ANALYSIS OF OLIGOSACCHARIDE BINDING TO LysM DOMAINS

AZURA MOHD NOOR

A thesis submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy

University of Sheffield

Department of Molecular Biology and Biotechnology

October 2019

Abstract

The LysM domain principally binds N-acetyl glucosamine (NAG) and N-acetyl muramic acid (NAM) sugars, major constituents of molecules such as peptidoglycan, chitin and other related molecules including lipochito-oligosaccharides such as the Nod-factor. This module is widely distributed in all domains of life. This thesis reports an analysis of the structure/function relationship of *M. avium* KEG15107, a protein that contains four LysM domains but with no covalently linked catalytic domains. Structure analysis revealed that KEG15107 formed a protease-resistant globular structure with the four LysM domains (Domains 1, 2, 3 and 4) tightly packed against each other. This tight packing resembles that seen between the multiple LysM domains in plant AtCERK1 and fungal Ecp6 but, whereas the interactions in these two proteins are stabilized by disulphide bridges, no such covalent interactions are found in KEG15107. This finding challenges the view proposed on the basis of studies of the multiple LysM domains in AtlA that non-covalently linked LysM domains are assembled like beads on a string.

Structural studies show that apo KEG15107 forms a tetramer whereas, in the presence of different oligosaccharides, the protein consistently formed a dimer. This difference in quaternary structures could be seen to be due to the binding of an oligosaccharide chain to Domain 1 thereby disrupting contacts in the tetramer involving this domain. Analysis of the structure of the dimer suggested that the tetramer must dissociate to allow the oligosaccharide to bind, indicating that, in solution, different quaternary structures must be in equilibrium with each other. An investigation by mass spectrometry suggested that in solution, the protein is a mixture of monomeric, dimeric and tetrameric forms.

Analysis of the high-resolution structure of complexes of KEG15107 with NAG_3, NAG_4 and NAG_5 suggest that there are five oligosaccharide binding sites (S0, S1, S2, S3 and S4). Analysis of the structures reveals that sites S1 and S3 exclusively recognises the smaller NAG moiety with the other sites being capable of binding either NAG or the larger sugar, NAM, in an orientation where its more bulky side chain faces the solvent.

Whilst the structural studies suggested that the oligosaccharide chains only bound to Domain 1, mass spectrometry indicated that all four domains are capable of binding oligosaccharide. Extension of these studies to the LysM binding region of Rv1288

provided support for the use of the mass spectrometry in the analysis of oligosaccharide binding. Modelling studies, combined with mass spectrometry analysis, suggested that in the monomer, the oligosaccharide binding sites on each of the LysM domains are exposed to the solvent suggesting that the protein is capable of recognizing multiple oligosaccharide chains at the same time. Consideration of the relative orientation and separation of the oligosaccharide binding sites on the different LysM domains favours the model where at least some of the chains are antiparallel and separated by ~40 Å. The latter distance is consistent with studies elsewhere on the separation of the peptidoglycan chains in the cell wall.

Acknowledgment

Bismillahirrohmanirrohim.

With the name of Allah, I would like to express my sincere gratitude to my supervisor, Prof David. W. Rice for his vital support, guidance, and encouragement until this thesis finished.

I would also like to express my gratitude to Dr. John Rafferty, Dr. Jaramey Craven, Dr. Patrick Baker, Dr. Claudine Bisson, Dr. Svetlana Sedelnikova, Fiona, Dr. Svetomir Tzokov, Mr Simon Thorpe, Dr. Adelina and Dr. George Turner for their endless support and guidance. For all my friends in the Crystallography lab and other Microbiology Lab, thank you so much for being there all the times I needed. Thank you so much for all the staffs in the Department of Molecular Biology and Biotechnology four the help.

To my wonderful husband, Syahrul Nizam Rosnan, I have no words to express how grateful I am by having you beside me. You are the one who is always there to comfort me, to pet me, to motivate me, to share all the sadness and happiness moments with me. I really appreciate it and I pray to Allah to keep us together forever till Jannah. To my beloved kids, Muhammad Danish, Muhammad Dhafir and Siti Aisyah, you are the precious things that Allah sends to me to face this life, and I dedicate this thesis to you. Thank you so much for being with me for every single of time. Barakallauhufik…

Abbreviation

Å	Angstrom
$A_{260/280}$	Absorbance
AC	Affinity chromatography
Ala	Alanine
AmpC	Ampicillin C
$AmSO_4$	Ammonium sulphate
AU	Asymmetric unit
ax	Axial
CC	Correlation coefficient
CFE	Cell-free extract
CH_2	Methyl
Da	Dalton
DNA	Deoxyribonucleic acid
EDTA	Ethylenediaminetetraacetic acid
EM	Electron microscope
EPE	*HEPES*/ethanesulfonic acid
eq	Equatorial
ESBL	Extended-spectrum beta lactamase
g	Gram
Gram-	Gram negative
Gram+	Gram positive
His-tag	Histidine tag
ht	Height
I	Intensity
IC	Inner cortex
IM	Inner membrane
IPTG	Isopropyl β-D-1-thiogalactopyranoside
IWZ	Inner wall zone
K	Lysine
Kav	Proportion of pores available to the molecule
kDa	kiloDalton

kL	Litter
LB	Luria bertani
Lipo	Lipoprotein
LTA	Lipoteichoic acid
MPD	2-methyl-2,4-pentanediol
mAU	Milli-absorbance unit
MDRO	Multi-drug resistant organism
Mem	Membrane
MES	2-(N-morpholino)ethanesulfonic acid
mL	Milliliter
mM	Millimolar
mm	Millimeter
MR	Molecular replacement
MRSA	Methicillin-resistance Staphylococcus
MS	Mass spectrometry
MW	Molecular weight
NaCl	Sodium chloride
NAG	N-acetyl glucosamine
NAM	N-acetyl muramic acid
NCBI	National Center for Biotechnology Information
OC	Outer cortex
OD	Optical density
OH	Hydroxyl
OM	Outer membrane
OWZ	Outer wall zone
PAMP	Pathogen-associated molecular pattern
PBP	Penicillin binding protein
PCR	Polymerase chain reaction
PDB	Protein database
PEG	Polyethylene glycol
PG	Peptidoglycan
Phe/F	Phenylalanine
pI	Isoelectric point

R	Arginine
rbs	Ribosome binding site
rSAP	Shrimp alkaline phosphatase
s	Second
S	Strand
H	Helix
SDS-PAGE	sodium dodecyl sulfate-polyacrylamide gel electrophoresis
Sec	Secretion
SEC	Size exclusion chromatography
SOB	Super minimal broth
SO_4	Sulphate
Tm	Melting temperature
Tris-HCL	Tris-Hydrochloride
UV	Ultra violet
V_o	Void volume
V_t	Total volume
w/w	Weight over weight
WTA	Wall teichoic acid
α	Alpha
β	Beta
µg	Microgram
µL	Microlitter
σ	Gamma
φ	Phi
Ψ	Psi

Table of contents

CHAPTER THREE
CLONING, EXPRESSION, PURIFICATION, CRYSTALLIZATION AND ELECTRON MICROSCOPE ANALYSIS OF Rv1288 FROM *MYCOBACTERIUM TUBERCULOSIS*

FIGURES

TABLES

CHAPTER 1

INTRODUCTION

1.0 Background

The recent emergence of multidrug-resistant bacteria is a major problem for all countries in the world. They now face large financial costs in order to combat superbugs. These can seriously affect global health conditions and increase mortality rates. According to the WHO global tuberculosis 2018 report, there were approximately 6.7 million cases. In 2017, it was also reported that there were 1.6 million deaths from tuberculosis (TB) infections (Figure 1.0). Of the total TB cases, approximately 180 000 cases of multidrug-resistant TB and rifampicin-resistant TB were recorded (World Day TB, 2017, WHO, 2018). An increment in latent TB cases worsens TB management (Getahun, Matteelli, Chaisson, & Raviglione, 2015). Diabetes mellitus has been reported as one of the co-factors for TB infection (Kibirige, Ssekitoleko, Mutebi, & Worodria, 2013).

It has been established that the highly complex structure of Mycobacterial cell envelopes with high lipid content can increase the impermeability of the cells to small molecules causing high resistance of *Mycobacterium tuberculosis* (MTB) to many drugs (Hett & Rubin, 2008, Yao *et al*., 2012). The available antibiotics for MTB treatment include Ethambutol, which inhibits the polymerization step of arabinogalactan; and Isoniazid, an activated KatG catalase that inhibits mycolic acid synthesis (Ying Zhang *et al*., 1992, Katarina et al., 1995). However, the emergence of isoniazid-resistant strain in MTB has affected the global tuberculosis control programs (Stagg *et al*., 2017). So far, recent advances in developing TB vaccines have been actively investigated (Vishwanath *et al*., 2015, Tang, Yam, & Chen, 2016, Kaufmann *et al*., 2017, Cardona, 2018).

There has been an increase in severe community and hospital-acquired infections caused by methicillin-resistant *Staphylococcus aureus* (MRSA) and extended-spectrum beta-lactamase (ESBL) Gram-negative bacteria. These can result in a higher risk of death among the population and requires higher health care expenditures (Moellering,

2012). Another disease that also contributes to serious health problems is melioidosis. The disease is caused by *Burkholderia pseudomallei*, a pathogen that can survive in the host latently up to twenty years and even longer. Treatment for melioidosis is frequently long and failure to take appropriate antibiotics increases the frequency of relapses. In addition, the mortality rate among melioidosis patients is relatively high due to difficulties in diagnosis resulting in increased death rates. Malaysia is one of the endemic countries for this disease and cases have been increasing annually (IMR report).

There has been growing global concern regarding other multidrug-resistant bacteria, the so-called 'multidrug-resistant organisms' (MDRO). The group consists of *Klebsiella pneumonia, Acinetobacter baumannii,* and *Enterobacter* species and all of them resistant to multiple antimicrobial agents (Chan et al., 2018). The most recently reported cases identify the existence of NDM-1 which stands for New Delhi Metallo-beta lactamase 1. It was discovered in a *K. pneumonia* strain isolated from a patient who acquired the organism in New Delhi, India. The NDM-1 strain is also found in the Enterobacteriaceae family and is widespread throughout India, Pakistan, and Bangladesh and even in Britain (Berrazeg et al., 2012). The spread of these organisms is a major global health problem as a number of them are resistant to all antimicrobial agents with the exception of polymyxin (Moellering, 2012).

Carbapenems are antibiotics used to treat patients who are suspected of being infected with multi-drug resistant bacteria. These antibiotics are members of the beta-lactam family including penicillin, cephalosporin and monobactam (Figure 1.1) (Nikolaidis *et al*., 2014). These agents kill bacteria by binding to the penicillin-binding proteins and inhibit the formation of peptidoglycan cross-links in the bacterial cell wall (Figure 1.2) (Nikolaidis *et al*., 2014). The carbapenems exhibit a broader spectrum of antimicrobial activities. Presently, however, they tend to be less effective in preventing the emergence of resistance in new superbugs (Baum *et al*., 2001, Jin *et al*., 2009, Gautam *et al*., 2011).

All ESBL producing organisms induce β-lactamase in the presence of a high concentration of drugs and these activities correlate with the process of peptidoglycan precursor synthesis. β-lactamase hydrolyzes the β-lactam ring of the antibiotic before it blocks the penicillin-binding protein active site (PBP). In the presence of β-lactam, the PBP is inhibited and affects the synthesis of peptidoglycan. Through the AmpG

permease protein anchored in the cell membrane, the muropeptides enter the cytoplasm and either activate the AmpG-AmpR-AmpC pathway or phosphorylation of BlrA to induce ampC gene. The latter is responsible for producing beta-lactamase enzymes in gram-negative organisms (Figure 1.3) (Nikolaidis *et al*., 2014). The Metallo-β-lactamase (MBL) organisms use a different mechanism for resisting the β-lactam antibiotics. These organisms contain one or two zinc ions within the active sites of their enzymes to hydrolyze the β-lactam ring. Investigation on the gene encoding this enzyme demonstrates that it can hydrolyze almost all known β-lactam antibiotics. It is commonly found in the *Pseudomonas aeruginosa*, *Acinetobacter baumannii* and Enterobacteriaceae family (Nikolaidis *et al*., 2014).

Glycopeptides are antimicrobial agents including vancomycin, teicoplanin, and telavancin and are the last resort antibiotics to treat infections by gram-positive cocci including *S. aureus*. Vancomycin binds tightly to the D-Ala:D-Ala cross-linking peptide of peptidoglycan, blocking the ability to form a cross-link with proper transpeptidation and transglycosylation activities of the peptidoglycan units (Chen *et al*., 2003, Pace & Yang, 200, Nikolaidis *et al*., 2014). The multidrug-resistant MRSA generates different peptidoglycan precursors carrying D-Ala:D-lac or D-Ala:D-Ser instead of D-Ala:D-Ala, and this phenomenon affects the vancomycin binding affinity to its target. There is an urgent need to find new antimicrobial candidates as well as vaccines in order to combat these aggressive superbugs. Therefore, an understanding of peptidoglycan architecture and its molecular recognition by other molecules in bacterial cells may lead to more accurate identification of new drug targets.

The molecular recognition of peptidoglycan is important as it regulates essential biological activities in living cells. These may be related to the pathogenicity of microorganisms. There are five main factors: cell adhesion of pathogens to infected hosts; cell-signaling; mediation of the immune response; response to bacterial and fungal infections and cell division activities. Given that cell walls are required for bacterial survival, and no homologs of peptidoglycan in human, interfering with the cell wall synthesis remains an attractive target. Peptidoglycan, in relation with other cell wall components such as cell wall-associated proteins, can provide clues for novel therapeutics.

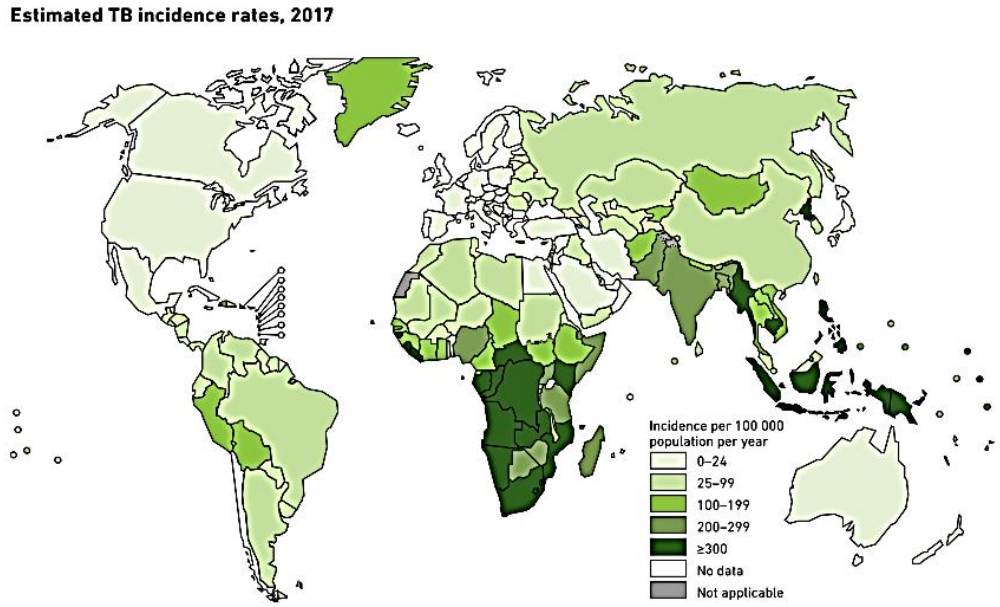


Figure 1.0: The distribution of TB cases around the world (WHO, 2018).

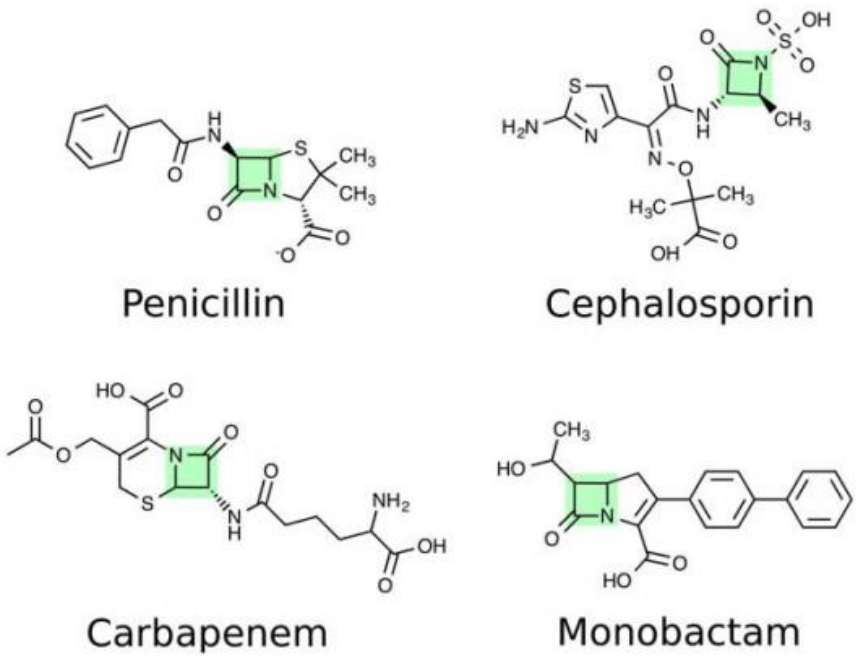


Figure 1.1: Chemical structures of B-lactam family drugs. The common beta-lactam ring is highlighted in green. The schematic diagram is taken from Nikolaidis *et al*., 2014.

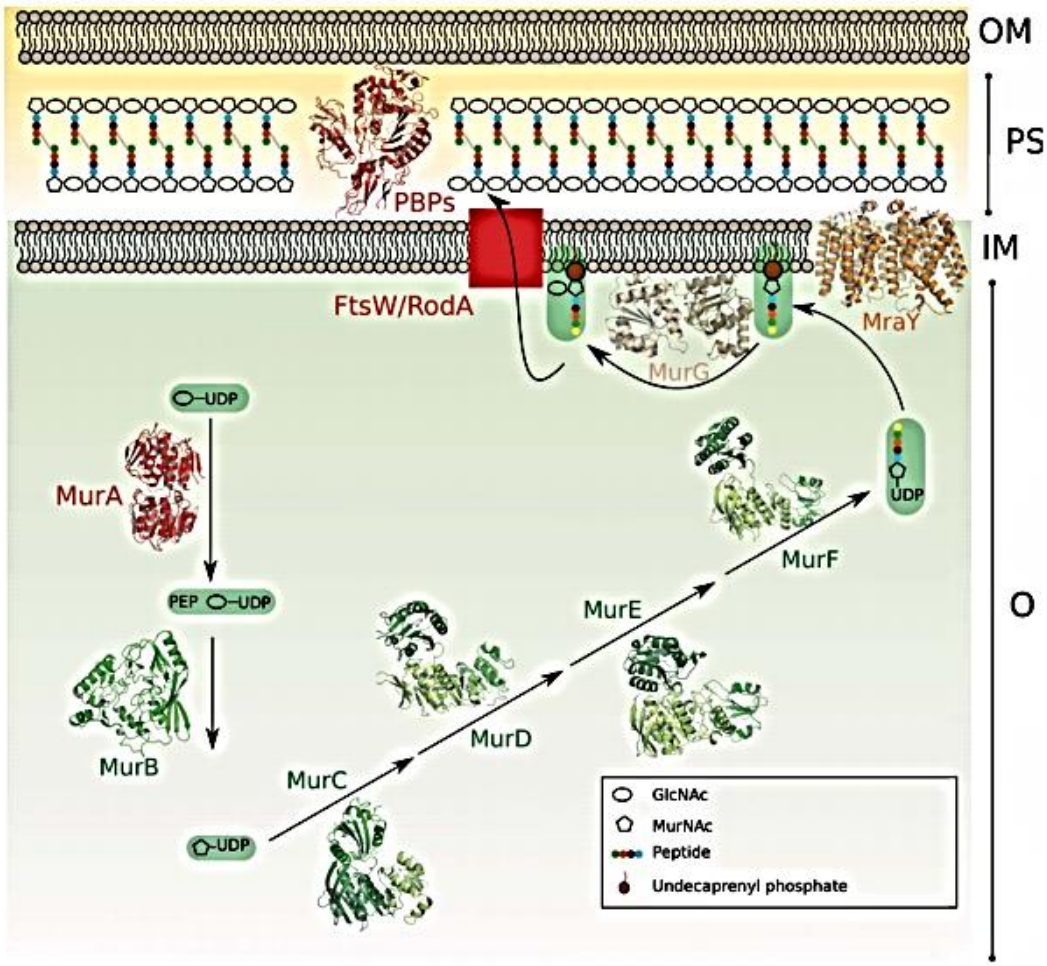


Figure 1.2: A diagram of the peptidoglycan biosynthetic pathway. The synthesis of peptidoglycan precursors involved different domains of Mur enzymes in the cytoplasm before they are transported across inner membranes through penicillin-binding proteins into the periplasm. PBP is the target of beta-lactam antibiotics. IM: inner membrane, O: cytoplasm, PS: periplasm, OM: outer membrane, PBP: penicillin-binding proteins. The figure is taken from Nikolaidis *et al*., 2014.

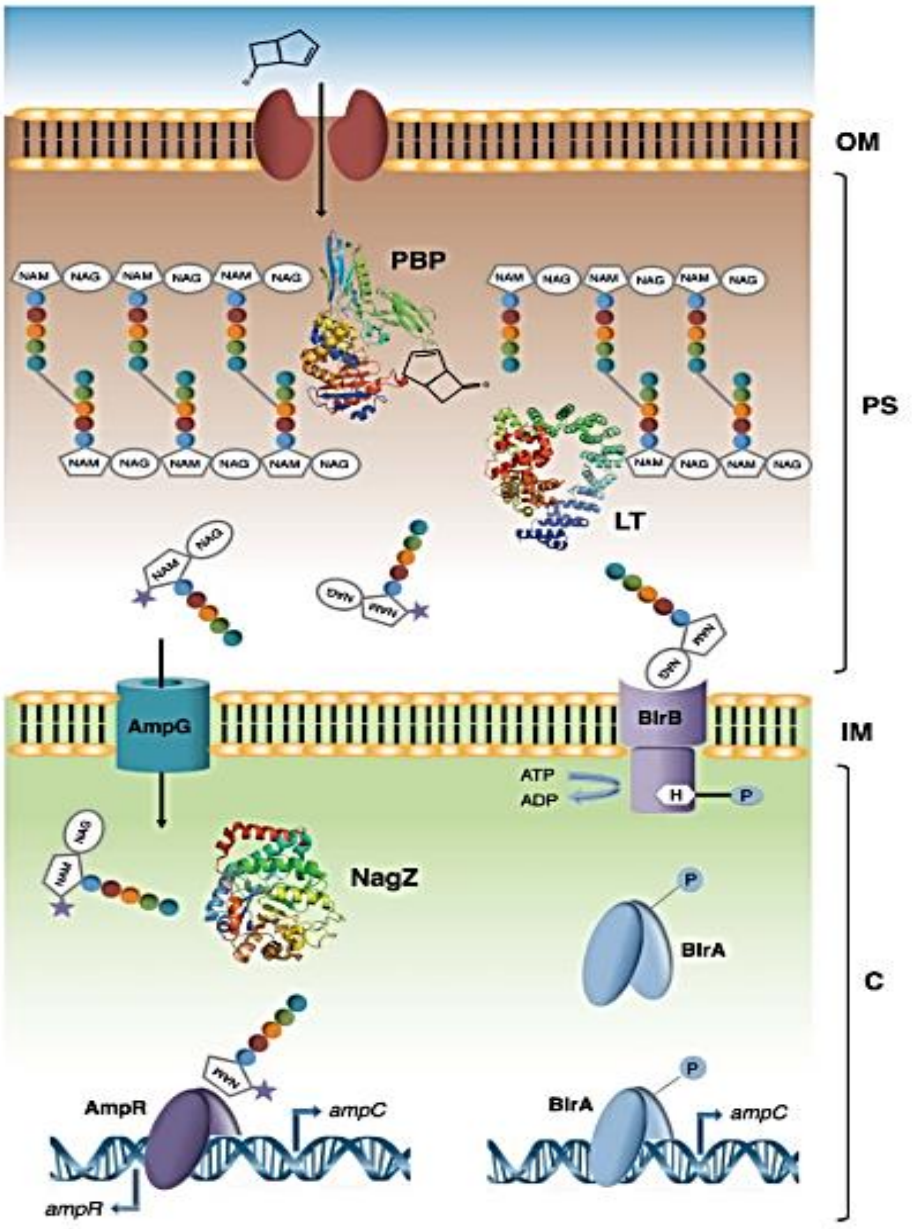


Figure 1.3**:** A schematic diagram of the AmpC b-lactamase induction in gram-negative organisms via two regulatory systems either the AmpG–AmpR–AmpC pathway or the phosphorylation by the BlrA gene. The presence of B-lactams results in the massive breakdown of the peptidoglycan networks and accumulates muropeptides in the cytoplasm which causes either the activation of the AmpR (AmpG–AmpR–AmpC pathway or the phosphorylation of BlrA. The activation of these regulators induces the ampC gene which is responsible for beta-lactamase enzymes production. OM: the outer membrane, PS: periplasm, IM: inner membrane, C: cytoplasm, PBP: penicillin-binding protein, LT: lytic transglycosylase. The schematic diagram is taken From Nikolaidis *et al*., 2014.

1.1 Cell wall-associated proteins

Cell wall-associated proteins are partially or fully secreted in bacterial cells, either covalently or noncovalently attached to membranes or peptidoglycan, or other cell wall components (Schneewind & Missiakas, 2012). These proteins usually contain binding domains such as LPXTG, CWBD1, CWBD2, LysM, GW and SLHD proteins (Figure 1.4) (Desvaux *et al.*, 2006). The LysM binding domain is the main interest in this study and is a peptidoglycan binding module.

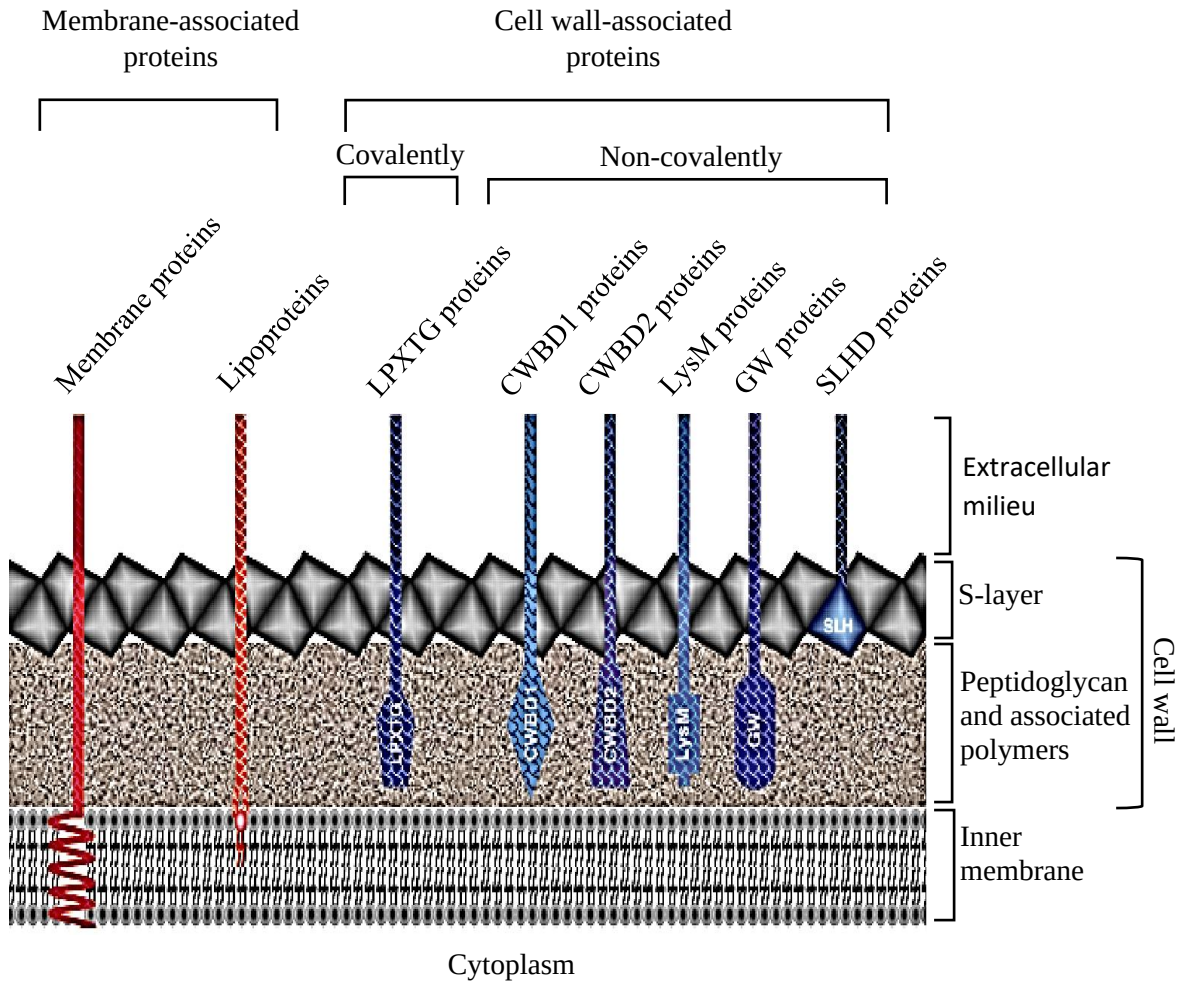


Figure 1.4: Cell wall binding proteins present in the cell envelopes of gram-positive bacteria. The schematic diagram is adopted by Desvaux *et al.*, 2006.

1.1.1 Lysin domain (LysM)

LysM domains are cell wall binding proteins that recognize N-acetylglucosamine, a component of peptidoglycan and chitin (Ohnuma *et al.*, 2008, Petutschnig *et al.*, 2010, Visweswaran *et al.*, 2011). It was first discovered and described from the lysozyme of Bacillus phage Φ29 by Garvey et al. (Garvey *et al*., 1986, Visweswaran *et al.*, 2014). The module is widely distributed in all domains of life (Steen *et al.*, 2003, Buist *et al*., 2008). To date, more than 4600 LysM homologs from species variants are found in the Pfam protein database (PF01476) (Finn et al., 2016).

The LysM domains vary in length from 44 to 65 amino acid residues (Desvaux *et al*., 2006) and protein containing the LysM domain frequently have a number of repeating LysM domains with these being located at either an N-terminal or a middle of two catalytic domains or a C-terminal end of the protein (Ruhland *et al.*, 1993, Bateman & Bycroft, 2000, Steen *et al.*, 2003, Percudani *et al*., 2005, Eckert *et al.*, 2006, Desvaux *et al*., 2006, Buist *et al.*, 2008, Visweswaran *et al*., 2011) (Figure 1.5). Sequence analysis of various LysM homologs of prokaryotes shows that the domains share high sequence similarities at the first 16 and the last 10 amino acid residues of the domain with YTVxxGDTLxxIA and GQxLxxP motifs, respectively (Bateman & Bycroft, 2000, Visweswaran *et al*., 2011, Buist *et al*., 2008).

The LysM domain exhibits a βααβ motif and each of the elements of secondary structure is separated by a loop which varies in length (Figure 1.6) (Buist *et al*., 2008, Visweswaran *et al*., 2014). LysM domains found in eukaryotes frequently possess disulfide bonds. LysM domains of eukaryotes contain conserved CXC motifs, located on the intervening regions of the domains. The conserved cysteine residues serve to link different LysM domains together presumably to increase stability and resistance of the domains to proteolysis activities of plants (Radutoiu *et al*., 2003, Steen *et al.*, 2003, Visweswaran *et al*., 2012, Liu *et al*., 2012, Sánchez *et al*., 2013). Unlike eukaryotic LysM domains, the LysM domains of prokaryotes do not possess disulfide bonds as the stability of the domains are supported by the extensive secondary structures and hydrogen bonding (Buist *et al.*, 2008b, Visweswaran *et al*., 2012).

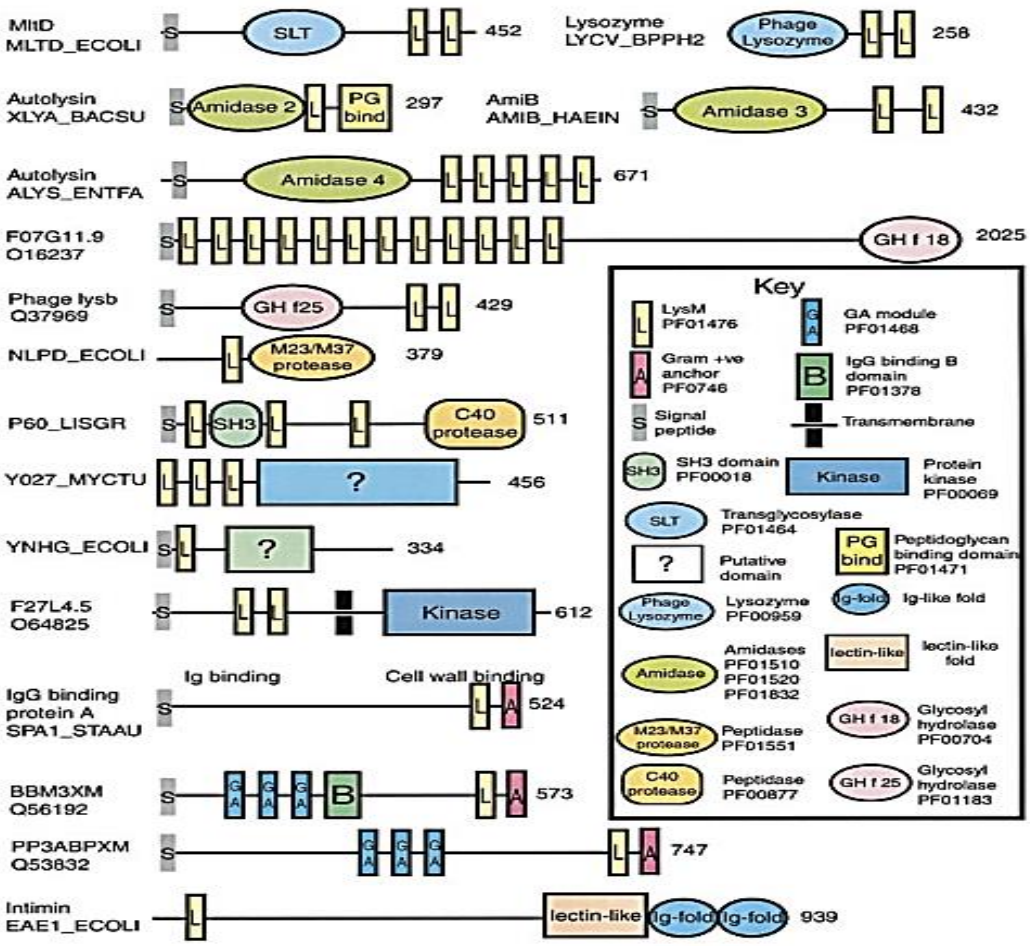


Figure 1.5: A schematic representation of the LysM domain architectures from various prokaryotic proteins. The diagram is taken from Bateman & Bycroft, 2000.

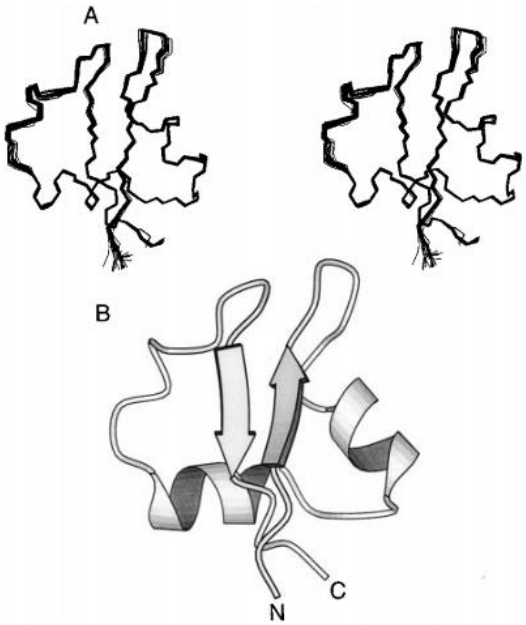


Figure 1.6: LysM domain containing βααβ secondary motif. A) Stereo view of NMR structure of LysM domain. B) Molscript representation of LysM domain. The schematic digram is taken from Kraulis, 2002, and Bateman & Bycroft, 2000.

1.1.1.1 LysM domains from species variants

The LysM domains are crucial in peptidoglycan management. They are the most common cell wall binding domain present in peptidoglycan hydrolases and also in chitinases, transglycosylases, peptidases, esterases, and nucleotidases (Desvaux *et al*., 2006). The LysM domain has also been observed in bacterial lysins (Béliveau *et al*., 1991, Bateman & Bycroft, 2000, Steen *et al.*, 2005, Buist *et al.*, 2008). The LysM domain is commonly fused to a catalytic domain which may play an active role in bacterial cell wall synthesis. As a cell wall binding protein, the LysM domains act to bind to peptidoglycan non covalently to assist catalytic domains anchoring their substrates. The domain is responsible for properly positioning the catalytic domains of the proteins in the vicinity of its substrate and increasing the enzyme concentration in the cell wall (Steen *et al.*, 2005, Buist *et al.*, 2008, Low *et al.*, 2011, Visweswaran *et al.*, 2014).

The LysM domains are one of the most common binding modules in bacterial cell surface proteins (Eckert *et al*., 2006, Dreisbach *et al*., 2011), are thought to be important for the bacterial pathogenesis (Jerse *et al.*, 1990, Ruhland *et al.*, 1993, Lenz *et al.*, 2003). In the example in Staphylococcus Protein A (SpA), the LysM domain is located at the N-terminal end of the protein. This is responsible for attaching the protein to peptidoglycan, while the SPA domain of the protein binds to IgG Fc of the host cell. The interaction between the protein and the IgG hinders phagocytosis (Uhlen *et al*., 1984, Bateman & Bycroft, 2000, Gómez *et al*., 2006).

In plants, LysM domains are involved in the NFR1 protein. In this case, a plant kinase receptor responds by sensing the presence of mycorrhizal fungi or rhizobial bacteria which are important for mediating symbiosis activities among the organisms *(*Radutoiu *et al*., 2003, Gust *et al*., 2012, Tanaka *et al.*, 2012, Miyata *et al.*, 2014, Carotenuto *et al.*, 2017). The LysM-containing proteins are also found in plant chitinases acting as a chitin receptor. These are critical for eliciting plants immunity during infection (Bateman *et al.*, 2000, Gust *et al*., 2012). The chitin receptors are also found in CEBiP, identified in *Oryza sativa* and its homolog, AtCERK1 from *Arabidopsis thaliana*, each of the proteins contains three LysM domains (Liu *et al*., 2012, Sánchez *et al*., 2013, Hayafune *et al*., 2014). In a swamp crayfish, *Procambarus clarkia*, the LysM domain acts as a cell receptor. The receptor senses the presence of an antigen as a pathogen-

associated molecular pattern (PAMP) from *Vibrio anguillarum*. It then activates the immune response to attack the pathogen (Shi et al., 2013).

The LysM domains are also reported present in Ecp6, a cell effector from *Cladosporium fulvum*, a tomato leaf pathogen. This protein acts as a scavenger of chitin fragments where it binds to the fragments. It also conceals it from the host immune system during the invasion and colonization of the pathogen in the host cells (Bolton *et al*., 2008, de Jonge *et al*., 2010, Sánchez *et al*., 2013). The reported proteins containing LysM domains from species variants of prokaryotes and eukaryotes, and the common sites in the cells where the proteins are localized, are listed in Table 1.0 and Figure 1.7.

