was also confirmed using CMC, on which the enzyme is known to be active. Digestion of SCB with CelMHS released a negligible quantity of reducing sugar. In comparison, the commercial enzyme released five times more reducing sugar (Fig. 3.23). This difference may possibly be attributed to the presence of inhibitors of CelMHS activity within the pre-treated SCB, or

that CelMHS had low accessibility to glycosidic bonds within the cellulose portion of SCB. The latter explanation is likely to be the case, as a digestion performed with the inclusion of CelMHS's preferred substrate, CMC, with SCB released a substantial amount of reducing sugar (Fig. 3. 23). The sugar released is presumably due to the hydrolysis of CMC, the lower activity in the presence of SCB would suggest the presence of inhibitors. Collectively, these results suggest that CelMHS may not be able to access a significant proportion of the digestible cellulose within SCB. Therefore, CelMHS may need to be used in conjunction with a cocktail of other cellulolytic enzymes for the efficient hydrolysis of lignocellulosic biomass, such as SCB.

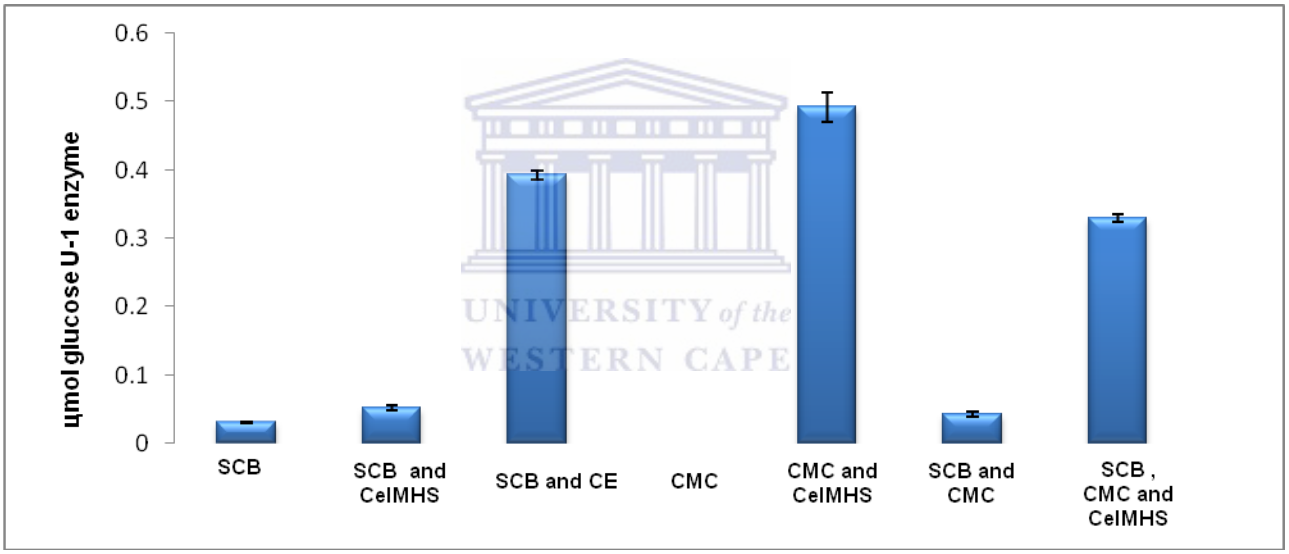


Figure 3.23 The reducing sugars liberated expressed as µmol glucose released per unit enzyme after 14 hrs at 60°C. Hydrolysis of sugarcane bagasse (SCB) or carboxymethyl cellulose (CMC) was performed in 100 mM MES buffer (pH 6). Data represent means ± SE (n=3). SCB was also hydrolysed using 15 U of a *T. reesei* cellulase available as a commercial enzyme (CE) under the same reaction conditions. The enzymes were mixed in various ratios of the total enzyme load.

Chapter 4

General Discussion and Conclusion

Liquid biofuels are currently regarded as the most feasible renewable resource to alleviate the world's dependence on fossil fuels (Demibas, 2008). Biofuels, such as bioethanol, can be produced from feedstocks comprising plant-derived agricultural waste products, otherwise referred to as lignocellulosic biomass (Margoet *et al.*, 2009). The production and use of biofuels, like bioethanol, have the potential to positively impact global economies as its use will reduce the consumption of crude oil and reduce greenhouse gas emissions (Margoet *et al.*, 2009). Lignocellulose, which is composed of the biopolymers cellulose, hemicellulose and lignin, is extremely recalcitrant to enzymatic degradation and its complete biodegradation requires the synergistic action of a suite of enzymes (Fuduka *et al.,* 2008; Hendriks and Zeeman, 2008; Kumar *et al.,* 2008). Furthermore, the complete hydrolysis of cellulose requires the concerted action of three types of cellulases: endoglucanase (EC 3.2.1.2), exoglucanase (EC 3.2.1.4) and β-glucosidase (EC 3.2.1.21) (Howard *et al.,* 2003). Lignocellulosic biomass is mainly comprised of cellulose and therefore, cellulolytic enzymes are some of the key components required for the enzymatic saccharification of lignocellulosic feedstocks to the fermentable sugars during industrial bioethanol production (Bayer *et al*., 1998).

The industrial hydrolysis of lignocellulose under high-temperature conditions, using thermostable enzymes, has numerous advantages over mesophilic processes which utilise mesophilic enzymes (Turner *et al.,* 2007). The most important of these advantages is that under thermophilic conditions the solubility of polymeric substrates is significantly enhanced which in turn increases the enzyme(s) penetration and substrate accessibility (Shao and Wiegel, 1995; Viikari *et al*., 2007). Furthermore, thermostable enzymes are more compatible than their mesophilic counterparts with the conditions experienced during industrial processes, such as the high-temperatures associated with steam pretreatment, which is designed to decrease the crystallinity of the cellulose polymer (Szjarto *et al*., 2008). In a thermophilic production process the costs associated with cooling of large industrial-scale bioreactors prior to enzymatic hydrolysis and upon end-product collection, may be decreased or be eliminated entirely by employing thermally active and stable enzymes (Turner *et al*., 2007; Viikari *et al*., 2007). Various

other advantages are also associated with thermogenic bioethanol production, as outlined by Turner *et al*. (2007), and a brief synopsis of these is now given. Streamlining the bioethanol production process, by reducing the cooling steps in production, may help to reduce the risks of volatile product formation. Mass transport costs within the production facility may also be decreased, due to reduced fluid viscosity and an associated reduction in the power requirement of pumps. In addition, risks of microbial and phage contamination are significantly reduced (but not entirely eliminated) at elevated incubation temperatures. Thermoenzymes may be stored at room temperature without the risk of inactivation, thereby reducing the need for additional refrigeration within a bioethanol production facility. Finally, the greatest advantage associated with the use of thermostable enzymes and the operation of a thermogenic process, is enhanced downstream processing of the volatile ethanol product during distillation following fermentation. The numerous advantages of employing thermostable enzymes within a high temperature production facility are significant as approximately half of the projected process-cost in biomass conversion is associated with enzymatic saccharification (lein-Marcuschamer, 2011). Therefore, the production benefits attributed to the use of thermostable enzymes may result in an improvement in the overall economy of the process as a whole.

Most of the thermostable microbial cellulases identified to date have been isolated from thermophilic microorganisms using a culture-based approach (Duan and Feng, 2008). However, the rate at which novel biocatalysts are discovered using traditional cultivation methods is extremely low (Daniel, 2004). Examples of thermostable cellulases identified through culture-based techniques include those isolated from *Acidothermus cellulolyticus* (Mohagheghi *et al*., 1986), *Clostridium thermocellum* (Ng *et al*., 1977), *Thermomonospora fusca* (Sakon *et al*., 1997), *Streptomyces albaduncus* (Harchand and Singh, 1997) and *Thermotoga neapolitana* (Bok, 1994). In comparison, the relatively recent developments in culture-independent approaches, which include metagenomics, have successfully been employed to identify a range of cellulases (Duan and Feng, 2008). Metagenomics is a culture-independent approach to gene discovery that enables us to access and exploit the tremendous genetic potential

contained within presently unculturable microorganisms (Handelsman *et al*., 2004; Kennedy *et al*., 2008; Simon and Daniels, 2009). Cellulases identified through functional metagenomic gene discovery programs have been isolated from various environmental sources, including soil (Kim *et al*., 2008; Jiang *et al*., 2009; Berlemont *et al*., 2011; Nacke *et al*., 2012), the hindgut of termites (Warnecke *et al*., 2007), compost (Pang *et al*. 2009), rabbit cecum contents (Feng *et al*., 2007), sludge from a biogas reactor (Jiang *et al*., 2010) the contents of an abalone digestive tract (Kim *et al*., 2011) and cow rumen (Hess *et al*., 2011).

Numerous investigations have demonstrated that metagenomic techniques, including both functional- and sequence-based screening, are an effective strategy for revealing the functional capacity of microorganisms which cannot be cultivated through traditional techniques (Handelsman *et al*., 2004; Duan and Feng, 2008; Simon and Daniels, 2009). Metagenomics allows access to the functional and phylogenetic composition of complex microbial communities as found in their natural habitats. Extremophilic microorganisms inhabiting extreme environments are likely to produce biocatalysts that function under harsh conditions thus improving their suitability for inclusion in industrial processes (Zhang *et al.,* 2006, Gerday and Glansdorf, 2007). Therefore, metagenomic libraries generated from extreme environmental samples potentially contain extremophilic microorganisms that increase the chance of identifying novel industrial biocatalysts capable of operating under extreme conditions. For example, in the search for a thermostable enzyme one should construct and screen a metagenomic library from a thermophilic environmental source as the microorganisms, and the enzymes they encode, found within high temperature environmental niches have evolved to operate optimally at elevated temperatures.

In this study, a novel thermostable cellulase was successfully identified from a metagenomic library constructed from an environmental sample characterized as being both high in temperature (72-78°C) and in decomposing organic matter. The results from this study clearly demonstrate the importance of targeting the correct

environmental source for the isolation of a particular enzyme with the desired process characteristics, in this case the thermogenic saccharification of lignocellulose for bioethanol production. Many reviews have emphasized the use of extreme environments, such as hot springs, for the construction of metagenomic libraries and the subsequent screening for, and isolation of, thermophilic enzymes (Schloss and Handelsman, 2003; Handelsman, 2004; Ferrer *et al*., 2007; Simon and Daniel, 2011; López-López *et al*., 2013). Although many cellulases have been identified from hot spring environments (Mawadza *et al*., 2000; Ibrahim and El-diwany, 2007; Graham *et al*., 2011) none of these were isolated using metagenomic approaches. Enzymes which have been identified from hot springs using metagenomic approaches include xylanases (Whang *et al*., 2012) and lypolytic enzymes (Rhee *et al*., 2005; Tirawongsaroj *et al*., 2008). The Malawian hot spring metagenomic library constructed and screened by XP Hu during her MSc project included the characterisation of XPgene12 (Hu, MSc thesis, 2010) and together with the current study, the characterisation of CelMHS, are to the best of our knowledge the first descriptions of cellulases from a hot spring environment identified using metagenomics.

In terms of screening for desired gene(s), metagenomic gene discovery investigations are characterized as being either sequence- or activity-based (Simon and Daniel, 2009). As previously discussed (see section 1.5), activity-based functional screening offers a number of advantages over sequence-based approaches (Rees *et al*., 2003; Heath *et al*., 2009; Simon and Daniel, 2009). One of these is that novel gene sequences may not share similarity to previously characterised genes of known function contained in publically available databases such as GenBank. In this study, we characterized a novel cellulase gene, *celMHS*, which had been identified following functional screening of a metagenomic library. The *celMHS* gene was identified from a fosmid clone displaying cellulolytic activity after subjecting it to 454 sequencing. Preliminary bioinformatic analyses of the fosmid sequence in June 2010 resulted in the identification of open reading frames (ORFs) with no previously characterised cellulolytic activity. However, one of these, later identified as *celMHS*, comprised of a single putative CBM domain

and no clearly defined catalytic domain. However, due to newly identified and characterised sequences constantly being deposited into the GenBank database, *celMHS* has subsequently been shown to have identity to a number of putative cellulase genes (section 3.3). Thus, there are short-comings associated with both activity- and sequence-based approaches to functional metagenomics, as in both instances researchers are reliant on the presence of domains with sequence similarity to previously identified genes of similar function for the identification of the gene of interest.

In some instances putative gene sequences may not even have partial sequence similarity to previously characterised genes and therefore a method to circumvent this problem may be employed. One such method is transposon mutagenesis and involves the random integration of a transposable element flanked by sequencing primer sites into a metagenomic library clone (Simon *et al*., 1985). Should the integration of the transposon take place within the gene encoding the desired enzymatic activity, it may result in the inactivation of the gene, leading to a loss of the phenotype. Subsequently, the gene responsible can be identified through sequencing using the transposon's sequencing primers. Methods to identify gene sequences encoding desired enzymatic activity may become even more important as the number of novel sequences being deposited onto public databases increase with the increased popularity of next-generation sequencing (NGS) technologies, together with the significant number of metagenomic gene discovery programs being conducted globally. These gene mining studies will need to be done in concert with functional and/or structural annotation of the novel genes to ensure a better understanding of the relationship between gene sequences, protein structures and functions of possibly entirely novel biocatalysts.

Metagenomic gene discovery involves the cloning of large fragments of metagenomic DNA and subsequent screening of the large metagenomic libraries in recombinant host systems. The successful heterologous expression of genes of interest contained within

these cloned fragments requires that the cloned genes contain the appropriate promoter elements for efficient transcription and subsequent translation of the ORFs from the chosen host expression system. This is often the limiting factor experienced during the functional screening of metagenomic libraries (Warren *et al*., 2008; Fakruddin *et al*., 2012). In the current study, the metagenomic library was expressed in an *E. coli* expression system and analysis of the sequence upstream of the gene encoding CelMHS revealed the presence of -10 and -35 sequences and a putative RNA polymerase transcription factor binding site (rpoD) upstream of the predicted transcription start site. These elements are conserved in *E. coli* promoter sequences and may explain why the protein was successfully expressed in *E. coli* thus allowing the identification of the positive cellulase fosmid clone during functional screening of the metagenomic library. A putative transcriptional start site was predicted within the promoter element of *celMHS,* however, this needs to be confirmed by means of primer extension analysis of the *celMHS* mRNA transcript.

Despite the numerous advantages of metagenomics as a research approach for novel gene discovery, there are unfortunately also a number of negative aspects that we have identified during the course of this study. Such a disadvantage is not having access to the microbial isolate from which this portion of metagenomic DNA originates. Being able to culture the wild-type microorganism would give us access to the total potential of the microbial gene pool. For example, it stands to reason that there may be additional thermostable gene(s) of interest within the genome of the microorganism from which this portion of metagenomic DNA, encoding a novel highly thermostable cellulase, originates. A follow-up study, employing a ‘reverse metagenomic’ approach, wherein the cultivation of this microorganism is attempted, may ultimately lead to the successful identification and isolation of the organism of interest (Tyson *et al*., 2005).

