

**Structural Analysis of Induced Mutagenesis A' Protein from
Mycobacterium tuberculosis and of a Thermophilic GH9 Cellulase**



2814902

**This thesis is submitted to the Department of Biotechnology, University of the Western
Cape, in fulfilment of the requirements for a Master of Science degree in Biotechnology**

Submission Date: November, 2014

Supervisor: Prof. Wolf-Dieter Schubert (University of the Western Cape)

Co-Supervisors: A/Prof. Digby Warner (University of Cape Town)

Preamble

The Thesis “Structural Analysis of Induced Mutagenesis A’ Protein from *Mycobacterium tuberculosis* and of a Thermophilic GH9 Cellulase” consists of two parts, I and II that share one abstract.



Abstract

The three-dimensional structures of proteins are important in understanding their function and interaction with ligands and other proteins. In this work, the structures of two proteins, ImuA' from *Mycobacterium tuberculosis* and GH9 C1 cellulase from a metagenomic library, were analysed using structural biological and modelling techniques. The gene encoding ImuA' was amplified by two-step PCR, cloned, and expressed in *E. coli*. The recombinant ImuA' produced was found to be largely insoluble. The insoluble protein was successfully solubilized in 8 M urea but refolding the protein to its native structure was unsuccessful. By homology modelling, a 3D model of ImuA' was obtained from a partly homologous protein RecA. In comparison to RecA, ImuA' appears to lack some loop amino acids critical for DNA binding. Hence ImuA' is postulated to not bind DNA. The second protein, GH9 C1 cellulase, was produced in *E. coli*. The protein was purified by chromatographic techniques and crystallized in a precipitant to protein ratio of 1:2 by hanging and sitting drop crystallization methods. The reservoir solution was made up of 15-30% (w/v) PEG 3350, 200 mM salt and 100 mM Tris-HCl pH 7.5-8.5. The protein crystals only diffracted X-rays to 4 Å resolution which could not be used to obtain a crystal structure of the protein. The diffraction data, however, showed the crystal to be monoclinic with space group P2. Homology modelling revealed GH9 C1 cellulase to be a two domain protein with a smaller N-terminal Ig-like domain and a larger catalytic domain. The catalytic domain retains two Ca²⁺ binding sites, which potentially stabilize the active site conformation and increase thermostability of the protein. Overall GH9 C1 cellulase is structurally similar to other GH9 cellulases, suggesting that its catalytic mechanism may be conserved.

Key Words

Mycobacterium tuberculosis

DNA damage

ImuA'-ImuB-DnaE2

GH9 cellulases

Crystallization

Homology modelling



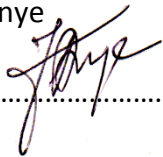
Declaration

I declare that “*Structural Analysis of Induced Mutagenesis A’ Protein from Mycobacterium tuberculosis and of a Thermophilic GH9 Cellulase*” is my own work, that it has not been submitted for any degree or examination at any other university, and that all the sources I have used or quoted have been indicated and acknowledged by complete references.

1st November, 2014

Valentine Anye

Signed.....



Date:



Acknowledgements

This academic endeavour would not have been successful without the unflagging, meticulous, insightful and financial support of my supervisor, Prof Wolf-Dieter Schubert. I owe a great deal to him for my academic development and success. Secondly, I acknowledge the inputs from Prof Trevor Sewell and Dr Brandon Weber of the Electron Microscopy Unit at the University of Cape Town. They allowed me to freely use their laboratory and continuously guided my research at the time my supervisor and colleagues relocated to University of Pretoria. I would not forget the contributions from my Co-supervisor Dr Digby Warner. I extend gratitude to Fru Azinwi, Jacob Cloette, Prudence Salame, Tephney Hutchinson and Thuso Mapotsane who edited sections of my thesis. My appreciation also goes to my colleagues in the Structural Biology of Infectious Diseases group, Donné Simpson, Rowan Julian, Clive Mketsu, Taariq Firfirey, Jeremy Boonzaier, Mujaahidah Mohamed, Lungelo Mandyoli, Mthawelanga Ndegane, Clifford Ntui, Zhuo Fang and Thuso Mapotsane for their cooperation in the laboratory and critique of my outputs during laboratory meetings. I also acknowledge the support of Portia Maumela of the NMR research group.

Special thanks goes to my newly found family, members of the Students' Christian Organisation, who laboured with me in prayers, gave emotion and spiritual support when things were tough. To my uncles Charles Fru Chi and Denis Ndoh Chi, and the rest of my family members especially my grandmother Mangie Trepina Sirri, mother Mangie Doris Nchang and brother Henry Nji. I thank you for always being there and believing that I could make it. The encouraging and loving support from, Miss Akwa Claudette, Siphokazi Makeleni, Nomcebo Zitha, Romeo Dube and Moffat Hassani cannot be forgotten.

Finally and most importantly, I give thanks and render all the glory for the successful completion of this work to God almighty through his son Jesus Christ for making me capable to do the work and for being there throughout.

Table of Content

Structural Analysis of Induced Mutagenesis A' Protein from <i>Mycobacterium tuberculosis</i> and of a Thermophilic GH9 Cellulase	i
Abstract	iii
Key Words	iv
Declaration	v
Signed.....	v
Acknowledgements	vi
Table of Content	vii
List of Abbreviations	xii
List of Figures	xiv
List of Tables	xvii
Part I	xviii
1. Introduction	1
1.1 History and Epidemiology of Tuberculosis	1
1.2 Tuberculosis Treatment	2
1.3 Management of Tuberculosis	3
1.4 Immunology of Tuberculosis Infection	3
1.5 Mycobacterium tuberculosis DNA Damage	4
1.6 DNA Replication	5
1.6.1 Eukaryotic and Prokaryotic DNA Polymerases	5
1.6.2 Prokaryotic Pol III Enzymes	6
1.6.3 The SOS Response Mechanism	6
1.6.4 Y-family Polymerases	8
1.6.5 Y-family Polymerases Versus Replicative Polymerases	8
1.6.6 <i>E. coli</i> Y-family Polymerases and their Orthologues in Other Bacteria	9
1.6.7 The DnaE Proteins.....	10
1.6.8 The DnaE2 (ImuC) Accessory Proteins ImuA and ImuB	10
1.6.9 The <i>Mtb imuA', imuB, imuC</i> Mutagenic Cassette	11
1.7 Aim and Objectives	12
2. Materials and Methods	13

2.1 General Chemicals and Enzymes	13
2.2 Stock Solutions, Buffers and Media	14
2.3 Plasmids and Bacterial Strains	15
2.3.1 Plasmids and their Properties.....	15
2.3.2 Bacterial Strains	15
2.4 Naming of <i>imuA'S</i> and <i>imuA'L</i>	16
2.5 Cloning	16
2.5.1 Primers for Amplification of <i>imuA'S</i> and <i>imuA'L</i> by Polymerase Chain Reaction (PCR).....	16
2.5.2 PCR Amplification of <i>imuA'S</i> and <i>imuA'L</i>	17
2.5.3 Agarose Gel Electrophoresis of DNA	18
2.5.4 Extraction and Purification of DNA from Agarose Gels	19
2.5.5 Restriction Enzyme Digestion of DNA.....	19
2.5.6 Ligation of DNA Molecules.....	19
2.5.7 Quantification of DNA.....	20
2.5.8 Preparation of Chemically Competent <i>E. coli</i> Cells.....	20
2.5.9 Antibiotic Selection	21
2.5.10 Bacterial Transformation	21
2.5.11 Colony Screening by Restriction Digest Analysis	21
2.5.12 Recombinant Plasmids.....	22
2.6 Recombinant Protein Production	22
2.6.1 Small Scale Protein Production Test	23
2.6.2 Harvesting Cells.....	23
2.6.3 Cell Rupture	24
2.6.4 Sodium Dodecyl Sulphate Polyacrylamide Gel Electrophoresis (SDS-PAGE) Analysis..	24
2.7 Optimization for Soluble Protein Production	25
2.7.1 Optimization of Culturing Temperature	25
2.7.2 Optimization of Gene Expression by IPTG	25
2.7.3 <i>ImuA'</i> Production in Optimized <i>E. coli</i> Strains	26
2.7.4 Optimization of Lysis Buffer pH	26

2.8 Solubilization of Inclusion Bodies and Protein Refolding	27
2.8.1 Inclusion Body Preparation.....	27
2.8.2 Denaturation of Inclusion Body Proteins.....	27
2.8.3 Protein Refolding by Dialysis.....	27
2.8.4 On-column Refolding	28
2.9 Co-transformation and Production of ImuA'S and ImuB	29
2.10 Modelling the Three-Dimensional Structure of ImuA'L	29
3 Results.....	31
3.1 Cloning <i>imuA'</i> into Protein Production Plasmids	32
3.1.1 Sequence Acquisition and Naming	32
3.1.2 Sequence Analysis for GC Content and Rare Codons	33
3.1.3 PCR Amplification of <i>imuA'S</i> and <i>imuA'L</i>	34
3.1.4 Colony Screening by Restriction Enzyme Double Digestion	36
3.1.5 Sequencing of Positive Recombinant Plasmids	38
3.2 Properties of ImuA'L and ImuA'S	40
3.3 Production of Recombinant ImuA' in <i>E. coli</i>	40
3.3.1 Production of GST-ImuA'S and GST-ImuA'L.....	42
3.3.2 Production of ImuA' from pGEX-6P-2 and pCOLD I Vectors.....	44
3.3.3 Optimization of Lysis Buffer	46
3.3.4 Protein Production in <i>E. coli</i> BL21-CodonPlus Cells.....	47
3.4 Solubilization and Refolding of GST-ImuA'S from Inclusion Bodies	48
3.4.1 On-column Refolding	49
3.5 Co-production of His₆-ImuA'L and GST-ImuB.....	50
3.6 Modelling the Three-Dimensional Structure of ImuA'L.....	51
3.6.1 Model Evaluation	57
4 Discussion	59
4.1 General Statement of Research Outcome	59
4.2 Analysis of the <i>Mtb imuA'</i>	59
4.3 Two-Step Amplification of <i>imuA'</i>	60
4.4 Production of ImuA' for Structural Studies.....	61

4.5	Insights from the model structure of ImuA'L.....	64
5	Conclusion and Outlook.....	65
	Part II	66
1.	Introduction.....	67
1.1.	Biofuels: Better Alternatives to Fossil Fuels	67
1.2.	Analysis of the Components of Lignocellulose.....	68
1.3.	Glycoside Hydrolase for Cellulose Degradation	69
1.3.1.	Classification of Glycoside Hydrolases.....	70
1.3.2.	Glycoside Family 9 (GH9) Enzymes.....	72
1.4.	Metagenomic Potential for Cellulose Deconstructing Enzymes.....	72
1.5.	The GH9 C1 Cellulase.....	73
2.	Materials and Methods.....	74
2.1.	Protein Production and Purification.....	75
2.1.1.	Starter Culture	75
2.1.2.	Main Culture	75
2.1.3.	Cell Harvesting and Sonication	75
2.2.	Protein Purification	76
2.2.1.	Nickel–nitrilotriacetic Acid (Ni ⁺² –NTA) Affinity Chromatography	76
2.2.2.	Anion Exchange Chromatography	76
2.2.3.	Size-exclusion Chromatography	77
2.2.4.	Protein Concentration	77
2.2.5.	Protein Quantification	77
2.2.6.	Sample Preparation for Crystallization	78
2.3.	Protein Crystallization Screen	78
2.3.1.	Data collection and Evaluation	79
2.4.	Modelling the Three-Dimensional Structure of GH9 C1 Cellulase.....	79
3.	Results and Discussion	81
3.1.	Production and Purification of GH9 C1 Cellulase	81
3.1.1.	Anion Exchange Chromatography (AEX).....	83
3.1.2.	Size Exclusion Chromatography (SEC).....	85
3.2.	Protein Crystallization	86

3.2.1. Initial Screening.....	87
3.2.2. Optimization of Crystallization	87
3.2.3. Diffraction Experiments	90
3.3. Homology Modelling of GH9 C1 Cellulase Structure	92
3.3.1. Overall Model Description	93
3.3.2. Metal Ion Binding.....	94
3.3.3. The Active Site Architecture	98
3.3.4. Mechanism of Action	100
4. Conclusion and Outlook	102
References.....	103



List of Abbreviations

Å	Ångström, 1 Å = 0.1 nm
AA	Amino acid
AEX	Anion exchange chromatography
Amp	Ampicillin
APS	Ammonium persulphate
BLAST	Basic local alignment search tool
bp	Base pair
CAZY	Carbohydrate active enzyme
CbhA	Cellobiohydrolase chain A
CCD	Charge-coupled device
CCP4	Collaborative computational project 4
C-terminus	Carboxyl terminus
CV	Column volume
DMSO	Dimethyl sulfoxide
DNA	Deoxyribonucleic acid
dNTP	Deoxyribonucleotide triphosphate
DOTS	Directly Observed Therapy - Short Course
<i>E. coli</i>	<i>Escherichia coli</i>
EDTA	Ethylene diamine tetraacetic acid
FPLC	Fast protein liquid chromatography
GH	Glycoside hydrolases
GH9 C1	Glycoside hydrolase family 9 C1
GS	Glutathione sepharose
GST	Glutathione S-transferase
HIV	Human immunodeficiency virus
Ig-like	Immunoglobulin-like
IMBM	Institute for microbial biotechnology and metagenomics

ImuA'	Induced mutagenesis A'
IPTG	Isopropyl β -D-1-thiogalactopyranoside
JCSG	Joint center for structural genomics
kb	Kilo base
kDa	Kilo Dalton
LB	Lysogeny broth
LOMETS	Local meta-threading-server
mAu	Milli-absorbance unit
MDR	Multi-drug resistant
<i>Mtb</i>	<i>Mycobacterium tuberculosis</i>
MW	Molecular weight
MWCO	Molecular weight cut off
NCBI	National center for biotechnology information
Ni ²⁺ -NTA	Ni ²⁺ -nitrilotriacetic acid
NMR	Nuclear magnetic resonance spectroscopy
NTA	Nitrilotriacetic acid
N-terminus	Amino terminus
OD ₆₀₀	Optical density measured at a wavelength of 600 nm
PBS	Phosphate buffered saline
PCR	Polymerase chain reaction
PDB	Protein data bank
pI	Isoelectric point
PMSF	Phenylmethanesulfonyl fluoride
Pol	Polymerase
RecA	Recombinase A
RNA	Ribonucleic acid
SDS PAGE	Sodium dodecyl sulphate polyacrylamide gel electrophoresis
SEC	Size exclusion chromatography
SOS	Save Our Souls

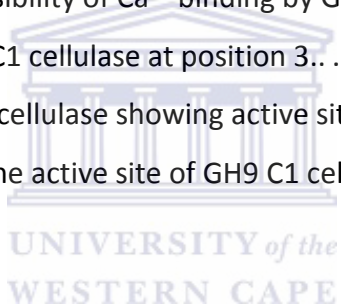
TAE	Tris-acetate-EDTA
TB	Tuberculosis
TBE	Tris-borate-EDTA
TBS	Tris buffered saline
TEMED	N,N,N',N'-Tetramethylethylenediamine
Tris	Tris(hydroxymethyl)aminomethane
UV	Ultraviolet
UWC	University of the Western Cape
WHO	World Health Organization
XDR	Extensively drug resistant
Yfp	Y-family polymerase



List of Figures

Figure 1: DnaE2-dependent mutagenesis in Mtb.....	32
Figure 2: Alignment of <i>imuA'S</i> and <i>imuA'L</i>	33
Figure 3: The DNA sequence for Mtb <i>imuA'L</i> highlighting rare codons.....	35
Figure 4: Agarose gels depicting the results of 'two-step' PCR	35
Figure 5: Screening colonies for positive <i>imuA'</i> clones by restriction digest	37
Figure 6: Nucleotide sequence alignment of pETM-30_ <i>imuA'L</i>	39
Figure 7: GST-ImuA'L produced in <i>E. coli</i> BL21 (DE3) cells.....	42
Figure 8: Production of GST-ImuA' from pETM-30_ <i>imuA'S</i> and pETM-30_ <i>imuA'L</i>	43
Figure 9: GST-ImuA'S produced from pETM-30_ <i>imuA'S</i> construct in <i>E. coli</i> at 15°C	44
Figure 10: Production of ImuA'S as a recombinant GST-fusion protein from pGEX-6P-2_ <i>imuA'S</i>	45
Figure 11: ImuA'L produced from pCOLD I_ <i>imuA'L</i> construct	46
Figure 12: Solubilization of ImuA'L with varying pH of lysis buffer.....	47
Figure 13: Production of GST-ImuA'S in BL21-CodonPlus cells.	48
Figure 14: On-column refolding of GST-ImuA'S from inclusion bodies.....	49
Figure 15: Co-production of His ₆ -ImuA'L and GST-ImuB.....	50
Figure 16: Protein sequences with significant sequence identity to ImuA'L.	52
Figure 17: Alignment of <i>M. smegmatis</i> RecA (PDB code: 1UBC) with ImuA'L	53
Figure 18: Aligned secondary structure of ImuA'L and structure 1UBC from <i>M. smegmatis</i>	54
Figure 19: Partial model for ImuA'L.....	55
Figure 20: Three potential domains of ImuA'L.	55
Figure 21: Modelling the N-terminal domain of ImuA'L.	56
Figure 22: Superposition of ImuA'L model (pink) with RecA/DNA complex	57
Figure 23: Model Quality Estimate for ImuA'L.	58
Figure 24: Secondary structure of ImuA'L predicted using PSIPRED (McGuffin <i>et al.</i> , 1999).....	58
Figure 25: Structural depiction of the composition of lignocellulose	69
Figure 26: pET21a_gh9_c1 is a pET21a based expression vector	74
Figure 27: Ni ⁺² -NTA affinity purification of GH9 C1 cellulase.....	82

Figure 28: GH9 C1 cellulase AEX chromatogram.....	84
Figure 29: SEC purification of the fractions collected from AEC peaks A and B in figure 4..	86
Figure 30: Selected crystallization hits from the PACT suite.....	89
Figure 31: Crystal hits and optimized crystal with the Hampton Crystal Screen reagent.....	89
Figure 32: Results of macro, micro and streak seeding.....	89
Figure 33: Crystal observed in a plate 5 months after setup.	90
Figure 34: Diffraction spots for GH9 C1 cellulase crystals to 4.1 Å.	91
Figure 35: Model structure of GH9 C1 cellulase generated using SWISS-MODEL..	92
Figure 36: The two modules of GH9 C1 cellulase.....	93
Figure 37: Superimposed structures of GH9 cellulases showing metal ion binding sites.....	95
Figure 38: Checking for Zn ²⁺ binding in GH9 C1 cellulase.....	96
Figure 39: Investigation of the possibility of Ca ²⁺ binding by GH9 C1 cellulase.	98
Figure 40: Binding of Ca ²⁺ by GH9 C1 cellulase at position 3..	99
Figure 41: Surface view of GH9 C1 cellulase showing active site cleft.....	99
Figure 42: Conserved residues at the active site of GH9 C1 cellulase.....	100



List of Tables

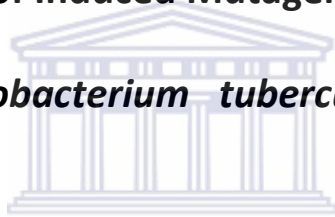
Table 1: Reagents and Suppliers.....	13
Table 2: Stock solutions and their compositions.....	14
Table 3: Names, features and suppliers of plasmids used in this study.....	15
Table 4: Bacterial strains used in this study	15
Table 5: Primers for Two-step Amplification of <i>imuA'</i>	18
Table 6: Recombinant Plasmids and Restriction Enzymes Used in this Study	22
Table 7: Molecular properties determined for <i>ImuA'L</i> and <i>ImuA'S</i>	40
Table 8: Arrangement of Cellulases into Families and Clans.....	71



Part I

Structural Analysis of Induced Mutagenesis A' Protein from

Mycobacterium tuberculosis



UNIVERSITY *of the*
WESTERN CAPE

1. Introduction

1.1 History and Epidemiology of Tuberculosis

Tuberculosis (TB) is a communicable disease ranked with the human immunodeficiency virus/acquired immunodeficiency syndrome (HIV/AIDS) and malaria as amongst the most lethal infectious diseases (Dye *et al.*, 1999). Tuberculosis claims over 2 million lives annually, primarily in developing countries. The disease has infected the human population throughout recorded human history. Modern evidence traces its existence back to the Pleistocene period, around 17 000 years ago and has been documented for the pre-dynastic and early dynastic periods in Egyptian civilization (Bedeir, 2004; Donoghue *et al.*, 2004). Evidence of TB has been confirmed in Egyptian mummies and human remains in India and China dating back to 3000 BC (Barnes, 2000; Daniel, 2006; El-Najjar *et al.*, 1996). Tuberculosis was described as “consumptive disease” or “consumption” (for consuming a patient’s body), “white plague” (causes pale white skin) and “phthisis pulmonary” (for wasting away a patient’s body). In 1865, TB was shown to be contagious by transmitting it to animals. About two decades later, Robert Koch isolated the bacillus *Mycobacterium tuberculosis (Mtb)* as the etiological agent of TB (Rosenthal & Fisher, 2013).

Mtb is a non-motile, acid-fast, rod-shaped bacillus with a Gram positive mycolic acid cell wall. In humans the bacterium typically attacks the lungs causing pulmonary TB. It can, however, also infect other body parts, known as extra-pulmonary TB (Bordbar *et al.*, 2010; Deb *et al.*, 2009). In its host, the bacterium may either be cleared by the host immune system, remain in a latent state, or multiply into active disease (Primm *et al.*, 2000). An estimated one-third of the world’s

population harbours latent *Mtb* and approximately 8-10 million new cases of active TB are recorded annually (McDonough *et al.*, 1993; WHO, 2003). Co-infection with HIV/AIDS, the emergence of multi-drug resistant (MDR) and extensive drug resistant (XDR) strains of *Mtb* contribute to the global difficulties in the control of TB (Barry & Blanchard, 2010; Sacchetti *et al.*, 2008).

1.2 Tuberculosis Treatment

Drug susceptible TB is treatable with a combination of four first line drugs: isoniazid, rifampicin, ethambutol and pyrazinamide (Sharma & Mohan, 2013). This treatment achieved a success rate of over 85 % until the emergence of MDR-TB strains. Multi-drug resistant TB refers to resistance towards isoniazid and rifampicin, the most effective of the first line anti-TB drugs (WHO, 2012). The primary causes of MDR-TB include poor patient management, non-compliance to prescribed regimen, poor national awareness programs or a combination of the three (Davies, 2001). Multi-drug resistant TB is, however, still treatable with a combination of first and second line drugs. Second line TB drugs include cycloserine, fluoroquinolones (ciprofloxacin and ofloxacin) and aminoglycoside injectables (kanamycin, amikacin or capreomycin). The main disadvantage of second line drugs is that they have more side effects than their first line counterparts (Kapoor *et al.*, 2013) and are difficult to administer. Treatment of MDR-TB extends for over 18 months with a daily dose of various medications (Iseman, 1993). This extended treatment period often results in patient non-compliance to treatment regimens, potentially the cause for extensive drug resistant TB. XDR-TB is defined as resistance to both isoniazid and rifampicin plus one fluoroquinolone and any of the second line injectables (Ehrt & Schnappinger, 2009).

1.3 Management of Tuberculosis

In 1995, the World Health Organization (WHO) introduced the Directly Observed Therapy - Short Course (DOTS). DOTS is a patient-centred health care management system with five key pillars: i) the detection of smear-positive pulmonary TB by sputum microscopy; ii) treatment with short-course chemotherapy during which the patient is directly-observed by a healthcare worker or family member in taking their TB treatment; iii) continuous drug supply; iv) government commitment to ensure TB control activities; and v) documentation of patient treatment outcomes (Obermeyer *et al.*, 2008; Obiri-Danso *et al.*, 2013; Volmink & Garner, 2007). The DOTS strategy recommends an intensive treatment with first-line drugs for a minimum period of six months. In this treatment, all four first line drugs are administered for two months, followed by four months of continuous treatment with isoniazid and rifampicin (Hall *et al.*, 2009; WHO, 2009). The DOTS TB management strategy has been implemented in most TB endemic countries, including China and India. Since its implementation, TB incidence worldwide has declined from 2000 onward (Zumla *et al.*, 2013). However, attributing the decline in TB incidences to DOTS alone is controversial as TB has also declined in areas without DOTS implementation. Further, direct supervision of patients leads to stigmatization and reduced acceptance in areas with high TB-HIV/AIDS co-infection and in MDR-TB affected areas (Out *et al.*, 2014; Volmink & Garner, 2007).

1.4 Immunology of Tuberculosis Infection

Mtb most commonly infects its host through aerosolized droplets, coughed up by actively shedding carriers of the bacillus (Rothman *et al.*, 2006). The droplets are inhaled by individuals

in the immediate vicinity and reach the lungs, the site of infection. Here, the bacterium infects individual endothelial cells, but more importantly encounters alveolar macrophages and dendritic cells - the first line of defence within the lungs (Raja, 2004). Alveolar macrophages engulf the bacilli and attempt to destroy them using an array of antimicrobial pathways. Additionally, alveolar macrophages and dendritic cells process the bacteria for presentation of peptides to naïve T-cells. The dendritic cells then transport the bacteria and their antigens to the draining lymph nodes to activate CD4⁺ and CD8⁺ T-cells (Silva & Lowrie, 2000). The T-cells migrate back to the lungs to activate more alveolar macrophages, stop the bacterial replication and remove the invader. While the host is thus initially able to control the spread of *Mtb*, about 5% of bacilli escape clearance and remain latent within the host (Ahmad, 2011). In 5 to 10% of latently infected individuals, bacilli will reactivate into active disease especially if the immune system is compromised (Flynn & Chan, 2001; Kaufmann & McMichael, 2005; Wolf *et al.*, 2007). The inability of the host to clear the infection completely has made *Mtb* such a successful pathogen (Chan *et al.*, 1991; Flynn & Chan, 2001).

1.5 Mycobacterium tuberculosis DNA Damage

The host immune response exposes *Mtb* to a range of DNA-damaging agents such as reactive oxygen and nitrogen intermediates (Stewart *et al.*, 2003). These chemical species can react and corrupt *Mtb* DNA. *Mycobacterium tuberculosis* survival is thus critically dependent on its ability to repair affected DNA (Mizrahi & Andersen, 1998). DNA damage leads to up-regulation of genes that encode DNA polymerases involved in DNA repair. One such polymerase, DnaE2, is a member of the C-family of DNA polymerases (Boshoff *et al.*, 2003). DnaE2-dependent DNA repair is linked

to a high rate of induced mutagenesis (Koorits *et al.*, 2007) mediated in bacteria such as *E. coli* by Y-family polymerases, members of which are also known for *Mtb* (Jarosz *et al.*, 2007; Kana *et al.*, 2010).

1.6 DNA Replication

1.6.1 Eukaryotic and Prokaryotic DNA Polymerases

To accurately copy both strands of DNA is the responsibility of DNA polymerases and accessory proteins, together constituting so-called replicases. Universally, replicases have three components: a polymerase (Pol III in prokaryotes, Pols δ and ϵ in eukaryotes), a sliding clamp processivity factor and a clamp loader. Without accessory proteins, replicative polymerases are largely indistinguishable from other cellular polymerases. With their accessory proteins, they become highly specialized (Fay *et al.*, 1981; Kornberg & Baker, 1992).

DNA replication commences with the ATP-dependent formation of an initiation complex by a DNA polymerase holoenzyme (Wickner & Kornberg, 1973). The clamp loader guides the polymerase to a sliding clamp where it proceeds to add bases to the DNA strand (Downey & McHenry, 2010).

Though similar, eukaryotic and prokaryotic replicative polymerases do differ. Firstly, B- and C-family DNA polymerases respectively replicate eukaryotic and prokaryotic DNA (Braithwaite & Ito, 1993). Secondly, eukaryotic clamp loaders rarely interact with polymerases whereas prokaryotic counterparts tightly interact with Pol III and helicases (Kim & McHenry, 1996). Low GC, Gram positive bacteria are intermediate with clamp loaders weakly interacting with Pol III but strongly with helicases (Rannou *et al.*, 2013). Thirdly, eukaryotic DNA replication involves

three replicative polymerases: the leading and lagging strand polymerases Pol ϵ and Pol δ , and Pol α , which adds dNTPs to RNA primers at the start of Okazaki fragments on the lagging strand before handing over to Pol δ (Nick McElhinny *et al.*, 2008). Gram negative prokaryotes, by contrast, have a single polymerase, the Pol III holoenzyme, whereas Gram positive bacteria have two Pol III enzymes. Some bacteria that lack the error-prone Pol V encode a third Pol III, which can induce mutagenesis (Afonso *et al.*, 2013; Bruck *et al.*, 2005; Gao & McHenry, 2001; Haroniti *et al.*, 2004).

1.6.2 Prokaryotic Pol III Enzymes

Low GC Gram positive bacteria have two Pol III enzymes, the homologues Pol C and DnaE differing by some domain rearrangements (Koonin, & Bork, 1996; Rannou *et al.*, 2013). Pol C has a Mg²⁺-dependent proof reading ability absent in DnaE. DnaE is more closely related to the lone *E. coli* Pol III holoenzyme (Afonso *et al.*, 2013) but functionally mimics eukaryotic Pol α (Sanders *et al.*, 2010) in using RNA primers to extend the lagging strand before handing over to Pol C (McHenry, 2011a; Sanders *et al.*, 2010). A third Pol III enzyme, DnaE2 or ImuC, is found in diverse bacterial phyla and is responsible for translesion synthesis. This polymerase is produced alongside ImuA and ImuB in response to DNA damage via the SOS response mechanism.

1.6.3 The SOS Response Mechanism

The “Save Our Souls” (SOS) response mechanism is a bacterial DNA repair system first described in *E. coli* in 1974 (Erill *et al.*, 2007; Patel *et al.*, 2010; Witkin, 1967). It up-regulates production of Y-family polymerases that synergize with bacterial repair proteins to repair damaged DNA. The SOS response has been implicated in bacterial integrase regulation and stress induced

mutagenesis as well as *Mtb* virulence and antibiotic resistance (Sanchez-Alberola *et al.*, 2012; Žgur-Bertok, 2013).

SOS response genes are regulated by the 27 kDa, dimeric transcriptional repressor LexA. LexA binds a palindromic “SOS box”, which mostly overlaps with RNA polymerase binding sites to prevent SOS regulon transcription and down-regulate SOS genes. At least 50 genes of this regulon have been identified (Galhardo *et al.*, 2007; Kim *et al.*, 1997; McKenzie *et al.*, 2000; Qiu & Goodman, 1997).

RecA, another critical protein, binds damage-induced, single-stranded DNA (ssDNA) to form a ssDNA/RecA nucleofilament denoted RecA* or “activated RecA”. This induces LexA auto-cleavage to release it from the SOS box (Lavery & Kowalczykowski, 1992). RNA polymerase then binds the free promoter and transcribes the SOS regulon. Products of this regulon repair damaged DNA, displacing and inactivating RecA*. LexA, itself SOS-regulated, accumulates to once more repress the regulon (Butala *et al.*, 2008).

Though highly conserved in bacteria, the SOS response differs with respect to the LexA recognition sequence and the set of genes under LexA control (da Rocha *et al.*, 2008). The *E. coli* SOS box CTGTN₈ACAG is found in most β and γ proteobacteria. Gram positive bacteria, non-sulphur bacteria and cyanobacteria by contrast share a GAACN₄GTTY SOS box. Other SOS boxes include GTTCN₇GTTTC in α proteobacteria and TTAN₆TACTA in Xanthomonadales (Erill *et al.*, 2006).

Bacteria such as *Geobacter sulfurreducens* have two *lexA* genes whose gene products bind the same SOS box (Jara *et al.*, 2003), whereas the two LexA proteins of *Pseudomonas* and *Xanthomonas* recognize unrelated SOS boxes. LexA1 of *Pseudomonas putida* in turn binds an

E. coli-like SOS box while its LexA2 shares an SOS box with LexA2 of *Xanthomonas* (Abella *et al.*, 2007). LexA2 is frequently co-transcribed with the common bacterial genes *imuA*, *imuB* and *dnaE2* implicated in DNA damage induced mutagenesis (Aravind *et al.*, 2007; Galhardo *et al.*, 2007; Sanchez-Alberola *et al.*, 2012). This gene cassette has undergone various reorganizations with three genes, two genes and a complete split but invariably remains LexA regulated (Abella *et al.*, 2007).

SOS regulon genes are expressed neither at the same level nor at the same time. Instead, expression levels and initiation time depends on the type of lesion and its severity (Žgur-Bertok, 2013). In case of severe DNA damage, the SOS gene product SfiA (or SulA) is produced (Burhans *et al.*, 2003; Campoy *et al.*, 2005), which participates in cell division arrest to allow for repair mechanisms to occur (Crowley & Courcelle, 2002; Janion *et al.*, 2002).

1.6.4 Y-family Polymerases

Specialized Y-family polymerases are DNA polymerases induced by the bacterial SOS response (Walsh *et al.*, 2011; Yang, 2003). Y-family polymerases are found in all domains of life and are of critical importance as they replicate DNA lesions in a damaged DNA tolerance process called translesion synthesis (Waters *et al.*, 2009), something replicative polymerases are unable to do (Ippoliti, 2012). Y-family polymerases have low fidelity and induce mutations when replicating undamaged DNA (Rattray & Strathern, 2003).

1.6.5 Y-family Polymerases Versus Replicative Polymerases

Y-family polymerases share structural features distinct from replicative polymerases. All share palm, finger, thumb and little finger domains. However, the little finger and thumb domains are

smaller in Y-family polymerases allowing them to accommodate bulky DNA lesions in their active site (Chandani *et al.*, 2010; Ling *et al.*, 2001; Washington *et al.*, 2010). Y-family polymerases further lack the intrinsic 3'-5' exonuclease proofreading and the 'O'-helix required for fidelity in replicative polymerases to accommodate damaged DNA (Patel *et al.*, 2001; Waters *et al.*, 2009).

1.6.6 *E. coli* Y-family Polymerases and their Orthologues in Other Bacteria

The two Y-family polymerase in *E. coli* with translesion synthesis activity, DNA Pol IV (DinB) and DNA Pol V (UmuD'2C), are regulated by UmuD (Beuning *et al.*, 2006) and all three genes *dinB*, *umuC* and *umuD* are co-regulated (Opperman *et al.*, 1996). Within 20 to 40 min after SOS induction, *umuD* primarily produces UmuD₂, a UmuD dimer (Smith & Walker, 1998). UmuD₂ interacts with RecA and ssDNA nucleoprotein filaments prompting auto-cleavage dependent conversion of UmuD to UmuD' by the removal of 24 N-terminal amino acids and the related conversion of Dimeric UmuD₂ to UmuD₂'. The latter combines with UmuC to form UmuD₂'C (Pol V), a Y-family polymerase with translesion synthesis activity (Hare *et al.*, 2006; Ippoliti *et al.*, 2012). UmuD₂ prevents mutations by DinB and UmuC whereas UmuD₂' facilitates mutagenesis through UmuD₂'C complex (Pol V). Cleavage of UmuD₂ is therefore a switch from a non-mutagenic to a mutagenic state in cells (Ippoliti *et al.*, 2012; Ollivierre *et al.*, 2010).

In *Mtb*, the genes *Rv1537* and *Rv3056* encode two Y-family polymerases, DinB1 (DinX) and DinB2 (DinP), both homologues of *E. coli* DinB. Surprisingly, their expression is independent of the SOS response (Cole *et al.*, 1998; Kana *et al.*, 2010; Rand *et al.*, 2003). DinB1 is produced in pulmonary TB and DinB2 upon exposure to novobiocin (Boshoff *et al.*, 2004). An unrelated protein DnaE2 (ImuC), a C-family polymerase, is produced in response to *Mtb* DNA damage (Ippoliti *et al.*, 2012).

1.6.7 The DnaE Proteins

DNA polymerase III (Pol III) is the main replicative polymerase in bacteria. Pol III is a C-family enzyme complex with ten subunits (Rachman *et al.*, 2006). Its α -subunit, DnaE or Pol C, is the replicative enzyme of the complex (Braithwaite & Ito, 1993; Ito & Braithwaite, 1991). Two or three forms of DnaE are encoded by *E. coli* from the same gene, *dnaE*, whereas *B. subtilis* encodes two forms of DnaE from two different genes, *dnaE* and *pol C* (Bruck *et al.*, 2003; Le Chatelier *et al.*, 2004).