Table 1.0: List of proteins containing LysM domains responsible for cell growth and pathogenesis in prokaryotes and eukaryotes

Catalytic enzymes	Pfam	Location*
Mannosyl-glycoprotein endo-beta-N-acetylglucosaminidase	PF01832	Sec
N-acetylmuramoyl-L-alanine amidase	PF01520	Sec
Phage lysozyme	PF00959	Sec
Glycosyl hydrolases family 18	PF00704	Sec
Membrane-bound lytic murein transglycosylase D	PF0674	Lipo
Transglycosylase SLT domain	PF01464	Mem
Glycosyl hydrolases family 25	PF01183	Sec
Plant self-incompatibility protein S1	PF05938	
NlpC/P60 family	PF00877	Sec/Lipo
Cystein,Histidine-dependent Aminohydrolases/Peptidase (CHAP)	PF05257	Sec
Peptidase family M23	PF01551	Sec/Lipo/Mem
Zinc carboxypeptidase	PF00246	Sec
Erfk/YbiS/YcfS/YnhG	PF03734	Sec/Lipo/Mem
Bacterial Ig-like domain (group 1)	PF02369	Mem
Bacterial Ig-like domain (group 2)	PF02368	Sec/Mem

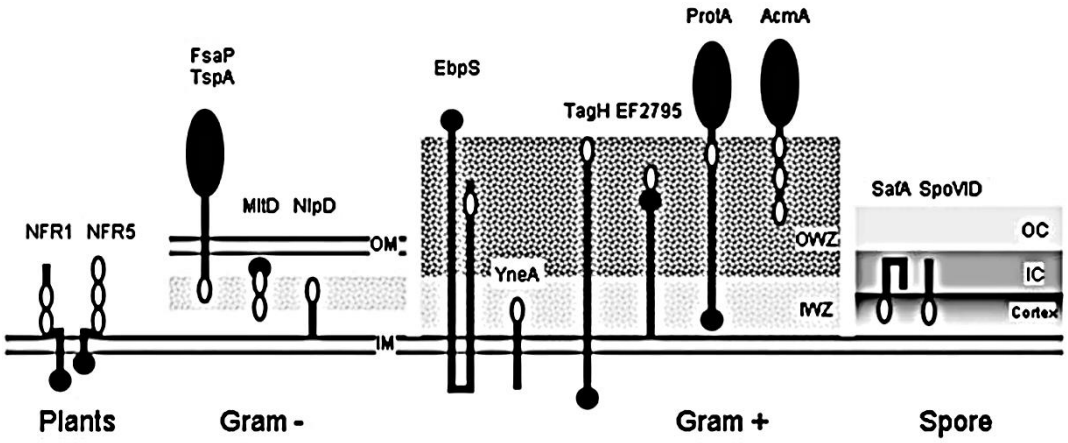


Figure 1.7: Cellular localization of the LysM domain in bacterial cells. OM, outer membrane, IM, inner membrane, OWZ, outer wall zone, IWZ, inner wall zone, OC, outer cortex, IC, inner cortex. The figure is taken from Steen *et al*., 2003.

1.1.1.2 LysM domains properties

The presence of multiple LysM domains in proteins of eukaryotes and prokaryotes remain unclear as one LysM domain is enough to recognize and bind to peptidoglycan such as CVNH from *Magnaporthe oryzae* (PDB ID: 5C8Q) (Koharudin *et al*., 2015). AtCERK1 (PDB ID: 4EBZ), a protein containing three LysM domains from the plant, *A. thaliana* showed that only one (LysM 2) of the LysM domains binds to various chitin fragments (Liu et al., 2012). In fungal effector protein, Ecp6 from *Cladosporium fulvum*, two out of its three LysM domains (LysMs 1 and 3) cooperatively bind to chitin fragments (Sánchez *et al*., 2013) as well as in NlpC from *Thermus thermophiles.* Its two symmetry-related LysM domains cooperatively bind chitohexaose with high binding affinity (Wong et al., 2015). However, in bacterial autolysin, AtlA from *Enterococcus faecalis*, all the six LysM domains of the protein, have been claimed to bind to peptidoglycan (Mesnage et al., 2014).

These findings have raised the question of whether all LysM domains can actually bind oligosaccharides or whether some of them are only responsible for maintaining the stability of the protein. A single LysM domain of AtlA may possibly recognize and bind the substrate. However, the binding affinity between the protein and peptidoglycan increase as the number of the LysM module rose. This observation is similar to the N-acetylglucosaminidase AcmA of *Lactococcus lactis*, a protein containing three LysM

domains, in which the protein shows regression in its enzymatic activities as one of the LysM domains is deleted (Steen *et al.*, 2005). Given that the LysM domains are important in peptidoglycan and chitin recognition, the presence of multiple LysM domains in particular proteins is probably related to the binding affinity of the proteins to it substrates. However, whether the affinity is contributed by the fact that multiple LysM domains cooperatively bind to the same oligosaccharide or each of the LysM domains bind to the oligosaccharide molecules remains unclear and needs to be investigated.

AtCERK1 is a globular protein. Its three LysM domains are tightly packed against each other mediated by three disulfide bridges which link all the LysM domains together (Liu et al., 2012). The structure of Ecp6 has a similar arrangement of the three LysM domains tightly packed to each other through disulfide bridges (Sánchez *et al.*, 2013). However, in the AtlA protein from *E. faecalis*, each of its six LysM domains does not interact with each other either in the absence or in the presence of peptidoglycan (Mesnage et al., 2014). An NMR structure analysis on the AtlA protein showed that the protein behaved in a similar way to beads on a string and did not form any specific quaternary structures for peptidoglycan binding. This tendency contradicts behavior observed in AtCERK1 and Ecp6 as the LysM domains of these proteins make protein contacts between the LysM domains.

The full-length AtCERK1 protein dimerizes in the presence of a long chitin fragment (NAG_8) as shorter fragments cause steric hindrance between the two peptides. The dimerization of the AtCERK1, a cell receptor, is important for cell sensing and signaling in the plant cells (Liu et al., 2012). In bacterial hydrolase protein, NlpC from *Thermus thermophilus* which contains two LysM domains that neighbor to the endopeptidase catalytic domain, the protein dimerization is mediated by either the catalytic domain or oligosaccharides (Wong et al., 2015). In contrast, Ecp6 showed dimerization when the protein is in the presence of chitin ligands (Sánchez *et al.*, 2013). These suggest that protein dimerization is important for regulating essential biological activities in living cells for growth. They are also important for pathogenesis which may include the adhesion of pathogens to infected hosts, cell-signaling, mediation of the immune response, response to bacterial and fungal infections and cell development

and division (Sánchez *et al.*, 2013). LysM domains from species variants have isoelectric points (PI) ranging

from 4 to 12. It has been proposed that these properties are linked to the adaptation of organisms to a wider range of pH at different environmental conditions, for example in the periplasm. (Turner *et al*., 2004, Buist *et al*., 2008a, Visweswaran *et al*., 2014).

1.2 Cell walls of Gram-negative and Gram-positive bacteria and *Mycobacterium* species

The cell wall of Gram-negative bacteria consists of a thin layer of peptidoglycan which is located in between the outer and inner cell membranes. The peptidoglycan layers serve as the substrate for the outer membrane, lipopolysaccharides, lipoproteins and porin proteins in the cell envelope of Gram-negative and Gram-positive bacteria (Neuberger & Deenen, 1994). In the cell envelopes of Gram-negative, the thickness of the peptidoglycan layer is ~7-8 nm (Frederick *et al*., 1990, Malanovic & Lohner, 2016). The thickness of peptidoglycan layers in cell envelopes of *E.coli* and *P. aeruginosa* are reported to be 2.41±0.54 nm and 6.35±0.53 nm respectively (Matias, Al-amoudi, Dubochet, & Beveridge, 2003).

In contrast, the cell envelope of the Gram-positive bacteria does not contain an outer membrane: its cell wall consists of a thick layer of peptidoglycan which is ~40-80 nm (Frederick *et al*., 1990, Malanovic & Lohner, 2016) as well as an inner membrane. This is referred to as a plasma membrane. In addition, the cell wall of Gram-positive bacteria contains two types of teichoic acids which are lipoteichoic acid (LTA) that binds to the inner membrane and wall teichoic acid (WTA) which binds to the surface of the peptidoglycan layers (Walsh, 2017). A schematic diagram of the cell wall of Gram-positive and Gram-negative bacteria is shown in Figure 1.8.

Mycobacterial cell envelopes contain an inner lipid membrane and peptidoglycan-arabinogalactan polymeric network occupying the periplasmic space with covalently linked-mycolic acids (Figure 1.9) (Hett & Rubin, 2008). The outer surface of mycobacterial cell envelopes are capsulated by a mixture of glycans, lipids, and proteins which contribute to the complexity of the Mycobacterial cell wall (Hett & Rubin, 2008, Yao *et al*., 2012). The glycan strands of mycobacterial cell walls contain a mixture of

muramic acid as some of it has an N-glycolyl group rather than an N-acetyl group (Mahapatra et al., 2005). The thickness of peptidoglycan layers of *Mycobacterium* sp is about 4 to 15 nm (Takade et al., 2003). Mycolic acids of the mycobacterial cell wall consist of a variety of short α-alkyl and β-hydroxyl fatty acids in a range of 60-90 carbons per chain. Most of the mycolic acid chains are linked to peptidoglycan by ester bonds (Crick *et al*., 2001).

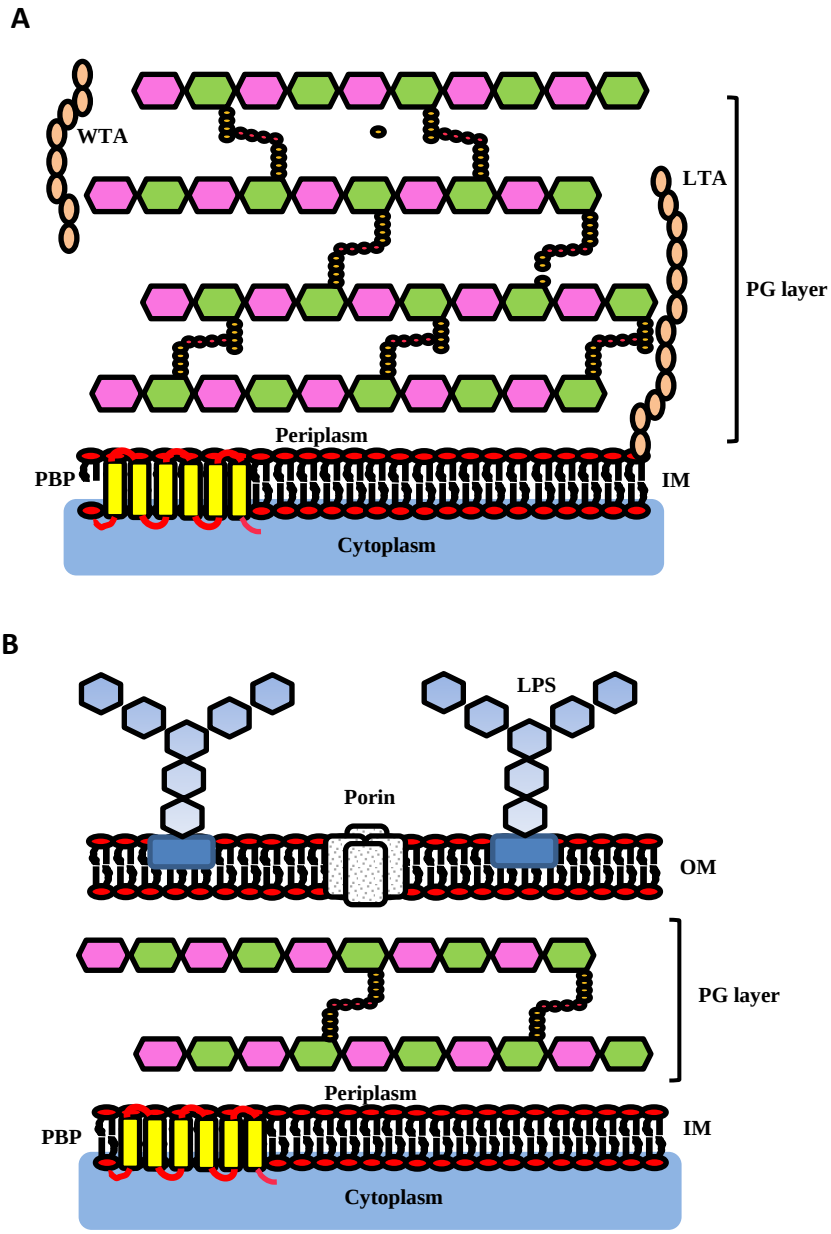


Figure 1.8: Cell envelopes of gram-positive and gram-negative bacteria. A) Gram-positive. B) Gram-negative. LPS: lipopolysaccharide, LTA: lipoteichoic acids, WTA: wall teichoic acids, PG: peptidoglycan, OM: the outer membrane, IM: inner membrane.

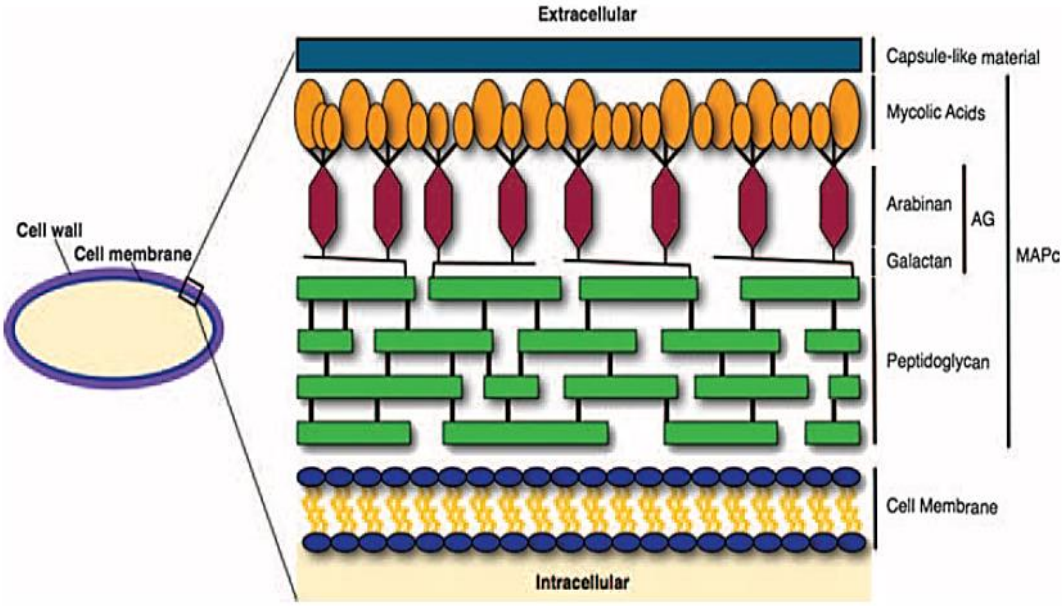


Figure 1.9: A schematic diagram of Mycobacterial cell wall taken from Hett & Rubin, 2008.

1.2.1 Peptidoglycan

Peptidoglycan which is also known as murein is the major component of the bacterial cell envelope for both, gram-negative and gram-positive bacteria. The main structural features of peptidoglycan are a linear oligosaccharide or a glycan strand interlinked by a short peptide (Heijenoort, 2001). The glycan strand is composed of N-acetylmuramic acid (NAM) and N-acetylglucosamine (NAG) linked by β(1-4) glycosidic bonds (Visweswaran *et al.*, 2011) and these repeating NAG-NAM residues serve as the backbones for the peptidoglycan chains (Figure 1.10). In Gram-positive bacteria, such as *Mycobacterium* and *Staphylococcus aureus*, their peptidoglycan is different from the other bacterial peptidoglycan, in which, the peptidoglycan backbone in *Mycobacterium* contains N-glycolylmuramic acid while in *S. aureus*, some of the muramic acids are not acylated (Singleton, 2004).

The short peptide strand of gram-positive bacteria contains amino acids which are L-alanine, D-glutamic acid, L-lysine and D-alanine, covalently attached to the NAM stem molecule. It also cross-links the glycan chains to form mesh-like layer peptidoglycan. In *E. coli*, however, instead of having the D-glutamic acid residue at the position three

in the short peptide, the peptide contains meso-DAP (Walsh & Timothy, 2017, Singleton, 2004). Different bacterial species may exhibit variations in their peptide chains. They are reflected through amidation, hydroxylation and, acetylation cleavage activities on the peptide residues (Vollmer *et al*., 2008).

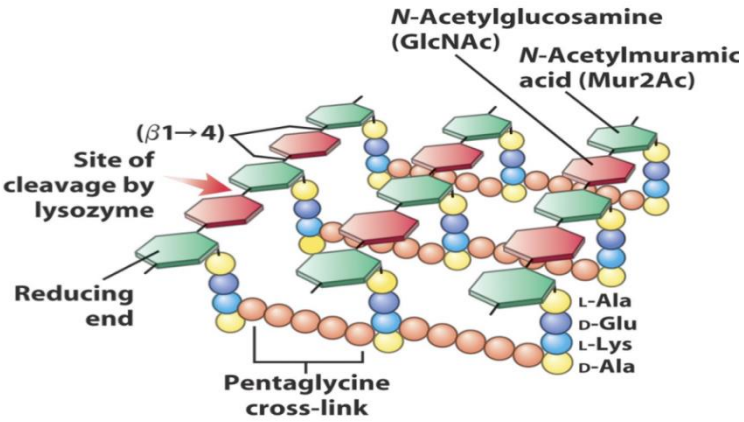


Figure 1.10: A schematic diagram of bacterial peptidoglycan layers in bacterial cell walls taken from DSM website (URL: dsm.com).

1.2.1.1 Glycan strands

The disaccharide units of NAGNAM are elongated to form a glycan strand by transglycosylase and transglycosylase-transferase. The strands are embedded in the bacterial cell envelopes (Heijenoort, 2001, Perlstein *et al*., 2007). The glycan strand is numbered conventionally from the non-reducing end toward the reducing end of the saccharide residues (Numbering, 1983, Varki *et al*., 2017). The disaccharides can be involved in more than two glycosidic linkages to form branched structures (Varki *et al*., 2017).

The length of a glycan chain of Gram-negative bacterial cell wall varies from 20-40 disaccharide units while in the Gram-positive bacilli, it ranges from 50-250 disaccharide units. The extracted glycan strand from *Staphylococcus aureus* shows it contains 18 disaccharide units (Walsh & Timothy, 2017). Lytic transglycosylase is a hydrolytic enzyme which is responsible for the chain-termination of the glycan strand during the peptidoglycan trimming and re-modeling (Lee et al., 2013).

The sugar ring and glycosidic linkage are the main features of the glycan strand. The NAGNAM disaccharide, both have a C_6 sugar ring which is a pyranose ring with the

hydroxyl (OH) groups oriented either above (beta) or below (alpha) the plane of the ring. The carbon atoms of the sugar ring are numbered in a clockwise direction from C1-C6 (Figure 1.11) (Numbering, 1983, Varki *et al.*, 2017). In solution, the pyranose sugar unit tends to form a chair conformation. This creates a stable sugar configuration in which the OH groups exist in equatorial positions (beta conformation) rather than the higher energy boat conformation (Varki *et al.*, 2017). In solution, about 36% of the hydroxyl groups show alpha isomers and 64% are beta isomers (Igor Tvaroska, 1989), and the interconversion occurs when the hemiacetal ring opens and recloses (Varki *et al.*, 2017).

The short peptide strand is crucial to cross-link the glycan chains to build a macromolecular meshwork of peptidoglycan that envelopes the whole exterior of bacterial cells. A transpeptidase enzyme is responsible for catalyzing the formation of an amide bond between the adjacent residues of the peptide strands and the linkage varies between bacterial species (Vollmer *et al.*, 2008, Walsh, 2017).

The orientation of glycan chains to the surface of the bacterium has been initially thought of being parallel (Vollmer et al., 2004, Dirk-Jan Scheffers, 2005). Recently, it has been suggested that the glycan chains are orthogonal to the bacterial surface with the chains composed by three NAG-NAM pairs per helix turn with a right-handed helical conformation (Meroueh et al., 2006).

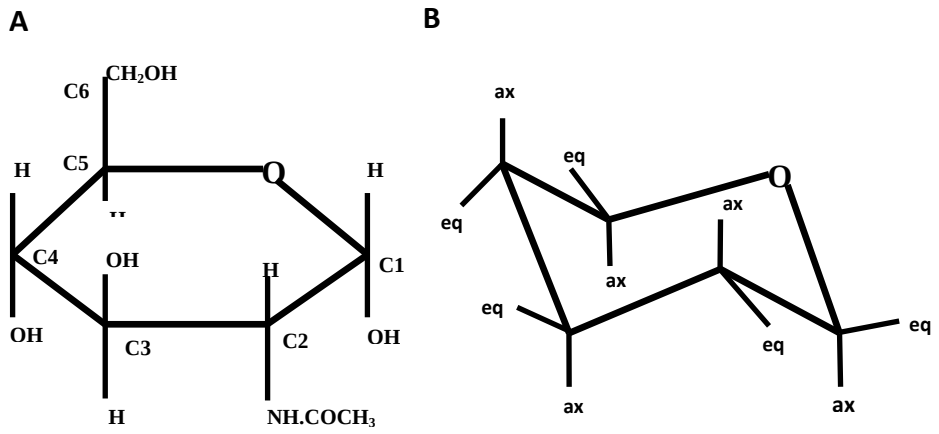


Figure 1.11: A schematic diagram of the pyranose ring conformations. A) A chemical structure of a C_6 pyranose ring with carbon atoms numbered in clockwise order. (B) The C_6 pyranose ring is in chair conformation. The hydroxyl groups (OH) are either at (axial) or (equatorial) positions. The schematic diagrams are adapted from Varki *et al.*, 2017.

1.2.1.2 Glycosidic linkage

The glycosidic linkage involves the anomeric hydroxyl group of the NAG and NAM molecules at C1 and C4 respectively or vice versa and the position/orientation of the glycosylated hydroxyl describes whether it is an alpha (α) or beta (β) linkage (Walsh, 2017). The torsion angle is used to define the orientation of the glycosidic linkage of two linked carbohydrate residues (Imberty, Delage, Bourne, Cambillau, & Pérez, 1991). The two torsion angles, phi (ϕ) and psi (ψ) in which the ϕ angle is determined from the bond of anomeric carbon to the oxygen that joins the two sugar residues (glycosylated oxygen). The ring oxygen of the first sugar residue is used as a reference atom (Numbering, 1983, Imberty *et al*., 1991). Whilst the ψ angle is determined from the bond of the glycosylated oxygen to the carbon (C_4) of the second sugar residue, uses the carbon atom one lower in numbering as a reference atom (Numbering, 1983, Imberty *et al*., 1991). The schematic diagram illustrating the torsion angle is shown in Figure 1.12.

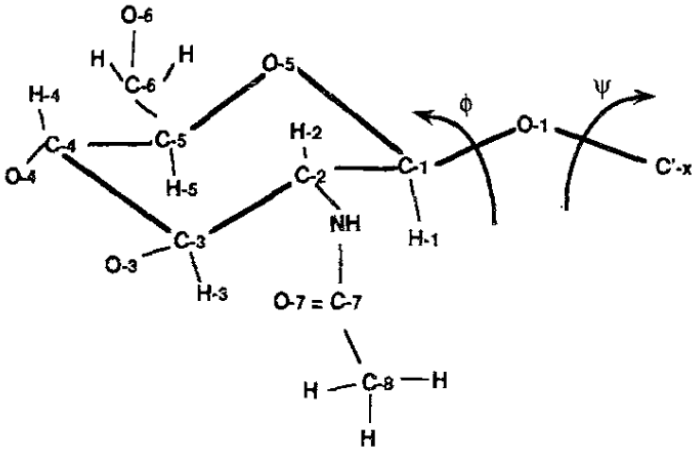


Figure 1.12: Torsion angles of phi (ϕ) and psi (ψ) of N-acetylglucosamine residue. The schematic diagram is taken from Imberty *et al*., 1991.

1.2.2 Peptidoglycan synthesis

Biosynthesis of the peptidoglycan occurs in two steps which primarily initiated in the bacterial cytoplasm then in the periplasm respectively. In the cytoplasm, all the peptidoglycan components including UDP-NAM, UDP-NAG, and tetrapeptide (L-Alanine--D-Glu--L-Lys—D-Ala) are synthesized, followed by the formation of Lipid 1 and Lipid 11 complexes. The final Lipid 11 complex containing NAG-NAM-

phosphate and tetrapeptide is bound to the membrane transporter bactoprenol which also is recognized as a peptidoglycan receptor. The second step in the PG synthesis occurs once the bactoprenol translocates the Lipid 11 across the membrane. The transglycosylase enzyme then catalyzes the formation of glycosidic bonds between NAG-NAM complexes to form new oligosaccharides. This reaction is referred to as glycosylation. In order to cross-link all the oligosaccharide chains, transpeptidation takes place where the transpeptidase enzyme hydrolyzes and re-joins the peptide bonds of the tetrapeptides, to form linked-layers of oligosaccharides. All the new oligosaccharide chains are arranged in layers. They are cross-linked by the tetrapeptide chains bound to the NAM molecules (Healing, 2009). A carboxypeptidase is a protease enzyme that hydrolyzes a peptide bond at the carboxyl-terminal (C-terminal) of the peptide. The schematic diagram of peptidoglycan synthesis is shown is Figure 1.13. The molecular structure of peptidoglycan is shown in Figure 1.14. Hydrolases are the most common enzymes involve in the peptidoglycan management including glycosidase (transglycosidase, lysozyme, lytic transglycosidase), amidase, peptidase (transpeptidase) as well as an esterase. These enzymes are vitally important to maintain the bacterial growth and once the enzymes regress in function, the bacteria lose their capabilities to propagate.

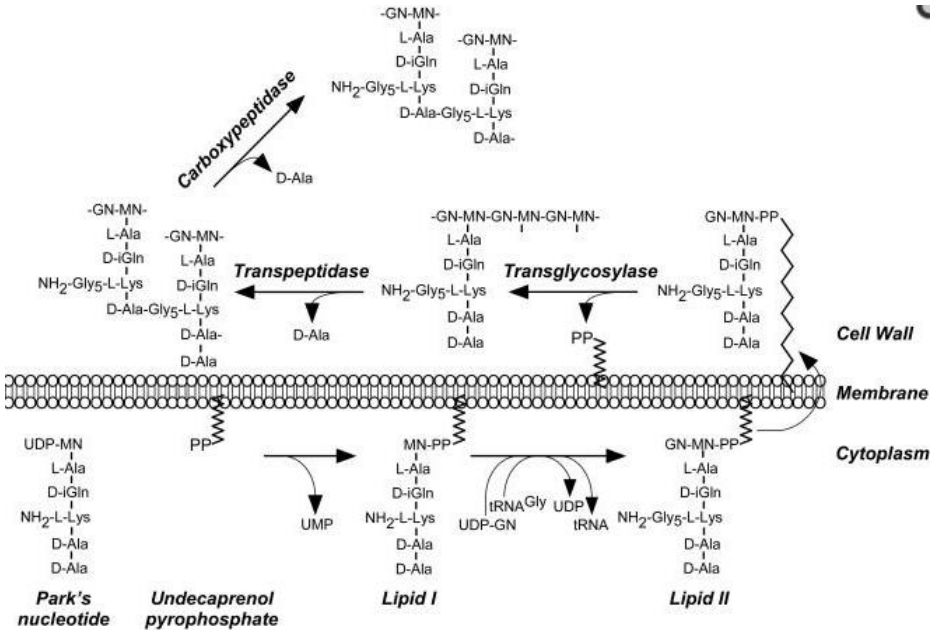


Figure 1.13: A schematic diagram of peptidoglycan synthesis in the bacterial cell. The figure is taken from Lovering *et al*., 2012.

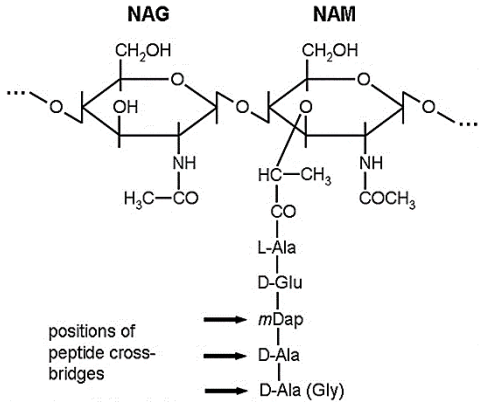


Figure 1.14: A chemical structure of peptidoglycan. Peptidoglycan is composed of repeating disaccharide units of N-acetylglucosamine and N-acetylmuramic acid. The two monosaccharides are linked by a β,1-4 glycosidic bond.

1.3 Aims and objectives

The central aim of this study is to determine the oligosaccharide recognition by multiple LysM domains. The study thus seeks to investigate the way in which multiple domains of LysM are arranged and interact with each other; rule out the function of multiple LysM domains by using X-ray crystallography method and probably to determine binding affinity between the LysM domains and oligosaccharides by using mass spectrometry analysis.

Specific objectives include the following

1. Identification of various proteins containing LysM domains from gram-positive and gram-negative bacteria.
2. All the selected genes will be cloned into suitable expression vectors then expressed in the appropriate competent *E. coli* cells.
3. Optimization of the expression condition will be carried out to enable purification of soluble fractions of the proteins. The proteins may be tagged with six histidines at an N or C terminal region.
4. The purified proteins will be crystallized in apo condition following by co-crystallizing the protein in various polyNAG ligands to produce crystals of the protein with and without an oligosaccharide.

5. Collect high-resolution data for the protein crystals at Diamond synchrotron, Oxford.
6. Determine the structures of the crystals in collaboration with other members in the Crystallography group.
7. Structure solution will be performed using molecular replacement with a known model, collaborated with a person in the crystallography group.
8. Rebuilding structures in Coot and refining them with REFMAC5, in the CCP4 suite (Collaborative Computational Project, Number 4, 2011).
9. Perform structure analyses for apo proteins and proteins in complex with various oligosaccharide substrates using PyMOL and other programs.
10. Perform mass spectrometry analysis on the purified protein in the absence and in the presence of oligosaccharides. The aim is to determine quaternary structures of the respective proteins and to use this approach to investigate the binding affinity between the proteins and the ligands if possible.

CHAPTER TWO

METHODOLOGY

2.0 Background to methods used in the thesis

This section of the chapter describes the principles and theories of the methods applied in the study from cloning, protein expression, protein purification, crystallization, X-ray data collection, structure determination, structure building, structure refinement and structure validation. Most of the figures in this chapter are taken and adapted from four standard textbooks including; '*Analysis of Genes and Genomes'* by Richard J Reece (Wiley, 2004), '*Gene Cloning & DNA Analysis: An Introduction'* by T.A Brown (Willey-Blackwell, 2010), while theories on DNA molecular are from '*Principles of Gene Manipulation'* by S.B Primrose, R.M Twyman and R.W Old (Blackwell Science, 2001) and '*Molecular Genetics of Bacteria'* by Lary Snyder et al. (American Society for Microbiology, 2013).

2.1 Primers

Primers are short oligosaccharides used to amplify genes of interest by PCR. Length of the oligosaccharide sequences should be between 17 to 30 bp. This provides a reasonable annealing temperature to allow the primers to anneal to DNA of the gene sufficiently. The forward and reverse primers ideally have less than 60% of GC content and have approximately the same melting temperature (Tm). The 5' end and 3' end of the oligosaccharides should not complement each other in order to avoid hairpin primer formation. Primers that are designed for a cloning experiment have restriction sites preceded by three to six nonsense nucleotides at the 5' ends of the primers to facilitate efficient cleavage by the restriction enzymes (Figure 2.1).

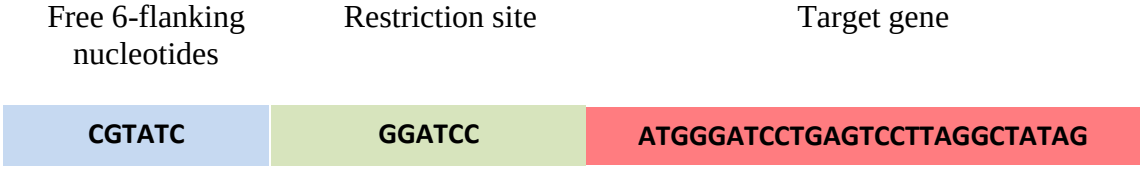


Figure 2.1: Primer design for gene cloning. The primer contains a sequence of a target gene following with sequence for restriction site and preceded by six extra nucleotides.

2.2 Polymerase chain reaction (PCR)

The polymerase chain reaction is a technique to amplify a specific gene in a short time using a thermo-cycler machine. This technique requires a DNA polymerase enzyme to synthesize new DNA copies of the gene of interest through repeating cycles of denaturation, annealing and extension steps. Denaturation and extension cycles are most often performed at 94 ^{o}C and 72 ^{o}C, respectively, while annealing temperature is in range of 50 ^{o}C – 65 ^{o}C. The short oligonucleotides or primers bind to complementary sequences on the strands during the annealing step and DNA polymerase then performs the synthesis of the new strands. The newly synthesized strands are elongated from the 5’ end to the 3’ end during extension cycles, and once the strands are complete, newly double-stranded DNA are produced (Figure 2.2). During the PCR amplification, dNTPs containing the sugar-phosphate as building blocks for DNA strands should be adequately supplied.

$MgCl_2$ is important for PCR reactions as Mg^{2+} ions are required to activate the DNA polymerase. The optimum concentration of magnesium in the PCR solution is 1.5 - 4.0 mM. At a low concentration of magnesium, the DNA polymerase is insufficiently active. The PCR reaction normally fails, while PCR reaction with a higher concentration of magnesium produces unspecific products. The polymerase enzyme possesses proofreading activity is important: it serves to remove any miss-incorporated nucleotides at 3’ end of newly synthesized DNA that might produce mutations into the DNA strands.

2.3 Agarose gel electrophoresis

An agarose gel electrophoresis separates PCR products by applying constant current between 80-100 volts to the system enabling the DNA to migrate from the negative electrode (cathode) to positive electrode (anode) (Figure 2.3). A smaller DNA fragment migrates faster than the larger DNA fragment, and the separated DNA is visualized under ultraviolet light at 260-300 nm.

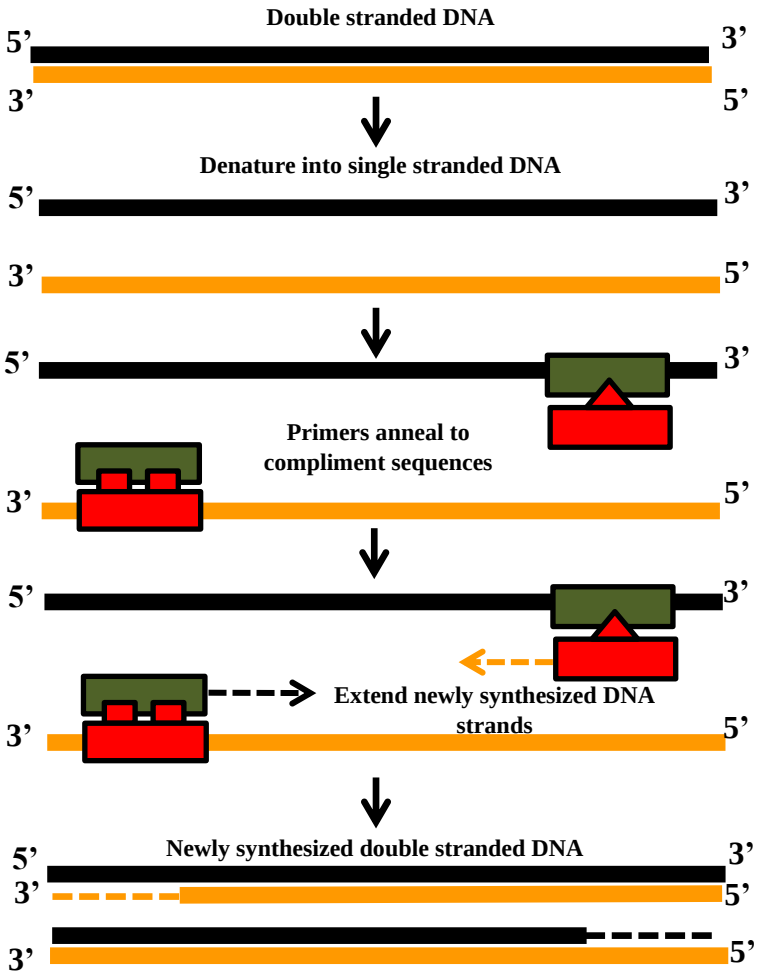


Figure 2.2: Polymerase chain reaction steps. Double-stranded of DNA is unwound during the denaturation step, and single-stranded DNA is produced. During the annealing step, primers bind to the complimented sequence of the DNA strands through hydrogen bonds. A DNA polymerase synthesizes new DNA strands by adding sugar-phosphate in 5' to 3' direction within the extension cycle.

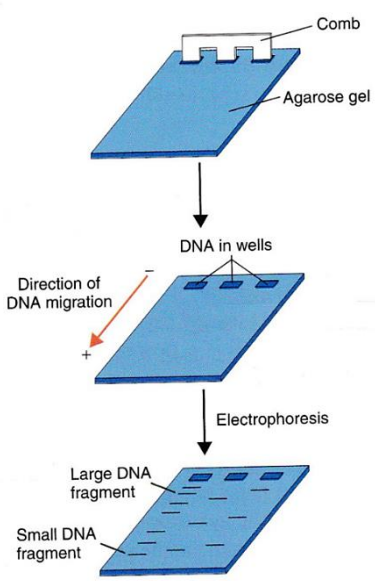


Figure 2.3: An agarose gel electrophoresis. The was taken from Richard J Reece, 2004.

2.4 DNA digestion

DNA digestion is performed by restriction enzymes at specific restriction sites, and the cleaved DNA contains 5'-phosphate and 3'-hydroxyl ends. Two different types of restriction enzymes are available to facilitate DNA digestion. One produces a sticky end product, while the other produces a blunt end product. Single DNA digestion is a method utilizing a single restriction enzyme to cut only one end of DNA, whereas a double DNA digestion applies two different enzymes to cut both ends of the DNA (Figure 2.4). After DNA digestion, the cleaved DNA products are treated with phosphatase enzyme to remove phosphate groups at the 5' and 3' ends of the DNA to avoid undesirable re-ligation products that commonly happens in a plasmid vector.

2.5 DNA Ligation

Joining of two different DNA fragments is called DNA ligation (Figure 2.5). A DNA Ligase enzyme catalyzes the formation of a phosphodiester bond between two linear fragments of DNA through a hydroxyl group at a 3' end and a phosphate group at a 5' end of the DNA strands. The common temperature for DNA ligation is in the range from 4 ^{o}C to 20 ^{o}C.

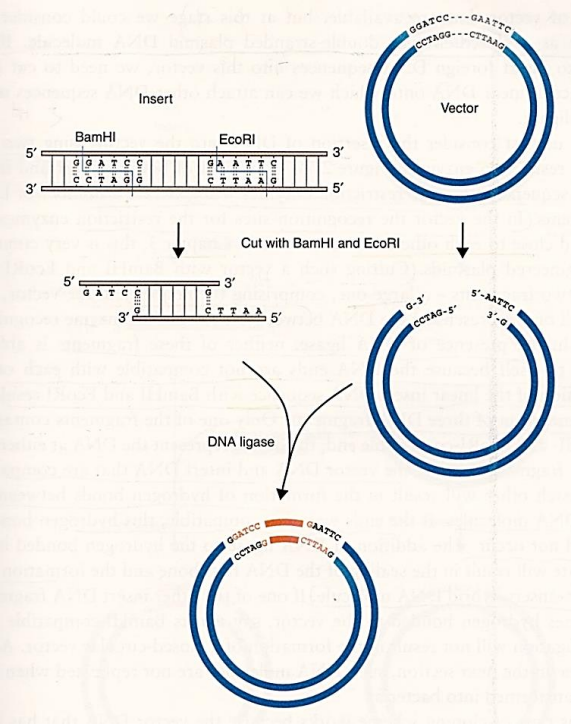


Figure 2.4: DNA double digestion of a gene of interest and plasmid. The figure was taken from Richard J Reece, 2004.