The specific combination of conserved domains proposed to be contained by the novel cellulase, CelMHS, identified in this study were quite different from those of previously

characterised cellulases (Fig. 1.7). CelMHS contains a putative transmembrane domain, a family 4 CBM, a truncated GH family 42 domain and a large N-terminal region that does not have homology to any previously described domains (Fig. 3.6) We have speculated that this region may contain an unclassified domain. Further experimentation, involving several truncations of the carboxy terminus can be performed, to determine whether these regions are critical for either endoglucanase or exoglucanase activity. The proteins with the greatest similarity to CelMHS, identified by blastp analysis, were shown to have similar domain architectures. However, to the best of our knowledge none of these putative cellulases have been characterized to date, making this the first data available for an enzyme with this unique domain architecture.

Proteins comprising a transmembrane domain are known to be involved in a variety of important biological functions, where they often play key roles as either receptors or transporters (Hicke and Dunn, 2003). On-going future research, conducted on CelMHS will attempt to characterize the domains identified in this study as well as the large uncharacterized N-terminal region. The transmembrane domain could be investigated to confirm if it is only involved in secretion of CelMHS or if it is in-fact an anchoring domain that results in CelMHS maintaining a close association with the cell. The transmembrane domain could be assessed by assembling a fusion protein construct in which the transmembrane domain is linked to a reporter protein such as green fluorescent protein (GFP). The presence of the protein and its interaction with the cell membrane may then be detected and analysed using fluorescent microscopy (Rolls *et al*., 1999).

CelMHS is also predicted to contain a family 4 CBM. To date, the role of cellulase CBMs in substrate affinity and enzyme activity has been determined through the construction of gene-truncation mutants (Zverlov *et al*., 2001), domain exchanges (Tomme *et al*., 1995; Srisodsuk *et al*., 1997) and site-directed mutagenesis of the CBMs (Simpson and Barras, 1999). Therefore, on-going analyses of the CelMHS_CBM should

include the construction of CelMHS mutants with substitution of the current CBM for CBMs of different families from other lignocellulosic enzymes, as well as site-directed mutagenesis of key conserved residues within the CBM substrate-binding region. In addition, it would be interesting to express a recombinant CelMHS_CBM, separately from the CelMHS catalytic domain(s), and investigate its substrate binding affinities. It may also facilitate the crystallization of the CBM, allowing us to confirm the modeled tertiary structure presented in this study. Furthermore, it would also be of interest to characterize and compare the binding potential of this domain to closely related CBMs, as this may reveal novel functionality within this domain.

The low level of structural homology between CelMHS and previously characterised enzymes makes this enzyme an important candidate for crystallization studies. Therefore, we propose that future investigations of this enzyme should consider including a structural aspect, during which the crystallization of CelMHS with and without a polymeric substrate should be attempted. The resulting crystal structure may help to identify the catalytic domain of CelMHS, as well as any additional CBM domains and the specific amino acid residues involved in substrate binding and catalysis. Furthermore, and although outside the scope of this study, it would be interesting to investigate the sequence and structural aspects of CelMHS that contribute to the extreme thermostability reported in this study. During future investigations of the thermostability of CelMHS the construction of truncated CelMHS mutants, as well as CelMHS mutations of conserved amino acid residues identified from multiple sequence alignments of related proteins, should be considered to determine the role(s) of the conserved domains (see Fig. 3.2). Identifying specific amino acid residues and/or structural aspects contributing to the thermostability of CelMHS may help future attempts to rationally engineer other biocatalysts for thermophilic applications. In addition, resolving an accurate crystal structure of CelMHS may aid in the engineering of this novel enzyme, to potentially enhance its suitability for inclusion in a thermogenic bioethanol production process.

The hydrolytic capacity of CelMHS was determined and the recombinantly expressed enzyme was shown to have the highest activity on amorphous cellulose substrates such as β-glucan and CMC. This suggests that this enzyme should be classified as an endoglucanase. However, the activity of CelMHS on a crystalline form of cellulose, namely avicel, suggests that this enzyme may be classified as a glucanase with both endo and exo activity. Bifunctional cellulase activity has previously been described for CelB, a dual acting cellulase from the extreme thermophile *Caldocellum saccrolyticum* (Saul *et al* 1990)*;* and CelA a dual acting cellulase from the extreme thermophile *Anaerocellum thermophilum* (Zverlov *et al*., 1998). In both cases the proteins displayed a multidomain structure, which at least two catalytic domains. As discussed, CelMHS contains a large N-terminal region that does not have homology to any previously characterized domains. As CelMHS has dual cellulase activity, this further strengthens our hypothesis that the unmapped regions may contain one or more catalytic domains.

Results of cello-oligosaccharide hydrolysis by CelMHS indicate that the shortest chain length CelMHS is able to act upon is cellotriose, however, cellotetraose was its preferred substrate as the substrate was completely hydrolysed to cellotriose, cellobiose and glucose. The dominant product released by CelMHS was observed to be cellotriose, and to a lesser extent cellobiose and glucose, whereas, the major product released from other metagenome derived cellulases have been found to be cellobiose (Voget *et al*., 2006; Feng *et al*., 2007). CelMHS appears to have an increased affinity for the hydrolysis of longer chain oligosaccharides, which may be due to the presence the family 4 CBM (Notenboom *et al*., 2001; Xie *et al*., 2001; Pell *et al*., 2003). With most CBM's an increase in oligosaccharide length correlates with an increase in binding affinity until the saccharide is of sufficient length to occupy the entire binding site, at which point the affinity ceases to improve (Boraston *et al.,* 2001). Studies have confirmed that family 4 CBMs employ a type B glucan chain binding mechanism which increases their affinity for binding longer chain oligosaccharides (Boraton *et al*., 2001; Notenboom *et al*., 2001; Xie *et al*., 2001; Pell *et al*., 2003; Boraston *et al*., 2004).

In vitro digestibility tests to evaluate the capacity of CelMHS to degrade a natural lignocellulose substrate, sugar cane bagasse (SCB), revealed that the enzyme had little to no activity on the complex polymeric substrate. Factors that may affect the hydrolysis of SCB by CelMHS include the possibility that the enzyme is binding irreversibly to the complex substrate. CelMHS was shown to bind irreversibly to crystalline cellulose, avicel (I. Ackerman, personal communication). Lignocellulosic enzymes have been shown to function in a synergistic manner to degrade complex polymeric substrates (Van Dyk *et al.,* 2012). This means that the synergistic activity of multiple enzymes on a substrate may be higher than the sum of the activities displayed by single enzymes. Future investigations should consider conducting multi-enzyme hydrolysis trials, by combining CelMHS with thermophilic exoglucanse as well as a β-glucosidase to investigate the synergistic effect of this suite of cellulolytic enzymes. Similar studies have shown that the inclusion of xylanase(s) improve the degree of degradation of lignocellulosic substrates (Garchia-Aparicio *et al*., 2006). Therefore, depending on the composition of the substrate under investigation, the inclusion of additional accessory enzymes, such as xylanases and other hemicellulolytic enzymes, within the cocktail of enzymes may help to improve the quantity of fermentable sugars liberated from the polymeric substrate.

As discussed in chapter 1, cellulases are the third largest group of industrial enzymes produced in terms of quantity. They have a broad range of applications in various industries including the textile industry, detergent industry, food industry and in the paper and pulp industries (Ibrahim and El-diwany, 2007). Industrial enzymes such as Pyrolase® a broad spectrum cellulase produced by Verenium Corporation, is able to efficiently hydrolyse guar gum and carboxymethyl cellulose. Pyrolase has a braod pH range (6.0-11.0) and a temperature optimum of 104 °C. When compared to the functional characteristics displayed by CelMHS (see section 3.6) and other metagenomic library derived cellulases (Table 4.1), it is clear that the characteristics of this enzyme are superior to that of other metagenome derived cellulases and that for the most part it is equivalent to that of industrial enzymes such as Pyrolase®. The

characteristics displayed by CelMHS may therefore be of more value to industries requiring robust thermally stable cellulases. With the completion of the additional experimental work proposed in this discussion, recommendations towards the possible engineering of CelMHS for these processes can be made.

Table 4.1. Metagenomic library derived cellulases and their enzymatic properties

Enzyme name	Meganomic source	pH	Temp. (°C)	Thermostability [a]	Substrate specificity	Mode of action	Reference
CelMHS	Hot spring soil	6.0	100	70°C for 60 min	β-glucan	Dual endo/exo	This study
Cel5A	Agrolytic enriched soil	6.5	40	40°C for 240 min	β-glucan	Endo-acting	Voget *et al*., 2006
CelM2	Koren soil	4.0	45	35°C for 60 min	β-glucan	Endo-acting	Kim *et al*., 2008
*bgl*1C	Alkalin polluted soil	8.0	42	28°C for 180 min	p-nitrophenyl -β-D-glucoside	Exo-acting	Jiang *et al*., 2009
Umcel9B	Compost soil	7.0	25	N/A	CMC	Endo-acting	Pang *et al*., 2009
CelAM11	Abolone digestive tract	7.0	37	N/A	CMC	Endo-acting	Kim *et al*., 2011
Cel01	Grassland soil	7.0	45-50	N/A	β-glucan	Endo-acting	Nacke *et al*., 2012

[a]Highest reported temperature and incubation time at which the enzyme retains 100% activity.
Kinetic parameters were not included in this table due to inconsistencies in the reporting of kinetic data.
N/A: Not applicable

In conclusion, this study has successfully demonstrated the potential of functional metagenomics for the discovery of novel thermophilic biocatalysts from environmental samples. In addition, it provides further evidence that the currently uncultured soil sediment microbiome from this Malawian hot spring represents a genetic reservoir of novel enzymes with potential applications in a variety of industries. Results from this investigation suggest that CelMHS may form part of a new glycoside hydrolase family, and on-going research of CelMHS at the Institute for Microbial Biotechnology and Metagenomics may verify the status of this potentially new family of glycoside hydrolases. Furthermore, investigations being conducted on this novel cellulase may

provide invaluable insight into its protein structure, particularly with respect to its extreme thermostability, catalytic domain(s) and functional mechanisms, as well as re-evaluate the functional classification of related proteins currently designated as being unclassified or hypothetical in databases.