Mycoplasma and Mycobacterium species encode two genes related to *E. coli dnaE* referred to as *dnaE1* or *Rv1547c* and *dnaE2* or *Rv3370c*. DnaE2, however, lacks the conventional 3'-5' exonuclease domain of C-family polymerases. DnaE2 proteins are generally not essential for replication but are required for damage-induced mutagenesis and translesion synthesis. They are typically complemented by either ImuA, ImuB or both (McHenry, 2011b). Note that this DnaE2 nomenclature conflicts with that of cyanobacteria where DnaE1 and DnaE2 denote two parts of a functional Pol C subunit formed after the excision of an intein (Liu & Yang, 2003). It was therefore proposed that DnaE2 proteins involved in damage induced mutagenesis and translesion synthesis be renamed "ImuC" as a logical extension of DnaE2 accessory proteins ImuA and ImuB (Galhardo *et al.*, 2007; Warner *et al.*, 2010).

1.6.8 The DnaE2 (ImuC) Accessory Proteins ImuA and ImuB

Induced mutagenesis protein A (ImuA) of proteobacteria and the partly homologous ImuA' of *Mtb* resemble *E. coli* and *M. smegmatis* LexA and RecA (Campoy *et al.*, 2005; Erill *et al.*, 2005).

Despite a sequence identity of 37%, ImuA' cannot complement a *recA*' mutant strain (Campoy *et al.*, 2005; Warner *et al.*, 2010). The function of ImuA' thus essentially remains unknown.

Induced mutagenesis B protein (ImuB) and its *Deinococcus deserti* homolog ImuY are both involved in translesion synthesis (Dulermo *et al.*, 2009). Although homologous to Y-family polymerases, ImuB lacks critical active site aspartic acids rendering it catalytically inactive (Warner *et al.*, 2010). DnaE2 (ImuC), by comparison, is catalytically active but lacks the β -clamp binding domain of ImuB. ImuB and ImuC thus complement each other to create a functional translesion synthesis polymerase (Warner *et al.*, 2010). The genetic arrangement of the *imuA*, *B* and *C* cluster is quite diverse. *Kineococcus radiotolerans*, for example, has a lone *imuC* gene, *Streptomyces coelicolor* has an *imuB-C* combination, while *Pseudomonas putida* has a complete *imuA-B-C* cassette. In *Mtb*, *imuA'-B* form a cassette but *imuC* is in a different locus. In extreme cases, all three genes are in separate loci with distinct SOS boxes (Erill *et al.*, 2007, 2006). Alternatively, the three genes may be spread over both strands of the chromosome or plasmids in either orientations (Abella *et al.*, 2004, 2007).

1.6.9 The *Mtb imuA'*, *imuB*, *imuC* Mutagenic Cassette

Expression of *Mtb imuC* is up-regulated 10-fold following mitomycin C (MMC) treatment or UV irradiation. The *imuC* gene is preceded by a LexA SOS box and is co-regulated with *recA*, *lexA*, *imuA'* and *imuB*. Deletion mutant strains of *Mtb imuC* ($\Delta imuC$) lose UV-induced mutagenesis and are less virulent in mice (Boshoff *et al.*, 2003; Davis *et al.*, 2002). Conversely, elevated levels of ImuC are not mutagenic without UV exposure. ImuC thus requires partners to function (Boshoff *et al.*, 2003). Furthermore, *Mtb* $\Delta imuA'$, $\Delta imuB$ or $\Delta imuC$ mutant strains or any combination

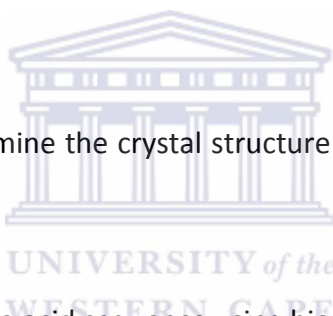
thereof are hypersensitive to MMC and UV. This links ImuA', ImuB and ImuC to a single UV/MMC resistance pathway in *Mtb* (Warner *et al.*, 2010).

ImuB binds both ImuA' and ImuC, but ImuA' and ImuC do not physically interact (Warner *et al.*, 2010). An ImuA'BC complex is therefore assumed to be essential for induced mutagenesis, *Mtb* virulence and DNA repair/survival after exposure to antibiotics or UV. Understanding these interactions at a structural level may allow the rational development of new *Mtb* drugs (Ndwandwe, 2013). As the determination of complex structure may be difficult and protracted, solving the structure of individual members could represent an alternative strategy.

1.7 Aim and Objectives

The aim of this project is to determine the crystal structure of ImuA' from *Mtb*. The objectives are to:

1. Analyse the ImuA' amino acid sequence using bioinformatics techniques.
2. Design primers to amplify *imuA'* by PCR and clone into protein production plasmids.
3. Produce ImuA' as a recombinant fusion protein with His₆- and/or GST-tags.
4. Purify the protein using chromatographic techniques.
5. Crystallize the protein.
6. Determine its crystal structure by X-ray crystallographic methods.



2. Materials and Methods

2.1 General Chemicals and Enzymes

All the reagents used in this study were of analytical grade:

Table 1: Reagents and Suppliers

40% 37.5:1 Acrylamide:bis-acrylamide	Bio-Rad
Agarose	Lonza
Ammonium persulphate (APS)	Merck
Ampicillin	Roche
Bacteriological agar	Merck
Bromophenol blue	Sigma
Coomasie brilliant blue R250	Sigma
Dithiothreitol (DTT)	Roche
DNase	Roche
Ethylene diamine tetra acetic acid (EDTA)	Merck
Ethanol	BDH
Glacial acetic acid	Merck
Glycine	BDH
Hydrochloric acid (HCl)	Merck
Isopropyl-D-1-thiogalactopyranoside (IPTG)	Roche
Kanamycin monophosphate	Roche
Lysozyme	Roche
Potassium chloride	Merck
Phenylmethylsulphonyl fluoride (PMSF)	Roche
Protein molecular weight standard	Fermentas
Reduced glutathione (GSH)	Sigma
Restriction enzymes	Fermentas
Sodium dodecyl sulphate (SDS)	Promega
Sodium chloride	Merck
Sodium hydroxide	Merck
T4 Ligase	Fermentas
TEMED (N, N, N', N'- Tetra methylethylene-diamine)	Promega
Tris (Tris[hydroxymethyl] aminoethane)	BDH
Triton X-100 (Iso-octylphenoxypolyethoxyethanol)	Roche
Tryptone	Merck
Tween 20 (Polyoxyethylene [20] sorbitan)	Merck
Yeast extract	Merck

2.2 Stock Solutions, Buffers and Media

Solutions were autoclaved at 120°C for 20 min unless otherwise indicated.

Table 2: Stock solutions and their compositions

Stocks Solutions	Composition
4 x Separating gel buffer	1.5 M Tris-HCl pH 8.8
4 x Stacking gel buffer	0.5 M Tris-HCl pH 6.8
5 x SDS running buffer	25 mM Tris pH 8.3, 0.1% SDS and 250 mM glycine,
6 x DNA loading dye	0.25% (w/v) bromophenol blue, 0.25% (w/v) xylene cyanol FF and 30% (v/v) glycerol.
10 x Phosphate-buffered saline (PBS)	80 g NaCl, 2 g KCl, 14.4 g Na ₂ HPO ₄ , 2.4 g KH ₂ PO ₄ , was dissolved by stirring in 800 mL deionized water. The pH of the solution was adjusted to 7.4 with HCl and filled to 1 L with deionised water
Ammonium persulphate	10% (w/v) stock solution was prepared in deionised water
Ampicillin	A 100 mg/mL stock solution was prepared in distilled water
Cell lysis buffer	150 mM NaCl, 50 mM Tris-HCl pH 7.5, 5.0 mM DTT, 200 g/mL lysozyme, 35 g/mL DNase, 1.0 mM PMSF, 5 mM EDTA, 2% (w/v) Triton X-100, 1 EDTA-free Complete Protease Inhibitor Cocktail Tablet (Roche)
Coomassie staining solution	0.25% (w/v) Coomassie Brilliant Blue R-250, 30% (v/v) ethanol, 10% (v/v) acetic acid
Destaining solution	40% (v/v) ethanol, 10% (v/v) acetic acid
DTT	A 1 M stock solution was prepared in 0.01 M sodium acetate pH 5.2 and sterilised by filtration
EDTA	A stock solution of 0.5 M, pH 8.0 was prepared in distilled water
IPTG	A 1 M stock solution was prepared in distilled water and sterilised by filtration
Luria bertani media (LB media)	10 g/L tryptone, 5 g/L yeast extract and 5 g/L NaCl. Solution was sterilised by autoclaving
LB agar plates	31 g nutrient agar, 5 g NaCl, 5 g yeast extract, and 10 g tryptone were dissolved in 1 L of water and sterilised by autoclaving. The medium was allowed to cool and 100 g/mL of ampicillin or kanamycin was added. The medium was poured into sterile plates in a laminar flow cabinet
Lysozyme	A 50 mg/mL of lysozyme solution was prepared in deionized water
PreScission 3C protease cleavage buffer	50 mM Tris-HCl pH 7.0, 150 mM NaCl, 5 mM DTT, 1.0 mM EDTA
10X TAE	48.4 g Tris-base, 10.9 g glacial acetic acid, 2.92 g EDTA

2.3 Plasmids and Bacterial Strains

2.3.1 Plasmids and their Properties

Table 3: Names, features and suppliers of plasmids used in this study

Plasmid	Size (bp)	Selection	Tag	Cleavage site	Supplier
pGEX-6P-2	4 985	Amp ^R	N-GST	PreScission protease	GE Healthcare
pETM-30	6 346	Kan ^R	N-His, N-GST C-His	TEV	EMBL
pCOLD I	4 407	Amp ^R	N-His	Factor Xa	Takara Bio Inc.

2.3.2 Bacterial Strains

Table 4: Bacterial strains used in this study

<i>E. coli</i> strain	Function	Supplier
Artic Express	Protein production at low temperature	Agilent Technologies
BL21 (DE3)	Production of proteins	Stratagene
BL21-CodonPlus	Protein production and supplies rare codons	Stratagene
DH5α	Plasmid DNA propagation	Stratagene

2.4 Naming of *imuA'S* and *imuA'L*

The gene sequence for *Mtb imuA'* (Rv3395c) was retrieved from three databases: TubercuList (tuberculist.epfl.ch/quicksearch.php?gene+name=rv3395c); the CMI JCVI (cmr.jcvi.org/cgi-bin/CMR/shared/GenePage.cgi?locus=NTL02MT03387) and the TB databases (genome.tdbb.org/annotation/genome/tbdb/GeneDetails.html?sp=S7000000635256581).

The *imuA'* sequence invariably consisted of 602 bp. However, the CMI JCVI database, provides a second open reading frame for *imuA'* with an earlier start codon adding 270 bp to the 5'-end. In this study, the 602 bp sequence is denoted *imuA'S* (S for short) and the longer gene as *imuA'L*. Both sequences were cloned and expressed in *E. coli*.

2.5 Cloning

2.5.1 Primers for Amplification of *imuA'S* and *imuA'L* by Polymerase Chain Reaction (PCR)

PCR primers (table 5) were designed with the aid of DNAMAN sequence analysis software (Lynnon Corp., Canada) taking into account primer lengths, GC-content, restriction enzyme cut sites, 5' and 3' overhangs for restriction enzyme cut site recognition, self-complimentary ends, melting temperatures and in-frame cloning.

Table 5: Primers for Two-step Amplification of *imuA'*

Name of primer	Primer sequence (5'...3')	T _m (°C)	GC %
pETM-30_ <i>imuA'</i> L Fd (<u>NcoI</u>)	GACT <u>CCATGGG</u> CGT GCCGTTAGTGCGATA	65	60
pETM-30_ <i>imuA'</i> S Fd (<u>NcoI</u>)	CGACT <u>CCATGGG</u> CAT GACTGCGGCCTT	58	63
pETM-30_ <i>imuA'</i> Rv (<u>XhoI</u>)	ACAGT <u>TCTGAG</u> CCGTCCACGCCCGTT	61	73
pCOLD I_ <i>imuA'</i> L Fd (<u>KpnI</u>)	AGACT <u>GGTACC</u> GT GCCGTTAGTGCGATA	56	56
pCOLD I_ <i>imuA'</i> L Rv (<u>EcoRI</u>)	ACAGTGAATTC CCGTCCACGCCCGTT	61	73
pGEX-6P-2_ <i>imuA'</i> S Fd (<u>HindIII</u>)	AAGACT <u>GGATCCA</u> T GACTGCGGCCTT	48	57

Restriction sites underlined, annealing sequence in bold.

2.5.2 PCR Amplification of *imuA'S* and *imuA'L*

2.5.2.1 Standard PCR Protocol

A standard protocol for gene amplification (three-step PCR) was initially used to amplify *imuA'* using the Roche FastStart Taq DNA Polymerase Kit. The reaction consisted of 200 ng template DNA (*Mtb H37Rv* genomic DNA), 1 x GC-RICH solution, 200 µM dNTP, 2 U Faststart Taq DNA polymerase (Roche Applied Science) and 10 pmol of forward and reverse primers, in a final volume of 50 µL. The amplification protocol was as follows: initial denaturation at 94°C for 5 min; 30 cycles consisting of denaturation at 94°C for 60 s, annealing for 30 s at 60°C and extension at 72°C for 60 s, a final extension at 72°C for 7 min. Amplified DNA fragments were analysed on a 1% agarose gel.

2.5.2.2 Gradient PCR

Nine 0.5 mL PCR tubes with reaction mixture as outlined for standard PCR protocol (2.3.2.1) were used with annealing temperature varying between 52 and 68°C at 2°C interval.

2.5.2.3 Two-step PCR

Two-step PCR DNA amplification eliminates the annealing temperature step of standard PCR protocols (three-step PCR). Here, the Phusion High Fidelity DNA Polymerase Kit (Finnzymes), was used to amplify *imuA'* by two-step PCR using *Mtb (H37Rv)* genomic DNA. Primers (Table 4) with 3'-terminal adenine (A) or thymine (T) instead of the generally recommended guanine (G) or cytosine (C) were used to increase primers specificity in a GC-rich gene like *imuA'*. The reactions were set up to a final volume of 50 µL containing: 1 x Phusion GC buffer (Finnzymes), 200 ng genomic DNA, 200 µM dNTP, 0.5 µM each primer, 3% DMSO and 2 U of Phusion Hf DNA polymerase. DNA amplification was performed using the following cycling parameters: denaturation at 98°C for 30 s, followed by 25-35 cycles of denaturation at 98°C for 10 s and extension at 72°C for 30 s/kbp. A final extension at 72°C for 7 min.

2.5.3 Agarose Gel Electrophoresis of DNA

DNA sizes were analysed on 1% agarose gels stained with 1 x GRGreen nucleic acid stain (Biotium Inc.). Agarose gels were prepared by boiling 1 g agarose in 100 mL of 1 x TAE buffer (Table 2). The dissolved agarose solution was allowed to cool to 60°C whereupon 5 µL GRGreen dye was added and mixed by careful swirling to avoid foaming. The stained solution was poured into gel plates, combs inserted, and solution allowed to set for 15 to 30 min. DNA size marker (mostly GeneRuler 1 kb DNA ladder, Thermo Scientific) was loaded alongside samples for size estimation. DNA

samples were stained with 1 x loading dye (Table 2) and loaded into wells on the agarose gel. Generally, gels were run at 100 V for 60 min, unless otherwise stated.

2.5.4 Extraction and Purification of DNA from Agarose Gels

PCR products or digested plasmid DNA in agarose gels were visualized using UV light. Essential DNA bands were manually excised from gel with a scalpel and transferred into clean Eppendorf tubes. DNA was purified using the GeneJet Gel Purification Kit (Thermo Scientific) according to manufacturer's instructions.

2.5.5 Restriction Enzyme Digestion of DNA

Standard restriction enzyme protocols were used to digest DNA using FastDigest restriction enzymes (Thermo Scientific Fermentas). Generally, 1 to 2 μg of plasmid DNA or 0.2 to 0.5 μg of PCR products were digested with 0.1 U/ μL each of the appropriate restriction enzyme in a final volume of 20 μL . The reactions were generally incubated at 37°C for 20 to 40 min. Digestion reactions were analysed on 1% agarose gels and fragments of interest were excised and purified using the GeneJet Gel Purification Kit (Thermo Scientific).

2.5.6 Ligation of DNA Molecules

Ligation reactions were carried out using a vector and an insert previously digested with restriction enzymes to produce matching sticky ends. Ligation reactions consisted of 50 to 100 ng of vector, 5 to 20 ng of insert DNA and 1 U of T4 DNA ligase (Thermos Scientific) in a final volume of 20 μL . Ligation reactions were allowed to proceed at room temperature for 3 h.

2.5.7 Quantification of DNA

DNA concentrations and purities were determined using a NanoDrop ND-1000 spectrophotometer (NanoDrop Technologies, Inc.) by comparing absorbances at 260 nm and 280 nm. A ratio of absorbance at 260 to 280 nm of ~1.8 was accepted as indicative of pure DNA. Appreciably lower ratios indicate contamination by proteins, phenols or other compounds absorbing at 280 nm, while higher ratios indicate contamination with RNA.

2.5.8 Preparation of Chemically Competent *E. coli* Cells

An appropriate strain of *E. coli* was plated on LB plates and incubated overnight at 37°C. A single colony from the plate was used to inoculate 5 mL of LB broth, which in turn was incubated for 16 h at 37°C in a shaker incubator shaking at 170 rpm. The culture was used to inoculate 100 mL LB broth and further incubated until the OD₆₀₀ was between 0.4 and 0.6. The culture was cooled on ice and centrifuged at 11 000 x g for 10 min at 4°C. The supernatant was discarded and the cell pellet resuspended in 35 mL ice-cold transformation buffer: 55 mM MnCl₂, 15 mM CaCl₂, 250 mM KCl and 10 mM PIPES pH 6.7. The cell suspension was incubated on ice for 15 min and centrifuged at 11 000 x g and 4°C. The supernatant was discarded and the cell pellet gently resuspended in 2 mL ice-cold transformation buffer. The cell resuspension was centrifuged at 4 500 x g at 4°C for 5 min using a bench top refrigerator centrifuge. The supernatant was discarded and cell pellet resuspended in 100 mM CaCl₂ solution containing 20% glycerol. Equal volumes (50 µL) of the cells were aliquotted into 1.5 mL Eppendorf tubes, flash cooled in liquid nitrogen and stored at -80°C.

2.5.9 Antibiotic Selection

In experiments with *E. coli* containing ampicillin or kanamycin resistant plasmids, transformed cells were plated on nutrient agar containing 100 µg/mL ampicillin or 25 µg/mL kanamycin. Selection was maintained during growth in liquid culture by adding appropriate antibiotics at the same concentration into the liquid culture.

2.5.10 Bacterial Transformation

Generally, 50 µl of competent *E. coli* cells were transformed with 25 to 50 ng of plasmid DNA. First, cells were thawed on ice and incubated with DNA for at least 30 min. Cells were then heat shocked at 42°C for 45 s and chilled on ice for 2 min. 800 µL of pre-warmed LB was added to the transformed cells and incubated with shaking at 37°C for 1 h. One tenth of the cells were plated onto LB agar plates with appropriate antibiotics and incubated overnight. As a control, untransformed competent cells were plated on agar plates with and without antibiotics and incubated alongside the experimental plates.

2.5.11 Colony Screening by Restriction Digest Analysis

Four to six individual colonies from each LB agar plate containing transformed *E. coli* cells (Section 2.5.10) were inoculated into separate 5 mL volumes of LB media containing appropriate antibiotics. Cells were cultured overnight at 37°C in a shaker incubator at 170 rpm. Cells were collected by centrifugation at 4 500 x g in a bench top centrifuge. Plasmid DNA was extracted using the GeneJET Plasmid Miniprep Kit (Thermo Scientific). The concentrations and purity of the extracted plasmids were determined on a NanoDrop ND-1000 spectrophotometer. The plasmids were subjected to restriction digest (Section 2.5.5) using the same restriction enzymes previously

used to digest the plasmid and insert. The digested DNA was separated by agarose gel electrophoresis and 10 μ L of all positive plasmid clones were sent for sequencing at Inqaba Biotech Inc. (Cape Town, South Africa). The plasmids were sequenced using the Sanger chain sequencing method and results were analysed using SnapGene (GSL Biotech, USA) and ClustalW2 multiple sequence alignment tools (Larkin *et al.*, 2007).

2.5.12 Recombinant Plasmids

Table 6: Recombinant Plasmids and Restriction Enzymes Used in this Study

Plasmid	Restriction Enzymes	Source
pETM-30_ <i>imuA'</i> L	NcoI, XhoI	This study
pETM-30_ <i>imuA'</i> S	NcoI, XhoI	This study
pETM-30_ <i>imuB</i>	NcoI, XhoI	Jeremy Boonzaier (Department of Biotechnology, UWC).
pCOLD I_ <i>imuA'</i> L	KpnI, EcoRI	This study
pGEX-6P-2_ <i>imuA'</i> S	HindIII, EcoRI	This study

2.6 Recombinant Protein Production

Plasmids with correct insert sequence as shown by the sequencing analysis were transformed into *E. coli* BL21 (DE3) cells for protein production.

2.6.1 Small Scale Protein Production Test

Five millilitres of sterile LB media were added to four test tubes each as well as appropriate antibiotics to a final concentration of 100 µg/mL for ampicillin or 25 µg/mL for kanamycin. Each test tube was inoculated with a colony of *E. coli* BL21 (DE3) containing a recombinant plasmid with resistance to the antibiotic in the test tube. The inoculated test tubes were incubated overnight at 37°C with shaking at 170 rpm.

Each overnight culture was used to inoculate 95 mL of LB media containing appropriate antibiotics in autoclaved Erlenmeyer flasks. The absorbance (OD₆₀₀) of each inoculate was measured in a spectrophotometer and samples were diluted to OD₆₀₀ = 0.1 with LB media. The cultures were incubated at 37°C and 170 rpm until the OD₆₀₀ was between 0.4 and 0.6. Two 1 mL samples from each culture flask were transferred to Eppendorf tubes and frozen for SDS-PAGE analysis and glycerol stock preparation. Isopropyl β-D-1-thiogalactopyranoside (IPTG) at 0.5 mM final concentration was added to each culture flask for induction of recombinant gene expression. The culture flasks were transferred to different temperatures for overnight culturing as follows: Flasks with pETM-30_ *imuA'*, pGEX-6P-2_ *imuA'*, and pCOLD I_ *imuA'* were respectively cultured at 30°C, 25°C and 15°C. After 3 and 6 h, 1 mL samples were collected and stored for SDS-PAGE analysis to record the rate of recombinant protein production.

2.6.2 Harvesting Cells

All 1 mL samples collected in section 2.6.1 were centrifuged in 1.5 mL Eppendorf tubes at 4 500 x g for 2 min in a bench top centrifuge. The overnight cultures were harvested in 50 mL centrifuge

tubes by centrifuging at 11 000 x g for 15 min at 4°C. Supernatants from centrifugation were discarded and the cells resuspended in lysis buffer.

2.6.3 Cell Rupture

The cells collected in Eppendorf tubes in section 2.6.2 were ruptured using BugBuster Protein Extraction Reagent (Merck Millipore). Cells were resuspended in BugBuster and incubated at room temperature for 20 min mixing gently on a roller mixer.

Cells from the overnight cultures were resuspended in cooled lysis buffer and ruptured by sonication on ice using six cycles of 30 s sonication separated by 30 s breaks to prevent protein heat denaturation. Ruptured cells were centrifuged at 16 000 x g for 60 min at 4°C to separate soluble cellular content from insoluble content (pellets).

2.6.4 Sodium Dodecyl Sulphate Polyacrylamide Gel Electrophoresis (SDS-PAGE)

Analysis

Soluble and insoluble fractions (pellets) were analysed by SDS-PAGE, which separates proteins by molecular weight in an electric field. Proteins were first denatured by boiling in the presence of a reducing agent and sodium dodecyl sulphate (SDS). Negatively charged SDS binds to the denatured proteins in proportion to their length resulting in an almost uniform charge-to-size ratio and allowing their separation within a porous polymer matrix. Protein samples are first concentrated in a stacking gel before being separated in a resolving gel with distinct pH, ionic strength and pore dimension (Laemmli, 1970). Ten microliters of each supernatant or soluble fraction (Section 2.6.3) was mixed with 2 µL of 8 x SDS-containing sample buffer and incubated at 95°C for 5 min to ensure complete protein denaturation. Pellets on the other hand were first

resuspended in lysis buffer containing 8 M urea to solubilize insoluble proteins and inclusion body proteins before adding 8 x SDS-containing sample buffer and boiling at 95°C for 5 min.

For electrophoresis, the samples were loaded in wells of the stacking portion of gels and a constant current of 40 mA per gel applied for 45 min. The gels were stained for 15 to 20 min in Coomassie Brilliant Blue R-250 solution and excess stain removed by incubating in destaining solution overnight at room temperature.

2.7 Optimization for Soluble Protein Production

2.7.1 Optimization of Culturing Temperature

Five millilitres of an *E. coli* BL21 (DE3) cell culture bearing either the plasmid pETM-30_ *imuA'S* or pETM-30_ *imuA'L* was used to inoculate 500 mL LB medium with 25 µg/mL kanamycin in a 2 L Erlenmeyer flask. The flask was incubated at 37°C and 170 rpm until the OD₆₀₀ was between 0.6 and 0.8. Each culture was divided into two 250 mL portions to test for protein production at 15°C and 25°C, respectively. Both cultured flask were induced for protein production with 0.5 mM IPTG and incubated overnight, one at 15°C and the other at 25°C in a shaker incubator at 170 rpm. Cells were recovered by centrifuging at 11 000 x g for 15 min. The supernatant was discarded, pellets resuspended in lysis buffer and prepared for SDS-PAGE analysis (Section 2.6.4).

2.7.2 Optimization of Gene Expression by IPTG

A 300 mL culture of *E. coli* BL21 (DE3) transformed with pETM-30_ *imuA'S* and cultured to OD₆₀₀ of 0.6 was divided into three equal fractions in 250 mL flasks and protein production induced by adding IPTG to a final concentrations of 0.05 mM, 0.1 mM and 0.25 mM, respectively. Flasks were incubated overnight at 20°C, centrifuged (11 000 x g for 15 min at 4°C), supernatant discarded

and pellets resuspended in lysis buffer. The resuspended pellets were lysed and analysed by SDS-PAGE.

2.7.3 ImuA' Production in Optimized *E. coli* Strains

The plasmid pETM-30_ImuA'S was transformed into two specialized *E. coli* strains: Artic Express (Agilent Technologies) for low-temperature protein production and BL21-CodonPlus (Stratagene) which provides additional tRNAs of normally rare codons. Cells were cultured and induced with 0.5 mM IPTG at OD₆₀₀ of 0.6. After induction, Artic Express cells were incubated at 10°C and BL21-CodonPlus cells at 15°C for 20 h at 170 rpm. Cells were separated from culturing media by centrifugation (11 000 x g for 15 min at 4°C) and resuspended in cell lysis buffer (Table 2). The resuspended cells were ruptured by sonication and analysed on SDS-PAGE gels.

2.7.4 Optimization of Lysis Buffer pH

E. coli BL21 (DE3) cells transformed with pETM-30_ImuA'L recombinant plasmids were cultured and induced with 0.5 mM IPTG at 15°C. The cells were pelleted by centrifugation (11 000 x g for 15 min at 4°C) and equal masses of pellets were resuspended and lysed in separate 15 mL tubes with Tris-HCl buffer at pH 7.0, 7.5, 8.0, 8.5, 9.0 and 9.5. The lysed cells were centrifuged at 16 000 x g for 60 min at 4°C to separate soluble from insoluble fractions. The soluble and insoluble fractions were analysed by SDS PAGE.

2.8 Solubilization of Inclusion Bodies and Protein Refolding

2.8.1 Inclusion Body Preparation

E. coli cells containing pETM-30_ *imuA*'S were cultured, centrifuged and sonicated. Pellets were separated from the supernatant by centrifugation at 16 000 x g for 60 min at 4°C. The pellets were washed twice with wash buffer I (20 mM Tris pH 8 and 50 mM NaCl), centrifuged at 16 000 x g for 1 h at 4°C. The supernatant was discarded and pellets were resuspended and washed twice with wash buffer II (50 mM Tris pH 8 50 mM NaCl, 2% Triton X-100, 1.5 mM β -mercaptoethanol and 1.6 mM urea). The pellets were washed once more with wash buffer I to remove triton X-100.

2.8.2 Denaturation of Inclusion Body Proteins

Washed inclusion bodies were dissolved in denaturing buffer (50 mM Tris pH 8, 50 mM NaCl, 10 mM β -mercaptoethanol, 8 M urea) and stirred for 20 min. The resulting solution was centrifuged at 16 000 x g and 4°C for 1 h. The supernatant with solubilized inclusion body proteins was stored at 4°C for SDS-PAGE analysis.

2.8.3 Protein Refolding by Dialysis

The solubilized inclusion body proteins solution (30 mL) was transferred to a SnakeSkin Pleated Dialysis Tube (Thermo Scientific) with a molecular weight cut off (MWCO) of 3.5 kDa. The dialysis tube was immersed in 2 L of refolding solution (20 mM Tris pH 8, 100 mM NaCl, 6 M Urea and 5 mM DTT) and stirred overnight with a magnetic stirrer at 4°C. The dialysis tube was transferred to a second (and third) refolding buffer with urea reduced to 4 (and 2) M urea and stirred for 10 h (overnight) at 4°C. The solution was clarified by centrifugation at 16 000 x g and 4°C for 1 h. The

supernatant was mixed with 2 mL of glutathione sepharose (GS) resin (GE Healthcare) previously washed with 2 CV of 2 M urea refolding solution. The resin-supernatant mixture was agitated overnight on a roller mixer at 4°C in a 50 mL Falcon tube. The mixture was transferred into an Econo-Pac chromatography column (Bio Rad, South Africa) and unbound proteins eluted. The resin was washed five times by refilling the column with wash buffer and allowing it to drain. All eluted fractions as well as a resin sample were analysed by SDS-PAGE to identify target protein potentially bound to resin or eluted.

2.8.4 On-column Refolding

Another sample of inclusion bodies was prepared and solubilized as described in Sections 2.8.1 and 2. As described, the urea concentration of the solubilized protein was lowered to 4 M by repeated dialysis against 2 L of refolding solution with decreasing urea concentrations. The resulting solution was mixed with 2 mL Ni²⁺-NTA resin (Qiagen), poured into an Econo-Pac chromatography column and washed with 4 M urea refolding buffer. The protein-resin mixture was agitated overnight on a roller mixer at 4°C. The resin was washed successively with five column volumes (CV) of 2 M urea-containing refolding solution and with 2 CV of 10 mM, 25 mM and 60 mM imidazole-containing refolding solution. The resin was further washed with 100 mM and 300 mM imidazole-containing refolding solution to elute proteins from resin. Samples of the wash fraction, the elution fraction and the resin after elution were analysed by SDS-PAGE. The resin was incubated overnight in 10 mL of 300 mM imidazole-containing refolding buffer and washed with 2 CV of 500 mM imidazole-containing refolding buffer to elute protein of interest that was still bound on the resin.

2.9 Co-transformation and Production of ImuA'S and ImuB

Plasmids of pCOLD I_ *imuA'L* (this work) and pETM-30_ *imuB* (provided by Mr Jeremy Boonzaier, Structural Biology Laboratory, University of the Western Cape) were co-transformed into *E. coli* BL21 (DE3) and cultured overnight on an LB plate containing ampicillin and kanamycin. A colony from the plate was used to inoculate 5 mL LB media containing 100 µg/mL ampicillin and 25 µg/mL kanamycin. The culture was incubated at 37°C overnight with shaking at 170 rpm. The culture was then transferred to 45 mL of LB with the same concentrations of antibiotics and incubated at 37°C with shaking at 170 rpm until the OD₆₀₀ was between 0.6 and 0.8. Protein production was induced by the addition of 0.5 mM IPTG and the culture incubated with shaking overnight at 15°C and 170 rpm. The next morning, the cell culture was centrifuged, the supernatant discarded and the pellets resuspended in lysis buffer. The resuspended cells were ruptured by sonication, centrifuged (11 000 x g for 1 h at 4°C) to separate soluble from insoluble fractions and both fractions analysed by SDS-PAGE.

2.10 Modelling the Three-Dimensional Structure of ImuA'L

Homology or comparative modelling is a technique used to infer the structure of a target protein based on the experimentally determined structure of a template by analogy. The technique generally involves four steps: i) template identification, ii) target–template alignment, iii) model building, and iv) model assessment. In identifying a template, a library of experimentally determined protein structures is searched using the sequence of the protein of interest (target) to identify proteins with significant sequence identity to the target. The protein with highest sequence identity and sequence coverage is selected and aligned with the target. Using the

sequence alignment and the coordinates of the template structure, the sequence of the target protein is threaded onto the existing backbone. Insertions and deletions are accommodated where possible. An energy minimization step allows high-energy artifacts from the modelling step to be eliminated. Finally the model quality is assessed by comparing backbone conformation and side-chain packing to chemical standards.

Protein sequences from the UniProtKB/Swiss-Prot database homologous to ImuA'L were identified using NCBI-BLAST (Ye *et al.*, 2006). The sequences of ImuA'L and the protein structure with highest sequence identity to ImuA'L were aligned using ClustalW2. The alignment was uploaded to the Workspace Alignment Mode on the SWISS-Model server (Kiefer *et al.*, 2009), to force the server pipeline to construct the model structure strictly on the alignment by segment matching or coordinate reconstruction. Minor modifications are made to accommodate deletions and insertions. The model was evaluated and plots of Anolea mean force potential, GROMOS empirical force field energy and Verify3D profile evaluation produced to judge the quality of model and template structures.

The graphics program Pymol was used to read protein coordinate files of model and template structures (The PyMOL Molecular Graphics System, Version 1.5.0.4 Schrödinger, LLC.).

3 Results

The working hypothesis at the inception of this project was that ImuB interacts with ImuA' and DnaE2 (ImuC) to form a molecular complex necessary for translesion synthesis in *Mtb*. This molecular complex would access DNA lesions through the interaction of ImuB and the β -clamp protein on the DNA (Figure 1) (Ndwandwe, 2013). The role of ImuA' in this complex, however, remains unknown. This study was aimed at structurally analysing ImuA' to help unravel its function during translesion synthesis. The strategy was therefore to design primers for PCR amplification of *imuA'* from *Mtb* genomic DNA, clone the PCR products into protein production plasmids for recombinant protein production in *E. coli*, purify the protein using chromatographic techniques, crystallize and solve the protein crystal structure using X-ray crystallographic methods.

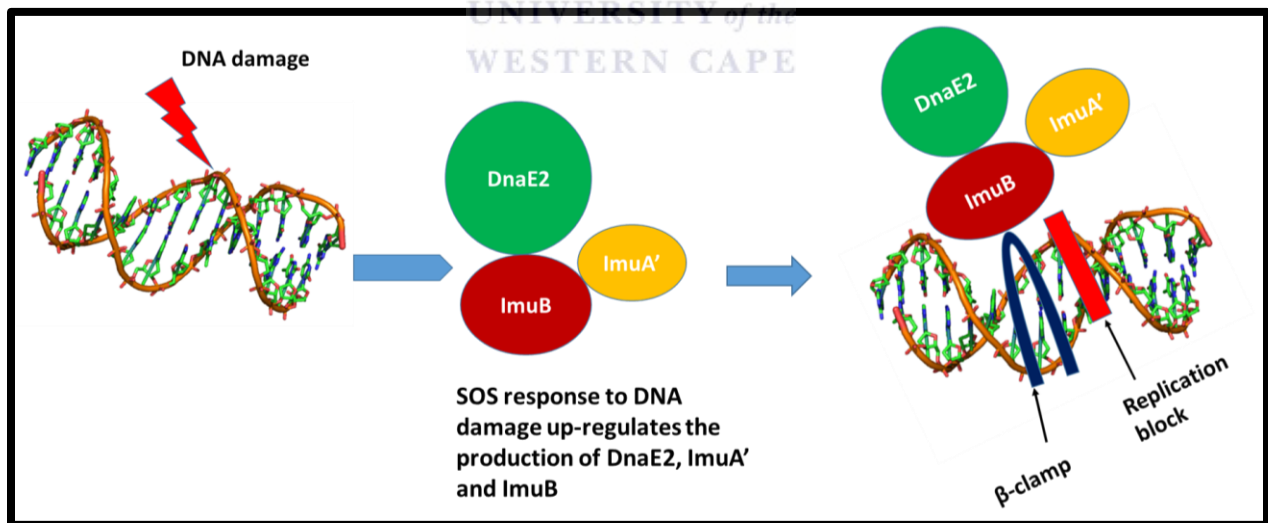


Figure 1: DnaE2-dependent mutagenesis in *Mtb*. According to this model, DNA damage causes normal replication to stall. In turn, expression of *dnaE2*, *imuB* and *imuA'* is up-regulated allowing the resulting proteins to be recruited to the site of the lesion. DnaE2 and ImuA' presumably access the DNA template by interacting with ImuB, which binds the β -clamp protein on the DNA. The complex once formed catalyses translesion synthesis across the lesion.