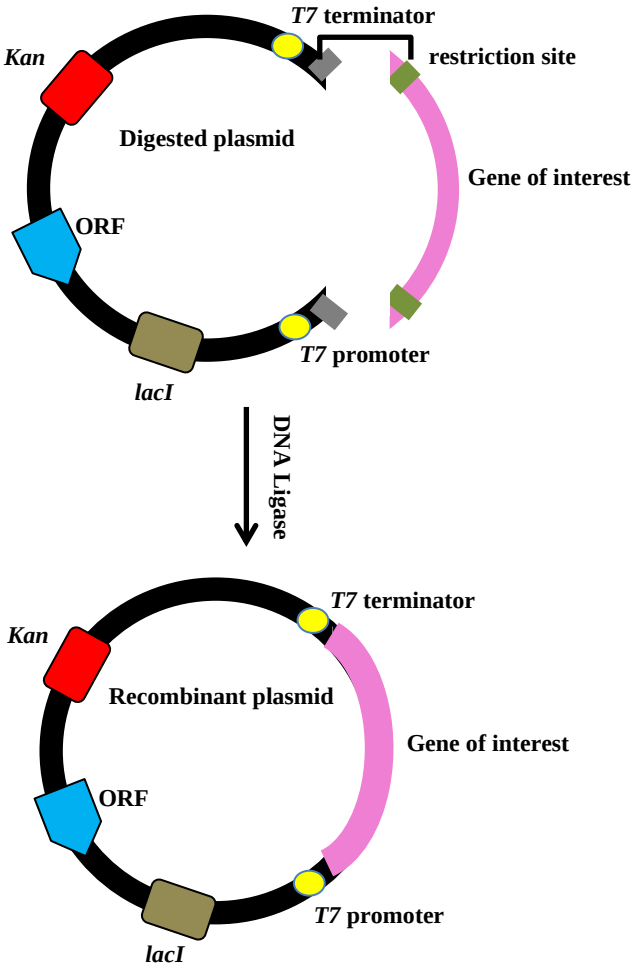


Figure 2.5: DNA ligation. The gene of interest is ligated into a digested plasmid to form a recombinant plasmid. The ligation is catalyzed by the DNA ligase enzyme that responsible to form a phosphodiester bond in between the adjustment DNA sequence of the insert and plasmid.

2.6 Transformation of recombinant plasmid

A principle of transformation is an uptake of naked DNA (plasmid) into bacterial cells and the transformation only occurs at a very low frequency. In 1970, it was found that soaking *E. coli* cells in calcium chloride ($CaCl_2$) solution on ice, enable the cells to take the naked DNA efficiently. The treated cells are called competent *E. coli* cells. Another characteristic of competent cells is that it is deficient in the restriction system to reduce degrading foreign DNA. It may, therefore, increase the efficiency of the transformation.

2.7 Colony PCR

The colony PCR assay is a screening method to verify the presence and orientation of the insert DNA in the recombinant plasmid. The primers used in the assay are universal

T7 primers and the oligonucleotides amplify DNA plasmids from a T7 promoter and end at a T7 terminator DNA sequence.

2.8 Plasmid purification

There are four major steps in extracting and purifying DNA plasmid including bacterial cell lysis, DNA binding, removing contaminants and DNA elution (Figure 2.6). The detergent sodium dodecyl sulfate (SDS) disrupts membrane cells of bacteria and denatures proteins including chromosomal DNA except for plasmids (Richard J Reece, 2004). In the presence of a high concentration of salt, the plasmids are extracted and eluted from other contaminants including cell debris.

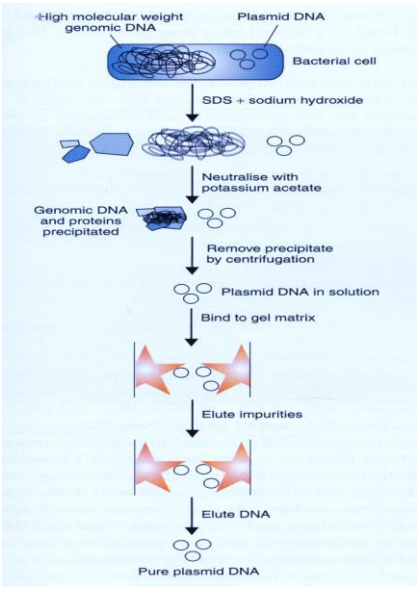


Figure 2.6: Principles of plasmid DNA purification. The illustration was taken from Richard J Reece, 2004.

2.9 Protein over-expression

The gene of interest is expressed in suitable host cells. The most common host cell is *E. coli* BL21 strain *DE3* and the cell uses *lacI* transcription regulator for protein translation. The *DE3* strain contains a *lac* promoter and T7 gene that encodes a T7 RNA polymerase. This is important for initiating the translation of the target genes encoded in the vector (Figure 2.7). An encoded protein of a gene of interest is expressed by inducing host cells by an appropriate amount of the lactose analog, Isopropyl β-D-1-thiogalactopyranoside (IPTG) under specific conditions (i.e: temperature, pH, nutrient and IPTG concentration). The *lacI* gene encodes a lac repressor which binds to a lac

operator and stops transcription and translation of proteins. In the presence of lactose, the repressor dissociates from the operator and an RNA polymerase can freely bind to the T7 lac promoter to initiate transcription of the T7 gene which is then translated into T7 RNA polymerase in the host cells. The T7 promoter of the plasmid vector free from lac repressor is bound by the T7 RNA polymerase acquired from the host cells. The transcription of the gene then occurs followed the translation activities to produce the target proteins.

2.10 Cell induction for protein expression

The induction of the cells serves to express target proteins encoded by genes of interest. The proteins encoded by the genes of interest which are carried by recombinant plasmids are expressed by inducing the host cells (*E.coli*) with IPTG when OD_{600} of the hosts is between 0.6 to 1.0. Optimization of the conditions of the cell growth to express the target protein as soluble fractions folows procedures including lowering growth temperature, increasing incubation time, using various concentration of IPTG and increasing aeration of the growth condition.

2.11 Measurement of protein

Two methods of measuring protein concentration, Bradford and UV assays were applied in this study. The Bradford assay is based on the principle of protein-dye binding in which, basic amino acid residues (lysine, arginine, and histidine) bind to Coomassie Brilliant Blue G-250 dye (Bio-Rad) under an acidic condition and change the dye color from brown to blue. The method is based on the proportional binding of the dye to proteins. As the protein concentration increases, the color of the test samples becomes darker. Coomassie absorbs at 595 nm and a total concentration of the protein can be calculated based on the formula of ((OD_{595}) x 15)/volume (µL). Prior to the measurement of protein concentration, a spectrometer is initially calibrated with standard proteins including Bovine Serum Albumin (BSA). It is issued to set a cut-off value for the assay analysis (i.e: reading of 0.1-0.7 has to be obtained for reliability, Walker, 2002).

The UV assay measures protein concentration based on the absorbance of UV light by aromatic rings of proteins (tryptophan, tyrosine, and phenylalanine) at 280 nm. The

total concentration of proteins can be calculated as (OD_{280}) x dilution factor (lid factor)/calculated extinction coefficient of a protein. Of the method, pH and ionic strength of buffers may affect the absorbance. The extinction coefficient of proteins is calculated using Protparam (Gasteiger et al., 2003).

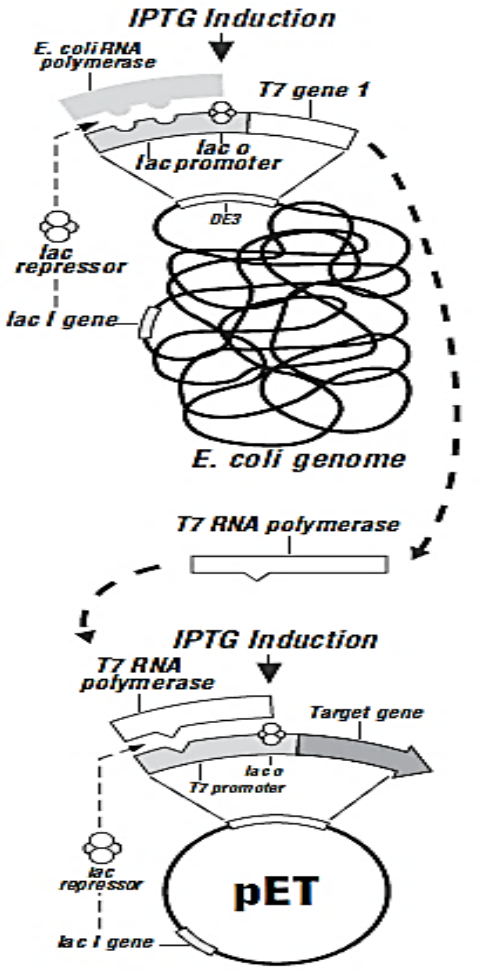


Figure 2.7: An illustration of T*7 lac* promoter in protein expression. The figure was taken from Novagen pET system manual.

2.12 SDS-PAGE electrophoresis

Sodium dodecyl sulfate-polyacrylamide (SDS-PAGE) electrophoresis is a method to separate proteins based on its molecular weight in an electric field. The protein molecules are denatured by SDS detergent to produce SDS-protein complex. The SDS-complex is a negatively charged molecule acquired from SDS ions and the complex migrates from a negative (cathode) to a positive (anode) electrode in the electrophoresis system containing 1x TAE buffer (pH 8.3) for separating the protein on a polyacrylamide gel.

The gel that serves as a medium for the protein separation contains two layers of gel: resolving (bottom layer) and stacking gel (top layer). The stacking gel is a medium to compress proteins. It contains a lower percentage of acrylamide (5%) and a lower pH (6.8) of Tris-HCl buffer. It can achieve larger pores for the gel as well as preparing its lower ionic strength. While the resolving gel, a medium for separating proteins as individual fractions, contains 12% of acrylamide and Tris-HCl buffer at pH 8.8 to decrease the size of the pores and to increase the ionic strength of the gel, respectively. The proteins are compressed and separated at constant current 80 and 200 volts for 15 and 42 minutes, respectively. The SDS-PAGE analyzes a relative molecular weight of the proteins and to determine the solubility and purity of the target proteins.

2.13 Affinity chromatography (Ni-NTA)

Affinity chromatography or Ni-NTA chromatography is the preferred method in protein purification because of its simplicity. It separates histag-proteins by a reversible interaction between a protein and an immobilized metal or ligand including antibodies incorporated into a column matrix. The technique shows high selectivity to the target proteins and capable of recovering abundance of that materials from high levels of contaminating substances.

The four main steps involved in the Ni-NTA chromatography are the application of samples to the column; absorption of the samples to the matrix; washing the unbound protein from the column and the elution by imidazole, a competitive ligand (Figure 2.11). During the elution process, the interaction between the immobilized ligand (Ni^{2+}) and histidine from the protein samples is disrupted by the imidazole. The histag-protein is dissociated from the immobilized ligand bound to the matrix (Figure 2.8). There are two types of immobilized resins (metal ions) incorporated into the column matrix that available for the Affinity chromatography; (1) nickel ion (Ni^{2+}) and (2) cobalt ion (Co^{2+}). Both of the ions bind to histidine but at a different level of binding affinity. The Ni^{2+} ion has a higher binding affinity to the histidine as compared to the Co^{2+} (Figure 2.9).

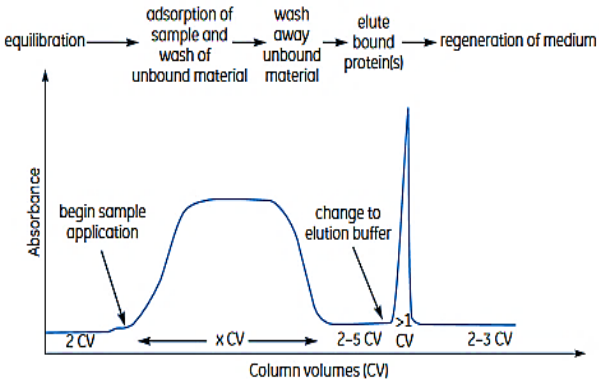


Figure 2.8: Affinity chromatography profiles. The schematic diagram of the Ni-NTA purification system using by AKTA machine taken from Amersham Biosciences Handbook.

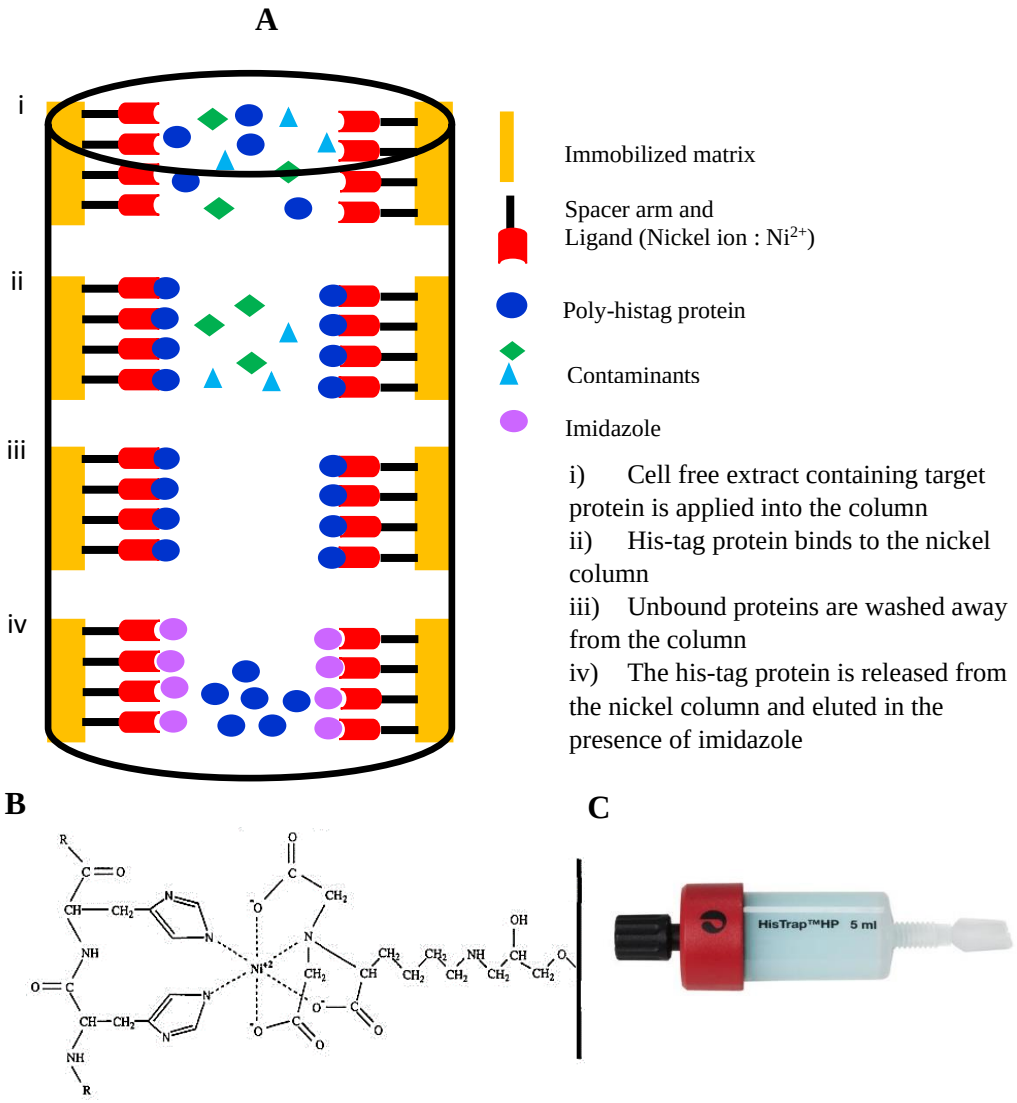


Figure 2.9: Schematic diagram for protein binding, washing and eluting by Affinity Chromatography. A) The poly-histag protein is eluted in the presence of Imidazole. Unbound proteins and any other contaminants are washed away from the column. B) Schematic figure of molecular interaction between histidine and nickel ions Ni^{2+}. C) The HisTrap column used in Ni-NTA chromatography. The illustration is initially adapted from Amersham Biosciences Handbook for protein purification.

2.14 Size exclusion chromatography

Size exclusion chromatography is a liquid-based chromatography that separates proteins based on sizes and compactness. Many types of porous supports or column matrix have been applied in SEC. A cross-linked carbohydrate-based matrix, such as dextran and agarose, are often used for experiments with biological samples such as proteins and nucleic acids. The porous support has an inert surface (inactive). Importantly, it does not react to the injected sample components, and therefore there is no weak or strong mobile phase in the SEC method.

One type of size exclusion chromatography is gel filtration chromatography in which the column used is a gel-based matrix. The mobile phase used in the system is water or aqueous phase. Gel filtration chromatography is applied in protein purification to estimate the molecular weight (apparent MW) of proteins or nucleic acids and for protein separation. In gel filtration chromatography, once the sample is applied to the column, small molecules enter the pores, while larger molecules pass through the column without entering the pores (Figure 2.10 A). The molecules entering the pores are separated. They are categorised according to their retention time or stationary phase. Ideally, when small molecules penetrate the beads, they remain longer in the matrix. They thus require a larger volume of liquid to wash away the molecules from the column.

Estimation of an apparent molecular weight for unknown protein may be performed by calibrating the size-exclusion column with proteins similar to the desired protein in shape and compactness or size with known molecular weight. A calibration plot is used to determine the apparent molecular weight of the desired proteins. It is plotted with logarithms of molecular weights of standard proteins versus their measured retention volumes (K_{av}) (Figure 2.10 B). The K_{av} is calculated based on Equation 1 (Figure 2.10 C). Given a column with the right size of pores, it is crucial for SEC to fully separate the proteins from any contaminants as well as from other proteins. Accordingly, the smaller the bead size, the lower the protein diffusion into the column during elution time. The target protein is then fractionated and more effectively eluted under a sharper peak of the chromatogram (Figure 2.10 D).

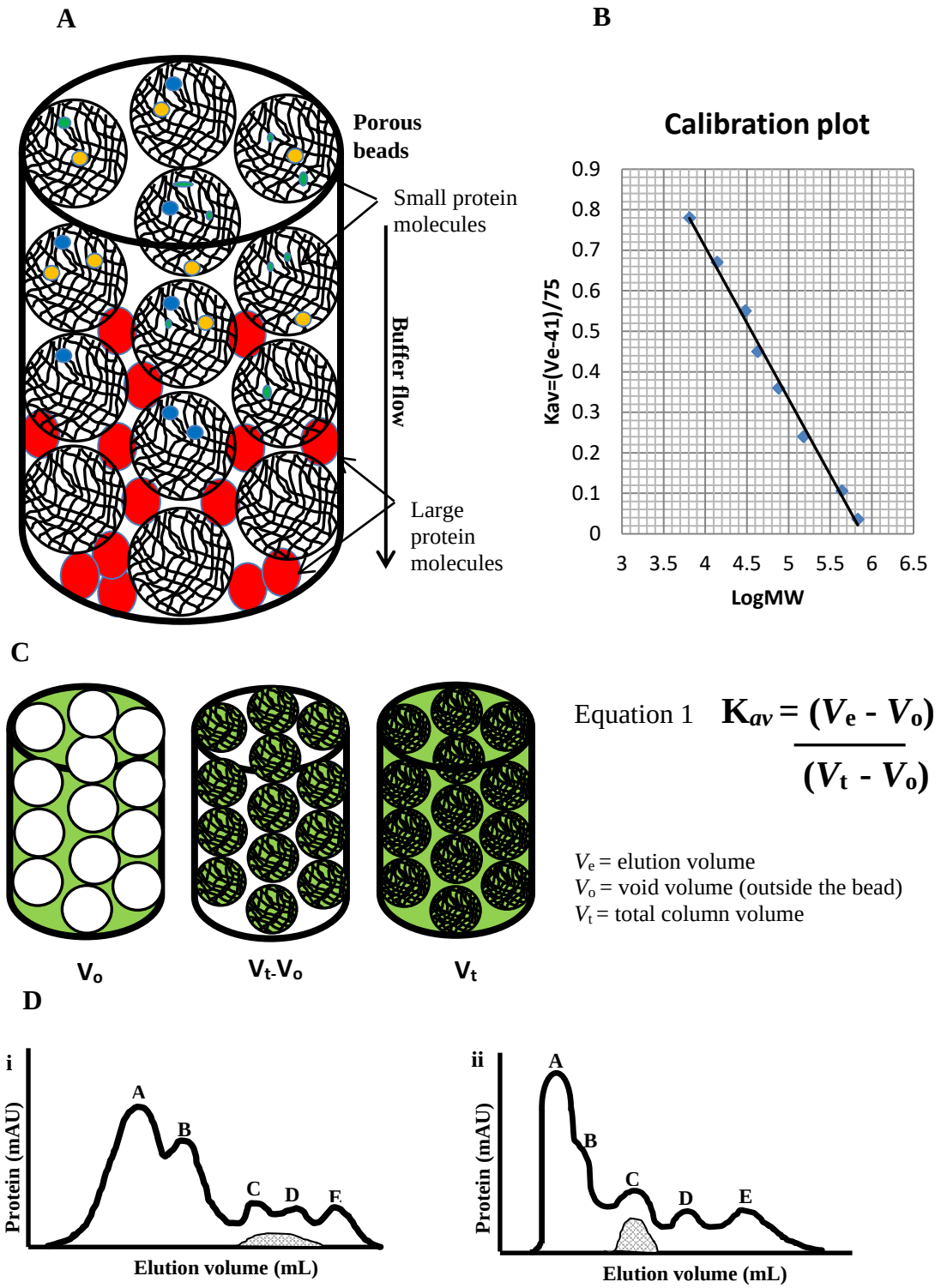


Figure 2.10: The gel filtration chromatography. A) The column used in the gel chromatography separated proteins based on their size and compactness. B) The calibration plot constructed by logarithms molecular weight versus time retention (K_{av}) values of calibrated proteins. The calibration plot is used to estimate an apparent molecular weight of the desired protein. C) The equation for calculating K_{av} values. D) Separation of a single protein from other proteins on the gel filtration column using different pore sizes of beads. The illustrations are initially adapted from Amersham Biosciences Handbook for protein purification.

2.15 Growing protein crystals

Proteins with at least >90% purity are suitable for crystallization trials. Many variables influence the formation of protein crystals including protein purity, protein concentration, precipitant amount, buffer pH, concentration and type of salt, temperature, the freshness of protein as well as the presence of ligands (Rhodes, 1993). The most common method used in protein crystallization is vapor diffusion. The protein solution is gradually concentrated to achieve a supersaturation condition. This allows the protein precipitate or crystallizes at a suitable condition. The solubility of proteins decreases by adding precipitants into protein drops. This is because they bind water molecules. When the water in the drops gradually evaporate, the equilibrium between the drop of protein and reservoir solution is achieved. When the protein approaches its solubility limit (highly supersaturated), it starts to produce a nucleation site. Under a consistent environment, it then grows larger crystals. The most common technique used for protein crystallization is sitting and hanging drop (Figure 2.11).

Derivative crystals are usually obtained by two methods: either co-crystallizing protein with ligands or soaking crystals in mother liquor solution containing ligands. Co-crystallization is an effective method for producing crystals of protein in complexes with the larger size of ligands such as nucleic acids, proteins or sugars (Rhodes, 1993). The soaking method, on the other hand, is utilized to produce heavy-atom derivatives and for micro-seeding crystals. The experimental setup for the seeding is the same as previously described. The difference is that each of the hanging droplets is seeded with a few, small, good quality crystals. (Rhodes, 1993).

2.16 Judging crystal quality

A good quality crystal can give good diffraction patterns that provide useful information about the molecules in the crystals. A brief inspection of the crystals prior to the crystal mounting initially carried out to choose the best crystals for data collection. Physically, under a low-power light microscope, the good-quality crystals have sharp or defined edges, smooth surface, and optical clarity. Under a polarized microscope, the high-quality crystals brighten and darken sharply (Rhodes, 1993). Another method to determine the quality of the crystals is based on its density (Rhodes, 1993). Three useful information sources may be obtained from the crystal density which

is the molecular weight of the protein, the ratio of the protein to numbers of water and number of the protein molecules in the asymmetric unit.

Protein crystals can be discriminated from salt crystals by staining the crystals with a dye such as methylene blue dye. In protein crystals, the dye will diffuse into the crystals through solvent channels. It binds to the protein and thus colors the protein crystals. However, salts crystals do not contain solvent channels then the crystals remain colorless.

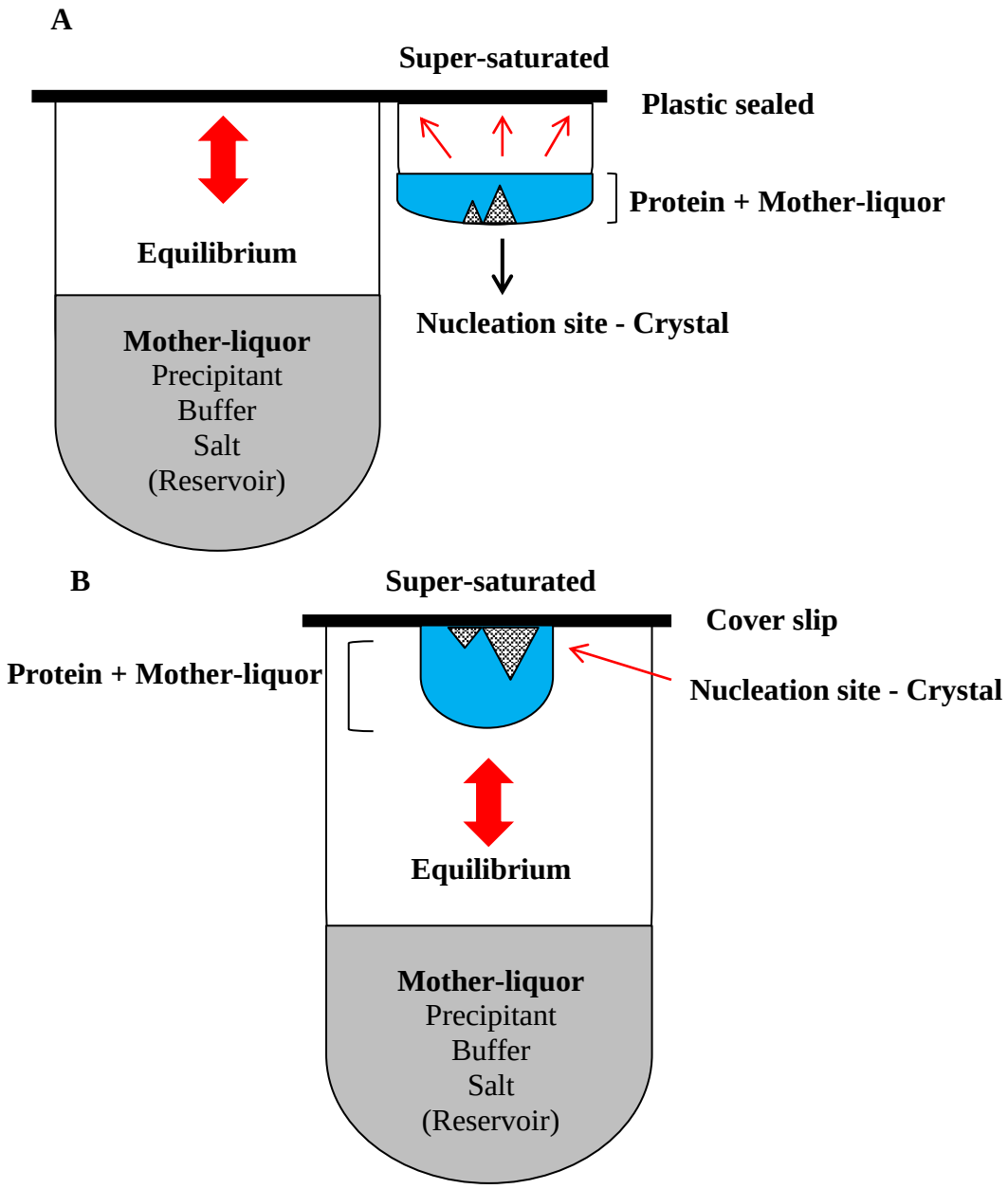


Figure 2.11: Two different crystallization techniques used in growing protein crystals. A) Sitting drop technique. B) Hanging drop technique.

2.17 Crystal mounting

Once good-quality crystals have been grown, the crystals are mounted by cryo-loops from their mother liquors then immediately transferred into their cryoprotectant solution followed by flash cooling the crystals in liquid nitrogen (Figure 2.12). The cryoprotectant solution contains the mother liquor where the crystals grow in addition with 20-25% of ethylene glycol or glycerol solution to preserve the crystals during transfer to the liquid nitrogen. The use of liquid nitrogen is used as a medium for storing and transportation during the data collection. It serves to preserve the morphology of the crystals and reduce the radiation damage to the crystals of X-ray.

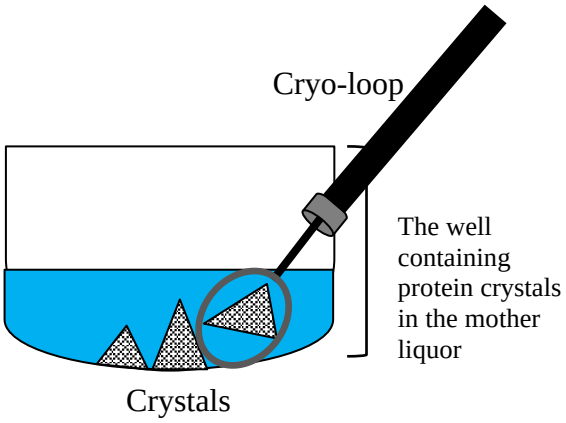


Figure 2.12: Crystal mounting

2.18 Structure determination, building, refinement, and validation

Estimating the number of molecules in the crystallographic asymmetric unit is one of the first steps in a macromolecular structure determination and the estimation is correlated to solvent content (Kantardjieff & Rupp, 2003). The solvent content in the protein crystals is primarily calculated using Matthews equition. It is possible to determine the asymmetric unit of the respective crystals and this analysis is carried out in CCP4 program (Dodson *et al*., 1997).

Most protein crystals contain about 40-60% solvent with the most common value being about 43%, and this estimation value is obtained based on the study of a variety globular proteins (116 proteins) which has a molecular weight of <70 kD (Matthews, 1968,

Kantardjieff & Rupp, 2003). The solvent content is calculated by plotting the volumes of the asymmetric unit (unit cells of the crystals) obtained from the X-ray diffraction measurements of the respective proteins against their molecular weight. The solvent content of a protein is defined as V_m which is crystal volume per unit of protein MW. The value also indicates a fractional volume of solvent in the crystal.

Matthews indicates that protein crystals with higher packing density have lower solvent content and lower V_m value, also known as the Matthews coefficient. This information is important as it serves as a guide in determining the contents of the crystallographic asymmetric unit (AU) (Matthews, 1968, Kantardjieff & Rupp, 2003). This can be an advantage in molecular replacement (MR) as the estimated number of subunits in the AU helps crystallographers conforming the non-crystallographic symmetry. The Matthews Probability calculator is currently available. It is used to estimate a possible number of subunits in the asymmetric crystallographic unit (http://www-structure.IInI.gov/mattprob/).

Structure determination of protein homologs applies MR using PhaserMR and isomorphous MR methods (Rhodes, 1993, McCoy *et al*., 2007). For the MR method, protein homologs with at least 35% sequence identity use a known protein structure as a search model to initially determine their unknown structures. Whilst the isomorphous MR method is applied for the protein structures which have highly similar cell dimension (99%) values. The successful structure determination is followed by model building in real space. This achieved by analyzing the electron density map using *Coot* and refinement with REFMAC5 (Emsley & Cowtan, 2004, Murshudov *et al*., 2011). A structure determination of proteins in complex with ligands follows a similar flow work as mentioned above. A ligand library (Jligand) which is available in *Coot* provides most of the ligand chemical structures used for ligand docking (Lebedev et al., 2012). For unavailable ligands in Jligand library can be imported from the PDB.

Structure validation is performed of the final refined structures as it will provide information about the probability of the model features. The MolProbity tool corresponds to Ramachandran analysis. Accordingly, bond length, angle, chirality, favored side chain rotamer positions and clashes are employed to verify any errors occurred within the refined structure models (Davis *et al*., 2007, Williams *et al*., 2018). Errors identified by MolProbity are then manually fixed by repeating the structure

building and structure refinement. A B-factor analysis is also performed to validate further the refined model structures. The final models of the protein structures are constructed using PyMOL and electron density maps of the structures were generated by FFT (McRee, 1999, Seeliger & de Groot, 2010).

Structure analysis is carried out on the final models of the protein structures by analyzing protein-protein interactions. The structures of the apo proteins are analyzed first and then followed by its complexes with ligands bound. A detailed analysis of the structures is further conducted by performing structure superposition using LSQkab tool in CCP4 and the results are observed in Coot. The final findings of the analysis are then displayed in PyMOL (The CCP4 Suite: Programs for Protein Crystallography, Number 11, Winn *et al*., 2011, Wlodawer *et al*., 2015). Further reading on the theory of X-ray crystallography can be made in standard textbooks. These include; '*Biomolecular crystallography: principles, practice, and applications to structural biology'* by Bernhard Rupp (Garland Science, Taylor & Francis Group, 2010), '*Crystallography made crystal clear'* by Gale Rhodes (Academic Press, 2006) and 'X-ray crystallography' by William Clegg (Oxford University Press, 2015).

2.19 Other methods

Additional techniques employed in this study include mass spectrometry analysis and EM analysis. These are performed in collaboration with respective persons in the Faculty of Science Mass Spectrometry Centre (ChemMS) and in the Electron Microscope Facility, Department of Biology and Molecular Biotechnology, the University of Sheffield.

2.20 Materials and Methods

This chapter describes all consumables, media, cloning and expression vectors, antibiotics, supplement and buffers, genomic DNA of *M. avium* and *M. tuberculosis* used in the experiments. The chapter also describes primer design, protein expression, protein purification, tryptic digest, crystallization, crystal validation and crystal mounting for KEG15107, Rv1288, and Trc1. X-ray data collection, mass spectrometry, and electron microscope analysis. These were all conducted at the Diamond synchrotron, the Mass spectrometry lab in the Faculty of Science, and the Electron microscope lab facilities in the Department of Molecular Biology and Biotechnology.

2.21 Consumables

All chemicals and growth media components including yeast extract, tryptone, and bacteriological agar were obtained from Aldrich, BDH, Glycon, Melford, Merck, Difco, and Oxoid. PCR reagents and DNA marker were purchased from NEB and Bioline. Restriction enzymes, ligase, shrimp alkaline phosphatase (rSAP) were purchased from New England Biolabs, Bioline, Novagen, and Invitrogen, and protein markers (Mark12^{TM}) were purchased from Invitrogen. The QuickChange ^{TM} Site-Directed Mutagenesis kit was purchased from NEB while DNA extraction kits, gel extraction kits, DNA miniprep kits were obtained from QIAgen. Oligonucleotides or primers were purchased from Eurofins Scientific. The crystal screening solutions, cryo-loops, 96-well MRC2 sitting-drop crystallization trays, and clear plastic tissue-culture trays were purchased from Molecular Cell Dimension, Hampton Research, and QIAgen while chitose oligomers (NAG) was purchased from BioSyntech.

2.22 *Escherichia coli* strains and plasmid

Strains of *E. coli* for gene cloning in this study were *E. coli* DH5α and DH5α silver, while *E. coli* BL21 (*DE3*) and BL21 (*DE3*) *placI* were used for gene expression. These strains were purchased from NEB. The bacterial cells were aliquoted into 50 µl batch stored in 1.5 mL centrifuge tubes at -80 ^{o}C to maintain the viability of the cells. The plasmids used in this study were based on the pET expression system, purchased from Novagen.

2.23 Luria-Bertani and SOC media

Luria-Bertani (LB) is a nutrient-rich growth media for propagating bacterial cells. SOC, in contrast, is an optimal media, containing potassium and glucose to recover *E. coli* cells after a heat shock treatment during the transformation of recombinant plasmids into the expression host cells. The LB and SOC media were sterilized at 15 psi for 20 minutes or at 120 ^{o}C for 15 minutes. The recipes for the LB and the SOC media are as in Table 2.1 (A-C).

Table 2.1 (A): A recipe for preparing Luria-Bertani (LB) agar (1L)

Tryptone	10 g
Yeast extract	5 g
Sodium chloride	10 g
Bacteriological agar	15 g
Distilled water	Add to give a final volume of 1L

Table 2.1 (B): A recipe for preparing Luria-Bertani (LB) broth (1L)

Tryptone	10 g
Yeast extract	5 g
Sodium chloride	10 g
Distilled water	Add to give a final volume of 1L

Table 2.1 (C): A recipe for preparing SOC media for 1L

SOB media	
Tryptone	20 g
Yeast extract	5 g
Sodium chloride	0.5 g
Potassium chloride (250 mM)	10 ml
Magnesium chloride (2 M)*	5 ml
Distilled water	Add to give a final volume of 1L
SOC media	
SOB media	980 ml
Glucose (I M)	20 ml

*Add before used

2.24 Antibiotics

Antibiotics are important in cloning experiments. As selectable markers, they help identify bacterial colonies that carrying copies of recombinant plasmids. Proper handling of the antibiotics is vital for successful cloning experiments. Prior to use, the freshly prepared antibiotic solution was filtered using 0.22 μm filters then were aliquot and stored at -20 ^{O}C. Freezing and thawing of the solution were avoided to maintain the efficiency of the antibiotics. Kanamycin and ampicillin are selectable antibiotic markers for pET24d and pET24a, respectively.

2.25 Supplement

Isopropyl-β-D-Thiogalactopyranoside (IPTG) is a lactose analog which is used as an inducer for gene expression in *E. coli* cells. The IPTG solution was freshly prepared in every experiment. Prior to use, the solution was filtered by a 0.22 μm filter.

2.26 Buffers

In this study, all buffer solutions were prepared using Milli-Q water to avoid contaminants from tap water including aluminum, chlorine, and copper. The buffers, especially for protein purification, were sterilized by filtration using a 0.22 μm filter. Recipes for Buffer A and B, 1xTAE and SDS running buffers are as in Table 2.2 (A-D).

Table 2.2 (A): Recipe for 1L Buffer A

Tris-HCl* (1M)	50 mL
NaCl (5M)	100 mL
MilliQ water	850 mL

*pH adjusted to pH 8.0 at 20^{o}C.

Table 2.2 (B): Recipe for 300 mL Buffer B

Buffer A	270 mL
Imidazole (5M)	30 mL

Table 2.2 (C): Recipe for 1L of 50X TAE running buffer

Components	Amount
Tris base	242.0 g
Acetic acid	7.1 mL
EDTA (0.5M, pH8.0)	100 mL

Table 2.2 (D): Recipe for 1L of 10X SDS running buffer

Components	Weight (g)
Tris	144.0
Glycine	30.0
SDS	10.0
MilliQ water	Add to give a final volume of 1L

2.27 Genomic DNA of *M. avium* strain Env 77 and *M. tuberculosis* strain H37Rv and target genes.

Genomic DNA of *M. avium* strain ENV77 was obtained from Prof. Adel Talaat, the Microbiology Laboratory, University of Wisconsin, USA, and details of the genome were obtained from the NCBI database (Hsu Wu and Talaat, 2011). The genomic DNA of *M. tuberculosis* strain H37Rv was obtained from Prof. Jeff Green (Molecular Biology and Biotechnology, the University of Sheffield). Details of the H37Rv genome are available in the Tuberculist or UniProt database.

Two target genes were selected for the study of *KEG15107* and *Rv128*8 with an additional truncated Rv1288 gene, Trc1. The protein encoded by the gene *KEG15107* is a homolog to *MSMEG3288* from *M. smegmatis* that was initially investigating LysM binding domains in the project. This study was carried out to further investigate other LysM homologs from species variants containing multiple LysM domains.