Supplementary Figures and Table

```
1       CCCATAAGGC GGCGAGGAGT GCAAGGCAAC CAAAATCCCC ACGCTCAAAA GCAGCGCCCC
61      CAAAACCAAA ACCGCATTAA GGGTCGGGTT CTGCACCGGT GTCAGCATCG CTCTCTCTGC
121     CCCCTCAAAG CCCAAAGAAT TGCAGCACGG CTGGAGGAAG GAATTTCTGG GTTAGGGTGA
181     CGGAAAAGAC GAGGGTTGCC AAGAGGATGC GGGAGATGTC CTGCCCCCAT ACGCTGCCCA
241     GCAAAATGGG CTCCCGGGTG AGGTAGGCGG TGACAGCGTA ATACTCTTCG CCAATCACGG
301     TGAAGTCACA CGCCGCAATG AAGAAGGGGA TTTGCAAGGT GGAAGGCGTG CCAGCGATTT
361     GGATGGCGCC CAGTGCCTGT CCCGTTTCCG CCAGAATGAG GGATTCGGCA GCGAAGGAGC
421     CGAAGAAGAA GTTGGCTGCC ACCCGCTCGC GGTTCATCAA CCCCACTACA CCAGAAGCGA
481     AGGCGAATTG GTCACCAGAA AGGAAGCGCA CATCCTCGGG ATGGAAGGCA TCGGGGGCGC
541     CTCCCTCGGC ATAGGCTTCC CGCAGCACCT GCTCCGCCAT TGTCAAGATA ACCGTATCCG
601     TAACGGGCAT CAGGATCCGG CTGCGGTAGC GAGCTGCGAG ACGAGCAATG TATCCTGCGA
661     TCGCAACAGC TTGGAAAGTC ACCACATCTA CGGCAGCCAA GCCGGGCACG AAGAGCATCG
721     GGCGCCCCAT CTCCGTTGCT CGCCCCACCG CTTCTTCTAT GGCATCAATG GCGGGGAGCC
781     GACGGATGCG CAGGGGCTTC CCTGAGGAAG CCAGTGCAAT CCCCAAACCG ATAAGGGCAG
841     TCAACACCCA AACCAGCCAC GGATAAAGGG TCGTCCAATC CGCATGCTTC ACCCTCTCAC
901     CCTCCTTTCA CAAAGATGCC TTCCCCACAA TCCGCCAGCC ATTGCCAAAA GCAATTAAAT
961     TTTGGCGCTT CGTGTCCGTC AGTGACGAAA GGGCTGCAAA AAAGCATCGC GTCTAAACCA
1021    TCGTTGTCGC GCCGCTGCGA AAGTCACGCC TTCCTTTCCC CCTCTGTCCT TCCAACCTTT
1081    CCTCGCAGGA TTGGTTGGCA AACCCGTATG GCAGGGCAGA AACTTTCTTG GTGCGCTATA
1141    ATGCATCCCG CGCCGAGGTA AAGAAGCGCT CATCCACGCA AACGCTTGTG CCAAGGAGGG
1201    ACATAGATGC CGCGAACGAA GTCAGCGTTA AAAGCCTTAC GGAGAGCGGA AAAACGGCGG
1261    ATGCGCAACC AACCTTATTA CACACGCATG CGCAACAGCA TCAAGCAGGC GCGCCAAGCT
1321    ATGGAAAGGA CTGCCAAAGG CACTGGGAGC TTGGAAGAAG CTCAAAAAGC GGTTTACGAA
1381    GCGTGCCGCA TCATTGACAA GACGGCGTCT AAAGGGGTCA TCAAAAAAGG TGCAGCTGCC
1441    CGCTATAAGT CCCGCTTAAT GCGCCGACTG AACGAGCTCA TCCATCAAAA GGCTGTCGTT
1501    GAAGCGGCGT AAGGGCTGTA GCAGTGTTGG GGTGATGAAA GATGCGCGGG TGGTGGCTGT
1561    TGGTGGCTTG TGCCATGTTG GCTGGTATGG GAGGATGGTG GACAATGCGA CAGGGAGCAT
1621    CCCAAGCAAC GCCGCAAACT GCTTCCCCAC TGTTCCCTTT CGTGCTCCCG TGGGATGACA
1681    ACGCACCGTC GGTGACCAAC CTCAACGGTT GGCTCCACCA ACCAGCGGGA AGGTTCGGGC
1741    ATGTGCGCGT GGGTGCGGAC GGGCACCTCT ATGTCGGTCA GCGGCGCATC CGCTTTTTCG
1801    GAGTCAACCT GTGCTTCGGC GCTTGCTTCC CCCGCAAAGA AGATGCCGAA AAGATTGCTG
1861    CCCGCATGGC GAAGTTCGGC ATCAACATCG TCCGCTTCCA CCACATGGAC ATGAACCCCT
1921    TCCCAAACGG GATCCGTGCC CGCAATGTCC CCCACACCCG CGACCTTGAC CCTGAAGCTT
1981    TAGATCGCTT GGACTACCTG ATTGCCCAAT TGAAACGCCA CGGCATCTAT GTCAACTTGA
2041    ACTTGTTGGT CTCTCGCCCC TTCAATGCTG CCGACGGGTT GCCGAAAGAG ATTGAGCAAC
2101    TCGGCTGGAA AGAACGGCAC ATTGTCGGCT TTTTCTATGA GCCATGTTTG GAGCTGCAGA
2161    AGGAGTATGC CCGCAAGCTG CTGACCCACC GCAACCCCTA CACAGGATTG ACTTATGCGG
2221    AAGCCCCCGT AGTCGCTTTC GTGGAAATCA ACAATGAAAA TGGCTTGTTG CACGCTTGGT
2281    TGGGAGGCGA TGTGGACCAA TTGCCGCCAG TGTTTCTGCA AGAATTGCAG CGGCAGTGGA
2341    ATGGGTGGCT CAAAGCCCGC TACGGCAGCA CGGCAAAGCT GCGGCAAGCA TGGGGCGTTC
2401    AGGAAGAACC GTTAGGCACC GAACTGCTGC AAAACACCAC CTTCGCGACA GGGTTGCAGG
2461    GATGGGTGTT GGAGCAACAC GCCAACGCCC AAGCGGCAGC GGAAGTGGTA GCCGACCCCA
2521    TTCCCGAGTT GAAAGGGATG CGCTTTGTGC GCATCACCGT CACCCGCCCA GGGACAGCAG
2581    GGTGGCATGT GCAATTTCAC CAACCAGGGC TCAAAGTTGC CCCCGATCGC CCTTACACCC
2641    TGTCCTTCTG GGCGAGAGCG GAGCGCCCTT GCCGCATCAC CGTCGGTCTC AGTCAAGCCC
2701    ACGAACCGTG GCAAAACTTG GGCTTCTCCG CCAATGTGGA CCTGACGCAA GAGTGGCGGG
2761    AGTATCGGTT CACCTTCGCC CTCGCCCGCG GGGACGACAA TGCGCGGGTT ATCTTCAGCA
2821    ATCTGGGTGC CCAGACAACC ACCTATTGGT TTGCCGCGCC TTCCTTGCGC CCTGGAGGGA
2881    TTGTGGGCTT GAAGGAAGAC GAACGGATAG AAGACACCAC CGTGCCTCCC TTCCGACGGG
2941    ACCGATTGGG CGAACGAACA CCGGAAGCGC AACGGGATTG GATGCGGTTC CTGTGGGAAA
3001    CGGAAGATCG CTACTGGCAG ACCCTCTACC GCTACTTGAA AGACGAGTTG AAAGTTAAGG
3061    CGTTGGTCAT GGGCACCATC GTTGGTTGCA GCACCCCCAA CTTAATGGCG AAACTGGACG
3121    GGGTGGACAC CCATGCCTAT TGGCAGCACC CCCTGTTTCC CACCCGCCCG TGGGACCCTG
3181    AAGACTGGAT TGTGCCGAAC CGCACAATGG TGAATGAGCG CGGGGGCACC TTGCCTGGGC
3241    TCGCTTTGCG GCGCGTGTTA GGCAAACCCC ATTCGTGCAC CGAATACAAC CATCCCGCTC
3301    CCAATACTTA CAGCAGTGAA GGCTTCTTGC TGTTAGCTGC TTATGCTGCC CTGCAGGATT
```

```
3361   GGGATGCTAT CTACGCTTTC TCCTACTCCC ACCGCCGTGA CGAATGGGAT TTACGCCGCA
3421   TCCCCAACTT CTTTGACATT GACCAGCACC CCACGAAGAT GGTGACGCTG GTCCCCGCCG
3481   CAGCACTGTT CTTGCGCGGG GATGTGAAAC CGGCGAAGCA GCAGGTGGTC GTGCGGCTGA
3541   CAAAGGAGCA AGAAGTGGAC TTGCTGCGCC GCAGTTGGGC GTGGGTCCTC GTGCATGCAG
3601   GGCATGTCGG GGTGCCGAAC GAGGTCGCGC TTGTCCATCG CGTCGCCATC GCAACGGAGG
3661   GCAAACGGGT GCCGCCCACA GCGTTGAAAC CTGAACAAGT GAAAATCGCA GGCGAGAAGT
3721   TTGTTTCAGA CACAGGCGAG CTCGTGTGGG ACTTGACCGA GAAGGGGCGG GGTGTCGTGA
3781   CGATCAACAC GCCCAACAGC AAAGCCGTCA TCGGTTACGG GGGCGGCAAA GTGTTCCCTT
3841   TGGGCAATGT GGTCATTGCG CCTGGTCAGA CCCTGCAAGA AGGGTGGTGC GCCATCACCC
3901   TGACGGTTAT GCAAGGCAAA TTGCCTGCGA AGATGACCTC CACCGTGCCC CGCAATTTGC
3961   GGTTGCTGAT CACCGCCACA GGCTATGCCG AGAACACCAA CATCGGTTGG AAAGAAGTGC
4021   CAGGGTATCC GCCCAAATCC AGCTGCGGTC GCAATTGGGG TTCCCCTCCG TCGCTGGTGG
4081   AAGGCATTCC CGCCCGCATC ACTTTGCCCC TGCCTGCCAA GCGCGTGAAA GCCTGGGCGT
4141   TGGACGAACG CGGGCACCGC AAAACCGCCT TGCCTGTCAG CGCCGATGCA AAGGGCAGCG
4201   CCGTGGTGGA AATCGGACCG CCGTATCAAA CGCTGTGGTA TGAAGTCGCT GTAGAGTGAC
4261   GACGGCTGCC TTCACTGTTC CACCCCCGCC AACAAGGGAT CGGCTGCCGT CAAAAACGAG
4321   CAAAGCGGCT CGTTGCTCTT GAGATGCAAA GGCGCCGCAG TTTTCGGAGG TGCCTGAAGA
4381   GGTGTGCTCC CTTTCGCTCC AACAGCCCCT CTTGCTCGCC TTTCTCAAGG TTCGGGTTGT
4441   CTTCTCCATC GCTCCGCCCC AAGAGGCACC ACCTTGCCTG CAAATCGCTC AAAACACCGT
4501   TGTCAAGAGA GCGTCGGCGA AGCGGTAACC CCCTTGTCGG TGGCACCGTC CTTCCCTTCT
4561   CCCATTGCTC CCCTCTCCGT GTTTGGCTCC TTGCGCTATA ACCTACGGGG CTGCTGAAGG
4621   CAAAGGAAAA GGGGGTTGGA GAAAGTGCTC TGGGGCTTTG CGGGACCTTG GTATGGGGAG
4681   TTTTTGAAAG ACCCGCAGGG ACGGGCGCCG AAGACCGAAT GGGAGCGCCT TGAGCGACAG
4741   CTGCAGTTTT TGGTGCGCTA TGGGCTAAAG GCGACAGGTT GTTCCCTTCA AACCATCACA
4801   GCACTGGACA AACACCAGCA AGACCAACTG GCAAATTGGC TCGCCCAACA TGACCTCCAC
4861   CTGGTGCTCC ATGTCGGCTA CGATTACTTG AACGCCGATG CGGAAACGAT ACAGCGGAAG
4921   ACGGAAGCAA TCCTTGCTGC CCTATCCCGC TACGCGCCTA TGATGCGGGT AGCGTTGGTC
4981   GCCACGACAG CCCACGCAGG ACACCGCTTT GACCGAACGA TGCCCGTTGC GGCGAAGCTG
5041   GAGCGCCTTG CGACGGCGCT CACCCCTTTA GCGCAAGGTT GCCGCGAATT GGGGCTTCCG
5101   TTCGGCATTG AAAACCACTG CGATTTTTAC ATCTCCGATT TGGTGGCTCT TTGTCGCTCC
5161   GTGCCTTTCT TGGGCATTTT GCTGGACACG GGCAACCCTT TCGTTATCGG TGAGAAGCCG
5221   CTCCCTGCTT TTGAAGAAGC GGCACCATAT GCGGTAGGCA CGCACTTTAA GGACCACTTC
5281   GTCCGCCCGC GCCCAGACCT GTTTCCGCTC TGCTTGGAAG TGGACGGTGC TCCGTTGGGA
5341   GAAGGGGATG TGCCTTTGCA GGAGTGCTAC GAACTCCTGA AGCGGAAAGC CCCGAATCCC
5401   ACTAAACTGG TTATGCTTGT TGAGATGGTG GCGCCAAAAG GCATGAGCCC GTTGGAGTGT
5461   TGGGAGAAGT CGTTGCAATT TGTGCGCGCT TTGATGGAGG GGGATTGACA TGGGTGGGCA
5521   AAAGTTGCGA GCGGCTGTTG TGGGCTGCCG CATGGGGGCT GAACACGCCC GCGCGATGGC
5581   GCAACTGGAA GACTACGAGG TCGTGGTGGT GTGCGACATT GATGAAAACC GTGCTCAAGC
5641   ACTTGCCAAA GAGCTGAAGA CTTGCCAGCC GTTTACAGAC TACGGGGCGA TGCTACGGGA
5701   CATTCAACCG GAAGTGGTTG CCGTTGCAAC CCCGACAAAC TTGCACGCAG AGCATGTTTT
5761   CCAAGCCATT GAGGCAGGGG TGAAAGCCAT TTGCTGCGAA AAGCCGATGG CGACGAACCT
5821   TGCGGAAGCG CGCCAGATGG TAGCGTTGTG CCGCGAGCGC GGTATCGCCT TGATCGTCAA
5881   CCACCAACGG CGGATGGGTT TGGAAATGGT GACAGCGCGA GAGCTGATAG AGCAGGGGGC
5941   AATCGGCGAG GTGCGCCTCA TTCGTGCCCA TTGCGCCGGT GACTTCCTTA CCGACGGCAC
6001   CCACGCTGTA GACAGCATCC TCTGGCTTCT CGGTGATCCC GCGATAGCTT GGGTCTTTGG
6061   GCAAGTGGAA GTGCCGGCAG AACCGAAGTG GCGTTACGGG CACCTGGTGG AGACAGCAGC
6121   GCTCGCGGTT TTTCAAACGG TGACCGGCGT TCGTGCAGAA ATTTTCTGCG GCGAAGCGCG
6181   GGAGCGCCAC CGCGCTTATC AAGACTACGA AGTTATCGGA ACACAGGGAC GGTTGTGGCG
6241   CACAGGGGAT AGCCCCCACC CCAACCTGTT TGTCCAGGAC GCTAAAGGCG GTTCGTGGGA
6301   TGCTGCTATT AGCAACGGTC AGTTGCGTCC TGTCCCTACG AAAAGCGCTG GGCTCTGGCG
6361   CCCTGTGGAC GCGCCTCCCG ATGATGCCTT AGGCGCAATG ACCCGCTCGT ATCAACTATT
6421   GGCTCAGGTA GTGCGGGGCA GTGTGCCCCT GCTCCAACAC CCCCTCGCGG ATGAAAAAGC
6481   ATTGCGGGGA TTTGAAGTGG TGATGGCGGT CTACGAATCA GCACGGTTGC GGCAGCGAAT
6541   ATCCCTCCCC TTGGAACAAG AACGCTTCCC GTTAGAATTG ATGGTGAGCG GAGAGGAGTA
6601   GTCTTCGGGA GAAGTCTGGG GGGATGGACA GTGAAGCGTT GGTGGCTGTT TGGTGCTTTC
6661   GCTGTATTGC TCAGCCTTGG TGTTGTGTTG GCAGCGCCCC GCACCCTACA GGATGATTTG
6721   GGGCGGACGG TGACCGTGCC CCCCACTGTT CGGCGCATCG TCTCTTTAGC CCCAAGCGGG
6781   ACGGAAAACT TGTTCGCTAT CGGAGCAGGG CAGCTGGTCG TTGGGGTCAC TTCCGCTTGC
```

```
6841   GATTACCCGC CGCAAGCGAA GAAGCTACCG CAGGTGGGCG ATTTCATGAA GCCCAGCTTG
6901   GAACGCATTG CAGGATTAAA CCCTGATTTG CTCCTTGTGG TCAGTTCTAC CATCCCGCAA
6961   ACCATTGCGG ATGAGTTGCA ATCTAAAGCG AAAGTGCCAG TGTTTGTGCT CCAGCCGCAA
7021   ACGGTGGAAG AGGTGCTGCG CAGCCTGTTA GTGCTGGGCG ATTTAACTAA CCGACGCAAA
7081   GCAGCTCGTG CGTTGGTGCA GCGGCTGGAA CGGCGATTGC AAGCAGTCCG CACGGGCAAA
7141   GCGAACAAAC CCCGCCCCAA AGTGGTCGTG GAAATTGCTC CGCCCCCTTC GCTGATGGTG
7201   GCAGGTCCGA AAAACTACAT TGACAGCGCT GTTCGTTGGG CGGGTGGCGA AAACGCGTTT
7261   GCCGATGCCC CGCAACCGTT TCCCACCGTC AGTTTGGAAG CGTTGGTGGC GAAAGACCCC
7321   GATGTCTACA TCGTGCCTGT CAAGGGCAAA ACGCACCAGC AAGTGTTAGC CGAAGTGCGC
7381   CAGCGCCCTG GCTATGGGCA CCTGCGGTGC GTTCGGGAAG GGCGGGTGTA TGCCGTTGAC
7441   CCTGACCTGC TGTTTCGCCC GACCCCTCGG CTTGTAGAAG GCGTAGAAAA GTTGGCAGCG
7501   CTGTTGCAGA AAGAGGCGCA GCCTATTTCC CCCAAGACGA TGCGCTATTG AGCTGGCTCT
7561   GCTTAGAACA ATGCCCTTTG ATCTGGCGCG GATTTCGGCG TTTCAGGGGG AAGTTGGCGC
7621   GCATAGAGGT GCGCTAAGCG GAGGGGTTCG GGGAGCCGTT GCCCTTCTTT CGTGCATTGG
7681   CGCACGAAGT TCACTGCTGT TTCCAAGGAC ACCTTGTGCC CAACGGAGAC ATAGACGGGC
7741   GGGCTGCCTT GCTTTGTCCG CACGACCATC CCTAAGACGC GCCCATCTAC AGGGTCTGAC
7801   AAAGGGGTCA TGGCGCCTGC TTCCTCCGCA ACGGGTTGAT GCCTACCGTA AAGACGACTC
7861   TTGGCACAGC CGATAGACGG CAAGTCCAAC AGCAGCCCCA TATGGCATGC CAAGCCAAAG
7921   TGGCGGGGAT GGGCGCGCCC CTGGGCATCA AACAGGGCAA GGTCTGGGTC ATGCCGCAGT
7981   TTCTCAAAAG CTGCCAACAC AGCGGGGCTC TCCCGAAAGG CGAGCAGCCC AGGGACATAG
8041   GGGAAAGTCA CGGGCACAAC CGCTCGCTGG ACTTCTAACA GCTCCAGTTC AGGAAACGAA
8101   AGGACACAAA CGACGGCTTG CATCTGGTCC CTGCCCACTT GGTGGACATC GGCACCTGCA
8161   ATATAGCGCA CCTTTTCTGG CAACTCATCG GCTTCTACAA CCCTTGTGGA CAACTCCCAC
8221   TGCAACCGTT GCGCTTCTTC CACTGTCATG TCCCAGCAAT GACTCAACTT CGCCCGCATA
8281   AAGAGACCCC CTCCGCATTT TCCATGAGCG TCATGCCACT TTTAGTTTAG CGACCGATGG
8341   TGGTATAATG GGAATGGTGC ACAAACTGCA GGCAATAGGA GGGATGGACG ATGCGAGTGG
8401   TGACTCGTCG TTGGGTTGGC TTTACCCTGA TTGAGCTGCT GGTGGTGATT GCGATAATCG
8461   CCATCCTGGC AGCCATTCTG TTCCCCGTGT TCTCCCGTGC TCGGGAGAAA GCCCGCCAAA
8521   CTTCCTGTCT GAGTAACCTG CGGCAATCGG CGACGGCGGT TAACCAGTAT GTCCAAGACT
8581   ACGACGAAGC GTTCCCCATG AGCGTGTATC TCGCCTTGAA CGCTTCCCTC CAACCCTGCG
8641   TCCTCACCAT GCTGGGTGCC GTTGAGCCGT ACATCAAAAA CAAGCAGATT TACCAGTGCC
8701   CATCGGAACC GCGGGCGATG GACCTTGATG CCGCCTTCCG TAATCTTGGA ATCCCTGGCG
8761   GTGAGTGTGG CGGGTTCCAG TATGGCAGCT ACCCCTTTAA CTTTGCGGTG TTTGAAGATG
8821   GCTATGTTCC AGGCTTGTTG AAGCCTGTCC CGCCTGTGCG GCTCCCCGAA ATCCAGTTCC
8881   CCGTAGAAAC TGCCATGAAC TACGACGGAA ACCTGGCAGC GCAAAGTGCT TGTGGCTTCA
8941   ACCTTTTTGA CTCGCCCGTG CAGGGGCGCC ATTCGGATTT TGTTGATGCC AGTTTTGTGG
9001   ATGGGCACGC AAAAGTCGTG AAGGCGGCTA AGTCTGGCTG CGTAGGCGCC AACATTAACC
9061   AGCAATCCTT AGAGCAATGG TGCCTCTTTC CCCCTTACAA TCGCTGGTGC AACAACCCCA
9121   ATCCCCCAGC CTGCAAGGAC GAATTGTGGG GCGTGGCAAT GCAAGACCAG TATGGCCTCT
9181   GCTACAGGGG CATTCGGTAG CCCACCGTTG GGTGCCTAAA GGCGATTTAT TCGCTGCAGC
9241   GATTTCATCG GATTTTCTGG GGAGGAATTC TTGATGACAA GACGAGATTT TATGCGGAAG
9301   GTGGGAGCCG CAATGGCTGT AGGGGTATCG GTAGCGCGGA TAGAGAGTCT GGATGCCCCG
9361   CAAGAAGGGT CGTTAGCCGA ACAACTGGGA TTCCGCAAAG AAGACCGCTT GCTGATCCTC
9421   CACGCCGACG ATCTCGGCAT GTGCCATGCC GTCAATGTCG CTTCCATCAA GGGGATGACG
9481   GAGGGGGTTG TGACCTGCGG GAGCGTCATG GTGCCCTGCC CTTGGTTCCC CGAAATTGCC
9541   GCATGGGCAC GGGAGCATCC CGAAGCGGAC TTGGGGTTGC ACCTGACCCT CAACAGCGAG
9601   TGGCGCTACT ACCGTTGGCG CCCCGTTGCC TCACCGGAGA AAGTCAAAGG GTTGCTGGAC
9661   GAGGAAGGGT TCTTGTGGCG GGATGTGGCT TCTGTCCGCC AACATGCCAC CGCCGCCGAA
9721   GTGGAGATAG AGCTGCGGGC GCAAATTGAG CGGGCGCTGC AGTTCGGGAT GAAACCCACC
9781   CACATAGACA CCCACATGGG CACCTGCTTC TCTCGCCCCG ATTTCTTCGA GGTCTATGTG
9841   CGGTTGGGAG CGGAATACGG TATCCTCCCG ATGCTCTTAG CGCCGACCCC CACTATCTTG
9901   GCAATGGCGA AGCAGCTCGG CATAGACTAT CAACCCATTG CGGTGAAGCT GCAACAGCAG
9961   CGGTTTGTGC TCTTAGACAA CCTCATCACG GGAGGCACGG GCGAAACACT GGAGGAGCGC
10021  CGCAAGAGTT ACTACGACCT TATCCGCAAC CTTTCCCCAG GCGTCACCGA GATCATTTTG
10081  CACTTGAGCA CCGACGACCC CGAAATCCGC GCCATCACCG GCGCTTGGCA TTATCGCTAC
10141  CACGAGTTTC TTATCTTCAC CGACCCGCAG ACCCGTCGGT TGATTGAGCA GGAAGGCATC
10201  AAACTCATCG GCTACAAACA CCTGGCTGCC CTGTGGCACC GCAAACGGCA ACGAGGCTAA
10261  ACCATAGGCG GCGTCATTTT TGCTTGCGCT GCCCAAAGAG CACAGTGCCC ACAGCGTTGG
```

```
10321  TAAGGCTGTA AAGCACGGTC ATCCCAAAGG CAGCGCCGAA GTCTTTAACA TGGAAGCCGC
10381  TGATCAATCC CGTTTTGCCG ATGGCAAAGA ACACGACAGC GTTCAGCACA AATCCTGCCA
10441  ACAGACCGAG GAGCCCAAGG GTGAGCAAGT TTAGGGGAAA GGTCGCCAGG CGGAAGAAAA
10501  ACAAAGCGGG GCGGACCAGA GCTTGGGCGA GCGCCAGCAC CAGCACGACT CCTGCTGCCG
10561  CACTGAACCC GCTGATTTCC ACCCGATTGC CGTAAACCCC TTTTGCCAGC AAAAAGAGCG
10621  CTAACGCGTT CCACAAGAAC CGATAAACAA GGCTGCGCAC AGTGTCCATG TTCAGGTGCC
10681  CCCTTTGCCT TTTTCCGCTC TCCAAAAGTG TGGCACGCTC CGTTGTTGTC TCAGCAGAAC
10741  TTCAAGCGCG AAGGGAAGAA GGCGGGGTGG GGAAGCGACG CTCTTGCGAA CCTTCTCAAG
10801  AGTTTGGTAG GAACTTCGTT TTTCGGGAGC GTCGGGACTT GTCCCCGACG GTCGTTCAAC
10861  ATCCGTCGCC TGCAAGAAGC GATGACGACT TAAACGGCCG TCGCCCACAA GGGGCGACGC
10921  TACCAACAAC TGTCGTGCAC GAATTGCCGC GGTAGCGTCG GGGCTTGTCC CCGATGTTTG
10981  TTCAACATCC ATCGCCTGCA AGGAGACGAC TTAAACGGCC GTCGCCCATA AGGGGCGATG
11041  CTACCGACGA CTTTCGTTCG TGAATTCTTG CGGTAGCGTC GGGGTTCATC CCCGACGGCC
11101  GTTTTTTGTC AACGGCCGTC GCCCACAAGG GGCGACGCTA CCAACAACTG TCGTGCACGA
11161  ATTGCCGCGG TAGCGTCGGG GCTTGTCCCC GACGGTTGTT TAACGCTCGT CGTCCATGAG
11221  AGGCGACGCG ACTAAAGGTT GTTCTTGCCC CTTGCAAGGT TGGGGACGCC CCATCATCAG
11281  GGACACCACG CAAC
```

