



UNIVERSITY *of the*
WESTERN CAPE

Identification and Characterisation of Novel Cellulolytic Genes Using Metagenomics

Xiao Ping Hu

A thesis submitted in partial fulfillment of the requirements for the degree of
MAGISTER SCIENTIAE (M.Sc)
Department of Biotechnology,
University of the Western Cape
Bellville

Supervisor: Prof. D.A. Cowan

September 2010

Declaration

Hereby I, *Xiao Ping Hu*, declare that “*Identification and Characterisation of Novel Cellulolytic Genes Using Metagenomics*” is my own original work and that all sources have been accurately reported and acknowledged, and that this document has not previously in its entirety or in part been submitted at any university in order to obtain an academic qualification.

Full name: Xiao Ping Hu

Date: September

2010

Signed:



Abstract

Metagenomics has been successfully used to discover novel enzymes from uncultured microorganisms in the environment. In this study, metagenomic DNA from a Malawian

hot spring soil sample was used to construct a fosmid library. This metagenomic library comprised of more than 10000 clones with an average insert size of 30 kb, representing more than 3.0×10^8 bp of metagenomic DNA (equivalent to approximately 100 bacterial genomes).

The library was screened for cellulase activity using a Congo red plate assay to detect zones of carboxymethylcellulose hydrolysis. This yielded 15 positive fosmid clones, of which five were further characterised for activity and thermostability using the 3, 5-dinitrosalicylic assay. Two of the five fosmids (XP008C2 and XP026G5) were selected for DNA pyrosequencing. The full sequence of the XP008C2 (29800bp) fosmid insert is presented in this study and genes thereon were chosen for further study.

Two putative cellulases XPgene12 (993 bp) and XP gene25 (1107 bp) were identified from fosmid XP008C2. They were independently cloned and heterologously expressed in *E.coli* Rosetta pLysS. XPgene12, corresponding to a 37 kDa recombinant cellulase was purified to homogeneity using Ni-chelation chromatography and kinetically characterized with carboxymethylcellulose as the substrate. The enzyme displayed optimum activity at 50°C and pH4. Through this characterization study, XPgene12 has been defined as a novel thermophilic and moderately acidophilic endoglucanase which has potential value for industrial applications.

Acknowledgments

I would like to express my appreciation to the following people for contribution:

1. Professor Don Cowan, for the honour to work under his supervision and within his laboratory and for creating an environment conducive to research activities.

2. Dr Marla Tuffin and Dr Rolene Bauer for their professional supervision, guidance, sound judgement and enthusiasm. Thank you Marla for all your gifts, you gave me a sense of belonging.
3. Dr Mark Taylor, thank you for almost everyday discussions, support and encouragement throughout the past two years. You are not only a good teacher, but also a good friend.
4. I can not thank Dr Heide Goodman enough. I would not have made it without your unconditional support. Your encouragement, enthusiasm and genuine caring attitude have done so much for me. Thank you for taking such good care of me.
5. To Moola, Bronwyn, Dom, Mesfin and Colin, thank you for your friendship, I really appreciate your continued support, love and encouragement. I am so proud to have you guys as my friends.
6. To Dr Rob Hubby and Mr. Lonnie van Zyl, thanks for random conversations and discussions.
7. To my family: I would like to thank my father and my brother for their love, support, advice and patience, my late mother for her love.
8. I would like to thank Professor Don Cowan for providing the environmental soil sample that was the source of the metagenomic DNA.
9. To Adeola Oluwaseyi Poroye who stood by me and supported me through everything.

In loving memory of my late mother

Table of contents

Declaration.....	i
Abstract.....	ii
Acknowledgments.....	iii
List of figures.....	vii
List of tables.....	ix
Abbreviations.....	xi
Chapter 1 Literature review.....	1
1.1 Introduction.....	1
1.2 Biofuel.....	3
1.2.1 Liquid fuels.....	4
1.2.2 Biofuel production from lignocellulosic biomass.....	6
1.2.3 The need for pretreatment.....	7
1.3 Lignocellulose degrading enzymes.....	8
1.3.1 Cellulases.....	8
1.3.2 Hemicellulases.....	10
1.4 Glycoside hydrolase family.....	11
1.4.1 Classification of glycoside hydrolases.....	11
1.4.2 Glycoside hydrolase family 8.....	12
1.5 Thermophiles and thermophilic enzymes.....	13
1.5.1 Features of thermophilic enzymes.....	14
1.5.2 Potential application of thermophilic enzymes for bioethanol production.....	15
1.6 Metagenomics and gene discovery.....	16
1.6.1 Metagenomic technologies.....	17
1.6.2 Gene discovery.....	18
1.7 Molecular techniques.....	20
1.7.1 Metagenomic DNA extraction from soil.....	20
1.7.2 Screening of the metagenomic DNA libraries.....	21
1.7.3 Metagenomic sequencing.....	23
1.8 Aims and objectives of the current study.....	25
Chapter 2 General materials and methods.....	25
2.1 General laboratory chemicals and reagents.....	25
2.2 Media.....	26
2.3 Bacterial strains and plasmids.....	28

2.4 DNA extraction	30
2.4.1 Alkaline cell lysis method for plasmid DNA extraction	30
2.4.2 Plasmid extraction for sequencing quality DNA	30
2.4.3 Fosmid extraction	31
2.5 Analytical procedures	32
2.5.1 Spectrophotometry.....	32
2.5.2 Fluorometry (Qubit™).....	32
2.5.3 Quantification of fosmid DNA.....	32
2.5.4 Agarose gel electrophoresis.....	32
2.6 Molecular biology techniques	33
2.6.1 Restriction enzyme digestion.....	33
2.6.2 DNA ligation	33
2.6.3 Preparation of competent cells	33
2.6.4 Transformation of competent <i>E. coli</i> cells.....	35
2.6.5 Sequencing.....	35
2.6.6 Polymerase chain reaction.....	36
2.7 Protein analysis techniques	38
2.7.1 Bradford assay	38
2.7.2 Nanodrop analysis.....	38
2.7.3 SDS-PAGE	38
2.7.4 Zymogram	40
Chapter 3 Metagenomic fosmid library construction and functional screening for cellulase activity.....	40
3.1 Introduction.....	40
3.2 Materials and methods.....	43
3.2.1 Sample information	43
3.2.2 DNA extraction from soil.....	43
3.2.3 Size fractionation and DNA purification.....	44
3.2.4 Cloning of high molecular weight DNA.....	45
3.2.5 Phage packing of the fosmid clones.....	46
3.2.6 Phage titering.....	47
3.2.7 Library verification	47
3.2.8 Prokaryotic diversity study.....	48
3.2.9 Preparation of glycerol stocks.....	49
3.2.10 Functional screening of the library for cellulase activities	50
3.2.11 Secondary screening for cellulase activity	50
3.2.12 Preliminary cellulase assay.....	51
3.3 Results and discussion	53
3.3.1 Metagenomic fosmid library construction	53
3.3.2 Library verification	54
3.3.3 Prokaryotic diversity study.....	59
3.3.4 Functional screening of the library	61
3.3.5 Preliminary cellulase assay	63
Chapter 4 Sequencing analysis and homology modeling	67
4.1 Introduction.....	67
4.2 Sequence analysis.....	69
4.3 Phylogenetic analysis.....	85

4.4 Homology modelling.....	86
Chapter 5 Cloning, expression and characterization of cellulolytic genes from a soil metagenomic library.....	97
5.1 Introduction.....	97
5.2 Materials and methods.....	98
5.2.1 Cloning of cellulolytic genes XPgene12 and XPgene25	98
5.2.2 Expression of the cellulolytic gene XPgene12.....	100
5.2.3 Cellulase activity assay.....	102
5.3 Results and discussion	104
5.3.1 Cloning of cellulase encoding genes XPgene12 and XPgene25	104
5.3.2 Expression and purification of XPgene12	107
5.3.3 Enzymatic characterization of the XPgene12 gene product.....	109
5.3.4 Thin layer chromatography.....	113
Chapter 6 General discussion	115
References	118



List of figures

Figure 1.1 A summary of second generation bioethanol production.....	6
Figure 1.2 The lignocellulosic complex has three major components: cellulose, hemicellulose and lignin.....	7
Figure 1.3 The three major types of cellulases and their hydrolysis reactions	9
Figure 1.4 The three major types of hemicellulases and their hydrolysis reactions (Kumar <i>et al.</i> , 2008).....	11
Figure 1.5 Multiple alignment of glycoside hydrolase family 8 enzymes	13
Figure 1.6 Temperature profiles of the lignocellulose-to-ethanol conversion processes.....	16
Figure 1.7 Construction of metagenomic libraries from environmental samples and methods for analyzing functions and sequences in the DNA library (Schmeisser <i>et al.</i> , 2007)	18

Figure 3.1 Agarose gel electrophoresis of extracted metagenomic DNA from the Mphizi hot spring site.....	54
Figure 3.2 Agarose gel electrophoresis of 12 restriction endonuclease digested randomly selected fosmid clones.....	55
Figure 3.3 PCR amplification of the 16S rRNA genes from the metagenomic library using universal bacterial PCR primers 341 F-GC and 534r.....	60
Figure 3.4 DGGE profile of 16S rRNA gene content of the Mphizi hot spring soil metagenomic library.....	60
Figure 3.5 Putative cellulase producing fosmid clones screened on CMC LB agar plates flooded with Congo red.....	62
Figure 3.6 Restriction profiles of the 17 transformants which produced zones of hydrolysis during screening of the metagenomic library on CMC LB agar indicator plates.....	63
Figure 3.7 DNS assay performed in the presence of culture supernatant and cell extract of fosmid clone 008C2.....	64
Figure 3.8 Glucose standard curve for DNS assay.....	66
Figure 3.9 Thermostability of 5 chosen fosmid clones using the DNS assay.....	67
Figure 4.1 Annotation of the fosmid clone 008C2 diagram using sequencer.....	71
Figure 4.2 Arrangement of the open reading frames identified in the insert of fosmid 008C2.....	77
Figure 4.3 Nucleotide and deduced amino acid sequences of XPgene12.....	80
Figure 4.4 Nucleotide and deduced amino acid sequences of XPgene25.....	81
Figure 4.5 Alignment of XPgene12, cellulase from <i>Enterobacter</i> sp.638 and endoglucanase from <i>Klebsiella</i> subsp. <i>rhinoscleromatis</i> ATCC 13884 showing conserved sequences.....	82
Figure 4.6 Alignment of XPgene 25, Endo-1, 4-D- glucanase from <i>Citrobacter rodentium</i> ICC168 and <i>Enterobacter cancerogenus</i> ATCC 35316 endoglucanase showing conserved sequences.....	82
Figure 4.7 Structure-based partial sequence alignment among characterized endoglucanases belonging to GH-8.....	83
Figure 4.8 Prediction of N-terminal signal peptide cleavage site in polypeptide XPgene12.....	84
Figure 4.9 Prediction of N-terminal signal peptide cleavage site in polypeptide XPgene25.....	84
Figure 4.10 Phylogenetic tree of XPgene12 and XPgene25 generated by the neighbour-joining method and on the CLC genomics work bench software (CLC Bio).....	86
Figure 4.11 Secondary structure for the amino acid sequence obtained for XPgene12.....	88
Figure 4.12 Secondary structure for the amino acid sequence obtained for XPgene25.....	90
Figure 4.13 Homology models of the XPgene12, XPgene25 and the <i>Acetobactactexylinum</i> endo-beta-1, 4-glucanase CMCAx gene built by the SWISS-MODEL server.....	91
Figure 4.14 Ramachandran plot analysis of XPgene12 for general, gly, Pre-Pro built by the SWISS-MODEL using RAMPAGE software.....	93
Figure 4.15 Ramachandran plot analysis of XPgene12 for general, gly, Pre-Pro built by 3D-JIGSAW using RAMPAGE software.....	94
Figure 4.16 Ramachandran plot analysis of XPgene25 for general, gly, Pre-Pro built by the SWISS-MODEL server using RAMPAGE software.....	95
Figure 4.17 Ramachandran plot analysis of XPgene25 for general, gly, pre-pro built by 3D-JIGSAW using RAMPAGE software.....	96
Figure 5.1 Cloning of XPgene12 and XPgene25 into pET 21a vector.....	105

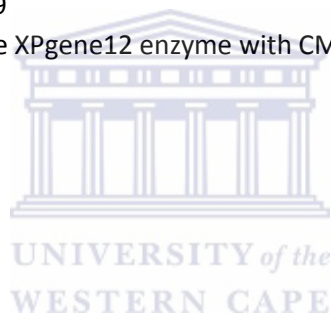
Figure 5.2 A XPgene12 <i>E. coli</i> Rosetta(DE3)pLysS transformant demonstrating a zone of clearance on a LB agar plate containing CMC(1%) after staining with Congo red.....	106
Figure 5.3 PCR amplification of XPgene12 and XPgene25 using gene specific primers (Table 2.5) for confirmation of cloning into the pET vectors.....	107
Figure 5.4 SDS-PAGE analysis of cell extracts of XPgene12-pET21a in <i>E. coli</i> Rosetta(DE3)pLysS.....	107
Figure 5.5 SDS-PAGE analysis of His-Tag purification of XPgene12-pet21a in <i>E. coli</i> Rosetta(DE3)pLysS	108
Figure 5.6 SDS-PAGE analysis of purified XPgene12 product (Lane 1) showing the zone of hydrolysis produced on a zymogram (Lane 2).	109
Figure 5.7 Effect of pH on XPgene12 protein activity with CMC as the substrate at 22°C.....	110
Figure 5.8 Effect of temperature on XPgene12 protein activity with CMC as substrate after 10mins incubation.....	111
Figure 5.9 The thermal inactivation profile of XPgene12 product at 80°C (▲), 70°C (■) and 60°C (◆).....	111
Figure 5.10 Activity of XPgene12 protein towards CMC, β-D-glucan, methyl-cellulose and xylan	112
Figure 5.11 Hydrolysis products of cello-oligosaccharides by the purified XPgene12 product.	114



List of tables

Table 1.1 First generation biofuels	4
Table 1.2 Fourteen glycoside hydrolase clans of related families	12
Table 1.3 Advantages and disadvantages of metagenome screening methods	23
Table 2.1 Growth media used in the study.....	26
Table 2.2 Stock and final concentrations of antibiotics used in the study	28
Table 2.3 Bacterial strains.....	29
Table 2.4 Plasmids used in the study.....	29

Table 2.5 Primers used in the study.....	37
Table 2.6 Preparation of 12% separating gels and 4% stacking gels for SDS-PAGE.....	39
Table 3.1 Location and characteristics of the Mphizi hot springs, Malawi	43
Table 3.2 Nucleotide end-sequences of selected fosmid clones and identities of the closest match.....	57
Table 3.3: DNS assay to determine reducing sugar generation by fosmid clones 008C2, 032B10, 026G5, 027B3, 032C10.....	65
Table 4.1 Nucleotide end-sequences of fosmid clones 008C2 and 026G5. The nucleotide identity of the closest match is indicated.....	72
Table 4.2 Predicted genes in fosmid 008C2.....	73
Table 4.3 Rare codons and their frequency in the nucleotide sequence of XPgene12 predicted by rare codon calculator.....	84
Table 4.4 Rare codons and their frequency in the nucleotide sequence obtained for XPgene25 predicted by rare codon calculator.....	85
Table 5.1 Recombinant plasmids constructed for expression studies.....	99
Table 5.2 : Kinetic parameters of the XPgene12 enzyme with CMC as a substrate	112



Abbreviations

Aa	Amino acid
APS	Ammonium persulphate
Bp	Base pair
BSA	Bovine serum albumin
CAPS	<i>N</i> -cyclohexyl-3-aminopropanesulfonic acid
CAM	Chloramphenicol
CAZY	Carbohydrate-Active Enzymes database
CMC	Carboxymethylcellulose sodium salt
CTAB	Cetyl-trimethyl-ammonium bromide
C-terminus	Carboxy terminus
Da	Dalton
ddH ₂ O	Deionised distilled water
DGGE	Denaturing gradient gel electrophoresis
DNA	Deoxyribonucleic acid
dNTP	Deoxynucleoside triphosphate
DTT	Dithiothreitol
EDTA	Ethylene diamine tetraacetic acid
EtBr	Ethidium bromide
EtOH	Ethanol
hr(s)	Hour(s)
IMBM	Institute for Microbial Biotechnology and Metagenomics
IPTG	Isopropyl β-D-thiogalactosidase
k_{cat}	Catalytic turnover
kDa	Kilo dalton
K_M	Michaelis-Menten constant
MES	2-(<i>N</i> -morpholino)ethanesulfonic acid
min(s)	Minute(s)
mM	Millimolar
μg	Microgram
μl	Microlitre
ml	Millilitre
MW	Molecular weight
Ng	Nanogram
N-terminus	Amino-terminus
OD	Optical density
ORF	Open reading frame
PAGE	Polyacrylamide gel electrophoresis
PBS	Phosphate buffered saline
PDB	Phage dilution buffer
PCR	Polymerase chain reaction

PVPP	Polyvinylpyrrolidone
SDS	Sodium dodecyl sulphate
sec(s)	Second(s)
sp.	Species
TAE	Tris acetate EDTA
TE	Tris EDTA
TEMED	N,N,N',N'-tetramethylethylenediamine
Tris-HCl	Tris (hydroxymethyl)methylamine hydrochloride
V_{\max}	Maximum velocity
X-gal	5-bromo-4-chloro-3-indolyl- β -D-galactoside



Chapter 1 Literature review

1.1 Introduction

Oil prices have fallen significantly since the 2007 peak of approximately \$100 per barrel (<http://futures.tradingcharts.com/chart/CO/M>, accessed 22 August 2010). Increased global fuel consumption and decreased crude oil production from politically and socially unstable countries has prompted the United States government to propose the use of 7.5 billion gallons of bioethanol be used to supplement fuel supplies by 2012 and the EU to state that 10% of all transport fuel must come from renewable sources by 2010. Similar targets have been proposed by South Africa, mandating the use of 10 000 GWh (0.8 Mtoe) renewable energy contribution by 2013.

The production of bioethanol as a renewable fuel has attracted a global interest (Hahn-Hagerdal *et al.*, 2006) with the increasing demand of economically competitive bioethanol derived from cheap and unlimited raw materials such as lignocellulose for transportation (Szczodrak & Fiedurek, 1996).

Lignocellulose is the major structural component of the plant biomass typically found in agricultural and municipal wastes. It represents a major source of renewable organic matter that can be degraded by certain microorganisms and deconstructed by their enzymes, collectively known as xylanases, ligninases and cellulases (Lopez *et al.*, 2002).

Cellulose is a major fraction of many lignocellulosic materials and the enzymatic conversion of cellulose to monomeric or polymeric variants of glucose is of great interest as a precursive step in fermentations to produce ethanol (Ohgren *et al.*, 2007). The cellulases can be sub classified as: a) endo-cellulases that mediate the cleavage of non-covalent interactions present in the crystalline structure of cellulose b) exo-cellulases that

mediate hydrolysis of the individual cellulose fibres to smaller sugar units and c) beta-glucosidases that hydrolyse the cleavage of cellobiose to monomeric glucose (Mussatto *et al.*, 2008).

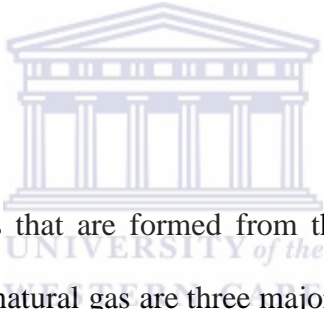
Many industrial enzymes are used at temperatures higher than 50°C and there is an increasing demand for the development of thermostable biocatalysts (Palomo *et al.*, 2004) that would in principle be more tolerant to fluctuations in process conditions and function at temperatures that facilitate sugar solubility and feedstock loading. In addition, certain economic savings associated with process heating and cooling cycles could be made by linking high temperature hydrolysis to a thermophilic fermentation process. Such processes are currently being developed by companies such as BioGasol (Denmark), Mascoma (USA) and TMO Renewables Ltd (UK). Thermophiles have a number of process advantages for ethanol production when compared to mesophilic organisms. These include a temperature associated increase in reaction rate, a decrease in the enzyme loading needed and an increased rate of substrate conversion to product (Haki & Rakshit, 2003; Koskinen *et al.*, 2007 de Vrije *et al.*, 2009). At high temperatures microbial contamination will decrease and gas solubility will be reduced facilitating the maintenance of a near anaerobic environment suitable for a fermentative process. Additionally, at high temperatures the solubility of sugars increases and crystalline/polymeric substrates become more accessible (Sommer *et al.*, 2004).

Thermostable cellulosic enzymes can be isolated from organisms living in various hot environments either through direct culturing and characterisation or via more sophisticated techniques such as metagenomics. In the search for thermozyms capable of deconstructing cellulosic biomass, corresponding thermal habitats rich in these materials would be the most productive sites (Blumer-Schuette *et al.*, 2008). One potential sample that forms part of the IMBM (University of the Western Cape, South Africa) collection is

from the Mphizi hot spring site, Chiweta (Malawi). The Mphizi hot spring site contains a number of geothermal sites wherein the temperatures fluctuates between 40 and 80°C. As a result of the geographical and environmental surrounds these thermal sites are rich in organic content such as decomposing grasses, plant materials, as well as human and animal waste.

The majority of microorganisms (99%) in the environment cannot be cultured using conventional laboratory techniques (Handelsman, 2004). Metagenomic tools can be used to mine the uncultivable and not yet cultured quotient of a suitable environmental sample to discover novel enzymes and biomolecules. In this project functional screening of metagenomic libraries for cellulase activity was performed.

1.2 Biofuel



Fossil fuels are natural resources that are formed from the organic remains of prehistoric plants and animals. Coal, oil and natural gas are three major forms of fossil fuels. Fossil fuels produce significant amounts of energy when they are burned, and a large percentage of the world's energy is supplied by fossil fuels. Up until 2004 the world was consuming 66.7% of its energy requirements in the form of coal, gas and oil. However fossil fuels are limited resources of energy and the consumption of fossil fuels is 100,000 times faster than its production. The Hubbert Peak Theory predicts that the supply of fossil fuels will be exhausted in the near future (Hubbert, 1956).

Thus a major challenge in the twenty-first century is the rate of excessive exploitation of the Earth's fossil energy (Kumar *et al.*, 2008). The negative impact of fossil fuels on climate change and of greenhouse gas emissions on the environment the dependence on non-renewable resources and the increased demand for energy for transportation, heating

and industrial processes are stimulating increased commercial interest in renewable energy technologies (Hahn-Hagerdal *et al.*, 2006). Biofuels are fuels produced from renewable biomass material which have the potential to replace the dependence on non-renewable fossil fuels.

1.2.1 Liquid fuels

1.2.1.1 First generation biofuels

First generation biofuels refer to the fuels that are made from food crops such as starch, sugar, animal fats and vegetable oil using conventional fermentation technology (<http://biofuel.org.uk/first-generation-biofuels.html>, accessed 22 August 2010). Table 1.1 describes some of the most popular types of first generation biofuels (After Gomez *et al.*, 2008a).

Table 1.1 First generation biofuels

Biofuel type	Specific name	Biomass feedstock	Production process
Bioethanol	Conventional bioethanol	Sugar beets, grains	Hydrolysis, fermentation
Pure vegetable oil	Pure plant oil (PPO)	Oil crops (e.g. rape seeds)	Cold pressing/ extraction
Biodiesel	Biodiesel from energy crops Rape seeds methyl (ester) (RME), fatty acid methyl/ethyl ester (FAME, FAEE)	Oil crops (e.g. rape seeds)	Cold pressing/ extraction, Transesterification
Biodiesel	Biodiesel from FAME/FAEE	Waste, cooking, frying oil	Transesterification
Biogas	Upgraded biogas	Biomass (wet)	Digestion
Bio ETBE		Bioethanol	Chemical synthesis

First generation biofuels have limitations with respect to their use as an oil-product substitute, and a stimulator of economic growth. These include the creation of competition for land, water and food resources and the total production costs which are expensive for energy security. Furthermore they have been accused of accelerating deforestation (Gomez *et al.*, 2008a). Concerned with the impact of these issues, researchers are increasingly looking to develop biofuels produced from non-food biomass (Tilman *et al.*, 2006). Feedstocks from lignocellulosic materials include cereal straw, bagasse, forest residues and purpose-grown energy crops such as vegetative grasses and short rotation forests (Sims *et al.* 2010).

1.2.1.2 Second generation biofuels

Second generation biofuels refer to the fuels that are made from non food crops such as lignocellulosic biomass. The biomass can include straw waste, cornstalks, wood chippings or other organic materials (Hahn-Hagerdal *et al.*, 2006). The second-generation technology is predicted to more than double bioethanol yields without interrupting the food chain since it allows the fuel to be produced from any organic material (Jeczmionek *et al.*, 2006). However, the majority of second-generation biofuel technologies are not at present commercially available.

Biofuels have the potential to reduce the emission of greenhouse gases when compared to conventional transport fuels. Life Cycle Analysis predicts that first generation biofuels can save up to 60% of carbon emissions compared to fossil fuels whereas second generation biofuels can save up to 80%. In addition, biofuel production is helping to deal with poverty alleviation around the world by increasing employment opportunities in rural areas (Koh *et al.*, 2009).

There are five stages to the production of a second generation ethanol using a biological approach. These are shown in figure 1.1.

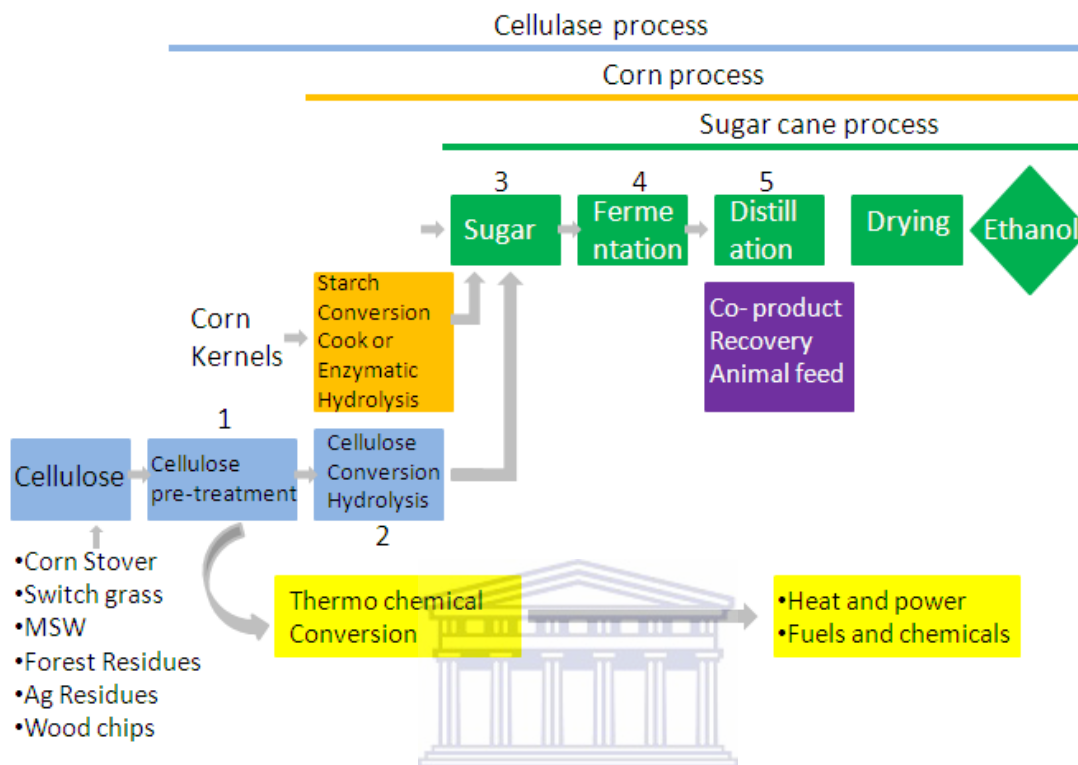


Figure 1.1 A summary of second generation bioethanol production

1) Pretreatment of lignocellulosic biomass amenable to hydrolysis; 2) Cellulosic enzyme hydrolysis to break down the molecules into sugars; 3) Separate sugar solution from the residual materials such as lignin; 4) Microbial fermentation of the sugar solution; 5) Distillation to produce pure alcohol and make use of co-product. (After <http://archive.energyfarms.net/blog/98?page=1>, accessed 22 August 2010).

1.2.2 Biofuel production from lignocellulosic biomass

Biomass represents a rich source of renewable natural biological material which may become important for the world's energy and chemical feedstock requirements (Gomez *et al.*, 2008a). Cellulose, hemicellulose, and lignin are the three major components of lignocellulosic biomass (Figure 1.2) (Gomez *et al.*, 2008b). Cellulose is the most abundant component and is composed of β -1, 4- linked glucose residues. Hemicellulose comprises 20-30% of typical biomass material and is a more complex structure of pentoses, hexoses etc. Lignin is a major component of plant cell walls and accounts for

approximately 30% of the terrestrial organic carbon fixed in the biosphere annually (Zhang *et al.*, 2006; Li *et al.*, 2008; Scheller & Ulvskov, 2010).

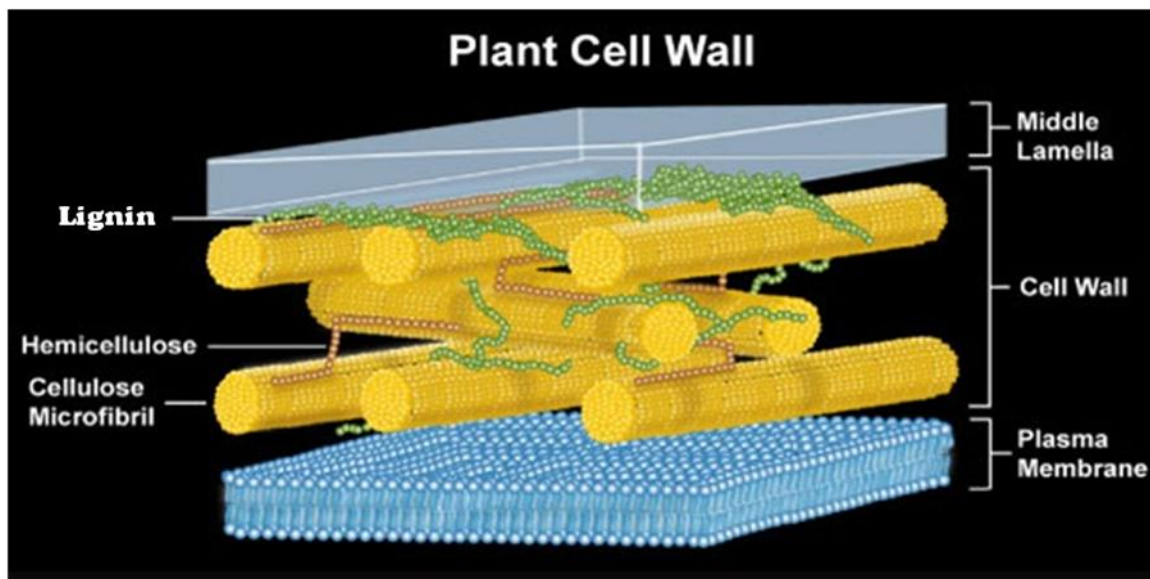


Figure 1.2 The lignocellulosic complex has three major components: cellulose, hemicellulose and lignin

The cellulose and hemicellulose fractions of lignocellulosic biomass can be converted into sugars which are fermented to produce bioalcohols such as bioethanol or biobutanol. Thermochemical processing and other biochemical processing are the two main methods for producing biofuels from biomass. Thermochemical processing converts biomass into products by thermal decay and chemical reformation. Biochemical processing converts biomass into sugars by enzymatic degradation and acid hydrolysis (Gomez *et al.*, 2008b).

1.2.3 The need for pretreatment

Lignocellulose is highly resistant to enzymatic degradation and pretreatment is needed to reduce the size of the lignocellulosic biomass to allow the hydrolytic enzymes to access the substrates (Mosier *et al.*, 2005). Several pretreatment methods have been developed: physical (mechanical comminution and pyrolysis), physico-chemical (steam explosion, ammonia explosion and CO₂ explosion), chemical (ozonolysis, acid hydrolysis, alkaline

hydrolysis, oxidative delignification and the organosolv process) and biological pretreatment using microorganisms (Sun & Cheng, 2002).

1.3 Lignocellulose degrading enzymes

Lignocellulose degrading enzymes are of interest for the hydrolysis of lignocellulosic biomass which can be utilized for bioethanol production. Two processes are involved in the conversion: hydrolysis of cellulose in the lignocellulosic materials to fermentable reducing sugars and fermentation of sugars to ethanol (Sun & Cheng, 2002; van Wyk, 2001).

1.3.1 Cellulases

Cellulases hydrolyze cleavage of the 1, 4 -beta-D-glycosidic bonds in cellulose and can be used to convert cellulose fibres to smaller units, primarily glucose (Parsiegla *et al.*, 2002). Cellulases have many biotechnological applications including in the production of bioethanol, textiles, detergents, food stuffs and animal feed. They are also used in the paper and pulp and pharmaceutical industries and in waste management (Bhat, 2000; Camassola & Dillon, 2007). Cellulases involved in the hydrolysis process have been classified on the basis of their action towards a substrate into three major groups: endoglucanases (EC 3.2.1.4), exoglucanases (EC 3.2.1.91) and β -glucosidases (EC 3.2.1.21) (Mussatto *et al.*, 2008) (Figure 1.3). Each type of cellulase hydrolyses a specific substrate.

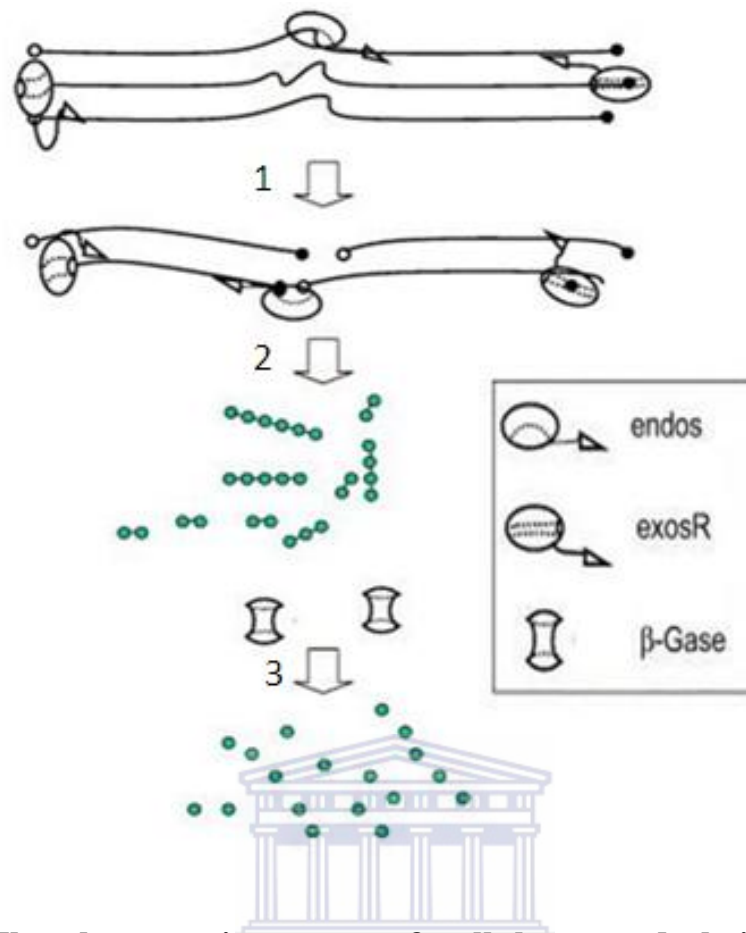


Figure 1.3 The three major types of cellulases and their hydrolysis reactions

1. Breakage of internal bonds to disrupt the crystalline structure of cellulose and expose individual cellulose polysaccharide chains (endoglucanase). 2. Hydrolysis of the individual cellulose fibres produce smaller sugars units (exoglucanase) units. 3. Hydrolysis of disaccharides and tetrasaccharides into individual monosaccharides (beta-glucosidase) (Zhang *et al.*, 2006).

The majority of reported cellulases have been isolated from cultured microorganisms (fungi and bacteria) and are able to catalyze the degradation of the cellulose complex. The cellulosome complex produced by anaerobic bacteria of the genera *Clostridium* and *Bacteroides* has also been identified (Lamed *et al.*, 1983; Bayer *et al.*, 1998; Schwarz, 2001). Cellulosomal enzymes carry a dockerin domain that incorporates the enzyme into the cellulosome complex, whereas non-cellulosomal enzymes usually include a carbohydrate-binding module for guiding the catalytic domain to the substrate (Schwarz, 2001).

Cellulases from specific microorganisms usually display activity which is specific to the environment from which they were isolated (Voget *et al.*, 2006). The limitation of traditional cultivation methods increases the attraction for using metagenomics to mine metagenomes for genes encoding novel cellulases from natural cellulase rich environments such as the soil, gut, cow rumen and biogas digesters which demonstrate highly hydrolytic activities (Schluter *et al.*, 2008; Morrison *et al.*, 2009; Wang *et al.*, 2009).