The DNA sequence of the *KEG15107* gene is annotated in the genome at positions 992-1639 (contig 345). Complete DNA sequence of the gene is 648 bp. *Rv1288* is a gene from *M. tuberculosis* found in UniProt database under identification number, P9WM39 or Y1288_MYCTU. The length of the complete DNA sequence of the *Rv1288* gene is

1371 bp, and gene sequences are annotated from position 1441348–1442718. Both of the genes encode uncharacterized proteins but sequence analysis suggests that the gene products contain multiple copies of LysM binding domain. Whilst Trc1 construct is the truncated gene containing only the N-terminal region of the *Rv1288* gene.

2.28 Cloning

This section describes primer design for KEG15107 and Rv1288 from *M. avium* and *M. tuberculosis*, respectively, PCR optimization, DNA digestion, gel extraction, DNA ligation, the transformation of recombinant plasmids into cloning and expression vectors, colony PCR and DNA sequencing for the recombinant genes.

2.28.1 Primers and PCR optimization

Primers are short oligonucleotides. They usually range from 17 to 30 nucleotides and they are used to amplify target genes by PCR. The amplified genes contain restriction sites for enzymes, preceded with six extra nucleotides on 5' end of a forward and a reverse primer to facilitate the gene cloning. The melting temperature (Tm) of the primers should not exceed 72 ^{o}C to avoid non-specific products.

The forward and reverse primers for the *KEG15107* gene amplification were designed based on the gene sequence under the locus tag KEG_RS0114950, extracted from the NCBI database (Table 2.3). The *Rv1288* gene, the primers were designed based on the gene sequence extracted from UniProt (Table 2.4). Another gene construct was Trc1, the truncated *Rv1288* gene containing only a C-terminal region of *Rv1288* which correspond to the three LysM binding domain. Primers for this gene were designed based on the truncated *Rv1288* gene sequence.

PCR conditions were optimized primarily focusing on an annealing temperature as the rest of the conditions including denaturation, elongation, and extension are universal for almost amplified genes in PCR assay. All the PCR samples contained DNA polymerase, dNTPs, magnesium chloride, Tris-HCl buffer (pH8), primers and DNA template. The end PCR products of the amplified genes were subjected to 1% agarose gel electrophoresis, and the results were visualized under UV light (~312 nm).

Table 2.3: DNA sequences of *KEG15107* from *M. avium* strain Env77

992-1639 (Contig345)

GTGAAGACCTACCAAGTCCAGCCGGGCGACACCCTGTTCGCCCTGGCCCGGCGCGAGTACGGTGACAGCA
CCCTGTACCCGGTGATCGCGCGGCAGAACCATCTCGCCAACCCGGATCTGATCGTGTCCGGGCAGCAGCT
GCTGATCCCGTACGTGACCTATCGACACCTGGTCGCCGCGGCCGATTCCACCGCGACCCGCAAGGAGATC
ACCCAGCACTACTACGGAACCGACGACACCAAAGTGCAGTTGATCTGGGAGATCGTCAACGGAGTAGCCC
AGCGGGAGATACAGCAGGGCAGCTGGCTGCACATCCCCGACCTGTCCAACGTCGGGCACCACACGATCGT
CGACGGAGAAAGCCTCGCGGGGCTGGCCGCCCGGTGGTACGGCGACGACCACCTCGCGATCGTGATCGGG
TTGGCGAACAACCTTCCCGCGAACACCGAACCGACCCCGGGCCAGGTGCTCATCGTCCCCGGCCTCAACC
GGCGCCGCCACATCGCCGGCGACACCCTGGTGTCACTGTGCCGCGAGGAATACGGCGATGCGGATCTGGA
CACCCGGACGTCGGTCGTCGCGGCCGCCAACCACATCGGCGAGCCGGCCGCGCTCTTCTCCAACCAGGTG
ATCTATTTCCCCTCCTAA (648 bp)

Table 2.4: Nucleotide sequence of *Rv1288* from *M. tuberculosis* strain H37Rv.

1441348-1442718

ATGGTCAGCACACATGCGGTTGTCGCGGGGGAGACGCTGTCGGCGTTGGCGTTGCGCTTCTATGGCGACG
CGGAACTGTATCGGCTGATCGCCGCCGCCAGCGGGATCGCCGATCCCGACGTCGTCAATGTGGGGCAGCG
GCTGATTATGCCTGACTTCACGCGATACACCGTTGTTGCCGGGGACACGCTGTCGGCGTTGGCGTTGCGC
TTCTATGGCGACGCGGAATTGAATTGGCTGATCGCCGCCGCCAGCGGGATCGCCGATCCCGACGTCGTCA
ATGTGGGGCAGCGGCTGATTATGCCTGACTTCACGCGATACACCGTTGTTGCCGGGGACACGCTGTCGGC
ATTGGCTGCGCGCTTCTATGGCGACGCCTCCCTATATCCGCTTATCGCCGCCGTCAATGGCATCGCCGAT
CCTGGCGTCATCGACGTCGGGCAGGTACTGGTCATATTCATCGGGCGTAGCGACGGGTTCGGCCTAAGGA
TCGTGGACCGCAACGAGAACGATCCCCGCCTGTGGTACTACCGGTTCCAGACCTCCGCGATCGGCTGGAA
CCCCGGAGTCAACGTCCTGCTTCCCGATGACTACCGCACCAGCGGACGCACCTATCCCGTCCTCTACCTG
TTCCACGGCGGCGGCACCGACCAGGATTTCCGCACGTTCGACTTTCTGGGCATCCGCGACCTGACCGCCG
GAAAGCCGATCATCATCGTGATGCCCGACGGCGGGCACGCGGGCTGGTATTCCAACCCGGTCAGCTCGTT
CGTCGGCCCACGGAACTGGGAGACATTCCACATCGCCCAGCTGCTCCCCTGGATCGAGGCGAACTTCCGA
ACCTACGCCGAATACGACGGCCGCGCGGTCGCCGGGTTTTCGATGGGTGGCTTCGGCGCGCTGAAGTACG
CAGCAAAGTACTACGGCCACTTCGCGTCGGCGAGCAGCCACTCCGGACCGGCAAGTCTGCGCCGCGACTT
CGGCCTGGTAGTGCATTGGGCAAACCTGTCCTCGGCGGTGCTGGATCTAGGCGGCGGCACGGTTTACGGC
GCGCCGCTCTGGGACCAAGCTAGGGTCAGCGCCGACAACCCGGTCGAGCGTATCGACAGCTACCGCAACA
AGCGGATCTTCCTGGTCGCCGGCACCAGTCCGGACCCGGCCAACTGGTTCGACAGCGTGAACGAGACCCA
GGTGCTAGCCGGGCAGAGGGAGTTCCGCGAACGCCTCAGCAACGCCGGCATCCCGCATGAATCGCACGAG
GTGCCTGGCGGTCACGTCTTCCGGCCCGACATGTTCCGTCTCGACCTCGACGGCATCGTCGCCCGGCTGC
GCCCCGCGAGCATCGGGGCGGCCGCAGAACGCGCCGATTAG (1371 bp)

2.28.2 DNA digestion of KEG15107, Rv1288, Trc1 and pET24d

The amplified DNA sequence of KEG15107, Rv1288, and Trc1 by PCR and the expression vector of pET24d were subjected to double digestion using compatible restriction enzymes, *NcoI* and *XhoI*. A total volume of 50 µl of digestion mixture was prepared in two separate sterile 1.5 mL tubes and each contained ~1 µg DNA of KEG15107 and ~1 µg DNA pET24d (Table 2.5). The mixture was incubated at 37ºC for 1 hour in a thermocycler machine or in a water bath followed by another five minutes incubation at 75 ºC to deactivate the enzymes. The digested plasmid was then treated with a shrimp alkaline phosphatase (rSAP) solution and the sample was further

incubated at 37 ºC for another an hour. The enzyme was deactivated at 75 ºC for five minutes of incubation. The digested products of KEG15107 and pET24d were subjected to a 1% agarose gel electrophoresis, and the results were visualized under UV light.

DNA digestion on Rv1288 and Trc1 was carried out following similar double digestion protocols as described for the DNA of KEG15107 (Table 2.5). The digested DNA was analyzed using a 12% SDS PAGE electrophoresis.

Table 2.5: A recipe and protocol for DNA digestion of the KEG15107, Rv1288 and Trc1 genes and the plasmid

Insert: *KEG15107*, *Rv1288* and Trc1 (Components)			Vector-pET24d (Components)		
DNA	1 µg		DNA	1 µg	
NcoI (10 U/µl)	2 µl		*NcoI*	2 µl	
XhoI (10 U/µl)	2 µl		*XhoI*	2 µl	
Cutsmart buffer (10x)	5 µl		Cutsmart buffer (10x)	5 µl	
Sterile Milli-Q water	make up to 50 µl		Sterile Milli-Q water	make up to 50 µl	
Incubation	(ºC)	(min)	Incubation	(ºC)	(min)
Incubation 1	37	60	Incubation 1	37	60
Incubation 2	75	5	add 2.0 µl rSAP		
			Incubation 2	37	60
			Incubation 3	75	5

2.28.3 Gel extraction

The digested DNA of *KEG15107* and pET24d were run on a 1% agarose gel and were excised using a scalpel under UV light (~312 nm). The cut gels containing the DNA of the gene inserts and pET24d were subjected to gel extraction (QIAquick® Gel Extraction, QIAgen). The extracted DNA of the gene inserts and pET24d were measured by a nanophotometer at 260 nm wavelength. The DNA was stored at -20 ºC.

2.28.4 Ligation of the target genes into the expression vector pET24d

Ligation is a technique to ligate a foreign gene of interest into a plasmid DNA and DNA ligase is the enzyme that responsible for ligating these two DNA fragments. The insert DNA of *KEG15107, Rv1288* or Trc1 were ligated into pET24d to produce recombinant plasmids containing the target genes. These were then transformed into expression cells for protein expression.

Twenty microliters of ligation mixtures containing DNA of the gene inserts, DNA of pET24d and T4 DNA ligase were prepared at a 1:3 of the insert to vector ratio, following Equation 1. The ligation mixture was incubated either on ice for overnight incubation or two hours incubation at room temperature (19 ^{o}C ±2) and successful ligated products were subjected to 1% agarose gel electrophoresis and the products were visualized under UV light.

$$\text{Equation 1} \quad \frac{\text{Concentration of insert } (\mu g/\mu L)}{\text{Insert size (bp)}} : \frac{\text{Concentration of vector } (\mu g/\mu L)}{\text{Vector size (bp)}} \text{ X3}$$

Table 2.6: A recipe and protocol for a DNA ligation to produce recombinant plasmids containing genes of interest

Components	Amount
T4 DNA Ligase buffer (10X)	2 µL
DNA of plasmid (~5 kb)	Based on DNA stock
DNA of gene inserts (~700 bp)	Based on DNA stock
T4 DNA Ligase (10U/µl)	1 µL
MilliQ water/nuclease free water	makeup to 20 µL
Protocol	**Duration**
Incubation at room température (19±2)	2 hours
Incubation on ice (0-4 ^{o}C)	overnight

2.28.5 Transformation of recombinant plasmids into *E. coli* cells.

Transformation is a method to introduce foreign DNA into bacterial cells. The recombinant plasmids that contained KEG15107, Rv1288 or Trc1 were transformed into *E. coli* DH5α silver cells by a heat-shock treatment. A total of 5 μL ligated product (~10 to ~20 ng) was gently mixed with 50 μL of *E. coli* cells, and the cells were incubated on ice for 30 minutes. The cells were heat-shocked at 42 ^{o}C in a water bath for 40 seconds and immediately transferred onto the ice. The cells were grown in SOC media at 37 ^{o}C, 250 rpm for 2 hours to recover *E. coli* cells from the heat-shocked treatment and to propagate the cells which contained the plasmids. The transformants were then plated on LB agar plates containing 50 μg/mL of kanamycin and the plates were incubated at 37 ^{o}C for overnight. Single colonies grown on the overnight plates were selected and were subjected to a colony PCR.

Table 2.7: A protocol for the transformation of the recombinant plasmid into *E. coli* cells by a heat-shock treatment.

Components	Amount
Plasmid DNA	5.0
Competent *E. coli* cells	50.0
Protocol	**Time**
Incubation 1 on ice	30 min
Heat shock at 42 ^{o}C	40 sec
Incubation 2 on ice	5 min

2.28.6 Colony PCR and DNA sequencing

Colony PCR and DNA sequencing are screening methods used to verify the presence and orientation of the insert DNA in recombinant plasmids. Bacterial colonies that were suspected containing the recombinant plasmids with the target genes were selected and boiled at 100 ^{o}C for 15 minutes in 60 mL Milli-Q water. The crude DNA was subjected to PCR and the recombinant plasmid containing the target genes were amplified by T7 primers (Table 2.8) under an optimized PCR condition (Table 2.9). The successfully

amplified PCR products observed on a 1% agarose gel electrophoresis were sent for sequencing. All the target genes of KE15107, Rv1288, and Trc1 that were ligated into pET24d plasmids were expected to have six extra codons of CAC encoding for six histidine residues as it served as a 6-histag to the target proteins.

Table 2.8: T7 primers for colony PCR

T7 forward primer	5' TAATACGACTCACTATAGG 3'
T7 reverse primer	5' GCTAGTTATTGCTCAGCGG 3'

Table 2.9: A recipe and protocol of PCR for a colony PCR

PCR mixture		
PCR components	Total volume (µL)	
Q5® Hot Start High-Fidelity 2X Master Mix	12.5	
Forward primer (20 pmol)	1.0	
Reverse primer (20 pmol)	1.0	
DNA template (25.0 – 100 ng)	-based on the stock concentration of the DNA	
milliQ water (5% DMSO)	-added to total volume	
Total volume	25.00	
PCR parameters		
Program	Temperature (^{o}C) / Time	
Pre-denaturation	98.0 / 5.0 min	
Denaturation	98.0 / 30 sec	
Annealing	56.0 / 30 sec	repeat the steps for
Extension	72.0 / 30 sec	25 cycles
Post-extension	72.0 / 5.0 min	
Hold	5.0 / ∞	

2.29 Protein expression of KEG15107, Rv1288, and Trc1

Prior to protein expression, the recombinant plasmids containing KEG15107 or Rv1288 or Trc1, confirmed by sequencing, purified by a mini prep kit were transformed into *E. coli* BL21 *DE3* cells. The cells were then plated onto LB plates containing 50 µg/mL of kanamycin as a selectable marker. After overnight incubation at 37 ^{o}C, a single bacterial colony grew on the plate was inoculated into 50 mL LB broth containing kanamycin and the inoculum was incubated at 37 ^{o}C, 250 rpm for overnight. The

overnight culture of no more than 16 hours was commonly selected for protein expression.

2.29.1 Small-scale protein expression

Small-Scale protein expression for KEG15107 or Rv1288 or Trc1 was initially carried out to obtain the best condition for *E. coli* BL21 *DE3* cells to express soluble proteins of the target genes. For this optimization, the experiment was performed in 50 mL LB broth.

Five milliliters of overnight culture (<16 hours) was transferred into 50 mL of LB broth containing 50 µg/mL of kanamycin and the culture was further incubated at 37 ^{o}C, 250 rpm for three hours or the cells OD_{600} reached 0.6. The cells were then induced with Isopropyl β-D-1-thiogalactopyranoside (IPTG) at three different concentrations, 0.1 mM, 0.5 mM, and 1.0 mM, then were further incubated at five different settings of incubation, at 250 rpm (Table 2.10). The cells were harvested by centrifugation.

Table 2.10: Five different incubation settings for a trial protein expression of KEG15107, Rv1288, and Trc1

Conditions	Temperature (^{o} C)	Incubation time (hr)
1	37	3
2	37	5
3	25	5
4	25	24
5	18	72

2.29.2 Harvesting the cells by centrifugation and cells lyses by sonication

E. coli cells from the small-scale of protein expression were harvested by centrifuging the cells at 4 ^{o}C, 10, 000 rpm for 20 minutes and the temperature during the centrifugation did not exceed 8 ^{o}C. Repeat the centrifugation if the supernatant is still cloudy. The supernatant was discarded and only the cell paste was kept at either -20 ^{o}C

or -80 ^{o}C for a short or a long storage respectively. The expressed proteins were obtained by lysing the bacterial cell walls through sonication method.

Prior to the cells lyses, the frozen cell pellets were dissolved in 50 mM Tris (pH8) at a 1:8 of the cells to the buffer (w/v) ratio. The cells were sonicated on ice for three bursts, each for 20 seconds to fully break up the bacterial cell walls to release the target soluble proteins into solution. The soluble proteins were separated from the insoluble fractions by centrifugation at 19,000 rpm or ~45,000g, 4 ^{o}C for 15 minutes. The soluble protein (cell-free extract) concentration was measured by a Bradford assay at 595 nm.

2.29.3 Determination of protein solubility by SDS PAGE analysis

The cell-free extracts containing the KEG15107 or Rv1288 or Trc1 proteins were subjected to a 12% SDS-PAGE electrophoresis to verify the solubility of the expressed protein. Approximately 10-30 μg of the cell-free extract was treated with two microliters of 2-mercaptoethanol (reducing agent) and five microliters of loading buffer (4% SDS, 20% glycerol, 100 mM Tris-Cl (pH 6.8), 2 mM EDTA and 200 mM DTT) with 0.1% of Bromophenol blue dye, followed by sample heating at 100 ^{o}C for two minutes (Table 2.11). Recipes for 12% of resolving and stacking gels of SDS PAGE gel are as in Table 2.12 (A-B).

Table 2.11: Sample preparation for the SDS-PAGE electrophoresis

Reagents	Volume/amount
Protein	10-30 μg
2X Loading buffer with Bromophenol blue dye	5 μL
2-mercaptoethanol	2 μL
Water	makeup to 30 μL

Table 2.12 (A): Recipe for a 12% resolving gel of SDS-PAGE gel

Reagents	Volume
30 % bis acrylamide	2.5 ml
1 M Tris-HCl (pH 8.8)	2.35 ml
MilliQ water	1.28 ml
10 % SDS	62.5 ml

TEMED	6.25 ml
10 % ammonium persulfate (APS)	62.5ml

Table 2.12 (B): Recipe for a 12% stacking gel of SDS-PAGE gel

Reagents	Volume
30 % bis acrylamide	0.63 ml
1 M Tris-HCl (pH 8.8)	0.47 ml
MilliQ water	2.60 ml
10 % SDS	37.5 ml
TEMED	3.75 ml
10 % ammonium persulfate (APS)*	37.5 ml

2.29.4 Large-scale protein expression

Large-Scale protein expression for the KEG15107, Rv1288 and Trc1 proteins were performed on the optimized conditions obtained from the small-scale expression. About 25 mL of overnight cultures (<16 hours) was transferred into 250 mL LB broth containing 50 µg/mL of kanamycin. The cells were induced with 1.0 mM IPTG when it reached 0.6 at OD_{600}. It was further incubated at 18 ^{o}C, 200-250 rpm for 72 hours. The cells were harvested once the incubation completed. The expressed proteins were obtained by lysing the cells through sonication and the cell-free extracts were subjected to a 12% SDS PAGE electrophoresis to determine the solubility of the protein.

2.30 Protein purification

The soluble KEG15107, Rv1288, and Trc1 proteins were initially purified by affinity chromatography (AC) and the proteins were further purified to increase the purity of the proteins by size exclusion chromatography (SEC). For the AC, a Ni-NTA column was used to fractionate and elute the $_{6his}$-KEG15107, $_{6his}$-Rv122 and $_{6his}$-Trc1 proteins. For the SEC, all the target proteins were fractionated and eluted by a HiLoad Superdex 200 pg column.

2.30.1 Affinity chromatography (AC)

Approximately 20 mL (70-120 mg/mL) of the cell-free extract containing the KEG15107 protein was applied into a 5mL HisTrap HP column (GE Healthcare Life Science). Prior to the protein purification, the column was equilibrated by Buffer B (pH 8), containing 500 mM sodium chloride, 50 mM Tris-HCl and 50 mM of imidazole. The KEG15107 protein was fractionated and eluted in Buffer B containing 500 mM imidazole under a linear gradient of imidazole concentration from 0 to 0.35 M (0-70%). The protein fractions were subjected to a 12% SDS PAGE gel to determine the purity of the eluted KEG15107 protein. Under a similar purification protocol and imidazole gradient of Buffer B, the Rv1288 and Trc1 proteins were fractionated and eluted by the HisTrap column accordingly.

2.30.2 Size exclusion chromatography

A 1.6x60 cm HiLoad, Superdex 200pg column was used to further purify the KEG15107, Rv1288 and Trc1 proteins which were primarily eluted by the affinity chromatography. Prior to the purification, a calibration plot (Figure 2.13) was initially developed for the Superdex 200pg column. This was achieved by calibrating the column with eight standard proteins including Thyroglobulin (669 kDa), Ferritin (440 kDa), Aldolase (158 kDa), Conalbumin (75 kDa), Ovalbumin (43 kDa), Carbonic anhydrase (29 kDa), Ribonuclease (13.7 kDa) and Aprotinin (6.5 kDa) (Table 2.13). The calibration plot for the respective column is useful in estimating molecular weights (apparent molecular weights) and quaternary structures of unknown proteins. This is because the plot is constructed based on details of standard proteins (shape and compactness or size) with known molecular weights. The plot compares the logarithms of the molecular weights of the standard proteins versus measured retention volumes (Kav) (Equation 1) (Figure 2.13).

The eluted fractions of the KEG15107 protein obtained from the nickel affinity chromatography were pooled in one tube to be polished by size exclusion chromatography (SEC) using a (1.6x60 cm HiLoad) Superdex 200pg column (GE Healthcare Life Sciences) in Buffer A. The eluted KEG15107 protein from the SEC column was analyzed by SDS-PAGE electrophoresis to determine the purity of the target protein. Given the elution volume (Ve) obtained from the SEC chromatogram

peak for the KEG15107 protein, the Kav was calculated. The value was then plotted on the calibration plot against LogMW of the calibrated proteins to determine the apparent molecular weight of the KEG15107 protein. The Rv1288 and Trc1 proteins were purified using a smilar protocol of the gel filtration to what had been applied for the KEG15107 protein and the purity of the proteins was determined by SDS PAGE gel.

Equation1: $K_{av} = (V_e - V_o)/(V_t - V_o)$

V_e = elution volume
V_o = void volume (outside the bead)
V_t = total column volume

Table 2.13: Eight standard proteins with calculated Kav values.

Protein	Molecular weight kDa (LogMW)	Calculated Kav
Thyroglobulin	669 (5.83)	0.036
Ferritin	440 (5.68)	0.11
Aldolase	158 (5.20)	0.24
Conalbumin	75 (4.88)	0.36
Ovalbumin	43 (4.63)	0.45
Cabonic anhydrase	29 (4.46)	0.55
Ribonuclease	13.7 (4.14)	0.67
Aprotinin	6.5 (3.81)	0.78

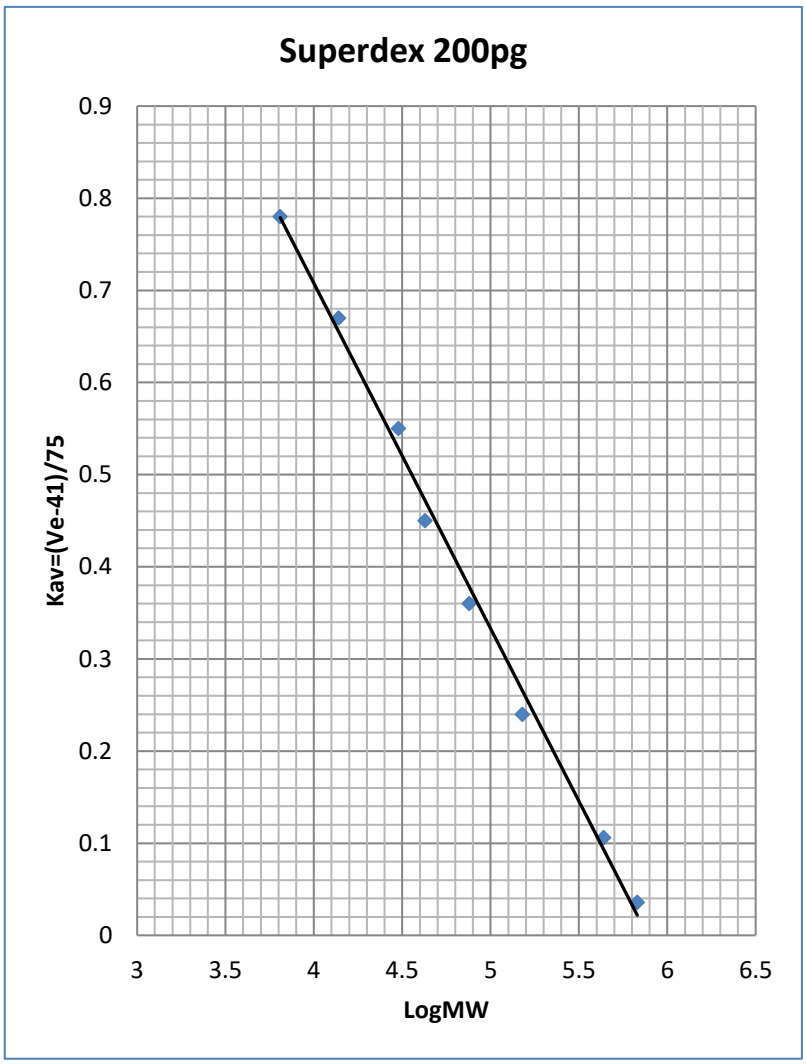


Figure 2.13: Calibration plot for a Superdex 200pg column in Buffer A using the eight standard proteins.

2.31 Tryptic digest analysis

Thirty-five microliters (10 µg) of the freshly purified KEG15107 protein was treated by 35 µL of trypsin (10 µg) at 1:1 of the protein to the enzyme (w/w) ratio. The mixture was incubated at four different incubation settings; 37 ^{o}C for 30 min, 1 hour, 2 hours and overnight (<24 hr). Each of the samples was kept at -20 ^{o}C once the incubation had finished deactivating the trypsin enzyme activity. Controls for the experiments were the untreated KEG15107 protein and the trypsin solution. These were from the same preparation for the experiment and these solutions were kept in -20 ^{o}C. The treated KEG15107 protein with trypsin was subjected to a 12% SDS PAGE electrophoresis together with the controls. Each of the tryptic digest samples was prepared by adding

two microliters of 2-Mercaptoethanol and five microliters of SDS solution into five microliters of the treated proteins (50 µg) and heated at 100 ^{o}C for one minute.

2.32 Crystallization

Prior to crystallization trials, the purified KEG15107 protein eluted from size exclusion chromatography with >95% purity analyzed by SDS PAGE was concentrated to 25 – 40 mg/mL by using a VIVASPIN device (10,000 MWCO, Sartorius) and the protein concentration was measured at 280 nm. The protein was desalted in 10 mM Tris (pH 8.0) by using a Zeba Spin Desalting column (Thermo Scientific) at 1000 g for two minutes and the protein concentration was taken once again. The protein was diluted to several concentrations for crystallization optimization.

Initial crystallization trials were performed on the apo KEG15107 with the protein concentration ranging from 10 mg/ml to 25 mg/ml by using a Matrix Hydra II PlusOne crystallization robot (BioMATRIX). The protein was dispensed into 96-well MRC2 sitting-drop crystallization trays at a 1:1 (v/v) of the protein to crystalline solution ratio, generating a 200 nL drop, which was allowed to equilibrate through vapor diffusion at 19 ^{o}C. Commercially available crystallization screens from Molecular Dimension including Morpheus, $AmSO_4$, JCSG, MPD, PACT, and ProPlex were used to identify conditions that formed crystals. Conditions that showed any crystal growth for the KEG15107 protein at a particular protein concentration were further optimized in order to grow bigger crystals for the protein. This was facilitated by using a hanging drop method in a 24-well tray.

For the hanging drop method, 500 µL reservoir was added into a well and 2 µL drops of protein and crystallization solution (1:1 (v/v)) was applied onto a siliconized coverslip. Optimization was initially performed by applying various amounts of precipitants in the reservoir (i.e: PEG solution) as well as varying pH of buffers and salts.

Further crystallization trials were performed on the KEG15107 protein in complex with its substrate which is oligosaccharides or chitose oligomers including NAG_3, NAG_4, and NAG_5 (Biosyntech). Prior to the crystallization trial, the KEG15107 protein at 25 mg/mL was mixed with the sugars at a 1:2, 1:10 and 1:100 of the protein to sugar molar ratio. Two hundred nanoliters of the KEG15107-NAG complexes were dispensed into

96 wells trays and the trays were sealed tightly before they were incubated at 17 ^{o}C. The trays were observed regularly for any crystal growth.

Crystallization trials for the full-length Rv1288 and truncated Trc1 proteins were performed following a similar sitting drop and hanging drop crystallization protocols as previously applied for the KEG15107 protein. The crystallization trials were initially focussed on apo protein at 1:2, 1:10 and 1:100 of the protein to sugar ratio. Any crystalline conditions suspected to be potential to grow crystals were further optimized using the hanging drop method. Different protein concentrations, various amounts of precipitants, a wider range of pH of buffers and salts concentration were applied during the optimization.

2.33 Crystal validation

The KEG15107 protein crystals grew from various crystallization conditions were run on a 12% SDS PAGE gel. The crystals were scooped from the original reservoir. They were quickly soaked into a freshly prepared cryoprotectant solution containing the mother liquor and 25% ethylene glycol. This washing was repeated three times in separated wells containing the same cryoprotectant solution. The crystals from the last well (third well) were scooped and dissolved in six microliters of Milli-Q water before they were transferred into a 0.2 µL tube. The dissolved KEG15107 crystals were further subjected to the SDS-PAGE gel. The SDS PSE gel containing the protein crystals of KEG15107 was dried out and the gel was sent to the Faculty of Science Mass Spectrometry Centre, the University of Sheffield for MS/MS analysis. The analysis was to verify that the crystals were of KEG15107.

2.34 Structure determination and structure solution of KEG15107

2.34 .1 Data collection and structure determination

The protein crystals of KEG15107 that were suitable for data analysis were harvested from their mother liquor and mounted on cryo-loops at cryogenic temperature. The crystals were protected by a cryo-protectant solution that was freshly prepared and the solution contained mother liquors of the crystals with additional 25% (*v/v*) of ethylene glycol or 20% (*v/v*) glycerol solution to avoid the formation of ice crystals from the

water molecules in the mother liquor during cryo-mounting. The crystals were flash-cooled in liquid nitrogen before being stored in a dry-shipping dewar filled with fresh liquid nitrogen. Accordingly, it was possible to reduce the effects of radiation damage from the ionizing X-ray photons. All the mounted crystals were sent to Diamond Light Source, Didcot, UK for the data collection.

The crystals underwent data collection either on the MX beamlines I04-I, I04 or I24 which are available at Diamond synchrotron, UK. The diffraction data were initially tested by taking 5, 0.1^{o} rotation, diffraction images at 45^{o} interval. In this way, it was possible to ensure the crystals lie in the X-ray beam and determine the quality of the diffractions at different spots of the crystals. It is important to identify clues on how the best diffraction (complete data with high intensities) can be collected for the respective crystals (Dauter, 1999). The tested images of the diffractions were then auto-indexed by MOSFLM (Battye, Kontogiannis, Johnson, Powell, & Leslie, 2011) in the ISPyB SynchWeb interface (Fisher, Levik, Williams, Ashton, & McAuley, 2015) with details of the possible space groups and unit cell dimensions. These can assist decision-making when formulating data collection strategies (Evans & Murshudov, 2013). The collected diffraction data were auto-processed for data integration, scaling and merging, and these were performed either through Fast DP, XIA or XIA2 programs (Winter, 2010, Evans, 2011, Winter and McAuley, 2011). Data with good qualities indicated by quality indicators including R_{merge}, R_{meas}, CC-half, completeness, and multiplicity were used for further analysis.

The apo structure of KEG15107 was initially determined by PhaserMR using the refined model of MSMEG3288 as a search model in a collaboration with Dr. Bisson. The structure of MSMEG3288 was primarily determined by mercury SAD (native crystals soaked with mercury phosphate) in a 2.5 Å experimentally phased map. Subsequently, the structure was used as a search model to determine the higher resolution structure of apo MSMEG3288 by a molecular replacement at 1.6 Å. The space group for the structure was $P2_12_12$ (cell dimension; a=91.2, b=61.9, c=87.0) with two molecules of MSMEG3288 in the asymmetric unit, arranged in a pseudo-2-fold symmetric dimer. The MSMEG3288 protein from *Mycobacterium smegmatis* strain MC2 155 shares approximately 80% of sequence similarity to the KEG15107 protein (Figure 2.14). The structure determination of KEG15107 complexes with polyNAG

substrates was performed using the apo KEG15107 monomer as a search model. The complexes possessing highly similar unit cell values used isomorphous replacement (MIR) for their structure determination.

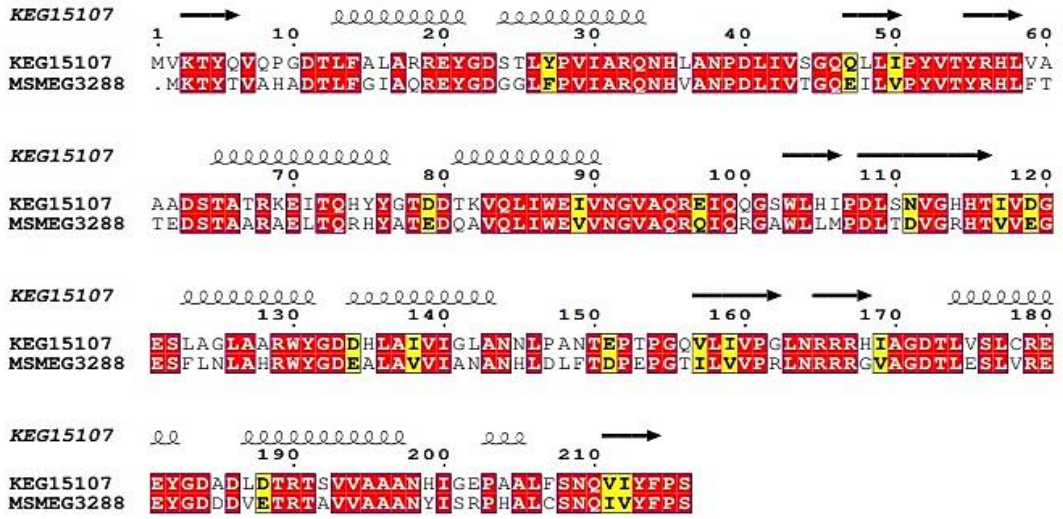


Figure 2.14: Multiple sequence analysis of KEG15107 from *M. avium* against MSMEG3288 from *M. smegmatis*. Protein residues encoded by the KEG15107 and MSMEG3288 genes shared ~80% sequence similarity. Positioned above the sequences are the secondary structures predicted based on the residues of the KEG15107 protein.

2.34.2 Structure building, refinement, and validation

Following the automatic interpretation of the KEG15107 structure in CCP4 using the phaserMR program, the apo structure of KEG15107 was manually rebuilt in *Coot* and refinement on the structure was carried out by REFMAC5. The subsequent solution of the structures of the complexes of KEG15107 with polyNAG substrates was determined by using MR with the monomer apo KEG15107 structure of the protein as a search model. The NAG molecule in which the molecular structure of the ligand was taken from the Jligand library. It was then docked in real space by analyzing the electron density map in *Coot*. The structure was further refined by REFMAC5 to improve the map. The final refined structures of KEG15107 were validated using MolProbity tool, available in *Coot*, and the analysis corresponds to Ramachandran analysis. The other parameters that were validated for the structures include bond length, angle, chirality,

favored side chain rotamer positions and clashes of the atoms. All the errors detected within the structures were manually fixed and refined. These processes were continuously repeated until good statistics generated by REFMAC5 were achieved. The structure refinement was monitored by *R* factors, in which, initial model structures commonly have R factor (0.4-0.5). The values can reduce to 0.2-0.07 for the final refined model structures (X-ray Crystallography by William Clegg, 2015, Oxford University Press).

2.35 Liquid chromatography-mass spectrometry analysis

The liquid chromatography-mass spectrometry (LC-MS) analysis was performed using a combination of LC instruments, Agilent 1260 Infinity, and Agilent 6530 Q-ToF. These instruments applied mobile phase system with two different types of solvent, A (0.1% formic acid) and B (Acetonitrile + 0.1% formic acid). The analytic samples were fractionated in a Phenomenex Aeris Widepore (3.6u XB-C18 50 mm x 2.1 mm) column under gradient (5% to 95%) of solvent B. The total amount of analyte (1.0 µL) was injected by loop injection into the column at a flow rate of 0.4 mL/min, and the analysis of the samples is performed using electrospray ionization (ESI). The experiment was conducted under optimized parameters; Drying Gas temperature 350 ^{o}C, 11 L/min, Nebuliser 45psig, Capillary voltage 4000v. The LC-MS analysis was carried out on studied proteins in a collaboration with Mr. Simon J. Thorpe, in the Faculty of Science Mass Spectrometry Centre (ChemMS), the University of Sheffield,

The Mass spectrometry analysis on the KEG15107, Trc1, and YgaU was performed in the study in order to investigate the proteins in complex with NAG_3, NAG_4, and NAG_5 molecules in the hope to determine an oligosaccharide binding by LysM domains of the respected proteins. The purity of the sugar molecules was initially determined. The experiment was initially carried out for the KEG15107 protein with NAG_5 at a 1:2 of the protein to the sugar ratio. Further analysis on the KEG15107-NAG complexes was performed for the protein with NAG_3 and NAG_4 at the higher sugar concentration (1:50, 1:100 and 1:200 of protein to sugar ratio). For a comparison of oligosaccharide binding by LysM domains, the Trc1 protein was recruited in the experiment. The protein was complexed with NAG_4 substrate at a 1:200 of protein to sugar ratio in the presence of 10 mM Tris (pH 8.0). A similar MS analysis as what had performed on the KEG15107-NAG complexes was applied to the Trc1-NAG_4 complex.

Another protein containing a single LysM domain which is YgaU was also employed in the study. It was used to compare the mode of oligosaccharide binding pattern between different types of LysM. The YgaU protein encoded from *Burkholderia pseudomallei* has a molecular weight 16332 Da is.

2.36 Electron microscope analysis

The purified Rv1288 protein was observed under an electron microscope. A sample for the EM was prepared by dropping a small amount of the protein on the carbon grid. The protein was then stained by negative staining containing with uranyl formate to differentiate the sample from the background. The experiment was carried out at the Electron Microscope Facility, Department of Biology and Molecular Biotechnology, the University of Sheffield (Dr. Svetomir B Tzokov).

CHAPTER THREE

CLONING, EXPRESSION, PURIFICATION, CRYSTALLIZATION, TRYPTIC DIGEST, EM ANALYSIS AND MASS SPECTROMETRY ANALYSIS OF RV1288 FROM *MYCOBACTERIUM TUBERCULOSIS* AND KEG15107 FROM *M. AVIUM*

3.0 Introduction

In this chapter, there are two sections, A and B, in which, Section A is a study on Rv1288 from *Mycobacterium tuberculosis* while Section B is a study on KEG15107 from *M. avium*. This section discusses the works on Rv1288.

M. tuberculosis is a pathogen that causes severe communicable tuberculosis disease which significantly affects the mortality and morbidity rates among global populations. Approximately 180 000 cases of multidrug-resistant TB and rifampicin-resistant TB were notified and relapse TB cases among HIV patients tremendously increased from 2004 to 2017 (WHO, 2018). Tuberculosis cases among elderly (>65 y) are the highest as compared to the other groups and high incidence of the disease is observed among healthcare workers (WHO, 2018). Latent TB is another significant contributor to the disease development in infected persons as it remains challenging for identifying and diagnosing the disease at an early stage of infections. Poorly understood biology of *Mycobacterium tuberculosis*, pathogenesis of the pathogen, undiagnosed primary and relapse cases as well as latent TB, limited anti-TB drugs, the emergence of XDR-TB had worsened TB treatments as it limits anti TB drugs and besides that, asymptomatic TB condition, long dormancy stage of causative organism *Mycobacterium tuberculosis* and co-infection with HIV further complicated the TB infection management (Ford *et al.*, 2016, Vilaplana *et al.*, 2017, Maan *et al.*, 2018). Therefore, under these circumstances, identification of new drug target for developing new anti TB drugs are urgently needed.