Supplementary Figure 1. Nucleotide sequence of Contig#23.

5’TAATACGACTCACTATAGGG 3’

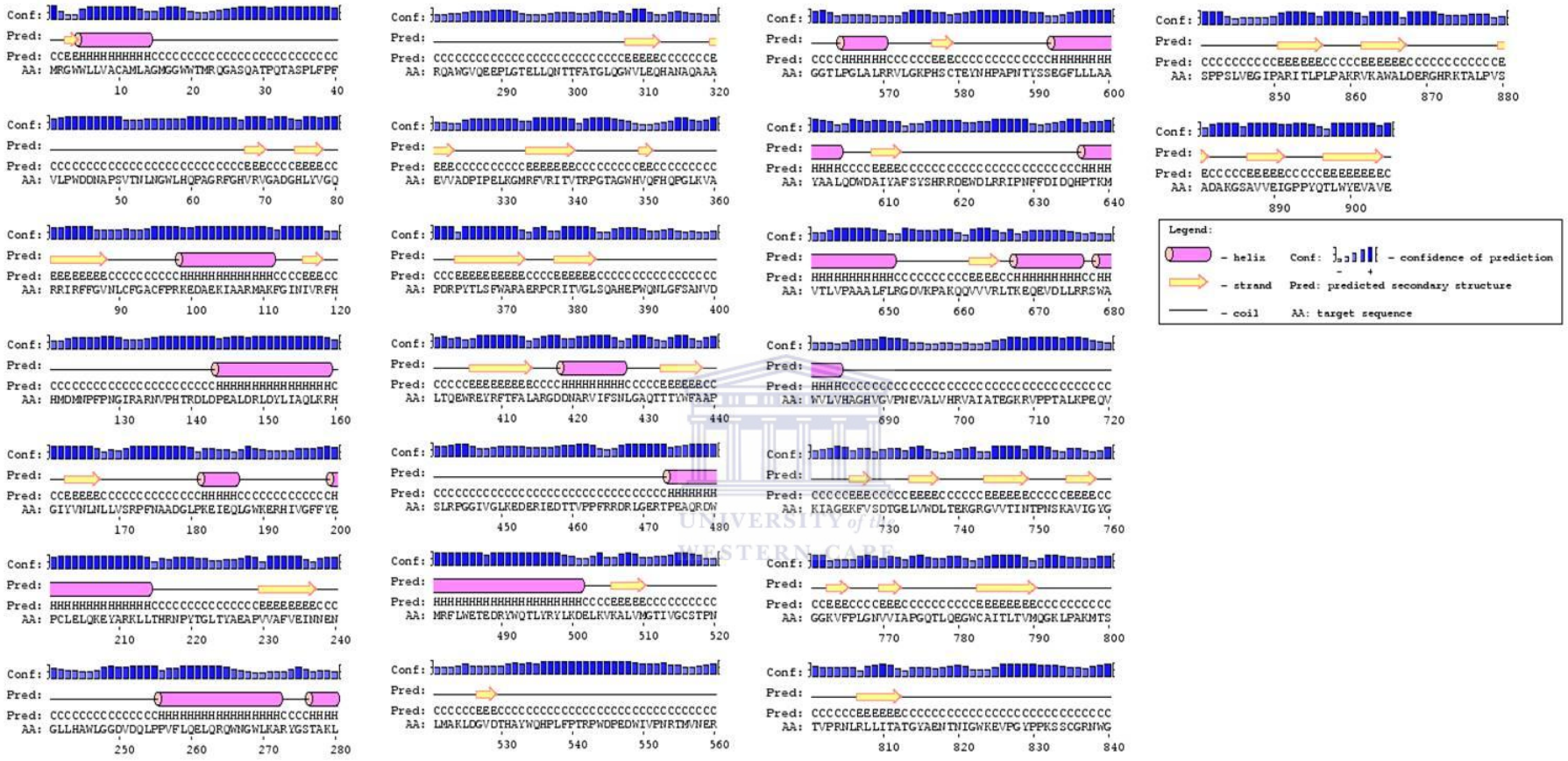
Supplementary Figure 2. Nucleotide sequence of T7 forward.

5' CTCGTATGTTGTGTGGAATTGTGAGC 3'

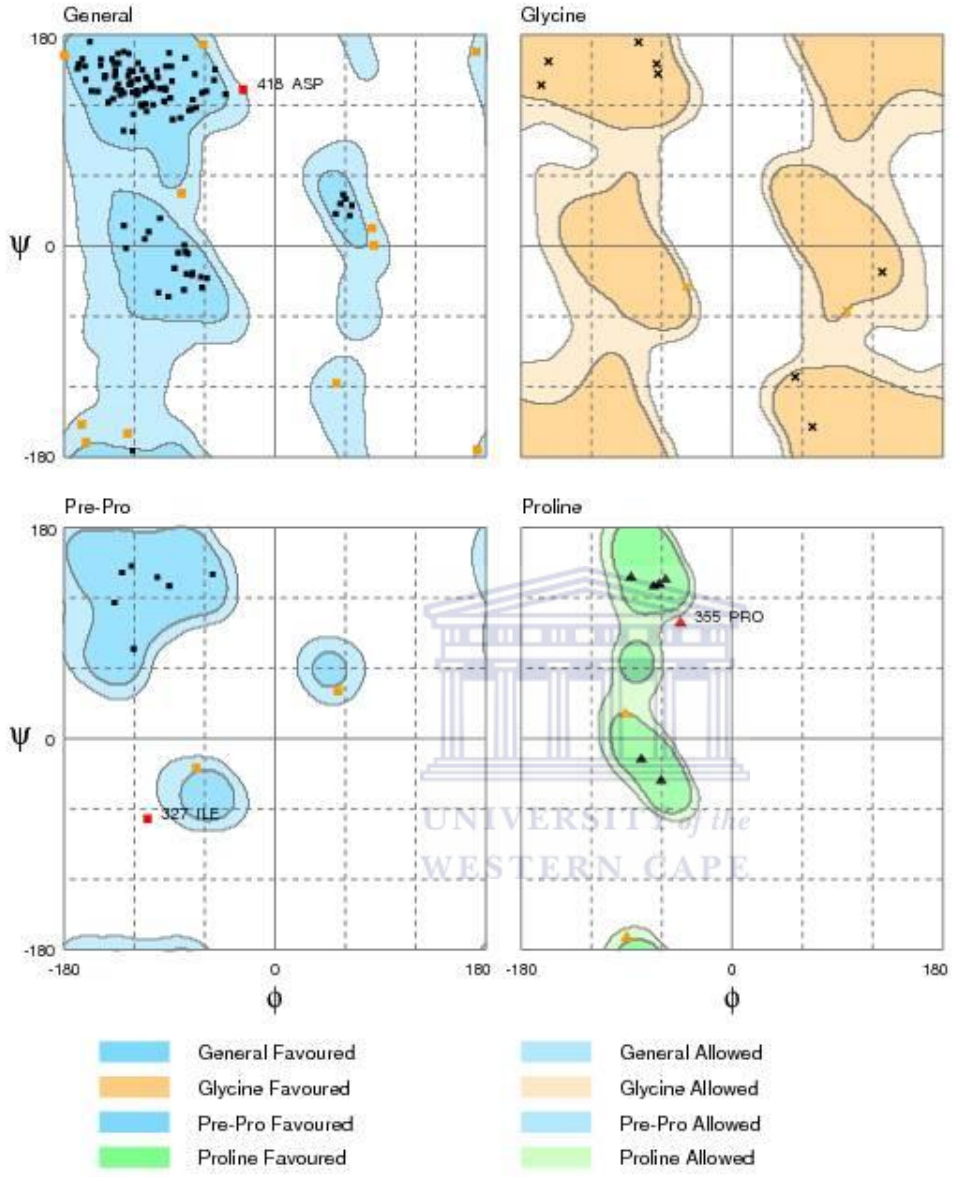
Supplementary Figure 3. Nucleotide sequence of pCC1Fos reverse

Supplementary Table 1 Rare codons and their frequency in the nucleotide sequence obtained for CelMHS predicted by rare codon calculator for expression in *E. coli*.