3.1 Cloning *imuA'* into Protein Production Plasmids

3.1.1 Sequence Acquisition and Naming

The CMI JCVI database (cmr.jcvi.org/cgi-bin/CMR/shared/GenePage.cgi?locus=NTL02MT03387), provided two sequences for *imuA'* differing with respect to the 5'-terminal start. Alternative promoters and start codons thus appear to exist for this gene. In the following, the two sequences are denoted *imuA'S* and *imuA'L*, where S and L indicate "short" and "long" respectively. The *imuA'L* sequence adds 270 bp to the 5'-end of *imuA'S* (Figure 2).



Figure 2: Alignment of *imuA'S* and *imuA'L*. The *imuA'L* sequence is 5'-terminally extended by 270 bp relative to *imuA'S*: bold nucleotides marked by the blue bracket and alternating with dashed lines. Aligned with ClustalW2 (Larkin *et al.*, 2007).

3.1.2 Sequence Analysis for GC Content and Rare Codons

As the *Mtb* genome is characteristically GC-rich, the two sequences for *imuA'* were analysed for their guanine and cytosine (GC) content and for rare codons. GC-rich genes often form inter- and intra-strand secondary structures (hairpins) due to increased hydrogen bonding between guanines and the N-7 rings of neighbouring guanines causing premature termination during PCR and correspondingly smaller bands in electrophoresis (Jesen *et al.*, 2010). High melting temperature overlaps between template and complement strands additionally cause mis-priming and -annealing in GC-rich genes. The respective GC-content of *imuA'S* and *imuA'L* is 70 and 68% (<http://www.endmemo.com/bio/gc.php>).

Organisms use 61 of 64 possible nucleotide codons to code for amino acids, two for stop codons and one for the N-terminal formyl-methionine. Depending on the organism, some codons for a particular amino acids are used more frequently (major codons) than others (rare codons). Fifteen codons classified as rare in *E. coli* were identified in *imuA'L* (<http://nihserver.mbi.ucla.edu/RACC/>) (highlighted in Figure 3).

```

gtg ccg gtt agt gcg ATA gtc gca acg gcc ggt agc tcg aac cca tcg gtg gtg ttg tcg gtg gcg aag agc tct
gcc ggc cgg cag gca ggc ccg cca ccg gtg gcc ggt ggg gcc gtc cct ggg cct aac AGG ccg caa aac agc AGG
gca gcc gcc agt acc gag gtg gtt tta cgc gat tgc aca AGG cag cct ctc atg acc ttg acg gac tcc aag gac
ggg tgt tta ctg act tcg aac ATA ttt tcg aac AGG AGG ctg gtc atg act gcg gcc ttc gcc tcc gac caa cgc
ctt gaa aat ggt gct gag cag ctc gaa tca CTA CGA cgg cag atg gct ttg ctg tcc gag aag gtg tcc ggg ggg
CCC agc cgt tcg ggc gac ctg gtg ccg gcg gga ccg gtg tcg ttg CCC ccg ggg acg gtg gga gtg ctg tcg ggt
gcg cgg tca ctg ctg ctg agc atg gtg gca tcg gtg acg gcg gcc ggg gga aac gcg gcc atc gtt ggc cag ccg
gat atc ggg ttg ctg gcc gcg gtg gag atg ggg gcg gat ctg agc cgg ctc gcg gtg ATA cca gat CCC ggg acc
gat ccg gtt gag gtg gcc gct gtg ctg atc gac ggc atg gat ctg gtg gtg ctc ggt ctg gga ggg cgc cgg gtg acg
cgg gcg cgg gcg cgg gca gtg gtg gcc cgt gcc cgt caa aaa ggc tgc acc ctg ctg gtc acc gac ggc gac tgg
caa ggc gtg tcg acg cgg ctt gcg gcc cgg gtc tgc ggc tat gag atc acc ccg gcc ctc AGG ggc gtg CCC acc
ccg ggg ttg ggg cgg atc agt ggg gtg cgg ctg cag atc aac ggg cgt gga cgg tga

```

Figure 3: The DNA sequence for *Mtb imuA'L* highlighting rare codons with respect to translation in *E. coli*. Rare codons for arginine, leucine, isoleucine, and proline are respectively coloured red, green, blue and orange

3.1.3 PCR Amplification of *imuA'S* and *imuA'L*

Amplification of *imuA'* proved a major challenge in this project. Initially, PCR primers for amplification of *imuA'S* using a standard three-step PCR method (section 2.5.2.1) were designed. These primers had one or two guanines (G) or cytosines (C) at their 3'-ends to ensure tight binding of primer to template. However, agarose gel electrophoresis showed truncated bands smaller than the anticipated size of *imuA'S*. The bands were isolated, sequenced and found to be truncated *imuA'S*. The sequenced gene also showed that the 3' G or C of the primers cause unspecific priming probably because of the high GC content of the gene. The specificity of the

primers was thus modified by ensuring they end in adenosine (A) or thymine (T) at the 3'-end (Table 3). However, standard PCR (section 2.5.2.1) still did not amplify the *imuA'S* gene using these modified primers. Using Phusion High Fidelity DNA polymerase (Finnzymes), designed for amplification of GC-rich genes, with the 3'-GC primers and 3'-AT primers in a standard PCR protocol again yielded truncated gene products for both sets of primers. Also repeating the amplification using annealing temperatures of 50, 56, 60 and 65°C or gradient PCR (section 2.5.2.2) with annealing temperature between 52 and 68°C at 2°C intervals did not yield the expected PCR products.

As mis-priming and mis-annealing can be caused by high primer melting temperatures, overriding the annealing temperature step was considered. Two-step PCR (section 2.5.2.3) in which the annealing temperature step is eliminated was therefore applied yielding full-length *imuA'S* and *imuA'L* products for the 3'-AT primers. The method was repeated to amplify *imuA'S* and *imuA'L* for cloning into pETM-30 and pCOLD I and pGEX-6P-2 vectors (Figure 4).

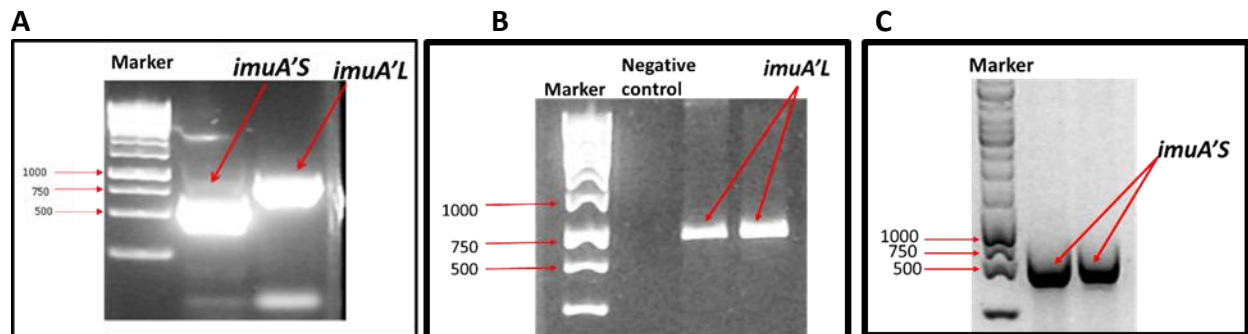


Figure 4: Agarose gels depicting the results of 'two-step' PCR amplification of *imuA'S* and *imuA'L* from *Mtb* genomic DNA. **A)** Amplification products of *imuA'S* and *imuA'L* for cloning into pETM-30: Lanes 2 and 3: *imuA'S* and *imuA'L* PCR products, respectively. **B)** Amplification products of *imuA'L* for cloning into pCOLD I: Lane 2: Negative control, PCR mix without the genomic DNA; Lane 3 and 4: PCR product of *imuA'L*. **C)** Amplification of *imuA'S* for cloning into pGEX-6P-2: Lanes 2 and 3: PCR products of *imuA'S*. Lanes 1 in A, B and C: GeneRuler 1 kb DNA ladder with sizes as indicated to the left of the gel.

3.1.4 Colony Screening by Restriction Enzyme Double Digestion

The PCR products (Figure 4) and the target vectors were digested using appropriate restriction enzymes (section 2.5.5). The plasmids and inserts were ligated using T4 DNA ligase, transformed into competent *E. coli* DH5 α cells and plated on agar plates with appropriate antibiotic. Only cells carrying the plasmids with an antibiotic resistance gene grew on agar plates while control plates with antibiotic and plated with un-transformed competent cells showed no growth. Occasionally, plasmids digested with two restriction enzymes may circularize without an inserted gene. Cells taking up such plasmids will grow on antibiotic plates as false positive colonies. To distinguish false and true positive colonies, plasmids were extracted and digested with appropriate restriction enzymes (Figure 5).



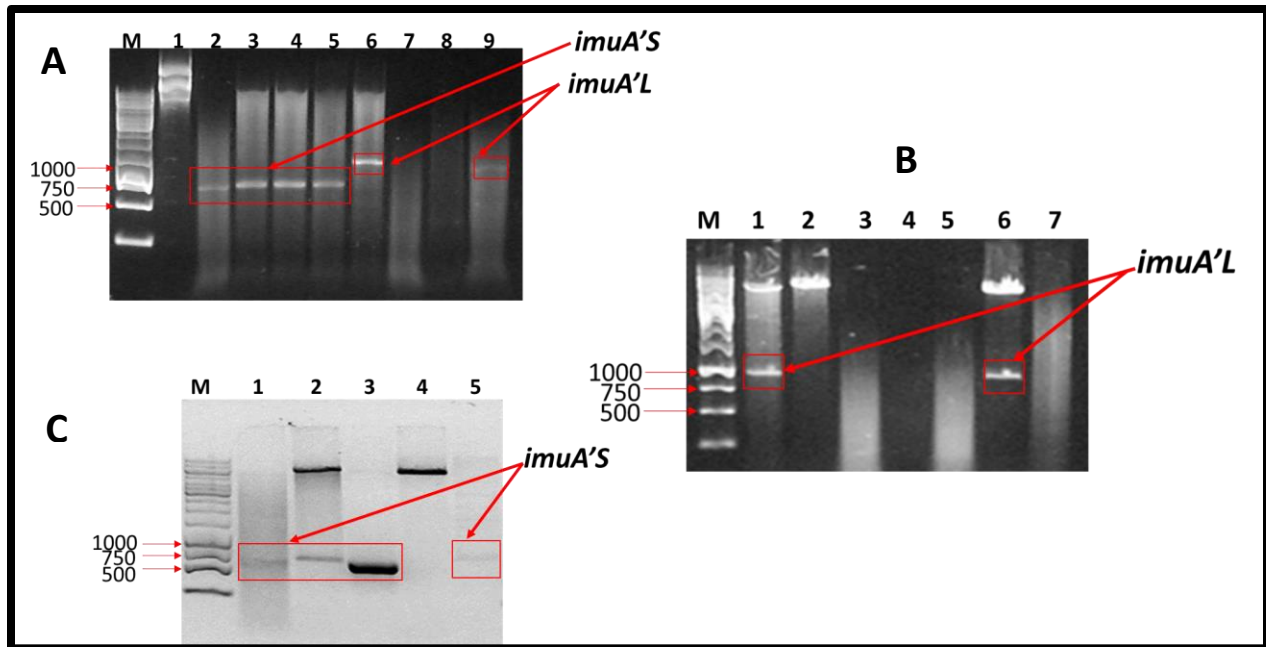
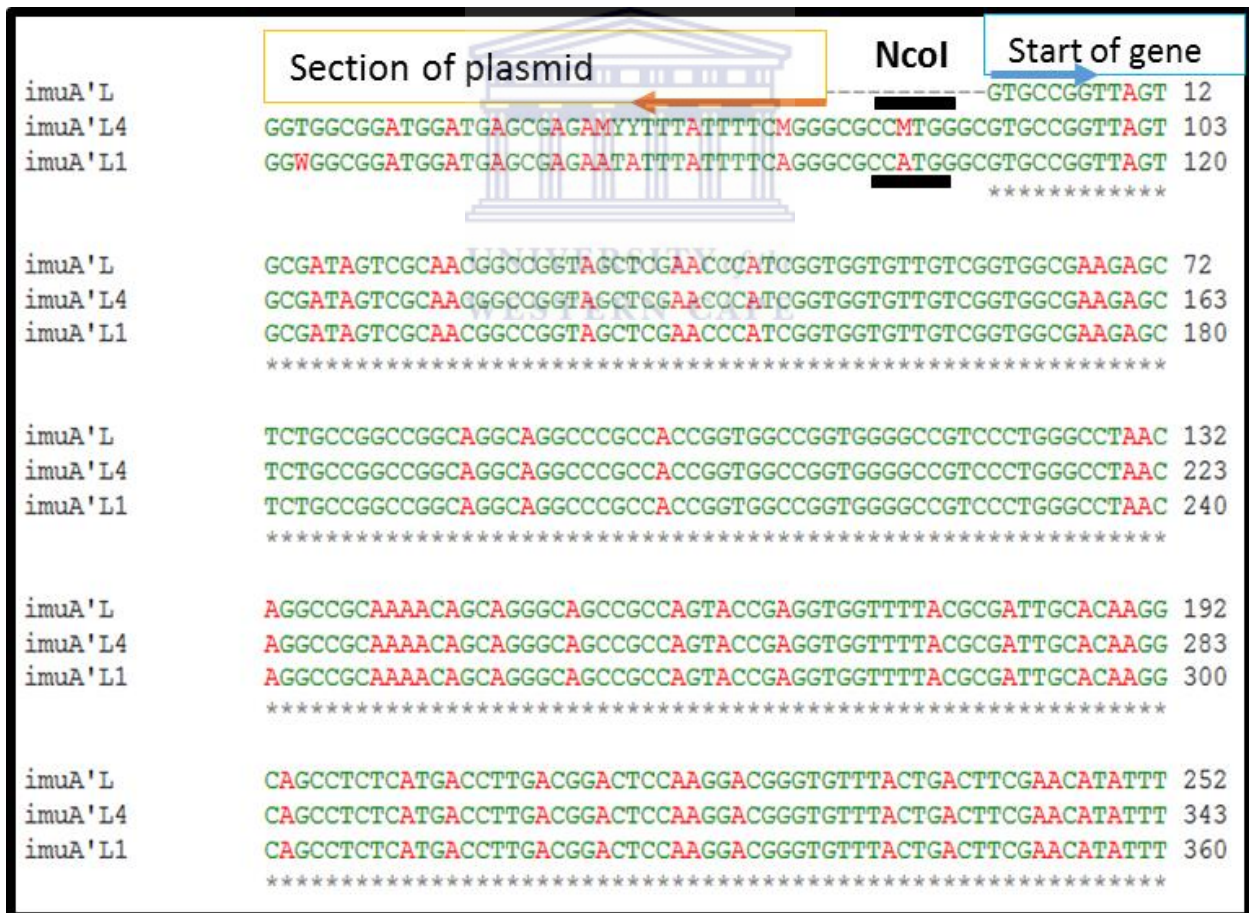


Figure 5: Screening colonies for positive *imuA'* clones by restriction digest of recombinant plasmids. Lanes M: GeneRuler 1 kb DNA ladder with sizes as indicated to the left of the gel. **A):** Restriction digest of recombinant pETM-30_*imuA'* plasmids isolated from transformed clones. Lane 1: Undigested plasmid; Lanes 2-5: XhoI and NcoI digest of pETM-30_*imuA'S* plasmids from four distinct colonies. Lanes 6-9: XhoI and NcoI digest of pETM-30_*imuA'L* from four distinct colonies. The undigested plasmid band in Lane 1 is characteristically smeared presumably due to plasmid DNA supercoiling. The original *imuA'S* insert can clearly be seen after restriction digest in lanes 2 to 5: (red rectangle) implying that all four colonies contain the insert *imuA'S*. The remaining plasmid, however, smeared out possibly due to shearing of the plasmid or high salts concentration in the digestion buffers used. Lanes 6 and 9 contain the *imuA'L* insert (red rectangles) while lanes 7 and 8 are false positive colonies. **B):** Lanes 1-3 and lanes 5-7: EcoRI and KpnI restriction digest of recombinant pCOLD I_*imuA'L* isolated from transformed clones. Lane 4: Empty. Lanes 1 and 6 show true positive clones with the *imuA'L* insert band indicated by red rectangles. Lane 2 shows a false positive clone with just the plasmid band and no insert. Lanes 3, 5 and 7 are smeared out possibly due to shearing of DNA or high concentration of salts in buffer, no band could be identified. **C):** Lanes 1, 2, 4 and 5: EcoRI and XhoI double restriction digest of pGEX-6P-2_*imuA'S* plasmids isolated from transformed clones. Lane 3: *imuA'S* PCR product as size marker. Lanes 1, 2 and 5 showed true positive clones as identified by the insert bands (red rectangles). The remaining plasmids in lanes 1 and 5, however, smeared out and could not be identified on the lane. Lane 4 shows a false positive colony as only the plasmid band was identifiable. Lane 3, the positive control band runs lower than the other insert bands. This could be due to the higher concentration of DNA in the control band. All true positive colonies were Sanger sequenced (Inqaba Biotech Inc.) and confirmed to be *imuA'* clones.

3.1.5 Sequencing of Positive Recombinant Plasmids

True positive recombinant plasmids (Figure 5) were further confirmed by sequencing the recombinant plasmids (Inqaba Biotech). The experimentally determined nucleotide sequences of pETM-30_ *imuA'*L corresponding to isolates seen in Figure 5A, Lanes 6 and 9 are compared to the theoretical *imuA'*L sequence in Figure 6. The experimental *imuA'*L nucleotide sequences aligned to the database sequence within the restriction cut sites as indicated on the Figure 6. All other colonies identified as true positives in Figure 5 were similarly sequenced and aligned. Constructs with mutations were discarded.



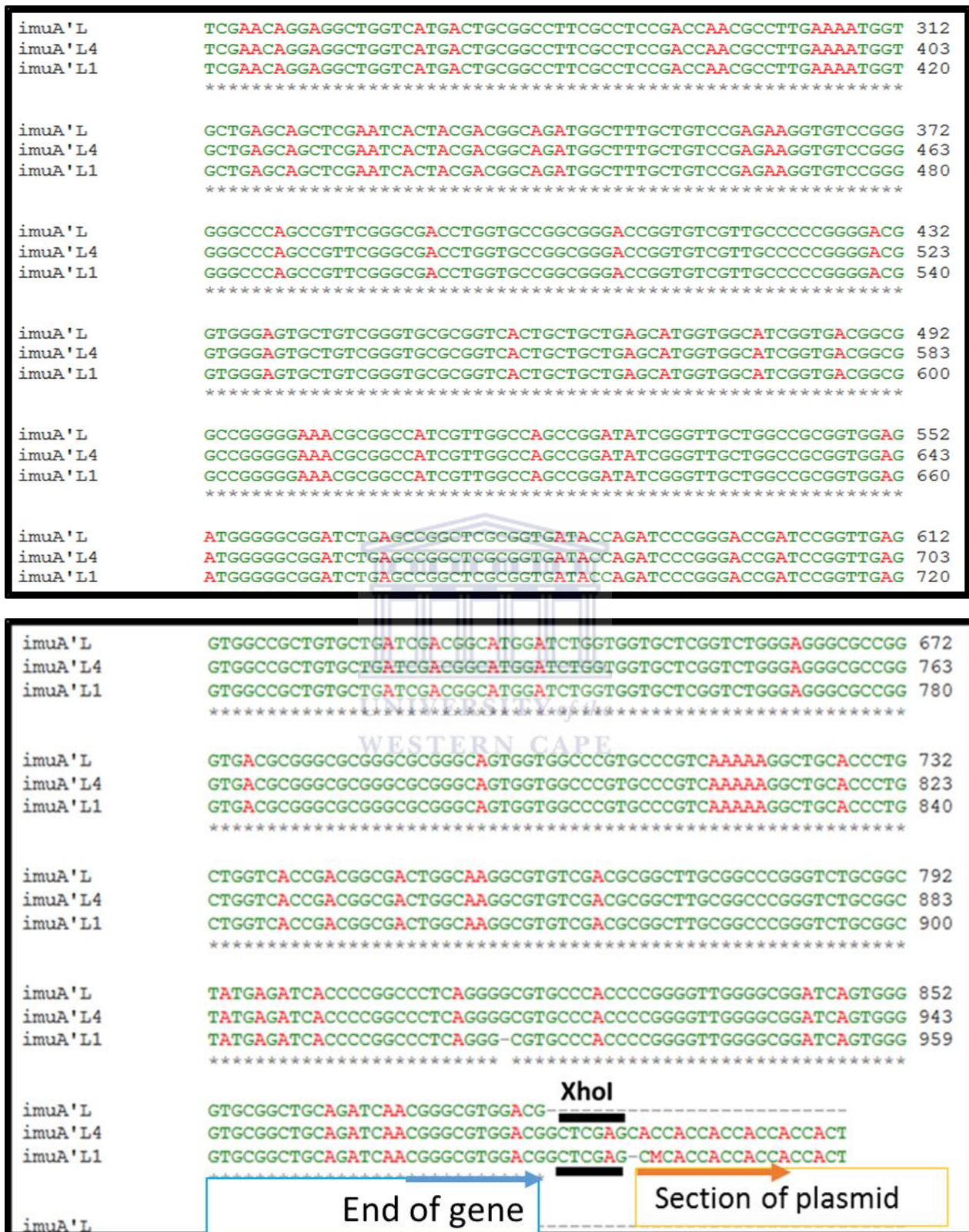


Figure 6: Nucleotide sequence alignment of pETM-30_ *imuA'L* from two true positive colonies (Figure 5B, Lanes 6 (L1) and 9 (L4)) with the database derived *imuA'L* sequence. Asterisks mark nucleotides identical between the experimental and database sequences.

3.2 Properties of ImuA'L and ImuA'S

Physical molecular properties calculated for ImuA'L and ImuA'S are contrasted in Table 7.

Table 7 : Molecular properties determined for ImuA'L and ImuA'S

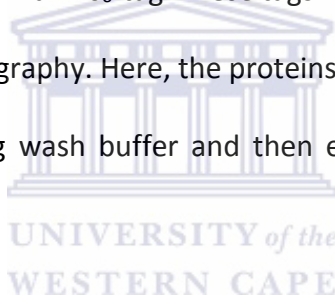
Property	ImuA'L	ImuA'S
Gene length (bp)	885	615
Protein length (AA)	294	204
Molecular weight (kDa)	29.9	20.8
pI	11.2	10.7
Instability index	44.1	34.0
Aliphatic index	99.5	106.6
% Solubility in <i>E. coli</i>	3.6	6.7

3.3 Production of Recombinant ImuA' in *E. coli*

Structural studies of proteins by X-ray crystallography generally require the production and purification of milligram amounts of the protein. In this study, four recombinant plasmids were constructed for overexpression of *imuA'* in *E. coli*. The plasmids: pETM-30, pCOLD I and pGEX-6P-2 were used to generate four recombinant constructs: pETM-30_*imuA'S*, pETM-30_*imuA'L*, pCOLD I_*imuA'L* and pGEX-6P-2_*imuA'S*. The pET plasmid has a T7 promoter

upstream of a multiple cloning site, into which the gene of interest is inserted. The promoter ensures high levels of gene expression in the presence T7 RNA polymerase inducible from the chromosomal DNA of *E. coli* BL21 (DE3) strains by inducers like IPGT. For its part, the pGex vector, transcribed by prokaryotic RNA polymerases contains a *tac* promoter under the control of a *lacI* operator inducible by IPTG. Additionally, each plasmid has an affinity tag, which allows the protein to be produced as a fusion protein with the tag at its either N- and/or C-terminus. The affinity tag is useful for downstream protein purification by affinity chromatograph.

The pETM-30 vector has a *gst* gene for the production of an N-terminally GST-tagged protein. It also encodes both an N- and C-terminal His₆-tag. These tags make it possible to purify the protein in two rounds of affinity chromatography. Here, the proteins were immobilised on Ni-NTA resin, washed with imidazole containing wash buffer and then eluted with high concentrations of imidazole (250 mM imidazole).



The vector pETM-30 does not encode a protease cleavage site for the C-terminal His₆-tag. The tag can potentially impair downstream processes such as crystallization. Hence *imuA'S* was also cloned into pGEX-6P-2 vector, which adds an N-terminal GST-tag separated from the protein by a PreScission Protease cleavage site.

Some proteins tend to be packaged in inclusion bodies when overexpressed in *E. coli*. The pCOLD I vector, a cold shock expression vector, was used to clone *imuA'L* for protein production at low temperatures. The low temperature slows the rate of protein production, potentially helping the protein to fold properly. The protein of interest is N-terminally His₆-tagged when produced from the pCOLD I vector.

The plasmid DNA constructs were sequenced and those with no mutations in the DNA were transformed into *E. coli* BL21 (DE3) cells for small scale protein production (Section 2.6.1).

3.3.1 Production of GST-ImuA'S and GST-ImuA'L

E. coli BL21 (De3) cells transformed with pETM-30_ImuA'S and pETM-30_ImuA'L were separately cultured in LB medium and induced with 0.5 mM IPTG for overnight protein production. The cultures were analysed the following day by SDS-PAGE (Figure 8). It was observed that both proteins were produced but the proteins were in the insoluble fraction.

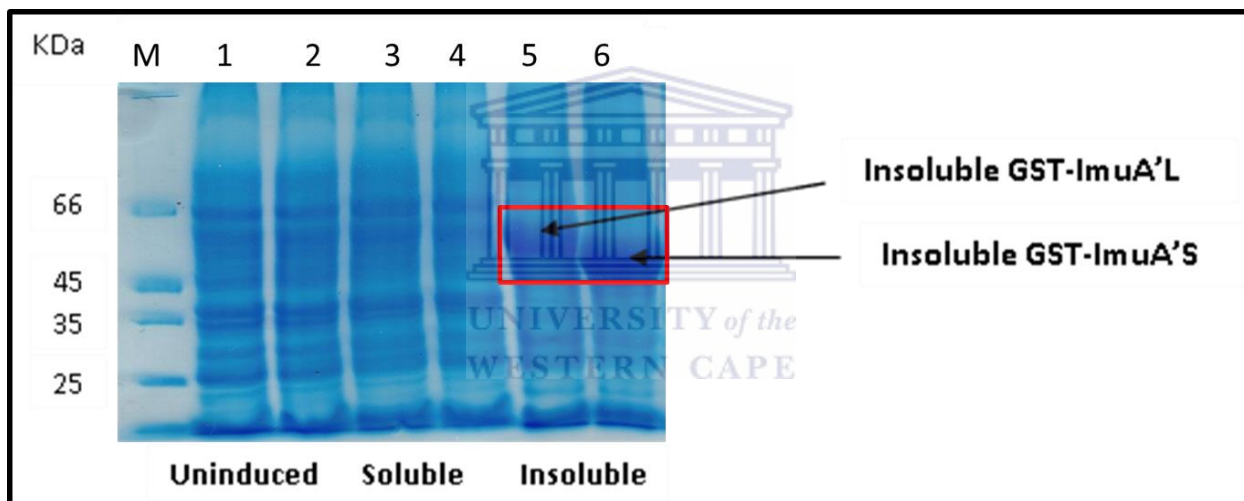


Figure 7: GST-ImuA'L produced in *E. coli* BL21 (DE3) cells following incubation at 30°C and induction with 0.5 mM IPTG. Lane M: Unstained Protein Molecular Weight Marker (Thermo Scientific); Lanes 1 and 2: Complete cells prior to addition of IPTG for protein production; Lanes 3 and 4: Soluble cell fractions of GST-ImuA'L and GST-ImuA'S after overnight induction; Lanes 5 and 6: Insoluble fractions of GST-ImuA'L and GST-ImuA'S. The protein produced is largely insoluble as indicated by the bands on lane 5 and 6 in the red box.

As the protein was predominantly insoluble, the induction conditions were varied to potentially increase the proportion of soluble protein (Section 2.7). First, the induction temperature was reduced to 25°C and 15°C, to enhance protein folding and solubilisation. The rate of protein

production at these lower temperatures was monitored by collecting samples after 2, 3, 4, and 6 h as well as overnight following induction with IPTG. However, temperature optimization did not visibly increase the proportion of soluble protein (Figure 9).

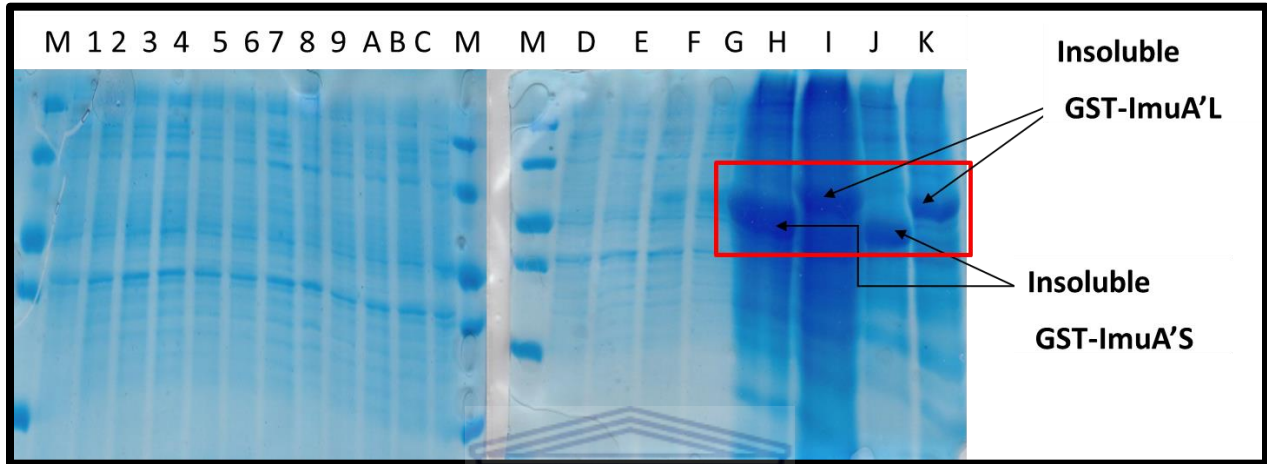


Figure 8: GST-ImuA' production from pETM-30_ *imuA'S* and pETM-30_ *imuA'L* plasmids in *E. coli* BL21 (DE3) cells at 15°C and 25°C, 0.5 mM IPTG. Lane M: Unstained Protein Molecular Weight Marker (Thermo Scientific); Lane 1: Uninduced fraction; Lanes 2 to F: soluble fractions, Lanes G to K: insoluble fractions; Lanes 2-5: GST-ImuA'S produced at 15°C, 2, 4, 6 and 15 h after induction; Lane 6-9: GST-ImuA'L produced at 15°C, 2, 4, 6 and 15 h after induction; Lanes A-C: GST-ImuA'S produced at 25°C and 3, 6 and 15 h after induction; Lane D-F: GST-ImuA'L produced at 25°C, 3, 6 and 15 h after induction. Lane G: Insoluble fraction before induction. Lane H: GST- ImuA'S after overnight induction at 15°C; Lane I: Insoluble fraction of GST- ImuA'L after overnight induction at 15°C; Lane J: Insoluble fraction of GST- ImuA'S after overnight induction at 25°C; Lane K: insoluble fraction of GST-ImuA'L after overnight induction at 25°C. The red box shows protein in the insoluble fractions.

The rate of induction was optimized by varying the IPTG concentration (Section 2.7.2) and temperature together. Protein production was induced at 15°C with 0.05, 0.1, and 0.25 mM IPTG and analysed after 6 h and overnight incubation (Figure 9). The protein was still observed to overwhelmingly occur in the insoluble fractions.

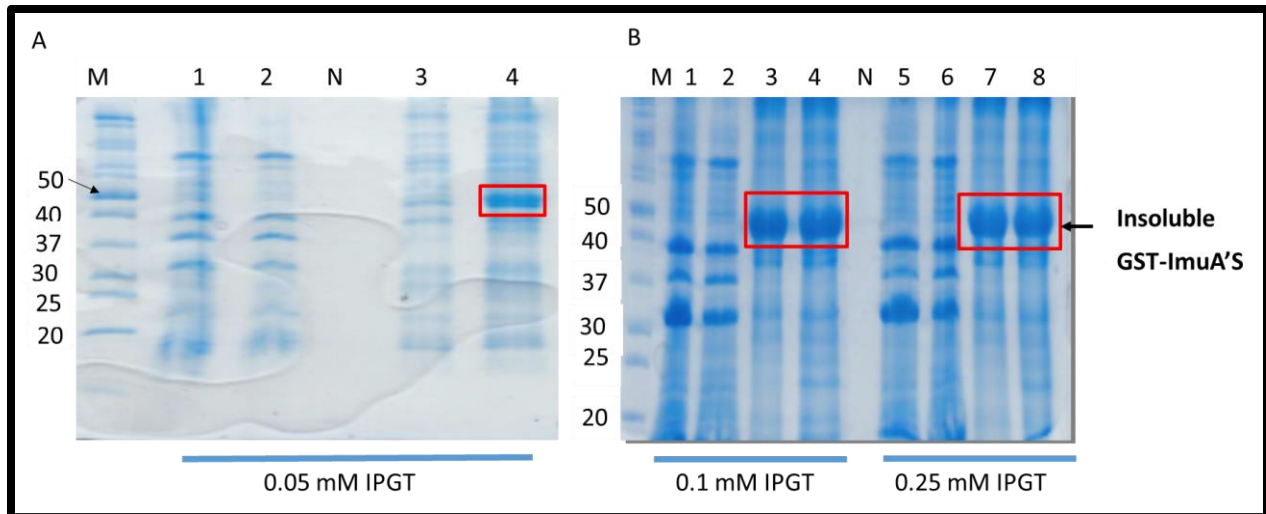


Figure 9: GST-ImuA'S produced from pETM-30_ImuA'S construct in *E. coli* at 15°C induced with varying IPTG concentrations. **A)** Lane 1: Soluble fraction before induction with IPTG; Lane 2: Soluble fraction after 0.05 mM IPTG induction; Lane 3: Insoluble fraction before induction; Lane 4: Insoluble fraction after induction with 0.05 mM IPTG. **B)** lane 1: Soluble fraction after 6 h induction with 0.1 mM IPTG; Lane 2: Soluble fraction after overnight induction with 0.1 mM IPTG; lane 3 and lane 4: Corresponding insoluble fractions of lane 1 and 2; Lane 6 and 7: 6 h and overnight soluble fractions induced with 0.25 mM IPTG; Lane 8 and 9: Corresponding insoluble fractions of lanes 6 and 7; N: Lanes that were not loaded with protein. GST-ImuA'S in the insoluble fractions are indicated with red rectangles. Lanes M: Odyssey Protein Molecular Weight Marker (Li-Cor).

UNIVERSITY of the
WESTERN CAPE

3.3.2 Production of ImuA' from pGEX-6P-2 and pCOLD I Vectors

The pGEX vector system is a classic system to produce proteins in *E. coli*. The vector encodes an N-terminal GST-tag that acts as a solubility partner to some insoluble proteins. Since the vector pETM-30 used initially for ImuA' production had both C- and N-terminal His₆-tags, which could interfere with protein folding and solubility, the genes of interest were cloned into pGEX-6P-2 and the resulting construct used to produce ImuA'S. The plasmid clone pGEX-6P-2_ImuA'S was transformed in *E. coli* BL21 (DES) cells, cultured at 25°C and induced for protein production with 0.5 mM IPTG. Protein production was analysed by SDS-PAGE (Figure 10).