1.3.2 Hemicellulases

Hemicellulases are a group of enzymes that hydrolyze hemicellulose. The same classification outlined above can be applied to hemicellulose hydrolysis. Enzymatic action involves the following reactions:

1. Endo-xylanase degradation of internal β -1,4-D-xylose linkages of the xylan backbone.
2. Exo-xylanase degradation of β -1,4-D-xylose linkages releasing xylobiose.
3. β -xylosidase that releases D-xylose from xylobiose and xylo-oligosaccharides (Saha, 2003).

The reactions are depicted in Figure 1.4.

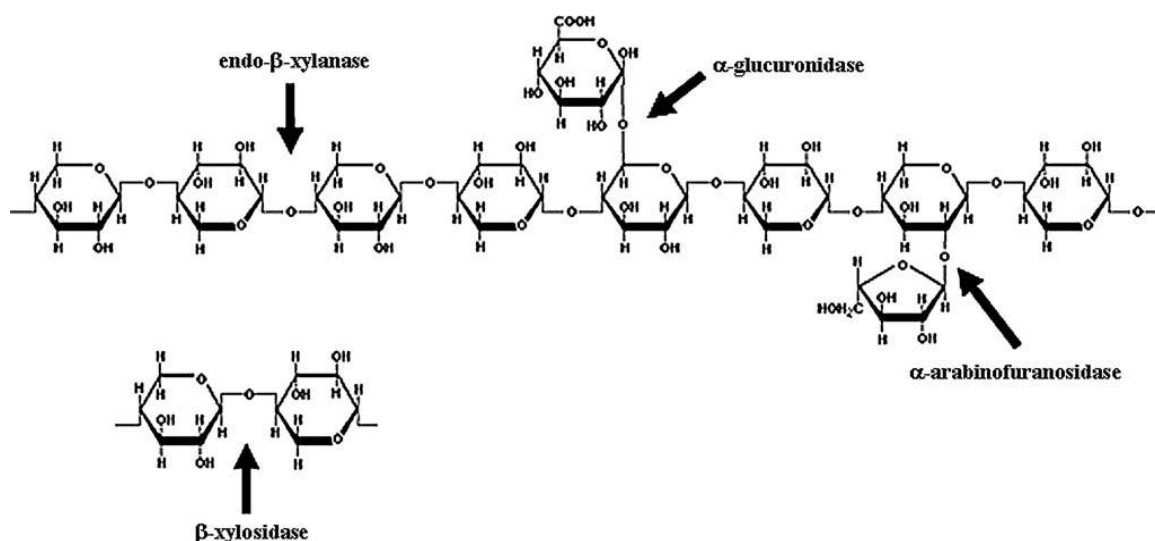


Figure 1.4 The three major types of hemicellulases and their hydrolysis reactions (Kumar *et al.*, 2008)

1.4 Glycoside hydrolase family

Glycoside hydrolases are groups of enzymes which catalyze the hydrolysis of the glycosidic linkages to produce two smaller sugar molecules. These are the most abundant enzymes in nature and can mediate the degradation of biomass (cellulose, hemicelluloses etc.), disrupt bacterial cell walls (lysozyme), be a drug target for the prevention of influenza infection (neuraminidase) and hydrolyse mannose (mannosidases) (Henrissat & Bairoch, 1996). A classification of glycoside hydrolases in families based on amino acid sequence similarity has been proposed (Henrissat, 1991; Henrissat & Bairoch, 1993). It reflects the structural features of these enzymes better than their substrate specificity, helps to reveal the evolutionary relationships between these enzymes, provides a convenient tool to derive mechanistic information (Henrissat, 1991; Henrissat & Bairoch, 1993) and explains the difficulty of deriving relationships between family membership and substrate specificity.

1.4.1 Classification of glycoside hydrolases

Based on their amino acid sequence similarities and according to a direct relationship between sequence and folding similarities, the Carbohydrate-Active Enzymes database (CAZy) (<http://www.cazy.org/>) was generated to aid the classification of members of the glycoside hydrolase family.

The CAZy database provides a continuously updated list of the glycoside hydrolase families. Because the folds of proteins are generally better conserved than their sequences, such families can be grouped into clans. There are 118 family members plus

one unclassified sequence in the glycoside hydrolase family. Fourteen glycoside hydrolase clans of related families exist in this database.

Table 1.2 Fourteen glycoside hydrolase clans of related families

GH-A	(β/α) ₈	1 2 5 10 17 26 30 35 39 42 50 51 53 59 72 79
GH-B	β -jelly roll	7 16
GH-C	β -jelly roll	11 12
GH-D	(β/α) ₈	27 31 36
GH-E	6-fold β -propeller	33 34 83 93
GH-F	5-fold β -propeller	43 62
GH-G	(α/α) ₆	37 63
GH-H	(β/α) ₈	13 70 77
GH-I	$\alpha+\beta$	24 46 80
GH-J	5-fold β -propeller	32 68
GH-K	(β/α) ₈	18 20 85
GH-L	(α/α) ₆	15 65
GH-M	(α/α) ₆	8 48
GH-N	β -helix	28 9

The table was taken from <http://www.cazy.org/Glycoside-Hydrolases.html>

1.4.2 Glycoside hydrolase family 8

The glycoside hydrolase family 8 (GH-8) proteins share a 6 barrel motif, which is a typical fold of enzymes in the GH-8 family. The family comprises several known enzyme activities including endoglucanase (EC: 3.2.1.4), lichenase (EC: 3.2.1.73) and chitosanase (EC: 3.2.1.132). These enzymes were formerly known members of the cellulase family D

(Henrissat *et al.*, 1989). GH-8 proteins have the most conserved region, a stretch of about 20 residues that contains two catalytic aspartates (Figure 1.5) (Alzari *et al.*, 1996). Of the forty eight characterized proteins, two from this family were from uncultured bacteria (CAZy database).

The GH-8 members share conserved catalytic triad residues (EDD), and aromatic residues forming sugar recognition subsites (Figure 1.5) (Yasutake *et al.*, 2006) .

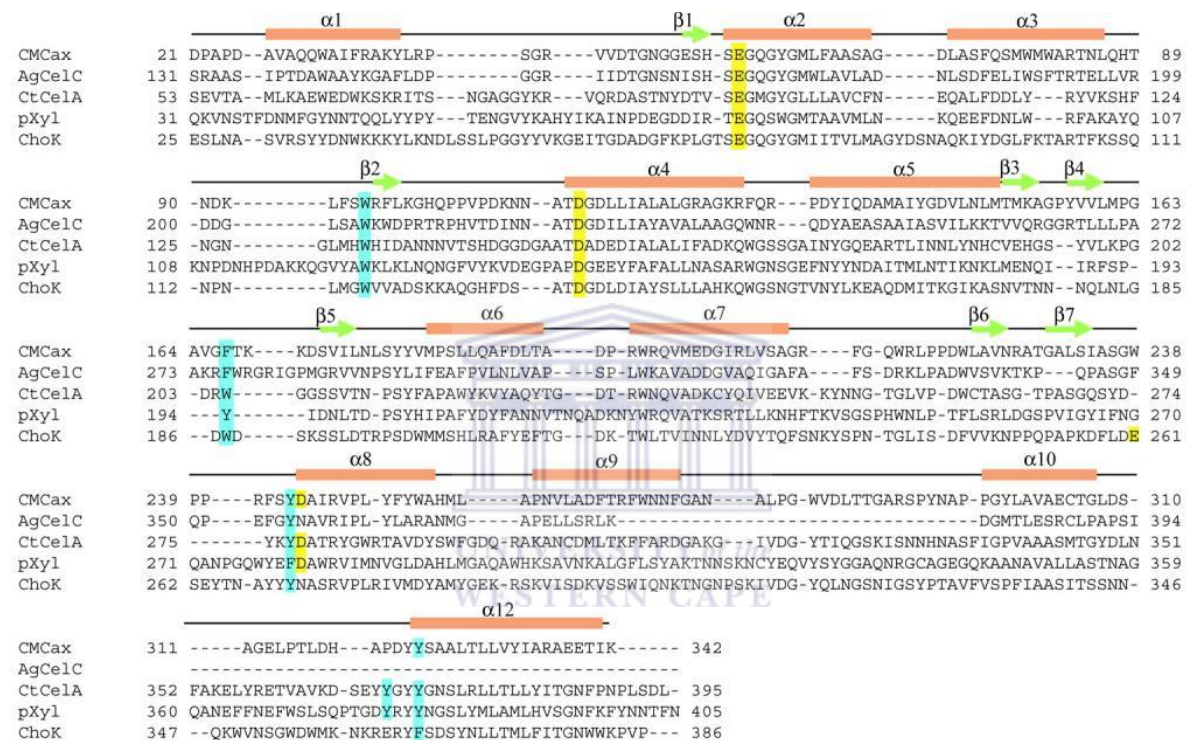


Figure 1.5 Multiple alignment of glycoside hydrolase family 8 enzymes (Yasutake *et al.*, 2006)

Secondary structure elements, conserved catalytic residues (highlighted in yellow) and the aromatic residues forming sugar recognition subsites (shown in blue) are shown (Yasutake *et al.*, 2006).

1.5 Thermophiles and thermophilic enzymes

A “normal environment” on earth is an anthropocentric term in that it refers to areas having a temperature range between 10-40°C, a pH close to neutrality, an atmospheric pressure close to one atmosphere, sufficient water and low levels of radiation. Higher

plants, animals and humans inhabit these areas. Microorganisms however have adapted to live in areas outside of this “normal environment” and may be found in the “extreme environments” on earth, such as areas with high and low temperatures, high and low pHs, high salt concentrations, high pressure, etc. Organisms that thrive in these types of environment are termed extremophiles. They are classified according to the different extreme habitats in which they exist.

One such group, the thermophiles, has an optimum growth temperatures between 45-80°C while hyperthermophiles have optimal growth temperatures of >80°C (Stetter, 1996). They inhabit various hot regions on the earth such as thermal springs and deep sea hydrothermal vents on the ocean floor. They can also live in biomass such as peat bogs and compost which can often reach temperature of >60°C (Madigan & Martinko, 2006). Thermophilic enzymes are of particular interest due to their potential application in biotechnology because of their perceived increased robustness under process conditions (Cowan & Daniel, 1996; Hough & Danson, 1999). Consequently several studies have been reported on the discovery of novel thermostable proteins such as xylanases (Pantazaki *et al.*, 2002; van den Burg, 2003), cellulases (Ando *et al.*, 2002; Kengen *et al.*, 1993) and DNA polymerases (Jones & Foulkes, 1989).

1.5.1 Features of thermophilic enzymes

Thermophilic enzymes are stable and active at elevated temperatures. These enzymes are useful in industrial processes because at elevated temperatures the solubility of many reaction components, in particular polymeric substrates, is significantly improved. The risk of contamination is reduced at high temperatures which avoids undesired complications (van den Burg, 2003). In bioethanol production, high temperatures eliminate the need for process cooling and the higher vapour pressure of ethanol at increased temperatures facilitates its removal by gas stripping (Taylor, 2007).

Thermophilic organisms possess heat stable enzymes as these enzymes have unique features which allow them to function at the elevated ambient temperatures. The structural features of thermophilic extremozymes have attracted much attention but are not well understood. Several three-dimensional structures have been solved by comparison with mesophilic counterparts. Analysis of the differences between the thermophilic and mesophilic homologues has highlighted factors that contribute toward protein thermostability (Sternier & Liebl, 2001; Vieille & Zeikus, 2001; van den Burg & Eijsink, 2002). These include greater hydrophobicity (more compact enzymes to exclude internal water), an increase in the number of amino acids with large branched and charged functional groups, smaller and fewer internal cavities, an increase in the number of residues in α helical conformation and the deletion and shortening of surface loops (De Simone *et al.*, 2001). There are also larger interfaces between subunits (Berezovsky & Shakhnovich, 2005). Lastly there is an increase in polar or charged interactions (hydrogen bonds and salt bridges) across the subunit interfaces and around active sites which contributes to the stability of thermophilic enzymes (Bae & Phillips, 2004).

1.5.2 Potential application of thermophilic enzymes for bioethanol production

Fermentation of lignocellulosic hydrolysates to produce ethanol is a temperature-dependent process (Figure 1.6). The process depends on the ability to utilize the high efficiency and specificity of enzyme catalysis to synthesize ethanol from a feedstock. The advantages of using thermophilic over mesophilic microorganisms for bioethanol production include higher growth and metabolic rates, decreased cellular growth yield, increased physicochemical stability of the catabolic enzymes and facilitated reactant activity and product recovery (Thomas *et al.*, 1981).

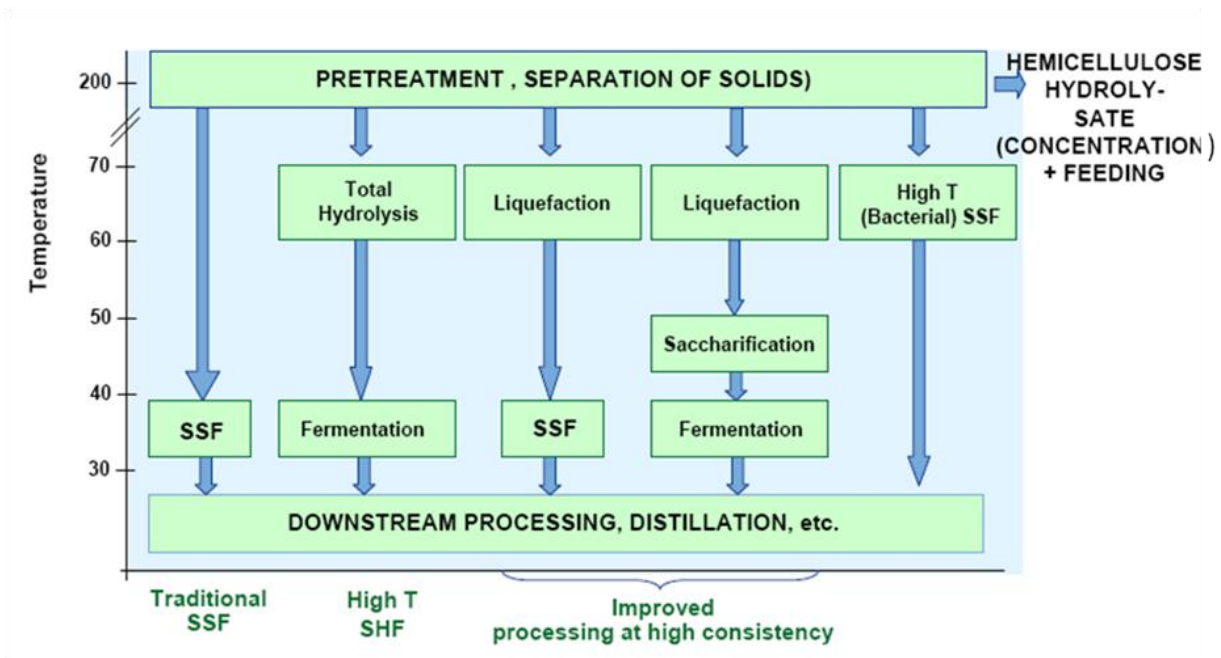


Figure 1.6 Temperature profiles of the lignocellulose-to-ethanol conversion processes

1.6 Metagenomics and gene discovery

There are a large number of microorganisms on the planet and the majority have not been cultured in the laboratory (Handelsman, 2004). Many approaches which are used to study the diversity and potential of microbial communities are biased due to the limitations of cultivation methods and physicochemical requirements such as temperature, pH, nutrient level, salinity etc. (Cowan *et al.*, 2005).

Metagenomics attempts to access the huge untapped resource of genetic material via culture-independent approaches (Steele & Streit, 2005). Furthermore, when coupled with protein evolution technologies, metagenomics can access new regions of protein sequence space and these techniques can ultimately be used to the search for the “ideal biocatalyst” (Cowan *et al.*, 2004).

1.6.1 Metagenomic technologies

The construction of a metagenomic library (Figure 1.7) is usually initiated by the extraction of total community DNA from an environmental sample (Daniel, 2005). Successful library construction depends on the efficiency of both the DNA extraction method employed (high molecular weight DNA and large yields are required) and the subsequent cloning techniques. After the isolation and purification of environmental DNA, the desired fractions are cloned into a suitable vector system, such as plasmids (Boubakri *et al.*, 2006; Lammle *et al.*, 2007), fosmids (Treusch *et al.*, 2004; Hardeman & Sjolting, 2007), cosmids (Voget *et al.*, 2006), bacterial artificial chromosomes (Beja *et al.*, 2000) and yeast artificial chromosomes (Beja, 2004). The advantage of the newer vectors (fosmids, cosmids and bacterial artificial chromosomes) is that they allow for the cloning of larger insert sizes which can include intact metabolic pathways, as has been reported for cloned gene clusters coding for the synthesis of valuable antibiotics (Brady *et al.*, 2001). *Escherichia coli* is the preferred host strain for library verification but recently *Streptomyces* species and *Bacillus* species have also been used as suitable hosts to identify genes of interest (Nakashima *et al.*, 2005).

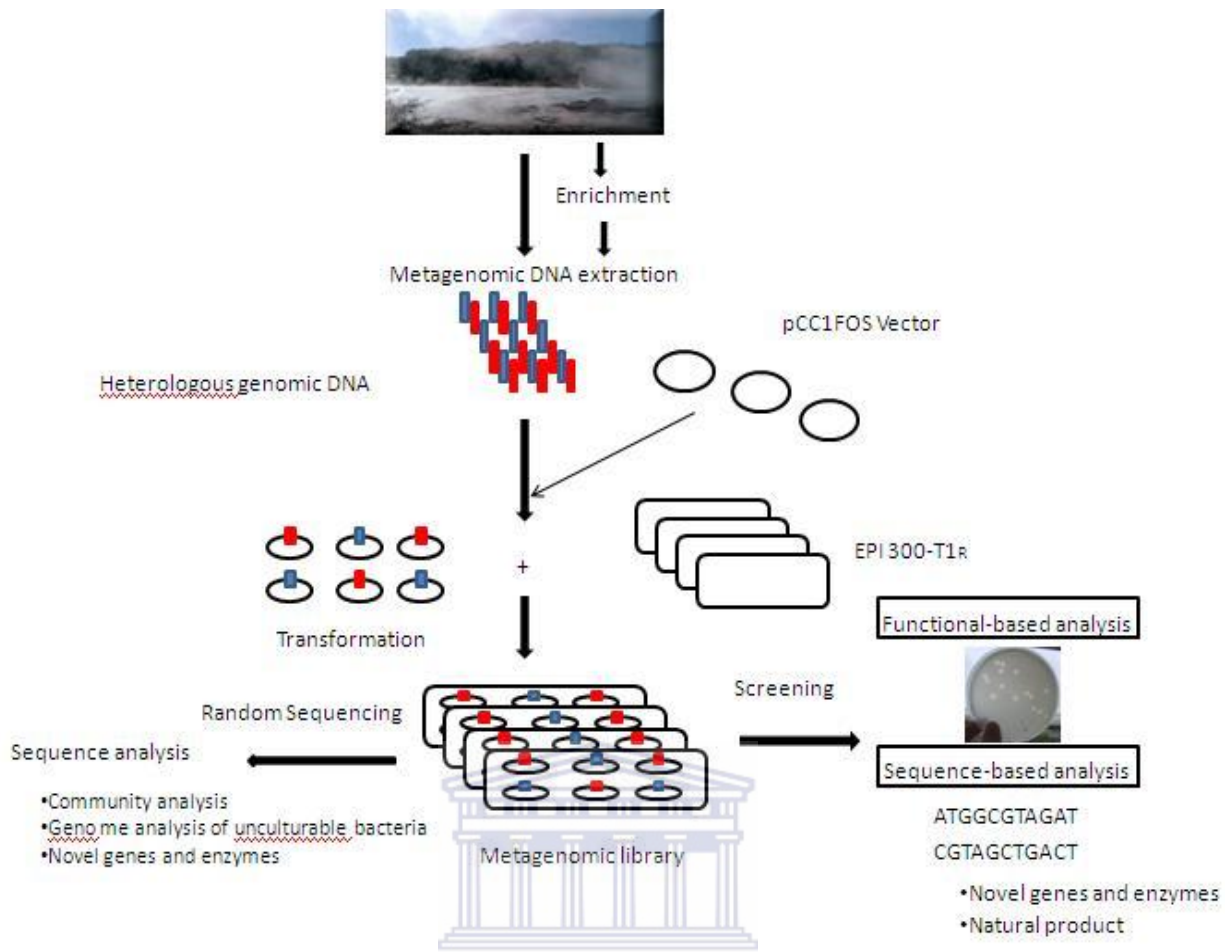


Figure 1.7 Construction of metagenomic libraries from environmental samples and methods for analyzing functions and sequences in the DNA library (Schmeisser *et al.*, 2007)

1.6.2 Gene discovery

The analysis of the genomes of uncultured microorganisms can not only explore the ecology of microbial communities, but can also be used in the discovery of novel biocatalysts and biomolecules (Schmeisser *et al.*, 2007). Metagenomes have been mined for a wide range of novel enzymes and biocatalysts, all of which have the potential for use in biotechnology and other industrial applications (baking, brewing, food and feed processes, detergents etc) (Lorenz & Schleper., 2002). The advantages of using biocatalysts obtained from natural habitats in industrial processes over chemical

counterparts include improved substrate specificity, lower cost of production and the ability to increase the sustainability of a process (Langer *et al.*, 2006).

A number of novel enzymes have been derived from metagenomes. The list includes lipases (Lee *et al.*, 2004), esterases (Elend *et al.*, 2006; Heath *et al.*, 2009), hydratases (Ferrer *et al.*, 2005), antibiotics such as turbomycin (Gillespie *et al.*, 2002) and even entire novel pathways for the degradation of xenobiotics (Boubakri *et al.*, 2006; Eyers *et al.*, 2004).

Several novel cellulolytic enzymes which have unique activities and/or sequences have been isolated, purified and characterized from metagenomic libraries (Rees *et al.*, 2003; Grant *et al.*, 2004; Feng *et al.*, 2007; Kim *et al.*, 2008; Duan *et al.*, 2009; Pang *et al.*, 2009). One of the earliest reported metagenome-derived cellulases was isolated from a thermophilic anaerobic digester fueled by lignocellulose (Healy *et al.*, 1995). A recently discovered cellulase derived from a soil metagenome is remarkably stable over a wide range of temperatures, pHs and in the presence of high salt concentrations (Voget *et al.*, 2006). Similarly several novel cellulase genes have been identified from different environmental genomic libraries (forest soil, dung of elephant, cow rumen and rotten tree remains) (Wang *et al.*, 2009). Sequence analysis of these environmental genomic libraries predicted that four endo- β -1,4-glucanases were members of the glycosyl hydrolase family5 (GHF5) and one endo- β -1,4-glucanase was a member of the glycosyl hydrolase family9 (GHF9). In addition two β -glucosidases belonged to glycosyl hydrolase family3 (GHF3) (Wang *et al.*, 2009).

1.7 Molecular techniques

1.7.1 Metagenomic DNA extraction from soil

Extraction of high molecular weight (HMW) metagenomic DNA from environmental samples is crucial for metagenomic library construction (Bertrand *et al.*, 2005). Large inserts decrease the number of clones needed to represent the community and provide better opportunity for recovery of full length open reading frames (Handelsman, 2005). Successful extraction depends on several parameters; however there are three major problems which need to be addressed. 1) DNA should represent the original microbial population from as broad a range of microorganisms as possible 2) the extracted DNA may shear and form high molecular weight chimeras and 3) the DNA must be pure enough to avoid contamination during downstream DNA processing such as restriction digestion and ligation (Schmeisser *et al.*, 2007). The choice of DNA extraction method is based on the type of sample and the purpose of the metagenomic study. The most widely used lysis methods are mechanical (bead beating or sonication) and chemical (detergents and enzymatic) lysis or a combination of both methods (Krsek & Wellington, 1999; Miller *et al.*, 1999).

Chemical lysis causes comparatively less DNA damage than mechanical methods. Nevertheless this method tends to be less effective for disrupting the soil matrix and exposing cells to the lysis buffer. Detergents such as sodium dodecyl sulphate (SDS) (Zhou *et al.*, 1996) or sarkosyl (Holben *et al.*, 1988) are used to aid cell membrane lysis. In addition, other compounds such as chelating agents (EDTA or Chelex 100) are added to inhibit nuclease activity and disperse the soil matrix (Miller *et al.*, 1999). Humic acid

complexing compounds such as polyvinylpyrrolidone (Gray & Herwig, 1996) and CTAB (Zhou *et al.*, 1996) are also used to increase DNA purity.

Phenol and /or chloroform extraction steps are used to recover the DNA from the soil and cell debris (Tebbe & Vahjen, 1993; Zhou *et al.*, 1996). After the DNA is recovered, ethanol, isopropanol and/or polyethyleneglycol (PEG) are used to precipitate DNA. Alcoholic precipitation may cause co-precipitation of humic acids which does not occur with PEG (Pang *et al.*, 2008). Good yields of DNA are achieved using isopropanol precipitation (Roose-Amsaleg *et al.*, 2001). Further purification maybe required after metagenomic DNA is isolated. Methods include caesium chloride density gradient ultracentrifugation, adsorption chromatography, agarose gel electrophoresis and in rare instances dialysis and filtration (Jacobsen & Rasmussen, 1992; Tebbe & Vahjen, 1993; Berthelet *et al.*, 1996; Stach *et al.*, 2001).

1.7.2 Screening of the metagenomic DNA libraries

Three methods have been used to screen metagenomic libraries 1) function/ activity-based screening 2) sequence-based screening and 3) substrate-induced gene-expression screening (SIGEX) (Yun & Ryu, 2005). They all have the potential for identification of clones carrying a specific gene (Daniel, 2005; Ferrer *et al.*, 2005).

Function/activity-based screening allows researchers to identify new classes of genes with useful functions. It is the only technique that enables scientists to discover new gene classes (Yun & Ryu, 2005). The advantage of functional screens is that they can be used to access single genes as well as multiple genes encompassing a complete metabolic pathway without prior knowledge of sequence data and thus may reveal novel genes and gene products unrelated to those currently known (Handelsman, 2004). However, activity based screening relies on the expression of genes in a heterologous host such as *E. coli*,

which may cause low detection incidence (Simon & Daniel, 2009). Host choice has been expanded to circumvent this problem and genetic tools have been developed for parallel studies in *Streptomyces lividans* and *Pseudomonas putida* to capture a wider range of expressed genes and proteins (Schmeisser *et al.*, 2007).

The sequence-based approach uses either PCR-based or hybridization-based procedures to detect genes homologous to those already known. It is a powerful tool for the identification of genes without the need to express the gene of interest in a host (Yun & Ryu, 2005). Sequence-based methods have been the driving force behind the development of many bioinformatics tools. However, as prior knowledge of the target sequence is required this approach is limited to the discovery of genes which are similar to those already known (Handelsman, 2005). Sequence-based metagenomics has driven the development of high-throughput sequencing technologies such as 454-pyrosequencing which has enabled entire communities to be sequenced (complete metagenome sequencing and assembly) (Schmeisser *et al.*, 2007).

SIGEX is an intracellular screening method, which is used to identify novel catabolic genes, particularly genes that are difficult to reveal using conventional gene-cloning methods. Operon-trap green fluorescence protein expression vectors have been introduced, into which environmental DNA is cloned. A library is then constructed in a liquid culture by transforming to a host strain such as *E. coli*. Positive clones will express green fluorescence protein when the target substrate is present (Kimura, 2006). Table 1.3 summarises the advantages and disadvantages of the metagenomic screening methods.

Table 1.3 Advantages and disadvantages of metagenome screening methods (Uchiyama & Watanabe 2008)

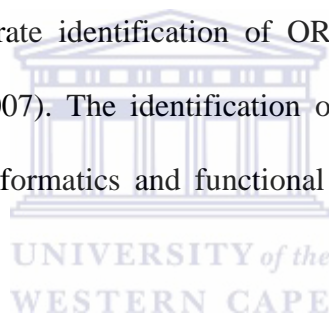
Method	Description	Advantage	Limitation
Nucleotide sequence-base screening	Primer and probes used for screening are designed from known gene sequences (mostly those cloned from easily cultivable bacteria)	High-throughput PCR cloning is possible	Only genes homologous to known genes can be obtained
Enzyme activity-based screening	An activity expressed by a transformed host cell (eg., an enzyme activity) is detected and used for selecting positive clones	Gene fragments that are sufficient to express enzymatic activities can be obtained	Many enzymes are difficult to be expressed in a heterogeneous host as an active form
Gene expression-based screening SIGEX	A gene-expression activity of a metagenome fragment in a cloning host is detected using an activity of co expressed marker encoded in a cloning vector	High-throughput fluorescence flow sorting is possible	It is generally laborious. Catabolic genes that are distant from a relevant transcriptional regulator cannot be obtained. Genes obtained may be partial

1.7.3 Metagenomic sequencing

DNA sequencing is the determination of the order of the nucleotide bases adenine, guanine, cytosine, and thymine in sample DNA. The Sanger method relies on random incorporation of chain terminating nucleotides in a capillary matrix. During the past three decades, Sanger sequencing has been used in large-scale production sequencing (Sanger

& Coulson, 1975; Hall, 2007). More recently, demand for faster and larger scale sequencing has led to the development of high-throughput sequencing methods or so called next-generation sequencing technologies. The 454 sequencing technology relies on detection of pyrophosphate release on nucleotide incorporation, which is based on the “sequencing by synthesis” principle. Using this 454 sequencing technology allows for the sequencing of 400-600 million base pairs with 400-500 base pair reading lengths (Wicker *et al.*, 2006).

A major goal of a metagenomic sequencing project is to identify novel genes. Metagenomic samples require fast and accurate sequencing methods. Some challenges encountered during the processing include the assembly and analysis of the short sequence reads (30-50bp), accurate identification of ORFs and assigning function to sequence fragments (Pachter, 2007). The identification of novel genes will be greatly influenced by advances in bioinformatics and functional genomics (Schmeisser *et al.*, 2007).



1.8 Aims and objectives of the current study

Broadly the aim of this project was to characterise novel thermophilic cellulolytic genes and enzymes for lignocellulose digestion.

The specific objectives of this study were:

- To construct a fosmid library from metagenomic DNA (40kb insert size).
- To screen for cellulase activities using a variety of functional screens.
- Based on the functional screen, genes of interest would be sequenced, cloned, expressed, purified and characterized



Chapter 2 General materials and methods

2.1 General laboratory chemicals and reagents

Unless otherwise specified, chemicals were supplied by Merck Chemicals and Laboratory Supplies (Darmstadt, Germany), Sigma Aldrich Chemical Company (Deisenhofen,

Germany) and Kimix Chemical and Laboratory Supplies (South Africa). Oxoid Ltd and Biolabs supplied culture media.

DNA size markers, protein size markers and all DNA modifying enzymes (polymerases and restriction endonucleases) were purchased from Fermentas Life Sciences Ltd (Vilnius, Lithuania).

Oligonucleotides for polymerase chain reaction (PCR) used in this study were synthesized by Inqaba Biotech (Johannesburg, South Africa).



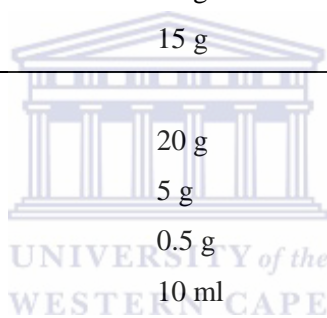
2.2 Media

The media used are listed in Table 2.1. The recipes are from Sambrook and Russel (2001) unless otherwise stated. All media was autoclaved at 121°C for 20 mins unless otherwise specified.

Table 2.1 Growth media used in the study

Constituent	1 litre final volume
LB Medium (Luria-Bertani Medium)	
Tryptone	10 g
Yeast extract	5 g

NaCl	10 g
2x YT Medium (pH 7.0)	
Tryptone	16 g
Yeast extract	10 g
NaCl	5 g
LB agar	
Tryptone	10 g
Yeast extract	5 g
NaCl	10 g
Agar	15 g
CMC LB agar	(Feng <i>et al.</i> , 2007)
CMC	10 g
Tryptone	10 g
Yeast extract	5 g
NaCl	10 g
Agar	15 g
SOB Medium	
Tryptone	20 g
Yeast extract	5 g
NaCl	0.5 g
KCl (250 mM)	10 ml
MgCl ₂ (2 M)	5 ml (filter sterilized and added before use)



SOC Medium	
Tryptone	20 g
Yeast extract	5 g
NaCl	0.5 g
KCl (250 mM)	10 ml
MgCl ₂ (2 M)	5 ml (filter sterilized and added before use)
Glucose (1 M)	20 ml (same with MgCl ₂)
M9 salt solution (pH 7.4)	
KH ₂ PO ₄	3 g
NaCl	0.5 g
Na ₂ HPO ₄ ·7H ₂ O	12.8 g
NH ₄ Cl	1.0 g
M9 Minimal Media	
M9 salt solution	200 ml
MgSO ₄	2 mM
Carbon source	20 ml of 20%
CaCl ₂	0.1 mM

The media were supplemented with antibiotics (Table 2.2) after autoclaving and cooling to 55°C where necessary.

Table 2.2 Stock and final concentrations of antibiotics used in the study

Antibiotics	Preparation
Carbenicillin (CAB)	50 mg/ml in distilled H ₂ O stock 50 µg/ml final concentration
Chloramphenicol (CAM)	34 mg/ml in 100% ethanol stock 34 µg/ml or 12.5 µg/ml final concentration
Kanamycin (KAN)	50 mg/ml in distilled H ₂ O stock 50 µg/ml final concentration
Ampicillin (AMP)	50 mg/ml in distilled H ₂ O stock 50 µg/ml final concentration

2.3 Bacterial strains and plasmids

The bacterial strains and plasmids used in the study are listed in Table 2.3 and Table 2.4.

Table 2.3 Bacterial strains

Bacterial strain	Relevant Genotype	Supplier
<i>E. coli</i> Gene Hog	F- mcrA Δ(mrr- hsdRMS- mcrBC) φ80lacZ M15 ΔlacX74 recA1 araD139 Δ(ara- leu 7697 galU galK rpsL (StrR) endA1 nupG	Invitrogen (USA)
<i>E. coli</i> Rosetta (DE3 pLysS	F- ompT hsdSB(rB- mB-) gal dcm (DE3)pLysSRARE (Cam ^R)	Novagen USA)
<i>E. coli</i> EPI300	F- mcrA Δ(mrr-hsdRMS- mcrBC) f80dlacZΔM15 ΔlacX74 recA1 endA1 araD139 Δ(ara, leu)7697 galU galK 1- rpsL nupG trfA	Epicentre Biotechnology (USA)

Table 2.4 Plasmids used in the study

Plasmid	Description	Source
pCC1FOS	Fosmid vector containing a chloramphenicol resistance gene, linearized at the <i>Eco</i> 72 I (blunt) site, dephosphorylated	Epicentre
pGEM-T Easy	Cloning vector containing an ampicillin resistance gene, with T overhangs at the insertion site	Promega
pET 21a	Expression vector containing an ampicillin resistance gene and a C- terminal His-tag	Novagen
pET28a	Expression vector containing a kanamycin resistance gene and N-terminal and C-terminal His-tags	Novagen

2.4 DNA extraction

2.4.1 Alkaline cell lysis method for plasmid DNA extraction

Single colonies were picked from agar plates and inoculated into 5 ml LB medium supplemented with the appropriate antibiotic(s). Inoculated cultures were incubated overnight at 37°C with shaking (150 rpm). Plasmid DNA was isolated from the overnight culture using an alkaline lysis method (Sambrook & Russell, 2001) with minor modifications. A volume of 2 ml of overnight culture was transferred into 2 ml microcentrifuge tubes and the cells were harvested by centrifugation at 5000 x g for 10 mins at room temperature. The supernatant was discarded and the pellet resuspended in 200 µl of solution 1 (50 mM glucose, 25 mM Tris-HCl pH8.0 and 10 mM EDTA pH8.0) containing RNase A to a final concentration of 20 µg/ml. A volume of 200 µl of solution 2 (1% [w/v] SDS and 0.2 M NaOH) was added and the tube contents were mixed by inversion and incubated for 5 mins at room temperature. Following the addition of 200 µl of 7.5 M ammonium acetate (pH5.5), the tubes were incubated on ice for 15mins and centrifuged at 13,000 x g for 20 mins at room temperature. The supernatant was transferred to new tubes and the plasmid DNA was precipitated by the addition of 0.7volume of isopropanol. The tubes were incubated at -20°C for 30 mins and centrifuged at 13,000 x g for 30 mins at 4°C. The DNA pellet was dried and resuspended in 1xTE buffer (10 mM Tris-HCl, 1 mM EDTA, pH8).

2.4.2 Plasmid extraction for sequencing quality DNA

Cultures were incubated overnight at 37°C with shaking (150 rpm) in LB medium (5 ml) in the presence of the appropriate antibiotic, typically 50 µg/ml CAB. Cells were harvested by centrifugation at 6000 x g for 10 mins. Plasmid DNA was extracted using the Invisorb Spin Plasmid Mini Two kit (Invitek, Germany). The plasmid isolation was

carried out according to the manufacturer's instructions. Plasmids were visualised by electrophoresis in 1% agarose gels (Section 2.5.4) prior to sequencing.