Amino acid	Rare Codon	Frequency of occurrence
Arginine	CGA	4
	CCG	16
	AGG	1
	AGA	1
Glycine	GGA	10
	GGG	24
Isoleucine	AUA	1
Leucine	CUA	0
Proline	CCC	25
Threonine	ACG	11



Supplementary Figure 4 Secondary structure prediction for the deduced CelMHS amino acid sequence. The figure was generated using PSIPRED (Jones, 1999; McGuffin *et al*., 2000).



Supplementary Figure 5 Ramachandran plot for the CBM of CelMHS CBM built by the Swiss model server using RAMPAGE.

References

- AGBOR, V. B., CICEK, N., SPARLING, R., BERLIN, A. & LEVIN, D. B. 2011. Biomass pretreatment: Fundamentals toward application. *Biotechnology Advances,* 29, 675-685.
- ALCALDE, M. 2007. Laccases: biological functions, molecular structure and industrial applications. In *Industrial Enzymes*. Springer Netherlands.
- ALEKLETT, K. & CAMPBELL, C. J. 2003. The peak and decline of world oil and gas production. *Minerals and Energy-Raw Materials Report,* 18**,** 5-20.
- AMANN RI, LUDWIG W, SCHLEIER KH: 1995. Phylogenetic identification and *in-situ* detection of individual microbial cells without cultivation. *Microbiological Reviews*. 59, 143–169.
- AMANN, R. I., LUDWIG, W., & SCHLEIFER, K. H. 1995. Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *Microbiological reviews*, 59, 143-169.
- ARANTES, V. & SADDLER, J. 2010. Access to cellulose limits the efficiency of enzymatic hydrolysis: the role of amorphogenesis. *Biotechnology for Biofuels,* 3, 4.
- AZZAM, A.M. 1987. Saccharification of bagasse cellulose pretreated with $ZnCl_2$ and HCl. *Biomass Bioenerg.* 12, 71-77.
- BAE, B., OHENE-ADJEI, S., KOCHERGINSKAYA, S., MACKIE, R. I., SPIES, M. A., CANN, I. K., & NAIR, S. K. 2008. Molecular basis for the selectivity and specificity of ligand recognition by the family 16 carbohydrate-binding modules from *Thermoanaerobacterium polysaccharolyticum* ManA. *Journal of Biological Chemistry*, 283, 12415-12425.
- BALAT, M .2007. An overview of biofuels and policies in the European Union countries. *Energy Sources*, 2,167–81
- BALAT, M. & BALAT, H. 2009. Recent trends in global production and utilization of bio-ethanol fuel. *Applied Energy,* 86, 2273-2282.
- BALAT, M. 2011. Production of bioethanol from lignocellulosic materials via the biochemical pathway: a review. *Energy Conversion and Management*, 52, 858-875.
- BALLESTEROS, I., OLIVIA. J., M. NAVARRO, A. A., GONZALEZ, A. CARRASCO, J., & BALLESTEROS, M. 2000. Effect of chip size on steam explosion pretreatment of softwood. *Twenty-First Symposium on Biotechnology for Fuels and Chemicals.* 97-110. Humana Press.

- BANERJEE, G., SCOTT-CRAIG, J. & WALTON, J. 2010. Improving Enzymes for Biomass Conversion: A Basic Research Perspective. *BioEnergy Research,* 3, 82-92.
- BAO, W. & RENGANATHAN, V. 1992. Cellobiose oxidase of *Phanerochaete chrysosporiumenhances* crystalline cellulose degradation by cellulases. *FEBS letters,* 302, 77-80.
- BAYER, E. A., CHANZY, H., LAMED, R., & SHOHAM Y. 1998. Cellulose, cellulases and cellulosomes. *Current opinion in structural biology*, 8, 548-557.
- BAYER, E. A., LAMED, R., & HIMMEL, M. E. 2007. The potential of cellulases and cellulosomes for cellulosic waste management. *Current Opinion in Biotechnology*, 18, 237-245.
- BEESON, W. T, PHILLIPS, C. M, CATE, J. H. D, MARLETTA, M. A: 2012. Oxidative cleavage of cellulose by fungal copper-dependent polysaccharide monooxygenases*. J Am Chem Soc,* 134:890–892
- BELDMAN, G., LEEWUEN, M. F., ROMBOUTS, F. M. & VORAGEN, F. G. J. 2005. The cellulase of Trichoderma viride. *European Journal of Biochemistry,* 146**,** 301-308.
- BERLEMONT, R., DELSAUTE, M., PIPERS, D., D'AMICO, S., FELLER, G., GALLENI, M., & POWER, P. 2009. Insights into bacterial cellulose biosynthesis by functional metagenomics on Antarctic soil samples. *The ISME journal*, 3, 1070-1081.
- BEUKES N., CHAN, H., DOI, R. H., & PLETSCHKE, B. I. 2008. Synergistic associations between *Clostridium cellulovorans* enzymes XynA, ManA and EngE against sugarcane bagasse. *Enzyme and Microbial Technology*, 42, 492-498.
- BEUKES, N., & PLETSCHKE, B. I. 2010. Effect of lime pre-treatment on the synergistic hydrolysis of sugarcane bagasse by hemicellulases. *Bioresource technology*, 101, 4472-4478.
- BEUKES, N., & PLETSCHKE, B. I. 2011. Effect of alkaline pre-treatment on enzyme synergy for efficient hemicellulose hydrolysis in sugarcane bagasse. *Bioresource technology*, 102, 5207-5213.
- BHATNAGAR, A. & SAIN, M. 2005. Processing of Cellulose Nanofiber-reinforced Composites. *Journal of Reinforced Plastics and Composites,* 24, 1259-1268.
- BIELECKI, S., KRYSTYNOWICZ, A., TURKIEWICZ, M. & KALINOWSKA, H. 2005. Bacterial cellulose. *Biopolymers Online*.
- BIELY, P., MISLOVICOVA, D., & TOMAN, R. 1985. Soluble chromogenic substrates for the assay of endo-1, 4-β-xylanases and endo-1, 4-β-glucanases. *Analytical Biochemistry*, 144, 142-146.
- BOK, J. D., GOERS, S. K., & EVELEIGH, D. E. 1994. Cellulase and xylanase systems of Thermotoga neapolitana. In *ACS Symposium Series*. Washington, DC: American Chemical Society.

- BOLAM, D. N., CIRUELA, A., MCQUEEN-MASON, S., SIMPSON, P., WILLIAMSON, M. P., RIXON, J. E., BORASTON, A., HAZLEWOOD, G. P. & GILBERT, H. J. 1998. Pseudomonas cellulose-binding domains mediate their effects by increasing enzyme substrate proximity. *Biochemical Journal,* 331, 775.
- BOLAM, D. N., XIE H., PELL, G., HOGG, D., GALBRAITH, G., HENRISSAT, B., & GILBERT, H. J. 2004. X4 modules represent a new family of carbohydrate-binding modules that display novel properties. *Journal of Biological Chemistry*, 279, 22953-22963.
- BORASTON, A. B., BOLAM, D. N., GILBERT, H. J. & DAVIES, G. J. 2004. Carbohydrate-binding modules: fine-tuning polysaccharide recognition. *Biochemical Journal,* 382, 769.
- BORASTON, A. B., MCLEAN, B. W., GUARNA, M. M., AMANDARON-AKOW, E. & KILBURN, D. G. 2001. A Family 2a Carbohydrate-Binding Module Suitable as an Affinity Tag for Proteins Produced in *Pichia pastoris*. *Protein Expression and Purification,* 21, 417-423.
- BORASTON, A. B., NURIZZO, D., NOTENBOOM, V., DUCROS, V., ROSE, D. R., KILBURN, D. G., & DAVIES, G. J. 2000. Differential oligosaccharide recognition by evolutionarily-related β-1, 4 and β-1, 3 glucan-binding modules. *Journal of molecular biology*, 319, 1143-1156.
- BOYER, M., CHAMBOST, J., MAGNAN, M. & CATTANEO, J. 1984. Carboxymethyl-cellulase from *Erwinia chrysanthemi* I. Production and regulation of extracellular carboxymethyl-cellulase. *Journal of Biotechnology,* 1, 229-239.
- CARAMELO, L., MARTINEZ, M. J., & MARTINEZ, Á. T. 1999. A Search for Ligninolytic Peroxidases in the Fungus *Pleurotus eryngii* Involving α-Keto-γ-Thiomethylbutyric Acid and Lignin Model Dimers. *Applied and environmental microbiology*, 65, 916-922.
- CARRARD, G., KOIVULA, A., SÖDERLUND, H. & BÉGUIN, P. 2000. Cellulose-binding domains promote hydrolysis of different sites on crystalline cellulose. *Proceedings of the National Academy of Sciences,* 97**,** 10342-10347.
- CHAKAR, F. S. & RAGAUSKAS, A. J. 2004. Review of current and future softwood kraft lignin process chemistry. *Industrial Crops and Products,* 20**,** 131-141.
- CHANG, L., DING, M., BAO, L., CHEN, Y., ZHOU, J., & LU, H. 2011. Characterization of a bifunctional xylanase/endoglucanase from yak rumen microorganisms. *Applied microbiology and biotechnology*, 90, 1933-1942.
- CHRISTENSEN, C. H., RASS-HANSEN, J., MARSDEN, C. C., TAARNING, E. & EGEBLAD, K. 2008. The Renewable Chemicals Industry. *ChemSusChem,* 1, 283-289.

- CLARKE, A. J. 1987. Essential tryptophan residues in the function of cellulase from *Schizophyllum commune*. *Biochimica et Biophysica Acta (BBA)-Protein Structure and Molecular Enzymology,* 912, 424-431.
- COWAN, D., MEYER, Q., STAFFORD, W., MUYANGA, S., CAMERRON, R., & WITTER, P. 2005. Metagenomic gene discovery: past, present and future. *TRENDS in Biotechnology*, 23, 321-329.
- DAVIES, G. & HENRISSAT, B. 1995. Structures and mechanisms of glycosyl hydrolases. *Structure,* 3, 853-859.
- DEMIRBAS, A. & KARSLIOGLU S. 2007. Biodiesel production facilities from vegetable oils and animal fats. *Energy Source,* 29,133–41.
- DEMIRBAS, A. 2008. Biofuels sources, biofuel policy, biofuel economy and global biofuel projections. *Energy Conversion and Management,* 49, 2106-2116.
- DEN HAAN, R. ROSE, S. H. LYND, L. R. & VAN ZYL, W. H. 2007. Hydrolysis and fermentation of amorphous cellulose by recombinant *Saccharomyces cerevisiae*. *Metabolic engineering,* 9, 87-94.
- DI LAURO, B., STRAZZULLI, A., PERUGINO, G., LA CARA, F., BEDINI, E., CORSARO, M. M., & MORACCI, M. 2008. Isolation and characterization of a new family 42 β-galactosidase from the thermoacidophilic bacterium *Alicyclobacillus acidocaldarius*: Identification of the active site residues. *Biochimica et Biophysica Acta (BBA)-Proteins & Proteomics*, 1784, 292-301.
- DIEN, B. S., COTTA, M. A., & JEFFRIES, T. W. 2003. Bacteria engineered for fuel ethanol production: current status. *Applied Microbiology and Biotechnology*, 63, 258-266.
- DREDGE, R. RADLOFF, S. E. VAN DYK, J. S. & PLETSCKE, B. I. 2011. Lime pretreatment of sugar beet pulp and evaluation of synergy between ArfA, ManA and XynA from *Clostridium cellulovorans* on the pretreated substrate.*3 Biotech*, 1, 151-159.
- DUAN, C. J., & FENG, J. X. 2010. Mining metagenomes for novel cellulase genes. *Biotechnology letters*, 32, 1765-1775.
- FENG, Y., DUAN, C. J., PANG, H., MO, X. C., WU, C. F., YU, Y., & FENG, J. X. 2007. Cloning and identification of novel cellulase genes from uncultured microorganisms in rabbit cecum and characterization of the expressed cellulases. *Applied microbiology and biotechnology*, 75, 319-328.
- FENGEL, D. & WEGENER, D. (eds.) 1984. *Wood:Chemistry, Ultrastructure, Reactions,* Berlin: De Gruyter.
- FERRER, M., GOLYSHINA, O., BELOQUI, A., & GOLYSHIN, P. N. 2007. Mining enzymes from extreme environments. *Current opinion in microbiology*, 10, 207-214.

- FISER, A., & SALI, A. 2003. ModLoop: automated modeling of loops in protein structures. *Bioinformatics*, 19, 2500-2501.
- FLINT, J., NURIZZO, D., HARDING, S. E., LONGMAN, E., DAVIES, G. J., GILBERT, H. J., & BOLAM, D. N. 2004. Ligand-mediated Dimerization of a Carbohydrate-binding Module Reveals a Novel Mechanism for Protein–Carbohydrate Recognition. *Journal of molecular biology*, 337, 417-426.
- FORSBERG, C. W., BEVERIDGE, T. J. & HELLSTROM, A. 1981. Cellulase and Xylanase Release from Bacteroides succinogenes and Its Importance in the Rumen Environment. *Applied and Environmental Microbiology,* 42, 886-896.
- FRENCH, A. D., BERTONIERE, N. R., BROWN, R. M., CHANZY, H., GRAY, D., HATTORI, K. & GLASSER, W. 2002. Cellulose. *Encyclopedia of Polymer Science and Technology.* John Wiley & Sons, Inc.
- FUKUDA, H., KONDO, A. & TAMALAMPUDI, S. 2009. Bioenergy: Sustainable fuels from biomass by yeast and fungal whole-cell biocatalysts. *Biochemical Engineering Journal,* 44**,** 2-12.
- GARCIA-APARICIO, M. P., BALLESTEROS, M., MANZANARES, P., BALLESTEROS, I., GONZALEZ, A., & NEGRO, M. J. 2007. Xylanase contribution to the efficiency of cellulose enzymatic hydrolysis of barley straw. In *Applied Biochemistry and Biotecnology* . Humana Press.
- GARROTE, G., DOMINGUEZ, H., & PARAJO, J. C. 1999. Hydrothermal processing of lignocellulosic materials. *European Journal of Wood and Wood Products*, 57, 191-202.
- GEBLER, J., GILKES, N. R., CLAEYSSENS, M., WILSON, D. B., BEGUIN, P. WAKARCHUK, W. W., & WITHERS, S. G. 1992. Stereoselective hydrolysis catalyzed by related beta-1, 4-glucanases and beta-1, 4-xylanases. *Journal of Biological Chemistry*, 267, 12559-12561.
- GENG, A., ZOU, G., YAN, X., WANG, Q., ZHANG, J. LIU, F., & ZHOU, Z. 2012. Expression and characterization of a novel metagenome-derived cellulase Exo2b and its application to improve cellulase activity in *Trichoderma reesei*. *Applied microbiology and biotechnology*, 96, 951-962.
- GEVER, J., KAUFMANN, R., SKOLE, D. & VOROSMARTY, C. 1985. Beyond oil: The threat to food and fuel in the coming decades.
- GOLDEMBERG, J. 2006. The promise of a clean energy. *Energy Policy*. 34, 2185-2190
- GOMES, E., AGUAIR, A. P., CARVALHO, C. C., BONFA, M. R. B., Silva, R. D., & Boscolo, M. (2009). Ligninases production by Basidiomycetes strains on lignocellulosic agricultural residues and their application in the decolorization of synthetic dyes. *Brazilian Journal of Microbiology*, 40, 31-39.
- GRAHAM, J. E., CLARK, M. E., NADLER, D. C., HUFFER, S., CHOKHAWALA, H. A., ROWLAND, S. E., & ROBB, F. T. 2011. Identification and characterization

of a multidomain hyperthermophilic cellulase from an archaeal enrichment. *Nature Communications*, 2, 375.