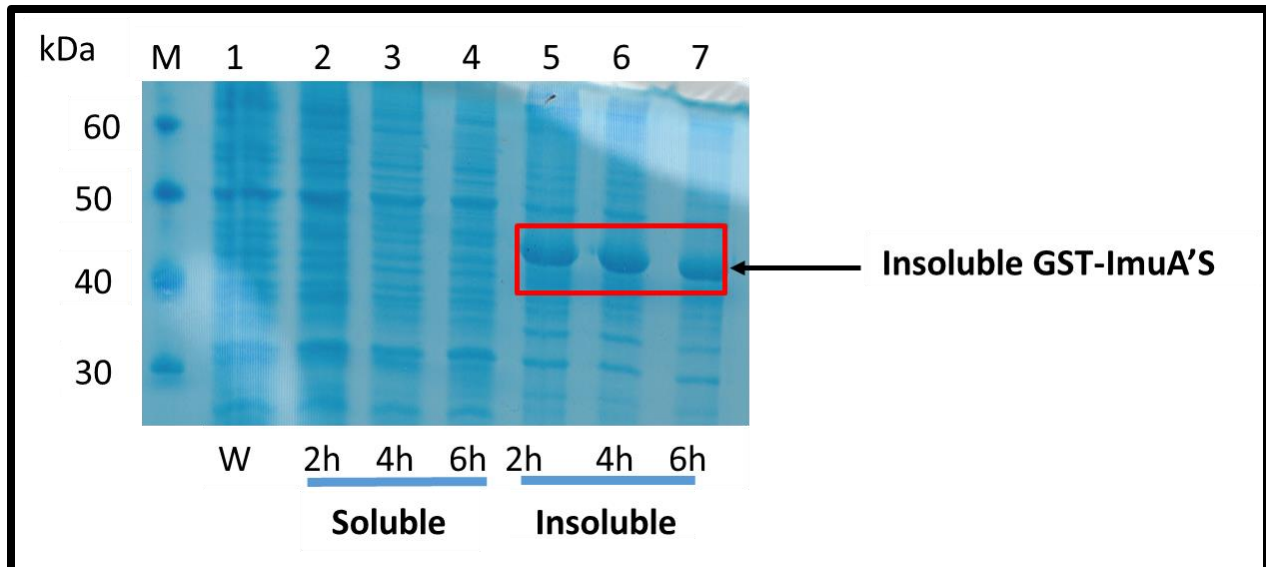


Figure 10: Production of ImuA'S as a recombinant GST-fusion protein from pGEX-6P-2_ImuA'S in *E. coli*. Lane M: PageRuler Unstained Protein Ladder (Thermo Scientific); Lane 1: Whole cell fraction of sample before induction; Lanes 2 to 4: Soluble fractions 3, 6 and 16 h after induction with 0.5 mM IPTG. Lane 5 to 7: Insoluble fractions corresponding to lanes 2 to 4. The red rectangle in lanes 5 to 7 marks the successful production of recombinant but insoluble GST-ImuA'S fusion protein.

As both pETM-30 and pGEX-6P-2 vectors successfully expressed *imuA'* but resulted in insoluble protein, the vector pCOLD I, a cold shock vector specially designed to increase solubility of proteins by producing them at low temperatures in *E. coli*, was used. The *imuA'L* gene was ligated into pCOLD I, transformed in *E. coli* BL21 (DE3) cells, cultured and induced with 0.5 mM IPTG for protein production at 15°C. Protein production was analysed by SDS-PAGE (figure 11).

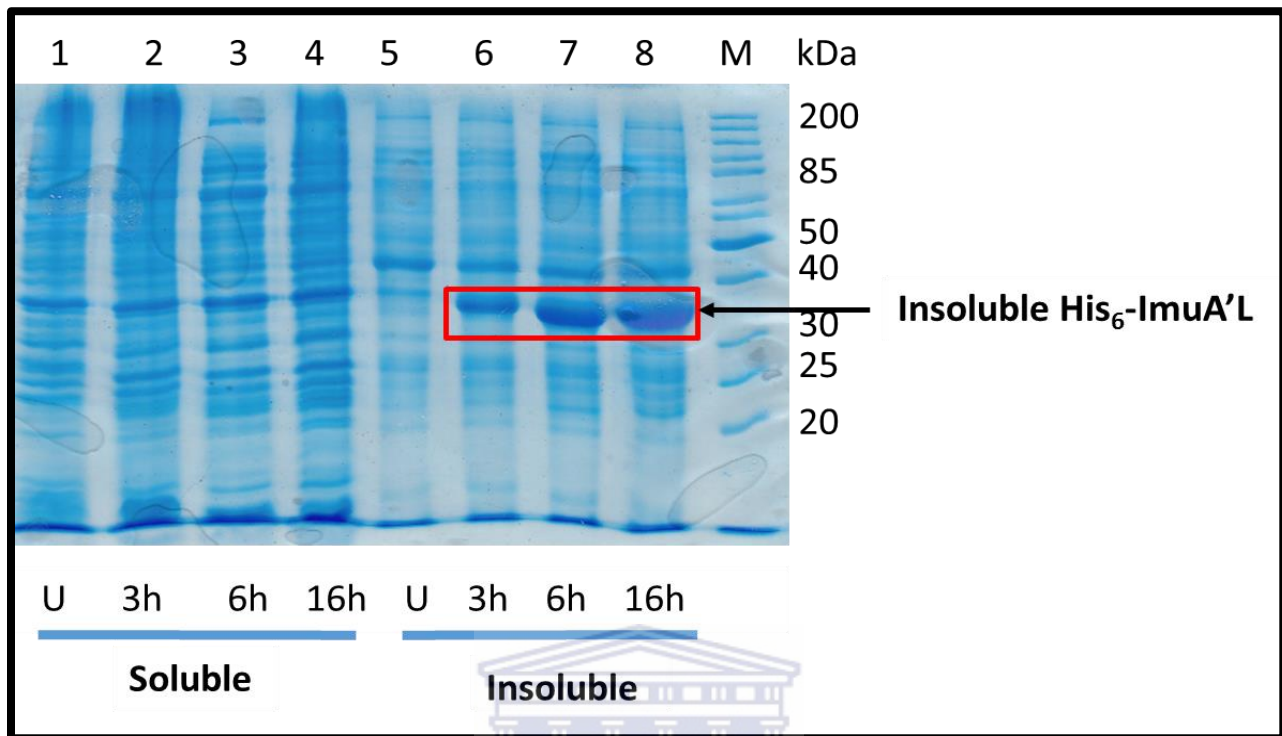


Figure 11: ImuA'L produced from pCOLD I_{imuA'L} construct as a recombinant His₆-tagged protein in *E. coli*. Lane 1: Sample before induction; Lane 2-4 Samples collected 3, 6 and 16 h after induction with 0.5 mM IPTG. Lane 6: Insoluble fraction before induction. Lane 7-9: Insoluble fractions corresponding to Lanes 2-4. Lane M: PageRuler Unstained Protein Ladder (Thermo Scientific). The protein band in the red box indicates insoluble His₆-ImuA'L.

3.3.3 Optimization of Lysis Buffer

The recombinant protein produced from all plasmids described up to this point using different induction conditions are essentially produced insolubly. Leibly (2012) hypothesized that a significant proportion of protein in insoluble fractions are originally soluble proteins that aggregate during cell lysis. Appropriate lysis buffers could prevent aggregation of soluble protein or solubilise partly aggregated protein. This hypothesis was investigated by using different lysis buffers to increase the yield of soluble *Mtb* proteins (Singh *et al.*, 2011). Buffer parameters tested included varying pH (6-9), the NaCl concentration (100-500 mM) and detergents to solubilise the

three insoluble *Mtb* proteins. Here, lysis buffers at different pH were used to support ImuA' solubility. However, the protein remained insoluble (Figure 12).

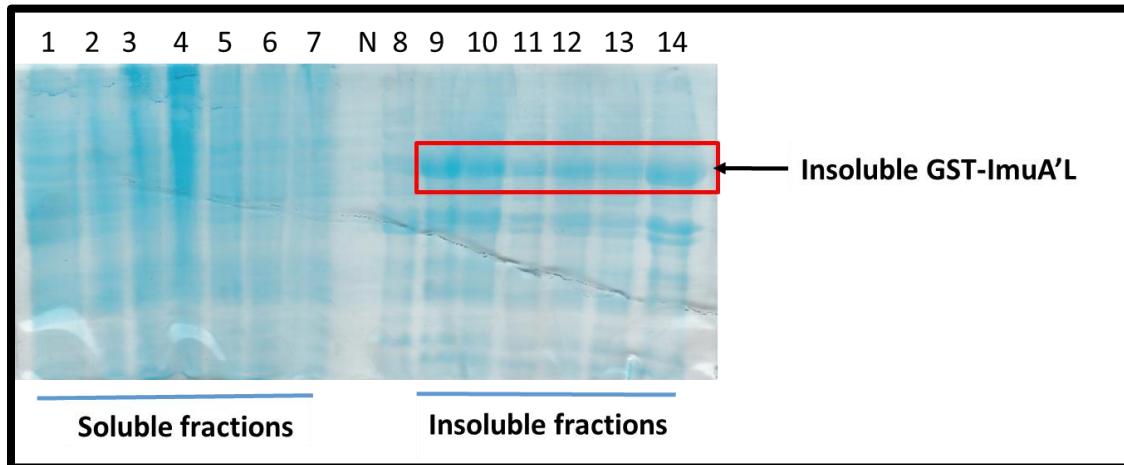
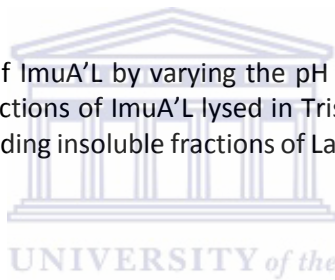


Figure 12: Attempted solubilisation of ImuA'L by varying the pH of the lysis buffer. Lane 1: Uninduced soluble fraction; Lane 2-7: Soluble fractions of ImuA'L lysed in Tris HCl buffer at pH 7, 7.5, 8, 8.5, 9, and 9.5 respectively; Lane 8-14: Corresponding insoluble fractions of Lane 2-7; Lane N: No protein. The protein remained insoluble (red rectangle).



3.3.4 Protein Production in *E. coli* BL21-CodonPlus Cells

Sequence analysis identified 15 *E. coli* rare codons in *imuA'* (Figure 3). The *E. coli* BL21-CodonPlus strain, optimised to provide tRNAs for rare *E. coli* codons, was used for ImuA'S production (Section 2.7.3). SDS-PAGE analysis, however, did not reveal either soluble or insoluble ImuA'S (Figure 13).

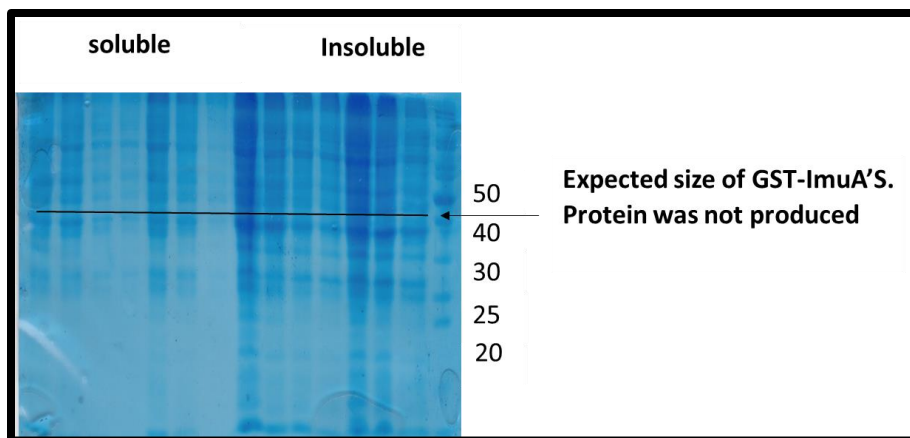


Figure 13: Production of GST-ImuA'S in BL21-CodonPlus cells. No visible GST-ImuA'S band is seen at the correct size (black line).

3.4 Solubilization and Refolding of GST-ImuA'S from Inclusion Bodies

To obtain functional protein from inclusion bodies, the target protein must first be solubilised and then refolded. First, the insoluble cell fraction encompassing inclusion bodies mixed with cell membranes is washed and proteins solubilised with buffers containing high chaotrope concentrations such as urea or guanidine HCl. Solubilized but unfolded proteins are then purified by liquid chromatography and allowed to refold to their original native conformation by reducing the denaturant concentration. Here, 8 M urea was used to solubilise GST-ImuA'S from inclusion bodies. Two refolding techniques were attempted for ImuA'S: refolding by dialysis and on-column refolding (section 2.8.3 and 4).

To refold a denatured protein by dialysis, the solubilised protein is dialyzed against a sequence of buffers with stepwise lower denaturant concentrations (Section 2.8.3). At 2 M urea concentration, white precipitate was observed in the protein solution indicating aggregation. The solution was centrifuged to remove the precipitate from potentially soluble protein that could

bind to GS-resin for purification by GST-affinity chromatography. GST-ImuA'S was, however, not observed to bind when analysing SDS PAGE, indicating that it had quantitatively precipitated.

3.4.1 On-column Refolding

For on-column refolding, the solubilised inclusion body solution was incubated with Ni-NTA resin to allow the protein to bind through its His₆-tag. The resin was then washed with buffers containing step-wise lower urea concentrations to potentially allow the protein to fold isolated from other protein molecules to prevent aggregation (Section 2.8.4). The protein was finally eluted with 2 M urea refolding buffer supplemented with 300 mM imidazole and analysed by SDS-PAGE (Figure 14).



Figure 14: On-column refolding of GST-ImuA'S from inclusion bodies. Lane 1: Insoluble GST- ImuA'L as size marker; Lane 2: slightly smaller GST-ImuA'S solubilized in 8 M urea and dialyzed to 4 M urea. Lane 3: Ni-NTA resin before adding the protein sample; Lane 4: Mixture of resin and protein sample; Lane 5-11: Wash fractions; Lane 12: Resin after removal of unbound protein. ImuA'S is seen to remain bound to the beads; Lane 13: Eluate after incubating the ImuA'-loaded resin in 300 mM imidazole overnight; Lane 14-15: Elution fractions with 500 mM imidazole; Lane 16: Resin after 500 mM imidazole wash to elute GST-ImuA'S. The GST-ImuA'S protein remains bound to the resin.

3.5 Co-production of His₆-ImuA'L and GST-ImuB

Based on the fact that ImuA' is known to interact with ImuB (Figure 1) (Warner *et al.*, 2010), we inferred that stability and hence solubility of ImuA' could be dependent on ImuB. To verify this hypothesis, two vector constructs, pCOLD 1_ImuA'L and pETM-30_ImuB (provided by Jeremy Boonzaier, fellow MSc. student, University of the Western Cape) were co-transformed into *E. coli* BL21 (DE3) cells. Protein production was induced with 0.5 mM IPTG for both proteins (Section 2.8). The resulting cells were lysed and soluble protein production was verified by Ni-NTA affinity chromatography (figure 15). The SDS-PAGE analysis revealed a protein band at the size of ImuA'L in the soluble fraction. However, this protein failed to bind to the Ni-NTA resin.

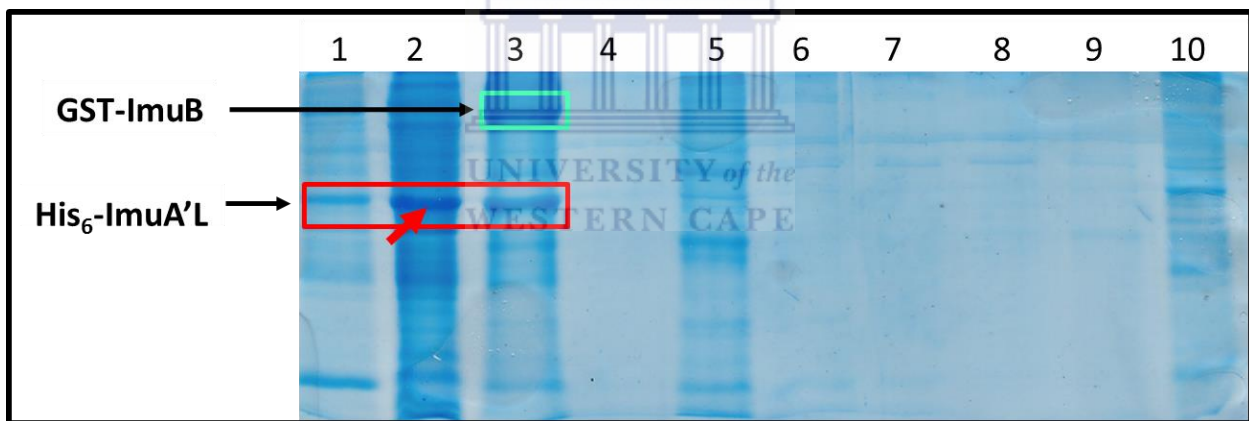


Figure 15: Co-production of His₆-ImuA'L and GST-ImuB: Lane 1: Whole cell fraction; Lane 2: Soluble fraction; Lane 3: Insoluble fraction; Lane 4: Ni-NTA resin prior to adding the soluble cell fraction; Lane 5: Ni-NTA resin plus soluble fraction; Lane 6-9: Wash fractions with 10, 20, 50, and 250 mM imidazole; Lane 10: Resin after washing with 250 mM imidazole. Co-producing ImuA'L and ImuB does produce soluble ImuA'L (red arrow). This protein, however, did not bind to Ni-NTA resin. Red rectangle: His₆-ImuA'L; green rectangle: insoluble GST-ImuB.

3.6 Modelling the Three-Dimensional Structure of ImuA'L

Three dimensional (3D) structures are invaluable sources of information for protein structure/function relationships. Though best determined experimentally by X-ray crystallography, nuclear magnetic resonance spectroscopy or electron microscopy, any one of the steps required to achieve the final aim, including protein production, purification and stability, may fail, preventing the desired outcome. In the absence of experimental structures, a homology model of the protein in question can still provide useful information. In this project, the experimental structure determination of ImuA' by X-ray crystallography was impeded by problems in folding and the resulting insolubility of ImuA'. A model structure of ImuA'L was instead generated using the SWISS-MODEL, a web-based protein modelling server. Four steps were required: First, a protein BLAST search of the UniProtKB/Swiss-Prot database with ImuA'L as the target returned 15 partly homologous protein sequences (Figure 16). ImuA'S is listed as the best hit with a sequence identity of 100% (narrow red line). However, the lack of an experimental structure precludes its use. The next hit, an ATP-dependent helicase with a sequence coverage of 54% and E-value of 0.22 was also excluded due to the lack of an experimental structure. The next three hits from the ATP-dependent hit were RecA proteins. Though without experimental structures themselves, crystal structures of homologues from *E. coli* and *M. smegmatis* have been determined. Their PDB codes are 3CMW and 1UBC, respectively. *M. smegmatis* RecA crystal structure was chosen for homology modelling of ImuA'L over the *E. coli* RecA for two reasons: first, ImuA'L and 1UBC are both mycobacterial proteins and may evolutionally be more closely related than the *E. coli* counterpart; second, the crystal structure of *E. coli* RecA is solved in complex with DNA potentially altering the solution structure

of the protein. *M. smegmatis* RecA and ImuA'L were aligned using the NCBI BLAST alignment tool. The alignment showed significant homology with an E-value of 7×10^{-5} and a sequence identity of 32%. The sequence coverage is only 23% implying that structural information is limited to around a quarter of ImuA'L.

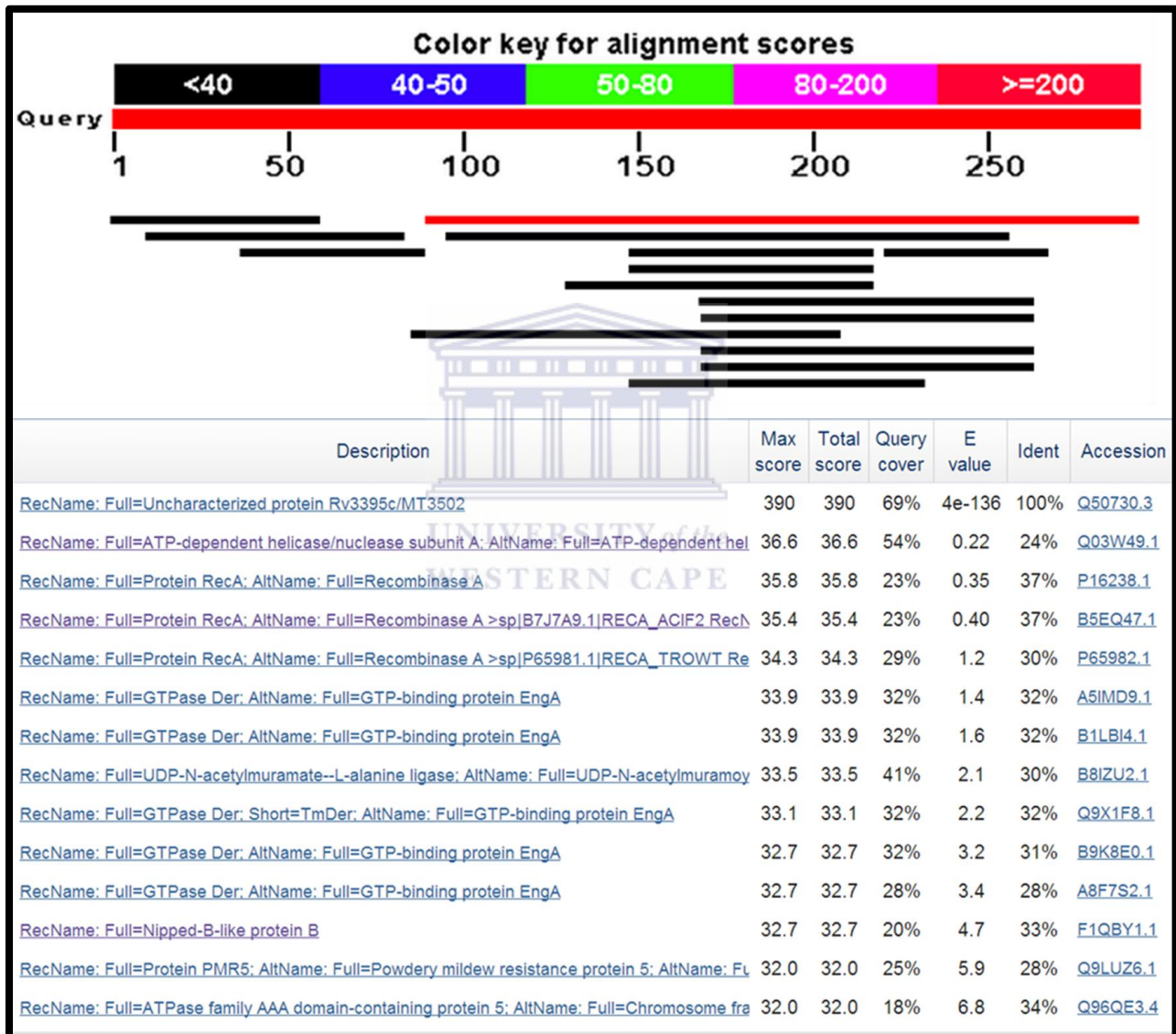


Figure 16: Protein sequences sharing significant sequence identity with ImuA'L. The red bar marks a protein sequence perfectly aligned to the target – here ImuA'S. Black lines indicate sequence alignments of $\leq 40\%$. Homology modelling requires templates with sequence identities $\geq 30\%$ for reliable model building. The crystal structure of RecA with a sequence identity of 37% would appear to represent a good starting point to model ImuA'L. A sequence coverage of only 23%, however, indicates that less than a quarter of the length of ImuA'L is covered.

in Figure 18. A comparative modelling package then adjusts side chain conformations to minimize collisions, and improves the model by energy minimization and/or molecular dynamics.

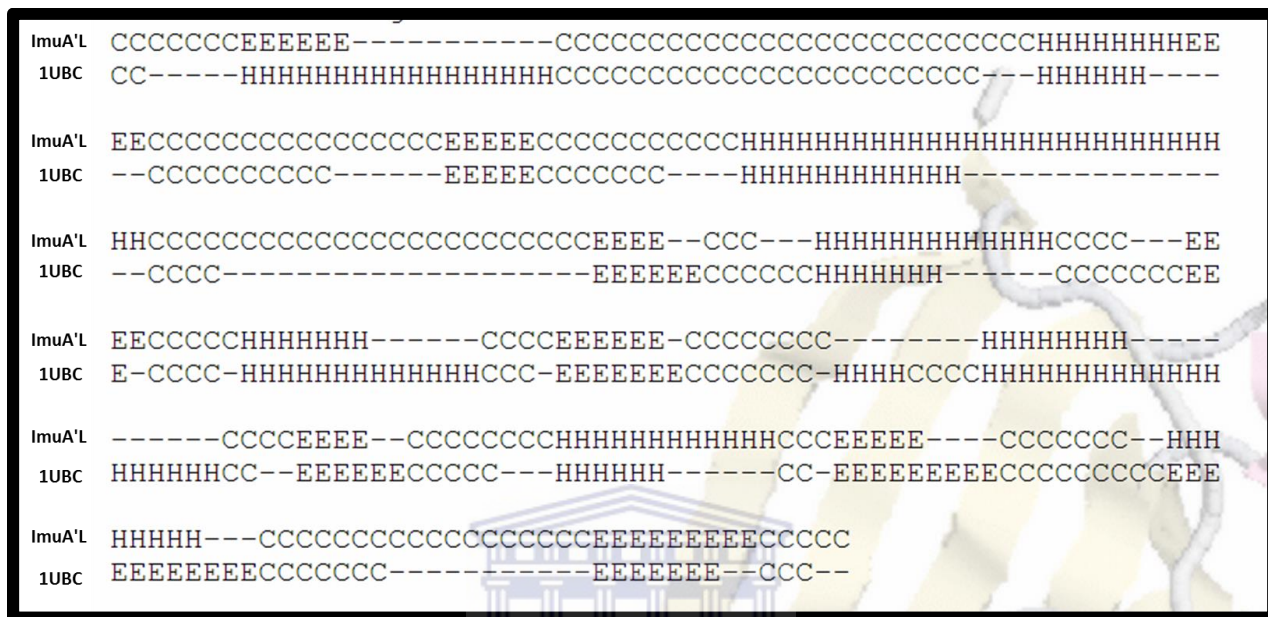


Figure 18: Aligned secondary structure of ImuA'L and structure 1UBC from *M. smegmatis*. H: α -helices, E: β -strands; C: loop regions; dashed lines insertions/deletions.

UNIVERSITY of the
WESTERN CAPE

ProMod-II of the SWISS-MODEL pipeline generates an all-atom model for the target using the sequence alignment by “modelling by segment matching” or “coordinate reconstruction” based on the observation that most protein hexapeptides cluster to around 100 structural classes. Models are constructed from a subset of guide template positions (conserved $C\alpha$'s) connected by short all-atom segments from known protein structures. Where loop modelling is not satisfactory, MODELLER is used to generate an alternative model.

The graphics program PyMol (PyMOL Molecular Graphics System, Version 1.5.0.4 Schrödinger, LLC.) was used to graphically compare the template and target model structures (Figure 19).

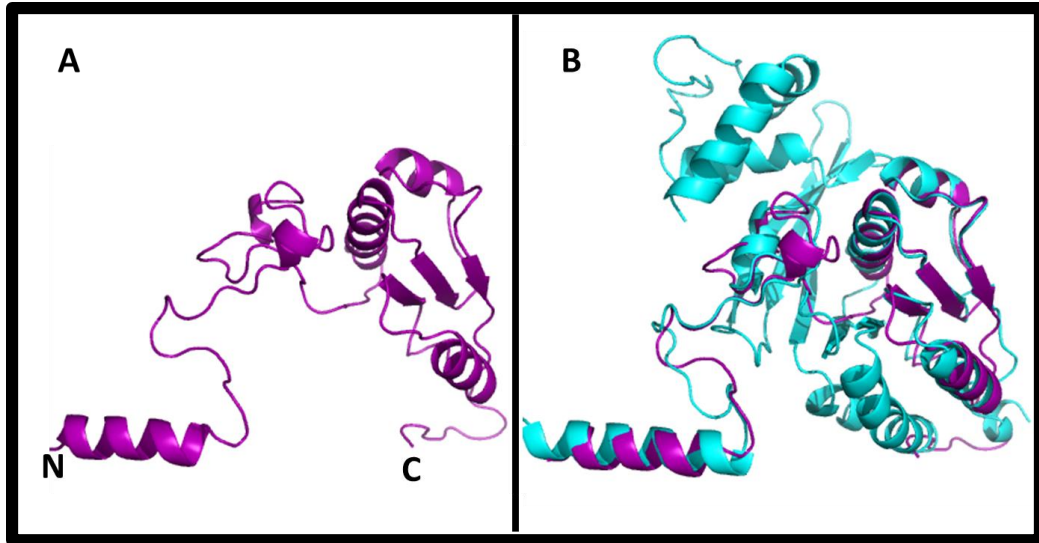


Figure 19: Partial model for ImuA'L. **A)** Structural model of ImuA'L generated by SWISS-MODEL; **B)** The structural model of ImuA'L superimposed on the experimental structure of RecA from *M. smegmatis* (PDB code 1UBC).

The ImuA' model accounts for residues 85 to 235, roughly covering the first domain of ImuA'S or the central domain of ImuA'L. The N-terminal domain of ImuA'L (residues 1 to 90) and the C-terminal domain of ImuA'S/L (235-294) are not covered.

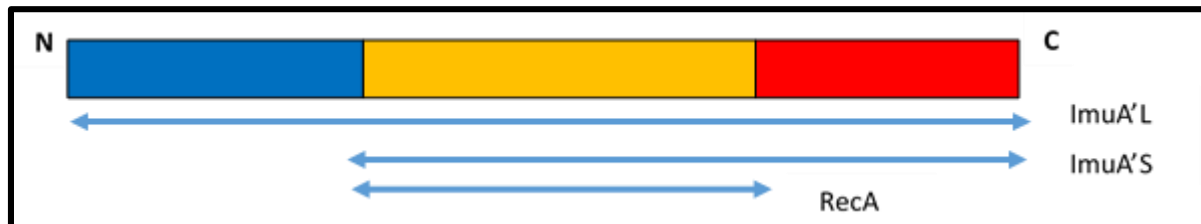


Figure 20: Three potential domains of ImuA'L. Blue: The N-terminal domain (residues 1 to 90); Orange: central domain (91 to 235) homologous to RecA; Red: C-terminal domain (235-294). ImuA'S maps to domains two and three of ImuA'L.

An equivalently derived model for the N-terminal domain reveals a $\beta\alpha\beta$ -motif in the first 39 residues with high sequence identity to a stretch of *E. coli* pyrimidine nucleoside hydrolase

(Figure 21). No significant identity to any known protein could be identified for the C-terminal domain, preventing a reliable structure-based model being derived.

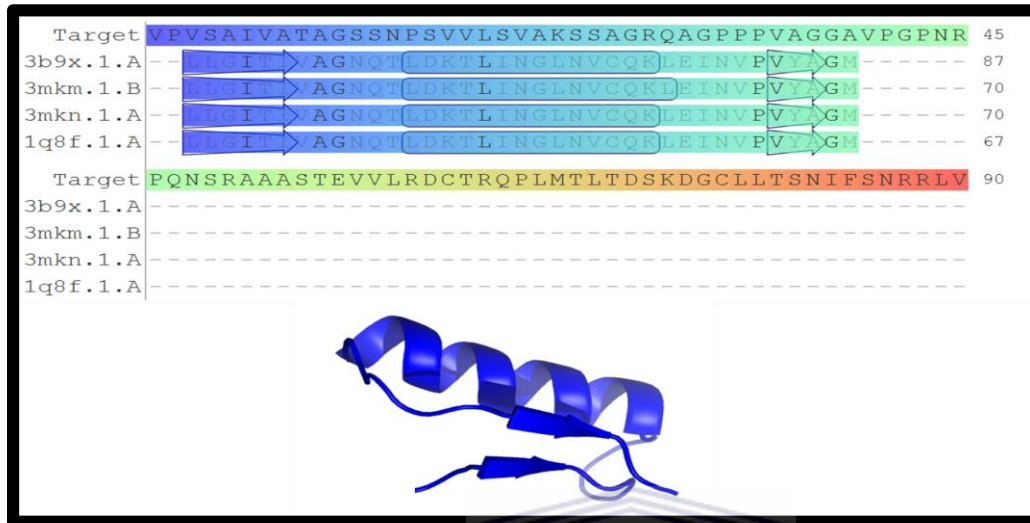
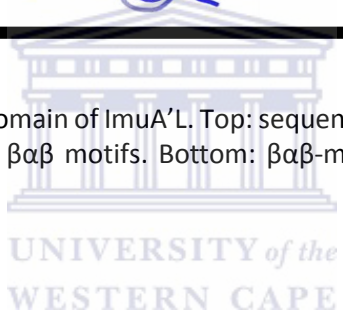


Figure 21: Modelling the N-terminal domain of ImuA'L. Top: sequence alignment of the ImuA'L N-terminal domain with four sequences sharing $\beta\beta$ motifs. Bottom: $\beta\beta$ -model structure for ImuA'L N-terminal domain.



To investigate DNA binding by ImuA'L, the crystal structure of RecA from *E. coli*, solved in complex with double stranded DNA (PDB code 3CMW) (Chen *et al.*, 2008), was superimposed on the ImuA'L model structure in PyMol. Recognition of the double stranded DNA by RecA involves the loops L1- α and L2- α of the latter (green in figure 22). Residues involved include Glu154, Ser172 and Arg176 in L1- α and Ile199, Met197, Lys198, Thr208, Gly200, Gly211, Gly212 and Asn213 in L2- α (Chen *et al.*, 2008). The L2- α loop is absent in ImuA'L implying that ImuA'L does not bind DNA like RecA. The L1- α loop is present in ImuA'L but lacks the α -helical portion (residues?). The L1 loop of ImuA'L consists of only hydrophobic residues, which could bind DNA. If ImuA'L were thus to bind DNA it would do so in a manner unrelated to RecA.

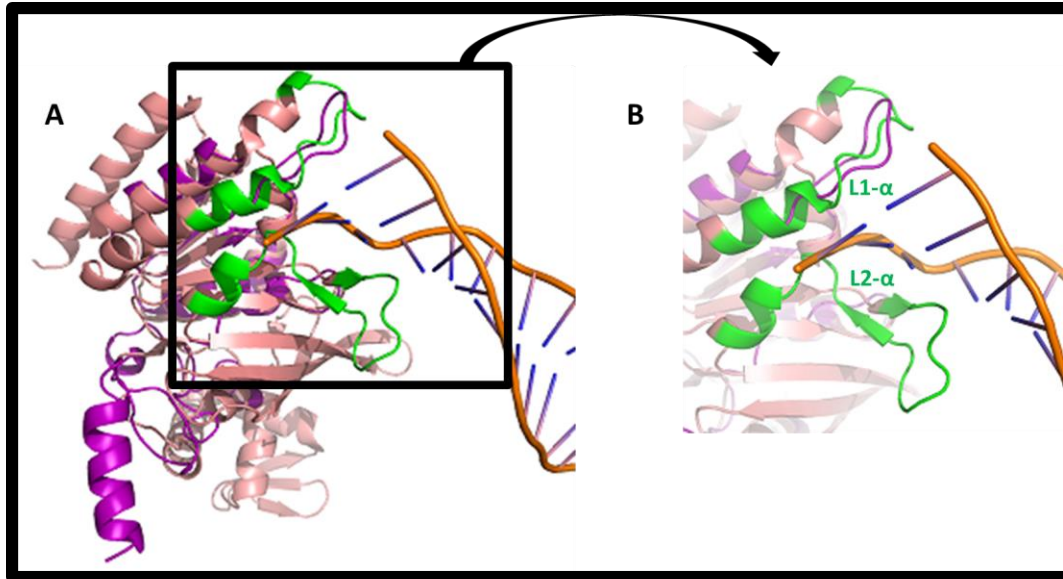
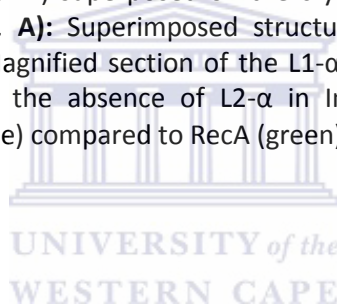


Figure 22: ImuA'L model structure (pink) superposed on the crystal structure of the *E. coli* RecA/DNA complex (PDB code: 3CMX; brown). **A):** Superimposed structures showing limited physical contact between the ImuA'L and DNA. **B):** Magnified section of the L1- α and L2- α loops of RecA/DNA (green) known to interact with DNA. Note the absence of L2- α in ImuA'L and the significantly different conformation of L1- α in ImuA'L (purple) compared to RecA (green).