2.4.3 Fosmid extraction

Selected fosmid-containing strains were inoculated into 5 ml LB CAM and incubated with shaking at 37°C overnight. A volume of 1 ml of the culture was inoculated into 9 ml LB medium containing 12.5 µg/ml CAM and 10 µl induction solution (EPICENTRE®). Cultures were grown at 37°C with agitation for 5 hrs, and centrifuged at 6000 x g for 30 mins at 4°C. The supernatant was discarded (blotting or recentrifugation was used to remove trace quantities of supernatant). Cells were resuspended in 400 µl of cooled GET buffer (50 mM glucose, 10 mM EDTA, 25 mM Tris-HCl, pH 8.0) and 10 µl of 10 mg/ml RNase A (Fermentas) was added. A volume of 400 µl of lysis solution (0.2 M NaOH, 1% w/v SDS) was added and the tubes were incubated at room temperature for 5 mins. A volume of 400 µl of 3 M potassium acetate (pH 5.5) was added and cells were incubated on ice for 20 mins. The tubes were centrifuged at 16000 rpm at room temperature (RT) for 20 mins. Supernatants were transferred into fresh tubes, 0.7 volumes of isopropanol was added and the solution incubated at -20°C for 30 mins to promote DNA precipitation. The tubes were centrifuged at 16000 rpm at room temperature for 30 mins, the supernatant was discarded and the DNA pellets were washed with 70% v/v ice-cold ethanol. Pellets were air dried in a laminar flow cabinet and the DNA was resuspended in 20 µl of 1xTE (pH 8).

2.5 Analytical procedures

2.5.1 Spectrophotometry

Direct DNA concentration and purity readouts were obtained from the NanoDrop® ND1000's (NanoDrop Technologies, USA) nucleic acid sample screen.

2.5.2 Fluorometry (Qubit™)

Plasmid DNA concentrations were measured using the Quant-iT™ dsDNA BR Assay Kit (Invitrogen) according to the manufacturer's instructions. All reagents for DNA assays were used at room temperature. Readings were taken using a Qubit™ fluorometer.

2.5.3 Quantification of fosmid DNA

Fosmid DNA was quantified by agarose gel electrophoresis using λ DNA as standards (100 ng, 75 ng, 50 ng, 20 ng and 10 ng of λ). A volume of 1 μ l of 6x loading dye (30% v/v glycerol, 0.25% w/v bromophenol blue) was added to 5 μ l of DNA standard and loaded onto an agarose gel. Fosmid DNA solutions were prepared in a similar manner and at suitable dilutions for comparison.

2.5.4 Agarose gel electrophoresis

Electrophoresis was used to separate nucleic acid fragments. Genomic and plasmid DNA and PCR amplicons were visualised by the addition of 6x loading buffer (30% v/v glycerol, 0.25% w/v bromophenol blue) and subsequent electrophoresis in 1% or 0.7% (w/v) agarose gels prepared in 1xTAE buffer containing 0.5 μ g/ml ethidium bromide (Sambrook and Russell, 2001). DNA molecular markers of an appropriate size distribution were used for molecular weight comparisons. Gel images were visualised and

photographed using a digital imaging system (AlphaImager 2000, Alpha Innotech, San Leadro, USA).

2.6 Molecular biology techniques

2.6.1 Restriction enzyme digestion

Restriction enzyme digestions were prepared in sterile 1.5 ml microcentrifuge tubes in 10 – 50 µl reaction volumes and were incubated at 37°C overnight. Approximately 1 U of enzyme was used per µg of plasmid or genomic DNA in the presence of the appropriate buffer as supplied by the manufacturer. Restriction enzymes were inactivated at 80°C for 20 mins.

2.6.2 DNA ligation

Ligations were carried out in 10 µl volumes. To each microcentrifuge tube insert DNA and an appropriate cloning vector in a 2:1 or 3:1 ratio were combined with 1 U of T4 DNA ligase and 1x ligation buffer (Sambrook & Russell, 2001).

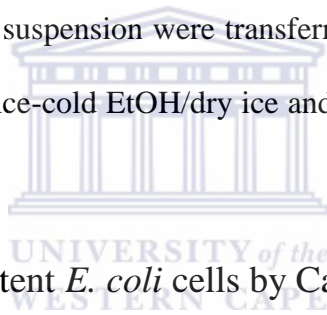
Reactions were incubated at 22°C overnight. Ligation reactions were transformed directly into host cells.

2.6.3 Preparation of competent cells

2.6.3.1 Preparation of electrocompetent *E. coli* cells

Glycerol stocks of appropriate *E. coli* cultures were streaked onto the surface of an LB agar plate. The plate was incubated for 24 hrs at 37°C. Pre-culturing was performed by transferring a single colony into 10 ml LB medium. The culture was incubated overnight at 37°C in a shaking incubator. A volume of 10 ml of the overnight culture was inoculated

into a 5 L flask containing 1 L 2xYT media and incubated with shaking for 3.5-4 hrs to an optical density at 600 nm of 0.6-0.9. The flask was placed on ice and the contents were divided into 4 equal volumes in ice-cold centrifugation bottles. The cultures in the centrifuge bottles were centrifuged at 4°C for 25 mins at 4000 rpm. The supernatant was discarded and the pellet resuspended in 200 ml sterile ice cold distilled water and centrifuged at 4°C for 25 mins at 4000 rpm. The previous step was repeated but the volume of ice cold distilled water was reduced to 100 ml. The supernatant was discarded and the pellets resuspended in 20 ml ice cold 15% v/v glycerol and 2% w/v sorbitol, and centrifuged at 4°C at 4000 rpm for 10 mins. The centrifuge tubes were placed on ice, the supernatant discarded and the pellet resuspended in 1ml ice cold 15% v/v glycerol and 2% w/v sorbitol. Aliquots of cell suspension were transferred into microcentrifuge tubes, snap frozen in liquid nitrogen or ice-cold EtOH/dry ice and stored at -80°C (Sambrook & Russell, 2001).



2.6.3.2 Preparation of competent *E. coli* cells by CaCl₂ treatment

Glycerol stocks of appropriate *E. coli* cultures were streaked onto the surface of an LB agar plate. The plate was incubated for 24 hrs at 37°C. Pre-culturing was performed by transferring a single colony into 5 ml LB medium. The culture was incubated overnight at 37°C in a shaking incubator and 500 µl of the overnight culture was inoculated into 100 ml 2xYT medium in a 1 L flask. The culture was incubated at 37°C until an optical density (OD at 600 nm) of 0.3-0.6 was attained. The flask was placed on ice and the culture was split into 4 equal volumes. Cells were kept on ice in all subsequent steps. The cultures were centrifuged at 4°C for 5 mins at 5000 rpm. The supernatant was discarded and the pellet was resuspended in 100 ml ice cold 0.1 M CaCl₂ and held on ice for 1 min. Cells were collected as before and resuspended in 50 ml of ice cold 0.1 M CaCl₂ and held

on ice for 90 mins. The cultures were centrifuged at 4°C at 5000 rpm for 5 mins and placed on ice. The supernatant was discarded and the pellet was resuspended in 10 ml ice cold 0.1 M CaCl₂. A volume of 10 ml of ice-cold sterile glycerol was added, the cells were resuspended, and aliquots were stored at -80°C (Sambrook & Russell, 2001).

2.6.4 Transformation of competent *E. coli* cells

2.6.4.1 Transformation by electroporation

Electrocompetent *E. coli* cells (Section 2.6.3.1) were transformed with 1-10 ng purified DNA. A microcentrifuge tube containing 50 µl of electrocompetent cells was removed from -80°C storage and allowed to thaw on ice. Ligation mixture (2 µl) (Section 2.6.2) was added to the thawed cells and gently mixed. The mixture was pipetted into a pre-chilled 0.1 cm electroporation cuvette (Biorad). Electroporation conditions using a BioRad Gene pulser were 1.8 KV, 15 µF and 200 Ω. After electroporation the cells were incubated in 1 ml of SOC medium for 1 hr at 37°C and 150 rpm. A volume of 100 µl of cells was plated onto CMC LB agar plates containing 12.5 µg/ml CAM and incubated at 37°C overnight.

2.6.4.2 Transformation by heat shock

Approximately 1-10 ng of purified DNA was added to 100 µl of chilled chemically competent *E. coli* cells (Section 2.6.3.2). The mixture was left on ice for 10 mins and heat shocked at 37°C for 5 mins. After incubation on ice for 1 min, the mixture was incubated in 1 ml 2xYT for 1 hr at 37°C (150 rpm) and the cells were plated onto CMC LB agar plates containing 12.5 µg/ml CAM and incubated at 37°C overnight.

2.6.5 Sequencing

Cloned insert DNA was sequenced using the sequencing facility (ABI PRISM 377 automated DNA sequencer) at the University of Stellenbosch, South Africa.

2.6.6 Polymerase chain reaction

Polymerase Chain Reaction (PCR) amplifications were performed in 0.2 ml thin-walled tubes using a thermocycler equipped with a heated lid (Thermo Hybrid PCR Sprint Temperature Cycling System). A standard 20 μ l reaction contained PCR buffer (20 mM Tris pH8.8, 10 mM KCl, 10 mM $(\text{NH}_4)_2\text{SO}_4$, 2 mM MgSO_4 , 0.1% Triton X-100), 0.4 mM each of dNTP (dATP, dCTP, dGTP and dTTP), 0.5 μ M of each primer, an appropriate amount of *Taq* polymerase and 25 or 100 ng of DNA as template. The primers used and the cycling conditions are given in Table 2.5.



Table 2.5 Primers used in the study

Genes Amplified	Primers (5' to 3')	PCR amplification cycle	Reference
Gene12FW	CATATGCGAAAACCCGTCTGCGC	94°C/4 mins	This study
Gene12XhoI RV	GCTCGAGTTGCTGGGCGATCCACACCAG	30x (94°C/30 s-63°C/30 s-70°C/105 s), 70°C/10 mins	
Gene25FW	CATATGAAAGCCTTTCGCTGGTG	94°C/4 mins	This study
Gene25XhoI RV	GCT CGA GCT GTG AAC TTG CGC	30x (94°C/30 s-57°C/30 s-70°C/105 s), 70°C/10 mins	
M13 FW	CCCAGTCACGACGTTGTAAAACG	94°C/10 mins	Messing, 1983
M13 RV	AGCGGATAACAATTCACACAGG	35x (94°C/30 s-64°C/30 s-72°C/60 s), 72°C/10 mins	
T7 Promoter	TAATACGACTCACTATAGGG	95°C/2 mins	Novagen 69348-3
T7 Terminator	GCTAGTTATTGCTCAGCGG	35x (95°C/30 s-50°C/30 s-72°C/3.5 mins), 72°C/5 mins	
Bacterial 16S rRNA	E9F GAGTTTGATCCTGGCTCAG	94°C/4 mins	Hansen <i>et al.</i> , 1998
	U1510R GGTTACCTTGTTACGACTT	30x (94°C/30 s-52°C/30 s-72°C/105 s), 72°C/10 mins	Reysenbach & Pace, 1995
	341F-GC CGCCCGCCGCGCGCGGGCGGGGCGGG	94°C/5 mins 20cycles: 94°C/45 s, 65°C touchdown to 55°C/30 s, 72°C/1 min	Muyzer <i>et al.</i> , 1993
	GGCACGGGGGGCCTACGGGAGGCAGCAG	20cycles: 94°C/30 s, 55°C/30 s, 72°C/1 min	
	534R ATTACCGCGGCTGCTGG	72°C/20 mins	

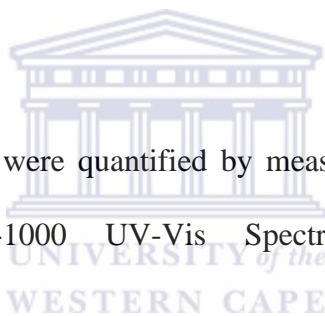
2.7 Protein analysis techniques

2.7.1 Bradford assay

The Bradford assay (Bradford, 1976) was used to determine the concentration of protein. Bovine serum albumin (Sigma-Aldrich) was used as a standard. The samples were prepared as follows: 10 μl of protein sample/standard and 790 μl of H_2O were added to a cuvette. A volume of 200 μl of Bradford reagent was added. The mixture was incubated at room temperature for 15 mins, and the absorbance at 595 nm was recorded. Distilled water was substituted for the protein/standard to serve as the blank.

2.7.2 Nanodrop analysis

Purified protein preparations were quantified by measuring the A_{280} absorbance using a NanoDrop ND-1000 UV-Vis Spectrophotometer (NanoDrop Technologies, USA).



2.7.3 SDS-PAGE

Sodium dodecyl sulphate polyacrylamide gel electrophoresis (SDS-PAGE) was used to separate proteins according to their size.

2.7.3.1 Gel preparation

SDS-PAGE gels were cast with a 12% separating gel and a 4% stacking gel (Table 2.6).

Table 2.6 Preparation of 12% separating gels and 4% stacking gels for SDS-PAGE

12% separating gels		4% stacking gels	
Reagents	Volume (ml)	Reagents	Volume (ml)
Sterile water	3.4	Sterile water	3.075
1.5 M Tris –HCL pH8.8	2.5	0.5 M Tris –HCL pH6.8	1.250
20% SDS (w/v)	0.05	20% SDS (w/v)	0.025
30% Acrylamide: 0.8% Bisacrylamide (w/v)	4.0	10% Ammonium persulfate	0.670
10% Ammonium persulfate	0.05	10% Ammonium persulfate	0.025
TEMED	0.005	TEMED	0.005

2.7.3.2 Preparation of protein samples

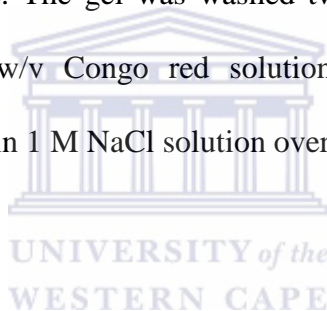
An equal volume of 2 x loading buffer (80 mM Tris-HCl pH6.8, 10% v/v mercaptoethanol, 2% w/v SDS, 10% w/v glycerine, 0.2% w/v bromophenol blue) was added to protein samples. The samples were heated at 95°C for 10mins and vortexed vigorously. The samples were either loaded immediately onto SDS-PAGE gels or stored at -20°C.

2.7.3.3 Electrophoresis of protein samples

Denatured samples (Section 2.7.3.2) were electrophoresed at a constant voltage of 60 V through the stacking gel and 120 V through the separating gel. On completion of the electrophoresis run the gel was stained for 30 mins in Coomassie staining solution (0.25% w/v Coomassie Brilliant Blue R250, 45% v/v methanol, 45% v/v distilled water, 1% v/v acetic acid) and destained with destaining solution (20% v/v methanol, 10% v/v acetic acid and 70% v/v distilled water) overnight (Sambrook & Russell, 2001).

2.7.4 Zymogram

A zymogram was prepared for the determination of cellulase activity, according to the method of Watson *et al.* (2009) with modifications. Samples were prepared and run as in Section 2.7.3.3 in SDS-PAGE gels containing CMC (1% w/v). Samples were run in duplicate and the gel cut into two portions after electrophoresis. One portion was stained and destained as described in Section 2.7.3.3. The other portion was washed twice in distilled water and incubated in 30ml of refolding buffer (20 mM piperazine-*N,N*-bis(2-ethanesulfonic acid (PIPES) buffer (pH6.8), 2.5% Triton X-100, 2 mM dithiothreitol, 2.5 mM CaCl₂) for 4hrs at room temperature. The gel was washed twice in distilled water and further incubated in 0.1% w/v Congo red solution for 2 hrs. Activity was visualized by soaking the gel in 1 M NaCl solution overnight with agitation.



Chapter 3 Metagenomic fosmid library construction and functional screening for cellulase activity

3.1 Introduction

Molecular analysis has revealed that the microbial diversity of the biosphere is more complex than previously imagined, with at least 99% of microbial species

present still being uncultured (Handelsman, 2004). Metagenomics can be defined as an approach to the genomic analysis of microbial communities present in a given habitat without the need for cultivation of the microorganisms (Amann *et al.*, 1995; Handelsman, 2005). Metagenomic analysis involves the direct extraction of total community DNA from environmental samples and further analysis of the metagenomic DNA using two approaches: 1) sequence based analysis which is used to study the phylogenetic diversity of complex microbial life and community structure within an environmental sample and 2) functional based analysis which is used to identify functional genes and enzymes by screening, cloning, expression and characterization. The availability of high accuracy, low cost, high throughput sequencing technologies has increased the ability to efficiently produce data for metagenomic sequencing projects for both of the approaches discussed above (Simon & Daniel, 2009).

Total community DNA extracted from a sample is fragmented and fragments of the desired size are cloned into appropriate vector systems. Effective cloning strategies and high cloning efficiencies are required due to the large size of the metagenomes and the need for coverage of the metagenome (Schmeisser *et al.*, 2007). Various vector systems including plasmids (Boubakri *et al.*, 2006; Lammle *et al.*, 2007), fosmids (Treusch *et al.*, 2004; Hardeman & Sjoling, 2007), cosmids (Voget *et al.*, 2006), bacterial artificial chromosomes (Beja *et al.*, 2000) and yeast artificial chromosomes (Beja, 2004) are used in metagenomic studies.

Fosmids are plasmids that use the F-plasmid origin of replication and partitioning mechanisms and allow the cloning of large DNA fragments. Copy control fosmid vectors contain both the bacterial F-factor single-copy origin of replication and an inducible high-copy ori-V. Copy control fosmid clones are initially grown at

single copy to ensure the stability of the insert DNA but can be induced to high-copy number as needed (CopyControl™ Fosmid Library Production Kit manual).

Cellulases have been studied comprehensively and more than a thousand cellulases have been identified from cultured microorganisms (Duan & Feng, 2010). Cellulases isolated from specific organisms tend to demonstrate traits of activity corresponding to the prevalent environmental conditions from where the organism was isolated. The temperature and pH optima of the enzyme resemble the ambient temperature and pH of the soil, water etc. from which they originate (Voget *et al.*, 2006). Metagenomic approaches have been widely used to isolate cellulases with varying properties from diverse environmental samples with corresponding properties. These include acidiphilic cellulases (active at low pH) which can be used for pre-treatment of lignocellulosic waste material and alkaliphilic cellulases (active at high pH) which can be used as detergent additives (Ferrer *et al.*, 2007; Fujinami & Fujisawa, 2010).

Functional based screening of metagenomic libraries relies on gene expression in heterologous host organisms and on simple and reliable protocols for screening of the libraries. Screening for cellulase producing microorganisms and clones is routinely done by flooding colonies grown on CMC LB agar plates with a solution of Congo red and fixing the colour with a solution of NaCl (Feng *et al.*, 2007). Congo red interacts strongly with intact polysaccharides such as (1-4) β -D-glucans, (1-3) β -D-glucans, and (1-4) β -D-xylans (Teather & Wood, 1982). Colonies capable of hydrolysing cellulose produce a clear zone around the colony. This chapter describes the construction of a fosmid library from the Mphizi hot springs, Malawi soil metagenomic DNA and the functional screening of the library with Congo red for cellulase activity.

3.2 Materials and methods

3.2.1 Sample information

Soil samples were collected from the Mphizi hot springs, Chiweta, Malawi during March 2009 by Professor Don Cowan. The temperature, pH and GPS co-ordinates of the sampling site were recorded (Table 3.1). All samples were stored below 0°C in the field and stored at – 80°C in the laboratory until analysed.

Table 3.1 Location and characteristics of the Mphizi hot springs, Malawi

Characteristic	Sample information
Temperature	72-78°C
pH	6.2
GPS coordinates	S° 10.682207 E° 34.185756

3.2.2 DNA extraction from soil

Soil samples were partially thawed at -20°C for at least 2 hrs and then thawed at 4°C overnight. Metagenomic DNA was extracted from the soil sample based on a method described previously (Zhou *et al.*, 1996) with the following modifications. Soil samples were suspended in an equal volume of extraction buffer containing protease K to a final concentration of 1 mg/ml. The suspension was incubated at 37°C with horizontal shaking (150 rpm) for 30 mins. Following the addition of 20% SDS (1% final volume), the tubes were incubated at 65°C in a water bath for 2 hrs with gentle inversion every 30 mins. After centrifugation at 6000 x g for 10 mins at room temperature the aqueous supernatant was transferred into microcentrifuge tubes. An equal volume of phenol: chloroform: isoamyl alcohol (25:24:1) was added and the tubes were centrifuged at 16,000 x g for 15 mins. The

aqueous phase was transferred to fresh microcentrifuge tubes and an equal volume of chloroform was added. After centrifugation at 16,000 x g for 10 mins, the DNA was precipitated with 0.6 volumes of isopropanol at room temperature overnight. The DNA pellet was obtained by centrifugation at 16,000 x g for 20 mins at 10°C, and was washed with 70% ethanol. The pellet was air dried in a laminar flow hood. Finally the DNA was resuspended in 50 µl of 1xTE buffer.

3.2.3 Size fractionation and DNA purification

3.2.3.1 Size selection

A 0.7% w/v agarose gel was prepared (Section 2.5.4) and a plug was cut out of the gel with a sterile surgical blade and filled with 0.7% low melting point agarose. The total extracted high molecular weight DNA, Lambda- *Hind*III molecular weight marker and fosmid control DNA (Epicentre, USA) were loaded onto the gel. The gel was electrophoresed at 30 V for 16 hrs. Both marker-containing lanes were cut out from the gel and stained in TAE-EtBr buffer (EtBr 5 mg/ml, 1xTAE) for 20 mins. Markers were viewed with a digital imaging system (AlphaImager 2000, Alpha Innotech, USA), and the 23 kb and 40 kb fragments were labelled with a sterile toothpick. The gel was reassembled and gel fragments of between 23 kb to 40 kb in size were excised from the gel with a scalpel blade and transferred to a pre-weighed 1.5 ml microcentrifuge tube and stored at 4°C.

3.2.3.2 Purification of DNA from the LMP agarose gel

The agarose plugs containing DNA fragments of the desired size (Section 3.2.3.1) were melted at 65°C for 10 mins, and transferred to a 42°C water bath for 5 mins.

One unit of agarase for every 100 mg of LMP agarose was added to each tube and the tubes were incubated at 42°C for 90 mins. The tubes were incubated at 70°C for 10 mins, chilled on ice for 5 mins and centrifuged at 10,000 x g for 10 mins. The supernatant was transferred to clean tubes and the DNA was precipitated with 0.1 volume of 3 M sodium acetate (pH7) and 2.5 volumes of ice cold absolute ethanol at -20°C overnight. All DNA was pooled into one tube.

3.2.4 Cloning of high molecular weight DNA

3.2.4.1 End repair of DNA

DNA was end-repaired to generate 5' – phosphorylated blunt-ended DNA fragments. The following reagents were combined on ice to a total volume of 80 µl and thoroughly mixed: sterile ddH₂O, 8 µl end-repair buffer (330 mM Tris-acetate [pH7.8], 660 mM potassium acetate, 100 mM magnesium acetate, 5 mM DTT), 8 µl 2.5 mM dNTP mix (2.5 mM each of dATP, dCTP, dGTP, dTTP), 8 µl 10 mM ATP, 4 µl end-repair enzyme mix (T4 DNA Polymerase and T4 Polynucleotide Kinase) and up to 20 µg insert DNA.

The reaction mixture was incubated at room temperature for 2 hrs and a volume of 320 µl of distilled H₂O was added. An equal volume of phenol: chloroform: isoamyl (25:24:1) was added and the tube was centrifuged at 13,000 x g for 2 mins. The aqueous phase was transferred to a sterile eppendorf tube, and an equal volume of chloroform was added. After centrifugation at 13,000 x g for 2 mins the DNA was precipitated with 1/10 volume of potassium acetate (pH4.8) and 2.5 volumes of 100% ethanol at -20°C overnight. The DNA pellet was obtained by centrifugation at 16,000 x g for 30 mins at 10°C and was washed with 70% ice-cold ethanol. The pellet was air dried in a laminar flow hood. Finally the

DNA was resuspended in 10 μ l of 1xTE (pH8) and quantified by fluorometry (Section 2.5.2).

3.2.4.2 DNA ligation

The ligation reaction was carried out as described in Section 2.6.2. A 250 ng volume of insert DNA and 1 μ l of CopyControl pCC1FOS vector (0.5 μ g/ μ l) were used in the ligation reaction. A volume of 1 μ l of 10 mM ATP was included in the reaction mixture.

3.2.5 Phage packing of the fosmid clones

The EPI300-T1^R plating strain from EpiFOSTM Fosmid Library Production Kit is supplied as a glycerol stock. Cells were streaked on to a LBA plate and incubated at 37°C overnight. A single colony was inoculated into 5 ml LB broth and the culture was grown at 37°C overnight. For the phage packaging, 5ml of the EPI300-T1^R overnight culture was inoculated into 50 ml of LB broth supplemented with 10 mM MgSO₄. This culture was incubated at 37°C until an OD₆₀₀ of 0.8-1.0 was obtained.

One tube of the MaxPlax Lambda packing extract per ligation reaction was thawed on ice. A 25 μ l volume was immediately transferred into a sterile 1.5 ml microcentrifuge tube and kept on ice. The rest was returned to -80°C. A volume of 10 μ l of the ligation reaction was added to thawed extract and mixed by pipetting. The reaction was incubated at 30°C for 90 mins after which time the remaining 25 μ l of extract was added. The reaction was incubated for a further 90 mins at 30°C. Phage dilution buffer (10 mM Tris-HCl, pH8.3, 100 mM NaCl and 10 mM MgCl₂) was added to a final volume of 1 ml. A volume of 25 μ l of chloroform was added and the reaction was stored at 4°C.

3.2.6 Phage titering

The phage was titered by making serial dilutions of the packaged phage particles in PDB. The three dilutions were:

1. 1:1 (10 μ l phage, no PDB)
2. 1:10² (10 μ l phage in 990 μ l PDB)
3. 1:10⁴ (10 μ l of 1:10² dilution in 990 μ l PDB)

Ten microliters of each dilution was added individually to 100 μ l of prepared EPI300-T1^R host cells and the reaction was incubated at 37°C for 20 mins. A volume of 100 μ l was plated onto LBA-CAM and incubated at 37°C overnight. The number of colony forming units was determined using the following equation:

$$\text{Titre} = \frac{(\# \text{ of colonies})(\text{dilution factor})(1000 \mu\text{l/ml})}{\text{Volume phage plated} (\mu\text{l})}$$

Based on the titer of the phage particles, phage particles were added to EPI300-T1^R cells at a ratio of 100 μ l of cells for every 10 μ l of phage particle and incubated at 37°C for 20 mins. Infected bacteria were plated on LBA-CAM plates and incubated at 37°C overnight to select for CopyControl fosmid clones.

3.2.7 Library verification

3.2.7.1 End- sequencing of fosmid clones

Fosmids were extracted from 6 randomly selected clones (Section 2.4.3). DNA was quantified by fluorometry (Section 2.5.2) and using lambda standards (Section 2.5.3). Fosmids were end-sequenced using the T7-promoter primer (5' TAATACGACTCACTATAGGG 3') at the University of Stellenbosch sequencing facility (ABI PRISM 377 automated DNA sequencer), South Africa.

3.2.7.2 Restriction analysis

Fosmid DNA was extracted from randomly selected clones. Five micrograms of sample DNA was digested with *EcoRI* and *HindIII* (Section 2.6.1). The reaction mixtures were incubated at 37°C overnight. Digested products were electrophoresed in a 0.7% agarose gel at 30 V for 16 hrs to estimate the average insert size of cloned DNA in the library (Section 2.5.4).

3.2.8 Prokaryotic diversity study

3.2.8.1 Fosmid DNA preparation

A 200 µl volume of infected bacteria was plated onto each of 30 LBA-CMC plates. After incubation at 37°C overnight, the colonies from the 30 plates were recovered using 15 ml (total) of ice-cold LB medium. A volume of 1 ml was inoculated into a 9 ml LB tube containing 12.5 µg/ml CAM and 10 µl induction solution (EPICENTRE®). After growth at 37°C with agitation for 2.5 hrs fosmids were extracted (Section 2.4.3). The DNA was quantified using fluorometry (Section 2.5.2).

3.2.8.2 PCR amplification of bacterial 16S rRNA

16S rRNA gene PCR was performed as described in Chapter 2, section 2.6.6 with minor variations. The primers used were E9F and U1510R (Table 2.5). PCR reactions contained 300 ng of template DNA and 40 µg of BSA was added. For DGGE analysis, a nested PCR using primers 341F-GC and 534R (Table 2.5) was subsequently conducted, containing 1 µl of the amplified DNA in a 50 µl reaction. After the final extension at 72°C for 20 mins, the reaction was held at 15°C.

3.2.8.3 DGGE (denaturing gradient gel electrophoresis) of PCR products

Low (30%) and high (70%) denaturing solutions were prepared from the '0%' solution [40% acrylamide : *N,N'* bis-acrylamide (37.5:1) and 1xTAE (40 mM Tris-HCl, 10 mM glacial acetic acid, 1 mM EDTA, pH8.0)] and the '100%' solution [40% acrylamide : *N,N'* bis-acrylamide (37.5:1) and 1xTAE (40 mM Tris-HCl, 10 mM glacial acetic acid, 1 mM EDTA, pH8.0, 7 M urea and 40% (v/v) deionised formamide] stock solutions for 9% polyacrylamide gels (Muyzer *et al.*, 1993). Urea-formamide gel denaturing gradients were developed using the Bio-Rad gradient former (Bio-Rad, Hercules, USA). A volume of 160 µl of 10% (w/v) APS and 16 µl of 0.02% (v/v) TEMED were added to each gel solution and swirled gently to mix. The gradient former was used to pour the gel which polymerised over 1-2 hrs. DGGE was performed using the Bio-Rad electrophoresis apparatus at a constant voltage of 100 V for 16 hrs in 1xTAE at 60°C.

A volume of 5 µl loading dye (30% v/v glycerol, 0.25% w/v orange G) was added to 20 µl of each sample and the sample was loaded into the wells. After electrophoresis, the gel was stained in 1xTAE containing ethidium bromide (0.5 mg/ml) for 20 mins, destained in 1xTAE for 15 mins and viewed. The image was captured under UV using the AlphaImager 3400 Imaging System (AlphaInnotech Corporation, San Leandro, CA).

3.2.9 Preparation of glycerol stocks

Approximately 10,000 clones were transferred into one hundred 96 well storage plates that contained LB medium supplemented with 12.5 µg/ml of

chloramphenicol. The plates were sealed with breathable strips and gently agitated at 37°C overnight. A volume of 50% [v/v] glycerol was added at a final concentration of 20% [v/v]. Replicates of the 96 well storage plates were made. Each plate was sealed and stored at -80°C.

3.2.10 Functional screening of the library for cellulase activities

For the direct screening of cellulase activity, CMC LB agar plates and CMC M9 minimal media agar plates (M9 minimal media containing 0.4% CMC as the carbon source and 1.5% purified agar) supplemented with CAM were prepared. A 96-pin replicator was used to transfer clones from each storage plate (Section 3.2.9) onto the CMC LB agar and CMC M9 minimal media agar plates. Plates were incubated at 28°C for 7 days and flooded with 0.1% (w/v) Congo red for 30 mins (Voget *et al.*, 2006). After discarding the Congo red the plates were flooded with 1 M NaCl for 30 mins and finally flooded with 1 M HCl for 10 secs (Kasana *et al.*, 2008). Colonies producing clear zones were selected for secondary screening.

3.2.11 Secondary screening for cellulase activity

Putative cellulolytic clones were inoculated from the original 96 well storage plates (Section 3.2.9) onto CAM LB agar plates containing CAM (12.5 µg/µl) and incubated at 37°C overnight. A single colony from each plate was inoculated into 5 ml of LB containing 12.5 µg/ml CAM and incubated at 37°C with shaking (150 rpm) overnight. A volume of 50% [v/v] glycerol was added to each culture to a final concentration of 20% [v/v]. The cultures were labelled and stored at -80°C. The plates were further incubated at 30°C for 5 days and flooded with Congo red,

NaCl and HCl as described previously (Section 3.2.10). Clones producing zones of hydrolysis were selected for further investigation.

3.2.12 Preliminary cellulase assay

3.2.12.1 Preparation of cell fractions using BugBuster extraction reagent

A single colony of a transformant producing a hydrolysis zone on CMC agar was inoculated into 5 ml LB medium containing 12.5 µg/µl CAM and incubated at 37°C with shaking (150 rpm) overnight. The culture was divided into two portions. One portion was lysed using 2 ml of BugBuster extraction reagent (Novagen, USA) (hereafter referred to as enzyme solution 1) as recommended by the supplier. The other portion was centrifuged at 4000 x g for 30 mins at room temperature and the supernatant was transferred to a sterile 15 ml Falcon tube (enzyme solution 2). The pellet was lysed using 2 ml of BugBuster extraction reagent and the supernatant was retained after centrifugation at 15000 x g for 5 mins at 4°C (enzyme solution 3).

3.2.12.2 Preparation of cell fractions using sonication

A single colony of a transformant producing a hydrolysis zone on CMC agar was inoculated into 5 ml of LB medium containing CAM (12.5 µg/µl) and incubated at 37°C with shaking (150 rpm) overnight. A volume of 2 ml of overnight culture was subcultured into 100 ml of LB medium containing CAM (12.5 µg/µl) and 0.1% of L- Arabinose and incubated at 37°C with shaking (150 rpm) until an OD₆₀₀ of 1 was achieved. The culture was centrifuged at 5000 rpm for 30 mins at 4°C. The supernatant was decanted and concentrated with a VIVASPIN 20

column (Sartorius Stedim Biotech) by centrifugation (3000 x g, 1 hr, at 4°C) (enzyme solution 4). The pellet was resuspended in 1 ml of ice cold phosphate buffered saline (140 mM NaCl, 2.7 mM KCl, 10 mM Na₂PO₄, pH7.3 in 1litre) and sonicated for 3 x 10 secs. During sonication the culture solution was kept on ice. The sonicated culture solution was centrifuged at 13000 rpm at 4°C for 5 mins and the supernatant retained (enzyme solution 5).

3.2.12.3 DNS assay and enzyme stability

Dinitrosalicylic acid (DNS) reagent (1, 4 – dinitrosalicylic acid: 10 g, phenol: 2 g, sodium sulfite: 0.5 g, sodium hydroxide: 10 g per liter) was prepared according to the method described by Miller (1959). D (+) Glucose monohydrate (2 mg/ml) was prepared as a stock and diluted to 1.33 mg/ml, 1.0 mg/ml, 0.5 mg/ml, 0.25 mg/ml, and 0.1 mg/ml with ddH₂O. All experiments were done in triplicate. DNS reagent (1.5 ml) was added to 1.5 ml of glucose sample in capped tubes. The mixture was boiled for 15 mins to develop a red-brown colour. A volume of 500 µl of a 40% potassium sodium tartrate solution was added to stabilize the colour. After cooling to room temperature, 10 ml water was added and the tube was inverted several times to obtain a homogenous solution. The absorbance was recorded with a spectrophotometer at 540 nm (Miller, 1959). Reducing sugar concentration was calculated from a standard curve using the equation $y=0.658x+0.021$ (where $y=OD_{540}$ and x =concentration of glucose).

Cellulase activity was measured in a total reaction mixture of 1.5 ml containing 0.75 ml of each enzyme solution (1 to 5) and 0.75 ml of 2% CMC dissolved in 0.05 M sodium citrate buffer (pH4.8). A volume of 0.75 ml of 2% CMC in sodium citrate buffer (pH4.8) was used as the enzyme blank. The reaction mixture

was incubated at 37°C overnight. 1.5 ml of DNS was added and the mixture boiled for 15 mins. A volume of 500 µl of a 40% potassium sodium tartrate solution was added immediately after boiling to stabilize the colour. After cooling to room temperature, 10 ml water was added and the absorbance recorded with a spectrophotometer at 540 nm.

Enzyme stability was determined by incubating the enzyme at 50°C or 60°C for an hour before addition of the substrate and incubation at 37°C overnight. The production of reducing sugars from the CMC in the growth medium was measured as above.

3.3 Results and discussion

3.3.1 Metagenomic fosmid library construction

Soil samples were collected from the Mphizi hot springs, Chiweta, Malawi and were used for extraction of total community DNA and for the construction of a fosmid library. When constructing large insert metagenomic libraries it is essential that the DNA from the environmental samples is of a particular size and not sheared as chimeric products form readily from smaller DNA products (Schmeisser *et al.*, 2007). In addition the DNA must be representative of the range of microorganisms present in the community and the DNA should be free of contaminating substances.