- GRANT, S., SOROKIN, D. Y., GRANT, W. D., JONES, B. E., & HEAPHY, S. 2004. A phylogenetic analysis of Wadi el Natrun soda lake cellulase enrichment cultures and identification of cellulase genes from these cultures. *Extremophiles*, 8, 421-429.
- GRAY, K.A. ZHAO, L. & EMPTAGE, M. 2006. Bioethnaol. *Current Opinion in Chemical Biology*. 10, 141-146.
- GROHMANN, K., TOGET, R. & HIMMEL, M. 1985. Optimisation of dilute acid pre-treatment of biomass. *Biotechnology and Bioengineering Symposium,* 15**,** 59-80.
- GROUS, W. R., CONVERSE, A. O., & GRETHLEIN, H. E. 1986. Effect of steam explosion pretreatment on pore size and enzymatic hydrolysis of poplar. *Enzyme and Microbial Technology*, 8, 274-280.
- GUILLEN, F., MARTINEZ, M. J., & GUTIERREZ, A. 2005. Biodegradation of lignocellu-losics: microbial, chemical, and enzymatic aspects of the fungal attack of lignin. *International Microbiology*, 8, 195-204.
- HACHEM, M. A., KARLSSON, E. N., BARTONEK-ROXA, E., RAGHOTHAMA, S., SIMPSON, P. J., GILBERT, H. J., & HOLST, O. 2000. Carbohydrate-binding modules from a thermostable *Rhodothermus marinus* xylanase: cloning, expression and binding studies. *Biochemical Journal*, 345, 53.
- HAMELINCK, C. N., & FAAIJ, A. P. 2006. Outlook for advanced biofuels. *Energy Policy*, 34, 3268-3283.
- HANDELSMAN, J. , RONDON, M. R., BRADY, S.F., CLARDY, J. & GOODMAN, R.M. 1998. Molecular access to the chemistry of the unknown soil microbes: a new frontier. *Chemistry and Biology*,5, 245-249.
- HANDELSMAN, J. 2004. Metagenomics: application of genomics to uncultured microorganisms. *Microbiology and Molecular Biology Reviews*, 68, 669-685.
- HARCHAND, R. K., & SINGH, S. 1997. Characterization of cellulase complex of *Streptomyces albaduncus*. *Journal of basic microbiology*, 37, 93-103.
- HARDEMAN, F., & SJOLING, S. 2007. Metagenomic approach for the isolation of a novel low-temperature-active lipase from uncultured bacteria of marine sediment. *FEMS microbiology ecology*, 59, 524-534.
- HARUN, R, & Danquah, M. K. 2011. Influence of acid pre-treatment on microalgal biomass for bioethanol production. *Process Biochemistry*. 46, 304-309
- HASAN, F., SHAH, A. A., & HAMEED, A. 2006. Industrial applications of microbial lipases. *Enzyme and Microbial Technology*, 39, 235-251.
- HAUG, R. T. (ed.) 1993. *The practical handbook of composting engineering,* Boca Raton: Lewis publishers.

- HEALY, F. G., RAY, R. M., ALDRICH, H. C., WILKIE, A. C., INGRAM, L. O., & SHANMUGAM, K. T. 1995. Direct isolation of functional genes encoding cellulases from the microbial consortia in a thermophilic, anaerobic digester maintained on lignocellulose. *Applied microbiology and biotechnology*, 43, 667-674.
- HEITZER, M., ECKERT, A., FUHRMANN, M., & GRIESBECK, C. 2007. Influence of codon bias on the expression of foreign genes in microalgae. In *Transgenic Microalgae as Green Cell Factories*. Springer New York.
- HENDRIKS, A. T. W. M. & ZEEMAN, G. 2009. Pretreatments to enhance the digestibility of lignocellulosic biomass. *Bioresource Technology,* 100, 10-18.
- HENRISSAT, B. & DAVIES, G. J. 2000. Glycosyltranferases: Families, modules and implications for genomics. *Plant Physiology,* 124, 1515-1519.
- HENRISSAT, B. 1991. A classification of glycoside hydrolases bases on amino acid sequence similarities. *The Biochemical Journal,* 280**,** 309-316.
- HERVÉ, C., ROGOWSKI, A., BLAKE, A. W., MARCUS, S. E., GILBERT, H. J. & KNOX, J. P. 2010. Carbohydrate-binding modules promote the enzymatic deconstruction of intact plant cell walls by targeting and proximity effects. *Proceedings of the National Academy of Sciences,* 107, 15293-15298.
- HICKE, L., & DUNN, R. 2003. Regulation of membrane protein transport by ubiquitin and ubiquitin-binding proteins. *Annual review of cell and developmental biology*, 19, 141-172.
- HIDAKA, M., FUSHINOBU, S., OHTSU, N., MOTOSHIMA, H., MATSUZAWA, H., SHOUN, H., & WAKAGI, T. 2002. Trimeric Crystal Structure of the Glycoside Hydrolase Family 42 β-Galactosidase from *Thermus thermophilus* A4 and the Structure of its Complex with Galactose. *Journal of molecular biology*, 322, 79-91.
- HILDÉN, L. & JOHANSSON, G. 2004. Recent developments on cellulases and carbohydrate-binding modules with cellulose affinity. *Biotechnology Letters,* 26, 1683-1693.
- HIMMEL MICHAEL, E., ADNEY WILLIAM, S., BAKER JOHN, O., ELANDER, R., MCMILLAN JAMES, D., NIEVES RAFAEL, A., SHEEHAN JOHN, J., THOMAS STEVEN, R., VINZANT TODD, B. & ZHANG, M. 1997. Advanced Bioethanol Production Technologies: A Perspective. *Fuels and Chemicals from Biomass.* American Chemical Society.
- HIMMEL, M. E., DING, S. Y., JOHNSON, D. K., ADNEY, W. S., NIMLOS, M. R., BRADY, J. W., & FOUST, T. D. 2007. Biomass recalcitrance: engineering plants and enzymes for biofuels production. *Science*, 315, 804-807.
- HOHL, H. R., HAMAMOTO, S. T. & DON, E. H. 1968. Ultrastructural Aspects of Cell Elongation, Cellulose Synthesis, and Spore Differentiation in *Acytostelium leptosomum*, a Cellular Slime Mold. *American Journal of Botany,* 55, 783-796.

- HOLMES, M. L., & DYALL-SMITH, M. L. 2000. Sequence and expression of a halobacterial β-galactosidase gene. *Molecular microbiology*, 36, 114-122.
- HORN, S.J.; VAAJE-KOLSTAD, G. WESTERENG, B. and EIJSINT, V.G.H. 2012. Novel enzymes for the degradation of cellulose. *Biotechnology for Biofuels*. 5:45, 1-12
- HOWARD, R. L., ABOTSI, E., JANSEN VAN RENSBURG, E. L. & HOWARD, S. 2003. Lignocellulose biotechnology: issues of bioconversion and enzyme production. *African Journal of Biotechnology,* 12, 602-619.
- HUBER, M. T. 2009. Energizing historical materialism: Fossil fuels, space and the capitalist mode of production. *Geoforum,* 40, 105-115.
- IBRAHIM, A. S. S., & EL-DIWANY, A. I. 2000. Isolation and identification of new cellulases producing thermophilic bacteria from an Egyptian hot spring and some properties of the crude enzyme. *Australian Journal of Basic and Applied Sciences*, 1, 473-478.
- IBRAHIM, A. S. S., & EL-DIWANY, A. I. 2007. Isolation and identification of new cellulases producing thermophilic bacteria from an Egyptian hot spring and some properties of the crude enzyme. *Australian Journal of Basic and Applied Sciences*, 1, 473-478.
- IFTIKHAR, T. E. H. R. E. E. M. A., NIAZ, M. U. B. A. S. H. I. R., ABBAS, S. Q., ZIA, M. A., ASHRAF, I., LEE, K. J., & HAQ, I. U. 2010. Mutation induced enhanced biosynthesis of lipases by rhizopus oligosporus var. microsporus. *Pakistan Journal of Botany*, 42, 1235-1249.
- ILLANES, A. 1999. Stability of biocatalysts. *Electronic Journal of Biotechnology*, *2*, -9.
- ILLMEN, M. DEN HAAN, R., BREVNOVA, E., MCBRIDE, J., WISWALL, E., FROEHLICH, A., & PENTTILA, M. 2011. High level secretion of cellobiohydrolases by *Saccharomyces cerevisiae*. *Biotechnol Biofuels*, 4, 30.
- ILMBERGER, N., MESKE, D., JUERGENSEN, J., SCHULTE, M., BARTHEN, P., RABAUSCH, U., & STREIT, W. R. 2012. Metagenomic cellulases highly tolerant towards the presence of ionic liquids—linking thermostability and halotolerance. *Applied microbiology and biotechnology*, 95, 135-146.
- IRWIN, D. C., SPEZIO, M., WALKER, L. P. & WILSON, D. B. 1993. Activity studies of eight purified cellulases: Specificity, synergism, and binding domain effects. *Biotechnology and Bioengineering,* 42, 1002-1013.
- ITO, S. 1997. Alkaline cellulases from alkaliphilic *Bacillus*: Enzymatic properties, genetics, and application to detergents. *Extremophiles,* 1, 61-66.
- JANG, H. D. & CHEN, K. S. 2003. Production and characterization of thermostable cellulases from Streptomyces transformant T3-1. *World journal of Microbiology and Biotechnology,* 19, 263-268.

- JEON, J. H., KIM, J. T., KANG, S. G., LEE, J. H., & KIM, S. J. 2009. Characterization and its potential application of two esterases derived from the arctic sediment metagenome. *Marine Biotechnology*, 11, 307-316.
- JIANG, C., HAO, Z. Y., JIN, K. E., LI, S. X., CHE, Z. Q., MA, G. F., & WU, B. 2010. Identification of a metagenome-derived β-glucosidase from bioreactor contents. *Journal of Molecular Catalysis B: Enzymatic*, 63, 11-16.
- JIANG, C., MA, G., LI, S., HU, T., CHE, Z., SHEN, P., & WU, B. (2009). Characterization of a novel β-glucosidase-like activity from a soil metagenome. *The Journal of Microbiology*, 47, 542-548.
- JOHNSON, P. E., TOMME, P., JOSHI, M. D., & MCINTOSH, L. P. 1996. Interaction of soluble cellooligosaccharides with the N-terminal cellulose-binding domain of *Cellulomonas fimi* CenC. 2. NMR and ultraviolet absorption spectroscopy. *Biochemistry*, 35, 13895-13906.
- JØRGENSEN, H., KRISTENSEN, J. B. & FELBY, C. 2007. Enzymatic conversion of lignocellulose into fermentable sugars: challenges and opportunities. *Biofuels, Bioproducts and Biorefining,* 1, 119-134.
- JULLANDER, I. 1947. Recent ultracentrifugal investigations on cellulose and cellulose derivatives. *Journal of Polymer Science*, 2, 329-345.
- JULLANDER, I. 1947. Recent ultracentrifugal investigations on cellulose and cellulose derivatives. *Journal of Polymer Science,* 2, 329-345.
- KÁDÁR, Z., SZENGYEL, Z. & RÉCZEY, K. 2004. Simultaneous saccharification and fermentation (SSF) of industrial wastes for the production of ethanol. *Industrial Crops and Products,* 20, 103-110.
- KIM, D., KIM, S. N., BAIK, K. S., PARK, S. C., LIM, C. H., KIM, J. O., & SEONG, C. N. 2011. Screening and characterization of a cellulase gene from the gut microflora of abalone using metagenomic library. *The Journal of Microbiology*, 49, 141-145.
- KIM, S. J., LEE, C. M., HAN, B. R., KIM, M. Y., YEO, Y. S., YOON, S. H., & JUN, H. K. 2008. Characterization of a gene encoding cellulase from uncultured soil bacteria. *FEMS microbiology letters*, 282, 44-51.
- KLEIN-MARCUSCHAMER, D., OLESKOWICZ-POPIEL, P., SIMMONS, B. A., & BLANCH, H. W. 2012. The challenge of enzyme cost in the production of lignocellulosic biofuels. *Biotechnology and Bioengineering*, 109, 1083-1087.
- KONIETZNY, U., & GREINER, R. 2002. Molecular and catalytic properties of phytate-degrading enzymes (phytases). *International journal of food science & technology*, *37*, 791-812.
- KRAULIS, P. J., CLORE, G. M., NILGES, M., JONES, T. A., PETTERSSON, G., KNOWLES, J., & GRONENBORN, A. M. 1989. Determination of the three-dimensional solution structure of the C-terminal domain of cellobiohydrolase I

from *Trichoderma reesei*. A study using nuclear magnetic resonance and hybrid distance geometry-dynamical simulated annealing. *Biochemistry*, 28, 7241-7257.

- KUMAR, P., BARRETT, D. M., DELWICHE, M. J. & STROEVE, P. 2009. Methods for Pretreatment of Lignocellulosic Biomass for Efficient Hydrolysis and Biofuel Production. *Industrial & Engineering Chemistry Research,* 48, 3713-3729.
- LAUREANO-PEREZ, L., TEYMOURI, F., ALIZADEH, H. & DALE, B. 2005. Understanding factors that limit enzymatic hydrolysis of biomass. *Applied Biochemistry and Biotechnology,* 124, 1081-1099.
- LEE, J. H., KIM, Y. S., YEOM, S. J., & OH, D. K. 2011. Characterization of a glycoside hydrolase family 42 β-galactosidase from *Deinococcus geothermalis*. *Biotechnology letters*, 33, 577-583.
- LEVY, I. & SHOSEYOV, O. 2002. Cellulose-binding domains: Biotechnological applications. *Biotechnology Advances,* 20**,** 191-213.
- LI, H. KIM, N. J. JIANG, M. KANG, J. W. & ChANG, H. N. 2009. Simultaneous saccharification and fermentation of lignocellulosic residues pretreated with phosphoric acid–acetone for bioethanol production. *Bioresource technology*. 100, 3245-3251.
- LIANG, Y., FENG, Z., YESUF, J., & BLACKBURN, J. W. 2010. Optimization of growth medium and enzyme assay conditions for crude cellulases produced by a novel thermophilic and cellulolytic bacterium, *Anoxybacillus* sp. 527. *Applied biochemistry and biotechnology*, 160, 1841-1852.
- LIN, Y. and TANAKA, S. 2006. Ethanol fermentation from biomass resources: current state and prospects. *Applied Microbiology and Biotechnology*. 69,627-642
- LINDER, M., TEERI, T. T. 1997. The role and function of cellulose binding domains. Journal of Biotechnology. 57, 15-28
- LIU, L., FENG, Y., DUAN, C. J., PANG, H., TANG, J. L., & FENG, J. X. 2009. Isolation of a gene encoding endoglucanase activity from uncultured microorganisms in buffalo rumen. *World Journal of Microbiology and Biotechnology*, 25, 1035-1042.
- LO, Y. C., SARATALE, G. D., CHEN, W. M., BAI, M. D. & CHANG, J. S. 2009. Isolation of cellulose-hydrolytic bacteria and applications of the cellulolytic enzymes for cellulosic biohydrogen production. *Enzyme and Microbial Technology,* 44, 417-425.
- LOPEZ-LOPEZ, O., CERDAN, M. E., & GONZALEZ-SISO, M. I. 2013. Hot Spring Metagenomics. *Life*, 3, 308-320.
- LORENZ, P., LIEETON, K., NIEHAUS, F., & ECK, J. 2002. Screening for novel enzymes for biocatalytic processes: accessing the metagenome as a resource of novel functional sequence space. *Current opinion in biotechnology*, 13, 572-577.