3.6.1 Model Evaluation

Estimating the quality of a modelled structure is essential as models more similar to the original crystal structure are generally more reliable for practical applications. SWISS-MODEL provides an estimate of the model quality based on a QMEAN potential, which compares geometrical features of the model such as pairwise atomic distances, torsion angles and solvent accessibility to statistical distributions from experimental structures. Residue scores range from 0 to 1 with higher scores indicating higher reliability. A global QMEAN, calculated to estimate the overall model quality, is provided as a Z-score relating the QMEAN to scores for X-ray structures. The QMEAN is combined with GMQE values from the target-template alignment to give a combined quality estimate for the model structure. Again, values near one indicate the most reliable model.

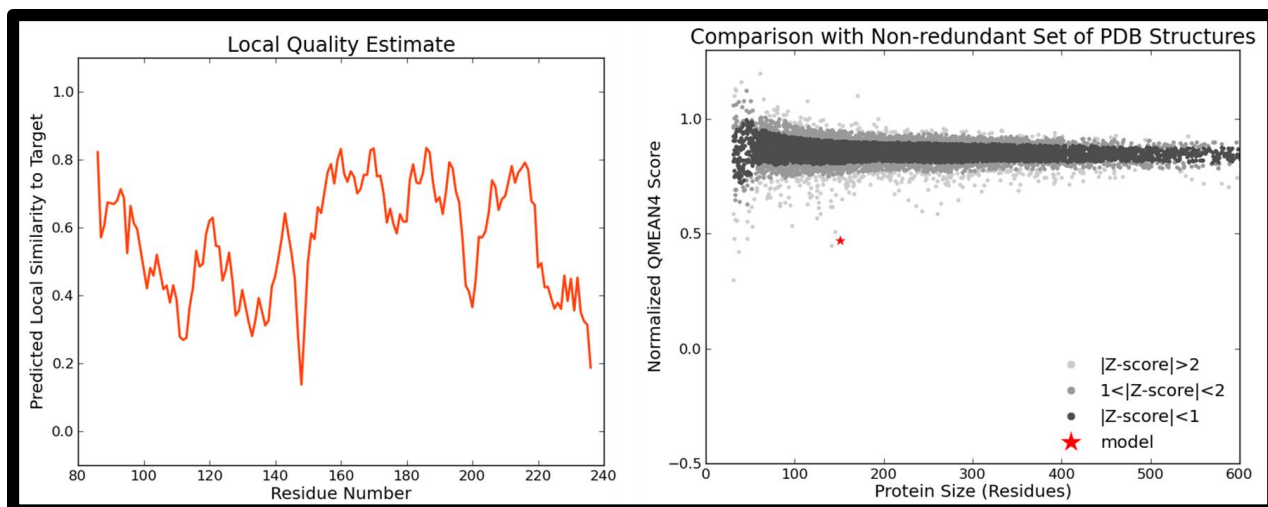
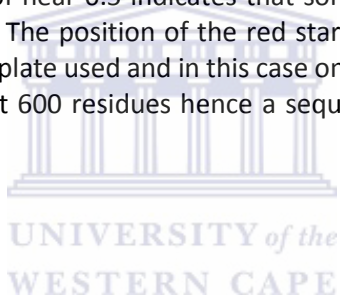


Figure 24: Model Quality Estimate for ImuA'L. Left: QMEAN estimate for each residue. Peaks are generally above 0.5 indicating fairly reliable residue positioning of model compared to template. Right: Overall estimate of model quality. The shaded region marks Z-scores of template structures. The red star indicates the quality of the model. A QMEAN of near 0.5 indicates that some useful information may be derived from the structural model of ImuA'L. The position of the red star along the X-axis indicates the overall model coverage compared to the template used and in this case only about 160 residues in the target are modelled on the template with about 600 residues hence a sequence coverage of less than 26% in the model.



4 Discussion

4.1 General Statement of Research Outcome

This study aimed at structurally analysing ImuA' from *Mtb*. This revealed that 1) *imuA'* of *Mtb* exist as two variants, *imuA'S* and *imuA'L*; 2) The amplification of *imuA'* was achieved by a two-step PCR but not by the conventional three step PCR procedure; 3) Large amounts of recombinant ImuA' are produced in *E. coli* albeit as insoluble protein; 4) Modelling of ImuA'L indicate its middle domain to be homologous to *M. smegmatis* RecA. The un-modelled C-terminal domain is known to be critical for the interaction of ImuA' with ImuB (Ndwandwe, 2013; Warner *et al.*, 2010).

Both ImuA' and ImuB are critical to the functioning of DnaE2 (ImuC) (Ndwandwe, 2013; Warner *et al.*, 2010), a C-family DNA damage-inducible polymerase implicated in translesion synthesis, virulence and the emergence of antibiotic resistance in *Mtb* (Boshoff *et al.*, 2003). DnaE2 lacks a β -clamp binding motif. The β -clamp is functionally replaced by the pseudo-polymerase ImuB (Warner *et al.*, 2010). The C-terminal loops of both ImuA' and ImuB are critical for their roles as DNAE2 accessory proteins (Ndwandwe, 2013) despite the precise role of ImuA' remaining elusive.

4.2 Analysis of the *Mtb imuA'*

The *Mtb imuA'* gene provides an example of uncertainties in *Mtb* genome annotation resulting from *Mtb* using transcription start codons other than ATG. Two start sites have been proposed as the origin of *imuA'*. The CMR JCVI database lists *imuA'* genes for both start sites. One, labelled *imuA'L* in this work, starts at a GTG codon 270 bp upstream of the ATG codon of the other, here denoted *imuA'S*. The first 270 bases of *imuA'L* are shared by another 5' end of another gene

encoded by the opposite strand. Databases such as the TBDB and TubercuList correspondingly only capture the *imuA'S* sequence. The model structure of ImuA'L generated in this study (Figure 19) and the findings of Ndwandwe (2013) suggest that the *imuA'S* and *imuA'L* are transcribed under different conditions, with the N-terminal domain of ImuA'L having a unique function. Inserting an artificial start site 399 bp downstream of the *imuA'L* start codon (or 129 bp downstream of the *imuA'S* start codon) still yields a protein that binds ImuB (Ndwandwe, 2013), indicating that the first 133 amino acids of ImuA'L are not required for its interaction with ImuB. The nearest ImuA'L homologue, RecA, is structurally related to the central domain of ImuA'L or the first domain of ImuA'S.

4.3 Two-Step Amplification of *imuA'*

The high GC content of both *imuA'S* (70%) and *imuA'L* (68%) proved challenging during amplification in that primers had to be shorter than the recommended 18 bp to achieve melting temperatures between 52 and 58°C. This, however, resulted in unspecific priming as the G or C at the 3'-end of the primers allow rapid, non-specific annealing to the template. To restore the specificity, primers were designed with T or A at their 3'-end. In addition, the annealing step was eliminated, to prevent non-specific primer binding and amplification of truncated *imuA'* due to different GC-contents of the 5'- and 3'-ends. With a much higher GC content at the 5'-end of the gene, primers had highly disparate melting temperatures potentially causing mis-annealing during the annealing step. Eventually the full-length gene was amplified using a two-step PCR. This method similarly proved successful for the amplification of *imuB*, *imuC* and *carD* from *Mtb* (personal communication: Jeremy Boonzaier, UWC, Simon Broadly, UCT), and would appear to be the method of choice to amplify genes from GC-rich genomes in general.

4.4 Production of ImuA' for Structural Studies

Protein structural studies often fail due to difficulties in producing soluble protein or crystallizing a protein (Arbing *et al.*, 2013; Christendat *et al.*, 2000; Terwilliger *et al.*, 2009). In this case, producing soluble, recombinant ImuA' in *E. coli* proved an insurmountable hurdle. Invariably, SDS-PAGE analysis would show the protein to be produced well but to be channelled to insoluble inclusion bodies. This observation is consistent with *in silico* prediction of the solubility of recombinant ImuA' in *E. coli*, which estimated the chances of producing soluble, recombinant ImuA' in *E. coli* at less than 7% (Table 7). Generally, *Mtb* proteins are difficult to produce solubly in *E. coli* (Chim *et al.*, 2011). The reasons may include 1) the high GC content and resulting secondary structures in the mRNA 5'-regions; 2) bias in codon usage between *Mtb* and *E. coli*; properties of protein such as protein size, hydrophobicity and acidity; and 3) the lack of essential chaperones (Arbing *et al.*, 2013; Allert *et al.*, 2010). The GC-content of 66% in the *Mtb* genome contrasts with 51% in *E. coli*. However, the high GC-content alone is unlikely to be the reason for the insolubility of ImuA' in *E. coli* as genes from other mycobacterial species with similar GC-contents are efficiently produced as soluble proteins in *E. coli* (www.webtb.org/Targets).

Differences in codon usage or codon usage bias between production host and gene source organism is another factor known to influence protein solubility (Angov, 2011). By altering the rate of mRNA translation, protein folding is affected (Fedyunin *et al.*, 2012). Rare or slow codons are normally located at regions in an mRNA corresponding to protein domain boundaries or at transition points between structured and unstructured regions of the polypeptide (Angov, 2011; Saunders & Deane, 2010; Zhang *et al.*, 2009). The associated slowing of transcription allows sufficient time for domain folding. Differences in host and source codon usage results in this

message being lost, affecting protein folding and insolubility. *E. coli* strains have been engineered to supply normally rare tRNAs (Baca & Hol, 2000; Redwan, 2006). The use of BL21-CodonPlus cells that provide additional tRNAs did not improve the solubility of ImuA'.

Physicochemical properties of a protein also significantly influence its solubility. Generally, low molecular weight (<60 kDa), low hydrophobicity and moderately acidic proteins are likely to be more soluble (Slabinski *et al.*, 2007). High pI values, by comparison, decrease solubility of proteins in *E. coli* (Mehlin *et al.*, 2006). The poor solubility of ImuA' in *E. coli* may be caused by a combination of these properties. With molecular weights of 20.8 kDa and 29.9 kDa for ImuA'S and ImuA'L, both proteins are within the range for production in *E. coli*. However, both are highly hydrophobic and have pI-values of 10.7 and 11.2, resulting in misfolding and protein insolubility. The solubility of protein may further be affected by the conditions of induction. The vectors into which *imuA'* was cloned had a *lac* promoter, which induce gene expression when lactose is present. Here, the lactose metabolite analogue isopropyl β -D-1-thiogalactopyranoside (IPTG) was used to induce gene expression. Varying induction conditions including IPTG concentrations, induction temperatures and induction time were explored for soluble production of ImuA'. However, varying these variables independently and in combination did not improve the protein solubility.

Insoluble proteins are mostly channelled to inclusion bodies (Williams *et al.*, 1982). However, a significant proportion of proteins may initially be produced in a soluble form only to aggregate after cell lysis (Leibly *et al.*, 2012). Subjected to the optimal lysis buffers, they may remain in solution. Conditions required include an ideal pH, ionic strength or various additives. Here, buffers with a range of pH values were tested for ImuA' solubilisation, however, ultimately

without success (Figure 12). The insolubility of ImuA' therefore does not appear to be linked to aggregation of soluble protein.

Proteins in inclusion bodies do not adopt a native fold but can be often be solubilised, purified and refolded to their native conformations (Vallejo & Rinas, 2004). Inclusion body proteins are further protected from by cellular proteases and contain target proteins in high concentration with little cellular material. Large quantities of fairly pure protein can therefore often be isolated from inclusion bodies (Li *et al.*, 2004; Singh & Panda, 2005). In this study, significant quantities of inclusion bodies were isolated, washed and solubilised in 8 M urea. However, refolding of solubilised protein was not successful. *Inter alia*, the urea concentration was gradually lowered by dialyzing against buffers with successively lower urea concentrations, to allow for protein to slowly fold back to a native conformation. However, protein precipitated as the urea concentration reached about 2 M indicating that folding did not succeed. In “on-column” refolding, the urea concentration was gradually lowered to 3 M after immobilising the protein on Ni-NTA resin – again to allow for slow refolding. However, eluting with 3 M urea buffer and increasing imidazole concentrations did not dislodge the protein from the resin even after overnight incubation (Figure 14) implying non-specific binding to the column due to improper folding. These outcomes confirm that “refolding is an empirical process, which needs to be optimized in each individual case to achieve reasonable yields of the protein in its functional form” (Tsumoto *et al.*, 2003). Optimal conditions have so far eluded us.

Yet a further strategy to produce soluble protein is to produce proteins together with their physiological partners. A colleague failed in an attempt to co-produce ImuB and ImuA' using the pETDuet system (Novagen) (Ndwandwe, 2013). Therefore, co-transformation of *imuA'* in the

pCOLD I vector (ampicillin resistance), and *imuB* in pETM-30 (kanamycin resistance) and co-expression in *E. coli* BL21 (DE3) cells was attempted in this study. Production of both proteins improved and an estimated 50% ImuA' was seen to be soluble. However, the resulting ImuA' protein failed to bind to Ni-NTA resin preventing its purification (Figure14). Clearly, however, the solubility of ImuA' appears to depend on ImuB being present, indicating that various routes of co-expression may need further refinement.

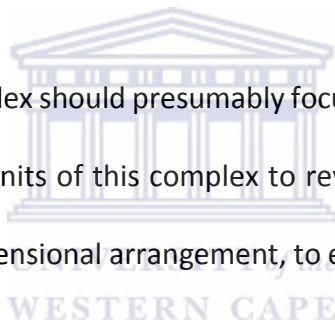
4.5 Insights from the model structure of ImuA'L

Sequence analysis of ImuA'L revealed the protein to consist of three domains: 1) An N-terminal domain, unique to ImuA'L, containing a $\beta\alpha\beta$ -motif similar to that of pyrimidine nucleoside hydrolysing enzymes; 2) a central domain related to the N-terminal domain of *M. smegmatis* RecA; and 3) a C-terminal domain of unknown fold. Despite homology to RecA, the central domain lacks critical loops for DNA binding losing most contacts observed between *E. coli* RecA and DNA. A possible exception is a C-terminal stretch of hydrophobic amino acids which could indicate some residual DNA binding (Figure 22). Possibly ImuA'L does not bind DNA itself but interacts with DNA-binding proteins such as ImuB during translesion synthesis. The C-terminal domain of ImuA'L is predicted to be highly disordered on its own but is critical for the interaction with ImuB (Ndwandwe, 2013). The details of this interaction, however, remain to be unravelled.

5 Conclusion and Outlook

The structural characterization of *Mtb* ImuA' has been carried out in this work. The genes were successfully amplified by two-step PCR, cloned and expressed in *E. coli*. The proteins were produced in large quantities in *E. coli* though were invariably channelled to inclusion bodies. Attempts to refold the protein similarly proved unsuccessful. Instead, structural information for ImuA' was derived from homology modelling. The structural model revealed the central domain of ImuA' to be related to N-terminal domain of RecA despite lacking some loop regions necessary for DNA binding. The C-terminus of ImuA' is disordered but binds ImuB by an unresolved mechanism.

Future studies on this critical complex should presumably focus on co-producing ImuA', ImuB and ImuC (DnaE2) plus any other subunits of this complex to reveal the exact stoichiometry of the subunits as well as their three-dimensional arrangement, to establish how the complex is able to restore DNA translesion.



Part II

Structural Analysis of a Gh9 C1 Cellulase from a Metagenomic Library



1. Introduction

1.1. Biofuels: Better Alternatives to Fossil Fuels

Cellulases are enzymes that catalyse the hydrolysis of cellulose. Cellulases are important in modern industry and hold great potential in biofuel production (Yan & Wu, 2013). The rapid depletion of fossil fuel deposits and the deteriorating impact of fossil fuel combustion on the environment necessitate the exploration of cleaner and renewable energy sources such as biofuels. Countries such as Brazil and the United States of America have made significant progress in producing and consuming biofuels. Overall, though, biofuels account for less than 10% of total world fuel usage (IEA, 2013).

Biofuels are energy sources derived from organic materials i.e. plant materials or animal waste. Biofuels are classified as first, second or third generation. First generation biofuels derive from food crops such as grains, sugar beet, oil seeds and sugar cane (Lee & Lavoie, 2013). The use of food crops for fuel production raises questions about food security in what is referred to as the 'food/fuel problem' (Tenenbaum, 2008). Attention has therefore shifted to biomass such as agricultural and forest residues for second generation biofuel production (Sims *et al.*, 2008). This biomass is rich in lignocellulose, the most abundant biomaterial on earth by mass, and constitutes an alternative to both first generation biofuels and fossil deposits (Sánchez & Cardona, 2008). Currently, however, cost-effective technologies to convert biomass to second generation biofuels on an industrial scale are lacking mainly due to the recalcitrance of lignocellulose to degradation (Xia *et al.*, 2013).

1.2. Analysis of the Components of Lignocellulose

Lignocellulose from plant cell walls consists of the components cellulose, hemicellulose and lignin. Cellulose, the most abundant constituent, is a β -(1–4)-linked polymer of glucose arranged in layers held together by hydrogen bonds. Two forms of cellulose, crystalline and amorphous cellulose often occur together in a 3:1 ratio respectively (Somerville *et al.*, 2004; Quiroz-Castañeda & Folch-Mallol, 2013).

Hemicellulose is the second most abundant component of lignocellulose. It is made up of different pentose and hexose sugars such as arabinose, galactose, glucose, mannose and xylose linked β -(1–4) to create a polysaccharide chain (Scheller & Ulvskov, 2010). The hemicellulose chain forms hydrogen bonds with cellulose microfibrils driving the overall assembly of cell wall structure. Counterintuitively, the hydrogen bonds cross-linking hemicellulose and cellulose prevent cellulose aggregation and mechanically weaken the cell walls to allow for cell wall expansion (Somerville *et al.*, 2004).

Lignin, the third constituent of lignocellulose, consists of three phenolic components, namely p-coumaryl (H), coniferyl (G) and sinapyl alcohols (S) that polymerize in different ratios in plants, wood tissue and cell wall layers. Lignin controls cell wall porosity, adhesion of adjoining cells and the ionic environment of the cell wall (Somerville *et al.*, 2004; Wi *et al.*, 2005).

Cellulose, hemicellulose and lignin come together into micro and macro-fibrils as depicted on figure 25 below (Rubin, 2008).

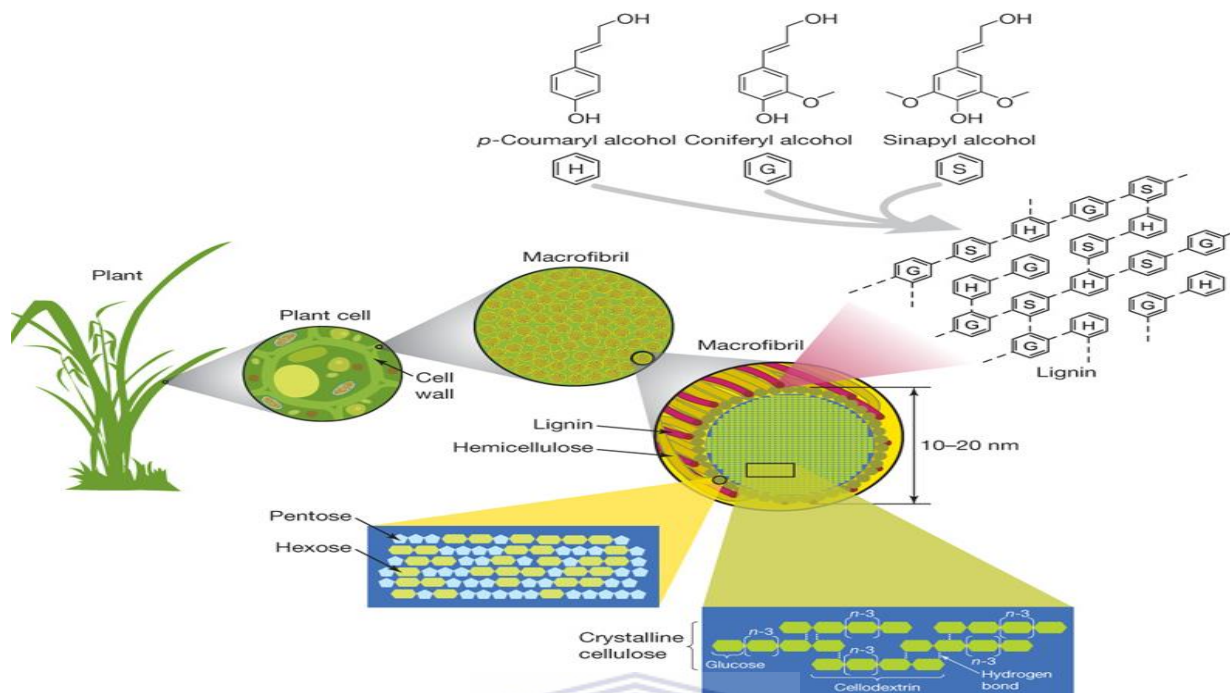


Figure 25: Structural depiction of the composition of lignocellulose - the most abundant biomass on earth. Image obtained with permission, from EM Rubin (2008).

UNIVERSITY of the
WESTERN CAPE

1.3. Glycoside Hydrolase for Cellulose Degradation

Lignocellulose degradation is a bottleneck for large-scale, industrial production of biofuels. Cellulolytic enzymes could overcome this hurdle (Wang *et al.*, 2009). Three cellulase types that synergistically convert cellulose to glucose are endoglucanases (EC 3.2.1.4), exoglucanases (EC3.2.1.91) and β -glucosidases (EC3.2.1.21) (Lynd *et al.* 2002).

Glucose is easily fermented to useful chemicals. However, hydrolysis of cellulose to glucose by cellulases is inefficient, making biofuels from lignocellulosic biomass uneconomical (Duan *et al.*, 2009). Understanding the biology of cellulases may be critical in overcoming this bottleneck in biofuel production.

Cellulases are glycoside hydrolases that hydrolyse β -(1–4)-glycosidic linkages of cellulose (Sandgren, 2003). Glycoside hydrolase (GH) enzymes are found in all domains of life but differ in enzymatic activity with marginal changes in primary structure significantly affecting substrate specificity. GHs have different domain structures with each domain having a unique evolutionary history (Naumoff, 2011).

Interest in glycoside hydrolases started during the first oil crisis as cellulose in plant biomass was recognized as a potential source of energy. The biotechnological application of GHs may remedy the current environmental crisis ensuring they remain a focus of research (Sulzenbacher *et al.*, 1996; Xia *et al.*, 2013).

1.3.1. Classification of Glycoside Hydrolases

GH enzymes are classified according to their sequences and structures. Currently the “Carbohydrate Active Enzyme” (CAZY) database recognizes 133 GH families. Family members share catalytic machinery, molecular mechanism and/or glycosidic bond geometry (Gebler *et al.*, 1992; Henrissat & Davies, 2000).

Mechanistically GH enzymes are “retaining” or “inverting” depending on the configuration of the anomeric oxygen before and after the reaction (Koshland, 1953). Retaining GHs use a general acid/base and a nucleophile during catalysis while inverting GHs require a catalytic acid and a catalytic base for catalysis (Vuong & Wilson, 2010). Variations from retaining and inverting mechanisms occur in some GH families, though members of one GH family mostly share a mechanism of action. Family GH97 is a notable exception in accommodating both retaining and inverting GH enzymes (Gloster *et al.*, 2008). Families GH4 and GH109 utilize the cofactor NAD in

glycosidic bond hydrolysis and cleave both α - and β -glycosidic bonds (Chakladar *et al.*, 2014; Yip *et al.*, 2004).

GH families are grouped into overarching 'clans'. Each clan comprises families with similar tertiary structure, catalytic residues and enzymatic mechanism. Clan members potentially derive from a common ancestor (Henrissat *et al.*, 1996).

Table 8: Cellulase Families and Clans

Clan	Common fold	Families within clan
GH-A	$(\beta/\alpha)_8$	1 2 5 10 17 26 30 35 39 42 50 51 53 59 72 79 86 113 128
GH-B	β -jelly roll	7 16
GH-C	β -jelly roll	11 12
GH-D	$(\beta/\alpha)_8$	27 31 36
GH-E	6-fold β -propeller	33 34 83 93
GH-F	5-fold β -propeller	43 62
GH-G	$(\alpha/\alpha)_6$	37 63
GH-H	$(\beta/\alpha)_8$	13 70 77
GH-I	$\alpha+\beta$	24 46 80
GH-J	5-fold β -propeller	32 68
GH-K	$(\beta/\alpha)_8$	18 20 85
GH-L	$(\alpha/\alpha)_6$	18 20 85
GH-M	$(\alpha/\alpha)_6$	8 48
GH-N	β -helix	28 49

1.3.2. Glycoside Family 9 (GH9) Enzymes

The much studied GH9 family mainly includes cellulases such as endoglucanases, cellobiohydrolases, β -glucosidases and *exo*- β -glucosaminidase (Guérin *et al.*, 2002). GH9 enzymes use an inverting catalytic mechanism and are mechanistically distinct from all other GH families precluding their assignment to any clan. Structurally they share an $(\alpha/\alpha)_6$ fold with aspartate and glutamate residues acting as catalytic nucleophile/base and catalytic proton donor respectively (Sakon *et al.*, 1997). Around 1620 GH9 enzyme-encoding genes have been identified in all domains of life with 844 eukaryotic, 561 bacterial, four archaeal and 211 unclassified genes. Eleven crystal structures of proteins from this family have been determined - ten bacterial and one eukaryotic (<http://www.cazy.org/GH9.html>).

1.4. Metagenomic Potential for Cellulose Deconstructing Enzymes

Despite much research into prokaryotic cellulases, efficient degradation of cellulose has not been achieved. Metagenomics, which involves the direct analysis of DNA fragments from environmental samples for novel genes and gene products, may yet yield better cellulose deconstructing enzymes for an industrial application (Xia *et al.*, 2013). Despite countless attempts to isolate and characterize cellulases from uncultured microbial communities, the full biotechnological potential of environmental cellulases remains to be realized (Ferrer *et al.*, 2005; Healy *et al.*, 1995; Kim *et al.*, 2008).

1.5. The GH9 C1 Cellulase

This study is geared towards the analysis of the structure of a GH9 cellulase from an uncultured environmental sample. The enzyme named GH9 C1 cellulase was isolated from straw based mushroom compost from Medallion mushroom farm, Stellenbosch, South Africa. Samples were collected over two periods and a library constructed from fresh and frozen compost. The metagenomic libraries were screened and putative gene cloned into protein expression vectors (personal communication Stephen Mackay, University of the Western Cape).

The *GH9 C1 cellulase* gene was cloned and expressed from a pET21a vector. Using chromatographic methods, the protein was purified and characterized. The protein showed maximum activity between pH 5.5 and 7.0 and a temperature optimum between 50 and 55°C. GH9 C1 cellulase has a low thermostability and loses activity at low temperatures. It has a high specific enzyme activity (personal communication Stephen Mackay). Low thermostability coupled to high enzyme activity and specificity may imply a fast turnover rate and rapid production in natural environments.

The aim of this project is to structurally analyse this novel GH9 C1 cellulase. The protein was to be produced, purified, crystallized and its crystal structure determined.

2. Materials and Methods

A recombinant plasmid containing a gene encoding GH9 C1 cellulase cloned by Stephen Mackay (Institute of Microbial Biotechnology and Metagenomics, IMBM, University of the Western Cape) was provided for this study. The vector was constructed as shown in Figure 26.

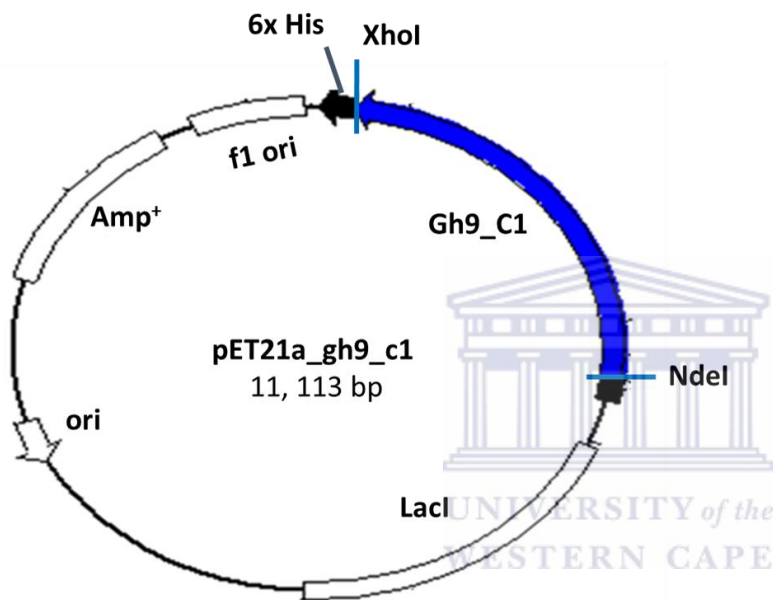


Figure 26: pET21a_gh9_c1 is a pET21a based expression vector with a His₆-tag encoding sequence at the C-terminus of the gene encoding Gh9_C1 (blue). The gene is cloned between a 5'-NdeI and 3'-XhoI restriction enzyme sites. The plasmid contains an ampicillin resistance gene (Amp⁺) for antibiotic selection.

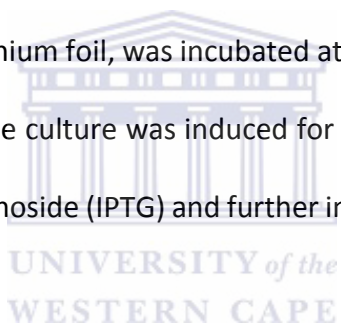
2.1. Protein Production and Purification

2.1.1. Starter Culture

A starter culture was prepared by inoculating 50 mL LB medium (100 µg/mL ampicillin) with pET21a_C1-SP transformed *E. coli* BL21 (DE3) and incubating at 37°C and 170 rpm shaking overnight.

2.1.2. Main Culture

The starter culture was used to inoculate 1 L of LB medium (100 µg/mL ampicillin) in a 5 L culture flask. The flask, capped with aluminium foil, was incubated at 37°C with shaking at 170 rpm until the OD₆₀₀ was between 0.6-0.8. The culture was induced for recombinant gene expression with 1 mM isopropyl-D-thiogalactopyranoside (IPTG) and further incubated overnight as before but at 30°C.



2.1.3. Cell Harvesting and Sonication

The final culture was transferred to two 500 mL centrifugation tubes and centrifuged at 11 000 x g and 4°C for 15 min. The supernatant was decanted from the cell pellet and the latter resuspended in chilled lysis buffer with 1 mM serine protease inhibitor PMSF. The cells were ruptured on ice by sonication at amplitude 28 by six cycles of 30 s sonication and 30 s resting time to prevent protein heat denaturation. 0.4 µg/mL DNase was added to the ruptured cells, incubated on ice for 20 min and centrifuged at 16 000 x g and 4°C for 1 h to separate soluble from insoluble cellular components.

2.2. Protein Purification

2.2.1. Nickel–nitrilotriacetic Acid (Ni⁺²–NTA) Affinity Chromatography

The *GH9 C1 cellulase* construct (Figure 26) was C-terminally His₆-tagged to assist purification of recombinant protein by immobilized metal-affinity chromatography (IMAC). Histidine readily binds to immobilized metal ions through its imidazole ring (Bornhorst & Falke, 2000).

Two millilitres of Ni⁺²-NTA were washed with 5 column volumes (CV) of water, equilibrated with 5 CV of lysis buffer and mixed with the soluble cellular component (see above). The mixture was agitated on a roller mixer overnight at 4°C and transferred to a gravity flow purification column. Unbound proteins were collected as the flow through. The matrix was washed with 5 CV of lysis buffer collected as wash fractions. Further washing followed with 2 CV each of 5, 10, 25, 50 and 100 mM imidazole enhanced lysis buffer. All remaining proteins were finally eluted with 250 mM imidazole. Samples were analysed by SDS PAGE (sections 2.3.5- 2.3.7).

2.2.2. Anion Exchange Chromatography

Fractions identified by SDS PAGE to contain GH9 C1 cellulase were pooled, sterile filtered using a 0.2 µm filter and further purified by anion exchange Hi-trap QHP chelating column (GE Healthcare). The column was equilibrated with 10 CV of buffer A (25 mM Tris-HCl pH 8.5 and 50 mM NaCl), rinsed with 1 CV of buffer B (25 mM Tris-HCl pH 8.5 and 1 M NaCl) to ensure all anion binding sites are occupied by Cl⁻ and followed by protein loading. The column was washed with buffer A until all unbound protein had eluted (OD₂₈₀ returned to baseline). Proteins

remaining bound to the column were eluted with a linear gradient (0-100%) of buffer B. Fractions within peaks of the chromatogram were analysed by SDS PAGE for protein.

2.2.3. Size-exclusion Chromatography

Samples from the anion exchange peak and confirmed by SDS PAGE to contain protein were pooled and concentrated to a final volume of 1 mL. The concentrated sample was injected onto a Superdex 200 16/60 column (GE Healthcare) with running buffer (25 mM Tris-HCl pH 8.5, 50 mM NaCl) and a flow rate of 1 mL/min. The elution profile was evaluated using the absorbance at 280 nm.

2.2.4. Protein Concentration

GH9 C1 cellulase fractions corresponding to the chromatographic peak and analysed by SDS-PAGE were pooled and concentrated by ultracentrifugation using a Vivaspin-6 concentrator (Sartorius, Germany) with a MWCO of 30 kDa. The absorption at 280 nm (A_{280}) was measured at intervals of centrifugation and the process terminated once the required concentration was achieved. To exchange the buffer or reduce the salt concentration, the concentrated protein was diluted with suitable buffer and re-concentrated to the desired volume.

2.2.5. Protein Quantification

Aromatic amino acids absorb UV light at a wavelength of 280 nm. The protein concentration of GH9 C1 cellulase was determined by measuring the A_{280} on a Nandrop ND-1000 spectrophotometer (PeqLab, Germany). For concentrated protein samples, the flow through of the

concentrator was used as a blank solution. The molar extinction coefficient ϵ_{280} was calculated from the amino acid composition. According to the Beer-Lambert equation

$$C = A_{280} / \epsilon \cdot D \quad \text{Equation 1}$$

the concentration C may be calculated if the values A , D , ϵ are known. The thickness of the measuring chamber D is mechanically adjusted by the Nanodrop spectrophotometer. This method does not distinguish individual proteins in a mixture. Hence the concentration of a protein is only accurate if contaminants are negligible.

2.2.6. Sample Preparation for Crystallization

The protein was concentrated to 10 mg/mL in a buffer containing 50 mM NaCl and 25 mM Tris-HCl pH 8.0. The concentrated sample was filtered using either a 13 mm syringe filter with a pore size of 0.20 μm or a micro-centrifuge tube with a molecular weight cut off of 10 kDa. The concentrated and filtered samples were stored at 4°C for crystallization screening.

2.3. Protein Crystallization Screen

Crystallization experiments were performed by sitting-drop vapour-diffusion method in 96-well crystallization plates (Hampton Research) using the Mosquito HTS (TTP Labtech, U.K.) dispensing robot. Commercial crystallization screens were used providing a range of precipitant, salt, buffer and pH formulations. Screens used include JCSG CORE I, II, III, IV and CORE+ suites, AmSO₄, MPD, PACT, PEGs and PEGs II Suites (Qiagen) and Crystal Screen, Crystal Screen II and Crystal Screen Lite (Hampton Research Corp). Crystallization drops consisted of 300 nL protein solution with 150 nL reservoir solution equilibrated against 80 μL of reservoir solution at 18°C.

Conditions producing lead crystal were manually optimized by varying the physicochemical parameters such as precipitant and protein concentrations, ionic strength and temperature on hanging drop vapour-diffusion 24-well crystallization plates (Hampton Research). Further optimization was attempted using 1) Additive Screen (Hampton Research, USA), 2) by micro, cross and streak seeding, 3) by varying the ratio of protein: reservoir volumes, and 3) by adding 0.5-3% glycerol.