In the process of the production of the fosmid library a large amount of DNA was lost. Therefore a large quantity of soil and an efficient chemical cell lysis and DNA extraction method i.e. the Zhou method (Zhou *et al.*, 1996) was employed. This method produced suitable quality and quantities of DNA for metagenomic analysis of the sample, limiting the shearing of DNA as compared to the more

robust mechanical methods available (Bertrand *et al.*, 2005). Electrophoretic analysis of extracted DNA showed that chemical lysis produced a band of high molecular weight DNA greater than 23 kb in size (Figure 3.1). The metagenomic library was constructed using the fosmid CopyControl pCC1FOS vector in the EPI300-T1^R *E. coli* host strain. A library of over 10,000 transformants was produced.

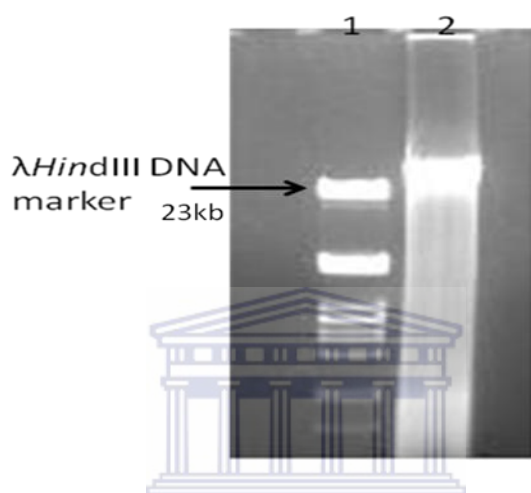


Figure 3.1 Agarose gel electrophoresis of extracted metagenomic DNA from the Mphizi hot spring site

Lane 1: λ *Hind*III DNA marker; Lane 2: Extracted HMW DNA.

3.3.2 Library verification

Fosmid extractions (Section 2.4.3) were performed on 12 randomly selected clones. Restriction endonuclease digestion of the 12 clones with *Eco*RI and *Hind*III (Figure 3.2) was used to calculate the average insert size of the metagenomic library. Inserts ranged from 19 kb to 43 kb with an average insert size of 30 kb (n=12). The library represents more than 3.0×10^8 bp of metagenomic DNA (equivalent to approximately 100 bacterial genomes). These compares well to metagenomic fosmid libraries constructed from environmental samples in other studies. Pang and coworkers constructed a fosmid library from forest topsoil containing 3624 fosmid clones with insert sizes ranging from 23.1-

40 kb (Pang *et al.*, 2009). Similarly a fosmid library prepared from a marine deep-sea sediment generated 39, 600 clones with inserts of 24-45 kb (Huang *et al.*, 2009).

A given environment is assumed to contain 2000 genomes with an average genome size of 3200 kb (Gabor *et al.*, 2003), thus the library constructed in this study represents 4.8% of the total metagenome.

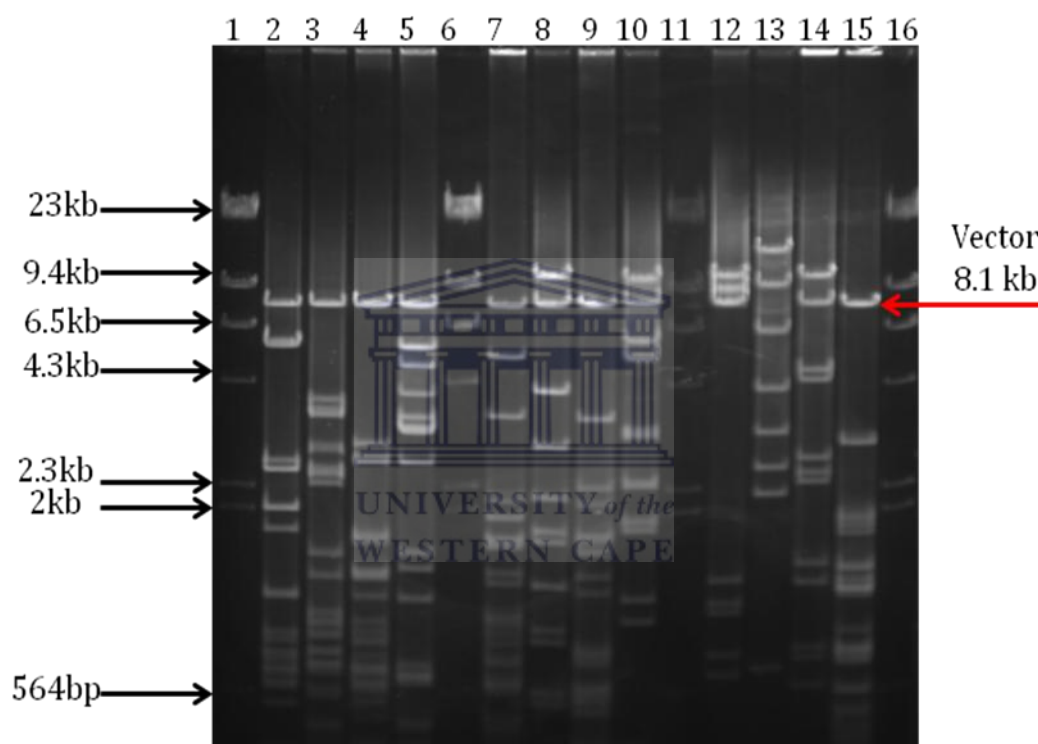


Figure 3.2 Agarose gel electrophoresis of 12 restriction endonuclease digested randomly selected fosmid clones

Recombinant fosmids were digested with *EcoRI* and *HindIII* to estimate average insert size. The band sizes of the λ *HindIII* molecular marker and the 8.1kb fosmid backbone are indicated. Lane 1, 6, 11 and 16: λ *HindIII* DNA marker. Lanes 2-5, 7-10 and 12-15: digested recombinant fosmid.

Fosmid end-sequencing analysis of 5 selected clones confirmed that the cloned fosmid DNA was of prokaryotic origin (Table 3.2). The bacterial groups represented in these sequences included the following strains: (1) *Thermotoga sp.* RQ2, a hyperthermophilic strain with an optimal growth temperature between 76

and 82°C isolated from the geothermally heated seafloor (Copeland *et al.*, 2008)
(2) *Thermotoga lettingae* TMO, a thermophilic strain isolated from a thermophilic sulfate-reducing bioreactor operated at 65°C (Zhaxybayeva *et al.*, 2009) (3) *Hydrogenivirga* sp. 128-5-R1-1, an aquificale isolated from the East Pacific Rise (104°C) and the Eastern Lau Spreading Center (76°C) (Reysenbach *et al.*, 2009) (4) *Marinitoga piezophila* KA3, a thermophilic (65°C) anaerobic, piezophilic, chemo-organotrophic sulfur-reducing bacterium isolated from a deep-sea hydrothermal chimney (Alain *et al.*, 2002) (5) *Caldicellulosiruptor saccharolyticus*, a thermophilic (70°C) strain isolated from a thermal spring (van de Werken *et al.*, 2008).

C. saccharolyticus is an extremely thermophilic, cellulolytic bacterium, which is able to hydrolyse a variety of polymeric carbohydrates (cellulose, hemicellulose, pectin, α -glucan (starch, glycogen), β -glucan (lichenan, laminarin), guar gum) (van de Werken *et al.*, 2008). Considering the organic matter rich thermal origin of the environmental sample used in this study, a wealth of lignocellulosic enzymes may be contained within the metagenomic library.

End-sequencing data suggests that the inserts in all 5 selected fosmid clones originate from thermophiles, and therefore the library is a likely to be a good representation of the Mphizi hot spring biomass which grows at 72 -78°C.

Table 3.2 Nucleotide end-sequences of selected fosmid clones and identities of the closest match

Cl on e na m e	Nucleotide sequence	Ide n t i t y %	I.D. of nearest match (Accession number)
M HS - XP 1	GGATTATCTCCTTGTAGCGGCTCACGAGCTCCGCGAGATCCTCGCGGTCAAATCCAATCGGAGGTAGCACCGCGCTCCCTTTAGTTCTCTCC ATGGCCCGGTGCGAAGGGGTCACATTCATCCTTGATGCCAGGACCTGACCCGTACATGAAAGGAGCCGACGGTAAGCGTCGTAGGCGAA GCGGGGGATGTACGGGACGCTAGGTGGGAAGAGCTTTTTCCGTCAAACCCAGGTTGAGGATGGTGTCCATCATTCCGGGCATGGAACTG GGGCGCCCGAGCGAACTGAAACGAGCAACGGATTCTTTGGTCCCAAACCTCCGACCGGTCTGGGCTTCGAGGAACTCGATCCCTCCCGAA CTTGCTCGGCCAGCCCTCCGGATATTTCCGGAGTTGGCGTAATAGTAGCGGCAAACCTCGGTGGTGTGGTGAATCCCGGGGGAACGGGGA TCCCAAAGACGCCATCTCCGCGAGGTTCCGCGCCTTTCCCAAGGAGGTCCTTCATCTTGGCG	63	pyruvate, phosphate dikinase [Thermotoga sp. RQ2] (YP_001738712)
M HS - XP 2	TGCGGTGCGCTGATTCGGATGCTCAAGACGTGAGCAACTATTACTTTAAGAGGTTCTCAAGAGTTATGTTTAGTCCAGCATCAAAGTTGAG GTCATTTCTGAAAGAGTGGTGGCGTATGCAGATCTTGAAACAAAGCTGTACGTGAGATAGAGGTACAAGATCTGCCTTTGCGTGTAAACAATC GATTCTCAGGGGATAACAAGGAAGTGGAAATTAAGTGAAGCTTGCCTTCATCTTTGGTACACGTCTGAAGTATAAAGTTGCACCTGTTTA TCTGTACTTTCAAAGGATGAACGAAACCGTTCTTGGCATTGCTACCGCGCAGCACGGGAAATGATGGACATGATGGTTCTGTTTTCCGTATG ACCCTTGATTACGACCTGAACATTATGGTTCAGAATCAATCTGAACTACGTGGTTTCTCGAGTAGTCAAGAATTGGAAGAGATTTGAAGG GGGAAAACCCGATCTTGATTTGTTCAAGGAGATACGACCACAGCTCTTGAAG	68	UDP-N- acetylglucosamine 2- epimerase [Thermotoga lettingae TMO] (YP_001470515)
M HS - XP 4	GATCCTGGCCGAGTTCCTCTTCGGCGACCAAGACGCCCTCATCCGCATCGACATGTCGAGTACATGGAGCGGTTCCGCGTGTCCGCTGGTG GGCGCTCCCCGGGCTACGTGGGCTACGAGGAGGCCGGCCAGCTCACCGAGCCGTGCGCGCCGCCCTACTCCGTCATCCTCTCGACGAG ATCGAAAAGGCCACCGCGACGTGTTCAACATCCTCCTCCAGGTGATGGACAACGGGATCCTCACCGACTCCAGGGCCGCAAGGTGGACTTC CGCAACACCATCCTCATCATGACCTCCAACGTGGGCGAGCAAACTCATCCGACATGAAGGCGGTGGGGTTACCACGGCCAGCATCGAATCC GAAGAACTACCAGGACATGAAAAAGCGGTTGGAGGGCGAGGTCAAGAAGGTGTTCTCCCGGAGTTCCTGAACCGGTTGGACGAC	69	ATPase [Caldicellulosiruptor saccharolyticus DSM 8903] (YP_001181136)
M HS - XP 5	TTACGCCCTAAAGTCTGACTACTATATGGCGCAAGCGTTGCCTTTGACAGGGAGTCACTCAAGCCCTTTACACCATACTGTATGGAAGCGTTG GGGAGAGTATGGCTTTGAAAGTTGCCAGAAGTGGCGGCTTCCCGAGGAGATAGTAAACATGGCTGAGGAGAGCATGCCGAGGACGCCAT AGACTACGCAAAGGCGAGGGAAAGCCTAAACCTGTACATAGAAGAGTACAGGGAAAAACTCAAAGAGGTAGAAGCTATAAAGCAAGACCTTA GCAAGCTAAAGGCTCAGGAAGAAAACTGCTTGCAGACCTTGAAGGCAAAGAGAGCAGATTCTCAAGAAGCTTTGAAGTCTGCACAAGAAT ACCTGGATAACCTTATGCGAGAGGCTGAGCAGCTTGTAGTTCTGCAAAGGAAAGACAGCGCTCAGGAGCTTTGAAGGGAAAAAGAGAAGA GAGATAGAGAAGGAGGTTGGAAGAGAGGAGATCAGAGTAGGAGACTACGTGGAGTTTATGGG	43	DNA mismatch repair protein MutS [Hydrogenivirga sp. 128-5-R1-1] (ZP_02177027)
M HS - XP	CCTACGGCTCCGCACCGGATGGTGTGCTTGCCTCGCACTGATGGCAACTCCCGGCTCATTCTCAAGAGGCACGCCGTACGGCAAATGCCCC GCAGTTGTTGCGGGGCATAGCCCTCCGACTGCTTGTAGGCACGCGGTTTTCAGGTTCTTTTCACTCGGCTTCCACCGTTCTTTTACCGTTCCCT CACGGTACTGGTTCGCTATCGGTACCAGGGGATTTAGCCTTAGAGGGTGGTCCCCCAGATTGCGCCAGGCTTTCACGGGGCCCGGGTACT TGGGAGCAGCACCCAGGAGATCCAGACCTTTCCGCTACGGGGCTCTACCCTCTGTGGCGACCGTTCCAGGTCACTTCGGCTAGGTCGGGA	39	conserved hypothetical protein [Marinitoga piezophila KA3] (ZP_05097862)

6	TTTTGTAACCTCCCTGAGGGGTCCGCAGCCCCCTCCGGTGCTGTCCACAACCCCGTTACCGCAACGCCTGCGGGCTTTTTCACGGTAACGGTTG GGCTGTGCCCTTCGCTCGCCACTACTCGGGGCATCGC		
---	---	--	--



3.3.3 Prokaryotic diversity study

3.3.3.1 PCR amplification of bacterial 16S rRNA

Bacterial diversity of the fosmid library was assessed by PCR amplification of the 16S rRNA gene followed by DGGE analysis. A 200 bp fragment of the 16S rRNA gene was amplified from fosmid DNA using universal bacterial PCR primers 341F-GC and 534R (Table 2.5) (Figure 3.3). These primers were designed specifically for conserved regions of the bacterial 16S rRNA gene at positions 341 and 534 for DGGE. No amplification was observed in the negative control.

3.3.3.2 Denaturing gradient gel electrophoresis (DGGE)

Denaturing gradient gel electrophoresis was performed using the 200 bp bacterial 16S rRNA gene PCR product (Fig. 3.4). At least eighteen different denaturation profiles are present in the DGGE profile. As each band theoretically represents a different bacterial population, the library contains at least 18 ribotypes. No sequence analysis was performed and the identities remain unknown.

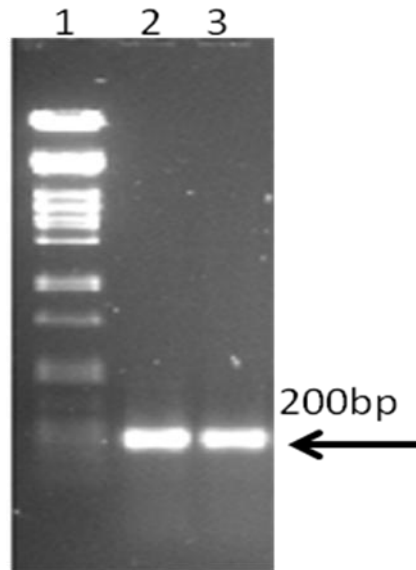


Figure 3.3 PCR amplification of the 16S rRNA genes from the metagenomic library using universal bacterial PCR primers 341 F-GC and 534r

Lane 1: Lambda *Pst*I molecular weight marker; Lanes 2-3: 16S rRNA product from the fosmid library.

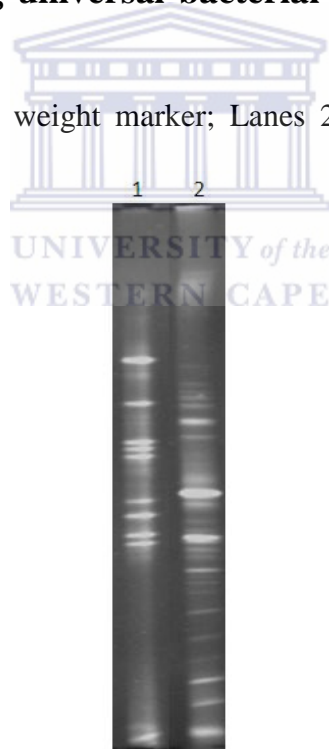


Figure 3.4 DGGE profile of 16S rRNA gene content of the Mphizi hot spring soil metagenomic library

Lane 1: DGGE marker. Lane 2: 16S rRNA gene component of the fosmid library.

3.3.4 Functional screening of the library

The metagenomic library was screened for cellulase activity on CMC LB agar plates as described in Section 3.2.10. A total of 6,000 colonies were screened for cellulase activity and of these, seventeen clones formed halos on the indicator plates. Each of the putative cellulolytic clones was streaked onto CMC LB agar supplemented with CAM (12.5 µg/ml). Screening on M9 minimal medium supplemented with CMC as the carbon source had no advantage over screening on CMC LB agar plates and as a consequence further screening was only done on CMC LB agar. Figure 3.5 shows the activity of all 17 putative cellulolytic clones growing on CMC LB agar stained with Congo red. Cellulase from *Trichoderma longibrachiatum* was used as positive control. The sizes of the zones of hydrolysis vary and five clones producing the largest zones were selected for further analysis.

Fosmids were extracted from these clones and double digested with *EcoRI* and *HindIII* restriction enzymes in order to reduce replicate clones and estimate the average insert sizes of the putative cellulase clones (Figure 3.6). The average insert size estimated for the halo-forming recombinant fosmid clones was 32.4 kb. Identical clones were found (Figure 3.6 lanes 3 and 4) indicating these two clones were replicates. The remaining 15 clones showed different patterns, indicating those clones contain different inserts.

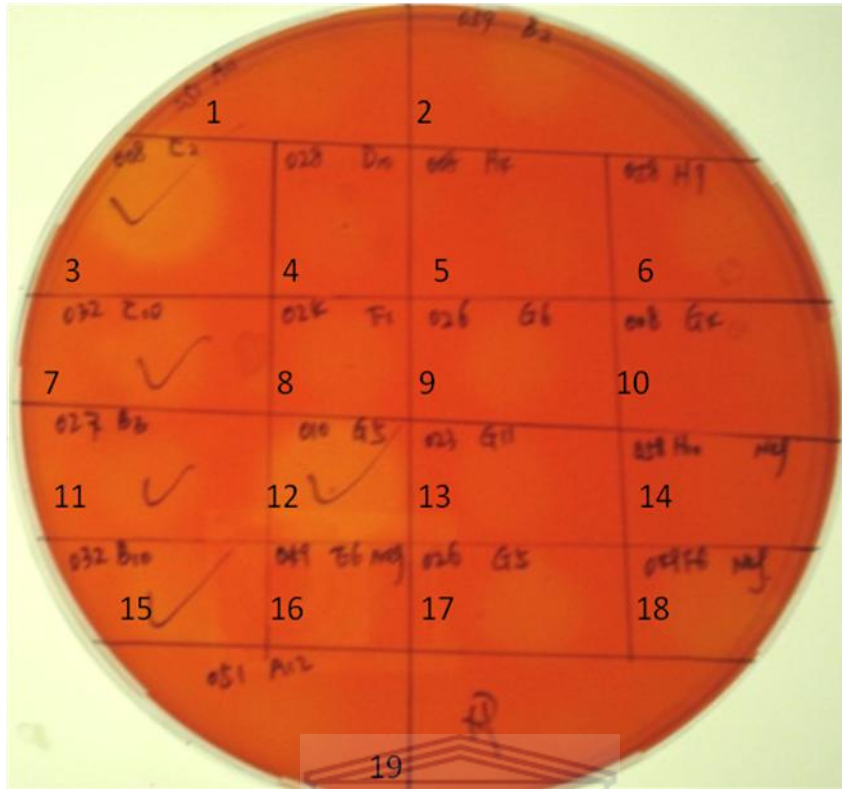


Figure 3.5 Putative cellulase producing fosmid clones screened on CMC LB agar plates flooded with Congo red

The clearing zone around the colonies indicates hydrolysis of CMC substrate. Number 12: Cellulase from *Trichoderma longibrachiatum* (positive control); Numbers 1-11 and 13-18: 17 positive putative cellulase fosmid clones; Number 19: Negative control.

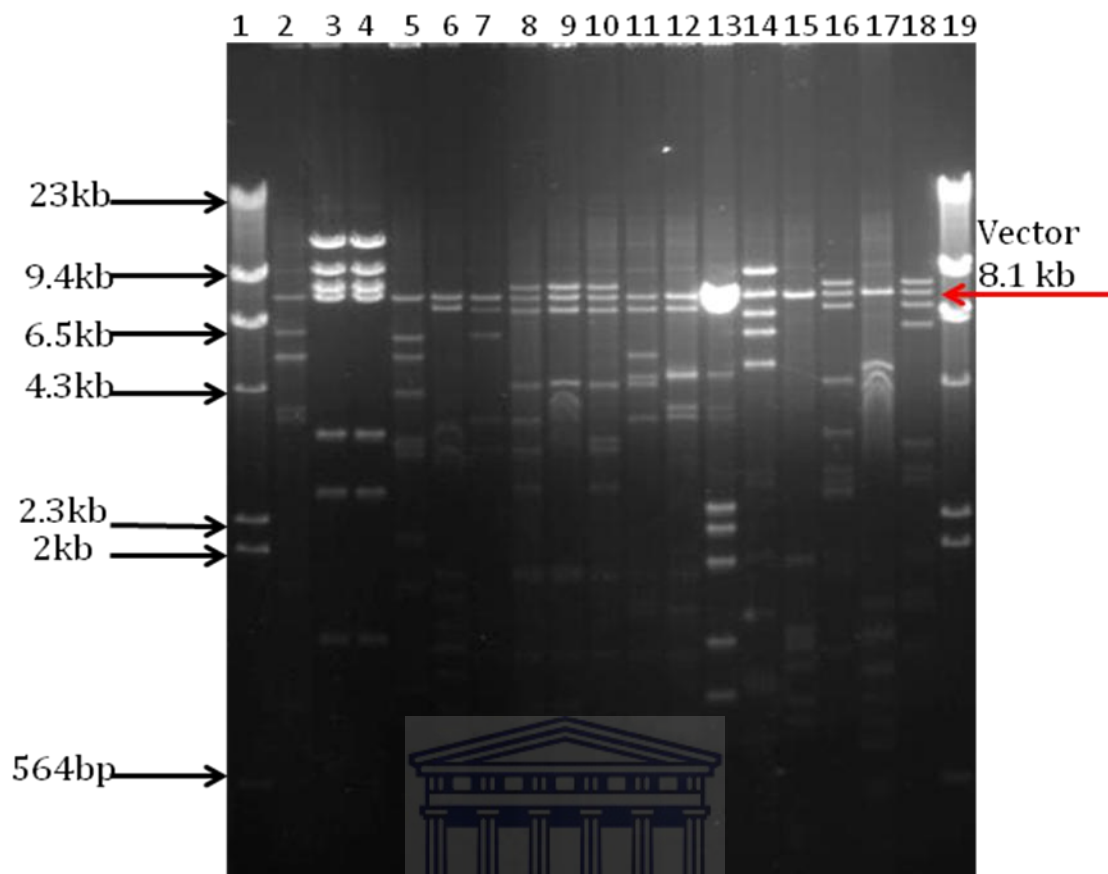


Figure 3.6 Restriction profiles of the 17 transformants which produced zones of hydrolysis during screening of the metagenomic library on CMC LB agar indicator plates

Lanes 1 and 19: λ HindIII DNA marker; Lanes 2-18: hydrolysis zone producing recombinant fosmids digested with *Eco*RI and *Hind* III.

3.3.5 Preliminary cellulase assay

The DNS assay was carried out to confirm cellulase activity and select cellulolytic clones for further analysis. Due to the low activity, a volume of 100 ml overnight culture was used for the assay. Reducing sugar was measured in intracellular fractions as well as in the culture supernatants. This was visually inspected for a change in colour after an overnight incubation at 37°C (Figure 3.7). The majority of cellulase activity was detected in cell extract (Figure 3.7, lane 9). No visible colour change was observed in the supernatant (Figure 3.7, lane 8).



Figure 3.7 DNS assay performed in the presence of culture supernatant and cell extract of fosmid clone 008C2

The reaction mixture was incubated overnight at 37°C. Tubes 1 and 7, blank; Tubes 2-6, glucose standards; Tube 8, culture supernatant; Tube 9, cell extract.

Five fosmid clones (008C2, 032B10, 026G5, 027B3, 032C10) that produced the largest cellulose degradation zones with plate assays were analysed using the DNS assay. Only intracellular fractions were subjected to analysis, and the reducing sugar content was measured before and after an incubation step at 60°C (Table 3.3). Generation of reducing sugar was calculated from a standard curve (Figure 3.8).

Table 3.3: DNS assay to determine reducing sugar generation by fosmid clones 008C2, 032B10, 026G5, 027B3, 032C10.

All assays were performed in triplicate, the average values are shown

Enzyme	OD _{540 nm}	Reducing sugar concentration (M)	After incubation at 60°C, OD _{540 nm}	Reducing sugar concentration(M)
008C2	0.488	0.53	0.432	0.48
032B10	0.462	0.50	0.278	0.34
026G5	0.411	0.46	0.465	0.50
027B3	0.200	0.27	0.237	0.31
032C10	0.261	0.33	0.318	0.38

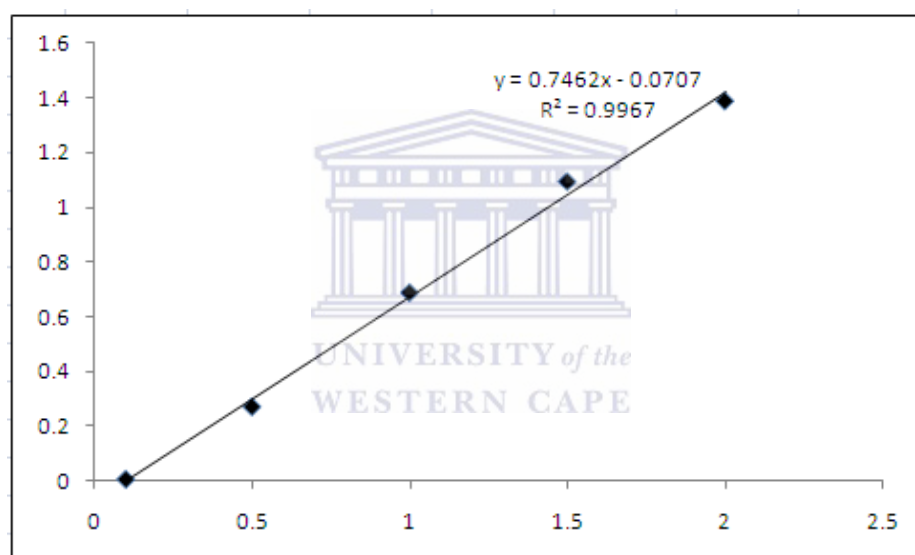


Figure 3.8: Glucose standard curve for DNS assay. x axis: glucose concentration M; y axis: OD_{540 nm}.

Enzyme stability was measured by exposing the cell fractions for 60 mins to 60°C. Fosmid clone 032B10 lost almost half its activity after incubation at 60°C for an hour. Fosmid clones 008C2 and 026G5 produced the most reducing sugar and the enzymes were stable for at least an hour at 60°C (Figure 3.9). These two fosmid clones were selected for FLX fosmid sequencing.

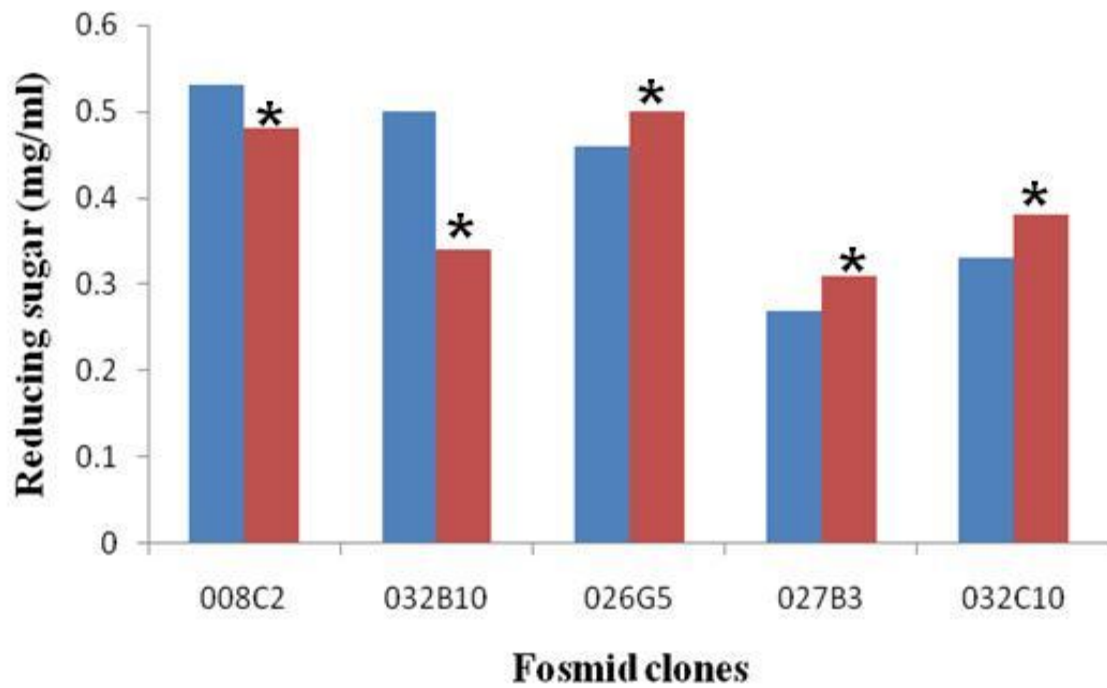
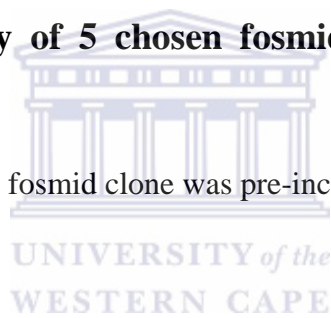


Figure 3.9 Thermostability of 5 chosen fosmid clones using the DNS assay

* Indicates the cell extract of the fosmid clone was pre-incubated at 60°C for an hour.



Chapter 4 Sequencing analysis and homology modeling

4.1 Introduction

Metagenomics is the genomic analysis of the entire genetic complement of microorganisms in a habitat by direct extraction and analysis of the entire community DNA. Metagenomics has a number of applications including the discovery of novel genes by functional screening methods, the analysis of the microbial diversity in a habitat and gene discovery by sequence analysis. The last two applications are reliant on metagenomic sequencing.

There are different methods that can be used to identify the sequence of a gene. A small insert library can be made. This involves the enzymatic digestion of the fosmid clone and cloning of 2-10 kb fragments into a suitable expression vector. The small insert library may be screened for clones with the desired activity and the corresponding small insert may be sequenced to obtain the sequence of the open reading frame (Heath *et al.*, 2009).

Another method for identifying the gene sequences of interest uses transposon mutagenesis. This involves the random mutation of putative positive clones to produce “loss –of – function” derivatives. These knock-out mutants can be selected and sequenced both upstream and downstream of the transposon insertion site with supplied primers. The full sequence of the open reading frame of interest may be obtained by primer walking (Hu & Coates, 2005).

Pyrosequencing involves the sequencing of all the independent clones. A sequencing primer is hybridized to a single stranded PCR amplicon template and incubated with a cocktail of enzymes, including DNA polymerase (Wicker *et al.*, 2006). Deoxyribonucleotide triphosphates (dNTPs) are added to the reaction in a stepwise manner. If complementary to the template, these dNTP's are incorporated onto the

template. This synthesis is accompanied by the release of pyrophosphate (PPi). In a cascade of enzymatic reactions, PPi is converted into visible light via enzymatic reactions and the amount of light emitted is proportional to the number of incorporated nucleotides. Of the three above mentioned approaches, pyrosequencing is considered to be the easiest and fastest for generation of large sequences and data sets.

One of the aims of a metagenomic sequencing project is to identify novel genes. GeneMark can be used to predict if the open reading frame is from the fosmid sequence. It is a heuristic approach, based on BLAST searches, and derives an adapted monocodon usage model from the GC-content of an input sequence. This model is utilised to compute the probability that an ORF encodes a protein (Lukashin & Borodovsky, 1998).

In order to study the structure of the protein, comparative protein structure modeling can be performed. This involves mapping of a translated ORF with unknown structure against known homologous proteins (Hilbert *et al.*, 1993). The accuracy of the model is dependent on both the quality of the sequence alignment and the template structure used for the alignment (Venclovas & Margelevicius, 2005).

This chapter describes the identification of two cellulase genes XPgene12 and XPgene25 from the fosmid clone designated 008C2, as well as the homology modeling of the two target proteins.

4.2 Sequence analysis

Fosmid clones 008C2 and 026G5 were sequenced using the Roche 454 GS FLX sequencing platform at Inqaba Biotechnical Industries (Pretoria, South Africa). The complete sequence of the cloned 008C2 fosmid insert was assembled from 8 contig sequences using CLC Genomics Workbench 3 and Sequencher. Alignment of the end sequences to the assembled contigs confirmed that the cloned inserts were completely assembled (Figure 4.1) (Table 4.1). The full sequence (29800 bp) was obtained (average mol% GC content of 51%) and ORFs within the insert were predicted using GENEMARK (v 2.6) (Besemer & Borodovsky, 1999). Although a putative cellulase coding gene was identified in fosmid clone 026G5, the fosmid insert was only partially sequenced. Consequently fosmid clone 008C2 was chosen for further study.

The majority of the ORF-encoded proteins on fosmid clone 008C2 had high sequence similarity to putative proteins of *Enterobacter* (Table 4.2) with 11 proteins showing more than 80% amino acid sequence identity. XPgene27 and XPgene28 had high sequence similarity to proteins of *Hydrogenobacter thermophilus*, while XPgene29 had high sequence similarity to the acetyl-CoA carboxylase protein of *Thermocrinis albus* (Table 4.2). Considering that the library was made from a Malawian hot spring soil sample (temperature 72-78°C), the occurrence of thermophilic species is expected and the similarity of proteins to proteins of their thermophilic homologues is not surprising. However 38% of proteins had high similarity to proteins from *Enterobacter*, a mesophilic organism. Since metagenomic DNA was used to construct the fosmid library, it is possible that DNA from mesophilic organisms was cloned.

A total of 29 ORFs, designated XPgene1- XP gene29 were predicted on the 008C2 fosmid insert (Table 4.2, Figure 4.2). The majority of the ORFs had high similarity to

sequences deposited in Genbank. Two putative cellulases, XPgene12 and XP gene25, were identified. Neighbouring genes XPgene13 and XPgene26, respectively, are situated downstream of the two putative cellulases, and possessed high sequence similarity to cellulose synthase genes. These may contribute to part of the cellulose biosynthesis pathway in bacterial species (Richmond & Somerville, 2000). The sequences obtained for the putative cellulases XPgene12 and XPgene25 and their nearest neighbours XPgene13 and XPgene26 had previously not been functionally characterized.



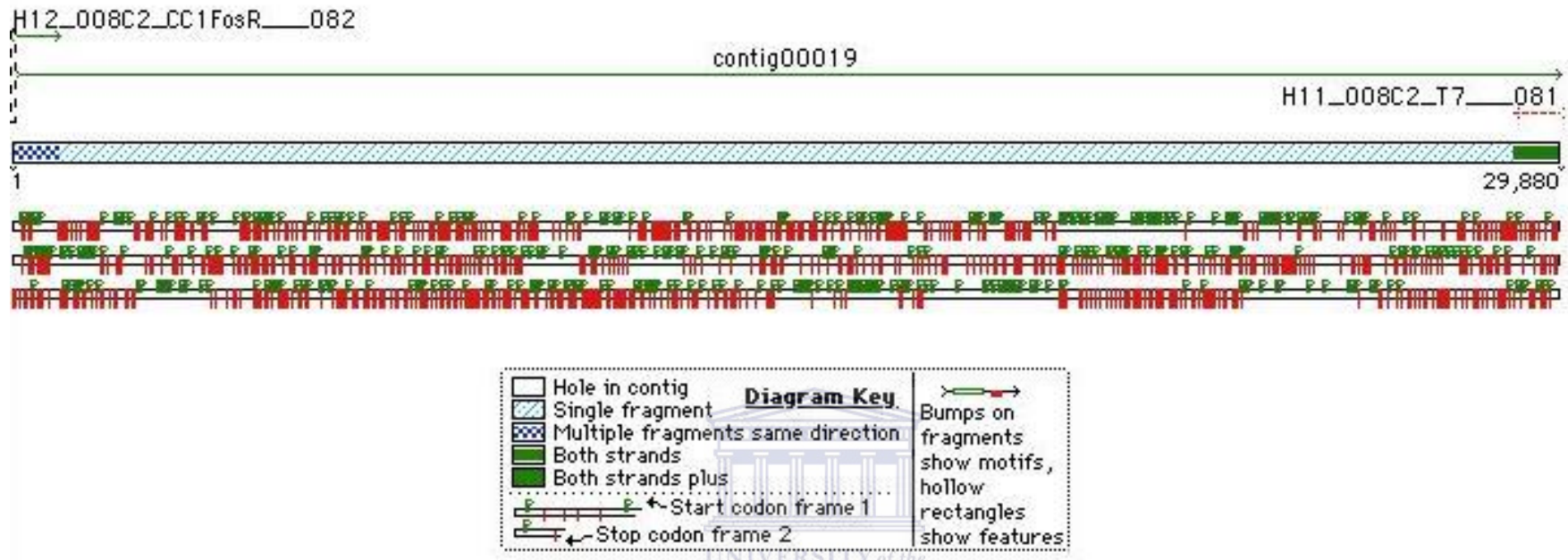


Figure 4.1 Annotation of the fosmid clone 008C2 diagram using sequencher

The data confirmed that the full length of fosmid clone 008C2 was obtained by 454 sequencing. The full length of fosmid clone 008C2 was aligned with end sequencing results using CC1FosR and T7 primers.