- MA, F. & HANNA, M. A. 1999. Biodiesel production: a review. *Bioresource Technology,* 70**,** 1-15.
- MACIEL, M. J. M., & RIBEIRO, H. C. T. 2010. Industrial and biotechnological applications of ligninolytic enzymes of the *basidiomycota*: A review. *Electronic Journal of Biotechnology*, 13, 14-15.
- MADSON, P. W., & LOCOCO, D. B. 2000. Recovery of volatile products from dilute high-fouling process streams. *Applied biochemistry and biotechnology*, 84, 1049-1061.
- MAKI, M., LEUNG, K. T., & QIN, W. 2009. The prospects of cellulase-producing bacteria for the bioconversion of lignocellulosic biomass. *International Journal of Biological Sciences*, 5, 500.
- MANSFIELD, S. D., MOONEY, C. & SADDLER, J. N. 1999. Substrate and Enzyme Characteristics that Limit Cellulose Hydrolysis. *Biotechnology Progress,* 15, 804-816.
- MARGEOT, A., HAHN-HAGERDAL, B., EDLUND, M., SLADE, R. & MONOT, F. 2009a. New improvements for lignocellulosic ethanol. *Current Opinion in Biotechnology,* 20, 372-380.
- MARGEOT, A., HAHN-HAGERDAL, B., EDLUND, M., SLADE, R. & MONOT, F. 2009b. New improvements for lignocellulosic ethanol. *Current opinion in biotechnology,* 20, 372-80.
- MAWADZA, C., HATTI-KAUL, R., ZVAUYA, R., & MATTIASSON, B. 2000. Purification and characterization of cellulases produced by two Bacillus strains. *Journal of Biotechnology*, 83, 177-187.
- MCMILLAN, J. D., JENNINGS, E. W., MOHAGHEGHI, A. & ZUCCARELLO, M. 2011. Comparative performance of precommercial cellulases hydrolyzing pretreated corn stover. *Biotechnology for biofuels,* 4, 1-17.
- MEINKE, A., GILKES, N. R., KILBURN, D. G., MILLER, R. C. & WARREN, R. A. 1991. Multiple domains in endoglucanase B (CenB) from *Cellulomonas fimi*: functions and relatedness to domains in other polypeptides. *Journal of Bacteriology,* 173, 7126-7135.
- MERINO, S. & CHERRY, J. 2007. Progress and Challenges in Enzyme Development for Biomass Utilization Biofuels. *In:* OLSSON, L. (ed.). Springer Berlin / Heidelberg.
- MILLER, E. M., & NICKOLOFF, J. A. 1995. Escherichia coli electrotransformation. In *Electroporation Protocols for Microorganisms*. Humana Press.
- MILLER, G. L. 1959. Modified DNS method for reducing sugars. *Analytical Chemistry*, 31, 426-428.
- MOHAGHEGHI, A., GROHMANN, K. M. M. H., HIMMEL, M., LEIGHTON, L., & UPDEGRAFF, D. M. 1986. Isolation and characterization of *Acidothermus*

cellulolyticus gen. nov., sp. nov., a new genus of thermophilic, acidophilic, cellulolytic bacteria. *International journal of systematic bacteriology*, 36, 435-443.

- MOHAN, D., PITTMAN, C. U. & STEELE, P. H. 2006. Pyrolysis of Wood/Biomass for Bio-oil: A Critical Review. *Energy & Fuels,* 20**,** 848-889.
- MORA, C., TITTENSOR, D. P., ADL, S., SIMPSON, A. G., & WORM, B. 2011. How many species are there on Earth and in the ocean?. *PLoS biology*, 9, e1001127.
- NACKE, H., ENGELHAUPT, M., BRADY, S., FISCHER, C., TAUTZT, J., & DANIEL, R. 2012. Identification and characterization of novel cellulolytic and hemicellulolytic genes and enzymes derived from German grassland soil metagenomes. *Biotechnology letters*, 34, 663-675.
- NAIK, S. N., GOUD, V. V., ROUT, P. K. & DALAI, A. K. 2010. Production of first and second generation biofuels: A comprehensive review. *Renewable and Sustainable Energy Reviews,* 14, 578-597.
- NAKAGAWA,H. HARADA, T. ICHINOSE, T. TAKENO, K. MATSUMOTO, S. KOBAYASHI, M. & SAKAI, M. 2007. Biomethnaol production and CO_2 emmision reduction from forage grasses, trees and crop reidues. *Japan Agricultural Research Quarterly*. 41, 173-180
- NG, T. K., WEIMER, P. J., & ZEIKUS, J. G. 1977. Cellulolytic and physiological properties of Clostridium thermocellum. *Archives of microbiology*, 114, 1-7.
- NGUYEN, N. H., MARUSET, L., UENGWETWANIT, T., MHUANTONG, W., HARNPICHARNCHAI, P., CHAMPREDA, V., & PONGPATTANAKITSHOTE, S. 2012. Identification and characterization of a cellulase-encoding gene from the buffalo rumen metagenomic library. *Bioscience, biotechnology, and biochemistry*, 76, 1075-1084.
- NIEDERBERGER, T. D., GOTZ, D. K., MCDONALD, I. R., RONIMUS, R. S., & MORGAN, H. W. 2006. *Ignisphaera aggregans* gen. nov., sp. nov., a novel hyperthermophilic crenarchaeote isolated from hot springs in Rotorua and Tokaanu, New Zealand. *International journal of systematic and evolutionary microbiology*, 56, 965-971.
- NOTENBOOM, V., BORASTON, A. B., CHUI, P., FREELOVE, A. C., KILBURN, D. G., & ROSE, D. R. 2001. Recognition of cello-oligosaccharides by a family 17 carbohydrate-binding module: an X-ray crystallographic, thermodynamic and mutagenic study. *Journal of molecular biology*, 314, 797-806.
- PANDEY, A., SELVAKUMAR, P., SOCCOL, C. R., & NIGAM, P. 1999. Solid state fermentation for the production of industrial enzymes. *Current science*, 77, 149-162.
- PANG, H., ZHANG, P., DUAN, C. J., MO, X. C., TANG, J. L., & FENG, J. X. 2009. Identification of cellulase genes from the metagenomes of compost soils

and functional characterization of one novel endoglucanase. *Current microbiology*, 58, 404-408.

- PANG, H., ZHANG, P., DUAN, C. J., MO, X. C., TANG, J. L., & FENG, J. X. 2009. Identification of cellulase genes from the metagenomes of compost soils and functional characterization of one novel endoglucanase. *Current microbiology*, 58, 404-408.
- PELL, G., WILLIAMSON, M. P., WALTERS, C., DU, H., GILBERT, H. J., & BOLAM, D. N. 2003. Importance of hydrophobic and polar residues in ligand binding in the family 15 carbohydrate-binding module from *Cellvibrio japonicus* Xyn10C. *Biochemistry*, 42, 9316-9323.
- PENG, F., REN, J.-L., XU, F., BIAN, J., PENG, P. & SUN, R.-C. 2009. Comparative Study of Hemicelluloses Obtained by Graded Ethanol Precipitation from Sugarcane Bagasse. *Journal of Agricultural and Food Chemistry,* 57, 6305-6317.
- PERCIVAL ZHANG, Y. H., HIMMEL, M. E., & MIELENZ, J. R. 2006. Outlook for cellulase improvement: screening and selection strategies. *Biotechnology advances*, 24, 452-481.
- POZZO, T., PASTEN, J. L., KARLSSON, E. N. & LOGAN, D. T. 2010. Structural and Functional Analyses of β-Glucosidase 3B from Thermotoga neapolitana: A Thermostable Three-Domain Representative of Glycoside Hydrolase 3. *Journal of Molecular Biology,* 397, 724-739.
- QUINLAN, R.J., SWEENEY, M.D., LO LEGGIO, L.,OTTEN, H., PAULSEN, J. C. N., JOHANESON, K. S. KROG, K. JORGENSEN, C. I. TORBORG, M., ANTHONSEN, A. 2011. Insights into the oxidative degradtion of cellulase by a copper metalloenzyme that expolits biomass components. *Proceedings of the National Academy of Sciences.* 108:15079-15084
- RAGHOTHAMA, S., SIMPSON, P. J., SZABÓ, L., NAGY, T., GILBERT, H. J. & WILLIAMSON, M. P. 2000. Solution Structure of the CBM10 Cellulose Binding Module from Pseudomonas Xylanase A†,‡. *Biochemistry,* 39, 978-984.
- RASHAMUSE, K., MABIZELA-MOKOENA, N., SANYIKA, T. W*., MABVAKURE, B. & BRADY, D. 2012a. Accessing Carb*oxylesterase Diversity from Termite Hindgut Symbionts through Metagenomics. *Journal of Molecular Microbiology and Biotechnology,* 22, 277-286.
- RASHAMUSE, K., MAGOMANI, V., RONNEBURG, T. & BRADY, D. 2009a. A novel family VIII carboxylesterase derived from a leachate metagenome library exhibits promiscuous β-lactamase activity on nitrocefin. *Applied microbiology and biotechnology,* 83, 491-500.

- RASHAMUSE, K., RONNEBURG, T., HENNESSY, F., VISSER, D., VAN HEERDEN, E., PIATER, L., LITTHAUER, D., MÖLLER, C. & BRADY, D. 2009b. Discovery of a novel carboxylesterase through functional screening of a pre-enriched environmental library. *Journal of applied microbiology,* 106, 1532-1539.
- RASHAMUSE, K., SANYIKA, W., RONNEBURG, T. & BRADY, D. 2012b. A feruloyl esterase derived from a leachate metagenome library.
- RASHAMUSE, K., VISSER, D., HENNESSY, F., KEMP, J., ROUX-VAN DER MERWE, M., BADENHORST, J., RONNEBURG, T., FRANCIS-POPE, R. & BRADY, D. 2012c. Characterisation of Two Bifunctional Cellulase–Xylanase Enzymes Isolated from a Bovine Rumen Metagenome Library. *Current Microbiology*, 1-7.
- REES, H. C., GRANT, S., JONES, B., GRANT, W. D., & HEAPHY, S. 2003. Detecting cellulase and esterase enzyme activities encoded by novel genes present in environmental DNA libraries. *Extremophiles*, 7, 415-421.
- REILLY, J. A. 1991. Celluloid objects: Their chemistry and preservation. *Journal of the American Institute of Conservation,* 30, 145-162.
- REN, N. Q., GUO, W. Q., WANG, X. J., XIANG, W. S., LIU, B. F., WANG, X. Z., & CHEN, Z. B. 2008. Effects of different pretreatment methods on fermentation types and dominant bacteria for hydrogen production. *International journal of hydrogen energy*, *33*, 4318-4324.
- RHEE, J. K., AHN, D. G., KIM, Y. G., & OH, J. W. 2005. New thermophilic and thermostable esterase with sequence similarity to the hormone-sensitive lipase family, cloned from a metagenomic library. *Applied and environmental microbiology*, 71, 817-825.
- RIESENFELD, C. S., SCHLOSS, P. D., & HANDELSMAN, J. 2004. Metagenomics: genomic analysis of microbial communities. *Annual Review: Genetics*, 38, 525-552.
- RUIZ, E., CARA, C., MANAZANARES, P., BALLESTEROS, M., & CASTRO, E. 2008. Evaluation of steam explosion pre-treatment for enzymatic hydrolysis of sunflower stalks. *Enzyme and Microbial technology*, 42, 160-166.
- RUYSENAAR, S. 2011. Rethinking the Food-versus-Fuel Debate: An Appraisal of International Perspectives and Implications for the South African Industrial Biofuels Strategy. *Environment and Society: Advances in Research,* 2, 124-148.
- SAKON, J., IRWIN, D., WILSON, D. B., & KARPLUS, P. A. 1997. Structure and mechanism of endo/exocellulase E4 from Thermomonospora fusca. *Nature Structural & Molecular Biology*, 4, 810-818.
- SÁNCHEZ, C. 2009. Lignocellulosic residues: Biodegradation and bioconversion by fungi. *Biotechnology Advances,* 27, 185-194.

- SANGWAN, P., CHEN, X., HUGENHOLTZ, P., & JANSSEN, P. H. 2004. *Chthoniobacter flavus* gen. nov., sp. nov., the first pure-culture representative of subdivision two, Spartobacteria classis nov., of the phylum *Verrucomicrobia*. *Applied and environmental microbiology*, 70, 5875-5881.
- SANNIGRAHI, P., PU, Y. & RAGAUSKAS, A. 2010. Cellulosic biorefineries—unleashing lignin opportunities. *Current Opinion in Environmental Sustainability,* 2, 383-393.
- SANNIGRAHI, P., RAGAUSKAS, A. J., & TUSKAN, G. A. 2010. Poplar as a feedstock for biofuels: a review of compositional characteristics. *Biofuels, Bioproducts and Biorefining*, 4, 209-226.
- SARATALE.G., CHEN.S, Y, L. & J, C. 2008. Outlook of biohydrogen produciton from lignocellulosic feedstock using dark fermentation - a review. *Journal of Scientific and Industrial Reasearch,* 67, 962-979.
- SAUL, D. J., WILLIAMS, L. C., GRAYLING, R. A., CHAMLEY, L. W., LOVE, D. R., & BERGUUIST, P. L. 1990. celB, a gene coding for a bifunctional cellulase from the extreme thermophile *Caldocellum saccharolyti*cum. *Applied and environmental microbiology*, 56, 3117-3124.
- SCHELLER, H. V. & ULVSKOV, P. 2010. Hemicelluloses. *Annual Review of Plant Biology,* 61, 263-289.
- SCHLOSS, P. D., & HANDELSMAN, J. 2003. Biotechnological prospects from metagenomics. *Current opinion in biotechnology*, 14, 303-310.
- SCHLOSS, P. D., & HANDELSMAN, J. 2003. Biotechnological prospects from metagenomics. *Current opinion in biotechnology*, 14, 303-310.
- SCHMEISSER, C., STEELE, H. & STREIT, W. 2007. Metagenomics, biotechnology with non-culturable microbes. *Applied Microbiology and Biotechnology,* 75, 955-962.
- SHAFIEE, S. & TOPAL, E. 2009. When will fossil fuel reserves be diminished? *Energy Policy,* 37, 181-189.
- SHAIKH, F. A., MULLEGGER, J., HE, S., & WITHERS, S. G. 2007. Identification of the catalytic nucleophile in Family 42 β-galactosidases by intermediate trapping and peptide mapping: YesZ from *Bacillus subtilis*. *FEBS letters*, 581, 2441-2446.
- SHALLOM, D., & SHOHAM, Y. 2003. Microbial hemicellulases. *Current opinion in microbiology*, 6, 219-228.
- SHERIDAN, P. P., & BRENCHLEY, J. E. 2000. Characterization of a Salt-Tolerant Family 42 β-Galactosidase from a Psychrophilic Antarctic *Planococcus* Isolate. *Applied and environmental microbiology*, 66, 2438-2444.
- SHOSEYOV, O., SHANI, Z. & LEVY, I. 2006. Carbohydrate Binding Modules: Biochemical Properties and Novel Applications. *Microbiology and Molecular Biology Reviews,* 70**,** 283-295.