A unique feature of *M. tuberculosis* is its high lipid-content cell wall. In every stage of TB infection, it was suggested that the pathogen modulates different types of lipids to

stay persistent to the host and therefore, any macromolecules that gets involved in the lipid management of the *M. tuberculosis* are important (Queiroz & Riley, 2017). Active studies on proteins which relates to lipids and lipid metabolism of *M. tuberculosis* has been conducted in many labs in many countries and one of the investigated protein was Rv1288 (Maan et al., 2018). The Rv1288 gene is found in the genome of *Mycobacterium tuberculosis* strain H37Rv (UniProt database under ID (P9WM39) or (Y1288_MYCTU)). The length of the complete DNA sequence of the Rv1288 gene is 1371 bp from position 1441348 – 1442718 in the genome. The gene encodes an uncharacterized protein, Rv1288 with 456 amino acids. Sequence analysis shows that the N-terminal region of Rv1288 contains three LysM domains each of approximately 50 residues (Figure 3.0). The C-terminal domain of Rv1288, with 300 amino acid residues, is annotated in the genome as a putative esterase domain.

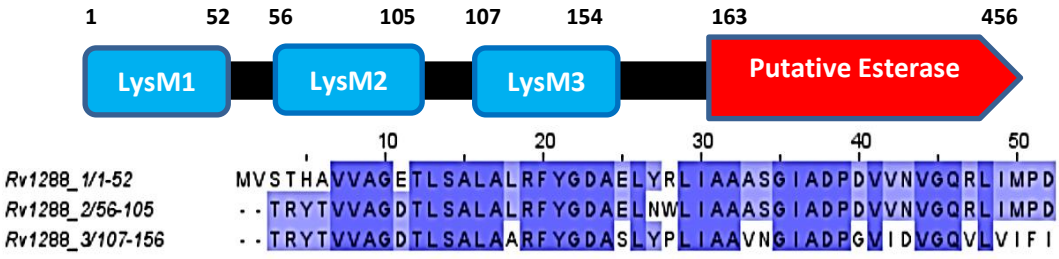


Figure 3.0: Domain structure of Rv1288 from *M. tuberculosis* strain H37Rv. A) The Rv1288 gene contains three tandem repeats of a LysM domain covalently linked to a putative esterase domain. B) Multiple sequence alignment of the three LysM domains of Rv1288.

The putative esterase domain of Rv1288 could potentially be involved in *M. tuberculosis* cell wall organization as many studies have reported that LysM is a peptidoglycan binding module. Also, the presence of multiple LysM modules in the Rv1288 protein raises the question as to why the protein needs a tandem repeat of three LysM domain as in the structure of AtCERK1 from *Arabidopsis thaliana* (Liu et al., 2012)? Does the presence of multiple domains of LysM stabilize the quaternary structure of the protein? Do the tandem repeats of LysM modules increase the binding affinity to peptidoglycan as suggested for the fungal protein Ecp6 from *Cladosporium fulvum* in which two out of its three LysM domains cooperate to provide higher binding affinity towards chitin (Sánchez-Vallet et al., 2013)? How are the LysM modules arranged? Do they have a globular arrangement in which each of the LysM domains tightly packed against each other as shown in the AtCERK1 structure, or, do they

behave like beads on a string as suggested in AtlA from *Enterococcus faecalis* (Mesnage et al., 2014)? Clearly, a three-dimensional structure of Rv1288 could provide an answer to some of these questions and might provide new insights in understanding the architecture of the *M. tuberculosis* cell wall which could generate leads for antibiotics discovery against tuberculosis.

In this study, two gene constructs were designed to overexpress full-length protein and a truncated variant of Rv1288 containing the three LysM tandem LysM domains alone, Trc1. The idea of the two different constructs of Rv1288 included in the study was to determine further how the multiple LysM domains pack to each other in the presence and the absence of a catalytic domain.

3.1 Analysis of the putative esterase domain of Rv1288

The sequence for the C-terminal domain of Rv1288 from *M. tuberculosis* was analyzed against all available sequences using BLAST. This showed that close homologs (>83% sequence identity) could only be identified in *Mycobacterium* species including *M. cannetti* (CCC63907.1), *M. bovis* (CCC63907.1, *M. abcessus* (CAM64874.1), and *M. smegmatis* (ABK74318.1), followed by *Saccharomonospora viridis* (ACU95748.1) (76% sequence identity) and *Actinomycetes sp* (EGF50189.1) (63% sequence identity). All the related genes from those species are annotated as putative esterases in their genome sequences. More distant homologs (<39% sequence identity) are found in *Streptomyces venezuelae* (CCA60450.1), *Burkholderia cenocepacia* (CAR56217.1), and *Mycococcus xanthus* (ABF89705.1) and the homologs also are assigned as putative esterases in the genome database (Figure 3.1). The protein sequence analysis showed that the C-terminal domain of Rv1288 shares a number of strongly conserved motifs (shaded boxes in Figure 3.1) including conserved motifs and catalytic triads colored in red, reflecting their similar protein folding and possibly their similar biological function.

A secondary structure analysis on Rv1288 showed the presence of three repetitive βααβ motifs on the N-terminal region of Rv1288, observed from residue methionine (M1) to residue isoleucine (I154), suggesting that the N-terminal Rv1288 protein contains three

LysM domains (Figure 3.2), followed by the catalytic domain of the protein from residues $D_{158}G_{159}$ to the last residue D_{456}. The absence of any low complexity sequence between the end of the third LysM domain and the first motif of the putative esterase domain suggests that these domains may be separated by a short linker ($G_{155}R_{156}S_{157}$). As LysM domains function to bind to peptidoglycan layers and in other cases are thought to assist the catalytic domains to position themselves in the vicinity of their substrates, this raises questions as to what is the biological relationship between the LysM domain and the putative catalytic esterase?

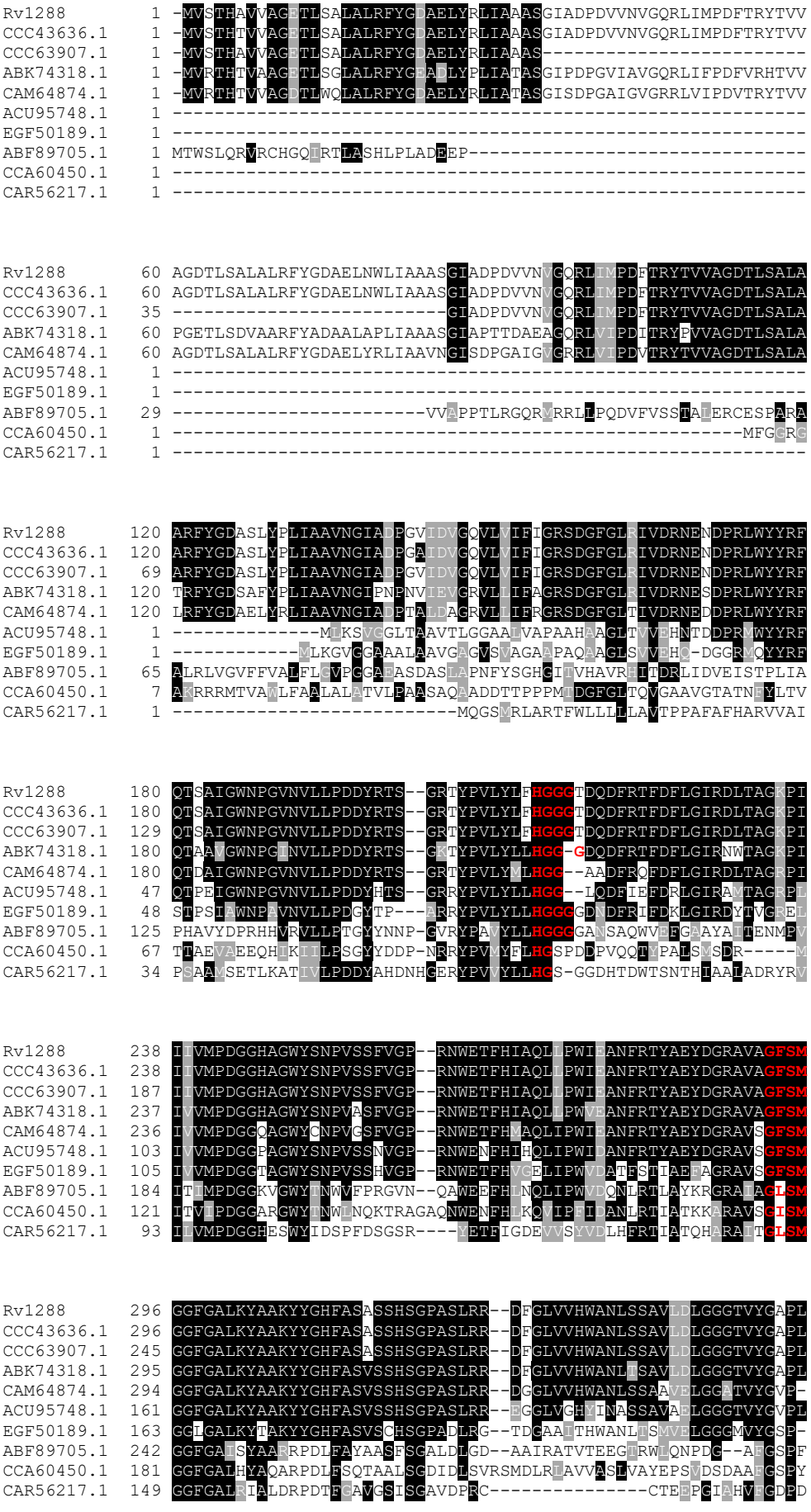
The key feature of the alpha/beta hydrolases fold of the esterase domain is the presence of eight beta strands for the protein core while five or six helices surround the strands (Heikinheimo *et al.*, 1999, Siew *et al*., 2005). Different types of hydrolases with this alpha/beta fold include carboxylic acid ester hydrolases, lipid hydrolases, haloperoxidases, and dehalogenases. Though these enzymes share low sequence similarity, their three-dimensional structures are highly conserved (Siew *et al*., 2005). Despite having a similar fold, the putative alpha/beta hydrolases often have large sequence insertions after the sixth beta-strand (β6) of the protein, and this feature is important for substrate recognition and specificity.

A key feature of the activity of numbers of the alpha/beta hydrolases superfamily is a catalytic triad which is involved in the mechanism of hydrolases with a serine as a nucleophile of the enzyme followed by aspartic acid/glutamic acid and histidine residues. The serine is located at the beginning of the short loop between β5 and α3, while the conserved histidine is located after β8. The acid residue of the catalytic triad which is either aspartic acid or glutamic acid in most hydrolases is generally found in a loop between β7 and α5. (Siew et al., 2005).

To further investigate the C-terminal domain of Rv1288, a structure-based analysis on four hydrolases obtained from the PDB including the aryl esterase from *Pseudomonas fluorescens* (1VA4), thermophilic esterase from *Archeaoglobus fulgidus* (5FRD), dipeptidyl amino peptidase IV from *Stenotrophomonas maltophilia* (2ECF) and chloroperoxidase from *P. fluorescens* (1A8S) was carried out first to examine the protein fold similarity and position of the catalytic triad before they were further used as a reference for analyzing the putative esterase domain of Rv1288. The analysis showed that the three-dimensional structures of the aryl-esterase, thermophilic esterase,

dipeptidyl amino peptidase, and chloroperoxidase were conserved including conservation of their catalytic triads (Figure 3.3). All the hydrolases shared an identical serine, aspartate acid and histidine at the respective positions and these catalytic triads form active sites for the enzymes.

Sequence alignment focussing on the putative esterase domain of Rv1288 was performed against the structure-based sequences of those four respective hydrolases. The analysis (Figure 3.4) showed that the C-terminal domain of Rv1288 has a conserved serine (S295) from motif 1 (GFSMGG) which is likely to be the nucleophilic serine for the protein. Possible catalytic residues are aspartate acid (D392) from motif 2 (DSVNE) and histidine (H426) from motif 3 (PGGH) suggested that the C-terminal domain of Rv1288 might possess an esterase function (Maan *et al*., 2018). Therefore, a structural study on the C-terminal domain of Rv1288 should be carried out to validate the esterase domain properties to shed light on its biological function.



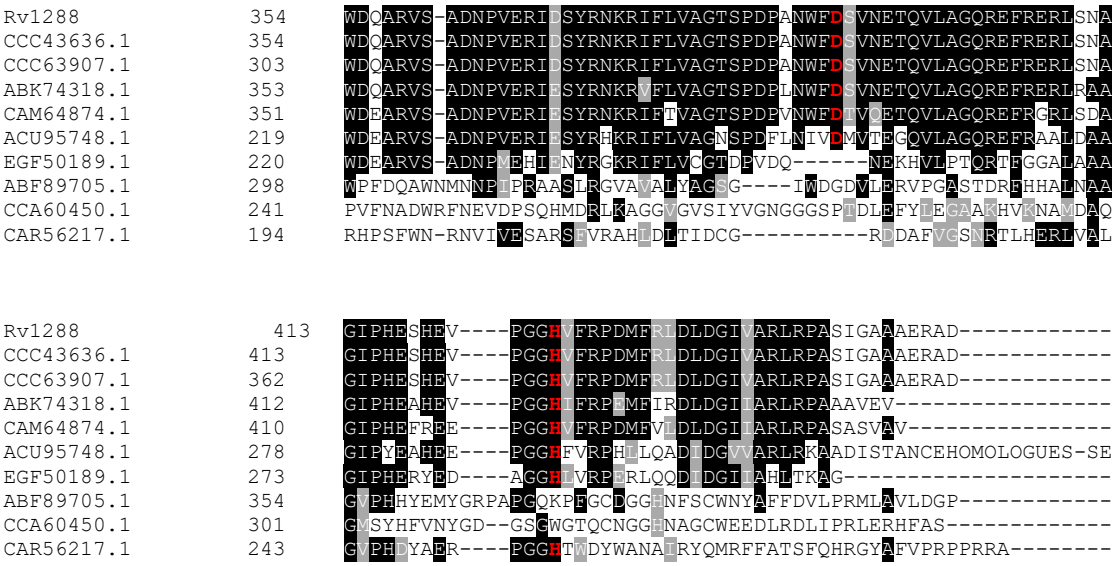
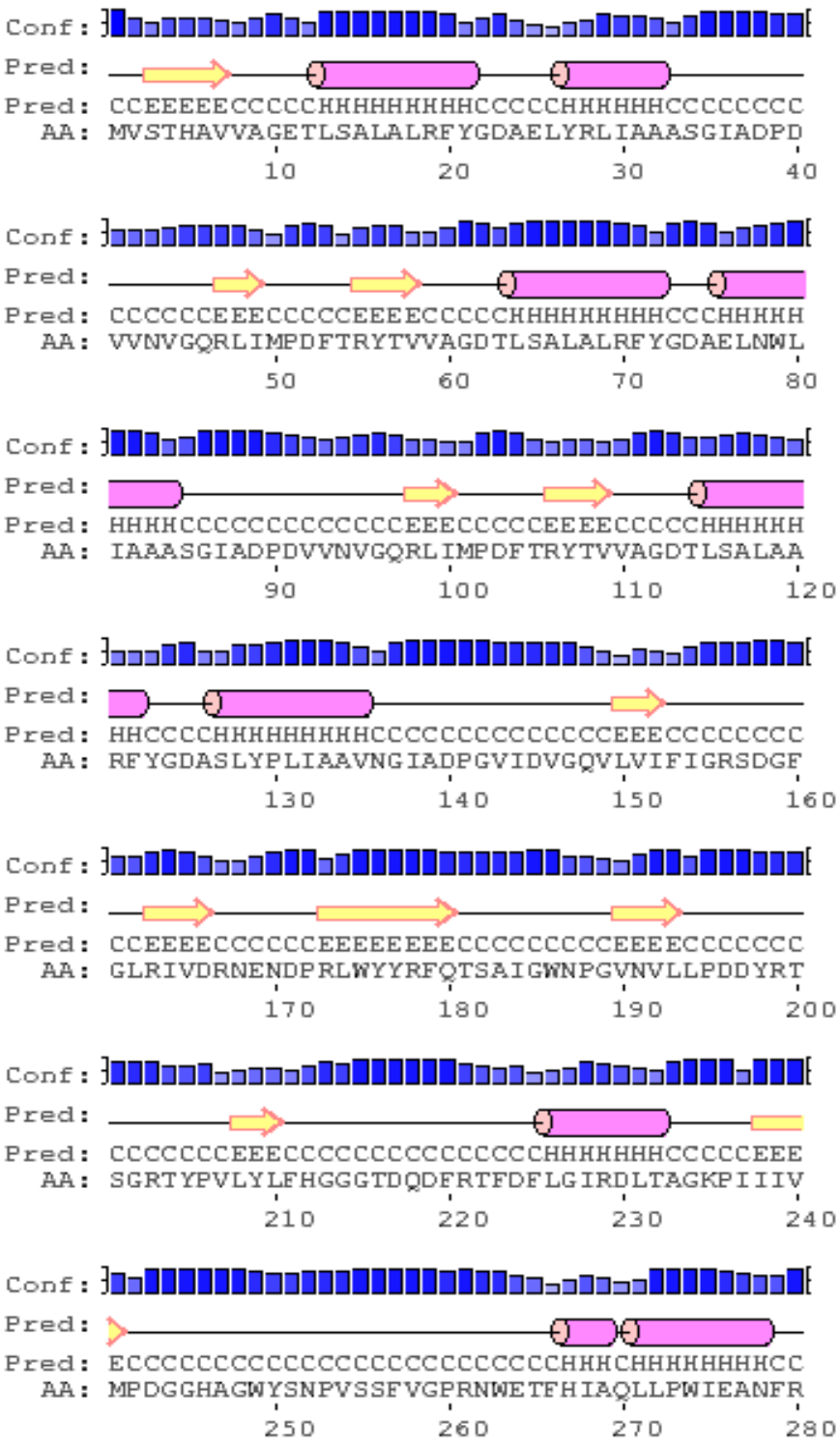


Figure 3.1: Multiple sequence alignment of Rv1288 against homologs containing putative esterase catalytic activity from species variants. Sequence analysis shows that the Rv1288 sequences are similar to the sequences of those homologs suggesting that the C-terminal domain of Rv1288 contains hydrolase activity. The homologs share strongly conserved motifs (shaded boxes) from the N-terminal to the C-terminal of the sequences suggesting that they possess similar protein fold as well as protein function. The residues colored in red are the key motifs and catalytic triads for an esterase. The homologues in the sequence alignment including the putative esterase from *B. cenocepacia* (CAR56217.1), putative esterase from *Actinomyces sp* (EGF50189.1), predicted esterase from *Saccharomonospora viridis* (ACU95748.1), putative esterase from *M. abcessus* (CAM64874.1), putative esterase from *M. canettii* (CCC63907.1), putative esterase from *M. smegmatis* (ABK74318.1), putative esterase from *M. bovis* (CCC63907.1, putative esterase from *Myxococcus xanthus* (ABF89705.1) and putative esterase from *Streptomyces venezuelae* (CCA60450.1).


Conf:
Pred:
Pred: CCEEEEECCCCCHHHHHHHHHCCCCCHHHHHHCCCCCCCC
AA: MVSTHAVVAGETLSALALRFYGDAELYRLIAAASGIADPD
10 20 30 40
Conf:
Pred:
Pred: CCCCCCEEECCCCCEEEECCCCCHHHHHHHHHCCCHHHHH
AA: VVNVGQRLIMPDFTRYTVVAGDTLSALALRFYGDAELNWL
50 60 70 80
Conf:
Pred:
Pred: HHHHCCCCCCCCCCCCCEEECCCCCEEEECCCCCHHHHHH
AA: IAAASGIADPDVVNVGQRLIMPDFTRYTVVAGDTLSALAA
90 100 110 120
Conf:
Pred:
Pred: HHCCCCHHHHHHHHHCCCCCCCCCCCCCCCEEECCCCCCCC
AA: RFYGDASLYPLIAAVNGIADPGVIDVGQVLVIFIGRSDGF
130 140 150 160
Conf:
Pred:
Pred: CCEEEECCCCCCEEEEEEEECCCCCCCCCCEEEECCCCCCC
AA: GLRIVDRNENDPRLWYYRFQTSAIGWNPGVNVLLPDDYRT
170 180 190 200
Conf:
Pred:
Pred: CCCCCCCEEECCCCCCCCCCCCCCCHHHHHHHCCCCCEEE
AA: SGRTYPVLYLFHGGGTDQDFRTFDFLGIRDLTAGKPIIIV
210 220 230 240
Conf:
Pred:
Pred: ECCCCCCCCCCCCCCCCCCCCCCCCHHHCHHHHHHHHCC
AA: MPDGGHAGWYSNPVSSFVGPRNWETFHIAQLLPWIEANFR
250 260 270 280

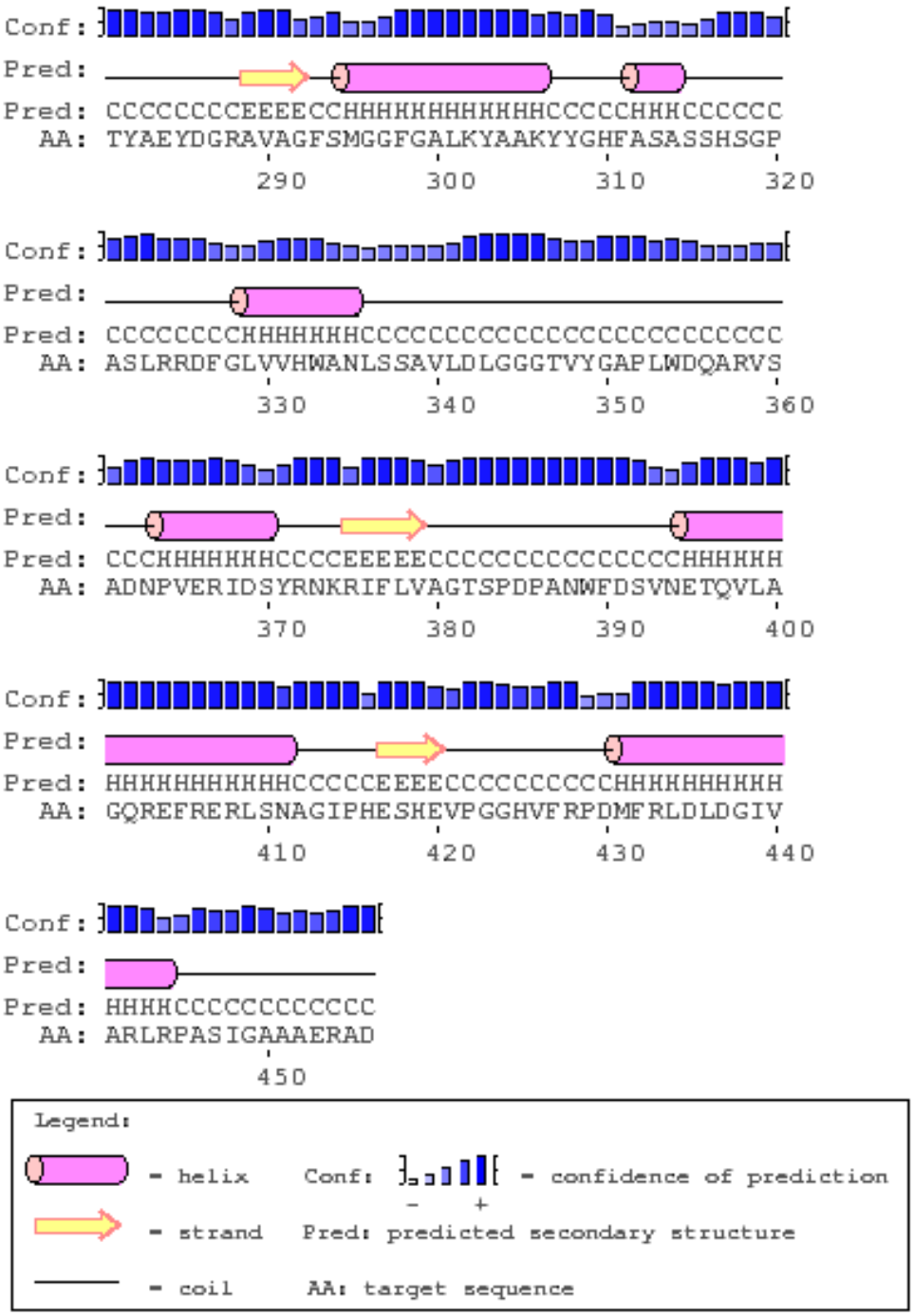


Figure 3.2: A predicted secondary structure of Rv1288. The PSI-blast secondary structure was performed by PSIPRED (Mcguffin *et al*., 2000). There are three LysM domains at the N-terminal end Rv1288 estimated from residue M1 to I154, while the remaining residues of the C-terminal of the protein are denoted for the putative catalytic esterase domain of Rv1288.

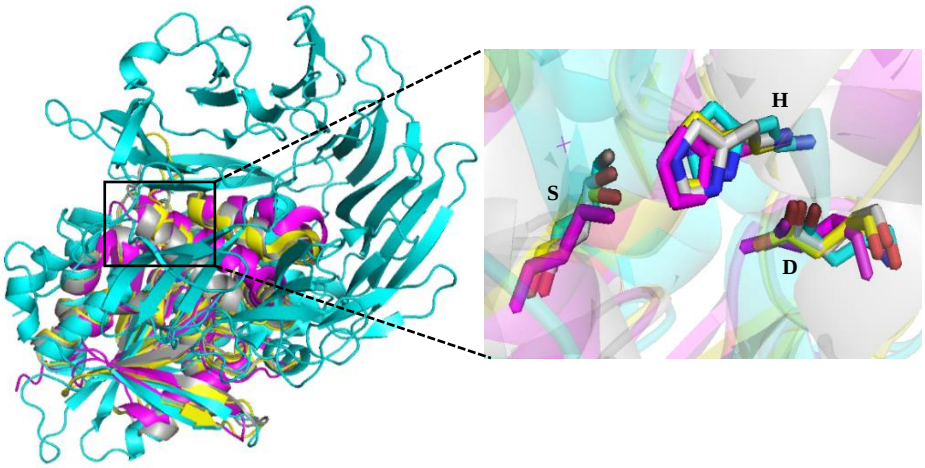


Figure 3.3: Structural analysis of multiple hydrolases from species variants in the PDB. All the hydrolases structures showed similar protein fold with an identical catalytic triad involving histidine (H), serine (S) and aspartic acid (D) residues. The aryl esterase from *P. fluorescens* (grey) (**1VA4**; catalytic triad: S94, D222, H251), chloroperoxidase from *P. fluorescens* (yellow) (**1A8S**; catalytic triad: S94, D224, H253), thermophilic esterase from *Archeaoglobus fulgidus* (pink) (**5FRD**; catalytic triad: S89, D200, H228), dipeptidyl amino peptidase IV from *Stenotrophomonas maltophilia* (cyan) (**2ECF**; S610, D685, H717).

```
                        β1          β2                  β3
1A8S                  TTFTTRDGTQIYYKDWGSGQ--------PIVFSHGWPLN------------  31
1VA4                  STFVAKDGTQIYFKDWGSGK--------PVLFSHGWLLD------------  31
5FRD                MLERVFIDVDGVKVSLLKGRERK---------VFYIHSSGS------------  32
2ECF                VEFGTLTAADGKTPLNYSVIKPAGFDPAKRYPVAVYVGGPASQTVTDSWP---  534
RV1288        GRSDGFGLRIVDRNENDPRLWYYRFQTSAIGWNP--GVNVLLPDDYRTSGRTYPVLYLF  211

                                  α1          β4
1A8S          ------------------ADSWESQMIFLAAQGYRVIAHDRRGHGRS------SQPWSGN  67
1VA4          ------------------ADMWEYQMEYLSSRGYRTIAFDRRGFGRS------DQPWTGN  67
5FRD          ----------------DATQWVNQLTAI---GGYAI--DLPNHGQS------D-TVEVN  64
2ECF          ---------------GRGDHLFNQYLAQQ--GYVVFSLDNRGTPRRGRDFGGALYGKQG  579
RV1288        HGGGTDQDFRTFDFLGIRDLTAGKPI-------IIVMPDGGHAGWYSNPVSSFVGPRNW  263

                    α2                  β5          α3              β6
1A8S          NDMDTYADDLAQLIEHL----DLRDAVLFGFSTGGGEVARYIGRHGTAR-VAKAGLISA  120
1VA4          NDYDTFADDIAQLIEHL----DLKEVTLVGFSMGGGDVARYIARHGSAR-VAGLVLLGA  120
5FRD          NSVDEYAYYASESLKKT-----VGKAVVVGHSLGGAVAQKLYLRNPEI--CLALVLV--  112
2ECF          GTVEVADQLRGVAWLKQQPWVDPARIGVQGWSNGGYMTLMLLAKASDSYACGVAG----  633
RV1288        WETFHIAQLLPWIEANFRTYAEYDGRAVAGFSMGGFGALKYAAKYYGH--FASASSHSG  320

                                  α4                                  α5
1A8S          VPPLMLKTEANPGGLPMEVFDGIRQASLADRSQLYKDLAS-GPFFGFNQPGAKSSAGMV  183
1VA4          VTPLFGQKPDYPQGVPLDVFARFKTELLKDRAQFISDF-N-APFYGINKGQVV-SQGVQ  181
5FRD          ------GTGARL-RVLPEILEGLKKEPEKA----VDLMLSMAFASK---------GEEY  155
2ECF          --------------APVT-------------------------------------------  636
RV1288        ------------PASLRR-------------------------------------------  326
```

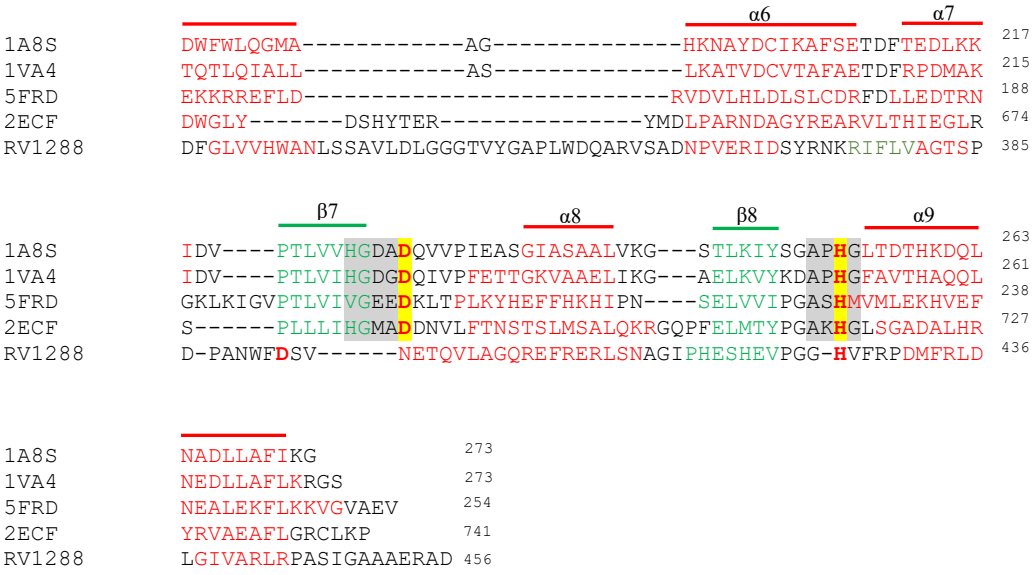


Figure 3.4: Structure-based sequence alignment of the related sequences in the PDB against the putative esterase domain of Rv1288. All the hydrolases including the aryl esterase from *P. fluorescens* (1VA4), thermophilic esterase from *Archeaoglobus fulgidus* (5FRD), dipeptidyl amino peptidase IV from *Stenotrophomonas maltophilia* (2ECF) and chloroperoxidase from *P. fluorescens* (1A8S), also exhibit an α/β hydrolase type. The key figure for these hydrolases that they shared identical catalytic triad involving histidine (H), serine (S) and aspartic acid (D) residues. Rv1288 might exhibit the α/β hydrolase type.

3.2 PCR product of Rv1288 and Trc1

The forward and reverse primers (Table 3.2) were designed to create a construct containing the full-length Rv1288 gene, followed by restriction sites for *NcoI* (CCATGG) and *XhoI* (CTCGAG) and six extra residues (ATATAT) for both ends of 5' and 3' of the nucleotides respectively (Figure 3.0). The gene was successfully amplified by PCR under the optimized PCR conditions (Table 3.1) using the designed primers at the annealing temperature of 72^{o}C, giving the total end PCR product size 1388 bp (Figure 3.5). The PCR amplicons of Rv1288 were detected on a 1% agarose gel visualized under UV light at the expected size (Figure 3.6).

A second construct, Trc1 was designed to express the three LysM domains alone and was successfully amplified by PCR under the optimized conditions using the designed primers at annealing temperature 67^{o}C (Table 3.0). The amplified product contained the truncated gene of Rv1288 followed by restriction sites for *NcoI* (CCATGG) and

XhoI (CTCGAG) and six extra residues (ATATAT) for both ends of 5' and 3' of the nucleotides respectively to give the total product size 485 bp (Figure 3.7). The amplified DNA of Trc1 on a 1% agarose gel at the expected size, visualized under UV light (Figure 3.8).

Table 3.0: Primers for the Rv1288 and Trc1* gene constructs

Forward primer	*5'ATATATCCATGG TCAGCACACATGCGGTTGT 3'*
Reverse primer	*5'ATATATCTCGAG ATCGGCGCGTTCTGCGGCC 3'*
Forward primer*	*5'TATATACCATGGTCAGCACACATGCGGTT 3'*
Reverse primer*	*5'TATATACTCGAGCCCGATGAATATGACCAGTACCT 3'*
Tm (°C)	72 and 74
Tm (°C)*	64 AND 68
Annealing (°C)	72
Annealing (°C)*	67
Restriction sites	*NcoI* (forward) and *Xho*I (reverse)
Restriction sites*	*NcoI* (forward) and *Xho*I (reverse)
Nucleotide	1371 bp
Nucleotide*	465 bp

Table 3.1: Optimized PCR conditions for Rv1288 and Trc1**

PCR mixture		
PCR components	Total volume (µL)	
Q5® Hot Start High-Fidelity 2X Master Mix	25.0	
Forward primer (20 pmol)	1.0	
Reverse primer (20 pmol)	1.0	
DNA template (25.0 – 100 ng)	-based on the stock concentration of the DNA	
miliQ water (5% DMSO)	-added to total volume	
Total volume	50.00	
PCR parameters		
Program	Temperature (°C) / Time	
Pre-denaturation	98.0 / 5.0 min	
Denaturation	98.0 / 30 sec	repeat the steps for 25 cycles
Annealing	72.0 / 30 sec	
Annealing**	67.0/ 30 sec	
Extension	72.0 / 1.0 min	
Post-extension	72.0 / 5.0 min	
Hold	5.0 / ∞	

ATATATCCATGGTCAGCACACATGCGGTTGTCGCGGGGGAGACGCTGTCGGCGTTGGCGTTGCGCTTCTATG
GCGACGCGGAACTGTATCGGCTGATCGCCGCCGCCAGCGGGATCGCCGATCCCGACGTCGTCAATGTGGGGC
AGCGGCTGATTATGCCTGACTTCACGCGATACACCGTTGTTGCCGGGGACACGCTGTCGGCGTTGGCGTTGC
GCTTCTATGGCGACGCGGAATTGAATTGGCTG
ATATATCCATGGTCAGCACACATGCGGTTGTCGCGGGGGAGACGCTGTCGGCGTTGGCGTTGCGCTTCTATG
GCGACGCGGAACTGTATCGGCTGATCGCCGCCGCCAGCGGGATCGCCGATCCCGACGTCGTCAATGTGGGGC
AGCGGCTGATTATGCCTGACTTCACGCGATACACCGTTGTTGCCGGGGACACGCTGTCGGCGTTGGCGTTGC
GCTTCTATGGCGACGCGGAATTGAATTGGCTGATCGCCGCCGCCAGCGGGATCGCCGATCCCGACGTCGTCA
ATGTGGGGCAGCGGCTGATTATGCCTGACTTCACGCGATACACCGTTGTTGCCGGGGACACGCTGTCGGCAT
TGGCTGCGCGCTTCTATGGCGACGCCTCCCTATATCCGCTTATCGCCGCCGTCAATGGCATCGCCGATCCTG
GCGTCATCGACGTCGGGCAGGTACTGGTCATATTCATCGGGCGTAGCGACGGGTTCGGCCTAAGGATCGTGG
ACCGCAACGAGAACGATCCCCGCCTGTGGTACTACCGGTTCCAGACCTCCGCGATCGGCTGGAACCCCGGAG
TCAACGTCCTGCTTCCCGATGACTACCGCACCAGCGGACGCACCTATCCCGTCCTCTACCTGTTCCACGGCG
GCGGCACCGACCAGGATTTCCGCACGTTCGACTTTCTGGGCATCCGCGACCTGACCGCCGGAAAGCCGATCA
TCATCGTGATGCCCGACGGCGGGCACGCGGGCTGGTATTCCAACCCGGTCAGCTCGTTCGTCGGCCCACGGA
ACTGGGAGACATTCCACATCGCCCAGCTGCTCCCCTGGATCGAGGCGAACTTCCGAACCTACGCCGAATACG
ACGGCCGCGCGGTCGCCGGGTTTTCGATGGGTGGCTTCGGCGCGCTGAAGTACGCAGCAAAGTACTACGGCC
ACTTCGCGTCGGCGAGCAGCCACTCCGGACCGGCAAGTCTGCGCCGCGACTTCGGCCTGGTAGTGCATTGGG
CAAACCTGTCCTCGGCGGTGCTGGATCTAGGCGGCGGCACGGTTTACGGCGCGCCGCTCTGGGACCAAGCTA
GGGTCAGCGCCGACAACCCGGTCGAGCGTATCGACAGCTACCGCAACAAGCGGATCTTCCTGGTCGCCGGCA
CCAGTCCGGACCCGGCCAACTGGTTCGACAGCGTGAACGAGACCCAGGTGCTAGCCGGGCAGAGGGAGTTCC
GCGAACGCCTCAGCAACGCCGGCATCCCGCATGAATCGCACGAGGTGCCTGGCGGTCACGTCTTCCGGCCCG
ACATGTTCCGTCTCGACCTCGACGGCATCGTCGCCCGGCTGCGCCCCGCGAGCATCGGGGCGGCCGCAGAAC
GCGCCGAT**CTCGAG****ATATAT** (1388 bp)

Figure 3.5: The nucleotide sequence of PCR amplicon for Rv1288. The Rv1288 gene was successfully amplified by PCR at an annealing temperature of 72 ºC. The restriction sites for *NcoI* (green) and *XhoI* (blue) are indicated and are followed with six extra nucleotides (red) at both ends of the gene sequence.

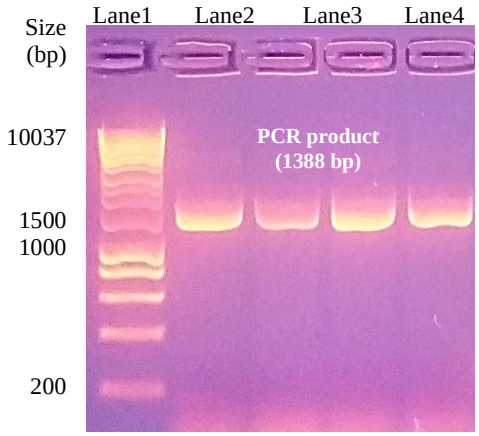


Figure 3.6: PCR product of Rv1288 on a 1% agarose gel. The amplified PCR product of Rv1288 was 1388 bp. Lane1: Marker, Lane2-Lane5: amplified Rv1288 product with 1388 bp in size.