Table 4.1 Nucleotide end-sequences of fosmid clones 008C2 and 026G5. The nucleotide identity of the closest match is indicated

Clone name	Nucleotide sequence	Identity (%)	I.D. of nearest match (Accession number)
008C2 _CC1F osR	AGCCCTTTTAATTCAAGTGTGTTTTCAAGAACATCCASCSCCTGGCCAAGACTGGCCACCCTCGCGGTATCAAGCTTGTTCCTGAGCCGGGAGC ATCGAACACCAGTTCCGCAATGCCATCTCCAGCCAGTTGAGGTACAGGGTGTGCGCTTTGTAGAGCATGTCAGTCTCCTGAATCCAGCAATTG GATCTGGTCATACCAGATGAATCGGAGTGTGGGATTTATGTTAATAAAATGCAAATTACTCATTAAAGAAAATGCTGCATTGATCACGGTCGGTG GAAATCACGCAGCCGGAGTGTGGTGTCTAAGATGCGATGACTTGAGGTCAAAAAACGAAAGGAAGAATGATGGACTCACTGGCTTCGCTT TATAAAAATCATATCGTTACCCTACAGGAACGTACCCGCGATGACTGGCCCGCTTAAAGCTGGATGCGCTGCTGATCCACTCCGCGGAGCTGA TGAATACCTTTCTGGATGACCACGCTTATC	402/481 (83%)	Human gut metagenome DNA, contig sequence: F1-U_000268, whole genome shotgun sequence (BAAW01000268.1)
008C2 _T7	AAGACAGAGGTATAAACTTTGTAGGACCACACTGGAAGGTTATAGMSCTAATGGGAGATAAAGGCTCGTTCTAAAGAGATAATGAAGAACTA GGTGTGCTACAGTTCCAGGAAGCGATGGTATACTCAAAGACCAGCAGGAAGCAAGGCACCTAGCAAAGGAGATAGGTTATCCTGTGCTTTTG AAAGCTCCGAGGTGGTGGCGGAAGGGGCATAAGGATATGCAGGGACKAAGAAGAACTAATTAATAAACTACGAGATGGCATAACAACGAAG CTCAAAAAGCCTTTGGAAGGGGAGACCTACTGCTAGAGAAGTACATAGAAAACCAAGACACATAGAGTTTCAAGTACTAGGAGACAAGTACG GCAACGTTATACACTTGGGCGAAAGGGACTGCTCCATACAGAGAAGAAACCAAAAGCTCGTAGAAATAGCACCTCTCTGTTACTTACACCTGG TAAAAGGGCTTACTACGGGAAATCGTAGCCCAAGCAGCTAAAGAAATAGG	0	No hits
026G5 _CC1F osR	ATCCACACTTTGACTTTGGGATTAAGTTGATGGGCGCTATAGGGGATAGTCGCTTGGGGTTCCTGAGCACCCACGACCTGCGCGACCACGAA CACACTTTCGTGCTCAACTTCGCCGCGCCTTCGGCAGCGAGCGCGCTTATTTGCGTTATGTGGACGCTTTCGCGCTGACAGTTTTACCG CGGCTTTCTGTGGGGCGCGAGATGCGCTTGGTCCCGCGACCGCTTCTGTTGCGAGGAAGTTACGCTCGCCATTTCTTCTGACGAAGCG CGGATGGCGGCGCTGGCGCGTTCGGTTCGGTTACGGTGATGACCGGTGGTTCCTCTCGTAGGGTGGACCGCTTTCAGCCGAACCTCCGC CCGTTGCTTAGTTACACTCCCGCACCGACTTCAAGCGTTTTCCCTTTCTTAAACCGCAACTTTCGCACCAAGGGGACAGAGTTTACCAGCAAC GCGGGTTTTGCGTGAAGTGGGCAAGAGGGGAACTTTCAAAGGCAAGTTTTTGAACACAGCTACGGCGCAATGTGTGGTCACTTTGCGC GATCAAACCGAAATCAGCTTGGGGTGAACCGCAATCGGCACGCGGAATACCATATTCGTGAGGAGCCTTTCAGTACTGCTTACTGGCTC TCGCTGGACTTAGGGGGCAACAACTTTGCGGGCAGACCTGCTTACCGATGGGGCGGGCGTTTGTGGGCGGTTCCGCCAACCC	0	No hits
026G5 _T7	TTGTGGTCATTGACAAATCCCCGCCGCACTTACGAAGAGATGGAGCGAGATGTGGCGAAAGTTTGGTGAAGTGGACCGGACGAAGCC GATGTCTTCTGCTTACGATGTCGGTTCGCCAGAGGACTTTCAATTGCGCCAAGCGCTTGGGATTACCTGCGGGAAGCCAAAGAGATGGGCT TGGTGAAAGCTATCGGGTGTCTACCTACACGGTGAAGGGGCACAGCTTGCCGCCGAGTTGCCGAAGCGGATTTGTGCAAGTCATCGTCA ACTTGACCGGTGCGGGGATTTGGACGGGGATGCGGCAGCGATGGAAGTAGCGCTGCGGCAGCTGAAGGACGACGCAAGACTTTTGGCG CATCAAACCGCTGGCGGGGGCATGTTGGATCGGGACCACTGGCAAGATGCCCTCGCTTTCGTTCAACCAACCCCTTGTAGACTGCGTCTGC GTGGGGATGAGCACCCATCAGGAAGTGAAGCGTTTTGCGCTTTCGCCAAACAGGAACCCATCAGCGCGCAACGAAGGAAAAGCTCCTGCG GCTGCCCGCCGGTGGTGTGATCTTACTGTTGACCGGTTGCGAAGCTGCTTCCAATGTGCCAACCAACGCTTTGTTCA	0	No hits

Table 4.2 Predicted genes in fosmid 008C2

Genes XPgene	O* 5- 3	Left end	Right end	Gene length (bp)	AA *	Closest match from Genebank (Identities)	Accession number	Organism
1	-	<3	206	204	68	Fatty oxidation complex (63/68, 92%)	ZP_06014071.1	<i>Klebsilla pneumonia subsp. Rhinoscleromatis</i> ATCC 13884
2	+	358	1692	1335	444	Xaa-pro dipeptidase (423/444, 97%)	ZP_05970965.2	<i>Enterobacter cancerogenus</i> ATCC 35316
3	+	1692	2306	615	204	IMPACT family member YigZ (192/204, 94%)	ZP_05970966.1	<i>Enterobacter cancerogenus</i> ATCC 35316
4	+	2296	2535	240	80	Unknown		
5	-	3122	3232	111	37	Unknown		
6	+	3810	4355	546	181	Protoporphyrinogen oxidase (165/179, 92%)	ZP_05970968.1	<i>Enterobacter cancerogenus</i> ATCC 35316
7	+	5949	6151	204	67	Yb1199 (76/67, 100%)	CAQ34320.1	<i>Escherichia coli</i> BL21 (DE3)
8	+	9043	9147	105	34	Conserved hypothetical protein (26/34, 76%)	ZP_00231457.1	<i>Listeria monocytogenes</i> str. 4bH7858
9	+	92	9410	156	51	Conserved hypothetical	ZP_04532935.1	<i>Escherichia</i> sp. 3_2_53 FAA

		55				protein (50/51, 98%)		
10	+	99 04	10932	1029	342	UDP-N-acetylmuramate dehydrogenase (308/342, 90%)	ZP_05969878.1	<i>Enterobacter cancerogenus</i> ATCC 35316
11	+	10 92 9	11891	963	320	Biotin-[acetyl-CoA-carboxylase] ligase (299/320, 93%)	ZP_05969877.1	<i>Enterobacter cancerogenus</i> ATCC 35316
12	+	11 90 8	12900	993	330	Cellulase (258/330, 78%)	YP_001178642.1	<i>Enterobacter</i> sp 638
13	-	12 94 9	14463	1515	504	Cellulose synthase operon protein YhjU (472/504, 93%)	ZP_05969876.1	<i>Enterobacter</i> sp 638
14	+	14 50 0	14601	102	34	Unknown		
15	-	14 81 2	16023	1212	403	Putative cytoplasmic protein (340/390, 85%)	ZP_05969874.1	<i>Enterobacter cancerogenus</i> ATCC 35316
16	-	16 03 7	16216	180	59	Hypothetical protein ECED1_5130 (57/58, 98%)	YP_002400873.1	<i>Escherichia coli</i> ED1a
17	-	16 38 8	16684	297	98	IS5 12KDa protein (92/98, 93%)	AAB08680.1	<i>Escherichia coli</i>
18	-	16 76	17017	255	84	Hypothetical protein	ZP_03834309.1	<i>Pectobacterium carotovorum</i>

		3				PcarcW_24216 (30/33, 90%)		subsp. <i>Carotovorum</i> wpp14
19	-	17 19 3	17426	234	77	Integrase core domain containing protein (19/58, 32%)	XP_001898735.1	<i>Brugia malayi</i>
20	-	17 59 7	17761	165	54	Unknown		
21	-	18 23 3	18970	738	245	Hypothetical protein BURPS1710b_A0627 (100/244, 40%)	YP_335786.1	<i>Burkholderia pseudomallei</i>
22	-	19 16 9	19672	504	167	Hypothetical protein BURPS1710b_A0627 (68/166, 40%)	YP_335786.1	<i>Burkholderia pseudomallei</i>
23	+	19 68 8	20191	504	167	Cellulose synthase catalytic subunit (166/167, 99%)	CAD56669.1	<i>Enterobacter sakazakii</i>
24	+	20 20 2	22622	2421	806	Cyclic di-GMP-binding protein (713/801, 89%)	ZP_05969870.2	<i>Enterobacter cancerogenus</i> ATCC 35316
25	+	22 62 9	23735	1107	368	Endoglucanase (311/368, 84%)	ZP_05969869.1	<i>Enterobacter cancerogenus</i> ATCC 35316
26	+	23 71	25942	2226	741	Putative cellulose synthase	ZP_05969868.1	<i>Enterobacter cancerogenus</i> ATCC 35316

		7				operon C protein (647/739, 87%)		
27	-	26 17 0	26640	471	156	Bacterioferritin compositional matrix adjust (111/154, 72%)	YP_003433321.1	<i>Hydrogenobacter thermophilus</i> TK-6
28	-	26 68 2	28646	1965	654	2-oxoglutarate carboxylase large subunit (567/649, 87%)	YP_003433044.1	<i>Hydrogenobacter thermophilus</i>
29	-	28 66 0	>297 99	1140	379	Acetyl-CoA carboxylase, biotin carboxylase (317/379, 83%)	YP_003473029.1	<i>Thermocrinis albus</i> DSM 14484

O*: Orientation

AA*:

amino

acid

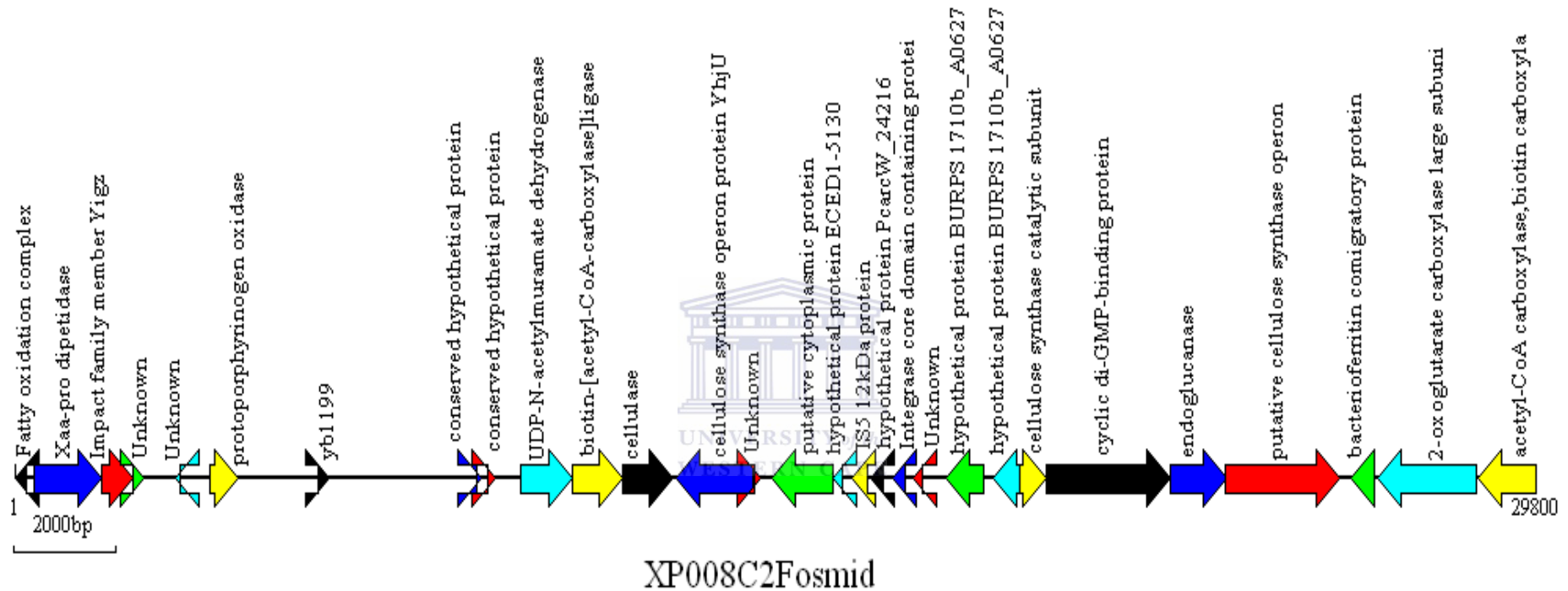


Figure 4.2 Arrangement of the open reading frames identified in the insert of fosmid 008C2

Arrows indicate the location and orientation of predicted open reading frames. Details of putative protein function and GeneBank accession numbers are given in Table 4.1.

The identified cellulase genes XPgene12 and XPgene25 encode proteins of 330 amino acids and 368 amino acids respectively (Figure 4.3 and 4.4). Multiple sequence alignments demonstrate the high levels of identity to the closest homologue (Figure 4.5 and 4.6).

These two genes both match to the family 8 glycosyl hydrolases according to the Pfam database, XPgene12 with an E-value of 1.4e-109, and XPgene25 with an E-value of 4.2e-95. This family of proteins share a common structure moiety composed of 6 α helices (Alzari *et al.*, 1996). When the XPgene12 and XPgene25 gene sequences were aligned with glycosyl hydrolase family 8 members they showed conserved catalytic residues (Figure 4.7) (Yasutake *et al.*, 2006).

Signal peptides were identified using the SignalP server. In XPgene12, a 24 amino acid N-terminal signal sequence was detected with the most likely cleavage site situated between amino acids 24 and 25 (HAD*RA) (Figure 4.8). A 22 amino acid N-terminal signal sequence was detected in XPgene25, with the most likely cleavage site situated between amino acids 22 and 23 (RAA*CT) (Figure 4.9). The Rare Codon Calculator predicted a total of 26 rare codons in XPgene12 (Table 4.3) and 29 in XPgene25 (Table 4.4). Although this accounts for less than 8% of the sequence, the high occurrence of proline (CCC) rare codons may lead to difficulties in expression of the genes in the heterologous *E. coli* host (Chumpolkulwong *et al.*, 2006).

1
ATGCGAAAACCCGTCTGCGCAACGCTGGCCGTCATGATGAGTGTGCTGTTTTTCGCCTCTC
1 M R K P V C A T L A V M M S V L F S P
L

61
TCTCATGCGGATCGGGCCTGGGAGAGTTACAAGCCCCGCTTTTTCAAACCGGAAGGCCGC
21 S H A D R A W E S Y K A R F F K P E G
R

121
ATTGTTGATAACCGGTAATGGCGGCGTGTCGCATACGGAAGGTCAGGGTTTTGCCATGCTG
41 I V D T G N G G V S H T E G Q G F A M
L

181
ATGGCGGTGGCTAACGACGATAAAGCCACGTTTCGATAAGCTCTGGCAATGGACAGACAGT
61 M A V A N D D K A T F D K L W Q W T D
S

241
CAGCTGAAGAACAAAGAAAATGGTCTGTTTTACTGGCGCTATAACCCCGCAGAGTCCAAC
81 Q L K N K E N G L F Y W R Y N P A E S
N

301
CCGGTCGCCCACAAAAACAACGCTGCAGATGGCGACGTGCTGATTGCCTGGGCGTTGCTG
101 P V A D K N N A A D G D V L I A W A L
L

361
AAAGCCGACGCCCCGCTGGCATGACAAGCGCTACAGCGCTGCATCGGATGCAATTACCAA
121 K A D A R W H D K R Y S A A S D A I T
K

421
GCACTGATTGACCACAGCGTGATCCGCTATGCCGGTTACCGCGTAATGCTGCCCGGCGTC
141 A L I D H S V I R Y A G Y R V M L P G
V

481
CAGGGGTTTAAGCTTGAGGGTGAAGTGGTCCTTAATCCTTCCTATTTTCGTGTTTCCGGCC
161 Q G F K L E G E V V L N P S Y F V F P
A

541
TGGCAGGCCTTCTCCAGACGCAGTCATTTGCCGGTCTGGCGGATTTGATTAAGGACGGG
181 W Q A F S R R S H L P V W R D L I K D
G

601
AAACGCCTGCTGGGGAAAATGGGCTCGGGTAAAGCGAATCTGCCCACTGACTGGGTTTCA
201 K R L L G K M G S G K A N L P T D W V
S

661
CTGGCATCAGGTGGAAAGCTGGCTCCCGCAAAGGCTGGCCGCGGAATGAGCTATGAC
221 L A S G G K L A P A K G W P P R M S Y
D

721
GCGATTTCGTGTTCCGCTGTACGTTGCCTGGTCTGATAAGCAAAGCCCCGCTGCTGACGCCG
241 A I R V P L Y V A W S D K Q S P L L T
P



781
 TGGAAGGCCTGGTTTCGGACAGTTTCCCCGGGAACAAACGCCCGCTGGGTTAACGTGACG
 261 W K A W F G Q F P R E Q T P A W V N V
 T


841
 ACCAACGAATATGCGCCCTACATGATGGAAGGCGGCCTGCTGGCTGTACGTGATTTCACT
 281 T N E Y A P Y M M E G G L L A V R D F
 T

901
 ATGGGGCAGTCTTCCGGTGAACCCGAAATCACCTCTAAAGACGACTATTATTTCGGCAAGT
 301 M G Q S S G E P E I T S K D D Y Y S A
 S

961 CTGAAAATGCTGGTGTGGATCGCCCAGCAATAA
 321 L K M L V W I A Q Q *

Figure 4.3 Nucleotide and deduced amino acid sequences of XPgene12

Total amino acid number: 330, MW=37011, Max ORF: 1-990



1
 ATGAAAGCCTTTTCGCTGGTGTGCATTAGCAGCGTTGATGCTGGCGGCGCTTCCTCTTCGC
 1 M K A F R W C A L A A L M L A A L P L
 R

61
 GCCGCCTGTACCTGGCCTGCCTGGGAGCAGTTTAAAAAAGGATTACATCAGTGAGGGCGGG
 21 A A C T W P A W E Q F K K D Y I S E G
 G

121
 CGTGTTCGTTGATCCCAGCGACACGCGCAAAATTACGACATCTGAAGGGCAAAGCTACGCC
 41 R V V D P S D T R K I T T S E G Q S Y
 A

181
 TTGTTCTTTGCCCTTTCGGCGAACGATCGCAGCGCGTTTGACCAGCTGCTGACCTGGACG
 61 L F F A L A A N D R S A F D Q L L T W
 T

241
 CGCGATAATCTTGCCAGCGCAATCTCAACGACCATCTGCCCGCCTGGCTATGGGGCCAG
 81 R D N L A S G N L N D H L P A W L W G
 Q

301
 AAAGATAAAGAGACGTGGGCGGTGATTGATACCAACTCCGCCTCTGACGCCGATGTCTGG
 101 K D K E T W A V I D T N S A S D A D V
 W

361
 ATCGCCTGGTCTCTGCTCGAAGCGGGCCGGTTGTGGAAAACATCCGGACTATACCCGCACG
 121 I A W S L L E A G R L W K H P D Y T R
 T

421
GGTAAGGCGCTGCTGAAACGCATTATCAGTGAGGAAGTGGTGAAAAGTGCCGGGGCTCGGC
141 G K A L L K R I I S E E V V K V P G L
G

481
GCAATGCTGCTTCCCGGTAAAGTCGGTTTTGCGGATGAAAACGTCTGGCGTTTTAACCCG
161 A M L L P G K V G F A D E N V W R F N
P

541
AGCTATCTTCCCTCCGCAGTTAGCGAGTTATTTTCACGCGCTTTGGTCCCCCGTGGACCCAG
181 S Y L P P Q L A S Y F T R F G P P W T
Q

601
CTTCGTGAAACCAATCAGCGTCTGCTGCTGGAGAGCGCGCCGAAAGGGTTTTCGCCGGAC
201 L R E T N Q R L L L E S A P K G F S P
D

661
TGGGTTTCAGTATCTGAAAAACAAAGGCTGGCGGTTACAGCAGGATAAAATCGCTGGTGGGG
221 W V Q Y L K N K G W R L Q Q D K S L V
G

721
GGCTACGACGCCATCCGCGTTTTATCTCTGGGTGGGAATGATGAGTGATAAAGATCCTCAG
241 G Y D A I R V Y L W V G M M S D K D P
Q

781
AAAGCCCGGCTGCTGACGCGCTTCCAGCCGATGGCGGCAAAGACAATGAAACGGGGTGTG
261 K A R L L T R F Q P M A A K T M K R G
V

841
CCGCCGGAGAAAGTGGATGTGGCGACGGGTAAACGCACCCGGAATGGCCCGGTCGGGTTG
281 P P E K V D V A T G K R T G N G P V G
F

901
TCTGCCGCCATGCTGCCGTTTTTACAACAACGTGATGCCAGGCGGTTTCAGCGCCAGCGC
301 S A A M L P F L Q Q R D A Q A V Q R Q
R

961
GTTGCGGACCATTTTTCCCGATAACAATGCCTATTACAGCTACGTGCTGACTCTCTTTGGG
321 V A D H F P D N N A Y Y S Y V L T L F
G

1021
CAAGGATGGGATCAGCATCGTTTTTCGCTTCACCGCAAAGGTGAATTAATACCGGATTGG
341 Q G W D Q H R F R F T A K G E L I P D
W

1081 GGCCAGGAATGCGCAAGTTCACAGTAA
361 G Q E C A S S Q *

Figure 4.4 Nucleotide and deduced amino acid sequences of XPgene25

Total amino acid number: 368, MW=41634, Max ORF: 1-1107.

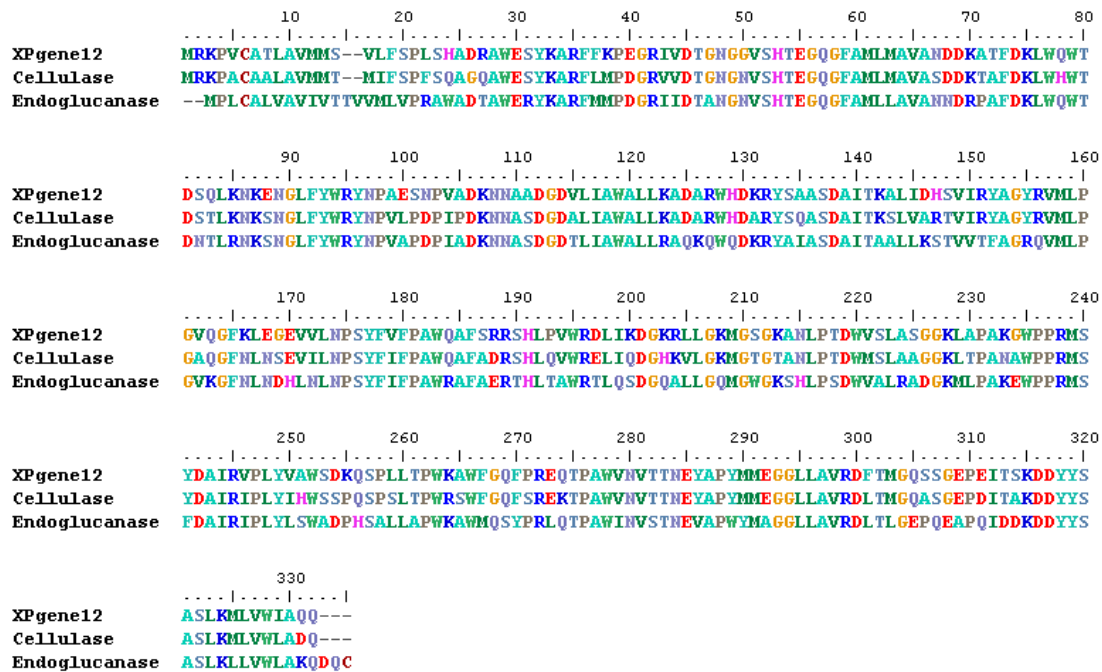


Figure 4.5 Alignment of XPgene12, cellulase from *Enterobacter* sp.638 and endoglucanase from *Klebsiella* subsp. *rhinoscleromatis* ATCC 13884 showing conserved sequences

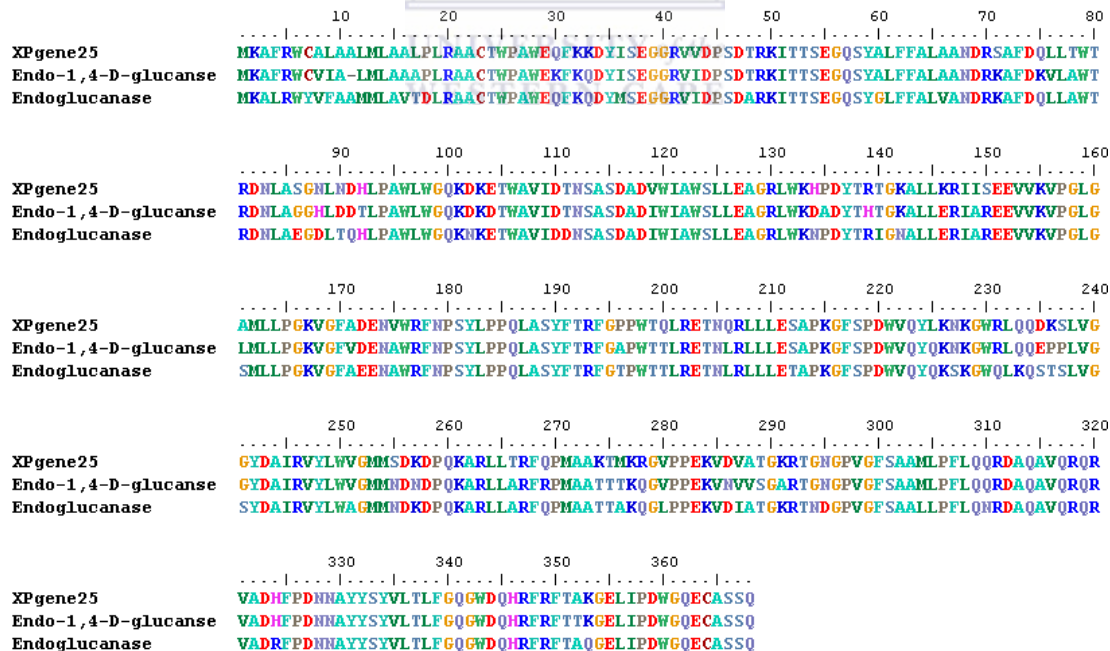


Figure 4.6 Alignment of XPgene 25, Endo-1, 4-D- glucanase from *Citrobacter rodentium* ICC168 and *Enterobacter cancerogenus* ATCC 35316 endoglucanase showing conserved sequences

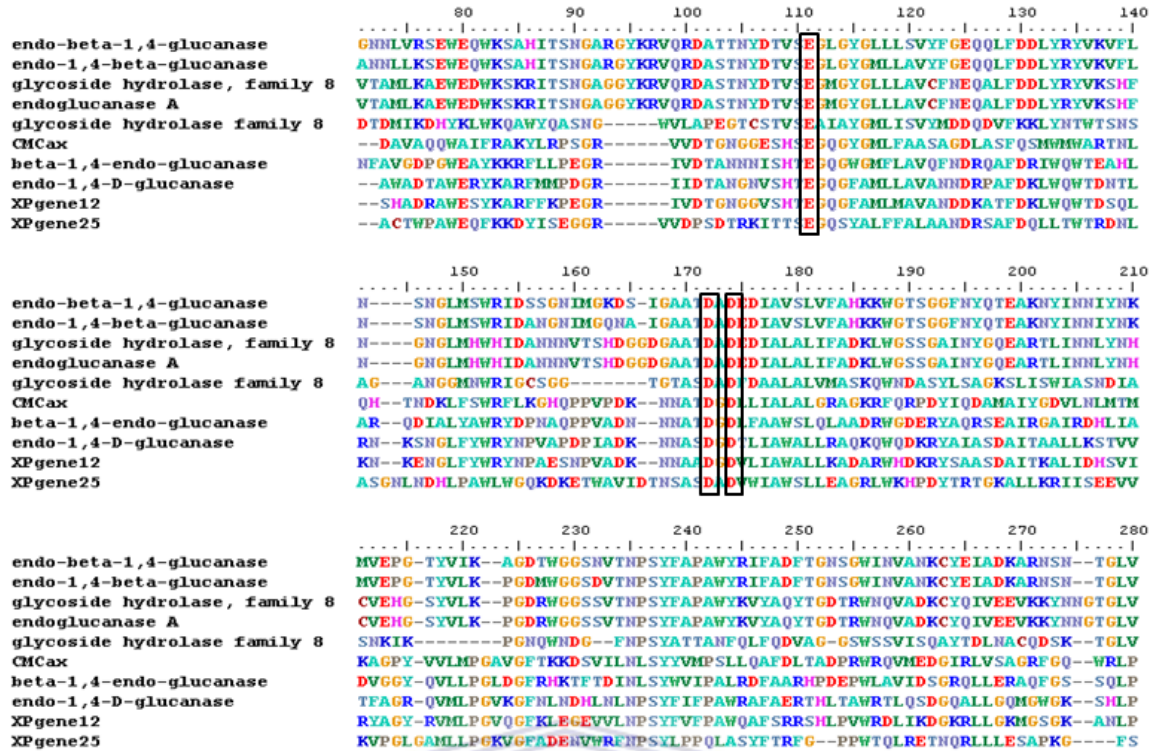


Figure 4.7 Structure-based partial sequence alignment among characterized endoglucanases belonging to GH-8

Protein sequences were retrieved from Carbohydrate-Active enzymes Database (CAZy) (<http://www.cazy.org/Home.html>). The alignment was performed with ClustalW. Accession numbers in the figure denote the following: AAA73867.1, endo-beta-1,4-glucanase precursor [*Clostridium cellulolyticum* H10 ATCC 35319]; BAA04078.1, endo-1,4-beta-glucanase [*Clostridium josui*]; ABN51508.1, glycoside hydrolase, family 8 [*Clostridium thermocellum* ATCC 27405]; AAA83521.1, endoglucanase A [*Clostridium thermocellum*]; ABU45499.1, glycoside hydrolase family 8 [*Fibrobacter succinogenes subsp. succinogenes* S85]; AAA16969.1, CMCax [*Gluconacetobacter xylinus*]; ACN29537.1, beta-1,4-endo-glucanase precursor [*Halomonas sp.* 1339]; ACO70964.1, endo-1,4-D-glucanase [*Klebsiella pneumonia*]. Catalytic residues are indicated.

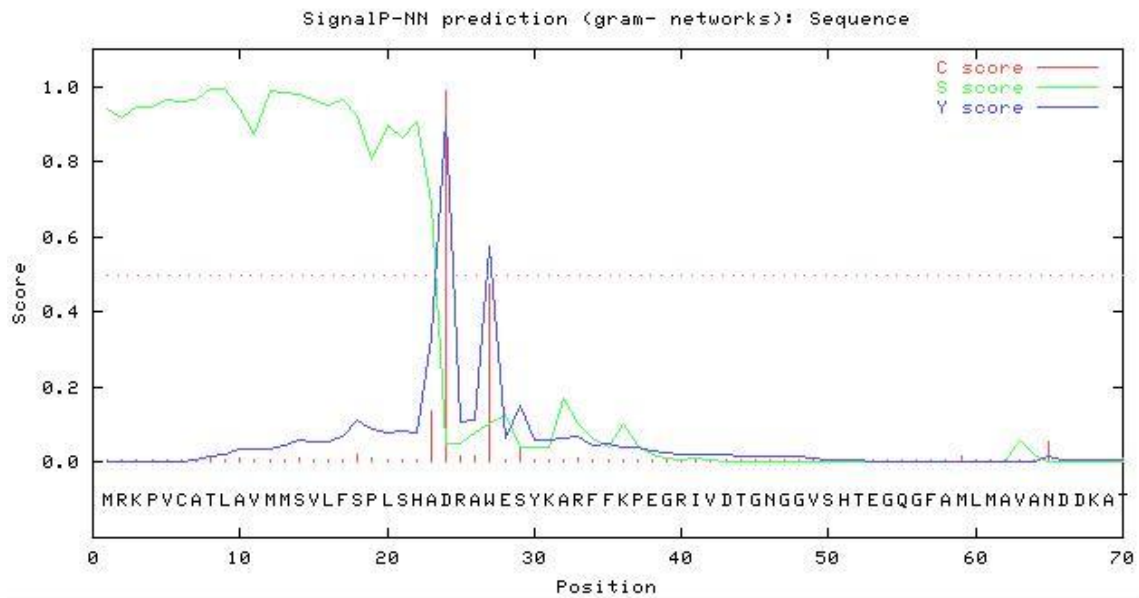


Figure 4.8 Prediction of N-terminal signal peptide cleavage site in polypeptide XPgene12

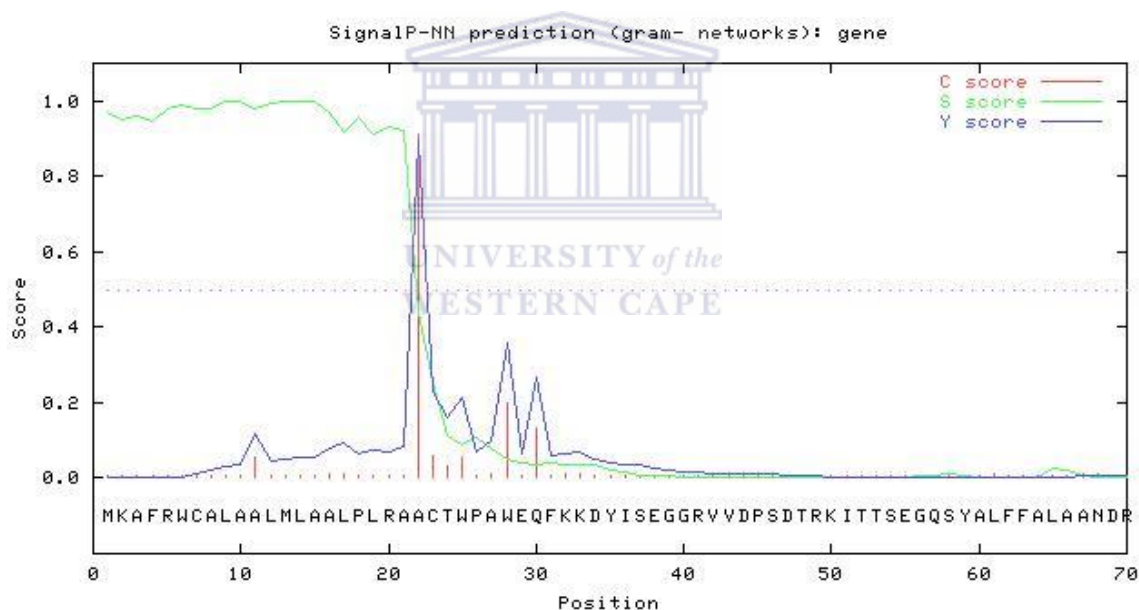


Figure 4.9 Prediction of N-terminal signal peptide cleavage site in polypeptide XPgene25

Table 4.3 Rare codons and their frequency in the nucleotide sequence of XPgene12 predicted by rare codon calculator

Amino Acid	Rare Codon	Frequency of Occurrence
Arginine	CGA	2
	CGG	2
	AGG	0
	AGA	1
Glycine	GGA	2
	GGG	4
Isoleucine	AUA	0
Leucine	CUA	0
Proline	CCC	9
Threonine	ACG	6

Table 4.4 Rare codons and their frequency in the nucleotide sequence obtained for XPgene25 predicted by rare codon calculator

Amino Acid	Rare Codon	Frequency of Occurrence
Arginine	CGA	0
	CGG	4
	AGG	0
	AGA	0
Glycine	GGA	2
	GGG	7
Isoleucine	AUA	1
Leucine	CUA	1
Proline	CCC	6
Threonine	ACG	8

4.3 Phylogenetic analysis

Comparisons of the XPgene12 and XPgene25 protein sequences with those in databases were conducted using the basic local alignment search tool, BlastP from NCBI, and a phylogenetic tree was generated (Figure 4.10).