2.3.1. Data collection and Evaluation

Single crystals were harvested from their mother liquor, rapidly transferred to cryoprotectant solution (mother liquor supplemented with 20-30% glycerol or 20-30% PEG 400) and flash-cooled either in liquid nitrogen or by mounting them in a liquid nitrogen stream at 100 K.

A home-source, rotating-anode X-ray generator (MicroMax-007HF, Rigaku, Japan) was ramped to an input current of 30 mA and a voltage of 40 kV resulting in a vacuum ion gauge value of 130-160. Diffraction data was collected using Cu-K α radiation ($\lambda = 1.54 \text{ \AA}$) using a Saturn 944HG CCD detector (Rigaku, Japan). The crystals were X-irradiated for 10 s and two test images 90° apart collected. Denzo and Scalepack of the HKL3000 suite (Minor *et al.*, 2006; Otwinowski & Minor, 1997) were used for data evaluation and scaling.

2.4. Modelling the Three-Dimensional Structure of GH9 C1 Cellulase

Comparative or homology modelling of GH9 C1 cellulase involved four steps: 1) template identification, 2) template-target alignment, 3) model building and 4) model quality evaluation.

A BLASTp search of the protein databank (PDB) using the amino acid sequence of GH9 C1 cellulase returned homologues of experimentally determined crystal structures. On the basis of a

sequence coverage of 96%, an E-value of 1×10^{-84} , and a sequence identity of 32%, the crystal structure of the cellobiohydrolase (CbhA) from *Clostridium thermocellum* (PDB ID: 1UT9) was chosen as the template to model GH9 C1 cellulase. The target and template amino acid sequences were aligned using ClustalW2 and submitted to the user interface of SWISS-MODEL running under alignment mode. The SWISS-MODEL pipeline uses the program ProModII to thread the target sequence onto the coordinates of the template structure. The model quality is fine-tuned using Gromos96, an energy minimization program. The output models were viewed and analysed using the molecular graphics program PyMol.



3. Results and Discussion

Three-dimensional structures of proteins contain valuable information about their function and, potentially, their evolutionary origins. This information can be gathered through structural techniques such as NMR and X-ray crystallography, which require extensive preparation ranging from protein production and purification as well as crystallization for X-ray crystallography. Generally, the protein purification strategy has to be rigorously optimised to yield a sample of highly homogenous and concentrated protein for crystallization. In this study, GH9 C1 cellulase, a 63 kDa protein obtained from a compost metagenomic library, was produced as recombinant protein in *E. coli*, extracted and purified by immobilized metal-affinity, ion exchange and size exclusion chromatography. The protein was crystallized and the crystals X-rayed for diffraction data. The resolution of the diffraction data was, however, too weak for use in structure determination by molecular replacement. The structure of GH9 C1 cellulase was therefore obtained by homology modelling and presented using the molecular graphics program PyMol.

3.1. Production and Purification of GH9 C1 Cellulase

The GH9 C1 cellulase was easily induced and overproduced as a His₆-tagged protein in *E. coli* BL21 DE3 cells. The protein was recovered in the soluble fraction with minimal protein lost in the insoluble fraction. Induction conditions with increasing IPTG concentrations at 30°C did not increase protein yield. GH9 C1 cellulase co-purified with a minor protein impurity running just below the major protein band. This contamination remained even after His₆-tag purification. Initially this band was suspected to be nickel-binding Hsp60 from *E. coli* produced under stress conditions. Alternatively it could represent a degradation product of the target protein.

Purification using protease inhibitor PMSF, protein production in a protease-free strain (Rosetta 2), and cold induction did not prevent the co-purification of the minor product, already visible in the crude extract. Ion exchange chromatography and size exclusion chromatography similarly did not remove the co-purified protein. Zymogram and Western blot analysis of the co-purified protein indicated cellulase cleaving activity and the presence of a His₆-tag (personal communication Stephen Mckay, IMBM, UWC). The co-purified protein is thus most likely to be an N-terminal degradation product of the target protein, retaining its enzymatic activity and its His₆-tag. Figure 27 represents a summary of the purification of GH9 C1 cellulase by Ni²⁺-NTA affinity chromatography where the lower band constitutes the N-terminal degradation product.

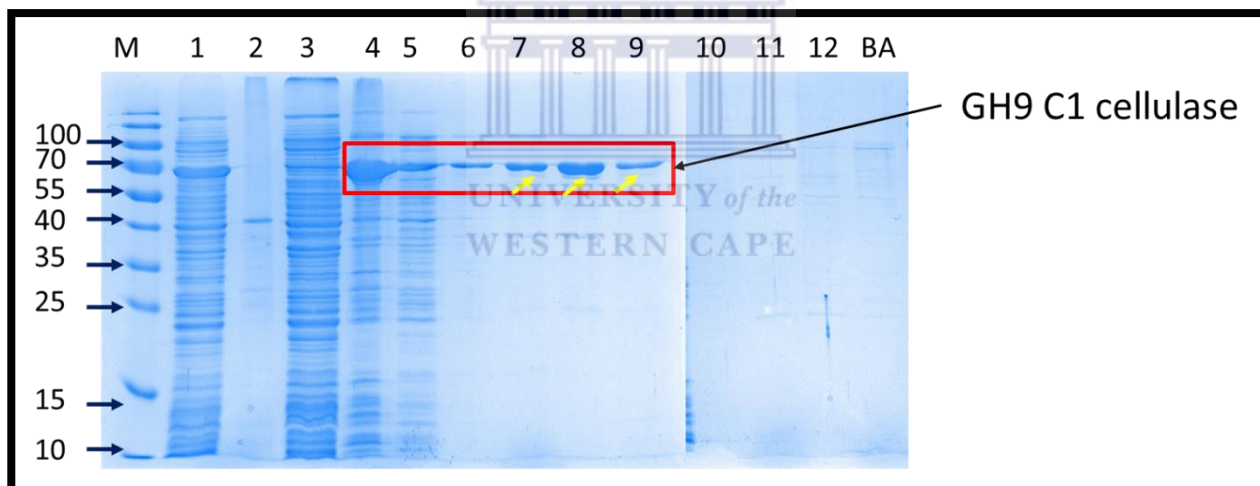


Figure 27: Ni²⁺-NTA affinity purification of GH9 C1 cellulase. Lane 1: Soluble cellular component; Lane 2: Ni²⁺-NTA resin before loading sample; Lane 3: Flow through after overnight agitation of sample on resin; Lane 4: Resin and protein sample; Lane 5-10: Wash fractions with increasing imidazole concentrations as follows: 0, 5, 10, 25, 50 mM and 75 mM respectively; Lane 11 and 12: Elution fractions with 100 and 250 mM imidazole respectively; Lane M: PageRuler Prestained Protein Ladder (Thermo Scientific) with sizes on the left in kDa. Lane BA: Resin sample after elution with 250 Mm imidazole. Red rectangle indicate the GH9 C1 cellulase and the yellow arrow indicate the N-terminally degrading band of GH9 C1 cellulase.

GH9 C1 cellulose eluted from the Ni-NTA beads even before imidazole was present in the wash buffer (lane 5). By step 5 with 50 mM imidazole all protein had effectively been eluted from the beads indicating that His₆-GH9 C1 cellulase has a low affinity for Ni²⁺-NTA matrix. The low imidazole concentration in the pooled fractions 6-9 was easily removed by dialysis. The sample was concentrated to a final volume of 30 mL before loading on a Q HP column.

3.1.1. Anion Exchange Chromatography (AEX)

The partially purified GH9 C1 cellulase from affinity chromatography (Lanes 6-9, Figure 27) was dialyzed to remove imidazole, concentrated to 30 mL and further purified by AEX. The theoretical isoelectric point (pI) of GH9 C1 cellulase was calculated to be 4.6 using the Expasy web-based server (Gasteiger *et al.*, 2005). As the pI represents the pH at which the protein has a net neutral electric charge, it will have a net negative charge for any pH above the pI and positive below. In AEX, the matrix will retain negatively charged molecules whereas cation exchange chromatography (CEX) binds positively charged molecules. The AEX chromatogram for GH9 C1 cellulase on an ÄKTA chromatography system (section 2.2.2) is shown in figure 28.

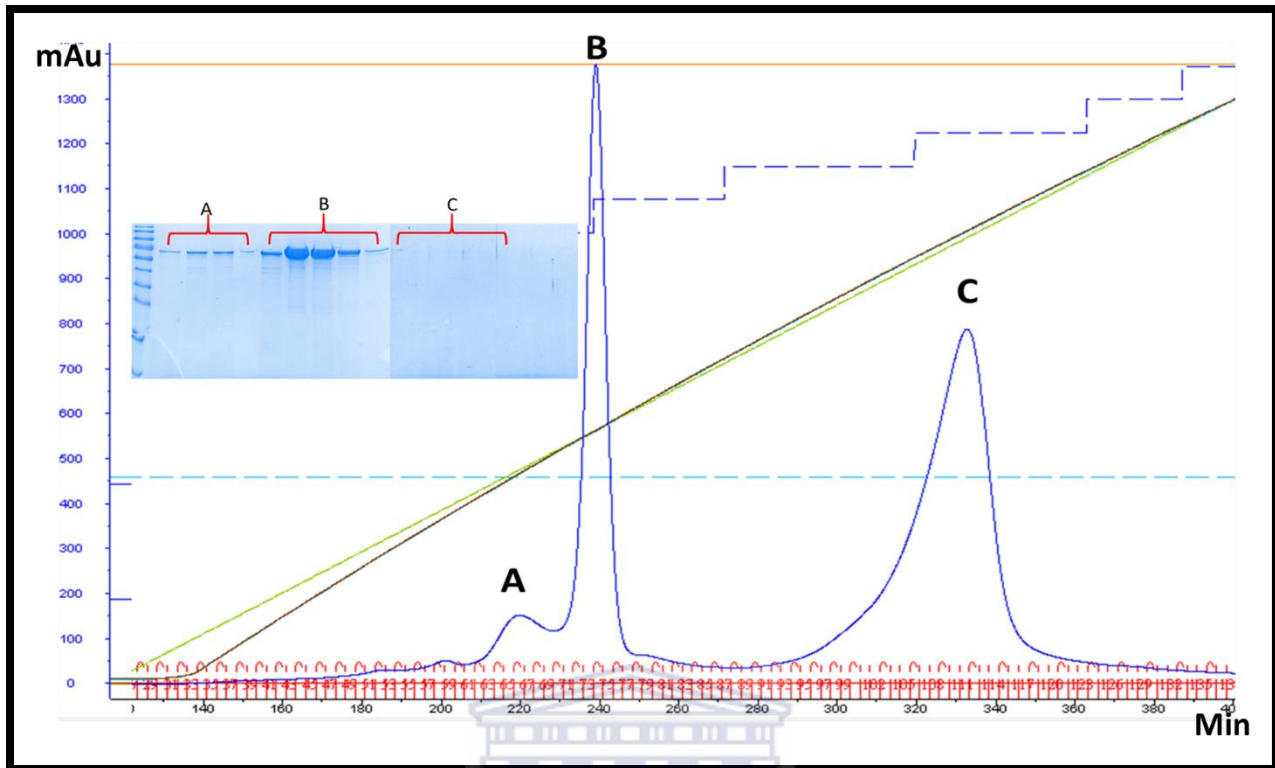


Figure 28: GH9 C1 cellulase AEX chromatogram illustrates a steadily increasing NaCl gradient (theoretical: green, actual: black line) in Tris-HCl pH 8.5. The peaks in the chromatogram are caused by eluting proteins. SDS PAGE analysis indicates proteins in peaks A and B to have the same size, which corresponds to the molecular mass of GH9 C1 cellulase. Fractions corresponding to Peak C did not reveal any proteinaceous bands in SDS PAGE indicating that it is likely to be due to remaining imidazole. Peak heights are in mAu. The chromatogram was produced with PrimeView program (GE Healthcare).

SDS-PAGE indicates that the proteins eluting as peak A and B have the same molecular weight. Both peaks A and B are due to GH9 C1 cellulase but with distinct overall charges. In fact, additional, smaller peaks prior to A indicate that the protein has a number of ionization states at pH 8.5 (protonation states of histidine). Peak C eluting at a high salt concentration (90% buffer B) and absorbing strongly is presumably due to imidazole as SDS-PAGE did not reveal any protein component associated with to this peak. Fractions of peaks A and B were separately pooled and concentrated to final volumes of 0.5 mL using an Amicon Concentrator (Millipore).

Simultaneously the high salt buffer from the elution was replaced by a low salt buffer: 25 mM Tris-HCl pH 8.5 and 50 mM NaCl.

3.1.2. Size Exclusion Chromatography (SEC)

Protein crystallization requires that the protein be as pure and homogenous as possible and SEC served as a final polishing step in the purification process. As SEC separates proteins by size, different oligomeric forms that could interfere with crystallization will be identified and separated. A S200 16/60 (GE Healthcare) column was used for SEC as outlined in section 2.2.3. Figure 29 shows the results for peaks A and B of figure 28 respectively. Interestingly, the sample from peak A of figure 28 results in two peaks on SEC indicating it to be a mixture of two molecular entities (Figure 29 peak A1). However, no protein band is associated with (A1) while the elution volume of peak A2 corresponds to the major peak in Figure 29B and SDS PAGE identifies proteins to be similar in size to the C1 cellulase. The single symmetrical peaks in 29A and B indicates the protein to be homogenous, a prerequisite for crystallization. Note that a small peak corresponding to peak A1 is also visible in figure 29B.

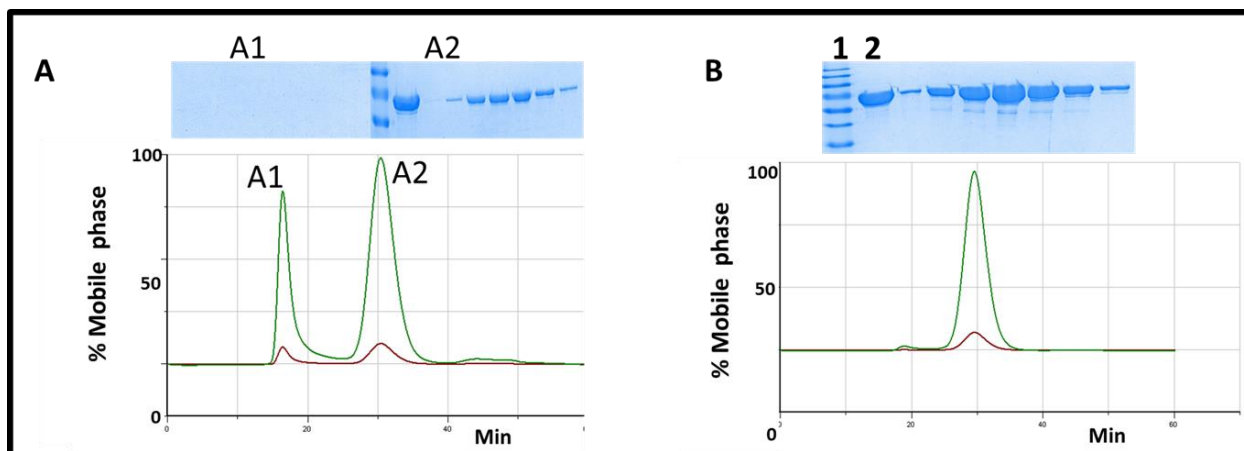


Figure 29: SEC purification of the fractions collected from AEC peaks A and B in figure 28. A): SEC chromatogram of AEC peak A. Two distinct peaks eluted revealing two entities binding with similar affinities to the anion exchange column. Peak A1 eluted in the void volume and did not reveal any protein band on SDS-PAGE gel. The second peak (A2) corresponds to the elution profile in B and SDS PAGE reveals both to be at same MW as GH9 C1 cellulase. B): SEC chromatogram of AEC peak B. The elution profile corresponds to GH9 C1 cellulase - as confirmed by SDS-PAGE. Lanes 1 and 2 contain the PageRuler Prestained Protein Ladder (Thermo Scientific) and the partly purified GH9 C1 cellulase as a control, respectively.

3.2. Protein Crystallization

Analysis of proteins by X-ray crystallography requires the target protein be crystallized. In this study, both sitting drop and hanging-drop vapour diffusion techniques were used to drive the equilibrium of protein-containing drops towards crystal formation. Fractions under the peaks in Figure 29B and A2 were separately pooled, concentrated to 10 mg/mL and 16 mg/mL respectively in 25 mM Tris-HCl pH 8.0 and 50 mM NaCl, and stored at 4°C for crystallization.

Initial crystal screening was conducted with commercial crystallization screens (Section 2.3) and crystal hits were manually optimized with varying buffer component and concentrations for diffraction size crystals.

3.2.1. Initial Screening

The commercial crystallization screens used for initial screening provided over 750 different buffer conditions for crystallization. Crystals were identified in wells 6, 15, 18, 28 and 41 of the Crystal Screen HT (Hampton Research); conditions 61-67, 74 and 77 of PACT Suite (Qiagen); and conditions 6, 18 and 41 of Crystal Screen Lite (Hampton Research). These conditions generally shared a similar precipitant (15-30% w/v polyethylene glycol [PEG] 3350 or 4000), similar salt concentrations (200 mM) and a neutral pH (6.5 to 8.5). These conditions were manually adjusted to optimize crystal morphology and diffraction.

3.2.2. Optimization of Crystallization

The aim of optimizing a particular crystallization condition is to increase the crystal size, to ensure growth of single crystals, to improve the crystals morphology, and to extend the resolution of diffraction. Crystals of GH9 C1 cellulase generally occurred as thin, intergrown plates. Optimization successfully increased the crystal size, though most crystals were still intergrown, indicating growth from a common crystallization nucleus. Crystals appeared within two weeks and increased to larger size and thickness within 24 hours.

Optimization of the lead crystals from the PACT Suite revealed them to be salt crystals. Crystals from Crystal Screen HT (Hampton Research) could not be reproduced except condition 6: 30% (w/v) PEG 4000, 0.1 M Tris-HCl pH 8.5, 0.2 M MgCl. Conditions 6 from Crystal Screen HT and Crystal Screen Lite (Hampton Research) differ only in PEG 4000 concentration with the reduced concentration (15% w/v) improving both crystal morphology and size.

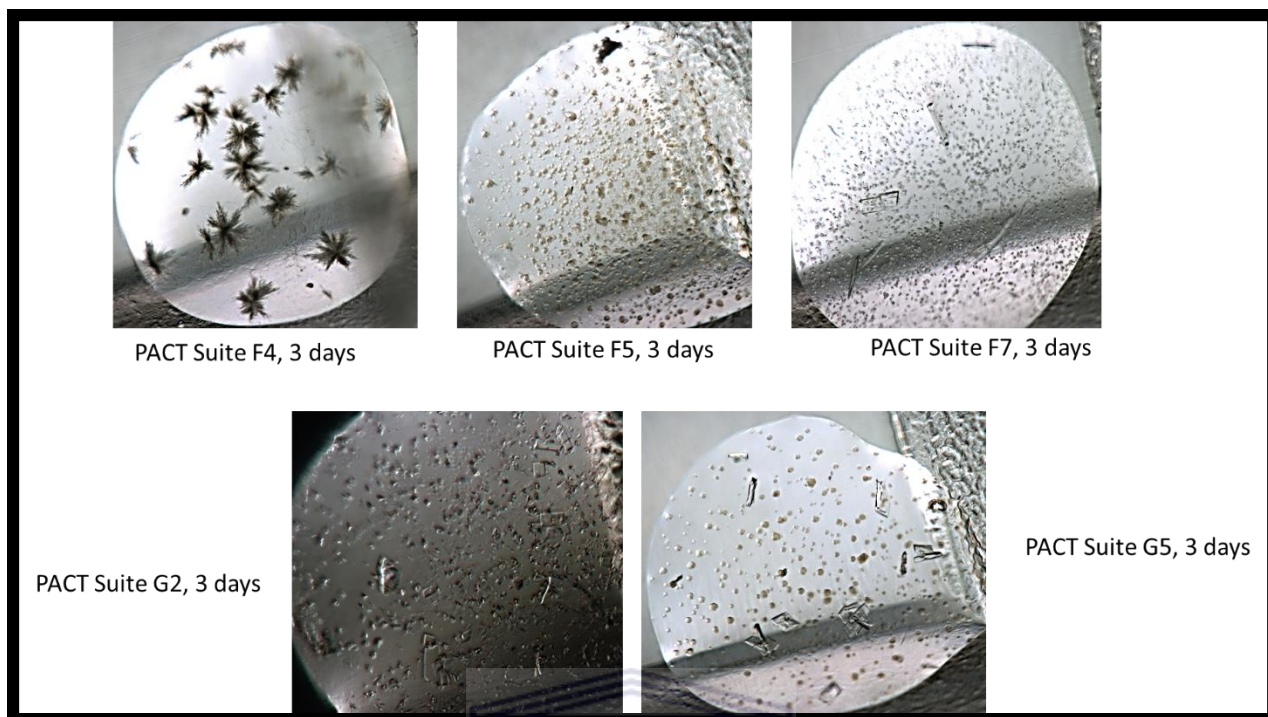


Figure 30: Selected crystallization hits from the PACT suite. F4: 0.1 M Bis-tris propane pH 6.5, 20% (w/v) PEG 3350, 0.2 M KSCN; F5: 0.1 M Bis-tris propane pH 6.5, 20% (w/v) PEG 3350, 0.2 M NaNO₃; F7: 0.1 M Bis-tris propane pH 6.5, 20% (w/v) PEG 3350, 0.2 M Sodium acetate; G2: 0.1 M Bis-tris propane pH 7.5, 20% (w/v) PEG 3350, 0.2 M NaBr; G5: 0.1 M Bis-tris propane pH 7.5, 20% (w/v) PEG 3350, 0.2 M NaNO₃.

Optimized crystals from Figure 30 observed with a light polarizing microscope showed high birefringence, suggesting that the crystals are potentially salt crystal. None of the crystals absorbed bromophenol blue confirming that the crystals are highly constituted of salts, rather than protein.

For Crystal Screen and Crystal Screen Lite Suites, crystals were grown from a protein to reservoir mixture of 2:1. Optimization of nucleation using micro-, streak- and macro-seeding resulted in the crystals shown in Figure 31 and 32).

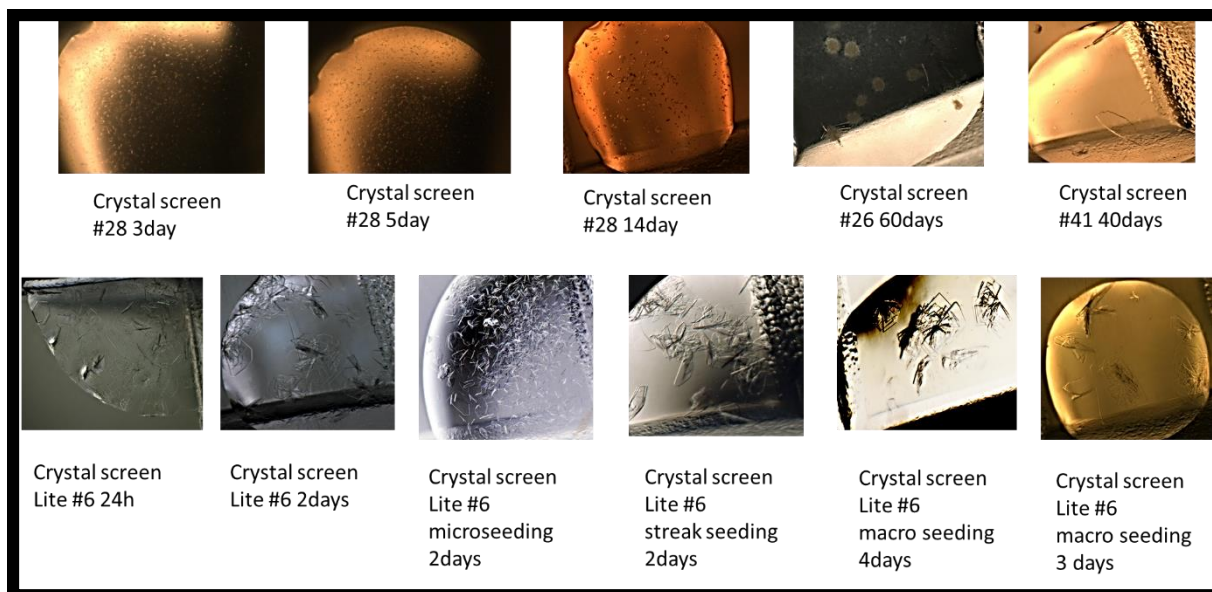


Figure 31: Crystal hits and optimized crystal with the Hampton Crystal Screen reagent.

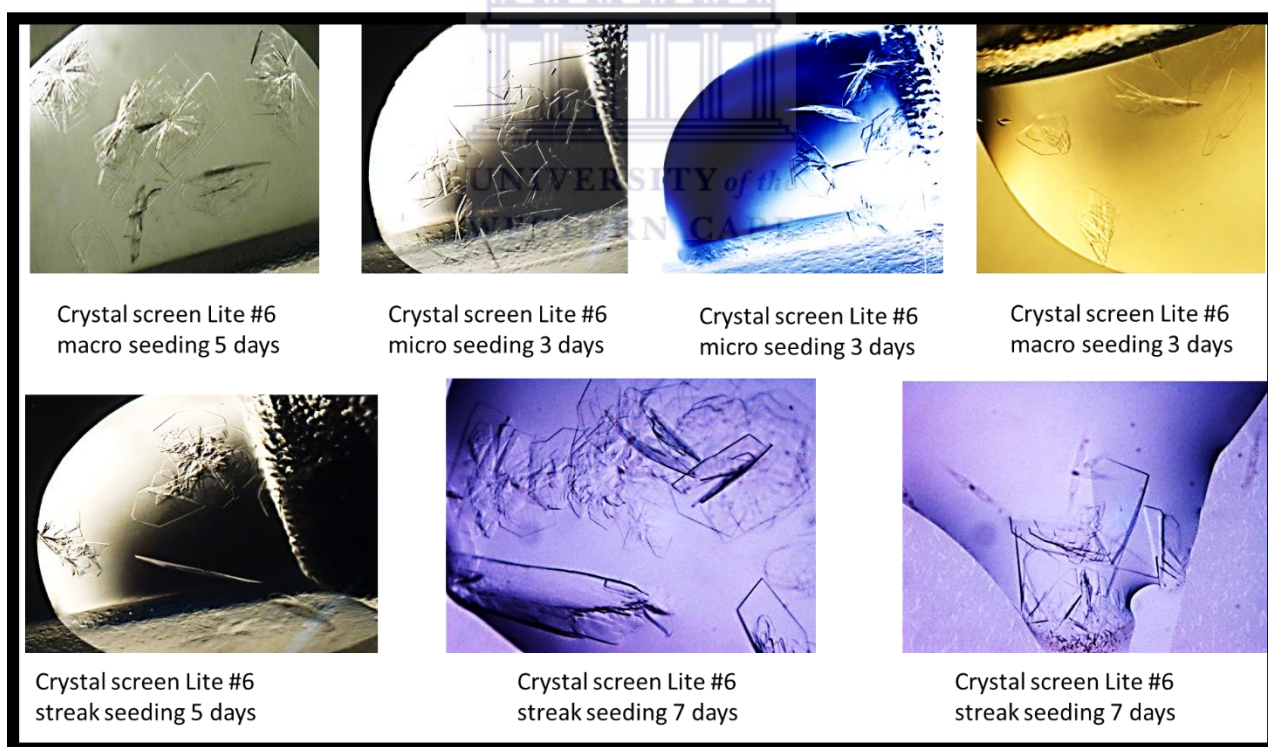


Figure 32: Results of macro-, micro- and streak-seeding using lead condition 6 of the Hampton Crystal Screen Lite.

Figure 33 below demonstrates an interesting observation after a crystallization plate had been left standing in the laboratory at 20°C. After five months, crystals were observed in condition 17 (Crystal Screen Suite). This condition (0.1 M Tris-HCl pH 8.5, 30% w/v PEG 4000, 0.2 M Li₂SO₄) differed only in its salt content (Li₂SO₄) from condition 6 for which crystals had previously been observed (Figure 31 and 32). This indicated monovalent cations to be critical for crystallization leading to further optimization using LiCl, NaCl and KCl. Crystals of similar morphology and size were obtained and used for diffraction experiments.



Figure 33: Crystal observed in a plate 5 months after setup. The three images are different portions of the same drop.

3.2.3. Diffraction Experiments

Nylon cryo-loops were used to transfer single crystals from the crystallization drops shown in figures 31, 32 and 33 to a cryo-protectant created by mixing mother liquor with 20-30% (v/v) glycerol or PEG 400. A cryo-protectant allows crystals to be flash cooled and maintained at 100 K throughout the diffraction experiment to prevent diffusion of free radicals and hence radiation damage. Cryo-protected crystals were flashed cooled in liquid nitrogen and mounted at the point

of intersection of a nitrogen stream at 100 K and the path of an X-rays beam on a diffractometer as described in section 2.3.1. The crystals diffracted X-rays to a moderately low resolution of 4 Å while reflections were quite weak and diffuse (Figure 34). Indexing of diffraction images using HKL3000 identified the most likely Bravais lattice as primitive monoclinic. The identified space group was P2 with unit cell dimensions $a = 74 \text{ \AA}$, $b = 109 \text{ \AA}$, $c = 84 \text{ \AA}$ and $\beta = -114^\circ$ while $\alpha = \gamma = 90^\circ$. Due to high crystal mosaicity of $>2^\circ$, the integrated dataset contained few fully recorded reflections. These images could not be reliably scaled nor reflections merged. As a result, statistics for data completeness and other parameter required to assess the data quality could not be determined. As a stop-gap measure, homology modelling was used to obtain structural information for GH9 C1 cellulase.

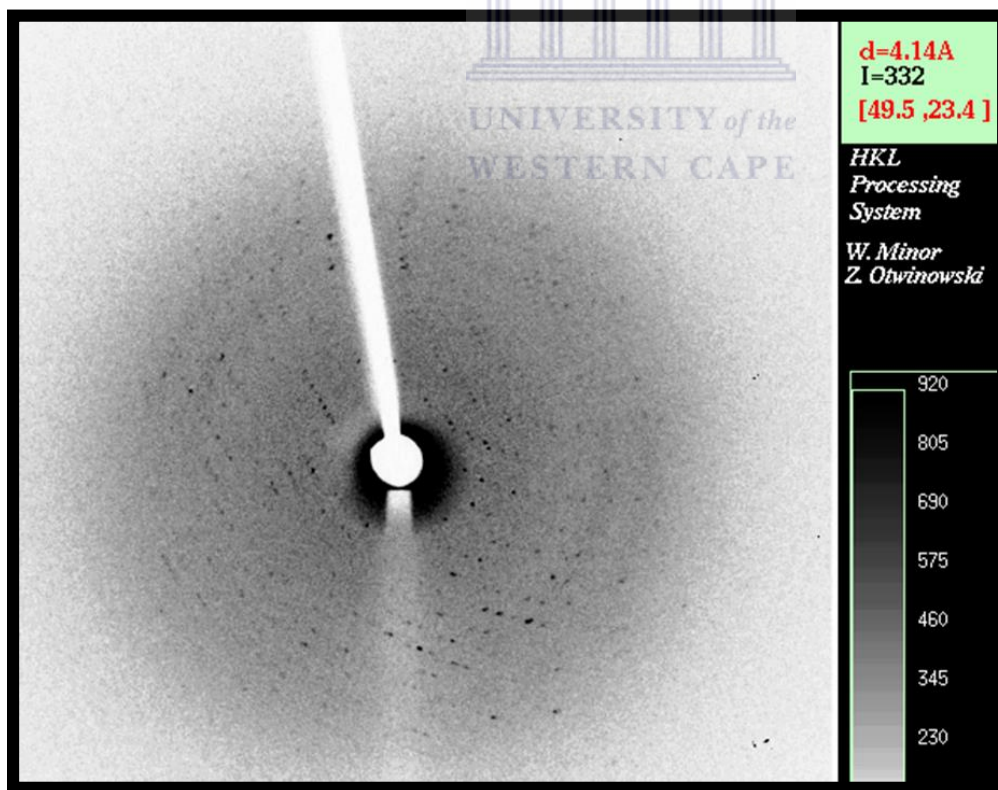


Figure 34: Diffraction pattern for GH9 C1 cellulase crystals. Limit of resolution is $\sim 4 \text{ \AA}$.

3.3. Homology Modelling of GH9 C1 Cellulase Structure

Modelling of protein structures provides an alternative route to obtaining three-dimensional structure information when experimental structures are not available. Homology or comparative modelling, the most reliable of these techniques, generates a model structure for a target protein based on the atomic coordinates of a template molecule sharing a sequence identity equal or larger than 30% to the target. In this project, a BLASTp search of the Protein Data Bank (PDB) identified cellobiohydrolase CbhA from *Clostridium thermocellum* as the most closely related protein (32% sequence identity) to GH9 C1 cellulase. Its crystal structure, PDB ID 1UT9, was thus used to construct a structural model for GH9 C1 cellulase. The amino acid sequences of GH9 C1 cellulase and CbhA were aligned using ClustalW2 and the alignment provided to SWISS-MODEL pipe line to generate a GH9 C1 cellulase model structure (Figure 35). The graphics program PyMol was used to analyse and depict the model for GH9 C1 cellulase.

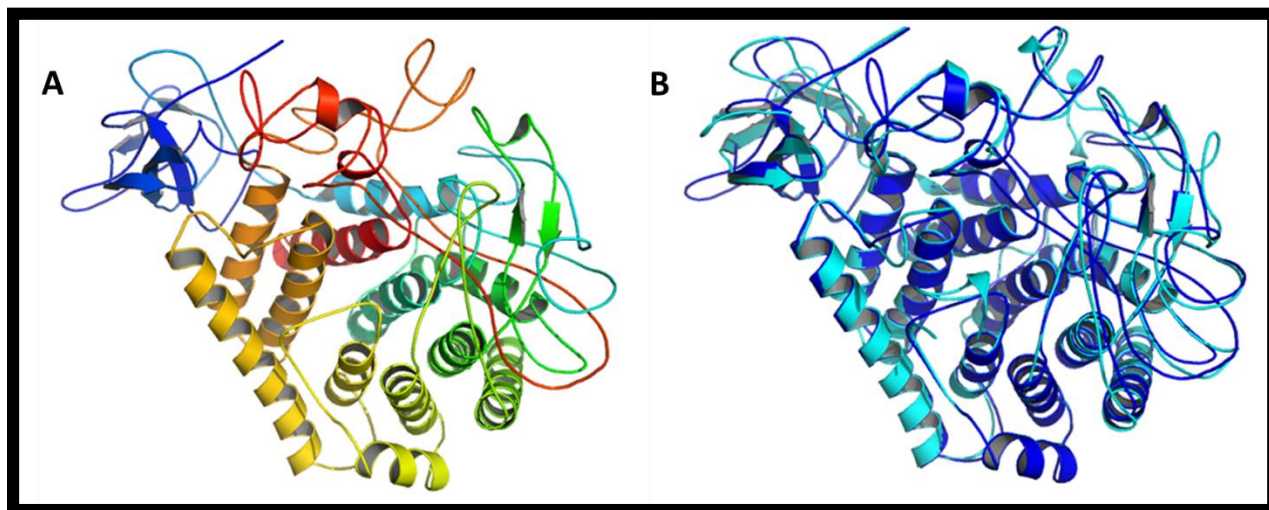
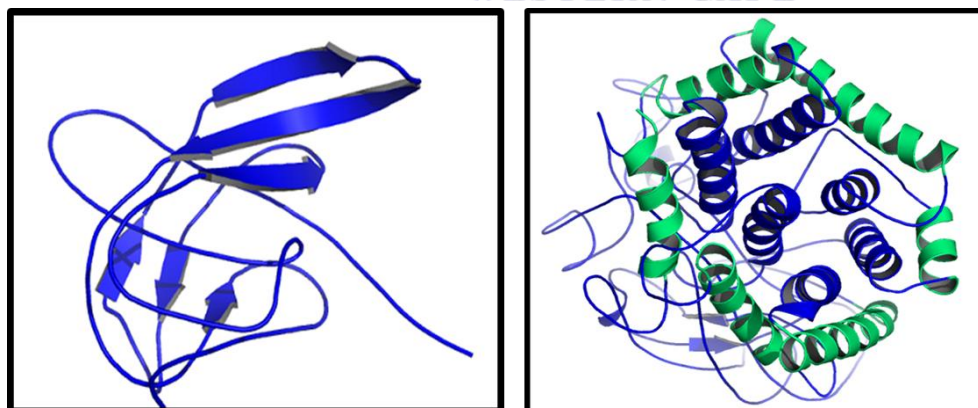


Figure 35: Model structure of GH9 C1 cellulase generated using SWISS-MODEL. A) GH9 C1 cellulase model structure presented as a ribbon diagram using PyMol. Using the rainbow colouring scheme, results in a blue N-terminus and a red C-terminus, with intermediate colours cyan, green, yellow and orange in between. B) The model of GH9 C1 cellulase (blue) superimposed in the crystal structure of CbhA (cyan) (PDB Code: 1UT9).