TATATACCATGGTCAGCACACATGCGGTTGTCGCGGGGGAGACGCTGTCGGCGTTGGCGTTGCGCTTCTATG
GCGACGCGGAACTGTATCGGCTGATCGCCGCCGCCAGCGGGATCGCCGATCCCGACGTCGTCAATGTGGGGC
AGCGGCTGATTATGCCTGACTTCACGCGATACACCGTTGTTGCCGGGGACACGCTGTCGGCGTTGGCGTTGC
GCTTCTATGGCGACGCGGAATTGAATTGGCTGATCGCCGCCGCCAGCGGGATCGCCGATCCCGACGTCGTCA
ATGTGGGGCAGCGGCTGATTATGCCTGACTTCACGCGATACACCGTTGTTGCCGGGGACACGCTGTCGGCAT
TGGCTGCGCGCTTCTATGGCGACGCCTCCCTATATCCGCTTATCGCCGCCGTCAATGGCATCGCCGATCCTG
GCGTCATCGACGTCGGGCAGGTACTGGTCATATTCATCGGG**CTCGAG****TATATA** (485 bp)

Figure 3.7: Nucleotide sequences of PCR amplicon for Trc1. The Trc1 construct was successfully amplified by PCR at an annealing temperature of 67 ºC. The restriction sites for *NcoI* (brown) and *XhoI* (orange) are indicated and are followed with six extra nucleotides (red) at both ends of the gene sequence.

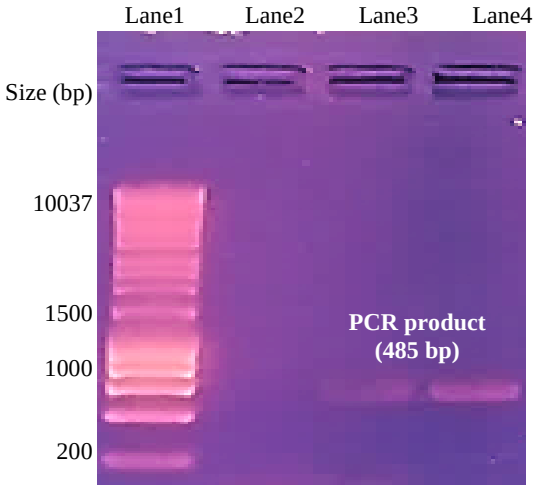


Figure 3.8: PCR product of Trc1 on a 1% agarose gel. The PCR amplicon of Trc1 was 485 bp. Lane1: Marker, Lane2: negative control, Lane3-Lane4: amplified Trc1 product.

3.3 Digestion product of Rv1288, Trc1 and pET24d

The PCR product of Rv1288, Trc1, and pET24d cloning vector were digested by using *NcoI* and *XhoI* restriction enzymes to provide sticky ends for the gene insert and the vector. After digestion, the size of the Rv1288 gene was 1374 bp, Trc1 was 471 while pET24d was ~5230 bp. The schematic diagram of the digested Rv1288 gene construct and the digested pET24d DNA with sticky ends were shown in Figure 3.9. The cut DNA of the Rv1288 gene and Trc1, both with vector pET24d were detected on a 1% agarose gel at the expected sizes shown in Figure 3.10 and Figure 3.11, respectively. The digested products of Rv1288, Trc1 and pET24d, were cleaned up from any contaminants using a gel extraction kit. The digested pET24d was further treated with phosphatase to remove the phosphate group at the end of the sticky ends of the vector DNA to avoid the re-ligation of the vector.

3.4 DNA ligation product of pET24d-Rv1288 and pET24d-Trc1

The DNA of the Rv1288 gene insert with the pET24d vector, and the Trc1 with the pET24d vector were subsequently ligated using DNA ligase. The ligated product of pET24d-Rv1288 and pET24d-Trc1 were successfully obtained from overnight incubation on ice (Figure 3.12).

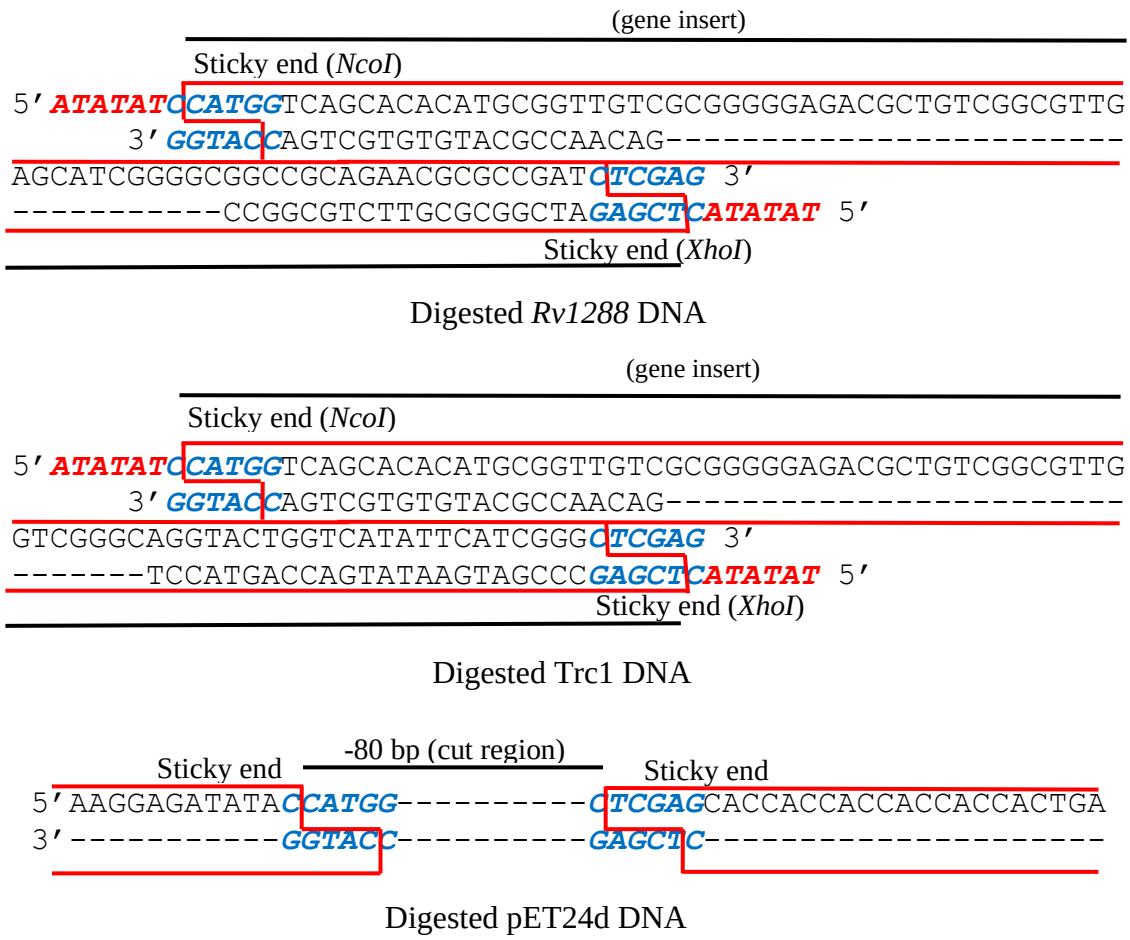


Figure 3.9: A schematic diagram of the digested Rv1288, Trc1, and pET24d constructs. The Rv1288 gene, Trc1, and pET24d DNA were treated with double digestion by *NcoI* and *XhoI* restriction enzymes to provide sticky ends to both ends of the products for DNA ligation experiment.

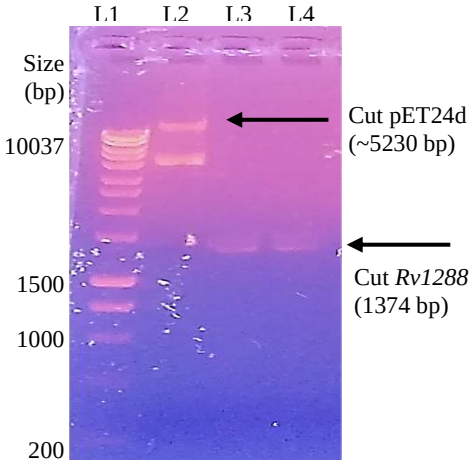


Figure 3.10: Digested DNA of Rv1288 and vector pET24d on a 1% agarose gel. Both, the Rv1288 gene and pET24d DNA were digested with *NcoI* and *XhoI* restriction enzymes. The final product for the cut Rv1288 gene was 1374 bp. Lane1: Marker, Lane2: cut vector pET24d, Lane3-Lane4: cut *Rv1288*.

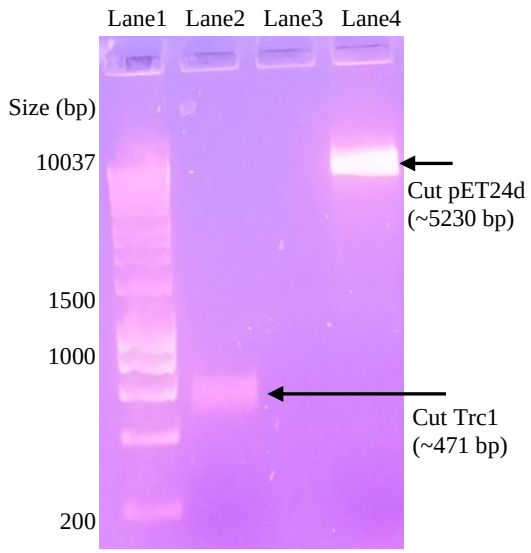


Figure 3.11: Digested DNA of Trc1 and vector pET24d on a 1% agarose gel. Both, the DNA construct of Trc1 and pET24d DNA were digested with *NcoI* and *XhoI* restriction enzymes. The final product for the cut was 471 bp. Lane1: Marker, Lane2: cut Trc1, Lane3: no sample, Lane4: cut pET24d.

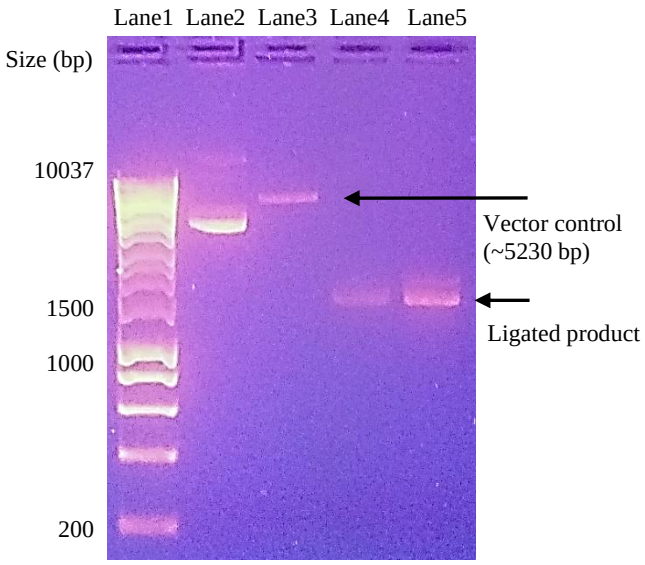


Figure 3.12: Ligation products of Rv1288 and pET24d on the 1% agarose gel. The ligation was carried out on ice for overnight. Lane1: Marker, Lane2: Uncut pET24d as a control, Lane3: cut pET24d, Lane4-5: Successful ligated product of pET24d-Rv1288.

3.5 Transformants

The recombinant plasmids containing the full length of Rv1288 gene and Trc1 were successfully transformed into *E. coli* cells by heat shock transformation. The transformants were initially inoculated into SOC media containing glucose and potassium which helps the recovery of *E. coli* cells after heat shock before plating onto LB media to get the higher efficiency of plasmid transformation into *E. coli* cells. The transformants grew on the LB agar supplemented with 50 µg/mL kanamycin, the selectable marker for pET24d containing pET-Rv1288 and pET-Trc1 respectively are shown in Figure 3.13 and Figure 3.14, respectively.

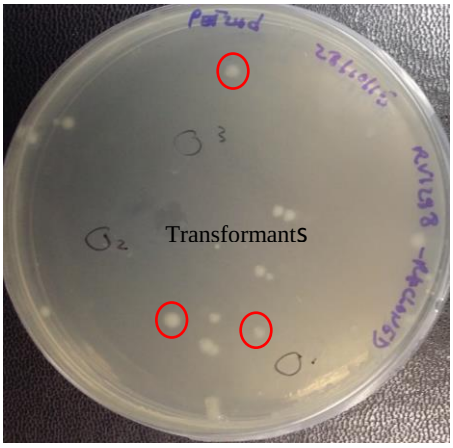


Figure 3.13: Transformants of recombinant plasmid pET24d-Rv1288. The transformants grew on LB media supplemented with 50 µg/mL kanamycin. Selected transformants for colony PCR assay, are highlighted with red circles.

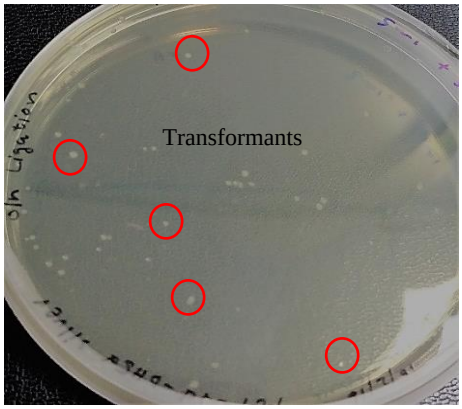


Figure 3.14: Transformants of recombinant plasmid pET24d-Trc1. The transformants grew on LB media supplemented with 50 µg/mL kanamycin. Selected transformants for colony PCR assay, are highlighted with red circles.

3.6 Colony PCR and sequencing analysis of pET24d-Rv1288 and pEt24d-Trc1

The transformants from two different transformations that grew on the LB media were randomly selected for a colony PCR assay to detect the presence of the ligated Rv1288 gene and the Trc1 gene constructs. T7 primers were utilized to amplify the Rv1288 gene and Trc1 gene construct as the vector pET24d utilizes a T7 promoter as a gene transcription regulator in *E. coli*. The amplified PCR product contained a vector sequence from the T7 promoter to the T7 terminator and the inserted genes. The expected end products of the colony PCR for Rv1288 was 1585 bp while 682 bp was for the Trc1 gene construct and the schematic diagram of the gene constructs for both Rv1288 and Trc1 in the recombinant plasmids is shown in Figure 3.15. DNA of pET24d-Rv1288 (Figure 3.16) and pET24d-Trc1 (Figure 3.17) were detected on 1% agarose gel at expected sizes

The DNA sequencing was carried out for the recombinant plasmid to analyze the inserted gene sequence and its orientation in the plasmid. The chromatogram obtained from the DNA sequencing showed that no mutations were observed in the sequence of the recombinant plasmids containing Rv1288 (Figure 3.18) or Trc1 (Figure 3.19), and the inserted genes were in the right orientation as the start codon of the gene of both constructs were next to the TATA box of the vector plasmids. The Rv1288 gene and the Trc1 gene construct were successfully cloned into the expression vector pET24d with six histidines at the C-terminus of the gene to facilitate purification. Subsequent sequence alignment showed that the DNA of the inserts were 100% identical to the gene sequence of Rv1288 (Figure 3.20) and Trc1 (Figure 3.21).

The nucleotide sequences of the Rv1288 gene (1392 bp) and His_6-Trc1 gene construct (489 bp) obtained from the sequencing analysis were translated into protein sequences by ExPASy translate tool. Rv1288 encodes a protein with 464 amino acid residues including eight extra residues which are leucine (L) and glutamic acid (E) as a linker followed by six histidine residues giving a total molecular weight of the protein of 50684.03 Da. The theoretical pI of the protein is 5.70, and the calculated extinction coefficient (Abs 0.1% (mg/mL)) for the protein at 280 nm is 1.57 M^{-1} cm^{-1} computed by the ExPASy protparam tool. While His_6-Trc1 contains 163 amino acid residues with a similar linker and six histidines giving a total molecular weight for the truncated protein is 17328.80 Da. The theoretical pI of the protein is 4.92, and the calculated extinction coefficient (Abs 0.1% (mg/mL)) for the protein at 280 nm is 0.91 M^{-1} cm^{-1} computed by the ExPASy protparam tool.

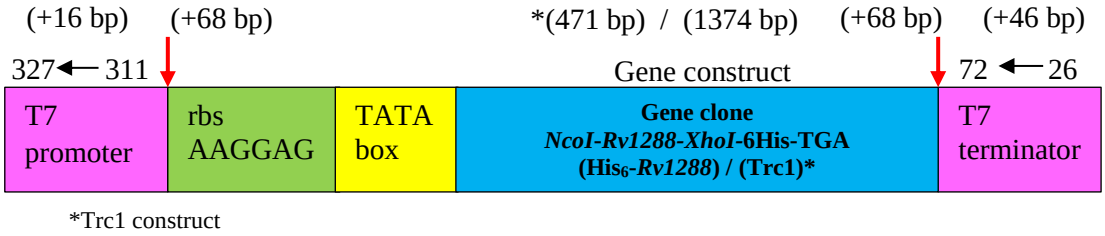


Figure 3.15: A schematic diagram of the recombinant plasmid pET24d-Rv1288 construct. The construct includes the T7 promoter to the T7 terminator of the recombinant plasmid, including the Rv1288 gene construct. The total size of the recombinant plasmid including the gene insert from the T7 promoter to T7 terminator position is 1585 bp. The two arrows indicate the position of the additional nucleotide sequence from the plasmid including Lac operator region in between the T7 promoter and rbs sites. The Trc1 product is 706 bp.

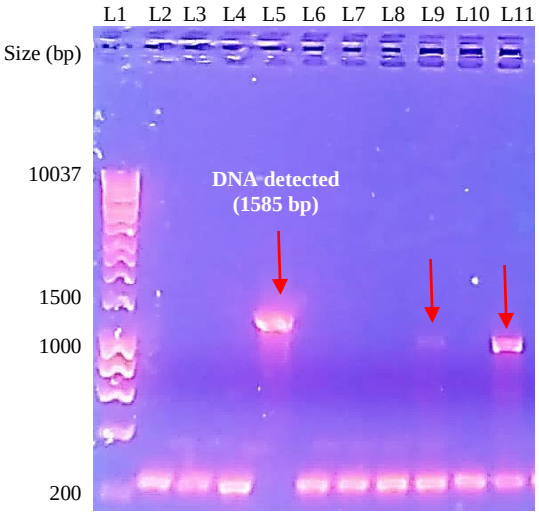


Figure 3.16: Colony PCR of the recombinant plasmid containing Rv1288 and Trc1 on the 1% agarose gels. The colony PCR using T7 primers was performed on 11 colonies of the transformants. L1: Marker, L2-L4, L6-L8 and L10: nonspecific PCR products, PCR products from L5, L9, and L11 (red arrows) were sent for sequencing.

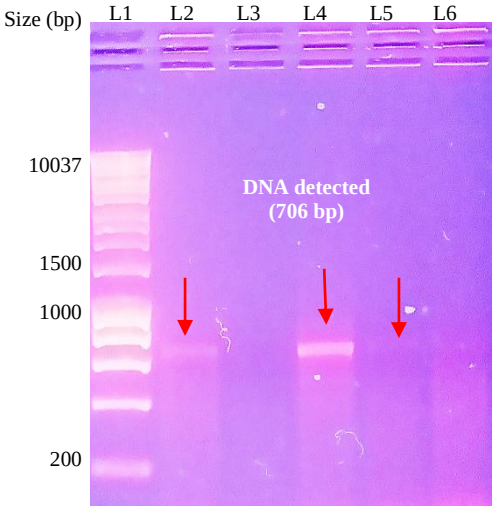
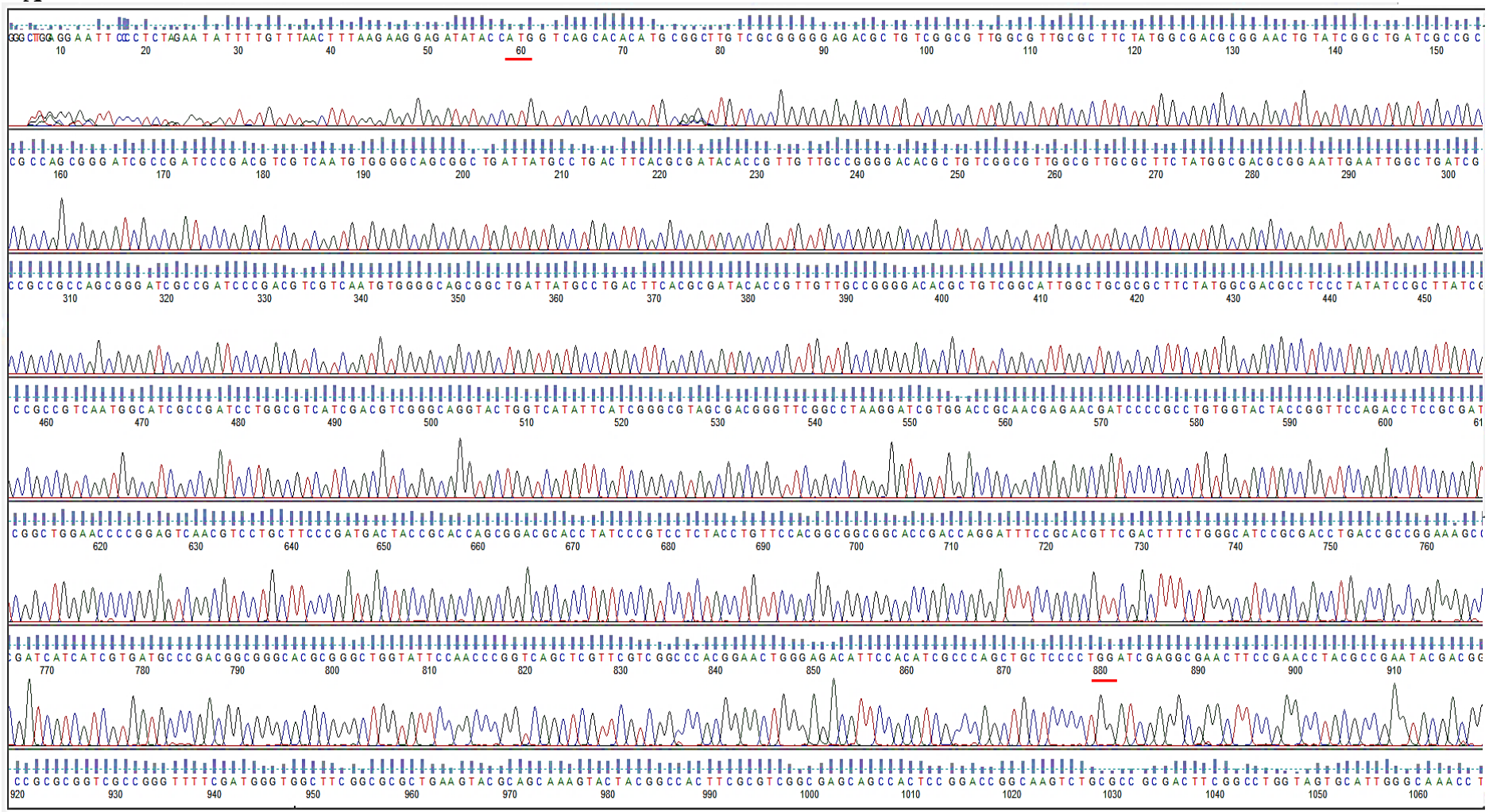


Figure 3.17: Colony PCR of the recombinant plasmid containing Trc1 on the 1% agarose gels. The colony PCR using T7 primers was performed on 5 colonies of the transformants. L1: Marker, L3 and L6: no PCR products were detected, PCR products from L2, L4, and L5 (red arrows) were sent for sequencing.

A



B

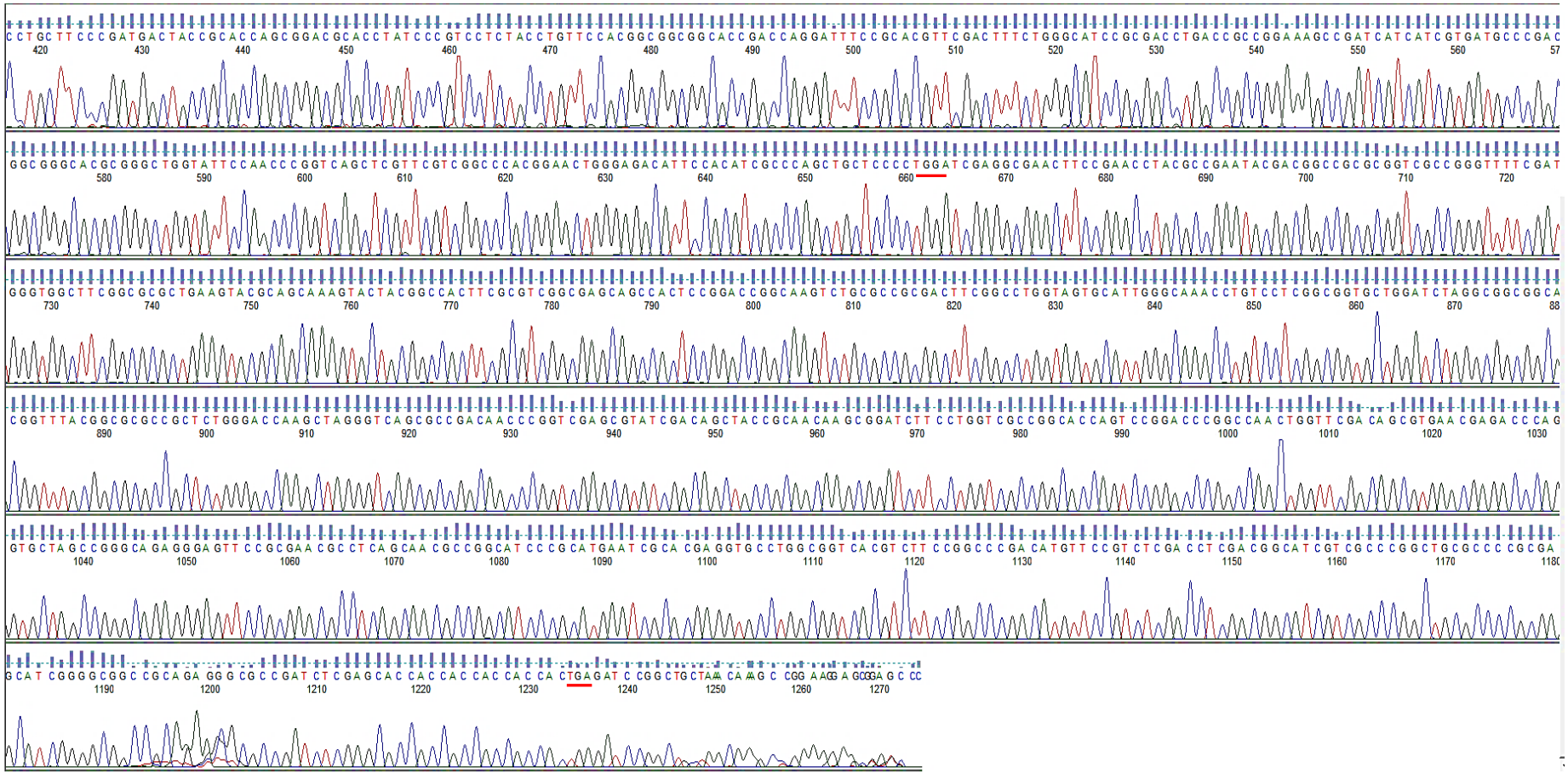


Figure 3.18 (A-B): Sequencing analysis of the recombinant plasmid pET24d-Rv1288 by T7 primers. The Rv1288 sequence was from position 59-880 (forward T7 primer) and 663-1236 (reverse T7 primer) indicated by red lines to give a total amplicon size 1395 bp (including the stop codon).

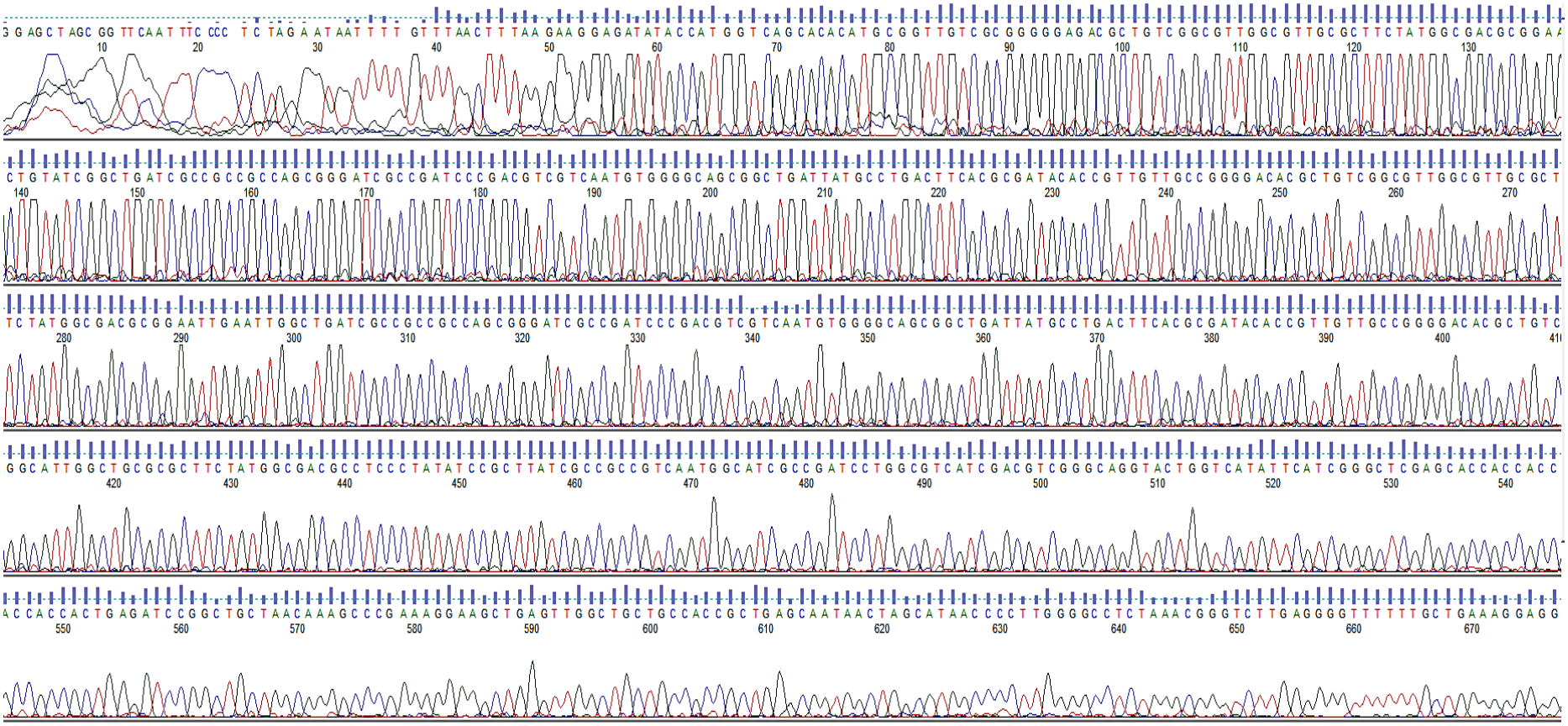


Figure 3.19: Sequencing analysis of the recombinant plasmid pET24d-Trc1 by T7 primers. The Trc1 sequence was from position 64-555 to give a total amplicon size 495 bp (including the stop codon).

```
55ED40/41       AAGGAGATATACCATGGTCAGCACACATGCGGTTGTCGCGGGGGAGACGC
Rv1288          AT-------------GGTCAGCACACATGCGGTTGTCGCGGGGGAGACGC
                *              **********************************

55ED40/41       TGTCGGCGTTGGCGTTGCGCTTCTATGGCGACGCGGAACTGTATCGGCTG
Rv1288          TGTCGGCGTTGGCGTTGCGCTTCTATGGCGACGCGGAACTGTATCGGCTG
                **************************************************

55ED40/41       ATCGCCGCCGCCAGCGGGATCGCCGATCCCGACGTCGTCAATGTGGGGCA
Rv1288          ATCGCCGCCGCCAGCGGGATCGCCGATCCCGACGTCGTCAATGTGGGGCA
                **************************************************

55ED40/41       GCGGCTGATTATGCCTGACTTCACGCGATACACCGTTGTTGCCGGGGACA
Rv1288          GCGGCTGATTATGCCTGACTTCACGCGATACACCGTTGTTGCCGGGGACA
                **************************************************

55ED40/41       CGCTGTCGGCGTTGGCGTTGCGCTTCTATGGCGACGCGGAATTGAATTGG
Rv1288          CGCTGTCGGCGTTGGCGTTGCGCTTCTATGGCGACGCGGAATTGAATTGG
                **************************************************

55ED40/41       CTGATCGCCGCCGCCAGCGGGATCGCCGATCCCGACGTCGTCAATGTGGG
Rv1288          CTGATCGCCGCCGCCAGCGGGATCGCCGATCCCGACGTCGTCAATGTGGG
                **************************************************

55ED40/41       GCAGCGGCTGATTATGCCTGACTTCACGCGATACACCGTTGTTGCCGGGG
Rv1288          GCAGCGGCTGATTATGCCTGACTTCACGCGATACACCGTTGTTGCCGGGG
                **************************************************

55ED40/41       ACACGCTGTCGGCATTGGCTGCGCGCTTCTATGGCGACGCCTCCCTATAT
Rv1288          ACACGCTGTCGGCATTGGCTGCGCGCTTCTATGGCGACGCCTCCCTATAT
                **************************************************

55ED40/41       CCGCTTATCGCCGCCGTCAATGGCATCGCCGATCCTGGCGTCATCGACGT
Rv1288          CCGCTTATCGCCGCCGTCAATGGCATCGCCGATCCTGGCGTCATCGACGT
                **************************************************

55ED40/41       CGGGCAGGTACTGGTCATATTCATCGGGCGTAGCGACGGGTTCGGCCTAA
Rv1288          CGGGCAGGTACTGGTCATATTCATCGGGCGTAGCGACGGGTTCGGCCTAA
                **************************************************

55ED40/41       GGATCGTGGACCGCAACGAGAACGATCCCCGCCTGTGGTACTACCGGTTC
Rv1288          GGATCGTGGACCGCAACGAGAACGATCCCCGCCTGTGGTACTACCGGTTC
                **************************************************

55ED40/41       CAGACCTCCGCGATCGGCTGGAACCCCGGAGTCAACGTCCTGCTTCCCGA
Rv1288          CAGACCTCCGCGATCGGCTGGAACCCCGGAGTCAACGTCCTGCTTCCCGA
                **************************************************

55ED40/41       TGACTACCGCACCAGCGGACGCACCTATCCCGTCCTCTACCTGTTCCACG
Rv1288          TGACTACCGCACCAGCGGACGCACCTATCCCGTCCTCTACCTGTTCCACG
                **************************************************

55ED40/41       GCGGCGGCACCGACCAGGATTTCCGCACGTTCGACTTTCTGGGCATCCGC
Rv1288          GCGGCGGCACCGACCAGGATTTCCGCACGTTCGACTTTCTGGGCATCCGC
                **************************************************

55ED40/41       GACCTGACCGCCGGAAAGCCGATCATCATCGTGATGCCCGACGGCGGGCA
Rv1288          GACCTGACCGCCGGAAAGCCGATCATCATCGTGATGCCCGACGGCGGGCA
                **************************************************


55ED40/41       CGCGGGCTGGTATTCCAACCCGGTCAGCTCGTTCGTCGGCCCACGGAACT
Rv1288          CGCGGGCTGGTATTCCAACCCGGTCAGCTCGTTCGTCGGCCCACGGAACT
                **************************************************

55ED40/41       GGGAGACATTCCACATCGCCCAGCTGCTCCCCTGGATCGAGGCGAACTTC
Rv1288          GGGAGACATTCCACATCGCCCAGCTGCTCCCCTGGATCGAGGCGAACTTC
                **************************************************

55ED40/41       CGAACCTACGCCGAATACGACGGCCGCGCGGTCGCCGGGTTTTCGATGGG
Rv1288          CGAACCTACGCCGAATACGACGGCCGCGCGGTCGCCGGGTTTTCGATGGG
                **************************************************

55ED40/41       TGGCTTCGGCGCGCTGAAGTACGCAGCAAAGTACTACGGCCACTTCGCGT
Rv1288          TGGCTTCGGCGCGCTGAAGTACGCAGCAAAGTACTACGGCCACTTCGCGT
                **************************************************
```

```
55ED40/41       CGGCGAGCAGCCACTCCGGACCGGCAAGTCTGCGCCGCGACTTCGGCCTG
Rv1288          CGGCGAGCAGCCACTCCGGACCGGCAAGTCTGCGCCGCGACTTCGGCCTG
                **************************************************

55ED40/41       GTAGTGCATTGGGCAAACCTGTCCTCGGCGGTGCTGGATCTAGGCGGCGG
Rv1288          GTAGTGCATTGGGCAAACCTGTCCTCGGCGGTGCTGGATCTAGGCGGCGG
                **************************************************

55ED40/41       CACGGTTTACGGCGCGCCGCTCTGGGACCAAGCTAGGGTCAGCGCCGACA
Rv1288          CACGGTTTACGGCGCGCCGCTCTGGGACCAAGCTAGGGTCAGCGCCGACA
                **************************************************

55ED40/41       ACCCGGTCGAGCGTATCGACAGCTACCGCAACAAGCGGATCTTCCTGGTC
Rv1288          ACCCGGTCGAGCGTATCGACAGCTACCGCAACAAGCGGATCTTCCTGGTC
                **************************************************

55ED40/41       GCCGGCACCAGTCCGGACCCGGCCAACTGGTTCGACAGCGTGAACGAGAC
Rv1288          GCCGGCACCAGTCCGGACCCGGCCAACTGGTTCGACAGCGTGAACGAGAC
                **************************************************

55ED40/41       CCAGGTGCTAGCCGGGCAGAGGGAGTTCCGCGAACGCCTCAGCAACGCCG
Rv1288          CCAGGTGCTAGCCGGGCAGAGGGAGTTCCGCGAACGCCTCAGCAACGCCG
                **************************************************

55ED40/41       GCATCCCGCATGAATCGCACGAGGTGCCTGGCGGTCACGTCTTCCGGCCC
Rv1288          GCATCCCGCATGAATCGCACGAGGTGCCTGGCGGTCACGTCTTCCGGCCC
                **************************************************

55ED40/41       GACATGTTCCGTCTCGACCTCGACGGCATCGTCGCCCGGCTGCGCCCCGC
Rv1288          GACATGTTCCGTCTCGACCTCGACGGCATCGTCGCCCGGCTGCGCCCCGC
                **************************************************

55ED40/41       GAGCATCGGGGCGGCCGCAGAACGCGCCGATCTCGAGCACCACCACCACC
Rv1288          GAGCATCGGGGCGGCCGCAGAACGCGCCGAT------------------
                *****************************

55ED40/41       ACCACTGA
Rv1288          -----TAG
                     *
```

```
MVSTHAVVAG ETLSALALRF YGDAELYRLI AAASGIADPD VVNVGQRLIM
PDFTRYTVVA GDTLSALALR FYGDAELNWL IAAASGIADP DVVNVGQRLI
MPDFTRYTVV AGDTLSALAA RFYGDASLYP LIAAVNGIAD PGVIDVGQVL
VIFIGRSDGF GLRIVDRNEN DPRLWYYRFQ TSAIGWNPGV NVLLPDDYRT
SGRTYPVLYL FHGGGTDQDF RTFDFLGIRD LTAGKPIIIV MPDGGHAGWY
SNPVSSFVGP RNWETFHIAQ LLPWIEANFR TYAEYDGRAV AGFSMGGFGA
LKYAAKYYGH FASASSHSGP ASLRRDFGLV VHWANLSSAV LDLGGGTVYG
APLWDQARVS ADNPVERIDS YRNKRIFLVA GTSPDPANWF DSVNETQVLA
GQREFRERLS NAGIPHESHE VPGGHVFRPD MFRLDLDGIV ARLRPASIGA
AAERADLEHH HHHH
```

Figure 3.20: Sequence comparison of the Rv1288 gene construct from the recombinant plasmid with genome sequence, and the amino acid sequences of the protein. The sequence comparison showed that the recombinant plasmid contained the full length of the Rv1288 gene (1392 bp) with a linker (leucine (L) and glutamic acid (E) followed by the six histag located at the C-terminus of the gene construct. The sequence shows no mutations compared to the original gene sequence. The Rv1288 gene sequence obtained from the sequencing analysis was translated into 464 amino acid residues by the ExPASy translate tool. The amino acid sequence for Rv1288 contains 464 residues.

```
38FF17          AAGGAGATATACCATGGTCAGCACACATGCGGTTGTCGCGGGGGAGACGC
Rv1288          AT-------------GGTCAGCACACATGCGGTTGTCGCGGGGGAGACGC
                *              **********************************

38FF17          TGTCGGCGTTGGCGTTGCGCTTCTATGGCGACGCGGAACTGTATCGGCTG
Rv1288          TGTCGGCGTTGGCGTTGCGCTTCTATGGCGACGCGGAACTGTATCGGCTG
                **************************************************

38FF17          ATCGCCGCCGCCAGCGGGATCGCCGATCCCGACGTCGTCAATGTGGGGCA
Rv1288          ATCGCCGCCGCCAGCGGGATCGCCGATCCCGACGTCGTCAATGTGGGGCA
                **************************************************

38FF17          GCGGCTGATTATGCCTGACTTCACGCGATACACCGTTGTTGCCGGGGACA
Rv1288          GCGGCTGATTATGCCTGACTTCACGCGATACACCGTTGTTGCCGGGGACA
                **************************************************

38FF17          CGCTGTCGGCGTTGGCGTTGCGCTTCTATGGCGACGCGGAATTGAATTGG
Rv1288          CGCTGTCGGCGTTGGCGTTGCGCTTCTATGGCGACGCGGAATTGAATTGG
                **************************************************

38FF17          CTGATCGCCGCCGCCAGCGGGATCGCCGATCCCGACGTCGTCAATGTGGG
Rv1288          CTGATCGCCGCCGCCAGCGGGATCGCCGATCCCGACGTCGTCAATGTGGG
                **************************************************

38FF17          GCAGCGGCTGATTATGCCTGACTTCACGCGATACACCGTTGTTGCCGGGG
Rv1288          GCAGCGGCTGATTATGCCTGACTTCACGCGATACACCGTTGTTGCCGGGG
                **************************************************

38FF17          ACACGCTGTCGGCATTGGCTGCGCGCTTCTATGGCGACGCCTCCCTATAT
Rv1288          ACACGCTGTCGGCATTGGCTGCGCGCTTCTATGGCGACGCCTCCCTATAT
                **************************************************

38FF17          CCGCTTATCGCCGCCGTCAATGGCATCGCCGATCCTGGCGTCATCGACGT
Rv1288          CCGCTTATCGCCGCCGTCAATGGCATCGCCGATCCTGGCGTCATCGACGT
                **************************************************

38FF17          CGGGCAGGTACTGGTCATATTCATCGGGCTCGAGCACCACCACCACCACC
Rv1288          CGGGCAGGTACTGGTCATATTCATCG-------------------------
                *************************

38FF17          ACTGA
Rv1288          ---GG
                   *
```

```
MVSTHAVVAG ETLSALALRF YGDAELYRLI AAASGIADPD VVNVGQRLIM
PDFTRYTVVA GDTLSALALR FYGDAELNWL IAAASGIADP DVVNVGQRLI
MPDFTRYTVV AGDTLSALAA RFYGDASLYP LIAAVNGIAD PGVIDVGQVL
VIFIGLEHHH HHH (163 residues)
```

Figure 3.21: Sequence comparison of the Trc1 construct from the recombinant plasmid with genome sequence, and the amino acid sequences of the recombinant protein. There were no mutations observed in the sequencing results of the Trc1 sequence, and the ExPASy translate tool analysis shows that the nucleotide sequences (489 bp) of the construct encode 163 amino acid residues with the linker (leucine (L) and glutamic acid (E)) followed by the six histag located at the C-terminus of the gene construct.