Extensive phylogenetic analyses of the identified genes were performed using the Blast algorithm. Multiple alignments were performed using the online tool, ClustalW (<http://www.ebi.ac.uk/clustalw>) and the phylogenetic trees were generated with CLC genomic workbench using the neighbour-joining method (Saitou & Nei, 1987). Both XPgene12 and XPgene25 are closely related to the glycosyl hydrolase family 8

endoglucanases (Figure 4.10). Although XPgene12 and XPgene25 have only 9% sequence identity (28/308), they both belong to glycosyl hydrolase family 8.

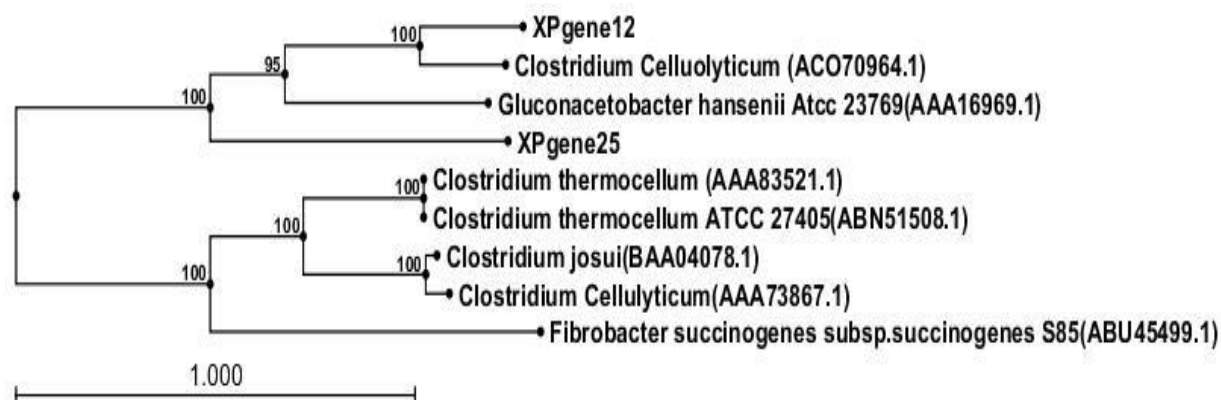
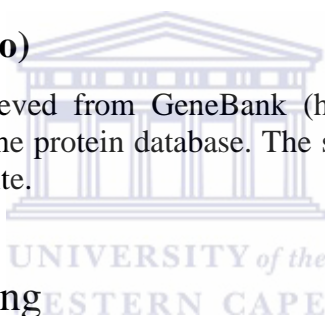


Figure 4.10 Phylogenetic tree of XPgene12 and XPgene25 generated by the neighbour-joining method and on the CLC genomics work bench software (CLC Bio)

Protein sequences were retrieved from GeneBank (<http://www.ncbi.nlm.nih.gov/>) by homology searching against the protein database. The scale bar indicates the number of substitutions per amino acid site.



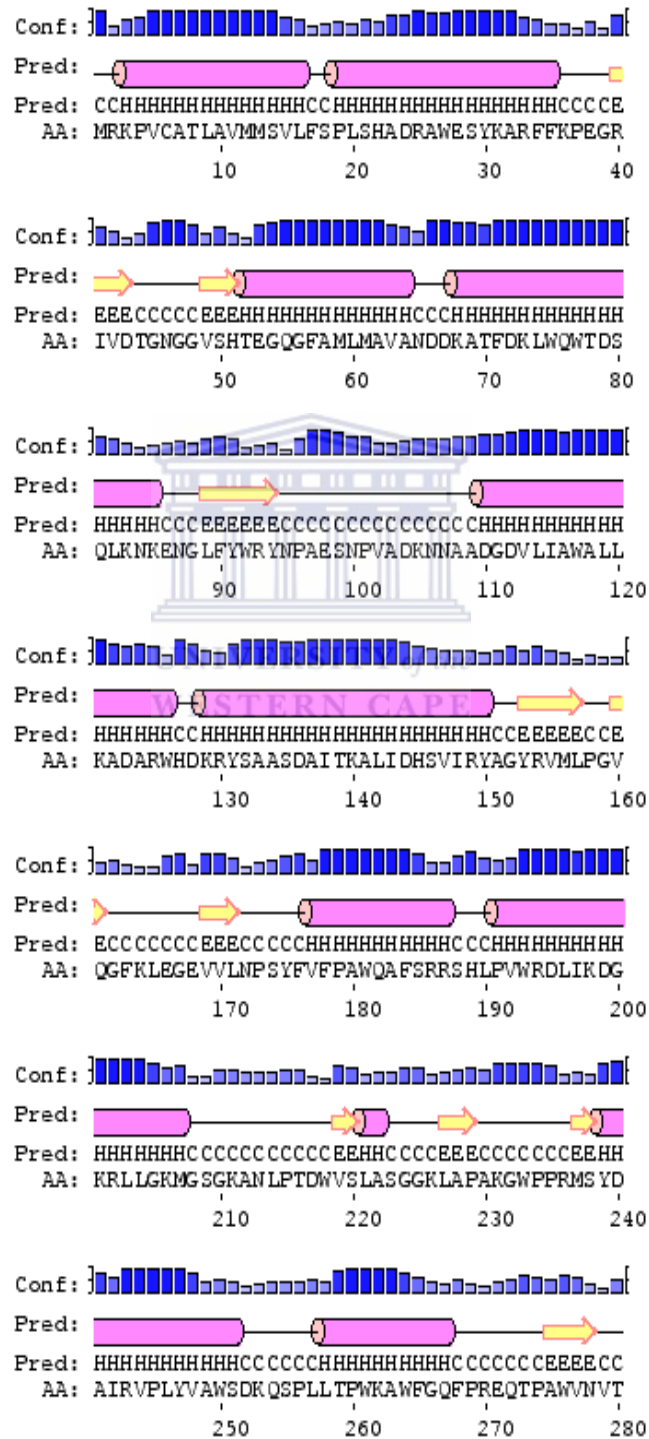
4.4 Homology modelling


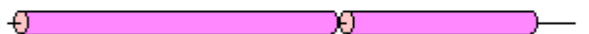
XPgene12 and XPgene25 are closely related to the glycoside hydrolase family 8. This group of enzymes is very diverse and have the ability to hydrolyse the glycosidic bond between two or more carbohydrates, or between a carbohydrate and a non-carbohydrate moiety (Henrissat & Bairoch, 1996).

Secondary structures of the proteins were predicted using PSIPRED VIEW. Many alpha helices, beta sheets and coiled regions were revealed and are depicted in Figure 4.11 and Figure 4.12. The figures also display the confidence levels for the occurrence of these secondary structures.


Sequence alignment analysis of XPgene12 and XPgene25 revealed 39.2% and 26.5% similarities, respectively, to the *Acetobactexylinum* endo- β -1,4-glucanase



CMCax (1wzza) (with E score=0.00E-1 for both) in the protein data bank (PDB). The sequence analysis and homology based 3-Dimensional structure prediction of XPgene12 and XPgene25 were carried out using 3D JIGSAW and Swiss-Model (Figure 4.13).



Conf: 
Pred: 
Pred: CHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHH CHHHHHHHHHHHHHHHH C C
AA: MKA FRWCALAA LMLAALP LRAAC TWP AWEQ FKKDYI SEGG
10 20 30 40

Conf: 
Pred: 
Pred: EEE CCCCC CCEEE HHHHHHHHHHHHHHH C C HHHHHHHHHHH
AA: RVV DPSDRKI ITT SEGQS YALFF ALA ANDRS AF DQLL TWT
50 60 70 80

Conf: 
Pred: 
Pred: HHHHH CCCCC CCEEE CCCCC CCEEE CCCCC CCHHHHH
AA: RDN LASGN LNDHLP AWLW GQDK ETV AV IDT NS ASDAD VW
90 100 110 120

Conf: 
Pred: 
Pred: HHHHHHHHHHHH C CHHHHHHHHH HHHHHHHHHHHHHHH C C C E
AA: IAW S LLEAG R L W K H P D Y T R T G R A L L K R I I S E E V V R V P G L G
130 140 150 160

Conf: 
Pred: 
Pred: EEE CCCCC CCCCC CCEEE CCCCC HHHHHHHHHHH C C C H H H
AA: AML LFGKV GFADENV WRFN P S Y L P P Q L A S Y F T R F G P P W T Q
170 180 190 200

Conf: 
Pred: 
Pred: HHHHHHHHHHHH C CCCCC CCEEE HHH C C C E E C C C C C C E
AA: LRE TN QRL L L E S A P K G F S P D W V Q Y L K N K G W R L Q Q D K S L V G
210 220 230 240

Conf: 
Pred: 
Pred: EHHHHHHHHHHH C C C C HHHHHHHHHHHHHHHHHHH C C
AA: GYDA IRVY LWVGMMS DKDPQ KAR LLTRF QPMAAK T M K R G V
250 260 270 280

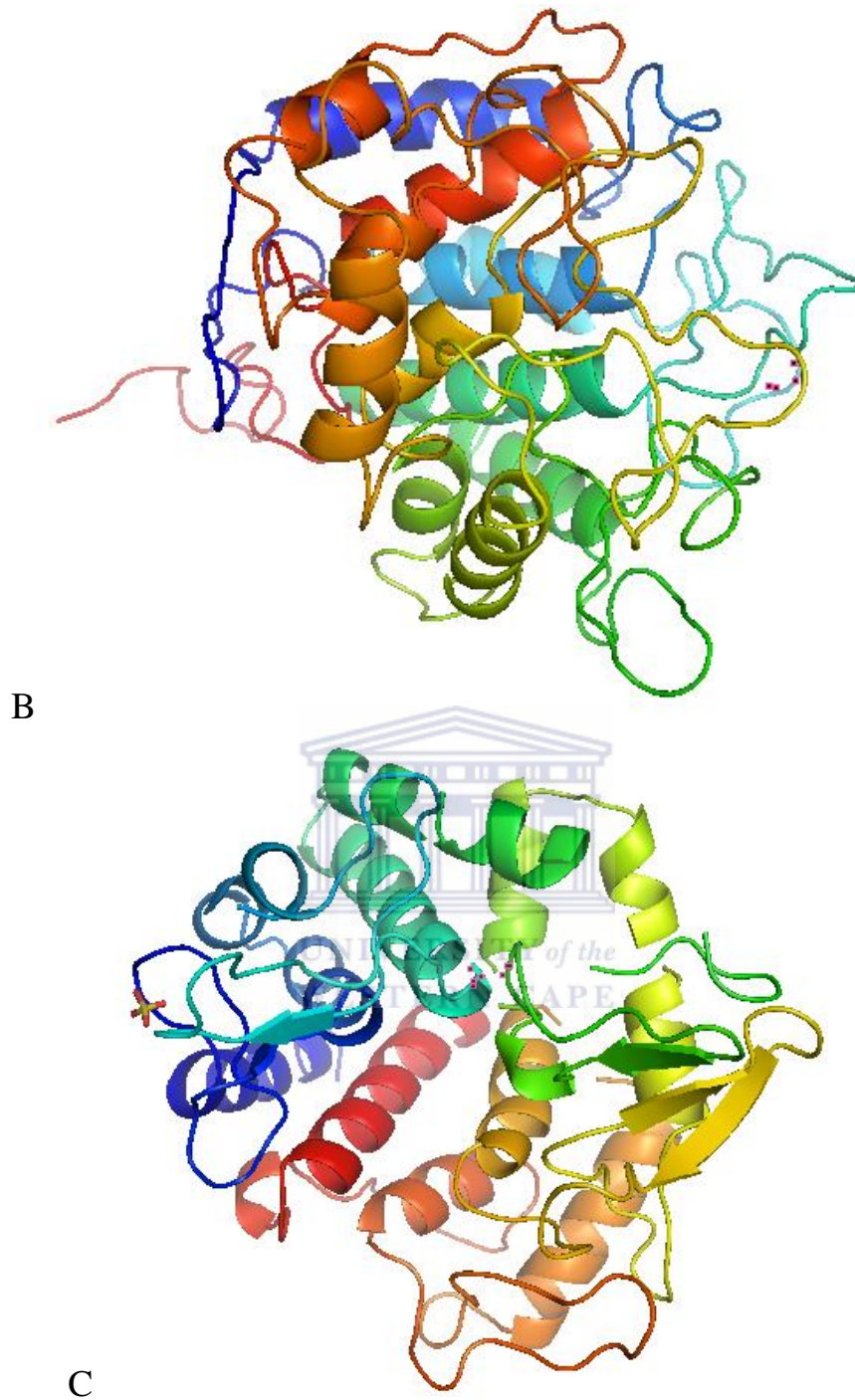


Figure 4.13 Homology models of the XPgene12, XPgene25 and the *Acetobactextylinum* endo-beta-1, 4-glucanase CMCAx gene built by the SWISS-MODEL server

XPgene12 [amino acids 25 to 330] and the Swiss model sever [amino acids 23 to 341].
B) XPgene25 [amino acids 26 to 346] and the Swiss model sever [amino acids 23 to 41].
C) The endo-beta-1, 4-glucanase CMCAx from *Acetobactextylinum* [1wzz] was used as a template.

Model accuracy was assessed using the RAMPAGE server, which considers dihedral angles ψ against ϕ of amino acid residues in XPgene12 and XPgene25 protein models. Based on this if the structure of the protein is reliable, most amino acid residues will occur in the favoured regions of the plot (Ramachandran *et al.*, 1963; Ramachandran & Sasisekharan, 1968).

With XPgene12, the Swiss model placed 284 amino acid residues (93.4%) in the favoured region, 14 residues (4.6%) in the allowed region and 6 residues (2.0%) in the outlier region (Figure 4.14). The model built by 3D JIGSAW only placed 74% residues in the favoured region, 16.1% residues in the allowed region and 10% residues in the outlier region (Figure 4.15).

With XPgene25, the Swiss model placed 284 amino acid residues (89%) in the favoured region, 19 residues (6.0%) in the allowed region and 16 residues (5%) in the outlier region (Figure 4.16). The model built by 3D JIGSAW only placed 75.4% residues in the favoured region, 14.8% residues in the allowed region and 9.8% in the outlier region (Figure 4.17).

These values suggest that the models generated by Swiss model were more accurate than 3D JIGSAW. Homology models generated by Swiss model server are depicted in Figures 4.14 and 4.16. A possible explanation for the higher accuracy of the Swiss model may be the number of residues used to generate the respective models. The Swiss model server utilized more protein residues for threading and model generation than were used by 3D JIGSAW. However, the accuracy of these two models remained relatively low. This could be expected, considering that both proteins showed a low sequence similarity to the template. The PDB database is far

smaller than the NCBI database, and only a few resolved structures suitable for use as a template for modelling were available.

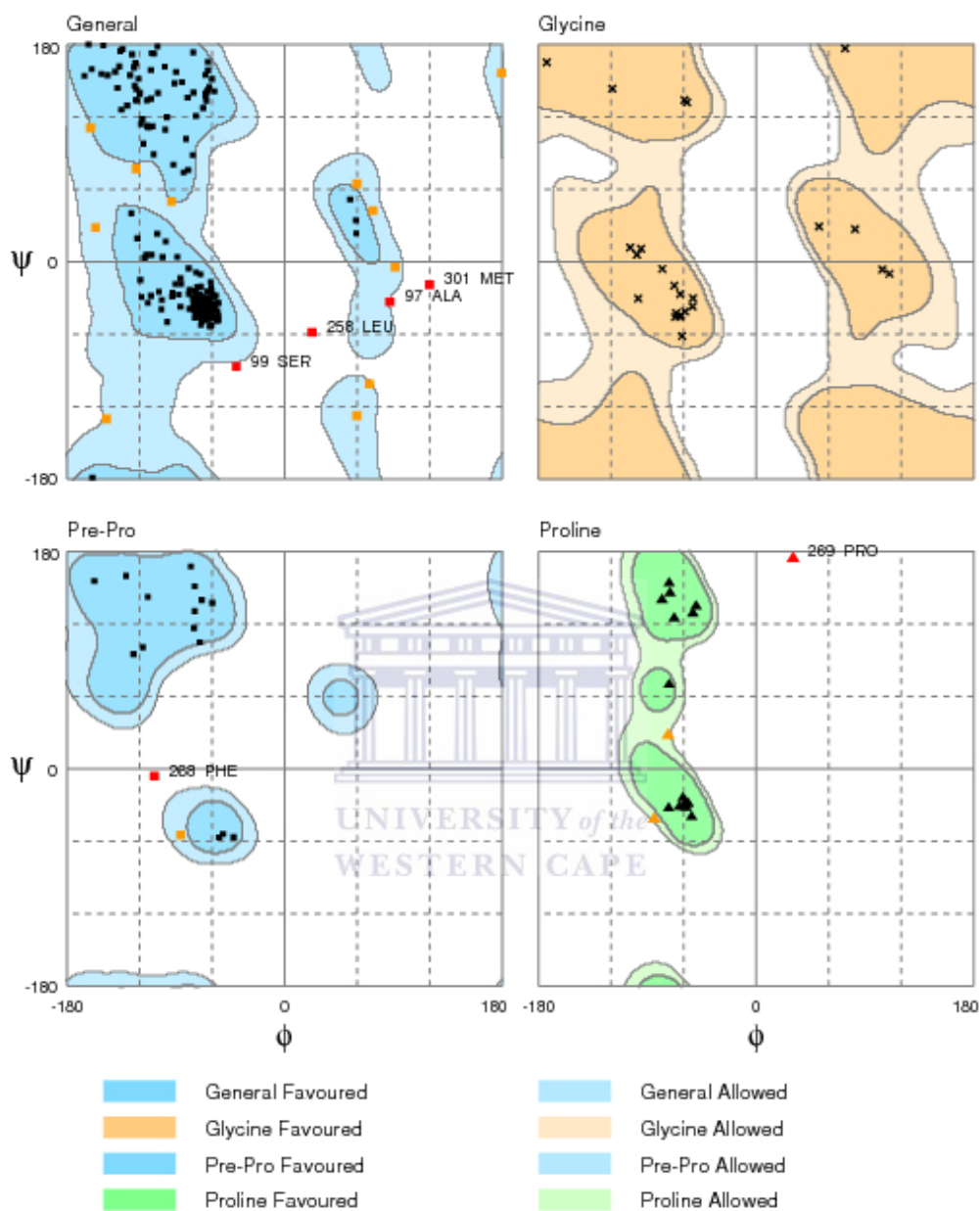


Figure 4.14 Ramachandran plot analysis of XPgene12 for general, gly, Pre-Pro built by the SWISS- MODEL using RAMPAGE software

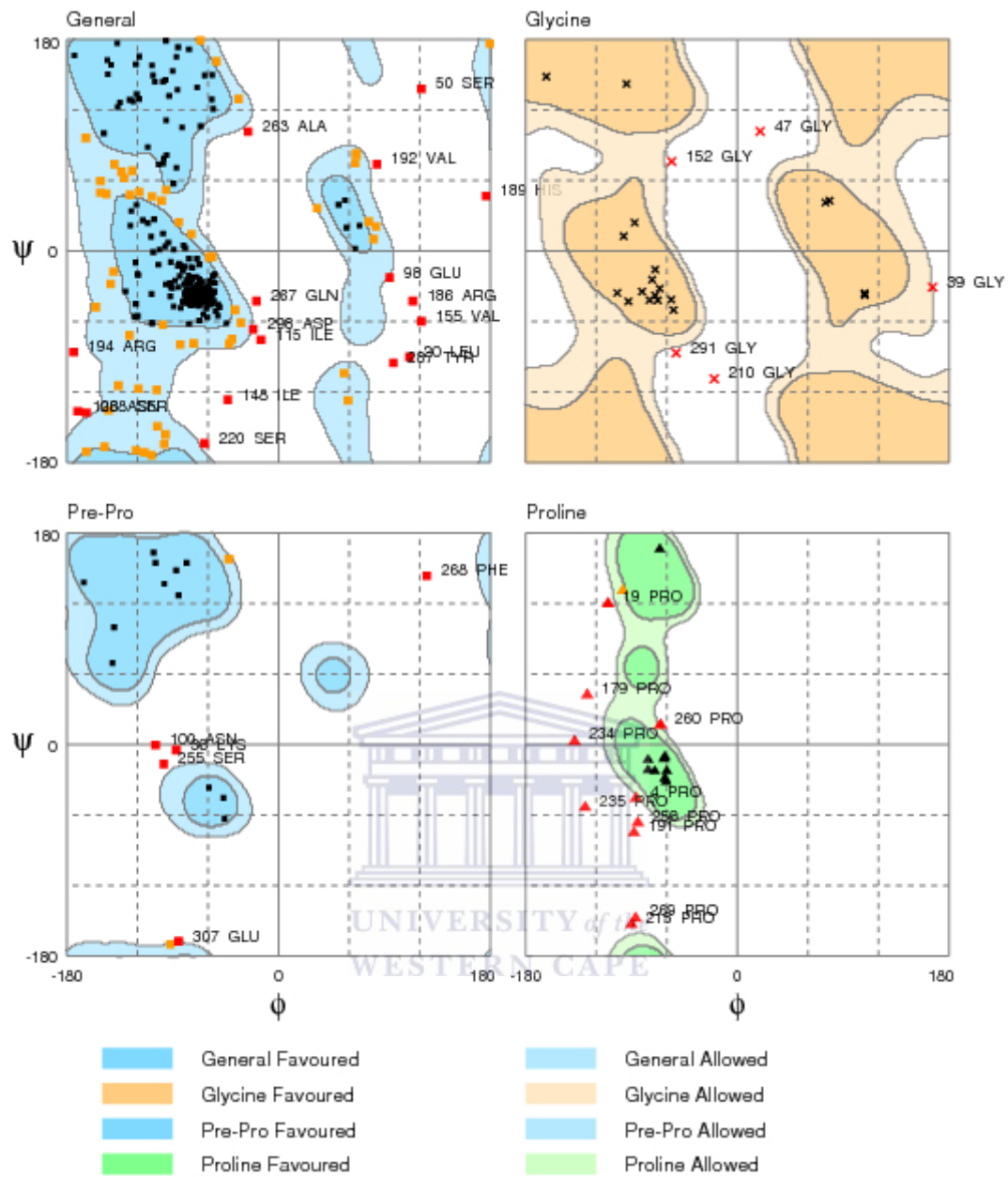


Figure 4.15 Ramachandran plot analysis of XPgene12 for general, gly, Pre-Pro built by 3D-JIGSAW using RAMPAGE software

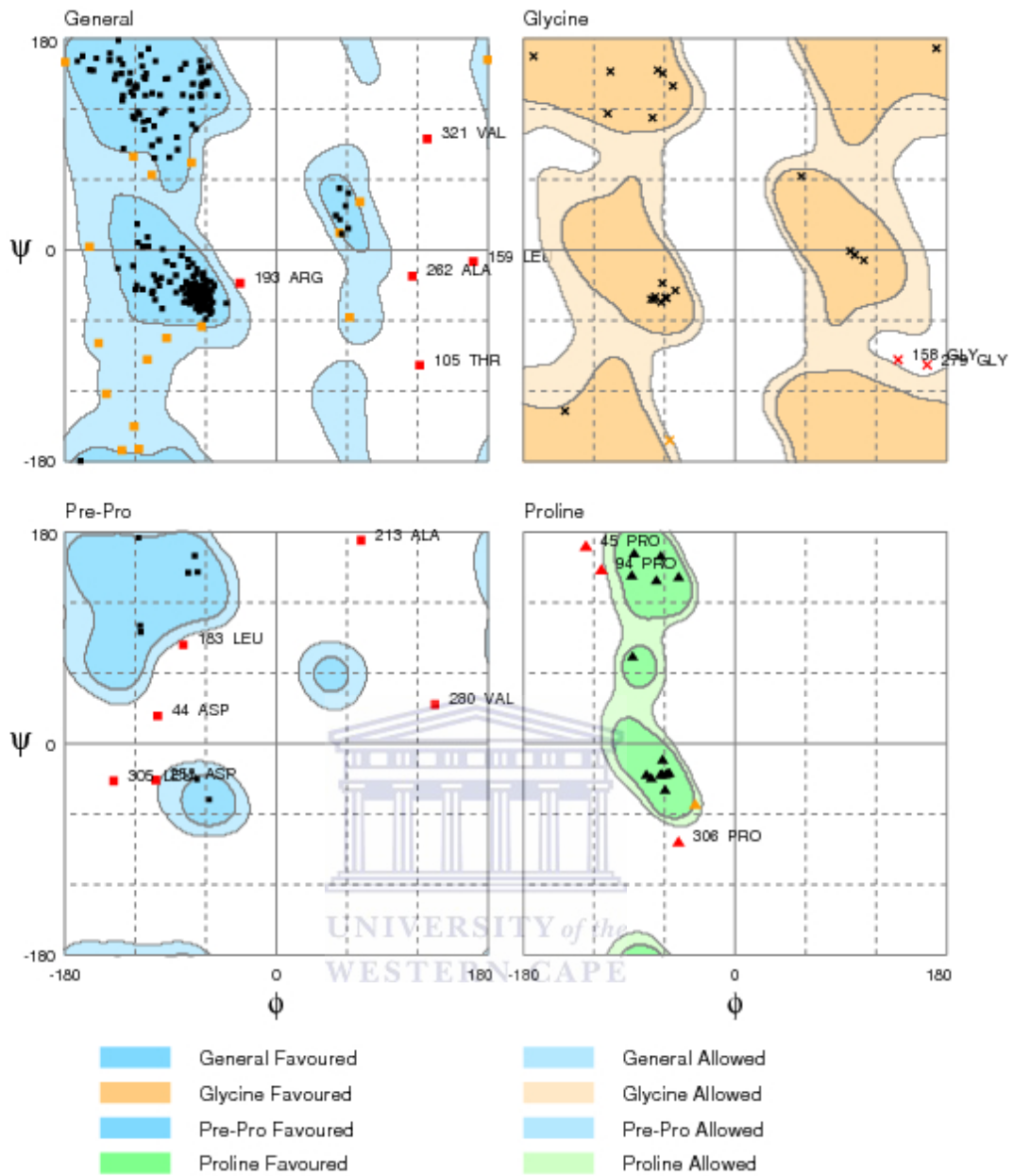


Figure 4.16 Ramachandran plot analysis of XPgene25 for general, gly, Pre-Pro built by the SWISS-MODEL server using RAMPAGE software

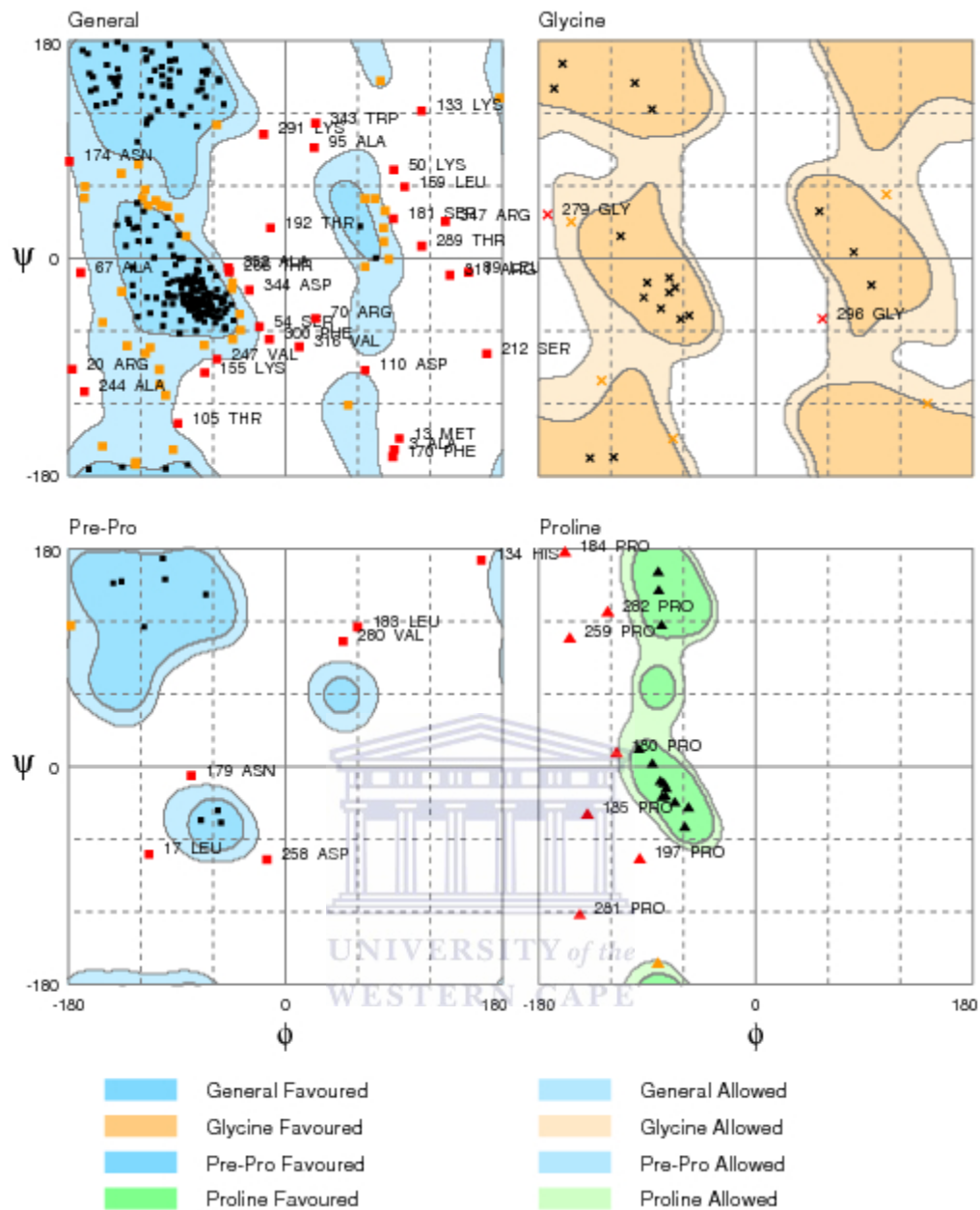


Figure 4.17 Ramachandran plot analysis of XPgene25 for general, gly, pre-pro built by 3D-JIGSAW using RAMPAGE software

Chapter 5 Cloning, expression and characterization of cellulolytic genes from a soil metagenomic library

5.1 Introduction

Two ORFs, XPgene12 (330 aa) and XPgene25 (368 aa), were identified from the fosmid library clone 008C2 as putative cellulases belonging to the endoglucanase subfamily (Chapter 4, Section 4.2). This family hydrolyzes the crystalline structure of cellulose and related cello-oligosaccharide derivatives releasing individual cellulose polysaccharide chains (Kim *et al.*, 2005). The amino acid sequences of XPgene12 and XPgene25 resemble the amino acid sequences of members of the glycosyl hydrolase family 8, formerly known as cellulase family D (Gilkes *et al.*, 1991; Henrissat *et al.*, 1989; Henrissat & Bairoch, 1993). Members of this family cleave cellooligosaccharide polymers that are at least five D-glucosyl subunits long (Alzari *et al.*, 1996). Numerous endoglucanases have been purified through direct cloning and heterologous expression in *E. coli*, often as His-tagged fusion products (Feng *et al.*, 2007; Pang *et al.*, 2009).

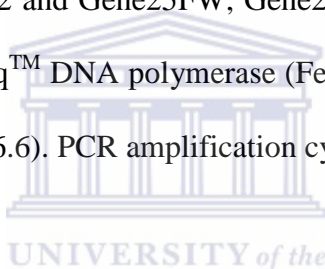
The aims of this section of the work were to determine whether XPgene12 and XPgene25 encode functional cellulase proteins, to establish substrate specificity and to confirm the identity of the group of cellulases to which these enzymes belong. The XPgene12 and XPgene25 genes were cloned and expressed using the pET vector expression system (Novagen Inc.) and a full kinetic analysis and characterisation of the recombinant His-tagged XPgene12 protein was obtained.

5.2 Materials and methods

5.2.1 Cloning of cellulolytic genes XPgene12 and XPgene25

5.2.1.1 Polymerase chain reaction (PCR) amplification of XPgene12 and XPgene25 DNA sequences

PCR primers containing sites for restriction enzymes *NdeI* and *XhoI* were designed for ORF XPgene12 and ORF XPgene25 to facilitate cloning into the pET 21a (+) and pET28a (+) vectors (Table 2.4) to express C-terminal and N-terminal His-tag fused proteins, respectively. Forward and reverse primers (Gene12FW, Gene12XhoIRV for XPgene12 and Gene25FW, Gene25XhoIRV for XPgene25) are listed in Table 2.5. Dream TaqTM DNA polymerase (Fermentas) (0.2 µl) was used in the PCR reactions (Section 2.6.6). PCR amplification cycles used are shown in Table 2.5.



5.2.1.2 Preparation of the amplified DNA sequences

The amplified gene sequences (Section 5.2.1.1) were loaded on a 1.5% agarose gel. After electrophoresis (Section 2.5.4) the DNA bands were excised and purified from the gel following the illustraTM GFXTM PCR DNA and Gel Band Purification kit protocol supplied by GE Healthcare (UK). The purified DNA fragments were ligated into the pGEM-T Easy vector (Table 2.4). Ligations were carried out using T4 DNA ligase as described in Section 2.6.2. The ligation reaction was directly transformed into electrocompetent *E. coli* Genehog cells (Section 2.6.4.1) and incubated overnight at 37°C.

Several white colonies were inoculated into 5 ml LB medium containing 50 µg/ml AMP and incubated overnight at 37°C with shaking (150 rpm). Plasmid DNA was isolated (Section 2.4.1; 2.4.2) and digested with restriction endonucleases *NdeI* and

*Xho*I in 50 µl reaction volumes at 37°C overnight (Section 2.6.1). These clones were sequenced using the M13 promoter and terminator oligonucleotide primers (Section 2.6.5).

5.2.1.3 Cloning of cellulolytic genes XPgene12 and XPgene25 for expression studies

The cloning vectors pET21a and pET28a (Table 2.4) were digested with restriction enzymes *Nde*I and *Xho*I in 20 µl reaction volumes (Section 2.6.1). Reactions were inactivated at 80°C for 20 mins and digests were stored at -20°C.

The restriction digested DNA (Section 5.2.1.2) and vectors (Section 5.2.1.3) were gel purified from a 1% agarose gel following the illustra™ GFX™ PCR DNA and Gel Band Purification Kit protocol supplied by GE Healthcare (UK). The fragments were ligated using T4 DNA ligase as described in section 2.6.2.

A volume of 1 µl of the ligation mix was transformed into 50 µl of pre-chilled electrocompetent *E. coli* Genehog cells via electroporation (Section 2.6.4.1). Recombinant plasmid was extracted (Section 2.4.2) from a single transformant and retransformed into *E. coli* Rosetta (DE3) pLysS (Section 2.6.4.2) (Table 5.1). The transformation mixture was plated on CMC LB agar plates (section 2.4.1) containing 50 µg/ml AMP and 34 µg/ml CAM for pET-21a, 50 µg/ml KAM and 34 µg/ml CAM for pET-28a. Plates were incubated at 37°C overnight.

Table 5.1 Recombinant plasmids constructed for expression studies

Clone	Description
XPgene12-pET21a	XPgene12 with C-terminal His-tag
XPgene25-pET21a	XPgene25 with C-terminal His-tag
XPgene12-pET28a	XPgene12 with N-terminal and C-terminal

5.2.1.4 Confirmation of cloning of the genes

Colony PCR was performed to confirm the cloning of the correct insert using primers for XPgene12 and XPgene25 (Table 2.5). A single colony (Section 5.2.1.3) which formed a halo on a CMC LB agar plate was suspended in 5 μ l dH₂O and used in a PCR reaction. Dream TaqTM DNA polymerase (Fermentas) (0.2 μ l) was used in the reaction (Section 2.6.5). PCR amplification cycles used are shown in Table 2.5.

All recombinant plasmids (Table 5.1) were restriction enzyme digested with *Nde*I and *Xho*I in a 10 μ l reaction as described in section 2.6.1. The inserts were sequenced using the T7 promoter and terminator primers by the University of Stellenbosch sequencing facility, South Africa.