- SILVERSTEIN, R. A., CHEN, Y., SHARMA-SHIVAPPA, R. R., BOYETTE, M. D., & OSBORNE, J. 2007. A comparison of chemical pretreatment methods for improving saccharification of cotton stalks. *Bioresource Technology*, 98, 3000-3011.
- SIMON, C., & DANIEL, R. 2009. Achievements and new knowledge unraveled by metagenomic approaches. *Applied microbiology and biotechnology*, 85, 265-276.
- SIMON, C., & DANIEL, R. 2011. Metagenomic analyses: past and future trends. *Applied and environmental microbiology*, 77, 1153-1161.
- SIMON, C., HERATH, J., ROCKSTROH, S., & DANIEL, R. 2009. Rapid identification of genes encoding DNA polymerases by function-based screening of metagenomic libraries derived from glacial ice. *Applied and environmental microbiology*, 75, 2964-2968.
- SIMON, R., PRIEFER, U., & PULHER, A. 1983. A broad host range mobilization system for in vivo genetic engineering: transposon mutagenesis in gram negative bacteria. *Nature Biotechnology*, 1, 784-791.
- SLEAT, R., MAH, R. A. & ROBINSON, R. 1984. Isolation and characterization of an anaerobic, cellulolytic bacterium, Clostridium cellulovorans sp. nov. *Applied and environmental microbiology,* 48, 88-93.
- SPIRIDONOV, N. A. & WILSON, D. B. 1998. Regulation of biosynthesis of individual cellulases in *Thermomonospora fusca*. *Journal of Bacteriology,* 180*,* 3529-3532.
- SRINIVAS, S., MALIK, R. K. & MAHAJANI, S. M. 2007. Fischer-Tropsch synthesis using bio-syngas and CO2. *Energy for Sustainable Development,* 11**,** 66-71.
- SRISODSUK, M., LEHTIO, J., LINDER, M., MARGOLLES-CLARK, E., REINIKAINEN, T., & TEERI, T. T. 1997. *Trichoderma reesei* cellobiohydrolase I with an endoglucanase cellulose-binding domain: action on bacterial microcrystalline cellulose. *Journal of biotechnology*, 57, 49-57.
- STAMM, A. J. 1977. Monomolecular adsorption and crystallite diameters of cellulose from structural and adsorption considerations. *Wood Science and Technology,* 11, 39-49.
- STAMM, W. P., & PHILLIPS, E. A. 1977. A cellulose acetate membrane precipitin (CAP) test for amoebiasis. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 71, 490-492.
- STEELE, H. L., & STREIT, W. R. 2005. Metagenomics: advances in ecology and biotechnology. *FEMS microbiology letters*, 247, 105-111.
- STEELE, H. L., JAEGER, K. E., DANIEL, R., & STREIT, W. R. 2008. Advances in recovery of novel biocatalysts from metagenomes. *Journal of molecular microbiology and biotechnology*, 16, 25-37.

- STEWART, B. J. & LEATHERWOOD, J. 1976. Derepressed synthesis of cellulase by Cellulomonas. *Journal of bacteriology,* 128, 609-615.
- STRIET, W. R., & SCHMITZ, R. A. 2004. Metagenomics–the key to the uncultured microbes. *Current opinion in microbiology*, 7, 492-498.
- SUKHARNIKOV, L. O., CANTWELL, B. J., PODAR, M., & ZHULIN, I. B. 2011. Cellulases: ambiguous non homologous enzymes in a genomic perspective. *Trends in biotechnology*, 29, 473-479.
- SUN, Y. & CHENG, J. 2002. Hydrolysis of lignocellulosic materials for ethanol production: a review. *Bioresource Technology,* 83, 1-11.
- TAHERZADEH, M. & KARIMI, K. 2007. Enzyme-based hydrolysis processes for ethanol from lignocellulosic materials- A review. *Bioresources*. 2, 707-738
- TAYLOR, M. P., ELEY, K. L., MARTIN, S., TUFFIN, M. I. BURTON, S. G., & COWAN, D. A. 2009. Thermophilic ethanologenesis: future prospects for second-generation bioethanol production. *Trends in biotechnology*, 27, 398-405.
- TAYLOR, M. P., ELEY, K. L., MARTIN, S., TUFFIN, M. I., BURTON, S. G. & COWAN, D. A. 2009. Thermophilic ethanologenesis: future prospects for second-generation bioethanol production. *Trends in Biotechnology,* 27, 398-405.
- TEATHER, R. M., & WOOD, P. J. 1982. Use of Congo red-polysaccharide interactions in enumeration and characterization of cellulolytic bacteria from the bovine rumen. *Applied and environmental microbiology*, 43, 777-78
- TEERI, T. T., LEHTOVAARA, P., KAUPPINEN, S., SALOVUORI, I. & KNOWLES, J. 1987. Homologous domains in Trichoderma reesei cellulolytic enzymes: Gene sequence and expression of cellobiohydrolase II. *Gene,* 51, 43-52.
- TEIXEREIRA, L. C., LINDEN, J. C., & SCHROEDER, H. A. 1999. Optimizing peracetic acid pretreatment conditions for improved simultaneous saccharification and co-fermentation (SSCF) of sugar cane bagasse to ethanol fuel. *Renewable Energy*, 16, 1070-1073.
- TETER, S. A. 2012. DECREASE Final Report: Development of a Commercial Ready Enzyme Application System for Ethanol. http://www.osti.gov/bridge/servlets/purl/1039767/1039767.pdf [accessed 30/08/2013].
- TIEN, M., KERSTEN, P. J., & KIRK, T. K. 1987. Selection and improvement of lignin-degrading microorganisms: potential strategy based on lignin model-amino acid adducts. *Applied and environmental microbiology*, 53, 242-245.
- TIJMENSEN, M. J. A., FAAIJ, A. P. C., HAMELINCK, C. N. & VAN HARDEVELD, M. R. M. 2002. Exploration of the possibilities for production of Fischer Tropsch liquids and power via biomass gasification. *Biomass and Bioenergy,* 23, 129-152.
- TIRAWONGSAROJ, P., SRIPRANG, R., HARNPICHARNCHAI, P., THONGARAM, T., CHAMPREDA, V., TANAPONGPIPATt, S., &

EURWILAICHITR, L. 2008. Novel thermophilic and thermostable lipolytic enzymes from a Thailand hot spring metagenomic library. *Journal of Biotechnology*, 133, 42-49.

- TOMME, P., CREAGH, A. L., KILBURN, D. G., & HAYNES, C. A. 1996. Interaction of polysaccharides with the N-terminal cellulose-binding domain of *Cellulomonas fimi* CenC. 1. Binding specificity and calorimetric analysis. *Biochemistry*, 35, 13885-13894.
- TOMME, P., KWAN, E., GILKES, N. R., KILBURN, D. G., & WARREN, R. A. 1996. Characterization of CenC, an enzyme from *Cellulomonas fimi* with both endo-and exoglucanase activities. *Journal of bacteriology*, 178, 4216-4223.
- TORSVIK, V., & OVREAS, L. 2002. Microbial diversity and function in soil: from genes to ecosystems.*Current opinion in microbiology*, 5, 240-245.
- TURNER, P., MAMO, G., & KARLSSON, E. N. 2007. Potential and utilization of thermophiles and thermostable enzymes in biorefining. *Microbial Cellular Factories*, 6, 1-23.
- TYSON, G. W., LO, I., BAKER, B. J., ALLEN, E. E., HUGENHOLTZ, P., & BANFIELD, J. F. 2005. Genome-directed isolation of the key nitrogen fixer *Leptospirillum ferrodiazotrophum* sp. nov. from an acidophilic microbial community. *Applied and Environmental Microbiology*, 71, 6319-6324.
- VAN DYK, J. S., SAKKA, M. SAKKA, K. & PLETSCHKE, B. I. 2009. The cellulolytic and hemi-cellulolytic system of *Bacillus licheniformis* SVD1 and the evidence for production of a large multi-enzyme complex.*Enzyme and Microbial Technology*, 45, 372-378.
- VAN DYK, J. S., SAKKA, M., SAKKA, K. & PLETSCHKE, B. I. 2010. Identification of endoglucanases, xylanases, pectinases and mannanases in the multi-enzyme complex of *Bacillus licheniformis* SVD1. *Enzyme and Microbial Technology*, 47, 112-118.
- VAN ZYL, W. H., ROSE, S. H., TROLLOPE, K. & GORGENS, J. F. 2010. Fungal β-mannanases: mannan hydrolysis, heterologous production and biotechnological applications. *Process Biochemistry,* 45, 1203-1213.
- VANDERHART, D. L. & ATALLA, R. H. 1984. Studies of microstructure in native celluloses using solid-state carbon-13 NMR. *Macromolecules,* 17, 1465-1472.
- VIIKARI, L., ALAPURANEN, M. PURANEN, T. VEHMAANPERA, J., & SIIKA-AHO, M. 2007. Thermostable enzymes in lignocellulose hydrolysis. *Biofuels*. Springer Berlin Heidelberg.
- VIIKARI, L., ALAPURANEN, M., PURANEN, T., VEHMAANPERÄ, J. & SIIKA-AHO, M. 2007. Thermostable Enzymes in Lignocellulose Hydrolysis Biofuels. *In:* OLSSON, L. (ed.). Springer Berlin / Heidelberg.

- VOGET, S., STEELE, H. L., & STREIT, W. R. 2006. Characterization of a metagenome-derived halotolerant cellulase. *Journal of biotechnology*, 126, 26-36.
- VRSANSKÁ, M. & BIELY, P. 1992. The cellobiohydrolase I from Trichoderma reesei QM 9414: action on cello-oligosaccharides. *Carbohydrate Research,* 227, 19-27.
- WALKER, J. 2006. Basic wood chemistry and cell wall ultrastructure Primary Wood Processing. Springer Netherlands.
- WANG, F., LI, F., CHEN, G., & LIU, W. 2009. Isolation and characterization of novel cellulase genes from uncultured microorganisms in different environmental niches. *Microbiological research*, 164, 650-657.
- WANG, G., MENG, K., LUO, H., WANG, Y., HUANG, H., SHI, P., & YAO, B. 2012. Phylogenetic Diversity and Environment-Specific Distributions of Glycosyl Hydrolase Family 10 Xylanases in Geographically Distant Soils. *PloS one*, 7, e43480.
- WARNECKE, F., LUGINBUHL, P., IVANOVA, N., GHASSEMIAN, M., RICHARDSON, T. H., STEGE, J. T., & LEADBETTER, J. R. 2007. Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite. *Nature*, 450, 560-565.
- WASCHKOWITZ, T., ROCKSTROH, S., & DANIEL, R. 2009. Isolation and characterization of metalloproteases with a novel domain structure by construction and screening of metagenomic libraries. *Applied and environmental microbiology*, 75, 2506-2516.
- WATANABE, H. & TOKUDA, G. 2010. Cellulolytic Systems in Insects. *Annual Review of Entomology,* 55, 609-632.
- WHITMAN, W. B. COLEMAN, D .C. WIEBE, W.J. 1998. Prokaryotes: The unseen majority. *Proceedings of the National Academy of Sciences of the United States of America*, 95, 6578-6583.
- WHITMAN, W. B., COLEMAN, D. C., & WIEBE, W. J. 1998. Prokaryotes: the unseen majority. *Proceedings of the National Academy of Sciences*, 95, 6578-6583.
- WILSON, D. B. 1992. Biochemistry and genetics of actinomycete cellulases. *Critical reviews in biotechnology,* 12, 45-63.
- WILSON, D. B. 2009. Cellulases and biofuels. *Current Opinion in Biotechnology*, 20, 295-299.
- WOOD, T. & MCCRAE, S. I. 1986. The cellulase of Penicillium pinophilum. Synergism between enzyme components in solubilizing cellulose with special reference to the involvement of two immunologically distinct cellobiohydrolases. *Biochemical Journal*, 243, 98.

- WOOD, T.1971.The cellulase of Fusarium solani. Purification and specificity of the β-(1-4)-glucanase and the β-d-glucosidase componenets. *Biochemical Journal,* 121, 353.
- WU, Z. & LEE, Y. 1998. Nonisothermal simultaneous saccharification and fermentation for direct conversion of lignocellulosic biomass to ethanol. *Applied biochemistry and biotechnology*, 70, 497-492.
- WYMAN, C. E., SPINDLER, D. D., & GROHMANN, K. 1992. Simultaneous saccharification and fermentation of several lignocellulosic feedstocks to fuel ethanol. *Biomass and Bioenergy*, 3, 301-307.
- YANG, J. C., MADUPU, R., DURKIN, A. S., EKBORG, N. A., PEDAMALLU, C. S., HOLSTETLER, J. B., & DISTEL, D. L. 2009. The complete genome of *Teredinibacter turnerae* T7901: an intracellular endosymbiont of marine wood-boring bivalves (shipworms). *PLoS One*, 4, e6085.
- YOOSAIN, S. & SORAPIPTANA, C. 2007. A study of ethanol production cost for gasoline substitution in Thailand and its competitiveness. *Thammasat International Journal of Science and Technology*, 12, 69-80.
- ZHANG, P. Y. H., HIMMEL, M. E. & MIELENZ, R. 2006. Outlook for cellulase improvement: Screening and selection strategies. *Biotechnology Advances*, 24, 452-481.
- ZHONG, L., MATTEWS, J. F., HANSEN, P.I., CROWLEY, M.F., CLEARLY, J.M., WALKER, R.C., NIMLOS, M.R., BROOKS III, C.L., ADNEY, W.S., HIMMEL, M.E. & BRADY, J.W. 2009. Computational simulations of the *Trichoderma reesi* cellobiohydrolase I acting on microcrystalline cellulose Iβ: the enzyme-substrate complex. *Carbohydrate Research*, 344-1984-1992.
- ZVERLOV, V. V., KELLERMANN, J., & SCHWARZ, W. H. 2005. Functional subgenomics of Clostridium thermocellum cellulosomal genes: identification of the major catalytic components in the extracellular complex and detection of three new enzymes. *Proteomics*, 5, 3646-3653.
- ZVERLOV, V., MAHR, S., RIEDEL, K., & BRONNENMEIER, K. 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.