3.7 Protein over-expression

Small and large-scale protein expressions were carried out on the recombinant plasmids containing Rv1288 and Trc1 in BL21 *DE3* cells. The small-scale expression was initially carried out to obtain the optimum conditions for Rv1288 and Trc1 to be expressed as a soluble protein in *E.coli* cells.

3.7.1 Small scale protein expression of Rv1288 and Trc1

Trial experiments of protein expression for the Rv1288 gene and Trc1 gene construct were carried out under a range of conditions. The Rv1288 gene was successfully expressed at all the trial conditions but with the different total amount of protein in the soluble fractions. The protein expressed best at either 25^{o}C for 24 hr or 18^{o}C for 42 hr for Rv1288. In contrast, the His_6-Trc1 gene construct expressed at all the setting conditions; 37^{o}C for 5 hr, 25^{o}C for 24 hr, 18^{o}C at either 48 hr or 72 hr. The results obtained from the experiments for both samples were visualized on a 12% SDS-PAGE gel as shown in Figure 3.22 and Figure 3.23.

3.7.2 Large-scale protein expression of Rv1288 and Trc1

Following the trial experiments, large-scale expression of Rv1288 was carried out using the optimal condition of 18^{o}C for 48 hr at 180 rpm. The 2L conical flask containing 500 mL of LB media were grown under this condition, and the cells were induced with 1 mM IPTG. The cells were harvested by centrifugation at either 10,000 rpm for 20 min or 5000 rpm for 30 min at 4^{o}C. SDS-PAGE gel analysis showed that the soluble Rv1288 protein was successfully expressed in the larger scale expression under the optimized condition with the ratio between soluble to insoluble fractions 40%: 60% (Figure 3.22 B). A similar protocol to that used to express the Rv1288 was applied for the Trc1 protein.

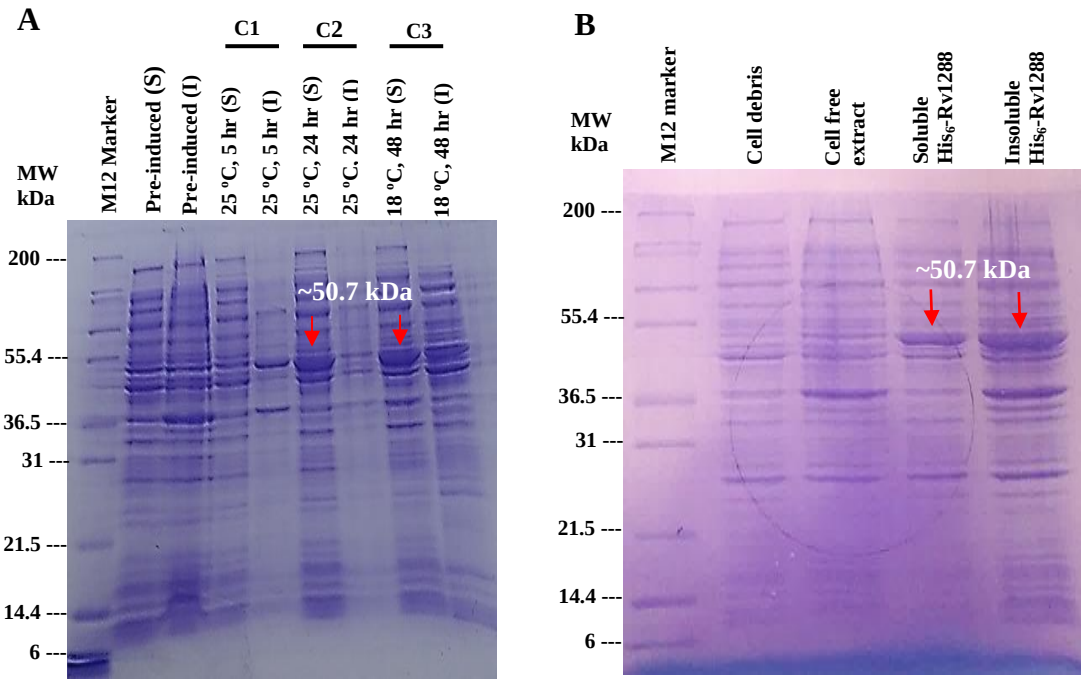


Figure 3.22: SDS-PAGE gels for small-scale and large-scale expressions of Rv1288. A)The small-scale of trial expression for His_6-Rv1288 was performed at three different conditions; (C1): 25 °C for 5 hr, (C2): 25 °C for 24 hr and (C3): 18 °C for 48 hr. SDS-PAGE gel analysis (left) shows that Rv1288 was successfully expressed at all the incubation settings with different amount of protein in the soluble fraction. The Rv1288 protein expressed best under conditions of C2 (25 °C, 24 hr) and C3 (18 °C, 48 hr). The position of the expressed soluble Rv1288 protein in a 12% SDS-PAGE gel is indicated by the arrows (MW ~50.7 kDa). B) The large-scale protein expression of Rv1288 was performed at 18 °C, 180 rpm for 48 hr with a final concentration of 1mM IPTG. SDS-PAGE gel analysis (right) shows that Rv1288 was successfully expressed, with different amount of protein in the soluble fraction indicated by the arrows.

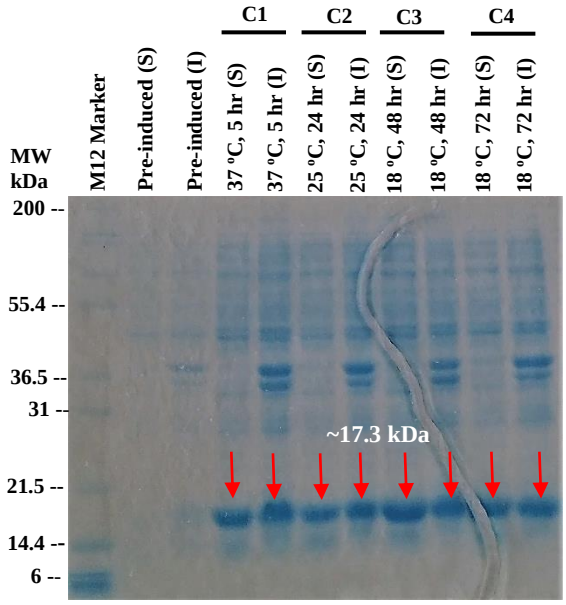


Figure 3.23: SDS-PAGE gel for small-scale expression of Trc1. The protein expression was performed at four different conditions; (C1): 37 °C for 5 hr, (C2) 25 °C for 24 hr, (C3): 18 °C for 48 hr and (C4) 18 °C for 72 hr. His_6-Trc1 was successfully expressed at all the incubation settings. The position of expressed Trc1 in the SDS-PAGE gel is indicated by the arrows at the MW ~17.3 kDa.

3.8 Protein purification

Before protein purification, the cell paste of the harvested cells was defrosted and resuspended in Buffer A containing 0.5 M NaCl and 50 mM Tris-HCl buffer (pH 8.0). The cell paste was disrupted by three bursts of sonication on ice, each for 20 seconds. The soluble Rv1288 and the His_6-Trc1 proteins in the cell-free extract (CFE) samples were separated from the insoluble fraction by centrifugation at ~45,000g at 4^{o}C for 15 min. The proteins were further purified and polished by affinity chromatography and size exclusion chromatography (SEC) steps.

3.8.1 Purification of Rv1288 and Trc1 by affinity chromatography

The Rv1288 protein was applied to a 5mL HisTrap HP nickel affinity column (GE Healthcare Life Science) and was eluted under a linear gradient 0-70% of 0.5 M imidazole (0-0.35 M) in Buffer A. The protein was fractionated and eluted from the column between a volume ~45 mL to ~55 mL corresponding to approximately 0.36 M imidazole (Figure 3.24 (A)). The Rv1288 protein fractions were subjected to the 12% SDS-PAGE gel to check the purity of the protein which was estimated at ~90% (Figure 3.24 (B)).

Under the same purification protocol of affinity chromatography that applied for Rv1288, the Trc1 protein was fractionated and eluted in the Buffer A containing 0.5 M imidazole from the column volume ~45 mL to ~55 mL corresponding to approximately 0.36 M imidazole (Figure 3.26 (A)) which was identical to the Rv1288 protein. The eluted His_6-Trc1 has 90% purity shown by the SDS-PAGE analysis (Figure 3.26 (B)).

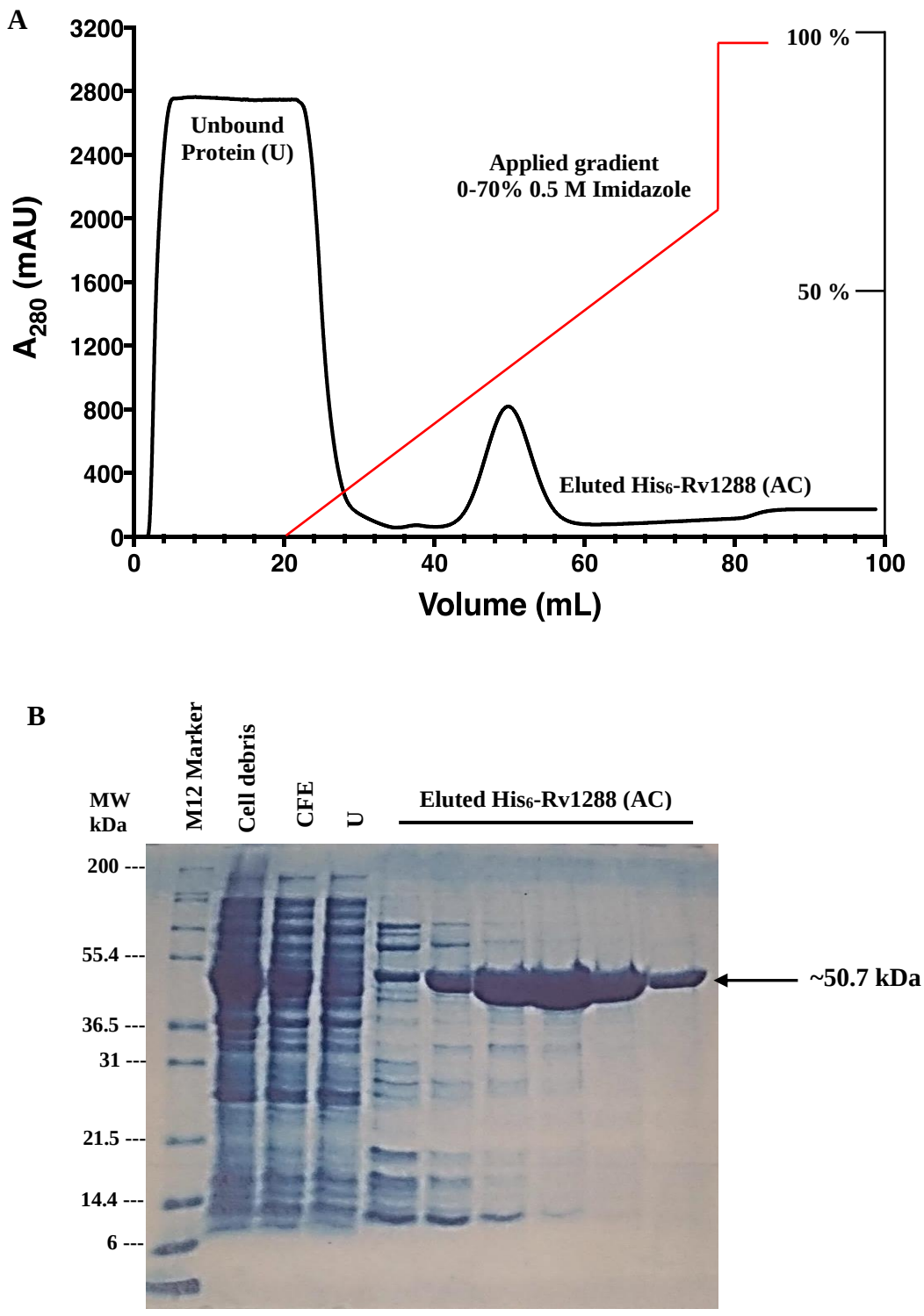


Figure 3.24: The His_6-Rv1288 purification by affinity chromatography. A) Chromatogram profile from the affinity chromatography using a 5mL HisTrap HP column. His_6-Rv1288 was fractionated under a linear gradient of 0.5 M imidazole (0-70%) in Buffer A, and the protein was eluted at a single peak at ~0.36 M imidazole. B) SDS-PAGE gel confirmed the presence of His_6-Rv1288 in the protein samples eluted under the single peak of the chromatogram at its expected molecular weight of ~50.7 kDa.

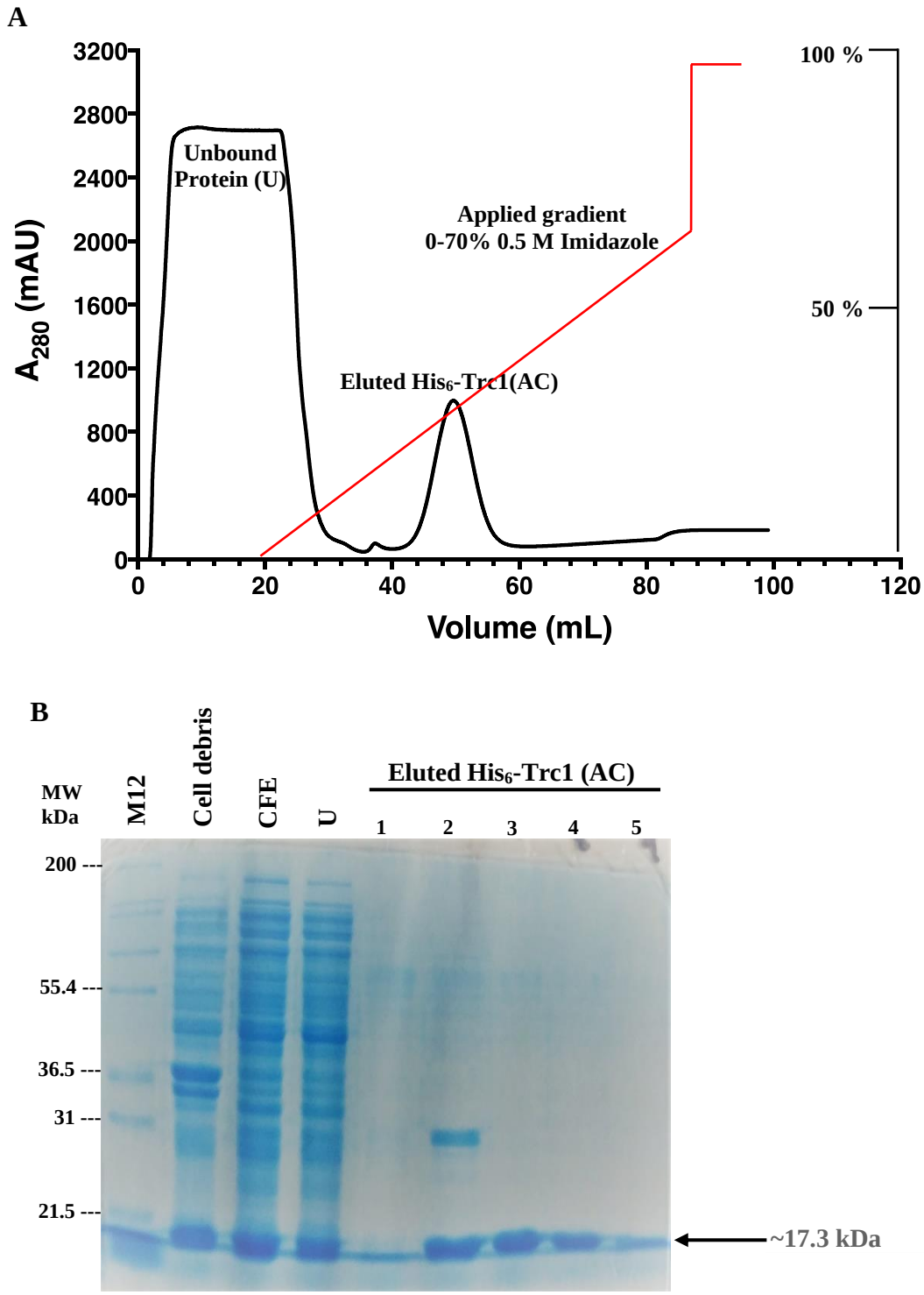


Figure 3.26: The His_6-Trc1 purification by affinity chromatography. A) Chromatogram profile from the affinity chromatography using a 5mL HisTrap HP column. His_6-Trc1 was fractionated under a linear gradient of 0.5 M imidazole (0-70%) in Buffer A, and the protein was eluted at a single peak at ~0.36 M imidazole. B) SDS-PAGE gel confirmed the presence of His_6-Trc1 in the protein samples eluted under the single peak of the chromatogram at its expected molecular weight of ~17.3 kDa.

3.8.2 Purification of Rv1288 and Trc1 by size exclusion chromatography

The eluted Rv1288 protein from the affinity chromatography column was further purified by size exclusion chromatography (SEC) by using Superdex-200pg (1.6x60 cm HiLoad) in Buffer A. The Rv1288 was successfully fractionated and eluted at an elution volume (Ve) of 50.9 mL (Figure 3.25 A). Given the calibrated void volume (Vo) and total volume (Vt) of the column (41 mL and 75 mL respectively), gave the Kav value for the Rv1288 protein ~0.13 to give an apparent molecular weight of the Rv1288 protein ~355 kDa (Figure 3.25 B). Given the subunit molecular weight of the protein (~50.7 kDa), this suggests that Rv1288 could be a hexamer. The SDS-PAGE gel (Figure 3.25 C) showed that, following SEC, the purity of the product was 90 %.

The Trc1 protein was further purified by SEC using the same column used for Rv1288, and the protein was eluted at Ve of 86.30, giving the Kav value 0.604 (Figure 3.27 A). The apparent molecular weight for the Trc1 protein based on the Kav against the calibration plot was 18 kDa (Figure 3.27 B). Given the subunit molecular weight of the protein (~17.3 kDa), this suggests that Trc1 is a monomer in solution. The SDS-PAGE gel (Figure 3.27 C) showed that, following SEC, the purity of the product was >95 %.

3.9 Mass spectrometry analysis on Trc1

Mass spectrometry analysis was performed on Trc1 to confirm the molecular weight of the expressed protein. One microliter of the Trc1 protein was utilized for mass spectrometry analysis. The mass spectrum results showed that the molecular mass of Trc1 was 17328.90 Da including a linker (leucine) and (glutamic acid) followed by the six histidine residues (Figure 3.28), similar to the computed MW for Trc1 by ExPASy protparam tool (Section 3.6).

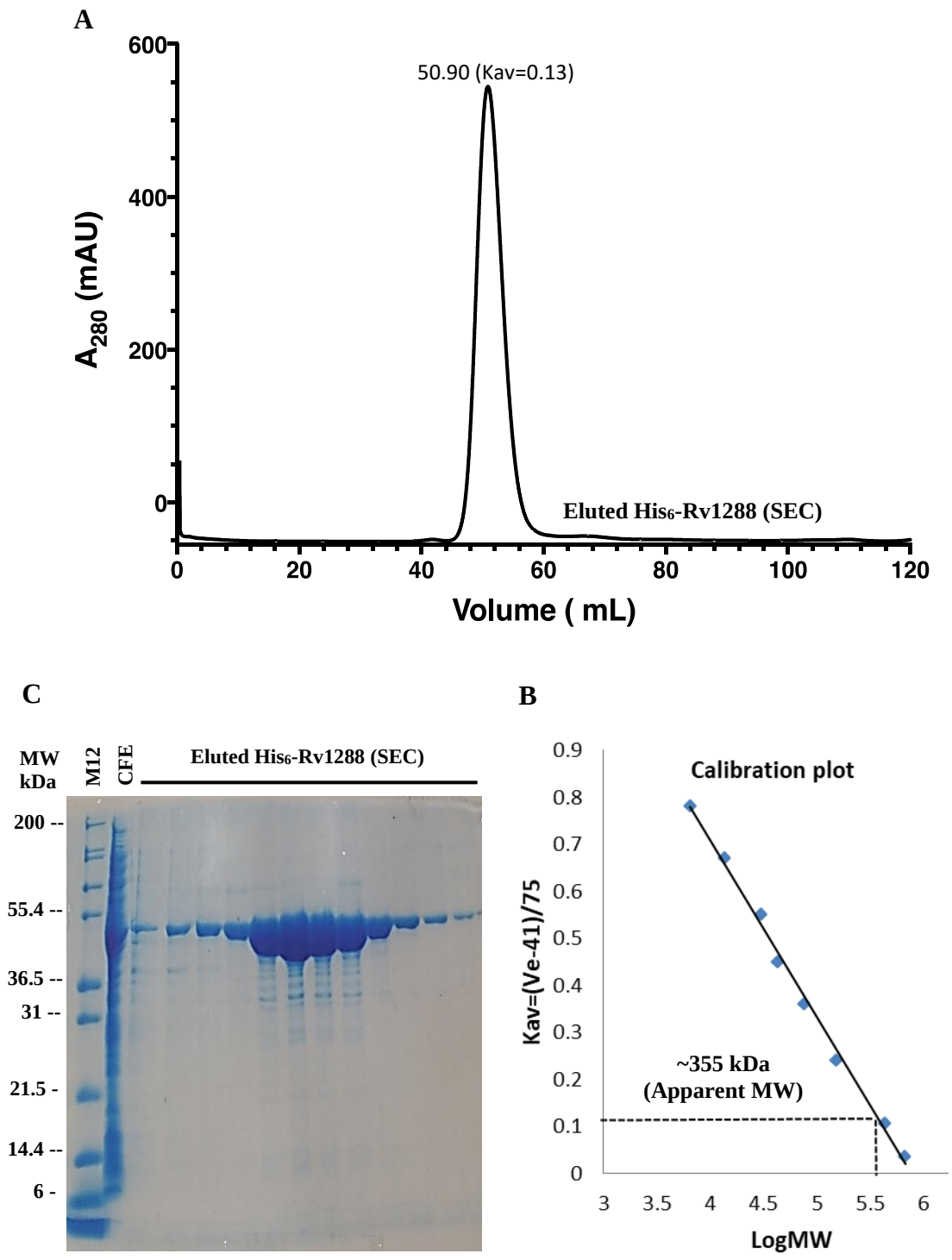


Figure 3.25: The Purification profile of His_6-Rv1288 by size exclusion chromatography. A) Chromatogram profile from SEC utilizing a Superdex-200pg column. His_6-Rv1288 was fractionated and eluted in Buffer A at elution column volume 50.90 mL, gave the Kav value ~0.13. B) The apparent MW of His_6-Rv1288 was ~355 kDa. The Kav (0.13) was plotted against the calibration plot to determine the apparent MW of the protein. C) The SDS-PAGE gel of purified His_6-Rv1288 by SEC. The purity of the final product was estimated to be ~90%.

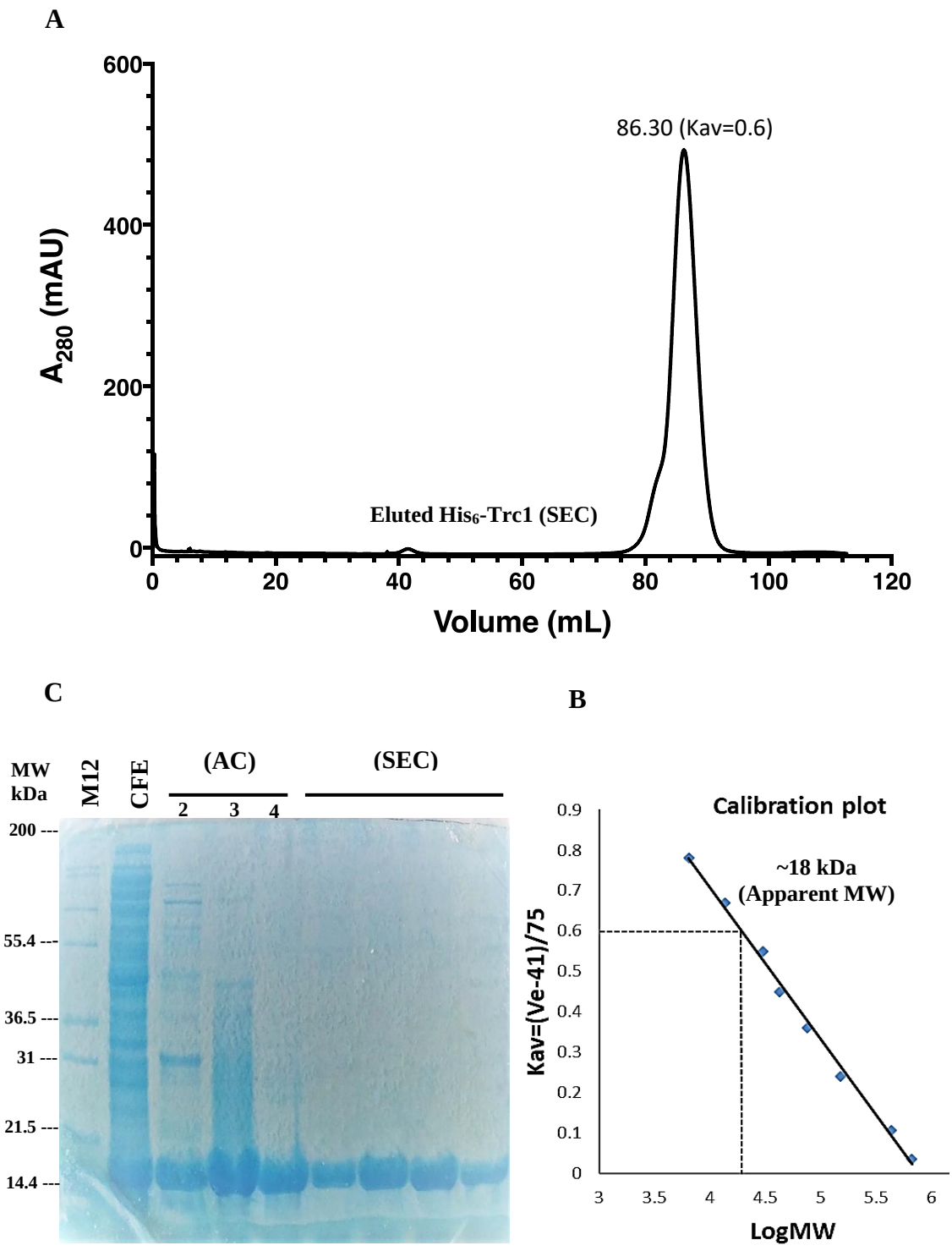


Figure 3.27: Purification profile of His_6-Trc1 by size exclusion chromatography. A) Chromatogram profile from SEC utilizing a Superdex-200pg column. His_6-Trc1 was fractionated and eluted in Buffer A at elution column volume 86.30 mL, gave the Kav value ~0.6. B) The apparent MW of His_6-Trc1 ~18 kDa. C) The SDS-PAGE gel of purified His_6-Trc1 by SEC. The purity of the final product was estimated to be ~90%.

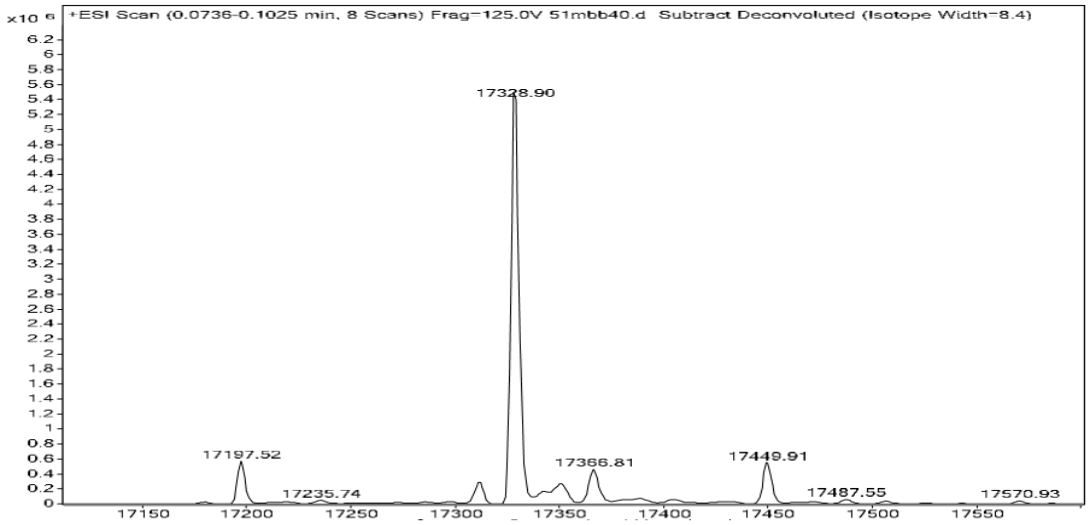


Figure 3.28: Mass spectrometry profile of His_6-Trc1. A mass spectrum peak obtained from the MS analysis showed the presence of His_6-Trc1 wit molecular mass (m/z) 17328.90 Da.

3.10 Analysis of Trc1 in solution by tryptic digest

A tryptic digest was carried out for the Trc1 protein to determine whether the LysM domains fold to form a globular protein as in the structure of AtCERK1, a plant protein containing three LysM domains, or whether the three LysM domains of Trc1 are connected by flexible linkers which might be susceptible to be attacked and cleaved by a protease. The amino acid sequence of Trc1 was analyzed to identify the position of lysine (K) and arginine (R) residues that might form the target for tryptic cleavage which could indicate they were exposed in the structure. There were eight possible cleavage sites for trypsin observed in the Trc1 sequence spread over each of the LysM domains and some of which are close to the inter-domain linkers.

The Trc1 was treated with 10 μg trypsin at 1:1 ratio (w/w). The treated protein was incubated at 37^{o}C for 30 min, 60 min, 120 min, 180 min and overnight. All the treated samples were kept in ice after the treatment to avoid further reaction between the enzyme and the protein. The digested Trc1 was subjected to the 12% SDS-PAGE gel (Figure 3.30). None of the treated samples showed evidence of cleavage. These suggest that the three LysM domains in the Trc1 protein are tightly packed and that, the lysine and arginine residues are not accessible to trypsin as in the structure of AtCERK1 from *A. thaliana* (PDB ID 4EBZ). Therefore, the idea that the 'individual functional LysM

module' in tandem repeats of LysM domains act as 'beads on a string' suggested in AtlA is not applicable to Trc1 of Rv1288.

MVSTHAVVAGETLSALALRFYGDAELYRLIAAASGIADPDVVNVGQRLIMPDFTR
YTVVAGDTLSALALRFYGDAELNWLIAAASGIADPDVVNVGQRLIMPDFTRYTVV
AGDTLSALAARFYGDASLYPLIAAVNGIADPGVIDVGQVLVIFIGLEHHHHHH

Figure 3.29: Possible cleavage sites for trypsin on the His6-Trc1 protein. The eight potential cleavage sites are highlighted in blue boxes. The residues colored in blue are Domain 1, green is Domain 2 and orange is Domain 3.

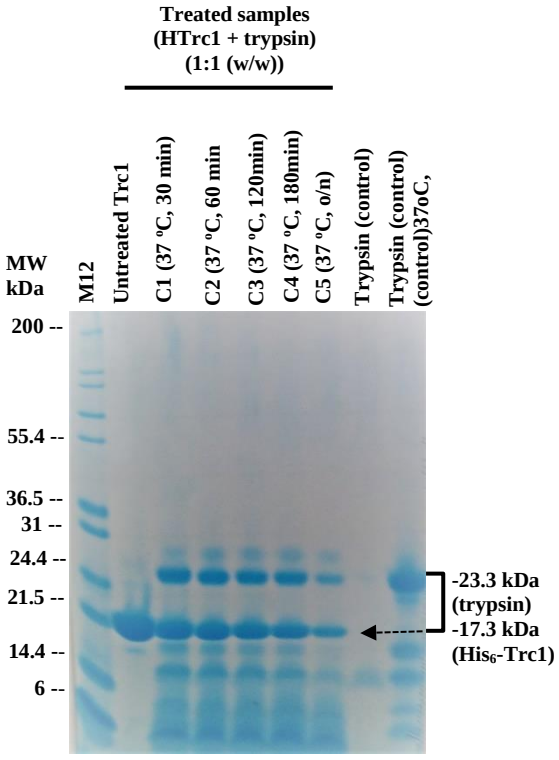


Figure 3.30: SDS-PAGE gel for trypsin-treated Trc1. The protein samples were treated with 10 µg trypsin (final concentration) at 1:1 (w/w). The experiment was carried out at five different incubation settings; C1 (37 °C, 30 min), C2 (37 °C, 60 min), C3 (37 °C, 120 min), C4 (37 °C, 180 min) and C5 (37 °C, overnight). The Trc1 protein was not digested by trypsin as there are no bands with smaller MW in treated samples for the bands which present in the trypsin control sample.

3.11 Crystallization trials on Rv1288 and Trc1

The purified Rv1288 protein was concentrated to 40 mg/mL by using a VIVASPIN device (30, 000 MWCO, Sartorius), and was desalted in freshly prepared of 10 mM Tris (pH 8.0) using a Zeba Spin Desalting column (Thermo Scientific). Crystallization trials were carried out on the protein using a Matrix Hydra II PlusOne crystallization robot (BioMATRIX), applying a sitting drop method. The protein was dispensed into 96-well MRC2 sitting-drop crystallization trays in 1:1 ratio of proteins, precipitant generating a 200 nL drop, which was allowed to equilibrate through vapor diffusion at

19^{o}C. Commercially available crystallization screens from Molecular Dimension (Morpheus, $AmSO_4$, JCSG, MPD, PACT, and ProPlex) were used to identify conditions that formed crystals. Unfortunately, none of the drops showed any crystal growth.

The purified Trc1 protein was concentrated to 40 mg/mL by using a VIVASPIN device (10, 000 MWCO, Sartorius), and was desalted in freshly prepared of 10 mM Tris (pH 8.0) using a Zeba Spin Desalting column (Thermo Scientific). The crystallization trials for Trc1 were set up using a similarapproach as for the Rv1288 protein. An apo His_6-Trc1 crystal was successfully grown in ProPlex solution (F11) containing 0.1 M sodium acetate (pH 5) and 1.5 M ammonium sulfate, and the crystal grew as a plate-shaped crystal cluster (Figure 3.31 A). The crystal was sent to the Diamond Light source to judge whether it was a protein or salt. On the Beamline IO4-I, the crystals showed diffraction spots to ~4 Å consistent with them being crystals of protein rather than salt as a result, the spots are close together indicating a large unit cell (Figure 3.32.). However, the attempt to optimize the conditions did not result in significant improvement and further trials are required. This gives hope that Trc1 can be crystallized and possibly a structure of the protein can be obtained.

Equivalent crystallization trials of Trc1 with NAG_5 under 1:1 (protein: sugar ratio) yielded small crystals of Trc1-NAG_5 in a few JCSG and ProPlex conditions (Figure 3.31). The crystals were mounted using tennis-typed loops and were sent to the Diamond Light source, but unfortunately, the crystals were too small and grid-scan (Beamline I24) failed to produce significant diffraction. Therefore, further optimization of the crystallization on the Trc1-NAG should be carried out in order to produce bigger crystals.