3.3.1. Overall Model Description

The model structure of the 63 kDa GH9 C1 cellulase reveals the latter to consist of two interconnected domains: A smaller N-terminal Ig-like domain spanning residues 1-126 (blue in Figure 35A) followed by a much larger catalytic domain (residues 127-524, cyan to red). This domain arrangement is typical of the subgroup C of GH9 cellulases (Schubot *et al.*, 2004) and the overall fold of the structure is similar to those of previously described family 9 cellulases (Chauvaux *et al.*, 1995; Pereira *et al.*, 2009; Schubot *et al.*, 2004). The N-terminal Ig-like domain consists of six antiparallel β -strands forming a two layered β -sandwich structurally linked to the catalytic domain. The function of the Ig-like domain is unclear though its deletion or extensive mutation progressively inactivates (Burstein *et al.*, 2009; Kataeva *et al.*, 2004) and limits the range of substrates of the catalytic module. This is similarly true for carbohydrate binding domains of related enzymes (Ravachol *et al.*, 2014).



Ig-like Module

Catalytic module

Figure 36: The domain structure of GH9 C1 cellulase. Left: Ig-like domain of GH9 C1 cellulase showing the six β -strand sandwich structure. Right: Catalytic domain showing the inner (blue) and outer (green) α -helices.

The fold of the GH9 C1 cellulase catalytic domain is similar to those of other GH9 family enzymes. The structure consists of twelve α -helices and two antiparallel β -strands. The twelve α -helices form a typical $(\alpha/\alpha)_6$ -barrel (Schubot *et al.*, 2004; Pereira *et al.*, 2009).

3.3.2. Metal Ion Binding

All GH9 cellulases bind metal ions (Chauvaux *et al.*, 1995). The metal-ion binding sites are limited to the catalytic domain and never occur in the Ig-like domain. Metal ion binding presumably enhances the cellulase thermal stability via intra-domain interactions (Juy *et al.*, 1992; Chauvaux *et al.*, 1995) and to stabilize the active conformation.

The endoglucanase CelD of *Clostridium thermocellum* binds three Ca^{2+} and one Zn^{2+} (Chitarra *et al.*, 1995); the endoglucanase Cel9A of *Alicyclobacillus acidocaldarius* binds two Ca^{2+} and one Zn^{2+} (Pereira *et al.* 2009), while the cellobiohydrolase CbhA of *Clostridium thermocellum* binds two Ca^{2+} (Schubot *et al.*, 2004). Ca^{2+} -coordinating residues of binding sites 1 and 2 (Figure 37) are conserved in GH9 cellulases. Ca^{2+} -binding site 3, involved in stabilizing the active site (Figure 37), is not conserved in CbhA (Schubot *et al.*, 2004).

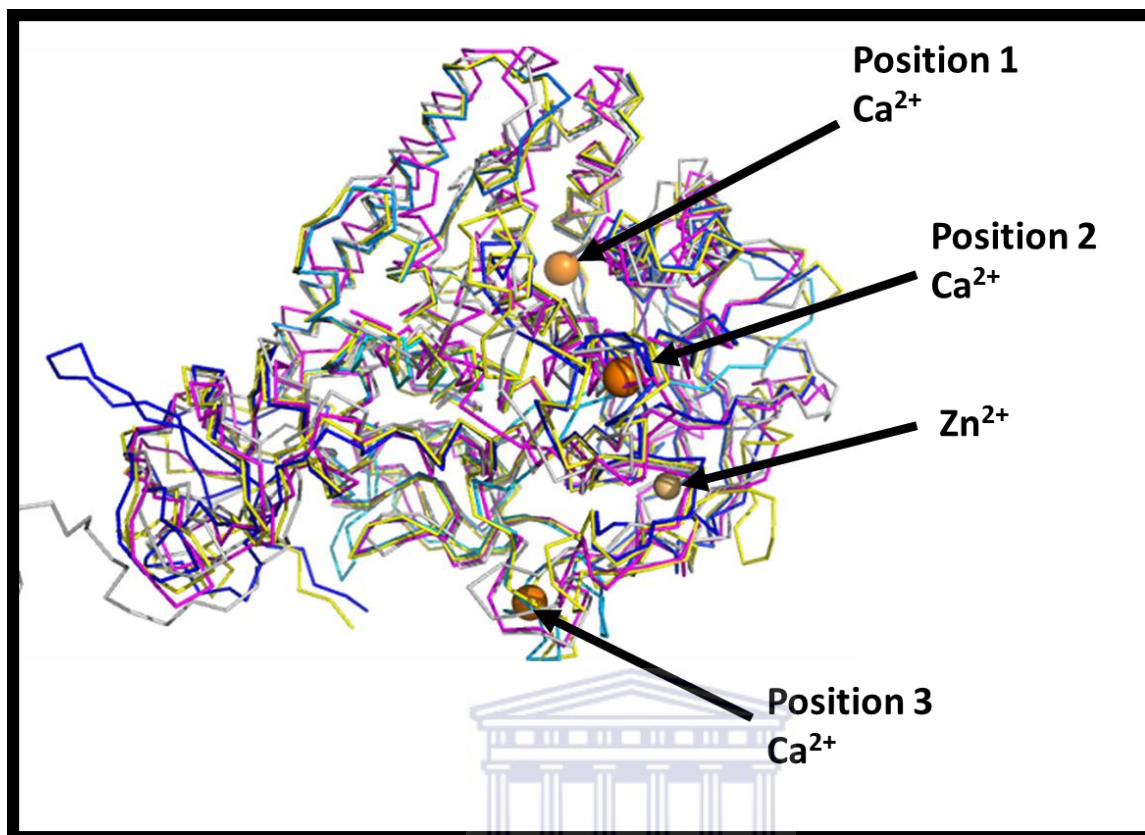


Figure 37: Superimposed structures of GH9 cellulases showing metal ion binding sites. Position 1, 2 and 3 are Ca^{2+} binding site (orange spheres). The light orange sphere indicates the Zn^{2+} binding site conserved in some GH9 cellulases. Blue: GH9 C1 cellulase, magenta: endoglucanase Cel9A of *Alicyclobacillus acidocaldarius* (3EZ8_Cel9A); grey: endoglucanase CelD of *Clostridium thermocellum* (1CLC_CelD); yellow: cellobiohydrolase CbhA of *Clostridium thermocellum* with substrate bound (1RQ5_CbhA); cyan: Cellobiohydrolase CbhA from *Clostridium thermocellum* without substrate (1UT9_CbhA).

3.3.2.1. Zn^{2+} -Binding Sites

To investigate Zn^{2+} binding by GH9 C1 cellulase, the crystal structures of CelA (PDB code: 3EZ8) and Endoglucanase D (PDB code: 1CLC) known to bind Zn^{2+} were superimposed on the modelled structure of GH9 C1 cellulase. The results showed that the amino acids involved in metal binding are conserved in the two crystal structures: 2 histidines and 2 cysteines each (magenta and light blue in Figure 38). These residues are replaced by tyrosine, histidine, isoleucine and cysteine in GH9 C1 cellulase indicating that the latter does not bind Zn^{2+} (blue in Figure 38).

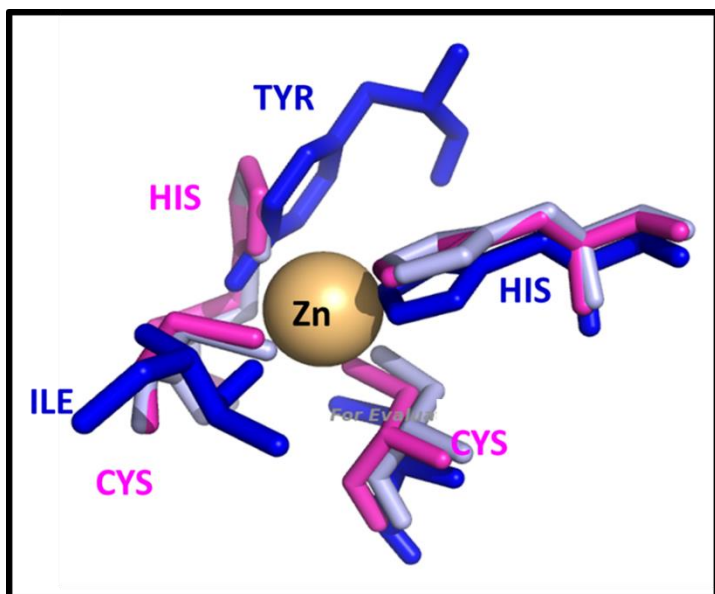
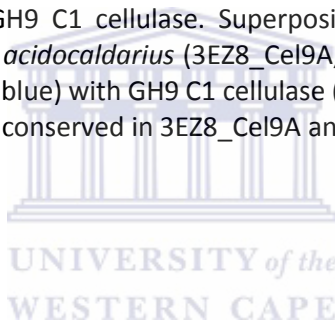


Figure 38: Lack of Zn^{2+} binding in GH9 C1 cellulase. Superposition of Zn^{2+} coordinating residues of endoglunase Cel9A of *Alicyclobacillus acidocaldarius* (3EZ8_Cel9A, magenta) and endoglucanase CelD of *Clostridium thermocellum* (1CLC, light blue) with GH9 C1 cellulase (blue). The superimposition shows that the residues for Zn^{2+} coordination are conserved in 3EZ8_Cel9A and 1CLC but not in GH9 C1 cellulase.

3.3.2.2. Ca^{2+} -Binding Sites



Calcium ions are generally considered hard metal ions. As a result they prefer similarly hard oxygen ligands most often from negatively charged glutamate and aspartate side-chains or alternatively from serine and threonine side chains and/or main-chain carbonyl atoms. The coordination sphere of Ca^{2+} is classically octahedral with Ca^{2+} -O distances around 2.4 to 2.5 Å. Within proteins this coordination sphere is frequently distorted. The cytoplasmic concentration of Ca^{2+} is maintained at very low levels by dedicated Ca^{2+} pumps. Cytoplasmic proteins therefore generally do not bin Ca^{2+} unless involved in Ca^{2+} -detection and -response. Extracellular Ca^{2+} -concentrations, by contrast, are much higher and many proteins in this environment use Ca^{2+} -sites for structural stabilization. The catalytic domains of GH9 cellulases share up to three Ca^{2+} binding sites with sites 1 and 2 more highly conserved than site 3.

Site 1 is near the enzyme active site indicating a stabilizing role for the catalytic domain. In Cel9A four residues are involved in coordinating Ca^{2+} . These are two aspartate residues, one alanine and one glutamate. In 1CLC the residues involved are two aspartate, one threonine and one serine residues. Analysis of Ca^{2+} binding site 1 shows that the above are replaced by two aspartate, one threonine and one serine residue implying that this site is indeed conserved in GH9 C1 cellulase (Figure 39A). The residues aspartate and glutamate coordinates Ca^{2+} through their side chain carbonyl oxygen whereas serine and threonine coordinate through the hydroxyl groups. Alanine, however, coordinates Ca^{2+} through its main chain carbonyl oxygen.

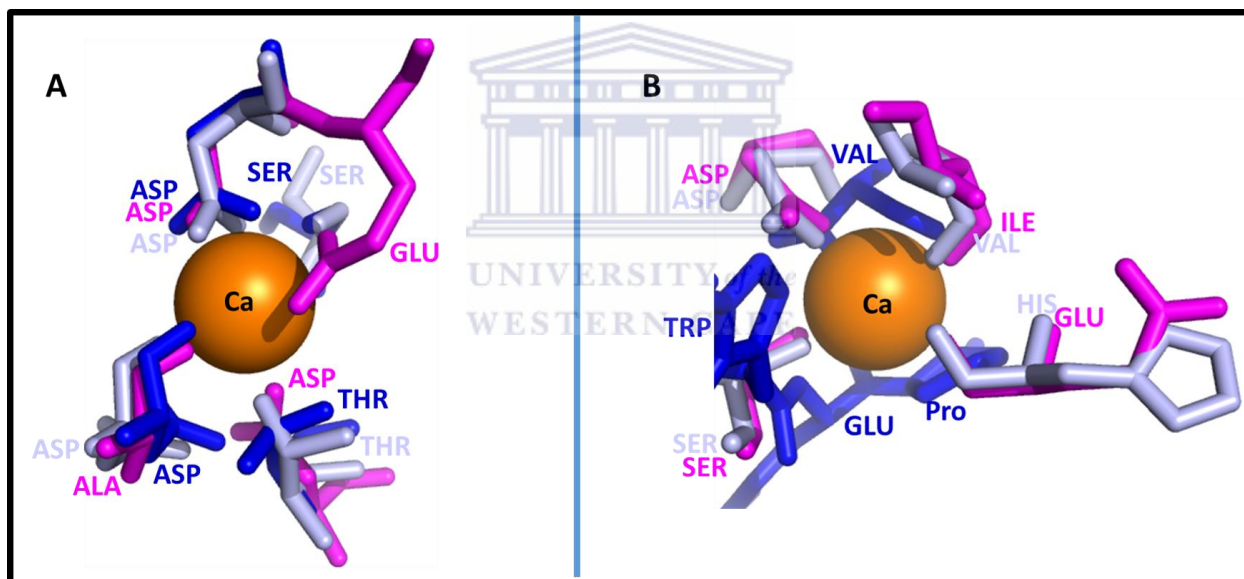


Figure 39: Ca^{2+} binding sites 1 (B) and 2 (A) in GH9 C1 cellulase. Ca^{2+} -coordinating residues of 3EZ8_Cel9A (magenta), 1CLC (light blue) and GH9 C1 cellulase (blue). The superposition shows that Ca^{2+} -binding residues of site 2 (A) are conserved, while those of site 2 (B) are not.

Comparing the possible Ca^{2+} -binding site 2 of GH9 C1 cellulase to equivalent sites in Cel9A and 1CLC shows aspartate, isoleucine, glutamate and serine are involved in Ca^{2+} -binding in Cel9A, and aspartate, valine, histidine and serine in 1CLC. These residues are replaced by valine, proline,

glutamate and tryptophan in GH9 C1 cellulase (Figure 39B). These residues do not allow for Ca²⁺-binding, implying site 1 not to be conserved in GH9 C1 cellulase.

To analyse Ca²⁺-binding site 3 in GH9 C1 cellulase, the site was compared to that of CelD (1CLC). Ca²⁺-coordination here is distorted octahedral with a water molecule at one vertex and protein groups providing the five remaining ligands (Chauvaux *et al.*, 2004). The protein ligands include two main chain carbonyls as well as one asparagine and two aspartic acid residue side chains. These residues are conserved in GH9 C1, indicating site 3 is conserved in GH9 C1 cellulase (Figure 40).

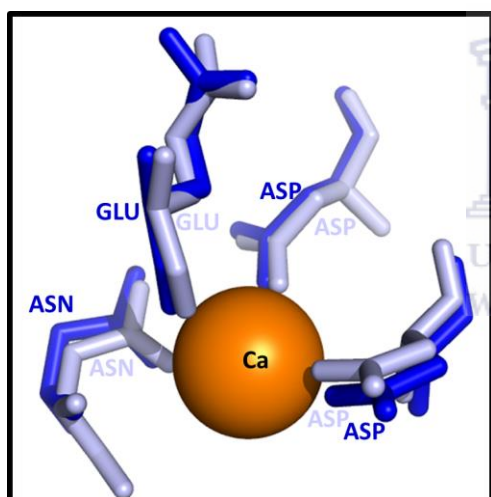


Figure 40: Binding of Ca²⁺ by GH9 C1 cellulase at position 3. Conserved residues between GH9 C1 cellulase (Blue) and 1CLC (Light blue) at the Ca²⁺ binding site of 1CLC. The main calcium ion coordinating residues are conserved in GH9 C1 cellulase, indicating that this site potentially binds calcium ions.

3.3.3. The Active Site Architecture

The active site of GH9 C1 cellulase is located between the loops at the N-terminus of the catalytic module and near the second calcium ion binding site. The active site is located in an open cleft which binds four to six glucose units of cellulose. Figure 41 shows a surface representation of GH9 C1 cellulase with a substrate molecule from the crystal structure 1RQ5 modelled into the active site.

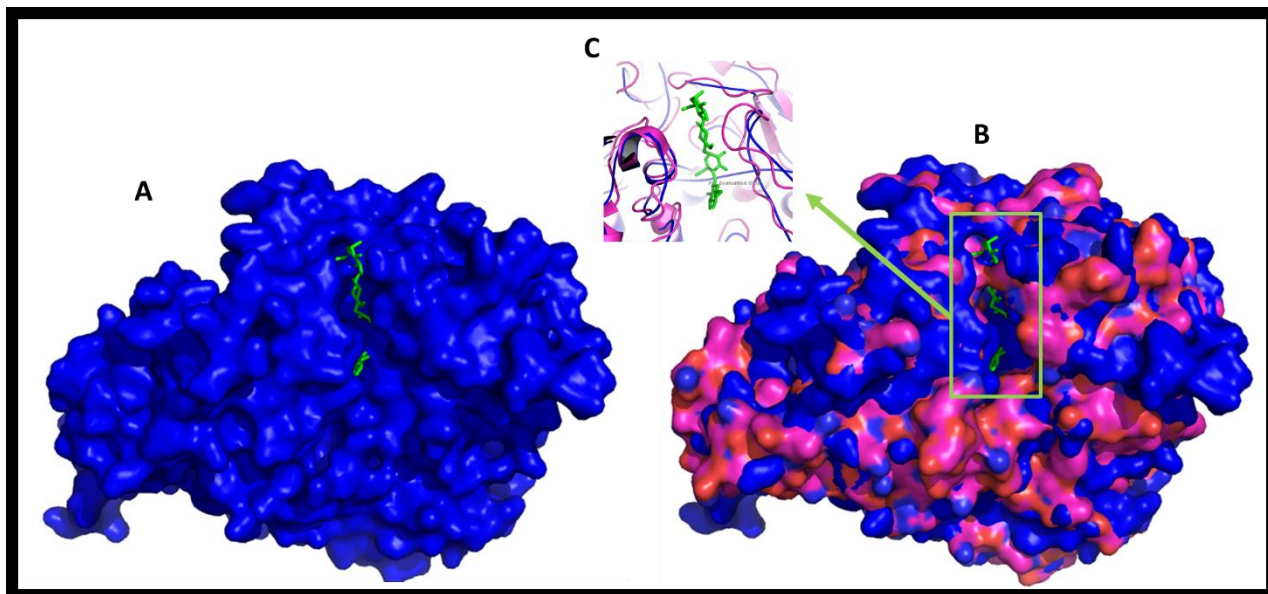


Figure 41: Surface view of GH9 C1 cellulase showing active site cleft. A) Surface view of GH9 C1 cellulase model structure (Blue) with a substrate molecule (Green) modelled into its active site. B) Surface view of GH9 C1 cellulase (Blue) superimposed on 1RQ5 (Magenta) and showing the active site containing a substrate molecule (Green). C): Cartoon image of the active site of B showing the secondary structure elements around the active site cleft.

A detailed analysis of the active site residues suggest that the conserved amino acids Phe266, Tyr344, Leu396, Trp447 and His506 mediate sugar binding through stacking van der Waals interactions with the glucose rings (Figure 42) as frequently observed for protein-sugar binding (Vyas, 1991). The interactions are flexible to allow sliding of the substrate across the active site. In addition, intricate hydrogen bonding network between cellulose and amino acid side chains are observed including Asp188, Asp191 and Arg508.

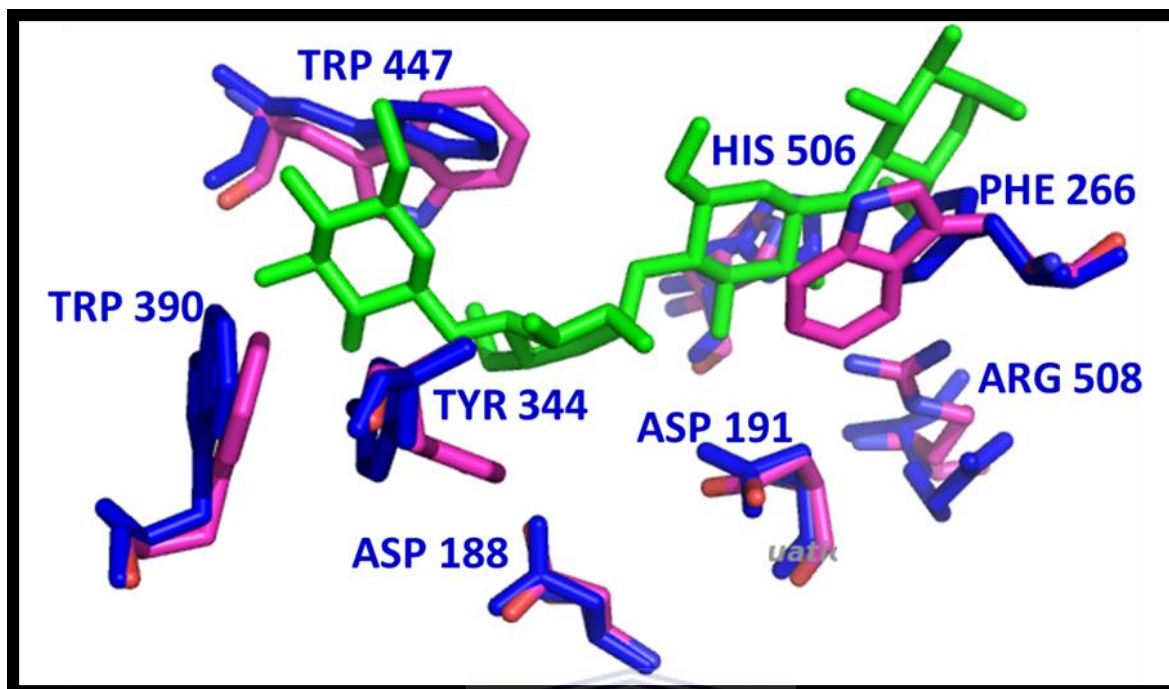
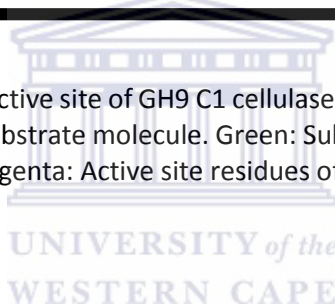


Figure 42: Conserved residues at the active site of GH9 C1 cellulase. The residues are involved in hydrogen bonding and base stacking with the substrate molecule. Green: Substrate molecule; Blue: Residues of the GH9 C1 cellulase model structure; Magenta: Active site residues of 1RQ5.

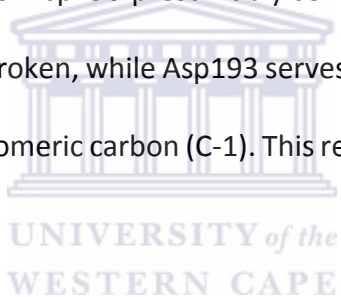


3.3.4. Mechanism of Action

Cellulases catalyse the hydrolysis of β -1,4-glycosidic bonds through general acid catalysis (Sinnott, 1990) with two acidic residues (glutamate or aspartate) being essential (Davies & Henrissat, 1995). Hydrolysis further retains or inverts the configuration of the substrate. To retain the configuration, one of the two catalytic residues nucleophilically attacks the substrate forming a covalent glycosyl-enzyme intermediate. The second catalytic residue, a general acid/base, then first protonates the leaving group following the formation of the covalent intermediate and then activates the incoming nucleophile, often a water molecule. The glycosyl-enzyme intermediate is both formed and hydrolysed via an oxocarbenium ion-like transition state.

In the inversion mechanism, one of the catalytic residues, a proton donor, protonates the glycosidic oxygen and promotes the leaving group to depart. The other catalytic residue acts as a general base, activating nucleophilic water by deprotonating it. The result is an oxocarbenium-ion-like transition state with electron density distributed over eight atoms resulting in three bonds being broken and formed in a concerted fashion (Withers, 2001).

GH9 C1 cellulase is an inverting enzyme. Asp190 and Asp193 constitute the conserved catalytic residues. Asp193 deprotonates a water molecule, which then nucleophilically attacks the anomeric carbon of the targeted glucose subunit. Another conserved residue Asn609 stabilizes the transition state during catalysis. Asp190 presumably serves as proton donor for the oxygen (O-4) in the glycosidic bond to be broken, while Asp193 serves as the catalytic base activating the nucleophilic water to attack the anomeric carbon (C-1). This results in the inversion of the original stereochemistry of the product.



4. Conclusion and Outlook

This study was aimed at structurally analysing the enzyme GH9 C1 cellulase obtained from a hot compost metagenomic library. The protein was readily produced in *E. coli* and purified using chromatographic techniques. The homogeneously pure protein was concentrated and used for crystallization. Crystallization conditions for GH9 C1 cellulase were established and refined to improve crystal size and morphology. The crystals, however, diffracted X-rays to a mere 4 Å resolution. Recorded diffraction data could not be used to solve the crystal structure of GH9 C1 cellulase. Instead the structure of GH9 C1 cellulase was modelled on the experimental structure of CbhA (PDB ID: 1UT9) from *Clostridium thermocellum* by homology modelling and analysed using the graphics program PyMol. Analysis revealed a similar fold to other GH9 family members included two linked domains. Two Ca²⁺-binding sites were found to be conserved in GH9 C1 cellulase and the amino acids for substrate binding and catalysis were identified.

Owing to time constraints, further optimisation of crystallization conditions for high resolution diffraction quality crystals could not be pursued. Future studies on this protein will involve the further improvement of the crystallization conditions for better diffraction quality crystals to allow the structure of GH9 C1 cellulase to be determined and refined experimentally.

References
Part I
Abella, M., Campoy, S., Erill, I., Rojo, F., & Barbé, J. (2007). Cohabitation of two different <i>lexA</i> regulons in <i>Pseudomonas putida</i> . <i>Journal of Bacteriology</i> , <i>189</i> , 8855–8862.
Abella, M., Erill, I., Jara, M., Mazón, G., Campoy, S., & Barbé, J. (2004). Widespread distribution of a <i>lexA</i> -regulated DNA damage-inducible multiple gene cassette in the Proteobacteria phylum. <i>Molecular Microbiology</i> , <i>54</i> , 212–222.
Afonso, J. P., Chintakayala, K., Suwannachart, C., Sedelnikova, S., Giles, K., Hoyes, J. B., Oldham, N. J. (2013). Insights into the structure and assembly of the <i>Bacillus subtilis</i> clamp-loader complex and its interaction with the replicative helicase. <i>Nucleic Acids Research</i> , <i>41</i> , 5115–5126.
Ahmad, S. (2011). Pathogenesis, Immunology, and Diagnosis of Latent <i>Mycobacterium tuberculosis</i> Infection. <i>Clinical and Developmental Immunology</i> .
Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., & Walter, P. (2002). The Initiation and Completion of DNA Replication in Chromosomes. Text. Retrieved March 18, 2014, from http://www.ncbi.nlm.nih.gov/books/NBK26826/ .
Allert, M., Cox, J. C., & Hellinga, H. W. (2010). Multifactorial determinants of protein expression in prokaryotic open reading frames. <i>Journal of Molecular Biology</i> , <i>402</i> , 905–918.
Angov, E. (2011). Codon usage: nature's roadmap to expression and folding of proteins. <i>Biotechnology Journal</i> , <i>6</i> , 650–659.
Aravind, L., Anand, S., & Iyer, L. M. (2013). Novel autoproteolytic and DNA-damage sensing components in the bacterial SOS response and oxidized methylcytosine-induced eukaryotic DNA demethylation systems. <i>Biology Direct</i> , <i>8</i> , 20.
Arbing, M. A., Chan, S., Harris, L., Kuo, E., Zhou, T. T., Ahn, C. J., ... Eisenberg, D. (2013). Heterologous Expression of Mycobacterial Esx Complexes in <i>Escherichia coli</i> for Structural Studies Is Facilitated by the Use of Maltose Binding Protein Fusions. <i>PLoS ONE</i> , <i>8</i> , e81753.
Artimo, P., Jonnalagedda, M., Arnold, K., Baratin, D., Csardi, G., de Castro, E., ... Stockinger, H. (2012). ExPASy: SIB bioinformatics resource portal. <i>Nucleic Acids Research</i> , <i>40</i> , W597–W603.
Baca, A. M., & Hol, W. G. (2000). Overcoming codon bias: a method for high-level overexpression of Plasmodium and other AT-rich parasite genes in <i>Escherichia coli</i> . <i>International Journal for Parasitology</i> , <i>30</i> , 113–118.
Barnes, D. S. (2000). Historical perspectives on the etiology of tuberculosis. <i>Microbes and Infection / Institut Pasteur</i> , <i>2</i> , 431–440.
Barry, C. E., & Blanchard, J. S. (2010). The Chemical Biology of New Drugs in Development for Tuberculosis. <i>Current Opinion in Chemical Biology</i> , <i>14</i> , 456–466.
Bedeir, S. A. (2004). Tuberculosis in Ancient Egypt. (In) M. M. Madkour (Ed.), <i>Tuberculosis</i> (pp. 3–13). Springer Berlin Heidelberg. Retrieved from http://link.springer.com/chapter/10.1007/978-3-642-18937-1_1 .

Beuning, P. J., Simon, S. M., Godoy, V. G., Jarosz, D. F., & Walker, G. C. (2006). Characterization of <i>Escherichia coli</i> translesion synthesis polymerases and their accessory factors. <i>Methods in Enzymology</i> , 408, 318–340.
Bordbar, A., Lewis, N. E., Schellenberger, J., Palsson, B. Ø., & Jamshidi, N. (2010). Insight into human alveolar macrophage and <i>M. tuberculosis</i> interactions via metabolic reconstructions. <i>Molecular Systems Biology</i> , 6, 422.
Boshoff, H. I. M., Myers, T. G., Copp, B. R., McNeil, M. R., Wilson, M. A., & Barry, C. E., 3rd. (2004). The transcriptional responses of <i>Mycobacterium tuberculosis</i> to inhibitors of metabolism: novel insights into drug mechanisms of action. <i>The Journal of Biological Chemistry</i> , 279, 40174–40184. .
Boshoff, H. I. M., Reed, M. B., Barry, C. E., 3rd, & Mizrahi, V. (2003). DnaE2 polymerase contributes to in vivo survival and the emergence of drug resistance in <i>Mycobacterium tuberculosis</i> . <i>Cell</i> , 113, 183–193.
Braithwaite, D. K., & Ito, J. (1993). Compilation, alignment, and phylogenetic relationships of DNA polymerases. <i>Nucleic Acids Research</i> , 21, 787–802.
Bruck, I., Georgescu, R. E., & O'Donnell, M. (2005). Conserved Interactions in the <i>Staphylococcus aureus</i> DNA PolC Chromosome Replication Machine. <i>Journal of Biological Chemistry</i> , 280, 18152–18162.
Bruck, I., Goodman, M. F., & O'Donnell, M. (2003). The essential C family DnaE polymerase is error-prone and efficient at lesion bypass. <i>The Journal of Biological Chemistry</i> , 278, 44361–44368.
Burhans, W. C., Weinberger, M., Marchetti, M. A., Ramachandran, L., D'Urso, G., & Huberman, J. A. (2003). Apoptosis-like yeast cell death in response to DNA damage and replication defects. <i>Mutation Research</i> , 532, 227–243.
Butala, M., Podlesek, Z., & Žgur-Bertok, D. (2008). The SOS response affects thermoregulation of colicin K synthesis. <i>FEMS Microbiology Letters</i> , 283, 104–111.
Campoy, S., Salvador, N., Cortes, P., Erill, I., & Barbe, J. (2005). Expression of Canonical SOS Genes Is Not under LexA Repression in <i>Bdellovibrio bacteriovorus</i> . <i>Journal of Bacteriology</i> , 187, 5367–5375.
Chan, J., Fan, X. D., Hunter, S. W., Brennan, P. J., & Bloom, B. R. (1991). Lipoarabinomannan, a possible virulence factor involved in persistence of <i>Mycobacterium tuberculosis</i> within macrophages. <i>Infection and Immunity</i> , 59, 1755–1761.
Chandani, S., Jacobs, C., & Loechler, E. L. (2010). Architecture of Y-Family DNA Polymerases Relevant to Translesion DNA Synthesis as Revealed in Structural and Molecular Modeling Studies. <i>Journal of Nucleic Acids</i> .
Chim, N., Habel, J. E., Johnston, J. M., Krieger, I., Miailau, L., Sankaranarayanan, R., ... Goulding, C. W. (2011). The TB Structural Genomics Consortium: a decade of progress. <i>Tuberculosis</i> , 91, 155–172.
Christendat, D., Yee, A., Dharamsi, A., Kluger, Y., Savchenko, A., Cort, J. R., ... Arrowsmith, C. H. (2000). Structural proteomics of an archaeon. <i>Nature Structural & Molecular Biology</i> , 7, 903–909.