5.2.2 Expression of the cellulolytic gene XPgene12

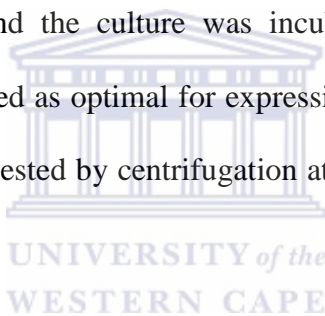
5.2.2.1 Small scale expression trials

A volume of 5 ml LB medium containing AMP (50 μ g/ml) and CAM (34 μ g/ml) was inoculated with a single colony of a *E. coli* Rosetta(DE3)PLysS XPgene12-pET21a (Section 5.2.1.3) and incubated at 37°C with shaking (150 rpm) overnight. A volume of 500 μ l of the overnight culture was subcultured into 12 ml LB medium containing AMP (50 μ g/ml) and CAM (34 μ g/ml) and incubated under the same conditions until an OD₆₀₀ of 0.6-1 was obtained. The culture was divided into two culture tubes. The culture in one tube was induced with 0.5 mM IPTG. The culture in the other tube was not induced and served as the negative control. The tubes were incubated with shaking (150 rpm) at 37°C. Cultures were sampled at 2 hrs intervals for up to 8 hrs for analysis on SDS-PAGE gels (Section 2.7.3). Each sample (1 ml) was

harvested by centrifugation at 6000 x g for 10 mins. The pellet was stored at -20°C and the supernatant at 4°C.

5.2.2.2 Large scale expression of cellulolytic gene XPgene12-pET21a

Large scale expression trials were performed on the *E. coli* Rosetta(DE3)pLysS transformed with a recombinant XPgene12-pET21a plasmid. Cells of *E. coli* Rosetta(DE3)pLysS XPgene12-pET21a from a glycerol stock were inoculated into 10 ml LB with antibiotics (50 µg/ml AMP, 34 µg/ml CAM) and incubated at 37°C with shaking (150 rpm) overnight. A 1 ml volume of culture was added to 50 ml LB with antibiotics (50 µg/ml AMP, 34 µg/ml CAM) and incubated at 37°C with shaking (150 rpm) until an OD₆₀₀ of 0.6-1 was obtained. A final concentration of 0.5 mM IPTG was added and the culture was incubated at 37°C with shaking (150 rpm) for 4 hrs (determined as optimal for expression by small-scale expression experiments). Cells were harvested by centrifugation at 6000 x g at 4°C for 10 mins and stored at -20°C.



5.2.2.2.1 Enzymatic lysis

Frozen cell pellets were resuspended in BugBuster extraction reagent (5 ml/g of pelleted cells) (Novagen, USA) containing benzonase nuclease (1 µl/ml) (Novagen, USA) and gently agitated at room temperature for 30 mins. The lysed cells were centrifuged at 10,000 x g for 10 mins and the supernatant was transferred to a sterile tube and stored at 4°C.

5.2.2.2.2 Mechanical disruption (Sonication)

Frozen cell pellets were resuspended in 1/50 volume of PBS buffer and sonicated on ice in cycles of 30 secs pulse and 30 secs pause for 5 mins per 50 ml of culture volume.

5.2.2.2.3 His-Tag affinity chromatography

The His-tag fused proteins from section 5.2.2.2.1 or 5.2.2.2.2 were purified by Ni-chelation chromatography with the His-Bind[®] Resin and Buffer kit (Novagen, USA). Purification was according to the manufacturer's instruction. The eluate was dialysed overnight in a 6 ml Slide-A-Lyzer Dialysis cassette (Thermo Fisher Scientific) against 50 mM Tris-HCl (pH7.0), 10% glycerol buffer. The His-tag purified protein was stored at 4°C. The fractions of each step of the Ni-chelation chromatography eluate were analysed by SDS-PAGE electrophoresis and zymograms were performed.

5.2.3 Cellulase activity assay

Enzyme activity was determined by measuring the amount of reducing sugar released by cellulase-catalyzed hydrolysis of the substrate CMC using dinitrosalicylic acid reagent (Miller, 1959). The reaction mixture consisted of 15 µl of CMC (2%) in 50 mM sodium acetate buffer, pH4 with 35 µl of sample. Unless otherwise specified, the enzyme reaction was performed after 10 mins of incubation at 50°C. Incubation at 22°C was used for pH profiling. The production of reducing sugar was measured by determining the absorbance at 510 nm in a Cary 50 Bio spectrophotometer (Varian, USA). Blank reactions (15 µl of CMC 2%) in 50 mM sodium acetate buffer, pH4 with 35 µl of water) were included with every measurement. One unit of enzyme activity (U) is defined as the amount of enzyme releasing 1 micromole of reducing sugar per min per milligram protein. All experiments were carried out in triplicate.

5.2.3.1 Determining the effect of pH on enzyme activity

To determine the optimum pH of the enzyme, the following buffers were used: 50 mM Glycine-HCl (pH1.0-3.0), 50 mM sodium acetate (pH3.0-5.0), 50 mM MES (pH5.0 - 7.0), 50 mM Tris-HCl (pH7.0 - 9.0) and 50 mM CAPS (pH9.0 - 11.0).

5.2.3.2 Determining the effect of temperature on enzyme activity

The optimal temperature for enzyme activity was determined for a temperature range of 10 to 90°C using the standard reducing sugar assay. Thermostability of the enzyme was determined at 60, 70 and 80°C. Enzyme samples were incubated for 15, 30, 45, 60 mins at each temperature and the residual activities were determined using the standard reducing sugar assay.

5.2.3.3 Substrate specificity

Substrate specificity of the enzyme was determined in a standard reducing sugar assay mixture of 15 µl of polysaccharide (2%) in 50 mM sodium acetate (NaOAc) buffer, pH4 with 35 µl of cell extract. Activity towards CMC was determined as above. The specificity assays were incubated for different time periods: 60 mins for β-D-glucan from barley (Sigma) and 120 mins for xylan from birchwood (Sigma) and methyl cellulose (Sigma). To determine enzyme activity towards p-nitrophenyl-β-D-glucopyranoside (pNPG) (Sigma) and p-nitrophenyl-β-D- cellobioside (pNPC) (Sigma), a volume of 10 µl enzyme (3.7 µg) was incubated with 100 µl of 10 mM pNPG or pNPC in sodium acetate buffer (pH4) at 50°C for 10 mins. The reaction was stopped by the addition of 1 ml of 0.6 M sodium carbonate, and the release of p-nitrophenol (p-NP) was measured by determining the absorbance at 410 nm using a Cary 50 Bio spectrophotometer (Varian, USA).

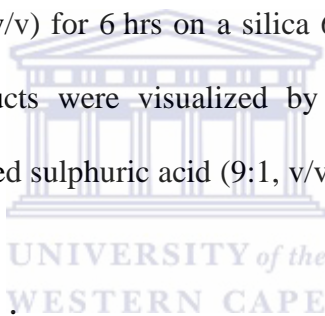
5.2.3.4 Determining catalytic efficiency

Enzyme kinetics were determined using CMC as the substrate at various substrate concentrations ranging from 10 mg/ml to 40 mg/ml (10.18 to 40.73 mM). The

standard enzyme assay (Section 5.2.3) in 50 mM NaOAc buffer (pH4) incubated at 30°C for 10 mins was performed. The data was analysed using the Lineweaver-Burk method (Lineweaver & Burk, 1934). Km and Vmax values were determined.

5.2.3.5 Thin-layer chromatography

Thin-layer chromatography was performed in order to detect hydrolysis by the purified XPgene12 gene products of the cello-oligosaccharides in 50 mM NaOAc, pH4. After 2 µg of enzyme was incubated with 10 µl sugar solutions containing 10 mg/ml of cellobiose (G2), cellotriose (G3), cellotetraose (G4) or cellopentaose (G5) at 30°C for 1 hr, chromatograms were developed using 1- propanol: nitromethane: H₂O (5:3:2, v/v/v) for 6 hrs on a silica 60 TLC plate (Merck KGaA, Darmstadt, Germany). Products were visualized by spraying the plates with a mixture of ethanol/concentrated sulphuric acid (9:1, v/v) (Voget *et al.*, 2006; Feng *et al.*, 2007).



5.3 Results and discussion

5.3.1 Cloning of cellulase encoding genes XPgene12 and XPgene25

DNAMAN was used to predict restriction enzyme recognition sites occurring in the XPgene12 and XPgene25 DNA sequences. Restriction enzymes *NdeI* and *XhoI* did not cut within the respective gene sequences. PCR primers (Section 5.1.1, Table 2.5) were therefore designed to introduce *NdeI* and *XhoI* sites at the 5' and 3' ends of the amplified products, respectively. Both amplified genes were successfully cloned into the expression vectors pET 21a (Figure 5.1) and pET 28a (data not shown).

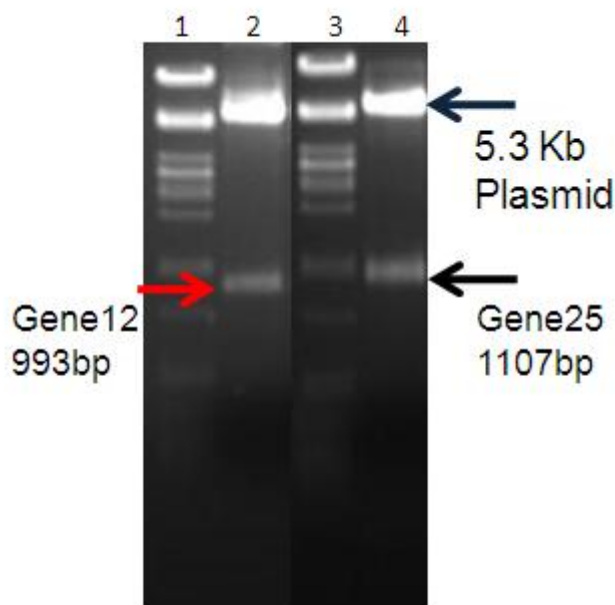


Figure 5.1 Cloning of XPgene12 and XPgene25 into pET 21a vector

Lanes 1 and 3: Lambda Pst I molecular weight marker; Lane 2: Recombinant plasmid XPgene12- pET 21a digested with *NdeI* and *XhoI*; Lane 4: Recombinant plasmid XPgene25- pET 21a digested with *NdeI* and *XhoI*; Red arrow: XPgene12; Black arrow: XPgene25; Blue arrow: pET 21a.

LB agar plates containing CMC (1% v/v) were used for cellulase activity screening. Plates were inoculated with *E. coli* Rosetta(DE3)pLysS transformed with plasmid (control) and recombinant clones (Table 5.1). Zones of clearance formed after flooding with Congo red indicated the production of CMC-hydrolases. Colonies on control plates harbouring vectors pET21a and pET28a did not form halos (data not shown). All four *E. coli* Rosetta(DE3)pLysS transformants (Table 5.1) showed zones of clearance on LB agar plates containing CMC. Figure 5.2 shows the zones of clearance produced by the XPgene12 gene product.



Figure 5.2 A XPgene12 *E. coli* Rosetta(DE3)pLysS transformant demonstrating a zone of clearance on a LB agar plate containing CMC(1%) after staining with Congo red

Following transformation, PCR analysis of the DNA from halo producing colonies of *E. coli* Rosetta(DE3)pLysS harbouring XpGene12-pET21a and XPgene25-pET21a confirmed the presence of the genes (Figure 5.3). Sequencing of the inserts using T7 primers confirmed the presence of the 993 bp (XPgene12) and 1107 bp (XPgene25) genes.

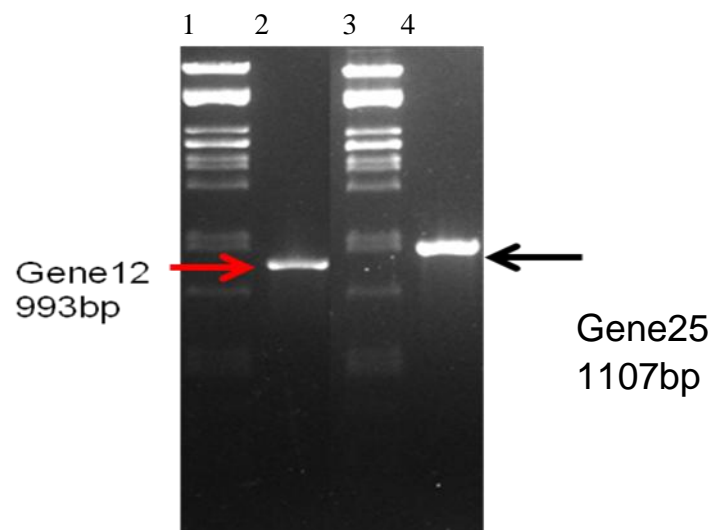
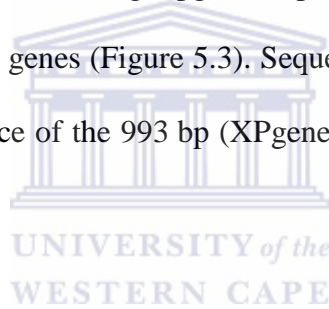


Figure 5.3 PCR amplification of XPgene12 and XPgene25 using gene specific primers (Table 2.5) for confirmation of cloning into the pET vectors

Lane 1 and Lane 3: DNA molecular marker Lambda *Pst*I digested DNA; Lane 2: PCR amplified gene of XPgene12 Lane 4: PCR amplified gene of XPgene25

5.3.2 Expression and purification of XPgene12

Cell extracts of *E. coli* Rosetta(DE3)pLysS XPgene12-pET21a (Section 5.2.1.3) were analysed on polyacrylamide gels in order to determine the expression and size of the XPgene12 gene product, and to monitor the degree of purity (Figure 5.4). A protein band migrating at ~35 kDa was present in cells which had been induced with IPTG (Figure 5.4: lanes 4 and 6). According to DNAMAN software, the calculated molecular mass of the gene product of XPgene12 is 37 kDa without the His-Tag and 38 kDa with the His-Tag.

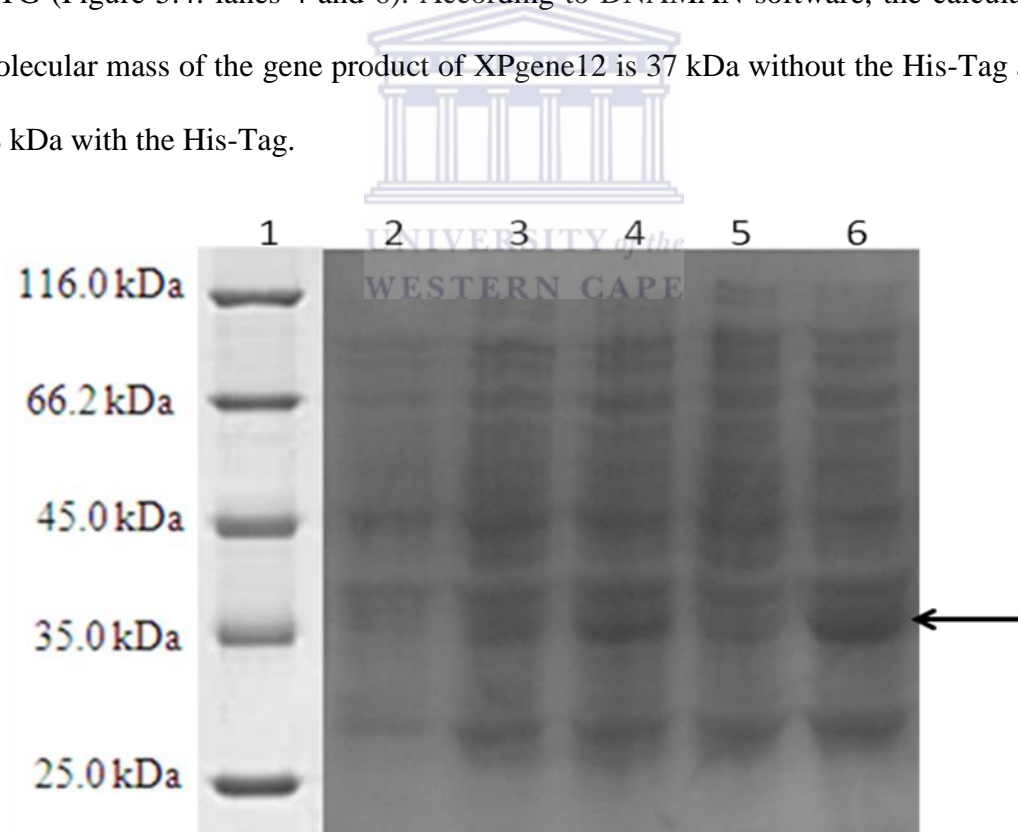


Figure 5.4 SDS-PAGE analysis of cell extracts of XPgene12-pET21a in *E. coli* Rosetta(DE3)pLysS

The protein band corresponding to a size of 38kDa is indicated. Lane 1: protein molecular weight marker (#SM0431 Fermentas), lane 2: uninduced total protein extract; Lanes 3 and 5: soluble fraction of XPgene12 after 2 hrs and 4 hrs uninduced;

lane 4: soluble fraction of XPgene12 after 2 hrs induction with IPTG; lane 6: soluble fraction of XPgene12 after 4 hrs induction with IPTG.

Cell extracts of XPgene12-pET21a *E. coli* Rosetta(DE3)pLysS were prepared by enzymatic methods (Section 5.2.2.2.1) or sonication (Section 5.2.2.2.2) and subjected to His-Tag affinity chromatography. The eluted fraction from XPgene12-pET21a *E. coli* Rosetta(DE3)pLysS showed a protein band of approximately 37 kDa (Figure 5.5, Lane 6). This was the expected size of the full length protein.

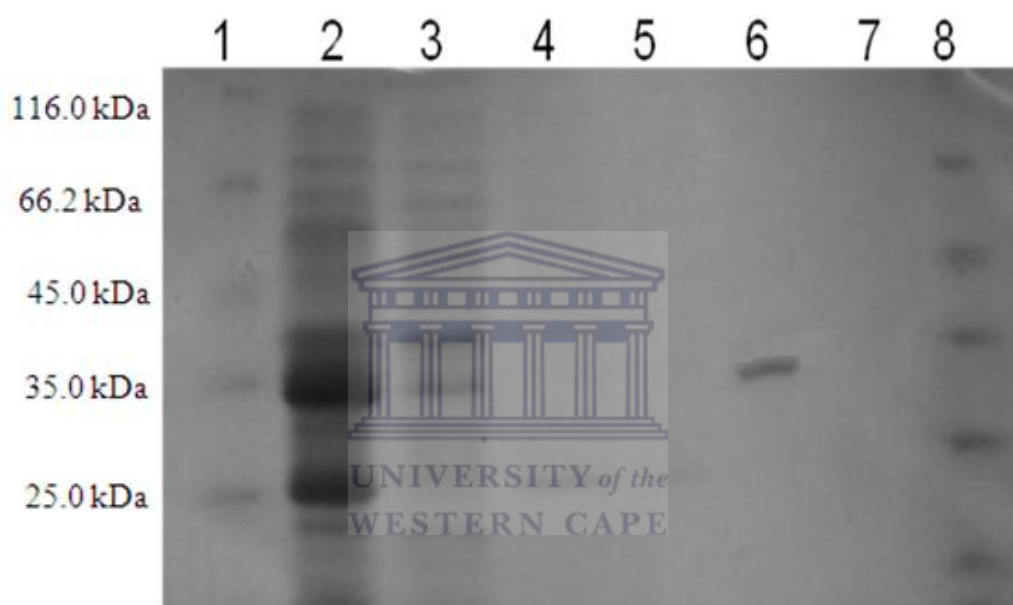


Figure 5.5 SDS-PAGE analysis of His-Tag purification of XPgene12-pet21a in *E. coli* Rosetta(DE3)pLysS

Lanes 1 and 8: protein molecular weight marker #SM0431 (Fermentas), lane 2: Total cell extract of XPgene12; Lane 3: Flow through eluate; Lane 4: Eluate from binding buffer; Lane 5: Eluate from washing buffer; Lane 6: Eluted XPgene12 protein (MW 37 kDa). Lane 7: Eluate from strip buffer.

Endoglucanase activity was confirmed by the presence of a zone of hydrolysis on a zymogram (Figure 5.6). The zone of clearance corresponded to the 37 kDa band on a SDS-PAGE gel run under the same conditions.

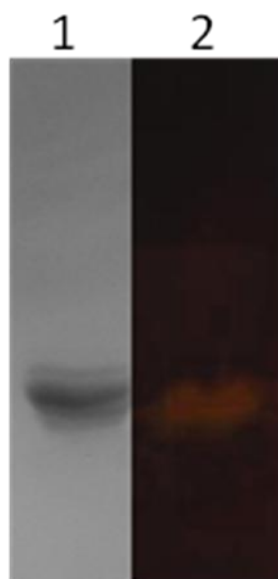
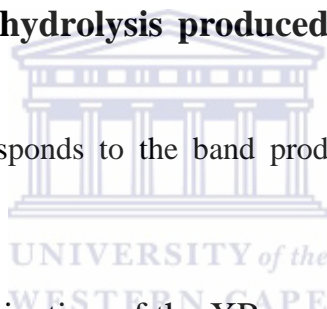


Figure 5.6 SDS-PAGE analysis of purified XPgene12 product (Lane 1) showing the zone of hydrolysis produced on a zymogram (Lane 2).

The zone of clearance corresponds to the band produced by the purified XPgene12 protein.



5.3.3 Enzymatic characterization of the XPgene12 gene product

Activity of the XPgene12 protein was tested over a pH range of 1-11 (Figure 5.7).

Little hydrolysis of the substrate CMC occurred under extreme acidic conditions.

The XPgene12 protein appears to be an acidophilic enzyme with an optimum activity at pH4. Relatively little activity was observed below pH2 and above pH6.

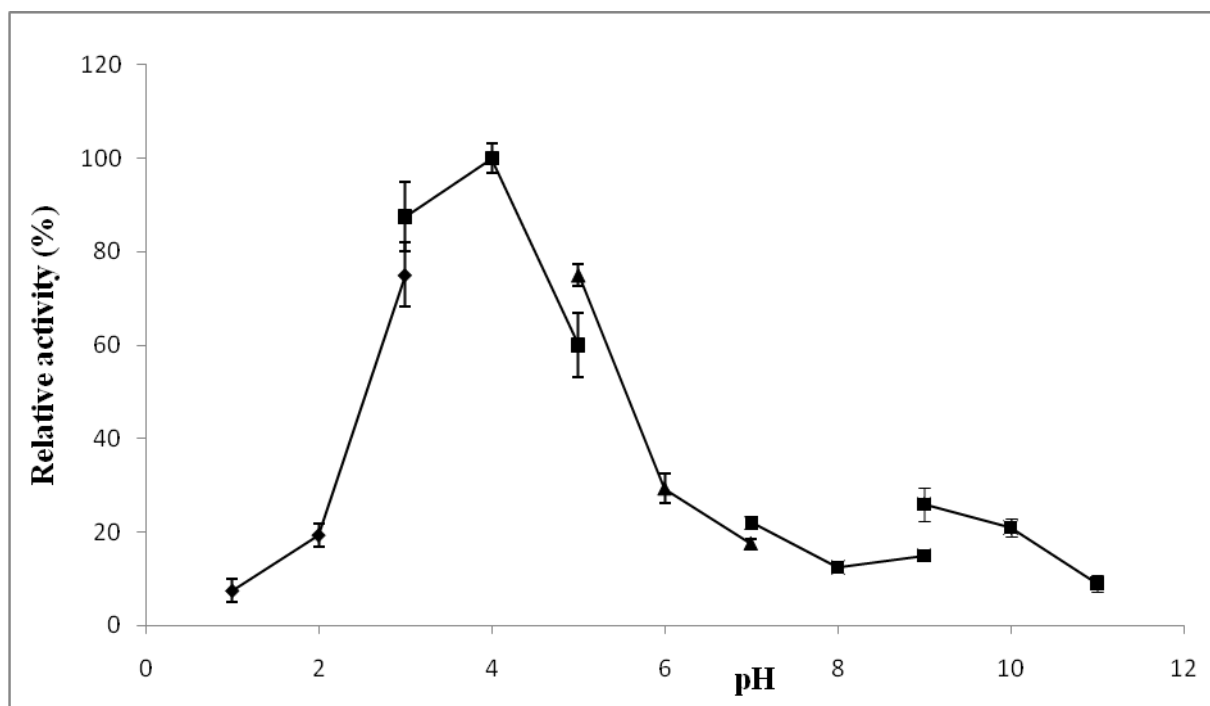


Figure 5.7 Effect of pH on XPgene12 protein activity with CMC as the substrate at 22°C

The maximum activity measured at pH4 was taken as 100%.

The effect of temperature on the activity of the XPgene12 protein with CMC as the substrate was determined at pH4. The enzyme was active over a wide temperature range (10-90°C) and displayed optimum activity at 50°C (Figure 5.8). The thermal inactivation profile of the enzyme is shown in Figure 5.9. XPgene12 protein was stable at 60°C, still exhibiting 67% activity after 1 hr pre-incubation at that temperature. Pre-incubation at 70°C for 15 mins stimulated activity. A similar profile was published for a thermophilic endoglucanase isolated from *Alicyclobacillus acidocaldarius* (Eckert & Schneider, 2003). After 1 hr pre incubation at 80°C, XPgene12 protein lost 70% activity (Figure 5.9).

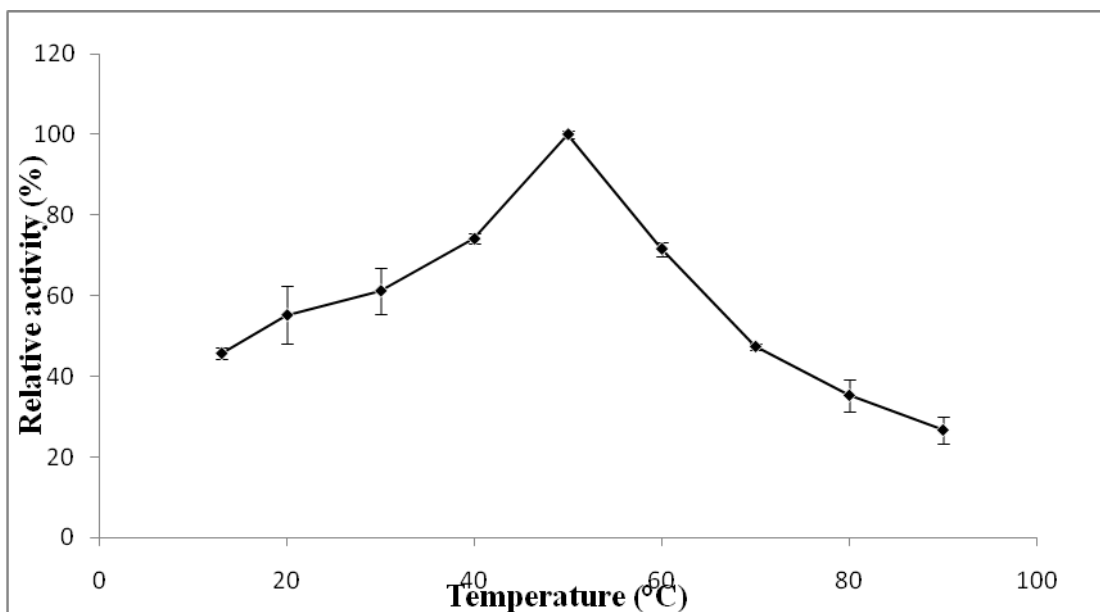


Figure 5.8 Effect of temperature on XPgene12 protein activity with CMC as substrate after 10mins incubation

The maximum activity at 50°C was taken as 100%.

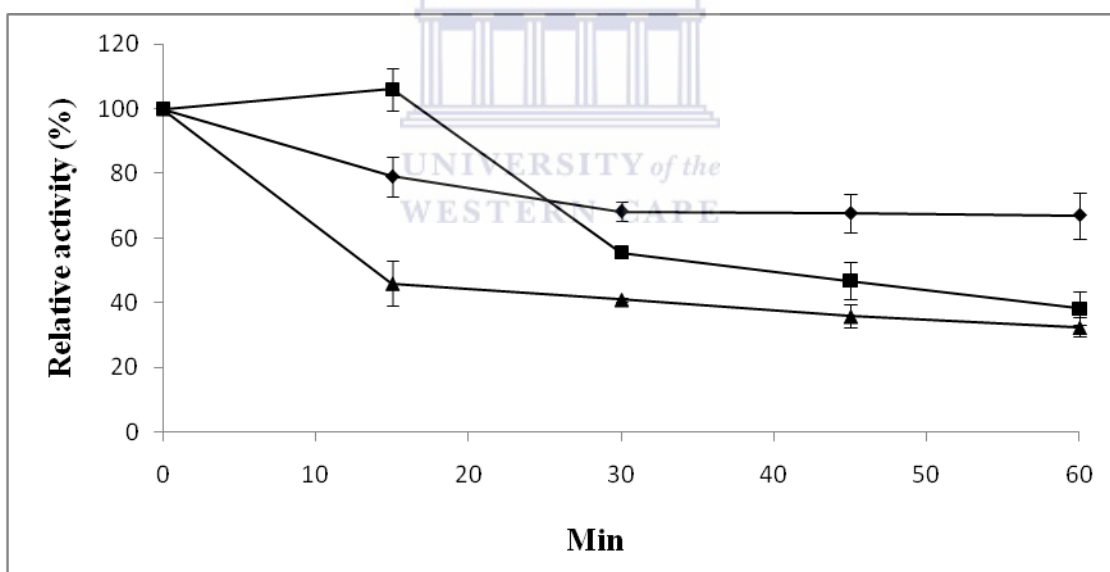


Figure 5.9 The thermal inactivation profile of XPgene12 product at 80°C (▲), 70°C (■) and 60°C (◆)

Activity was measured under optimum conditions (sodium acetate buffer of pH4, 50°C, and 10 mins) after the incubation of the enzyme at the indicated temperature for up to 60 mins.

XPgene12 protein studies were conducted using CMC as substrate, as it has been shown that endoglucanases prefer β -1, 4-linked glucans such as CMC (Voget *et al.*, 2006). Substrate specificity was tested with various other substrates (Figure 5.10).

Highest activity was observed with β -D-glucan from barley. Very low activity was observed with methyl-cellulose, while no activity could be measured with p-nitrophenyl- β -D-glucopyranoside (pNPG) and p-nitrophenyl- β -D-cellodioside (pNPC) as substrates. These results confirm that XPgene12 is an endo-1, 4-glucanase (Voget *et al.*, 2006; Feng *et al.*, 2007; Pang *et al.*, 2009).

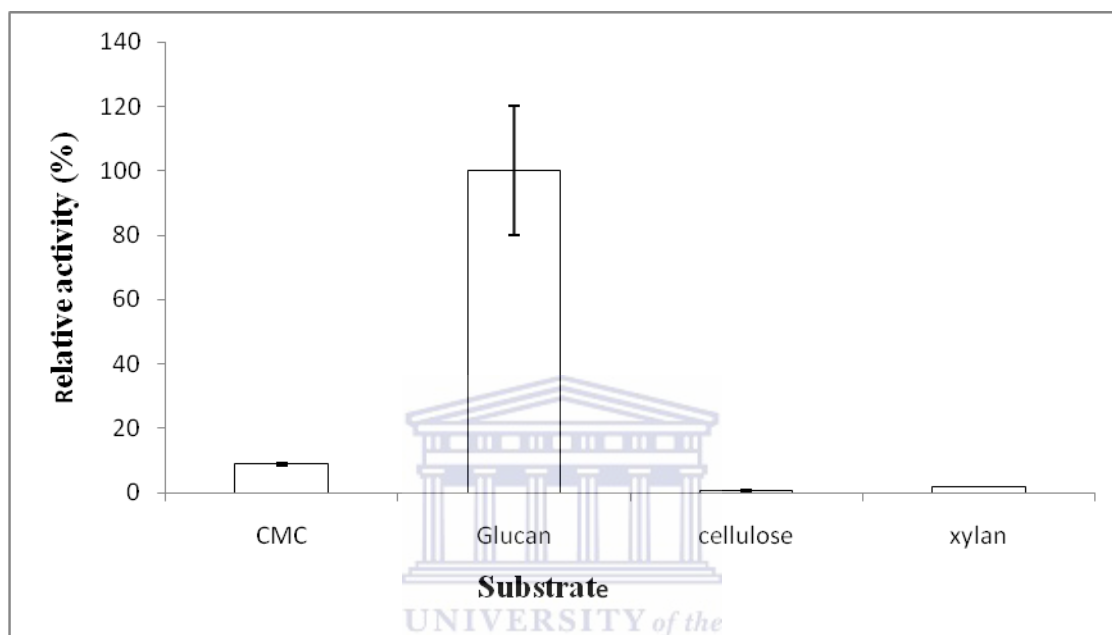


Figure 5.10 Activity of XPgene12 protein towards CMC, β -D-glucan, methyl-cellulose and xylan

The maximum activity measured with β -D-glucan as substrate was taken as 100%.

The preliminary kinetic parameters of recombinant XPgene12 protein were determined. Values of K_{cat} , K_m and V_{max} are shown in Table 5.2. The Lineweaver-Burk constants determined at optimal assay conditions with CMC as the substrate resulted in a V_{max} of 1085 U/mg enzyme and a K_m of 30.74 mM (4.2 mg/ml).

Table 5.2 : Kinetic parameters of the XPgene12 enzyme with CMC as a substrate

Substrate	K_m (mM)	V_{max} (Umg ⁻¹)	K_{cat} (S ⁻¹)	K_{cat}/K_m (mM ⁻¹ S ⁻¹)
CMC	30.74	1085	8.7	0.28

5.3.4 Thin layer chromatography

Hydrolysis of cello-oligosaccharides by the XPgene12 gene product was analyzed by TLC chromatography (Figure 5.11). No hydrolysis of G2, G3 and G4 cello-oligosaccharides was observed (Figure 5.11, Lanes 2, 4, 6). G5 was digested by the purified XPgene12 product. However the G5 control (lane 7) did not migrate to a single locus on the gel, possibly because of dissociation of G5 to a G5/G1-4 mixture. Incubation of G5 with the XPgene12 product demonstrated digestion although it was not clearly apparent to which products. The chromatogram shown in figure 5.11 was photographed approximately 1-2 hours after visualisation of the carbohydrate spots. Immediately after the visualisation process, a spot that was analogous to G3 (Lane 5) was more clearly visible although the corresponding G2 spot could not be readily identified. Based on figure 5.11 and other incubations of the enzyme and G5 (TLC, data not shown), a digestion event is likely to be occurring with G5, although the nature of the digestion is as yet not clearly defined.

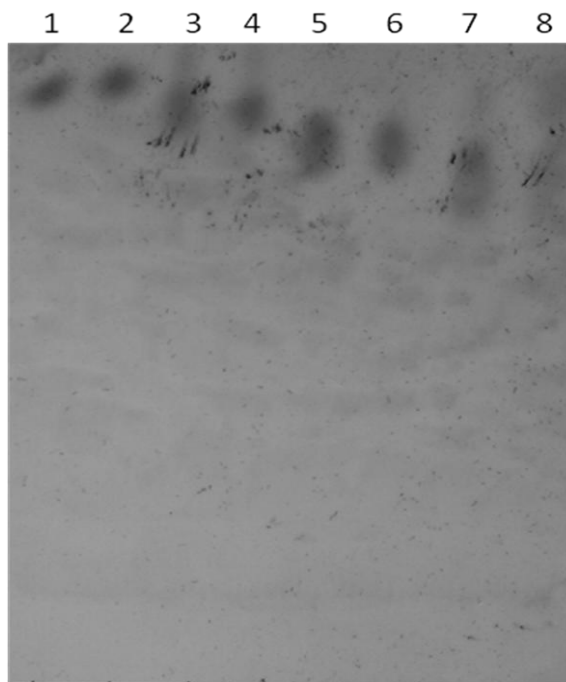


Figure 5.11 Hydrolysis products of cello-oligosaccharides by the purified XPgene12 product.

Lanes 1: standard sugar marker cellobiose (G2); Lane 3: standard sugar marker cellotriose (G3); Lane 5: standard sugar marker cellotetraose (G4); Lane 7: standard sugar marker cellopentaose (G5). Lanes 2, 4, 6, 8: purified XPgene12 product (2 μ g) hydrolysis of cellooligosaccharides after 24 hrs incubation at 50°C. Lane 8 is for G5.

WESTERN CAPE

Chapter 6 General discussion

The need for an environmentally sustainable energy source is stimulating increased commercial interest in renewable energy technologies. The most common renewable transport fuel today is bioethanol produced from corn and sugarcane. However, researchers are increasingly focusing on processes based on lignocellulosic substrates (second generation technology). The development of a flexible, efficient and cost effective hydrolysis step, combining multiple hydrolytic enzymes, plays a crucial role in developing a process-friendly fermentable feedstock. Thermophilic enzymes offer several advantages not commonly offered by mesophilic enzymes for the fermentation of cellulosic derivatives in biomass to ethanol (Thomas *et al.*, 1981).