<p>Cole, S. T., Brosch, R., Parkhill, J., Garnier, T., Churcher, C., Harris, D., ... Barrell, B. G. (1998). Deciphering the biology of <i>Mycobacterium tuberculosis</i> from the complete genome sequence. <i>Nature</i>, <i>393</i>, 537–544.</p>
<p>Crowley, D. J., & Courcelle, J. (2002). Answering the Call: Coping with DNA Damage at the Most Inopportune Time. <i>Journal of Biomedicine and Biotechnology</i>, <i>2</i>, 66–74.</p>
<p>Da Rocha, R. P., de Miranda Paquola, A. C., do Valle Marques, M., Menck, C. F. M., & Galhardo, R. S. (2008). Characterization of the SOS Regulon of <i>Caulobacter crescentus</i>. <i>Journal of Bacteriology</i>, <i>190</i>, 1209–1218.</p>
<p>Daniel, T. M. (2006). The history of tuberculosis. <i>Respiratory Medicine</i>, <i>100</i>, 1862–1870.</p>
<p>Datta, S., Krishna, R., Ganesh, N., Chandra, N. R., Muniyappa, K., & Vijayan, M. (2003). Crystal Structures of <i>Mycobacterium smegmatis</i> RecA and Its Nucleotide Complexes. <i>Journal of Bacteriology</i>, <i>185</i>, 4280–4284.</p>
<p>Davies, P. D. O. (2001). Drug-resistant tuberculosis. <i>Journal of the Royal Society of Medicine</i>, <i>94</i>, 261–263.</p>
<p>Davis, E. O., Dullaghan, E. M., & Rand, L. (2002). Definition of the Mycobacterial SOS Box and Use To Identify LexA-Regulated Genes in <i>Mycobacterium tuberculosis</i>. <i>Journal of Bacteriology</i>, <i>184</i>, 3287–3295.</p>
<p>Deb, C., Lee, C.-M., Dubey, V. S., Daniel, J., Abomoelak, B., Sirakova, T. D., ... Kolattukudy, P. E. (2009). A Novel In Vitro Multiple-Stress Dormancy Model for <i>Mycobacterium tuberculosis</i> Generates a Lipid-Loaded, Drug-Tolerant, Dormant Pathogen. <i>PLoS ONE</i>, <i>4</i>, e6077.</p>
<p>Dervyn, E., Suski, C., Daniel, R., Bruand, C., Chapuis, J., Errington, J., ... Ehrlich, S. D. (2001). Two essential DNA polymerases at the bacterial replication fork. <i>Science</i>, <i>294</i>, 1716–1719.</p>
<p>Diaz, A. A., Tomba, E., Lennarson, R., Richard, R., Bagajewicz, M. J., & Harrison, R. G. (2010). Prediction of protein solubility in <i>Escherichia coli</i> using logistic regression. <i>Biotechnology and Bioengineering</i>, <i>105</i>, 374–383.</p>
<p>Donoghue, H. D., Spigelman, M., Greenblatt, C. L., Lev-Maor, G., Bar-Gal, G. K., Matheson, C., ... Zink, A. R. (2004). Tuberculosis: from prehistory to Robert Koch, as revealed by ancient DNA. <i>The Lancet Infectious Diseases</i>, <i>4</i>, 584–592.</p>
<p>Downey, C. D., & McHenry, C. S. (2010). Chaperoning of a replicative polymerase onto a newly assembled DNA-bound sliding clamp by the clamp loader. <i>Molecular Cell</i>, <i>37</i>, 481–491.</p>
<p>Dulermo, R., Fochesato, S., Blanchard, L., & de Groot, A. (2009). Mutagenic lesion bypass and two functionally different RecA proteins in <i>Deinococcus deserti</i>. <i>Molecular Microbiology</i>, <i>74</i>, 194–208.</p>
<p>Dye, C., Scheele, S., Dolin, P., Pathania, V., & Raviglione, M. C. (1999). Consensus statement. Global burden of tuberculosis: estimated incidence, prevalence, and mortality by country. WHO Global Surveillance and Monitoring Project. <i>JAMA: The Journal of the American Medical Association</i>, <i>282</i>, 677–686.</p>
<p>Ehrt, S., & Schnappinger, D. (2009). Mycobacterial survival strategies in the phagosome: defence against host stresses. <i>Cellular Microbiology</i>, <i>11</i>, 1170–1178.</p>

El-Najjar, M., Al-Shiyab, A., & Al-Sarie, I. (1996). Cases of tuberculosis at 'Ain Ghazal, Jordan. <i>Paléorient</i> , 22, 123–128.
Erill, I., Campoy, S., & Barbé, J. (2007). Aeons of distress: an evolutionary perspective on the bacterial SOS response. <i>FEMS Microbiology Reviews</i> , 31, 637–656.
Erill, I., Campoy, S., Mazon, G., & Barbé, J. (2006). Dispersal and regulation of an adaptive mutagenesis cassette in the bacteria domain. <i>Nucleic Acids Research</i> , 34, 66–77.
Fay, P. J., Johanson, K. O., McHenry, C. S., & Bambara, R. A. (1981). Size classes of products synthesized processively by DNA polymerase III and DNA polymerase III holoenzyme of <i>Escherichia coli</i> . <i>The Journal of Biological Chemistry</i> , 256, 976–983.
Fedyunin, I., Lehnhardt, L., Böhmer, N., Kaufmanna, P., Zhanga, G., Ignatova, Z. (2012). tRNA concentration fine tunes protein solubility. <i>FEBS Letters</i> , 586, 3336–40.
Flynn, J. L., & Chan, J. (2001). Tuberculosis: Latency and Reactivation. <i>Infection and Immunity</i> , 69, 4195–4201.
Frouin, I., Montecucco, A., Spadari, S., & Maga, G. (2003). DNA replication: a complex matter. <i>EMBO Reports</i> , 4, 666–670.
Galhardo, R. S., Hastings, P. J., & Rosenberg, S. M. (2007). Mutation as a stress response and the regulation of evolvability. <i>Critical Reviews in Biochemistry and Molecular Biology</i> , 42, 399–435.
Gao, D., & McHenry, C. S. (2001). Tau binds and organizes <i>Escherichia coli</i> replication through distinct domains. Partial proteolysis of terminally tagged Tau to determine candidate domains and to assign domain V as the alpha binding domain. <i>The Journal of Biological Chemistry</i> , 276, 4433–4440.
Gasteiger, E., Gattiker, A., Hoogland, C., Ivanyi, I., Appel, R. D., & Bairoch, A. (2003). ExPASy: the proteomics server for in-depth protein knowledge and analysis. <i>Nucleic Acids Research</i> , 31, 3784–3788.
Hall, R. G., Leff, R. D., & Gumbo, T. (2009). Treatment of Active Pulmonary Tuberculosis in Adults: Current Standards and Recent Advances. <i>Pharmacotherapy</i> , 29, 1468–1481.
Hare, J. M., Perkins, S. N., & Gregg-Jolly, L. A. (2006). A Constitutively Expressed, Truncated <i>umuDC</i> Operon Regulates the <i>recA</i> -Dependent DNA Damage Induction of a Gene in <i>Acinetobacter baylyi</i> Strain ADP1. <i>Applied and Environmental Microbiology</i> , 72, 4036–4043.
Haroniti, A., Anderson, C., Doddridge, Z., Gardiner, L., Roberts, C. J., Allen, S., & Soutanas, P. (2004). The Clamp-loader-Helicase Interaction in <i>Bacillus</i> . Atomic Force Microscopy Reveals the Structural Organisation of the DnaB-Complex in <i>Bacillus</i> . <i>Journal of Molecular Biology</i> , 336, 381–393.
Ippoliti, P. (2012). DNA damage specificity of <i>Escherichia coli</i> DNA polymerase DinB. <i>Chemistry Master's Theses</i> . Retrieved from http://iris.lib.neu.edu/chemistry_theses/26 .
Ippoliti, P. J., DeLateur, N. A., Jones, K. M., & Beuning, P. J. (2012). Multiple Strategies for Translesion Synthesis in Bacteria. <i>Cells</i> , 1, 799–831.
Iseman, M. D. (1993). Treatment of multidrug-resistant tuberculosis. <i>The New England Journal of Medicine</i> , 329, 784–791.

Ito, J., & Braithwaite, D. K. (1991). Compilation and alignment of DNA polymerase sequences. <i>Nucleic Acids Research</i> , <i>19</i> , 4045–4057.
Janion, C., Sikora, A., Nowosielska, A., & Grzesiuk, E. (2002). Induction of the SOS response in starved <i>Escherichia coli</i> . <i>Environmental and Molecular Mutagenesis</i> , <i>40</i> , 129–133.
Jara, M., Núñez, C., Campoy, S., Henestrosa, A. R. F. de, Lovley, D. R., & Barbé, J. (2003). <i>Geobacter sulfurreducens</i> Has Two Autoregulated <i>lexA</i> Genes Whose Products Do Not Bind the <i>recA</i> Promoter: Differing Responses of <i>lexA</i> and <i>recA</i> to DNA Damage. <i>Journal of Bacteriology</i> , <i>185</i> , 2493–2502.
Jarosz, D. F., Beuning, P. J., Cohen, S. E., & Walker, G. C. (2007). Y-family DNA polymerases in <i>Escherichia coli</i> . <i>Trends in Microbiology</i> , <i>15</i> (2), 70–77. doi:10.1016/j.tim.2006.12.004
Jesen, M., Fukushima, M., & Davis, R. (2010). DMSO and Betaine Greatly Improve Amplification of GC-Rich Constructs in De Novo Synthesis. <i>PLoS ONE</i> , <i>5</i> .
Kana, B. D., Abrahams, G. L., Sung, N., Warner, D. F., Gordhan, B. G., Machowski, E. E., ... Mizrahi, V. (2010). Role of the DinB homologs <i>Rv1537</i> and <i>Rv3056</i> in <i>Mycobacterium tuberculosis</i> . <i>Journal of Bacteriology</i> , <i>192</i> , 2220–2227.
Kapoor, N., Pawar, S., Sirakova, T. D., Deb, C., Warren, W. L., & Kolattukudy, P. E. (2013). Human Granuloma in Vitro Model, for TB Dormancy and Resuscitation. <i>PLoS ONE</i> , <i>8</i> , e53657.
Kaufmann, S. H. E., & McMichael, A. J. (2005). Annulling a dangerous liaison: vaccination strategies against AIDS and tuberculosis. <i>Nature Medicine</i> , <i>11</i> , S33–44.
Kiefer, F., Arnold, K., Künzli, M., Bordoli, L., & Schwede, T. (2009). The SWISS-MODEL Repository and associated resources. <i>Nucleic Acids Research</i> , <i>37</i> , D387–392.
Kim, D. R., & McHenry, C. S. (1996). In vivo assembly of overproduced DNA polymerase III. Overproduction, purification, and characterization of the alpha, alpha-epsilon, and alpha-epsilon-theta subunits. <i>The Journal of Biological Chemistry</i> , <i>271</i> , 20681–20689.
Kim, S., Maenhaut-Michel, G., Yamada, M., Yamamoto, Y., Matsui, K., Sofuni, T., ... Ohmori, H. (1997). Multiple pathways for SOS-induced mutagenesis in <i>Escherichia coli</i> : an overexpression of <i>dinB/dinP</i> results in strongly enhancing mutagenesis in the absence of any exogenous treatment to damage DNA. <i>Proceedings of the National Academy of Sciences of the United States of America</i> , <i>94</i> , 13792–13797.
Koonin, E., & Bork, P. (1996). Ancient duplication of DNA polymerase inferred from analysis of complete bacterial genomes. <i>Trends Biochem Sci</i> , <i>21</i> , 128–129.
Koorits, L., Tegova, R., Tark, M., Tarassova, K., Tover, A., & Kivisaar, M. (2007). Study of involvement of ImuB and DnaE2 in stationary-phase mutagenesis in <i>Pseudomonas putida</i> . <i>DNA Repair</i> , <i>6</i> , 863–868.
Kornberg, A., & Baker, T. (1992). <i>DNA Replication</i> (2nd Ed.). New York: W. H. Freeman and Co.
Laemmli, U. K. (1970). Cleavage of Structural Proteins during the Assembly of the Head of Bacteriophage T4. <i>Nature</i> , <i>227</i> , 680–685.
Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., ... Higgins, D. G. (2007). Clustal W and Clustal X version 2.0. <i>Bioinformatics</i> , <i>23</i> (21), 2947–2948.

Lavery, P. E., & Kowalczykowski, S. C. (1992). Enhancement of <i>recA</i> protein-promoted DNA strand exchange activity by volume-occupying agents. <i>Journal of Biological Chemistry</i> , 267, 9307–9314.
Le Chatelier, E., Bécherel, O. J., d' Alençon, E., Canceill, D., Ehrlich, S. D., Fuchs, R. P. P., Jannièrè, L. (2004). Involvement of DnaE, the second replicative DNA polymerase from <i>Bacillus subtilis</i> , in DNA mutagenesis. <i>The Journal of Biological Chemistry</i> , 279, 1757–1767.
Leibly, D. J., Nguyen, T. N., Kao, L. T., Hewitt, S. N., Barrett, L. K., & Van Voorhis, W. C. (2012). Stabilizing Additives Added during Cell Lysis Aid in the Solubilization of Recombinant Proteins. <i>PLoS ONE</i> , 7, e52482.
Li, M., Su, Z.-G., & Janson, J.-C. (2004). In vitro protein refolding by chromatographic procedures. <i>Protein Expression and Purification</i> , 33, 1–10.
Ling, H., Boudsocq, F., Woodgate, R., & Yang, W. (2001). Crystal structure of a Y-family DNA polymerase in action: a mechanism for error-prone and lesion-bypass replication. <i>Cell</i> , 107, 91–102.
Liu, X.-Q., & Yang, J. (2003). Split <i>dnaE</i> Genes Encoding Multiple Novel Inteins in <i>Trichodesmium erythraeum</i> . <i>Journal of Biological Chemistry</i> , 278, 26315–26318.
McDonough, K. A., Kress, Y., & Bloom, B. R. (1993). Pathogenesis of tuberculosis: interaction of <i>Mycobacterium tuberculosis</i> with macrophages. <i>Infection and Immunity</i> , 61, 2763–2773.
McGuffin, L. J., Bryson, K., & Jones, D. T. (2000). The PSIPRED protein structure prediction server. <i>Bioinformatics</i> , 16, 404–405.
McHenry, C. S. (2011a). Bacterial replicases and related polymerases. <i>Current Opinion in Chemical Biology</i> , 15, 587–594.
McHenry, C. S. (2011b). Breaking the rules: bacteria that use several DNA polymerase IIIs. <i>EMBO Reports</i> , 12, 408–414.
McKenzie, G. J., Harris, R. S., Lee, P. L., & Rosenberg, S. M. (2000). The SOS response regulates adaptive mutation. <i>Proceedings of the National Academy of Sciences of the United States of America</i> , 97, 6646–6651.
Mehlin, C., Boni, E., Buckner, F. S., Engel, L., Feist, T., Gelb, M. H., ... Hol, W. G. . (2006). Heterologous expression of proteins from <i>Plasmodium falciparum</i> : Results from 1000 genes. <i>Molecular & Biochemical Parasitology</i> , 148, 144–160.
Méndez, J., & Stillman, B. (2003). Perpetuating the double helix: molecular machines at eukaryotic DNA replication origins. <i>BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology</i> , 25, 1158–1167.
Mizrahi, V., & Andersen, S. J. (1998). DNA repair in <i>Mycobacterium tuberculosis</i> . What have we learnt from the genome sequence? <i>Molecular Microbiology</i> , 29, 1331–1339.
Ndwandwe, D. E. (2013, July 29). <i>Mechanisms of mutagenesis in Mycobacterium tuberculosis: structural and functional characterisation of the DNA polymerase accessory factors encoded by Rv3394c and Rv3395c</i> (Thesis). Retrieved from http://wiredspace.wits.ac.za/handle/10539/12918 .
Nick McElhinny, S. A., Gordenin, D. A., Stith, C. M., Burgers, P. M. J., & Kunkel, T. A. (2008). Division of Labor at the Eukaryotic Replication Fork. <i>Molecular Cell</i> , 30, 137–144.

Obermeyer, Z., Abbott-Klafter, J., & Murray, C. J. L. (2008). Has the DOTS Strategy Improved Case Finding or Treatment Success? An Empirical Assessment. <i>PLoS ONE</i> , 3, e1721.
Obiri-Danso, K., Acheampong, L., & Edoh, D. (2013). Cure rate of Tuberculosis patients using DOTS programme in Kumasi metropolis, Ghana. <i>The Internet Journal of Pulmonary Medicine</i> , 11.
Ollivierre, J. N., Fang, J., & Beuning, P. J. (2010). The Roles of UmuD in Regulating Mutagenesis. <i>Journal of Nucleic Acids</i> .
Opperman, T., Murli, S., & Walker, G. C. (1996). The Genetic Requirements for UmuDC-Mediated Cold Sensitivity Are Distinct from Those for SOS Mutagenesis. <i>Journal of Bacteriology</i> , 178, 4400–4411.
Otu, A., Umoh, V., Habib, A., & Ansa, V. (2014). Prevalence and clinical predictors of drug-resistant tuberculosis in three clinical settings in Calabar, Nigeria. <i>The Clinical Respiratory Journal</i> .
Patel, M., Jiang, Q., Woodgate, R., Cox, M. M., & Goodman, M. F. (2010). A New Model for SOS-induced Mutagenesis: How RecA Protein Activates DNA Polymerase V. <i>Critical Reviews in Biochemistry and Molecular Biology</i> , 45, 171–184.
Patel, P. H., Suzuki, M., Adman, E., Shinkai, A., & Loeb, L. A. (2001). Prokaryotic DNA polymerase I: evolution, structure, and “base flipping” mechanism for nucleotide selection. <i>Journal of Molecular Biology</i> , 308, 823–837.
Primm, T. P., Andersen, S. J., Mizrahi, V., Avarbock, D., Rubin, H., & Barry, C. E., 3rd. (2000). The stringent response of <i>Mycobacterium tuberculosis</i> is required for long-term survival. <i>Journal of Bacteriology</i> , 182, 4889–4898.
Qiu, Z., & Goodman, M. F. (1997). The <i>Escherichia coli polB</i> locus is identical to <i>dinA</i> , the structural gene for DNA polymerase II. Characterization of Pol II purified from a <i>polB</i> mutant. <i>The Journal of Biological Chemistry</i> , 272, 8611–8617.
Rachman, H., Strong, M., Schaible, U., Schuchhardt, J., Hagens, K., Mollenkopf, H., ... Kaufmann, S. H. E. (2006). <i>Mycobacterium tuberculosis</i> gene expression profile within the context of protein networks. <i>Microbes and Infections</i> , 1, 747–757.
Raja, A. (2004). Immunology of tuberculosis. <i>The Indian Journal of Medical Research</i> , 120, 213–232.
Rand, L., Hinds, J., Springer, B., Sander, P., Buxton, R. S., & Davis, E. O. (2003). The majority of inducible DNA repair genes in <i>Mycobacterium tuberculosis</i> are induced independently of RecA. <i>Molecular Microbiology</i> , 50, 1031–1042.
Rannou, O., Le Chatelier, E., Larson, M. A., Nouri, H., Dalmais, B., Laughton, C., ... Soultanas, P. (2013). Functional interplay of DnaE polymerase, DnaG primase and DnaC helicase within a ternary complex, and primase to polymerase hand-off during lagging strand DNA replication in <i>Bacillus subtilis</i> . <i>Nucleic Acids Research</i> , 41, 5303–5320.
Rattray, A. J., & Strathern, J. N. (2003). Error-Prone Dna Polymerases: When Making a Mistake is the Only Way to Get Ahead. <i>Annual Review of Genetics</i> , 37, 31–66.
Redwan, E.-R. M. R. (2006). The optimal gene sequence for optimal protein expression in <i>Escherichia coli</i> : principle requirements. <i>Arab J. Biotech</i> , 9, 493–510.

Rosenthal, M., & Fisher, B. (2013). Tuberculosis: Ancient History, Modern Scourge. <i>Journal of Ancient Diseases & Preventive Remedies</i> , 1.
Rothman, R. E., Hsieh, Y.-H., & Yang, S. (2006). Communicable respiratory threats in the ED: tuberculosis, influenza, SARS, and other aerosolized infections. <i>Emergency Medicine Clinics of North America</i> , 24, 989–1017.
Sacchettini, J. C., Rubin, E. J., & Freundlich, J. S. (2008). Drugs versus bugs: in pursuit of the persistent predator <i>Mycobacterium tuberculosis</i> . <i>Nature Reviews Microbiology</i> , 6, 41–52.
Sanchez-Alberola, N., Campoy, S., Barbé, J., & Erill, I. (2012). Analysis of the SOS response of <i>Vibrio</i> and other bacteria with multiple chromosomes. <i>BMC Genomics</i> , 13.
Sanders, G. M., Dallmann, H. G., & McHenry, C. S. (2010). Reconstitution of the <i>B. subtilis</i> replisome with 13 proteins including two distinct replicases. <i>Molecular Cell</i> , 37, 273–281.
Saunders, R., & Deane, C. M. (2010). Synonymous codon usage influences the local protein structure observed. <i>Nucleic Acids Research</i> , 38, 6719–6728.
Sharma, S. K., & Mohan, A. (2013). Tuberculosis: From an incurable scourge to a curable disease - journey over a millennium. <i>The Indian Journal of Medical Research</i> , 137, 455–493.
Silva, C. L., & Lowrie, D. B. (2000). Identification and Characterization of Murine Cytotoxic T Cells That Kill <i>Mycobacterium tuberculosis</i> . <i>Infection and Immunity</i> , 68, 3269–3274.
Singh, S. M., & Panda, K. A. (2005). Solubilization and Refolding of Bacteria Inclusion Body Proteins. <i>Journal of Biomedicine and Bioengineering</i> , 99, 303–310.
Singh, V., Chandra, D., Srivastava, B. S., & Srivastava, R. (2011). Biochemical and transcription analysis of acetohydroxyacid synthase isoforms in <i>Mycobacterium tuberculosis</i> identifies these enzymes as potential targets for drug development. <i>Microbiology</i> , 157, 29–37.
Slabinski, L., Jaroszewski, L., Rodrigues, A. P. C., Rychlewski, L., Wilson, I. A., Lesley, S. A., & Godzik, A. (2007). The challenge of protein structure determination—lessons from structural genomics. <i>Protein Science</i> , 16, 2472–2482.
Smith, B. T., & Walker, G. C. (1998). Mutagenesis and More: <i>umuDC</i> and the <i>Escherichia coli</i> SOS Response. <i>Genetics</i> , 148, 1599–1610.
Stewart, G. S., Wang, B., Bignell, C. R., Taylor, A. M. R., & Elledge, S. J. (2003). MDC1 is a mediator of the mammalian DNA damage checkpoint. <i>Nature</i> , 421, 961–966.
Takeda, D. Y., & Dutta, A. (2005). DNA replication and progression through S phase. <i>Oncogene</i> , 24, 2827–2843.
Terwilliger, T. C., Stuart, D., Yokoyama, S. (2009). Lessons from structural genomics. <i>Annual Review of Biophysics</i> , 38, 371–383.
Tsumoto, K., Umetsu, M., Yamada, H., Ito, T., Misawa, S., & Kumagai, I. (2003). Immobilized oxidoreductase as an additive for refolding inclusion bodies: application to antibody fragments. <i>Protein Engineering</i> , 16, 535–541.
Vallejo, L. F., & Rinas, U. (2004). Strategies for the recovery of active proteins through refolding of bacterial inclusion body proteins. <i>Microbial Cell Factories</i> , 3, 11.
Volmink, J., & Garner, P. (2007). Directly observed therapy for treating tuberculosis. <i>The Cochrane Database of Systematic Reviews</i> , 1, CD003343.

Walsh, J. M., Hawver, L. A., & Beuning, P. J. (2011). <i>Escherichia coli</i> Y family DNA polymerases. <i>Frontiers in Bioscience</i> , <i>16</i> , 3164–3182.
Warner, D. F., Ndwandwe, D. E., Abrahams, G. L., Kana, B. D., Machowski, E. E., Venclovas, C., & Mizrahi, V. (2010). Essential roles for <i>imuA</i> '- and <i>imuB</i> -encoded accessory factors in DnaE2-dependent mutagenesis in <i>Mycobacterium tuberculosis</i> . <i>Proceedings of the National Academy of Sciences of the United States of America</i> , <i>107</i> , 13093–13098.
Washington, M. T., Carlson, K. D., Freudenthal, B. D., & Pryor, J. M. (2010). Variations on a theme: eukaryotic Y-family DNA polymerases. <i>Biochimica et Biophysica Acta</i> , <i>1804</i> , 1113–1123.
Waters, L. S., Minesinger, B. K., Wiltrout, M. E., D'Souza, S., Woodruff, R. V., & Walker, G. C. (2009). Eukaryotic Translesion Polymerases and Their Roles and Regulation in DNA Damage Tolerance. <i>Microbiology and Molecular Biology Reviews : MMBR</i> , <i>73</i> , 134–154.
WHO (2003). WORLD TB DAY 2003 HIGHLIGHTS REPORT. Retrieved from http://www.stoptb.org/assets/documents/events/world_tb_day/2003/wtbd_2003_highlight_s.pdf
WHO (2009). Global Tuberculosis Control 2009 Epidemiology Strategy Financing. <i>World Health Organisation</i> . Retrieved from http://reliefweb.int/sites/reliefweb.int/files/resources/878BDA5E2504C9F449257584001B5E60-who_mar2009.pdf
WHO (2012). GLOBAL TUBERCULOSIS REPORT 2012. Retrieved from http://www.who.int/tb/publications/global_report/gtbr12_main.pdf
Wickner, W., & Kornberg, A. (1973). DNA Polymerase III Star Requires ATP to Start Synthesis on a Primed DNA. <i>Proceedings of the National Academy of Sciences of the United States of America</i> , <i>70</i> , 3679–3683.
Williams, D. C., Van Frank, R. M., Muth, W. L., & Burnett, J. P. (1982). Cytoplasmic inclusion bodies in <i>Escherichia coli</i> producing biosynthetic human insulin proteins. <i>Science</i> , <i>215</i> , 687–689.
Witkin, E. M. (1967). The radiation sensitivity of <i>Escherichia coli</i> B: a hypothesis relating filament formation and prophage induction. <i>Proceedings of the National Academy of Sciences of the United States of America</i> , <i>57</i> , 1275–1279.
Wolf, A. J., Linas, B., Trevejo-Nuñez, G. J., Kincaid, E., Tamura, T., Takatsu, K., & Ernst, J. D. (2007). <i>Mycobacterium tuberculosis</i> infects dendritic cells with high frequency and impairs their function in vivo. <i>Journal of Immunology</i> , <i>179</i> , 2509–2519.
Wu, S., & Zhang, Y. (2007). LOMETS: A local meta-threading-server for protein structure prediction. <i>Nucleic Acids Research</i> , <i>35</i> , 3375–3382.
Yang, W. (2003). Damage Repair DNA Polymerases Y. <i>Current Opinion in Structural Biology</i> , <i>23</i> –30.
Ye, J., McGinnis, S., & Madden, T. L. (2006). BLAST: improvements for better sequence analysis. <i>Nucleic Acids Research</i> , <i>34</i> , W6–W9.
Žgur-Bertok, D. (2013). DNA Damage Repair and Bacterial Pathogens. <i>PLoS Pathog</i> , <i>9</i> , e1003711.

Zhang, G., Hubalewska, M., & Ignatova, Z. (2009). Transient ribosomal attenuation coordinates protein synthesis and co-translational folding. <i>Nature Structural & Molecular Biology</i> , 16, 274–280.
Zumla, A., Raviglione, M., Hafner, R., & Fordham von Reyn, C. (2013). Tuberculosis. <i>New England Journal of Medicine</i> , 368, 745–755.
Part II
Bornhorst, J. A., & Falke, J. J. (2000). Purification of proteins using polyhistidine affinity tags. <i>Methods in Enzymology</i> , 326, 245–254.
Chakladar, S., Abadi, S. S. K., & Bennet, A. J. (2014). A mechanistic study on the α -N-acetylgalactosaminidase from <i>E. meningosepticum</i> : a family 109 glycoside hydrolase. <i>MedChemComm</i> .
Chauvaux, S., Souchon, H., Alzari, P. M., Chariot, P., & Béguin, P. (1995). Structural and functional analysis of the metal-binding sites of Clostridium thermocellum endoglucanase CelD. <i>The Journal of Biological Chemistry</i> , 270, 9757–9762.
Chitarra, V., Souchon, H., Spinelli, S., Juy, M., Béguin, P., & Alzari, P.M. (1995). Multiple crystal forms of endoglucanase CelD: signal peptide residues modulate lattice formation. <i>Journal of Molecular Biology</i> , 248, 225–32.
Davies, G., & Henrissat, B. (1995). Structures and mechanisms of glycosyl hydrolases. <i>Structures</i> , 3, 853–859.
Duan, C.-J., Xian, L., Zhao, G.-C., Feng, Y., Pang, H., Bai, X.-L., ... Feng, J.-X. (2009). Isolation and partial characterization of novel genes encoding acidic cellulases from metagenomes of buffalo rumens. <i>Journal of Applied Microbiology</i> , 107, 245–256.
Ferrer, M., Golyshina, O. V., Chernikova, T. N., Khachane, A. N., Reyes-Duarte, D., Santos, V. A. P. M. D., ... Golyshin, P. N. (2005). Novel hydrolase diversity retrieved from a metagenome library of bovine rumen microflora. <i>Environmental Microbiology</i> , 7, 1996–2010.
Gasteiger, E., Hoogland, C., Gattiker A., Duvaud, S., Wilkins, M. R., Appel, R. D. & Bairoch A. (2005). <i>Protein Identification and Analysis Tools on the ExPASy Server</i> ; (In) John M. Walker: The Proteomics Protocols Handbook. <i>Humana Press</i> .
Gebler, J., Gilkes, N. R., Claeysens, M., Wilson, D. B., Béguin, P., Wakarchuk, W. W., ... Withers, S. G. (1992). Stereoselective hydrolysis catalyzed by related beta-1,4-glucanases and beta-1,4-xylanases. <i>Journal of Biological Chemistry</i> , 267, 12559–12561.
Gloster, T. M., Turkenburg, J. P., Potts, J. R., Henrissat, B., & Davies, G. J. (2008). Divergence of catalytic mechanism within a glycosidase family provides insight into evolution of carbohydrate metabolism by human gut flora. <i>Chemistry & Biology</i> , 15, 1058–1067.
Guérin, D. M. A., Lascombe, M. B., Costabel, M., Souchon, H., Lamzin, V., Béguin, P., & Alzari, P. M. (2002). Atomic (0.94 Å) resolution structure of an inverting glycosidase in complex with substrate. <i>Journal of Molecular Biology</i> , 316, 1061–1069.

<p>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. <i>Applied Microbiology and Biotechnology</i>, 43, 667–674.</p>
<p>Henrissat, B. & Bairoch, A. (1996). Updating the sequence-based classification of glycosyl hydrolases. <i>Biochemistry</i> 316, 695-6.</p>
<p>Henrissat, B., & Davies, G. J. (2000). Glycoside Hydrolases and Glycosyltransferases. Families, Modules, and Implications for Genomics. <i>Plant Physiology</i>, 124, 1515–1519.</p>
<p>Henrissat, B., Callebaut, I., Fabrega, S., Lehn, P., Mornon, J. P., & Davies, G. (1995). Conserved catalytic machinery and the prediction of a common fold for several families of glycosyl hydrolases. <i>Proc Natl Acad Sci USA</i>, 92, 7090-4.</p>
<p>Henrissat, B., Callebaut, I., Fabrega, S., Lehn, P., Mornon, J. P., & Davies, G. (1995). Conserved catalytic machinery and the prediction of a common fold for several families of glycosyl hydrolases. <i>Proc Natl Acad Sci USA</i>, 92, 7090-4.</p>
<p>http://www.iea.org/publications/freepublications/publication/WEO_Special_Report_2013_Redrawing_the_Energy_Climate_Map.pdf</p>
<p>IEA (2013). Redrawing the energy-climate map. World energy outlook special reports International Energy Agency, Paris.</p>
<p>Juy, M., Amrt, A. G., Alzari, P. M., Poljak, R. J., Claeysens, M., Béguin, P., & Aubert, J.-P. (1992). Three-dimensional structure of a thermostable bacterial cellulase. <i>Nature</i>, 357, 89–91.</p>
<p>Kim, S. J., Lee, C. M., Han, B. R., Kim, M. Y., Yeo, Y. S., Yoon, S. H., ... Jun, H. K. (2008). Characterization of a gene encoding cellulase from uncultured soil bacteria. <i>FEMS Microbiology Letters</i>, 282, 44–51.</p>
<p>Koshland, D. E. (1953). Stereochemistry and the mechanism of enzymatic reactions. <i>Biological Reviews</i>, 28, 416–436.</p>
<p>Lee, R. A., & Lavoie, J.-M. (2013). From first- to third-generation biofuels: Challenges of producing a commodity from a biomass of increasing complexity. <i>Animal Frontiers</i>, 3, 6–11.</p>
<p>Lynd, L. R., Weimer, P. J., van Zyl, W. H., & Pretorius, I. S. (2002). Microbial cellulose utilization: fundamentals and biotechnology. <i>Microbiology and Molecular Biology Reviews: MMBR</i>, 66, 506–577.</p>
<p>Minor, W., Cymborowski, M., Otwinowski, Z., & Chruszcz, M. (2006). HKL-3000: the integration of data reduction and structure solution--from diffraction images to an initial model in minutes. <i>Acta Crystallographica. Section D, Biological Crystallography</i>, 62, 859–866.</p>
<p>Naumoff, D. G. (2011). Hierarchical classification of glycoside hydrolases. <i>Biochemistry. Biokhimiia</i>, 76, 622–635.</p>
<p>Otwinowski, Z., & Minor, W. (1997). Processing of X-ray Diffraction Data Collected in Oscillation Mode. <i>Methods in Enzymology</i>, 276, 307–326.</p>

Pereira, J. H., Sapra, R., Volponi, J. V., Kozina, C. L., Simmons, B., & Adams, P. D. (2009). Structure of endoglucanase Cel9A from the thermoacidophilic <i>Alicyclobacillus acidocaldarius</i> . <i>Acta Crystallographica Section D: Biological Crystallography</i> , 65, 744–750.
Quiroz-Castañeda, R. E., Folch-Mallol, J. L. (2013). Hydrolysis of biomass mediated by cellulases for the production of sugars. (In) <i>Sustainable degradation of lignocellulosic biomass—Techniques, applications and commercialization</i> ; Chandel, A. K. & Silva, S. S., p. 275.
Rubin, E. M. (2008). Genomics of cellulosic biofuels. <i>Nature</i> , 454(7206), 841–845.
Sakon, J., Irwin, D., Wilson, D. B., & Karplus, P. A. (1997). Structure and mechanism of endo/exocellulase E4 from <i>Thermomonospora fusca</i> . <i>Nature Structural & Molecular Biology</i> , 4, 810–818.
Sánchez, Ó. J., & Cardona, C. A. (2008). Trends in biotechnological production of fuel ethanol from different feedstocks. <i>Bioresource Technology</i> , 99, 5270–5295.
Sandgren, M., Gualfetti, P. J., Shaw, A., Gross, L. S., Saldajeno, M., Day, A. G., Jones, T. A. and Mitchinson, C. (2003). Comparison of family 12 glycoside hydrolases and recruited substitutions important for thermal stability. <i>Protein Science</i> , 12, 848-866
Schubot, F. D., Kataeva, I. A., Chang, J., Shah, A. K., Ljungdahl, L. G., Rose, J. P., & Wang, B. C. (2004). Structural basis for the exocellulase activity of the cellobiohydrolase CbhA from <i>Clostridium thermocellum</i> . <i>Biochemistry</i> , 43, 1163–1170.
Sims, R., Taylor, M., Saddler, J., & Mabee, W. (2008). From 1 st –to 2 nd -generation biofuel technologies: an overview of current industry and RD&D activities. <i>International energy agency and organisation for economic cooperation and development</i> .
Sinnott, M. L. (1990). Catalytic mechanism of enzymic glycosyl transfer. <i>Chemical Reviews</i> , 90, 1171–1202.
Somerville, C., Bauer, S., Brininstool, G., Facette, M., Hamann, T., Milne, J., ... Youngs, H. (2004). Toward a systems approach to understanding plant cell walls. <i>Science</i> , 306, 2206–2211.
Sulzenbacher, G., Driguez, H., Henrissat, B., Schülein, M., & Davies, G. J. (1996). Structure of the <i>Fusarium oxysporum</i> Endoglucanase I with a Nonhydrolyzable Substrate Analogue: Substrate Distortion Gives Rise to the Preferred Axial Orientation for the Leaving Group. <i>Biochemistry</i> , 35, 15280–15287.
Tenenbaum, D. J. (2008). Food vs. fuel: Diversion of crops could cause more hunger. <i>Environmental Health Perspectives</i> , 116, A254–A257.
Vuong, T. V., & Wilson, D. B. (2010). Glycoside hydrolases: Catalytic base/nucleophile diversity. <i>Biotechnology and Bioengineering</i> , 107, 195–205.
Wang, F., Li, F., Chen, G., & Liu, W. (2009). Isolation and characterization of novel cellulase genes from uncultured microorganisms in different environmental niches. <i>Microbiological Research</i> , 164(6), 650–657.

<p>Wi, S. ., Singh, A. P., Lee, K. H., & Kim, Y. S. (2005). The pattern of distribution of pectin, peroxidase and lignin in the middle lamella of secondary xylem fibres in alfalfa (<i>Medicago sativa</i>). <i>Annals of Botany</i>, 863–868.</p>
<p>Withers, S. G. (2001). Mechanisms of glycosyl transferases and hydrolases. <i>Carbohydrate Polymers</i>, 44, 325–337.</p>
<p>Xia, Y., Ju, F., Fang, H. H. P., & Zhang, T. (2013). Mining of novel thermo-stable cellulolytic genes from a thermophilic cellulose-degrading consortium by Metagenomics. <i>PLoS ONE</i>, 8,</p>
<p>Yan, S., & Wu, G. (2013). Secretory pathway of cellulase: a mini-review. <i>Biotechnology for Biofuels</i>, 6, 177.</p>
<p>Yip, V. L. Y., Varrot, A., Davies, G. J., Rajan, S. S., Yang, X., Thompson, J., ... Withers, S. G. (2004). An unusual mechanism of glycoside hydrolysis involving redox and elimination steps by a family 4 β-glycosidase from <i>Thermotoga maritima</i>. <i>Journal of the American Chemical Society</i>, 126, 8354–8355.</p>
<p>Yip, V. L., Thompson, J. & Withers, S. G. (2007). Mechanism of GlvA from <i>Bacillus subtilis</i>: a detailed kinetic analysis of a 6-phospho-alpha-glucosidase from glycoside hydrolase family 4. <i>Biochemistry</i>, 46, 9840-52.</p>