Metagenomics is a rapidly growing field of research that has been successfully employed as a powerful tool for the discovery of enzymes with novel biocatalytic activities from unculturable microbial communities (Kennedy *et al.*, 2008). Coupled to this technique, highly efficient and low cost high-throughput screening techniques have been developed to facilitate the recovery of a large number of new biocatalysts and small molecules (Simon & Daniel, 2009).

In this study, high molecular weight DNA was isolated from the Mphizi hot springs Chiweta, Malawi. This metagenomic DNA was size selected, purified, cloned into a fosmid vector, and a metagenomic library was created. The coverage of the library was calculated to be 3.0×10^8 base pairs, equivalent to 100 bacterial genomes. End-sequencing analysis of fosmid clones (Chapter 3) confirmed that the library is a good representation of the Malawian hot spring prokaryotic biomass.

The library was grown on media containing CMC as a substrate and screened by subsequent flooding with Congo red to identify fosmid clones conferring cellulase activity. Seventeen clones were identified and two of these (008C2 and 026G5) were selected for sequencing. Selection criteria included high activity displayed under thermophilic conditions as determined by the DNS assay. The complete sequence of the insert of fosmid 008C2 was assembled from various contig sequences. A total of 29 predicted opening reading frames were identified which included two putative endoglucanases. Based on identification of conserved residues and motifs, these two genes were classified as members of the glycoside hydrolase family 8 (cellulase family D) (Henrissat *et al.*, 1989). Family members cleave cellooligosaccharide polymers that are at least five D-glucosyl subunits long (Alzari *et al.*, 1996).

Two genes (XPgene12 and XPgene25) were cloned into pET 21 a (+) and heterologously expressed in the *E. coli* expression strain Rosetta (DE3) pLysS. XPgene12 was expressed in the soluble fraction and purified to homogeneity using Ni-chelation chromatography (Section 5.2.4). XPgene25 was not fully characterised due to the time limitation. It is impossible without further analysis to compare activity contributions of the two enzymes derived from the fosmid XP008C2.

XPgene12 displayed maximum activity at pH4, suggesting that it is an acidophilic enzyme. However the enzyme was active over a wide pH range (pH3-pH9), making it a potentially useful resource for industrial applications. XPgene12 is more stable than other metagenome-derived cellulases characterized thus far (Healy *et al.*, 1995; Fontes *et al.*, 1997; van Solingen *et al.*, 2001; Voget *et al.*, 2006). XPgene12 displayed a broad thermal activity range with 61.2% of its maximal activity at 30°C and 71.6% at 60°C, with the maximum activity at 50°C. This result suggest that XPgene12 is a thermophilic enzyme (Taylor & Vaisman, 2010).

The preference of the XPgene12 protein for β -1, 4-linked glucans, and lack of activity with p-nitrophenyl- β -D-glucopyranoside and p-nitrophenyl- β -D-cellodioside as substrates confirm that the enzyme is an endo-1, 4-glucanase (MacLeod *et al.*, 1992; Chen *et al.*, 1994). A K_m value of 30.74mM and V_{max} of 1085U/mg was obtained for the XPgene12 protein, which compares favourably with the published values for a cellulase from uncultured microorganisms in rabbit cecum (V_{max} of 56.56 U/mg) (Feng *et al.*, 2007) and an endoglucanase from a metagenome library (V_{max} of 390 U/mg) (Voget *et al.*, 2006). XPgene12 protein has a greater specific activity and these characteristics, together with the pH and temperature optima and the thermostability of the enzyme make it a good candidate for industrial applications.

The low level of structural homology exhibited by the XPgene12 enzyme makes this enzyme an interesting candidate for crystallisation studies. Such endeavours may reveal subtle differences between related proteins that have not been solved structurally, since XPgene12 is a novel endoglucanase.

It would be interesting if the XPgene 25 and the nearest neighbour genes to XPgene 12 can be characterised during future studies.

Congress contributions (National and International)

CEDAD-IMPRS symposium. 18- 21/July/ 2010. Muenster, Germany. Oral presentation. Identification and Characterisation of Novel Extremophilic Genes Using Metagenomics

Cape Biotechnology Forum. 24-26/ March/ 2010. Cape Town, South African. Poster. Screening for Lignocellulosic Enzymes using a Metagenomic Approach

Thermophiles 2009 Conference. 16-22/ Sep/ 2009. Bei jing, China. Poster. Searching for Novel Thermophilic Cellulytic Genes Using Metagenomics

Publication

Hu. X.P., Taylor. M.P., Bauer. R., Tuffin, M & Cowan. D.A. Identification and characterization of a novel acidophilic cellulase using metagenomics (In preparation).



References

Alain, K., Olagnon, M., Desbruyeres, D., Page, A., Barbier, G., Juniper, S. K., Querellou, J. & Cambon-Bonavita, M. A. (2002). Phylogenetic characterization of the bacterial assemblage associated with mucous secretions of the hydrothermal vent polychaete *Paralvinella palmiformis*. *FEMS Microbiol Ecol* **42**, 463-476.

Alzari, P. M., Souchon, H. & Dominguez, R. (1996). The crystal structure of endoglucanase CelA, a family 8 glycosyl hydrolase from *Clostridium thermocellum*. *Structure* **4**, 265-275.

Amann, R. I., Ludwig, W. & Schleifer, K. H. (1995). Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *Microbiol Rev* **59**, 143-169.

Ando, S., Ishida, H., Kosugi, Y. & Ishikawa, K. (2002). Hyperthermostable endoglucanase from *Pyrococcus horikoshii*. *Appl Environ Microbiol* **68**, 430-433.

Bae, E. & Phillips, G. N., Jr. (2004). Structures and analysis of highly homologous psychrophilic, mesophilic, and thermophilic adenylate kinases. *J Biol Chem* **279**, 28202-28208.

Bayer, E. A., Shimon, L. J., Shoham, Y. & Lamed, R. (1998). Cellulosomes-structure and ultrastructure. *J Struct Biol* **124**, 221-234.

Beja, O., Suzuki, M. T., Koonin, E. V., Aravind, L., Hadd, A., Nguyen, L.P., Villacorta, R., Amjadi, M., Garrigues, C., Jovanovich, S.B., Feldman, R.A., DeLong, E.F. (2000). Construction and analysis of bacterial artificial chromosome libraries from a marine microbial assemblage. *Environ Microbiol* **2**, 516-529.

Beja, O. (2004). To BAC or not to BAC: marine ecogenomics. *Curr Opin Biotechnol* **15**, 187-190.

Berezovsky, I. N. & Shakhnovich, E. I. (2005). Physics and evolution of thermophilic adaptation. *Proc Natl Acad Sci U S A* **102**, 12742-12747.

Berthelet, M., Whyte, L. G. & Greer, C. W. (1996). Rapid, direct extraction of DNA from soils for PCR analysis using polyvinylpolypyrrolidone spin columns. *Fems Microbiology Letters* **138**, 17-22.

Bertrand, H., Poly, F., Van, V. T., Lombard, N., Nalin, R., Vogel, T. M. & Simonet, P. (2005). High molecular weight DNA recovery from soils prerequisite for biotechnological metagenomic library construction. *J Microbiol Methods* **62**, 1-11.

Besemer, J. & Borodovsky, M. (1999). Heuristic approach to deriving models for gene finding. *Nucleic Acids Res* **27**, 3911-3920.

Bhat, M. K. (2000). Cellulases and related enzymes in biotechnology. *Biotechnol Adv* **18**, 355-383.

Blumer-Schuetz, S. E., Kataeva, I., Westpheling, J., Adams, M. W. & Kelly, R. M. (2008). Extremely thermophilic microorganisms for biomass conversion: status and prospects. *Curr Opin Biotechnol* **19**, 210-217.

Boubakri, H., Beuf, M., Simonet, P. & Vogel, T. M. (2006). Development of metagenomic DNA shuffling for the construction of a xenobiotic gene. *Gene* **375**, 87-94.

Bradford, M.M. (1976). A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. *Anal Biochem* **72**: 248-254.

Brady, S. F., Chao, C. J., Handelsman, J. & Clardy, J. (2001). Cloning and heterologous expression of a natural product biosynthetic gene cluster from eDNA. *Org Lett* **3**, 1981-1984.

Camassola, M. & Dillon, A. J. (2007). Production of cellulases and hemicellulases by *Penicillium echinulatum* grown on pretreated sugar cane bagasse and wheat bran in solid-state fermentation. *J Appl Microbiol* **103**, 2196-2204.

Chen, H., Li, X. & Ljungdahl, L. G. (1994). Isolation and properties of an extracellular beta-glucosidase from the polycentric rumen fungus *Orpinomyces* sp. strain PC-2. *Appl Environ Microbiol* **60**, 64-70.

Chumpolkulwong, N., Sakamoto, K., Hayashi, A. & other authors (2006). Translation of 'rare' codons in a cell-free protein synthesis system from *Escherichia coli*. *J Struct Funct Genomics* **7**, 31-36.

Copeland A., Lucas S., Lapidus A., Barry K., Glavina del Rio T., Dalin E., Tice H., Bruce D.B., Goodwin L., Pitluck S., Saunders E., Brettin T., Detter J.C., Han C., Schmutz J., Larimer F., Land M., Hauser L., Kyrpides N., Mikhailova N., Nelson K., Gogarten J.P., Noll K., Richardson P. (2008) Complete sequence of *Thermotoga* sp. RQ2. Submitted to the EMBL/GenBank/DDBJ databases.

Coughlan, M.P. (1985). The production of fungal and bacterial cellulases with comment on their production and application. *Biotechnology & Genetic Engineering Reviews* **13**:39–109.

Cowan, D., Meyer, Q., Stafford, W., Muyanga, S., Cameron, R. & Wittwer, P. (2005). Metagenomic gene discovery: past, present and future. *Trends Biotechnol* **23**, 321-329.

Cowan, D. A. & Daniel, R. M. (1996). Rapid purification of two thermophilic proteinases using dye-ligand chromatography. *J Biochem Biophys Methods* **32**, 1-6.

Cowan, D. A., Arslanoglu, A., Burton, S. G., Baker, G. C., Cameron, R. A., Smith, J. J. & Meyer, Q. (2004). Metagenomics, gene discovery and the ideal biocatalyst. *Biochem Soc Trans* **32**, 298-302.

Daniel, R. (2005). The metagenomics of soil. *Nat Rev Microbiol* **3**, 470-478.

De Simone, G., Menchise, V., Manco, G., Mandrich, L., Sorrentino, N., Lang, D., Rossi, M. & Pedone, C. (2001). The crystal structure of a hyper-thermophilic carboxylesterase from the archaeon *Archaeoglobus fulgidus*. *J Mol Biol* **314**, 507-518.

de Vrije, T., Bakker, R. R., Budde, M. A., Lai, M. H., Mars, A. E. & Claassen, P. A. (2009). Efficient hydrogen production from the lignocellulosic energy crop *Miscanthus* by the extreme thermophilic bacteria *Caldicellulosiruptor saccharolyticus* and *Thermotoga neapolitana*. *Biotechnol Biofuels* **2**, 12.

Duan, C. J. & Feng, J. X. Mining metagenomes for novel cellulase genes. *Biotechnol Lett*. DOI 10.1007/s10529-010-0356-z. Published electronically prior to print.

Duan, C. J., Xian, L., Zhao, G. C., Feng, Y., Pang, H., Bai, X. L., Tang, J. L., Ma, Q. S. & Feng, J. X. (2009). Isolation and partial characterization of novel genes encoding acidic cellulases from metagenomes of buffalo rumens. *J Appl Microbiol* **107**, 245-256.

Eckert, K. & Schneider, E. (2003). A thermoacidophilic endoglucanase (CelB) from *Alicyclobacillus acidocaldarius* displays high sequence similarity to arabinofuranosidases belonging to family 51 of glycoside hydrolases. *Eur J Biochem* **270**, 3593-3602.

Elend, C., Schmeisser, C., Leggewie, C., Babiak, P., Carballeira, J. D., Steele, H. L., Reymond, J. L., Jaeger, K. E. & Streit, W. R. (2006). Isolation and biochemical characterization of two novel metagenome-derived esterases. *Applied and Environmental Microbiology* **72**, 3637-3645.

Eyers, L., George, I., Schuler, L., Stenuit, B., Agathos, S. N. & El Fantroussi, S. (2004). Environmental genomics: exploring the unmined richness of microbes to degrade xenobiotics. *Appl Microbiol Biotechnol* **66**, 123-130.

Feng, Y., Duan, C. J., Pang, H., Mo, X.C., Wu, C.F., Yu, Y., Hu, Y.L., Wei, J., Tang, J.L., Feng, J.X. (2007). Cloning and identification of novel cellulase genes from uncultured microorganisms in rabbit cecum and characterization of the expressed cellulases. *Appl Microbiol Biotechnol* **75**, 319-328.

Ferrer, M., Golyshina, O. V., Chernikova, T. N., Khachane, A.N., Reyes-Duarte, D., Santos, V.A., Strompl, C., Elborough, K., Jarvis, G., Neef, A., Yakimov, M.M., Timmis, K.N., Golyshin, P.N. (2005). Novel hydrolase diversity retrieved from a metagenome library of bovine rumen microflora. *Environ Microbiol* **7**, 1996-2010.

Ferrer, M., Golyshina, O., Beloqui, A. & Golyshin, P. N. (2007). Mining enzymes from extreme environments. *Curr Opin Microbiol* **10**, 207-214.

Fontes, C. M., Clarke, J. H., Hazlewood, G. P., Fernandes, T. H., Gilbert, H. J. & Ferreira, L. M. (1997). Possible roles for a non-modular, thermostable and proteinase-resistant cellulase from the mesophilic aerobic soil bacterium *Cellvibrio mixtus*. *Appl Microbiol Biotechnol* **48**, 473-479.

Fujinami, S. & Fujisawa, M. (2010) Industrial applications of alkaliphiles and their enzymes—past, present and future. *Environ Technol* **31**, 845-856.

Gabor, E. M., de Vries, E. J. & Janssen, D. B. (2003). Efficient recovery of environmental DNA for expression cloning by indirect extraction methods. *FEMS Microbiol Ecol* **44**, 153-163.

Gilkes, N. R., Henrissat, B., Kilburn, D. G., Miller, R. C., Jr. & Warren, R. A. (1991). Domains in microbial beta-1, 4-glycanases: sequence conservation, function, and enzyme families. *Microbiol Rev* **55**, 303-315.

Gillespie, D. E., Brady, S. F., Bettermann, A. D., Cianciotto, N. P., Liles, M. R., Rondon, M. R., Clardy, J., Goodman, R. M. & Handelsman, J. (2002). Isolation of antibiotics turbomycin A and B from a metagenomic library of soil microbial DNA. *Appl Environ Microbiol* **68**, 4301-4306.

Gomez, L. D., Steele-King, C. G. & McQueen-Mason, S. J. (2008a). Sustainable liquid biofuels from biomass: the writing's on the walls. *New Phytol* **178**, 473-485.

Gomez, L. D., Steele-King, C. G. & McQueen-Mason, S. J. (2008b). Sustainable liquid biofuels from biomass: the writing's on the walls. *New Phytologist* **178**, 473-485.

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, J. P. & Herwig, R. P. (1996). Phylogenetic analysis of the bacterial communities in marine sediments. *Appl Environ Microbiol* **62**, 4049-4059.

Hahn-Hagerdal, B., Galbe, M., Gorwa-Grauslund, M. F., Liden, G. & Zacchi, G. (2006). Bio-ethanol--the fuel of tomorrow from the residues of today. *Trends Biotechnol* **24**, 549-556.

Haki, G. D. & Rakshit, S. K. (2003). Developments in industrially important thermostable enzymes: a review. *Bioresour Technol* **89**, 17-34.

Hall, N. (2007). Advanced sequencing technologies and their wider impact in microbiology. *J Exp Biol* **210**, 1518-1525.

Handelsman, J. (2004). Metagenomics: application of genomics to uncultured microorganisms. *Microbiol Mol Biol Rev* **68**, 669-685.

Handelsman, J. (2005). Sorting out metagenomes. *Nat Biotechnol* **23**, 38-39.

Hansen, MC., Tolker-Nielsen, T., Givskov M & Molin, S. (1998). Biased 16S rDNA PCR amplification caused by interference from DNA flanking the template region. *FEMS Microbiol Ecol* **26**:141-149.

Hardeman, F. & Sjoling, S. (2007). Metagenomic approach for the isolation of a novel low-temperature-active lipase from uncultured bacteria of marine sediment. *FEMS Microbiol Ecol* **59**, 524-534.

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. *Appl Microbiol Biotechnol* **43**, 667-674.

Heath, C., Hu, X. P., Cary, S. C. & Cowan, D. (2009). Identification of a novel alkaliphilic esterase active at low temperatures by screening a metagenomic library from antarctic desert soil. *Appl Environ Microbiol* **75**, 4657-4659.

Henrissat, B., Claeysens, M., Tomme, P., Lemesle, L. & Mornon, J. P. (1989). Cellulase families revealed by hydrophobic cluster analysis. *Gene* **81**, 83-95.

Henrissat, B. (1991). A classification of glycosyl hydrolases based on amino acid sequence similarities. *Biochem J* **280**, 309-316.

Henrissat, B. & Bairoch, A. (1993). New families in the classification of glycosyl hydrolases based on amino acid sequence similarities. *Biochem J* **293**, 781-788.

Henrissat, B. & Bairoch, A. (1996). Updating the sequence-based classification of glycosyl hydrolases. *Biochem J* **316**, 695-696.

Hilbert, M., Bohm, G. & Jaenicke, R. (1993). Structural relationships of homologous proteins as a fundamental principle in homology modeling. *Proteins* **17**, 138-151.

Holben, W. E., Jansson, J. K., Chelm, B. K. & Tiedje, J. M. (1988). DNA probe method for the detection of specific microorganisms in the soil bacterial community. *Appl Environ Microbiol* **54**, 703-711.

Hough, D. W. & Danson, M. J. (1999). Extremozymes. *Curr Opin Chem Biol* **3**, 39-46.

Hu, Y. & Coates, A. R. (2005). Transposon mutagenesis identifies genes which control antimicrobial drug tolerance in stationary-phase *Escherichia coli*. *FEMS Microbiol Lett* **243**, 117-124.

Huang, Y., Lai, X., He, X., Cao, L., Zeng, Z., Zhang, J. & Zhou, S. (2009). Characterization of a deep-sea sediment metagenomic clone that produces water-soluble melanin in *Escherichia coli*. *Mar Biotechnol (NY)* **11**, 124-131.

Hubbert, M.K. (1956). "Nuclear Energy and the Fossil Fuels 'Drilling and Production Practice'" . Spring Meeting of the Southern District. Division of Production. American Petroleum Institute. San Antonio, Texas: Shell Development Company. pp. 22–27.

Jacobsen, C. S. & Rasmussen, O. F. (1992). Development and application of a new method to extract bacterial DNA from soil based on separation of bacteria from soil with cation-exchange resin. *Appl Environ Microbiol* **58**, 2458-2462.

Jeczminek, L., Oleksiak, S., Skret, I. & Marchut, A. (2006). The second-generation biofuels. *Przemysl Chemiczny* **85**, 1570-1574.

Jones, D. T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol* **292**, 195-202.

Jones, M. D. & Foulkes, N. S. (1989). Reverse transcription of mRNA by *Thermus aquaticus* DNA polymerase. *Nucleic Acids Res* **17**, 8387-8388.

Kasana, R. C., Salwan, R., Dhar, H., Dutt, S. & Gulati, A. (2008). A rapid and easy method for the detection of microbial cellulases on agar plates using gram's iodine. *Curr Microbiol* **57**, 503-507.

Kengen, S. W., Luesink, E. J., Stams, A. J. & Zehnder, A. J. (1993). Purification and characterization of an extremely thermostable beta-glucosidase from the hyperthermophilic archaeon *Pyrococcus furiosus*. *Eur J Biochem* **213**, 305-312.

Kennedy, J., Marchesi, J. R. & Dobson, A. D. (2008). Marine metagenomics: strategies for the discovery of novel enzymes with biotechnological applications from marine environments. *Microb Cell Fact* **7**, 27.

Kim, J. Y., Hur, S. H. & Hong, J. H. (2005). Purification and characterization of an alkaline cellulase from a newly isolated alkalophilic *Bacillus sp.* HSH-810. *Biotechnol Lett* **27**, 313-316.

Kim, S. J., Lee, C. M., Han, B. R., Kim, M. Y., Yeo, Y. S., Yoon, S. H., Koo, B. S. & Jun, H. K. (2008). Characterization of a gene encoding cellulase from uncultured soil bacteria. *FEMS Microbiol Lett* **282**, 44-51.

- Kimura, N. (2006).** Metagenomics: Access to unculturable microbes in the environment. *Microbes and Environments* **(21)** 201-215.
- Koh, L. P., Levang, P. & Ghazoul, J. (2009).** Designer landscapes for sustainable biofuels. *Trends Ecol Evol* **24**, 431-438.
- Koskinen, P. E. P., Lay, C.-H., Beck, S. R., Tolvanen, K. E. S., Kaksonen, A. H., Årlygsson, J. h., Lin, C.-Y. & Puhakka, J. A. (2007).** Bioprospecting thermophilic microorganisms from Icelandic hot springs for hydrogen and ethanol production. *Energy & Fuels* **22**, 134-140.
- Krsek, M. & Wellington, E. M. (1999).** Comparison of different methods for the isolation and purification of total community DNA from soil. *J Microbiol Methods* **39**, 1-16.
- Kumar, R., Singh, S. & Singh, O. V. (2008).** Bioconversion of lignocellulosic biomass: biochemical and molecular perspectives. *J Ind Microbiol Biotechnol* **35**, 377-391.
- Lamed, R., Setter, E. & Bayer, E. A. (1983).** Characterization of a cellulose-binding, cellulase-containing complex in *Clostridium thermocellum*. *J Bacteriol* **156**, 828-836.
- Lammle, K., Zipper, H., Breuer, M., Hauer, B., Buta, C., Brunner, H. & Rupp, S. (2007).** Identification of novel enzymes with different hydrolytic activities by metagenome expression cloning. *J Biotechnol* **127**, 575-592.
- Langer, M., Gabor, E. M., Liebeton, K., Meurer, G., Niehaus, F., Schulze, R., Eck, J. & Lorenz, P. (2006).** Metagenomics: an inexhaustible access to nature's diversity. *Biotechnol J* **1**, 815-821.
- Lee, S. W., Won, K., Lim, H. K., Kim, J. C., Choi, G. J. & Cho, K. Y. (2004).** Screening for novel lipolytic enzymes from uncultured soil microorganisms. *Appl Microbiol Biotechnol* **65**, 720-726.
- Li, X., Weng, J. K. & Chapple, C. (2008).** Improvement of biomass through lignin modification. *Plant J* **54**, 569-581.
- Liebeton, K. & Eck, J. (2004).** Identification and Expression in *E. coli* of Novel Nitrile Hydratases from the Metagenome. *Engineering in Life Sciences* **4**: 557-562.
- Lineweaver, H and Burk, D. (1934).** The determination of enzyme dissociation constants. *Journal of the American Chemical Society* **56**, 658-666.
- Lopez, M. J., Vargas-Garcia, M. D., Suarez-Estrella, F., Nichols, N. N., Dien, B. S. & Moreno, J. (2006).** Lignocellulose-degrading enzymes produced by the ascomycete *Coniochaeta ligniaria* and related species: Application for a lignocellulosic substrate treatment. *Enzyme and Microbial Technology* **40**, 794-800.
- Lorenz, P & Schleper, C. (2002).** Metagenome - a challenging source of enzyme discovery. *Journal of Molecular Catalysis B: Enzymatic* **19-20**: 13-19.
- Lukashin, A. V. & Borodovsky, M. (1998).** GeneMark.hmm: new solutions for gene finding. *Nucleic Acids Res* **26**, 1107-1115.
- MacLeod, A. M., Gilkes, N. R., Escote-Carlson, L., Warren, R. A., Kilburn, D. G. & Miller, R. C., Jr. (1992).** *Streptomyces lividans* glycosylates an exoglucanase (Cex) from *Cellulomonas fimi*. *Gene* **121**, 143-147.

- Madigan, M.T. & Martinko, J.M. (2006).** Brock Biology of Microorganisms, 11th edition. Pearson Prentice Hall, Upper Saddle River, NJ. ISBN 0-13-144329-1.
- McGuffin, L. J., Bryson, K. & Jones, D. T. (2000).** The PSIPRED protein structure prediction server. *Bioinformatics* **16**, 404-405.
- Messing, J. (1983).** New M13 vectors for cloning. *Methods Enzymol.* **101**: 20–78
- Miller, G.L. (1959)** Use of dinitrosalicylic acid reagent for determination of reducing sugar. *Analytical Chemistry.* **31**, 426-428.
- Miller, D. N., Bryant, J. E., Madsen, E. L. & Ghiorse, W. C. (1999).** Evaluation and optimization of DNA extraction and purification procedures for soil and sediment samples. *Appl Environ Microbiol* **65**, 4715-4724.
- Morrison, M., Pope, P. B., Denman, S. E. & McSweeney, C. S. (2009).** Plant biomass degradation by gut microbiomes: more of the same or something new? *Curr Opin Biotechnol* **20**, 358-363.
- Mosier, N., Wyman, C., Dale, B., Elander, R., Lee, Y. Y., Holtzapple, M. & Ladisch, M. (2005).** Features of promising technologies for pretreatment of lignocellulosic biomass. *Bioresour Technol* **96**, 673-686.
- Mussatto, S. I., Fernandes, M., Milagres, A. M. F. & Roberto, I. C. (2008).** Effect of hemicellulose and lignin on enzymatic hydrolysis of cellulose from brewer's spent grain. *Enzyme and Microbial Technology* **43**, 124-129.
- Muyzer, G., de Waal, E. C. & Uitterlinden, A. G. (1993).** Profiling of complex microbial populations by denaturing gradient gel electrophoresis analysis of polymerase chain reaction-amplified genes coding for 16S rRNA. *Appl Environ Microbiol* **59**, 695-700.
- Nakashima, N., Mitani, Y. & Tamura, T. (2005).** Actinomycetes as host cells for production of recombinant proteins. *Microb Cell Fact* **4**, 7.
- Nobutada. K. (2006).** Metagenomics: Access to unculturable microbes in the environment. *Microbes and Environments* **21**, 201-215.
- Ohgren, K., Bura, R., Saddler, J. & Zacchi, G. (2007).** Effect of hemicellulose and lignin removal on enzymatic hydrolysis of steam pretreated corn stover. *Bioresour Technol* **98**, 2503-2510.
- Pachter, L. (2007).** Interpreting the unculturable majority. *Nat Methods* **4**, 479-480.
- Palomo, J. M., Segura, R. L., Fernandez-Lorente, G., Pernas, M., Rua, M. L., Guisan, J. M. & Fernandez-Lafuente, R. (2004).** Purification, immobilization, and stabilization of a lipase from *Bacillus thermocatenuatus* by interfacial adsorption on hydrophobic supports. *Biotechnol Prog* **20**, 630-635.
- Pang, M-F., Abdullah, N., Lee, C-W & Ng, C-C. (2008).** Isolation of high molecular weight DNA from forest topsoil for metagenomic analysis. *Asia Pacific Journal of Molecular Biology and Biotechnology.* **16 (2)**, 35-41.
- 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. *Curr Microbiol* **58**, 404-408.

Pantazaki, A. A., Pritsa, A. A. & Kyriakidis, D. A. (2002). Biotechnologically relevant enzymes from *Thermus thermophilus*. *Appl Microbiol Biotechnol* **58**, 1-12.

Parsiegla, G., Belaich, A., Belaich, J. P. & Haser, R. (2002). Crystal structure of the cellulase Cel9M enlightens structure/function relationships of the variable catalytic modules in glycoside hydrolases. *Biochemistry* **41**, 11134-11142.

Percival Zhang, Y. H., Himmel, M. E. & Mielenz, J. R. (2006). Outlook for cellulase improvement: screening and selection strategies. *Biotechnol Adv* **24**, 452-481.

Ramachandran, G. N., Ramakrishnan, C. & Sasisekharan, V. (1963). Stereochemistry of polypeptide chain configurations. *J Mol Biol* **7**, 95-99.

Ramachandran, G. N. & Sasisekharan, V. (1968). Conformation of polypeptides and proteins. *Adv Protein Chem* **23**, 283-438.

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.

Reysenbach, A. L., Hamamura, N., Podar, M., Griffiths, E., Ferreira, S., Hochstein, R., Heidelberg, J., Johnson, J., Mead, D., Pohorille, A., Sarmiento, M., Schweighofer, K., Seshadri, R., Voytek, M.A. (2009). Complete and draft genome sequences of six members of the Aquificales. *J Bacteriol* **191**, 1992-1993.

Reysenbach, A.-L. & Pace, N. R. (1995). Reliable amplification of hyperthermophilic archaeal 16S rRNA genes by the polymerase chain reaction. In *Archaea: a Laboratory Manual – Thermophiles*. Edited by F. T. Robb & A. R. Place. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory.

Richmond, T. A. & Somerville, C. R. (2000). The cellulose synthase superfamily. *Plant Physiol* **124**, 495-498.

Roose-Amsaleg, C. L., Garnier-Sillam, E. & Harry, M. (2001). Extraction and purification of microbial DNA from soil and sediment samples. *Applied Soil Ecology* **18**, 47-60.

Saha, B. C. (2003). Hemicellulose bioconversion. *Journal of Industrial Microbiology & Biotechnology* **30**, 279-291.

Saitou, N. & Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* **4**, 406-425.

Sanger, F. & Coulson, A. R. (1975). A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J Mol Biol* **94**, 441-448.

Scheller, H. V. & Ulvskov, P. (2010). Hemicelluloses. *Annu Rev Plant Biol* **61**, 263-289.

Schluter, A., Bekel, T., Diaz, N. N. & other authors (2008). The metagenome of a biogas-producing microbial community of a production-scale biogas plant fermenter analysed by the 454-pyrosequencing technology. *J Biotechnol* **136**, 77-90.

Schmeisser, C., Steele, H. & Streit, W. R. (2007). Metagenomics, biotechnology with non-culturable microbes. *Appl Microbiol Biotechnol* **75**, 955-962.

Schwarz, W. H. (2001). The cellulosome and cellulose degradation by anaerobic bacteria. *Appl Microbiol Biotechnol* **56**, 634-649.

Simon, C. & Daniel, R. (2009). Achievements and new knowledge unraveled by metagenomic approaches. *Appl Microbiol Biotechnol* **85**, 265-276.

Sims, R. E. W., Mabee, W., Saddler, J. N. & Taylor, M. (2010). An overview of second generation biofuel technologies. *Bioresour Technol* **101**, 1570-1580.

Sommer, P., Georgieva, T. & Ahring, B. K. (2004). Potential for using thermophilic anaerobic bacteria for bioethanol production from hemicellulose. *Biochemical Society Transactions* **32**, 283-289.

Stach, J. E. M., Bathe, S., Clapp, J. P. & Burns, R. G. (2001). PCR-SSCP comparison of 16S rDNA sequence diversity in soil DNA obtained using different isolation and purification methods. *Fems Microbiology Ecology* **36**, 139-151.

Steele, H. L. & Streit, W. R. (2005). Metagenomics: advances in ecology and biotechnology. *FEMs Microbiology Letters* **247**, 105-111.

Sterner, R. & Liebl, W. (2001). Thermophilic adaptation of proteins. *Critical Reviews in Biochemistry and Molecular Biology* **36**, 39-106.

Stetter, K. O. (1996). Hyperthermophiles in the history of life. *Evolution of Hydrothermal Ecosystems on Earth (and Mars?)* **202**, 1-18.

Sun, Y. & Cheng, J. (2002). Hydrolysis of lignocellulosic materials for ethanol production: a review. *Bioresour Technol* **83**, 1-11.

Szczodrak, J. & Fiedurek, J. (1996). Technology for conversion of lignocellulosic biomass to ethanol. *Biomass & Bioenergy* **10**, 367-375.

Taylor, M. P. (2007). Metabolic engineering of *Geobacillus* species for enhanced ethanol production. Unpublished doctoral dissertation. Imperial College of Science, Technology and Medicine, University of London, UK.

Taylor, T. J. & Vaisman, II (2010). Discrimination of thermophilic and mesophilic proteins. *BMC Struct Biol* **10 Suppl 1**, S5.

Teather, R. M. & Wood, P. J. (1982). Use of Congo red-polysaccharide interactions in enumeration and characterization of cellulolytic bacteria from the bovine rumen. *Appl Environ Microbiol* **43**, 777-780.

Tebbe, C. C. & Vahjen, W. (1993). Interference of humic acids and DNA extracted directly from soil in detection and transformation of recombinant DNA from bacteria and a yeast. *Appl Environ Microbiol* **59**, 2657-2665.

Thomas, K. N. G., Arie, B. B & Zeikus, J. G. (1981). Ethanol production by thermophilic Bacteria: fermentation of cellulosic substrates by cocultures of *Clostridium thermocellum*

and *Clostridium thermohydrosulfuricum*. *Applied and Environmental Microbiology* **41**: 1337-1343.

Tilman, D., Hill, J. & Lehman, C. (2006). Carbon-negative biofuels from low-input high-diversity grassland biomass. *Science* **314**, 1598-1600.

Treusch, A. H., Kletzin, A., Raddatz, G., Ochsenreiter, T., Quaiser, A., Meurer, G., Schuster, S. C. & Schleper, C. (2004). Characterization of large-insert DNA libraries from soil for environmental genomic studies of Archaea. *Environ Microbiol* **6**, 970-980.

van de Werken, H. J., Verhaart, M. R., VanFossen, A. L., Willquist, K, Lewis DL, Nichols JD, Goorissen HP, Mongodin EF, Nelson KE, van Niel EW, Stams AJ, Ward DE, de Vos WM, van der Oost J, Kelly RM, Kengen SW. (2008). Hydrogenomics of the extremely thermophilic bacterium *Caldicellulosiruptor saccharolyticus*. *Appl Environ Microbiol* **74**, 6720-6729.

van den Burg, B. (2003). Extremophiles as a source for novel enzymes. *Current Opinion in Microbiology* **6**, 213-218.

van den Burg, B. & Eijsink, V. G. (2002). Selection of mutations for increased protein stability. *Curr Opin Biotechnol* **13**, 333-337.

van Solingen, P., Meijer, D., van der Kleij, W. A., Barnett, C., Bolle, R., Power, S. D. & Jones, B. E. (2001). Cloning and expression of an endocellulase gene from a novel streptomyces isolated from an East African soda lake. *Extremophiles* **5**, 333-341.

van Wyk, J. P. (2001). Biotechnology and the utilization of biowaste as a resource for bioproduct development. *Trends Biotechnol* **19**, 172-177.

Venclovas, C. & Margelevicius, M. (2005). Comparative modeling in CASP6 using consensus approach to template selection, sequence-structure alignment, and structure assessment. *Proteins* **61**, 99-105.

Vieille, C. & Zeikus, G. J. (2001). Hyperthermophilic enzymes: sources, uses, and molecular mechanisms for thermostability. *Microbiol Mol Biol Rev* **65**, 1-43.

Viikari, L & Siila-aho, M. (2006). " New thermophilic enzyme systems for biorefineries". *European conference on biorefinery research*. Helsinki, Lund University.

Voget, S., Steele, H. L. & Streit, W. R. (2006). Characterization of a metagenome-derived halotolerant cellulase. *Journal of Biotechnology* **126**, 26-36.

Wang, F., Li, F., Chen, G. & Liu, W. (2009). Isolation and characterization of novel cellulase genes from uncultured microorganisms in different environmental niches. *Microbiol Res* **164**, 650-657.

Watson, B. J., Zhang, H., Longmire, A. G., Moon, Y. H. & Hutcheson, S. W. (2009). Processive endoglucanases mediate degradation of cellulose by *Saccharophagus degradans*. *J Bacteriol* **191**, 5697-5705.

Wicker, T., Schlagenhauf, E., Graner, A., Close, T. J., Keller, B. & Stein, N. (2006). 454 sequencing put to the test using the complex genome of barley. *BMC Genomics* **7**, 275.

Yasutake, Y., Kawano, S., Tajima, K., Yao, M., Satoh, Y., Munekata, M. & Tanaka, I. (2006). Structural characterization of the *Acetobacter xylinum* endo-beta-1,4-glucanase CMCax required for cellulose biosynthesis. *Proteins* **64**, 1069-1077.

Yun, J. & Ryu, S. (2005). Screening for novel enzymes from metagenome and SIGEX, as a way to improve it. *Microb Cell Fact* **4**, 8.

Zhang, Y. H-P., Himmel, M. E. & Mielenz, J. R. (2006). Outlook for cellulase improvement: screening and selection strategies. *Biotechnol Adv* **24**, 452-481.

Zhaxybayeva, O., Swithers, K. S., Lapierre, P. & other authors (2009). On the chimeric nature, thermophilic origin, and phylogenetic placement of the Thermotogales. *Proc Natl Acad Sci U S A* **106**, 5865-5870.

Zhou, J., Bruns, M. A. & Tiedje, J. M. (1996). DNA recovery from soils of diverse composition. *Appl Environ Microbiol* **62**, 316-322.