A

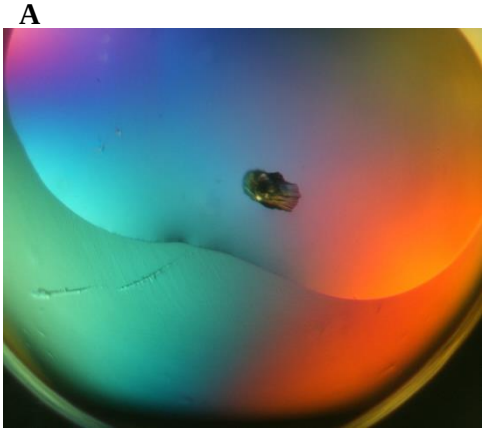
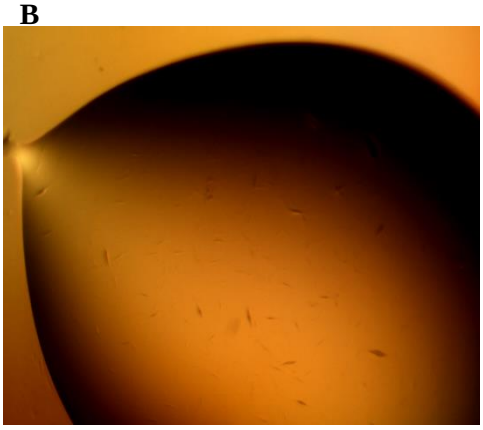


Plate-like apo crystal of Trc1
0.1 M sodium acetate (pH 5)
1.5 M ammonium sulfate
ProPlex (F11)

B

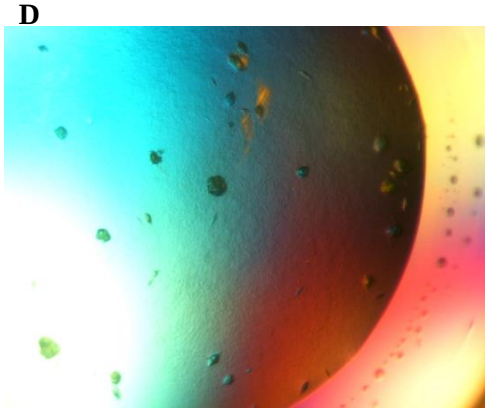


Needle-like crystals of Trc1-NAG_5
1.6 M Sodium citrate tribasic dihydrate
(pH 6.5). JCSG (B11)

C

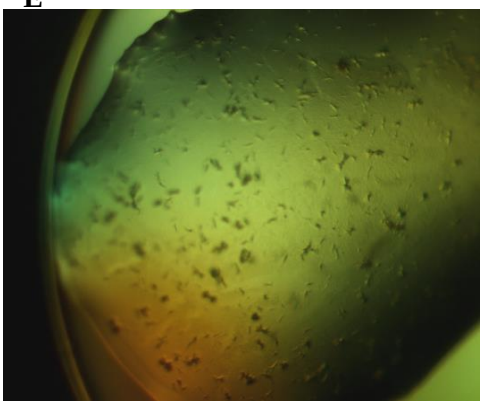
Needle-like crystals of Trc1-NAG_5
0.2 M Magnesium chloride hexahydrate
0.1 M Sodium cacodylate (pH 6.5)
50% v/v PEG 200. JCSG (B10)

D



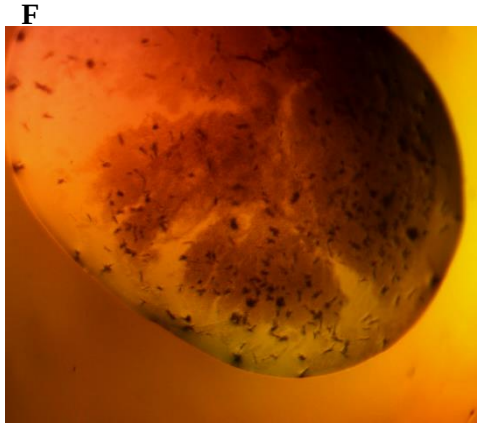
Crystals of Trc1-NAG_5
0.1 M HEPES (pH 7.5)
70% v/v MPD. JCSG (D5)

E



Needle-like crystals of Trc1-NAG_5
3.2 M Ammonium sulfate
0.1 M citrate (pH 5.0)
JCSG (F2)

F



Needle-like crystals of Trc1-NAG_5
2.4 M Sodium malonate dibasic
monohydrate (pH 7.0)
JCSG (F9)

Figure 3.31: Crystals of the apo Trc1 and in complex with NAG_5 oligomers. The crystals were grown at protein concentration ~40 mg/mL. A) A plate-shaped crystal of the apo Trc1 was grown in ProPlex solution. B-F) Crystals of the Trc1-NAG_5 were grown in various conditions of JCSG screening solutions. The crystals of the complex were grown at a concentration of protein to sugar 1:10 (molar ratio).

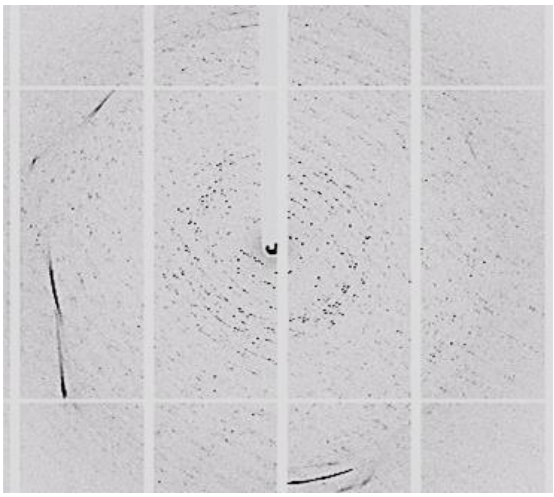


Figure 3.32: X-ray diffraction pattern of the apo Trc1 crystal at ~4 Å. The crystal that grew in a ProPlex solution containing 0.1 M sodium acetate (pH 5) and 1.5 M ammonium sulfate was a protein crystal indicated by the X-ray diffraction pattern obtained from the Diamond Light source, Oxfordshire.

3.12 EM analysis

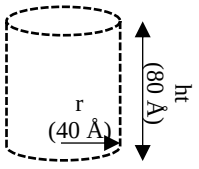
The full-length Rv1288 protein containing the putative esterase domain and the three LysM domains eluted from the size exclusion chromatography with an apparent MW consistent with a possible hexameric quaternary structure. While the Trc1 protein which contains only the three LysM domains eluted as a monomeric species. This suggests that the assembly of the higher-order quaternary structure in Rv1288 mainly involved residues from the putative esterase domain.

As the full-length Rv1288 failed to crystallize, the protein solution was further analyzed under an electron microscope with the hope that the analysis would give some insight into the quaternary structure of the protein, and ultimately how it is folded. The freshly purified Rv1288 (0.05 mg/mL) was imaged in the Electron Microscopy Laboratory in the Department of Molecular Biology and Biotechnology, the University of Sheffield. The protein sample was layered on the coated carbon grid which was initially treated with negative stain solution containing uranium for a dark background. The grid which contained the sample was dried in a drying chamber before it was utilized by the electron microscope. Thirty-six 2D class averages of the Rv1288 were obtained using

single-particle image processing. The averages were of different views of the complex, top or end views, as well as side views.

The Rv1288 molecule appeared as doughnut-shaped from the top and end views (Panel 4: 22-30) giving a diameter for the particle of approximately 80 Å, while from the side view (Panel 1: no. 1-7, Panel 2: 15, Panel 3: 16-19 and Panel 5: 33), the molecule appears as a cylinder with a height of 80 Å. The top and end views provide evidence for rotational symmetry of the particles, but the nature of the symmetry is unclear. In addition, the side views (Panel 1 (no. 1-7), Panel 2 (15), Panel 3 (16-19) and Panel 5 (33)) (Figure 3.33), revealed two identical units are packing against each other. Based on the volume of the particle (assuming a cylinder of radius 40 Å and height 80 Å), the calculated mass of the whole particle of Rv1288 is approximate ~$5.2x10^{-19}$ g, assuming an average density for protein as 1.35 x 10^6 g/m^3 (Andersson, K. & Hovmöller, 1998). Given the molecular mass of one subunit of the protein (~$8x10^{-20}$ g), this suggests that the whole particle of the Rv1288 protein, imaged under EM contains six subunits. This result is consistent with the gel filtration analysis in which the protein (~51 kDa) was eluted with apparent MW ~355 kDa suggesting that it is a hexamer.

Equation 1 **Volume (($Å^3$): πr^2 x ht)**



The cylinder represents a 2D average image of *Rv1288*

Volume (m^3): 22/7 x 40Å x 40Å x 80Å
: 400, 000 $Å^3$
: 4 x 10^5 $Å^{3*}$
: 4 x 10^5 x 10^{-30} m^3
*1Å=10^{-10}m

Equation 2 **Mass: Density x Volume**

1. Mass of *Rv1288* in an image (Figure 3.14)

Average protein density (MW dependent)
: 1.35 x 10^6 g/m^3

Mass (g): Density x volume
: (1.35 x 10^6) 4 x 10^5 x 10^{-30}
: 5.2 x 10^{-19} g

2. Estimation of number subunits in an image of *Rv1288*

A) Mass of 1 particle: 51000 / 6 x 10^{23}
: (8.5 x 10^{-20} g)

B) Total number of particle/subunits per-image
: (5.2 x 10^{-19}) / (8.5 x 10^{-20}) g
: 6.1 particles/subunits

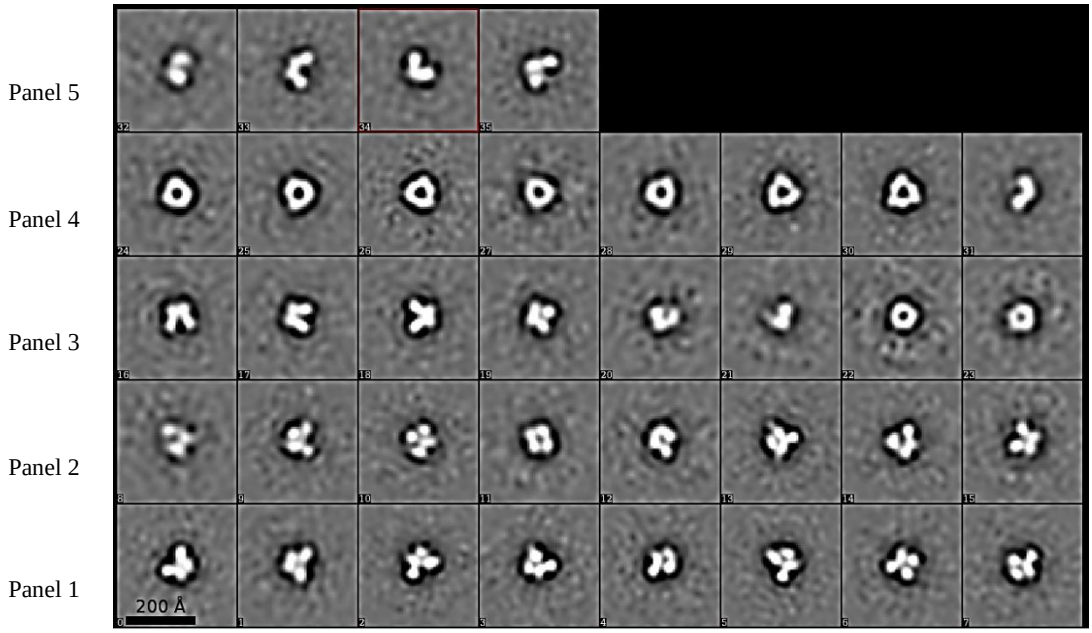


Figure 3.33: 2D class averages of negatively stained Rv1288 observed under the electron microscope at 100kV. The complex averages were from three different views; top and end view (Panel 4 (no.22-30), and side views (Panel 1 (no. 1-7), Panel 2 (15), Panel 3 (16-19) and Panel 5 (33)) (Figure 3.14).

3.13 Second target

Following the study of Rv1288 and its truncated gene, Trc1 from *M. tuberculosis*, the second target gene which was KEG15107 from *M. avium* was recruited in this study. Section B describes the works and results of KEG15107 from *M. avium.*

3.14 Target selection and cloning

The interest of the laboratory in proteins containing the LysM binding domain initially involved studies on *MSMEG3288* from *M. smegmatis*. This study was carried out by Prof. David W. Rice and Dr. C. Bisson to address questions concerning the arrangement of multiple LysM domains in a protein and to study the interaction between the protein and oligosaccharide moieties. However, the crystals of MSMEG3288 did not diffract to high resolution and therefore the three-dimensional structure of the protein lacked the detail that could be provided by such a study. Therefore, in order to provide this level of detail particularly in oligosaccharide recognition by multiple LysM domains, other proteins containing such multiple domains were identified for structural studies.

3.15 BLAST Analysis

MSMEG3288 contains four tandem LysM domains which, unusually, are not covalently attached to a catalytic domain. BLAST searches were performed using a LysM sequence motif (YTVXXGDTLXXIA) or the entire protein sequence of *MSMEG3288* to identify possible targets. This led to the identification of a number of LysM containing proteins that could form interesting molecules for structural studies. Of particular interest were LysM containing proteins from *Mycobacterium* species, *Nocardia* species, and *Rhodococcus* species, all of which are acid-fast Gram-positive bacteria from the corynebacteria genus. Other proteins containing LysM domains were identified from BLAST searches were proteins containing multiple LysM domains from *B. pseudomallei* (*BPSL1345)*, *S. aureus* (*V070_01208*) and *A. baumannii* (*KCZ33729*). The LysM domains of the latter proteins have transglycosylase (*B. pseudomallei*), esterase (*M. tuberculosis*), or amidase (*S. aureus*) activity or, in the case of *A. baumannii*, a SafA domain that is involved in spore coat assembly. Multiple sequence alignment on these selected targets shows that all the LysM homologs share low sequence similarity to each other (Figure 3.34). One feature of interest was the presence of a YGD motif observed in the LysM sequences from *Mycobacterium* sp, *Nocardia* sp, and *Rhodococcus* sp. This raises questions as to what the motif does, why it is only observed among these species and is it of importance for the biological function of the proteins? The architecture of the multiple LysM domains of the selected proteins is displayed in Figure 3.35.

The homolog of MSMEG3288 from *M. smegmatis* identified in the blast search that forms a major focus for the work described in the thesis is KEG15107 from *M. avium*. This protein is identical in domain architecture to MSMEG3288 and shares 65% sequence identity. The following sections discuss aspects of the work carried out on KEG15107.

3.16 KEG15107

The genome sequence of *M. avium* strain Env 77 derived by whole-genome ShotGun-based sequence consists of 772 contigs (Hsu, Wu, and Talaat, 2011). The KEG15107 gene was identified in the genome of *M. avium* strain ENV 77 under a locus_tag (KEG_RS0114950), from contig345 (NZ_AGAQ01000345), with a complete DNA

sequence (cds) from 992 – 1639. The gene contains 648 nucleotides and encodes a protein with 215 amino acids including a start codon and a stop codon with protein_id (WP_019737356). This gene, which is assigned as a putative peptidoglycan binding protein, is located close to genes encoding peptidase and amidase activities (Figure 3.36 (A)). The domain architecture of KEG15107 is shown in Figure 3.36 (B). Each of the four LysM domains of KEG15107 shares low sequence identity (~16~36 %) to each other but domains 1, 3 and 4 shares an identical short motif YGD (Domain 2 has YGT) at the position 21-23 (Figure 3.36 (C)).

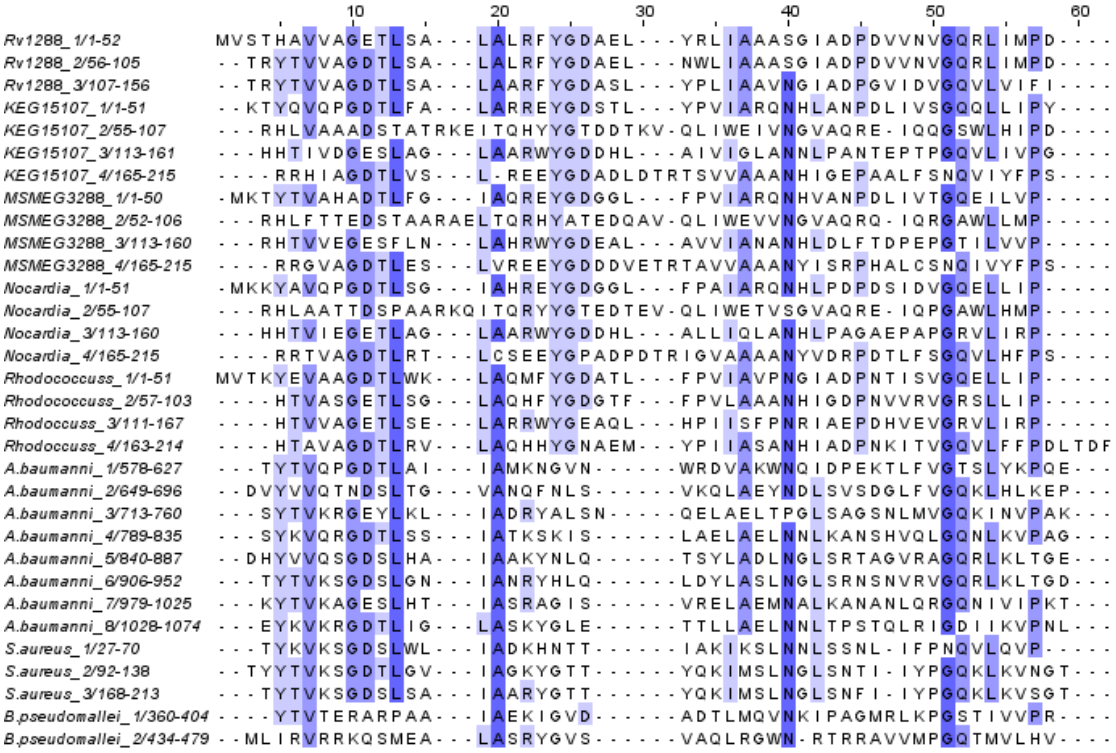


Figure 3.34: Multiple sequence alignment of a subset of proteins containing multiple LysM domains identified by BLAST including *M. tuberculosis* Rv1288 (3x LysM with an attached C-terminal esterase domain), *M. avium* KEG15107 (4x LysM domains), *M. smegmatis* MSMEG3288 (4x LysM domains), *Nocardia* sp (4x LysM domains), *Rhodococcus* sp (4x LysM domains), *A. baumanii* (8x LysM with an attached N-terminal SafA domain), *S aureus* (3x LysM with attached C-terminal amidase domain) and *B. pseudomallei* (2x LysM with an attached N-terminal transglycosylase domain).

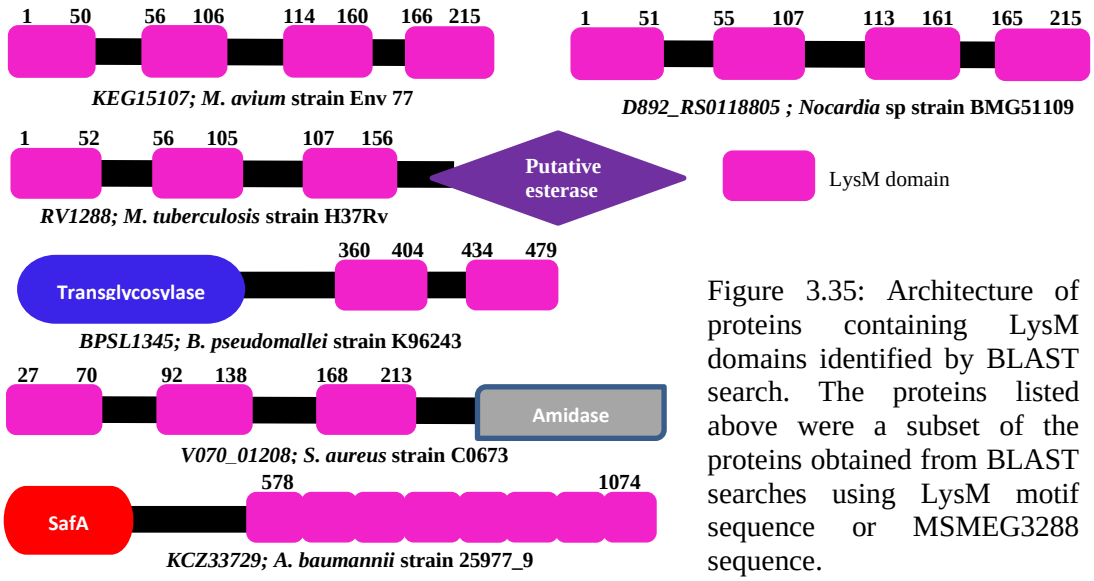


Figure 3.35: Architecture of proteins containing LysM domains identified by BLAST search. The proteins listed above were a subset of the proteins obtained from BLAST searches using LysM motif sequence or MSMEG3288 sequence.

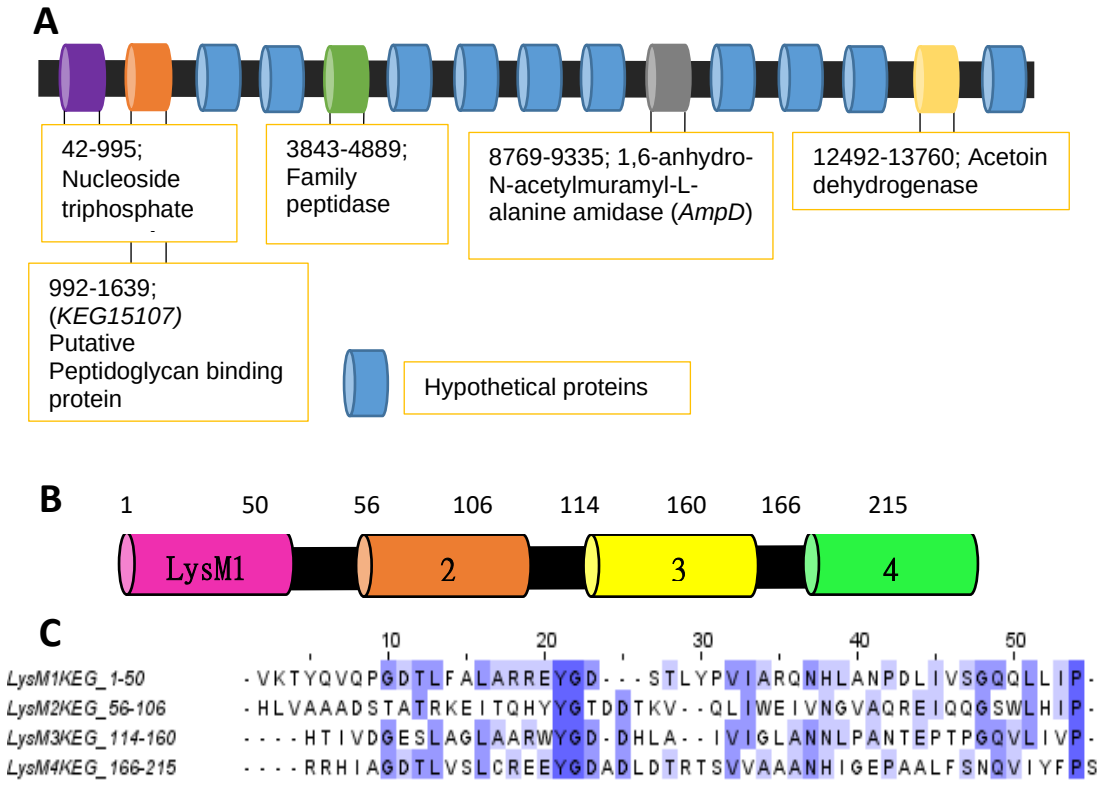


Figure 3.36: A schematic diagram to show the position of the gene that encodes the LysM containing protein KEG15107 and its domain architecture. A) The KEG15107 gene (KEG_RS0114950) in *M. avium* strain Env 77 encodes a putative peptidoglycan binding protein LysM (WP_019737356), and the cds of the gene are located from position 992-1639. The gene is close to other genes that encode a peptidase and an amidase, enzymes that might be responsible for peptidoglycan hydrolysis. B) KEG15107 encodes four repeating LysM domains (~50 amino acids), each of it is separated by a short segment of sequence. C) Multiple sequence alignment of the four LysM domains of KEG15107. Identical residues of the four LysM domains are highlighted in dark blue color with lower level similarities showed in light blue.

3.17 PCR product of KEG15107

Based on the complete DNA sequence of KEG15107 obtained from the NCBI database, the first three nucleotides of the gene sequence encode the unusual start codon which is valine (GTG) instead of methionine (ATG). Therefore, primers for KEG15107 were designed based on the DNA sequence which included the addition of an extra methionine residue prior to the valine. The forward and reverse primers (Table 3.2) were designed to create a construct containing the KEG15107 gene, followed by restriction sites for *NcoI* (CCATGG) and *XhoI* and six extra residues (ATATAT) for both the 5' and 3' ends of the nucleotides respectively. The full-length KEG15107 gene was successfully amplified by PCR using a set of primers at the annealing temperature of 66 ^{o}C giving a total product size 668 bp (Figure 3.37). The PCR product of the KEG15107 gene was subjected to gel electrophoresis, and the expected bands were visualized under UV light (Figure 3.38). The optimized PCR conditions are described in Table 3.3.

Table 3.2: Primers for the KEG15107 gene construct

Forward primer	*5'ATATATCCATGG TGAAGACCTACCAAGTCCA 3'*
Reverse primer	*5'ATATATCTCGAG GGAGGGGAAATAGATCACCTG 3'*
Tm (^{o}C)	68 and 64
Annealing (^{o}C)	66
Restriction sites	*NcoI* (forward) and *Xho*I (reverse)
Nucleotides	645 bp (GTG is a start codon)

Table 3.3: Optimized PCR conditions for KEG15107

PCR mixture		
PCR components	Total volume (μL)	
Q5® Hot Start High-Fidelity 2X Master Mix	25.0	
Forward primer (20 pmol)	1.0	
Reverse primer (20 pmol)	1.0	
DNA template (25.0 – 100 ng)	-based on the stock concentration of the DNA	
miliQ water (5% DMSO)	-added to total volume	
Total volume	50.00	
PCR parameters		
Program	Temperature (^{o}C) / Time	
Pre-denaturation	98.0 / 5.0 min	
Denaturation	98.0 / 30 sec	
Annealing	66.0 / 30 sec	repeat the steps for
Extension	72.0 / 1.0 min	25 cycles
Post-extension	72.0 / 5.0 min	
Hold	5.0 / ∞	

5′*ATATATCCATGG*TGAAGACCTACCAAGTCCAGCCGGGCGACACCCTGTTCGCCCTGGCCCGGCGCGA
GTACGGTGACAGCACCCTGTACCCGGTGATCGCGCGGCAGAACCATCTCGCCAACCCGGATCTGATCGT
GTCCGGGCAGCAGCTGCTGATCCCGTACGTGACCTATCGACACCTGGTCGCCGCGGCCGATTCCACCGC
GACCCGCAAGGAGATCACCCAGCACTACTACGGAACCGACGACACCAAAGTGCAGTTGATCTGGGAGAT
CGTCAACGGAGTAGCCCAGCGGGAGATACAGCAGGGCAGCTGGCTGCACATCCCCGACCTGTCCAACGT
CGGGCACCACACGATCGTCGACGGAGAAAGCCTCGCGGGGCTGGCCGCCCGGTGGTACGGCGACGACCA
CCTCGCGATCGTGATCGGGTTGGCGAACAACCTTCCCGCGAACACCGAACCGACCCCGGGCCAGGTGCT
CATCGTCCCCGGCCTCAACCGGCGCCGCCACATCGCCGGCGACACCCTGGTGTCACTGTGCCGCGAGGA
ATACGGCGATGCGGATCTGGACACCCGGACGTCGGTCGTCGCGGCCGCCAACCACATCGGCGAGCCGGC
CGCGCTCTTCTCCAACCAGGTGATCTATTTCCCCTCC*CTCGAGATATAT* 3′ (668 bp)

Figure 3.37: The nucleotide sequence of KEG15107. The KEG15107 gene was successfully amplified by PCR at an annealing temperature 66°C. The restriction sites for *NcoI* (blue) and *XhoI* (green) are indicated and preceded by extra nucleotides (red) to provide the sticky ends.

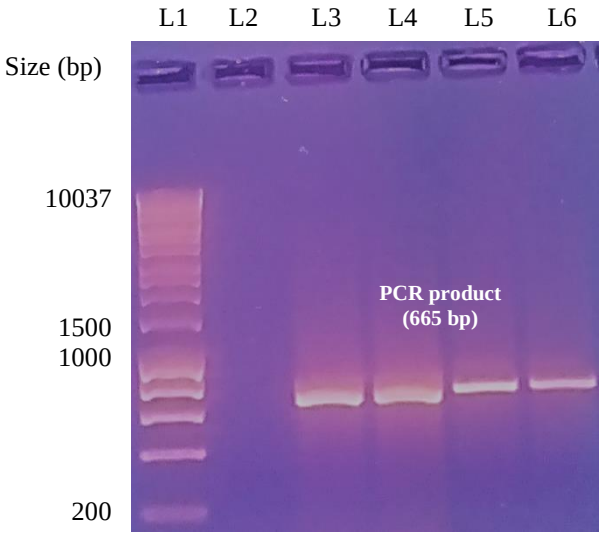


Figure 3.38: PCR product of KEG15107 on a 1% agarose gel. The amplified PCR product was 668 bp in size. L1: Marker; L2: negative control; L3-L6: amplified KEG15107 product

3.18 Digestion product of KEG15107 and pET24d

The PCR product of KEG15107 and pET24d cloning vector were digested by using *NcoI* and *XhoI* restriction enzymes to provide sticky ends for the gene insert and the vector. After digestion, the size of the KEG15107 gene was 654 bp while pET24d was ~5230 bp. The cut DNA of the gene and the vector, both were detected on the agarose gel at the expected sizes (Figure 3.39). The schematic diagram of the digested KEG15107 gene construct and the digested pET24d DNA with sticky ends is shown in Figure 3.40. The digested products of KEG15107 and pET24d were cleaned up from any contaminants using a gel extraction kit. The digested pET24d was further treated with a phosphatase enzyme to remove the phosphate group at the end of the sticky ends of the vector DNA to avoid the re-ligation of the vector. The digested KEG15107 gene contained ATG (methionine), a part of its sticky end which acts as a start codon instead

of GTG (valine), the original start codon annotated in the genome of *M. avium* strain Env 77, to enable the gene to be translated and expressed in *E. coli* cells.

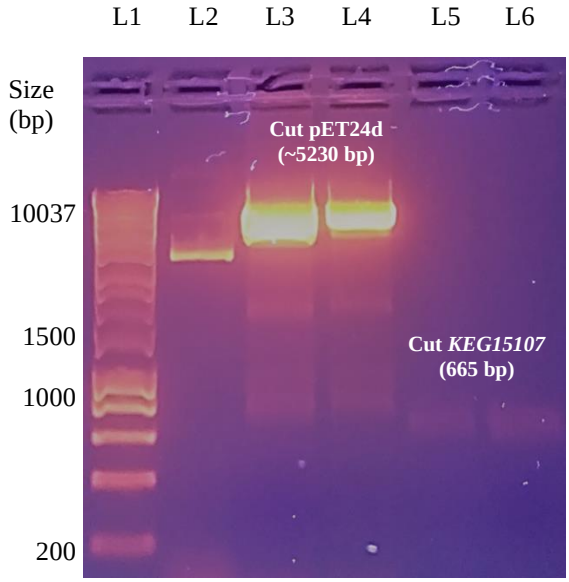
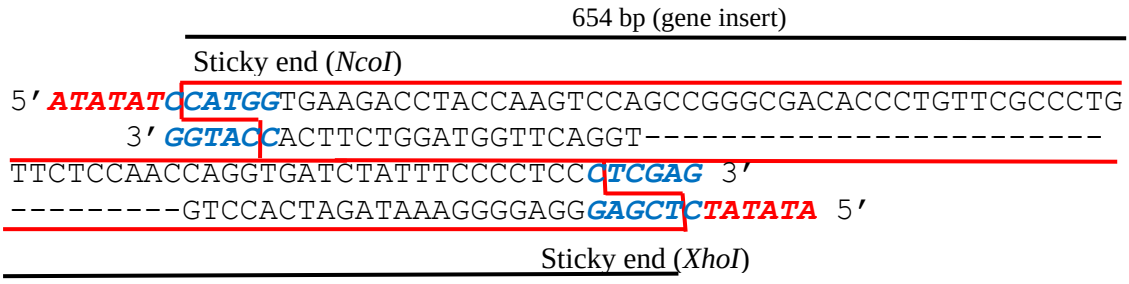
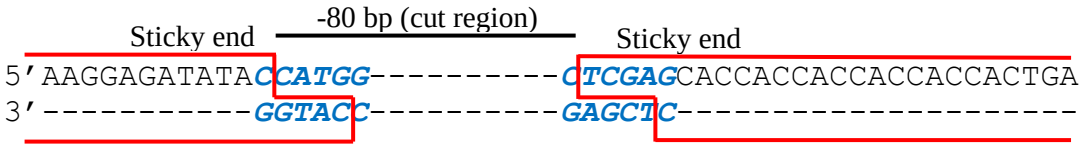


Figure 3.39: Digested DNA of KEG15107 and pET24d on a 1% agarose gel. Both, the KEG15107 gene and pET24d DNA were digested with *NcoI* and *XhoI* restriction enzymes. The final product for the cut KEG15107 gene was 665 bp. Lane1: Marker, Lane2: uncut vector (pET24d) as a control, Lane3-Lane4: cut vector (pET24d), Lane5-Lane6: cut KEG15107.



Digested *KEG15107* DNA



Digested pET24d DNA

Figure 3.40: A schematic diagram of the digested KEG15107 and pET24d constructs. Both, the KEG15107 gene and pET24d were treated with double digestion by *NcoI* and *XhoI* restriction enzymes to provide sticky ends to both ends of the products for DNA ligation experiment.

3.19 DNA ligation product of pET24d-KEG15107

The DNA of the KEG15107 gene insert and the pET24d vector were subsequently ligated using DNA ligase described in Chapter 2 (section 2.8.4). The ligated product of pET24d-KEG15107 was successfully obtained from overnight incubation on ice while no product was detected from samples incubated at room temperature as shown in Figure 3.41.

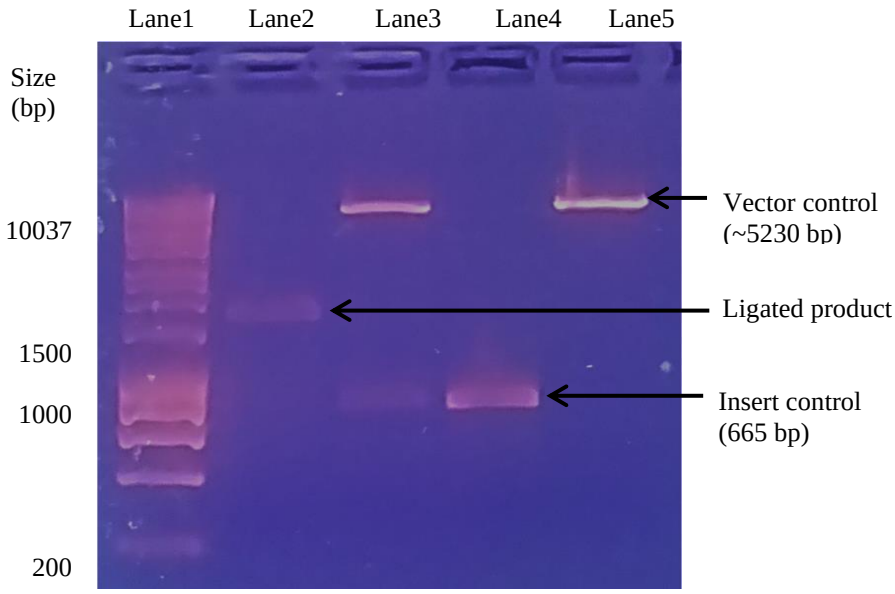


Figure 3.41: Ligation products of KEG15107 and pET24d on the 1% agarose gel. The ligation was carried out at two different incubations, on ice for overnight and 2 hours at room temperature. Lane1: Marker; Lane2, a successful ligation product containing KEG15107 and pET24d from overnight incubation; Lane3: unsuccessful ligation product from 2 hr incubation containing the DNA of digested pET24d and digested KEG15107; Lane4: the DNA of digested KEG15107 as a positive control; Lane5: the DNA of digested pET24d as a positive control.

3.20 Transformants

The recombinant plasmids containing KEG15107 gene were successfully transformed into *E. coli* cells by heat shock transformation. The transformants were initially inoculated into SOC media containing glucose and potassium which helps the recovery of *E. coli* cells after heat shock before plating onto LB media to get the higher efficiency of plasmid transformation into *E. coli* cells. The transformants on the LB agar

supplemented with 50 µg/mL kanamycin, the selectable marker for pET24d is shown in Figure 3.42

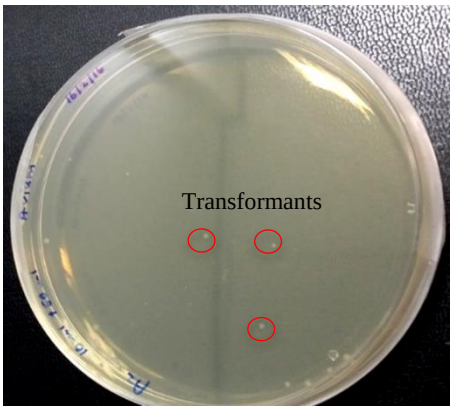


Figure 3.42: Transformants of recombinant plasmid pET24d-KEG15107. The transformants grew on LB media supplemented with 50 µg/mL kanamycin. Selected transformants for colony PCR assay, are highlighted with red circles.

3.21 Colony PCR and sequencing analysis of pET24d-KEG15107

The transformants that grew on the LB media were randomly selected for a colony PCR assay to detect the presence of the ligated KEG15107 gene. T7 primers were utilized to amplify the KEG15107 gene as the vector pET24d utilizes a T7 promoter as a gene transcription regulator in *E. coli*. The amplified PCR product contained a vector sequence from the T7 promoter to the T7 terminator and the KEG15107 gene, giving the total product size 885 bp as shown in Figure 3.43. The end product of the colony PCR was subjected to a 1% agarose gel and the result was visualized under UV light. Three bands appeared on the gel of which two were at the expected size as shown in Figure 3.44. The method of the colony PCR is described in Chapter 2 (section 2.8.6).

The DNA sequencing was carried out for the recombinant plasmid to analyze the inserted gene sequence and its orientation in the plasmid. The results obtained from the DNA sequencing showed that the DNA of the insert was 100% identical to the gene sequence of KEG15107 taken from GenBank database (Figure 3.46). No mutations observed in the sequence of the recombinant plasmid pET24d-KEG15107 vector. The gene sequence for the recombinant plasmid was in the right orientation as the start codon of the gene was next to the TATA box of the vector plasmid. The KEG15107 gene was successfully cloned into the expression vector pET24d with six histidines at the C-terminus of the gene to facilitate purification.

The KEG15107 gene sequence was translated into a protein sequence by ExPASy translate tool, and the results are shown in Figure 3.46. KEG15107 encodes a protein with 224 amino acids including nine extra residues which are methionine (Met) acting as a start codon for the protein in *E. coli*, leucine (L) and glutamic acid (E) as a linker followed by six histidine residues giving a total molecular weight of the protein of 24797.82 Da. The theoretical pI of the protein is 5.87, and the calculated extinction coefficient (Abs 0.1% (mg/mL)) for the protein at 280 nm is 1.3 M^{-1} cm^{-1} computed by the ExPASy protparam tool.

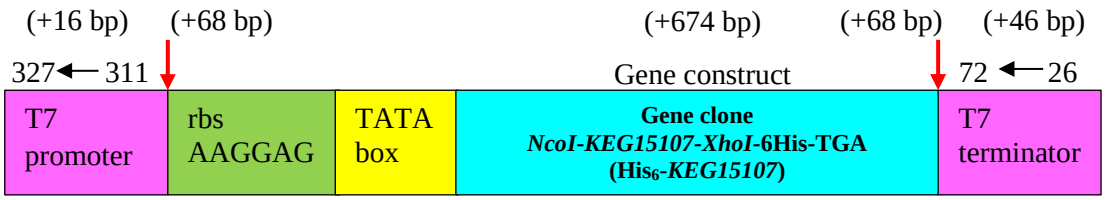


Figure 3.43: A schematic diagram of the recombinant plasmid pET24d-KEG15107 construct. The construct includes the T7 promoter to the T7 terminator of the recombinant plasmid, including the KEG15107 gene construct. The total size of the recombinant plasmid including the gene insert from the T7 promoter to T7 terminator position is 885 bp. The two arrows indicate the position of the additional nucleotide sequence from the plasmid including Lac operator region in between the T7 promoter and rbs sites.

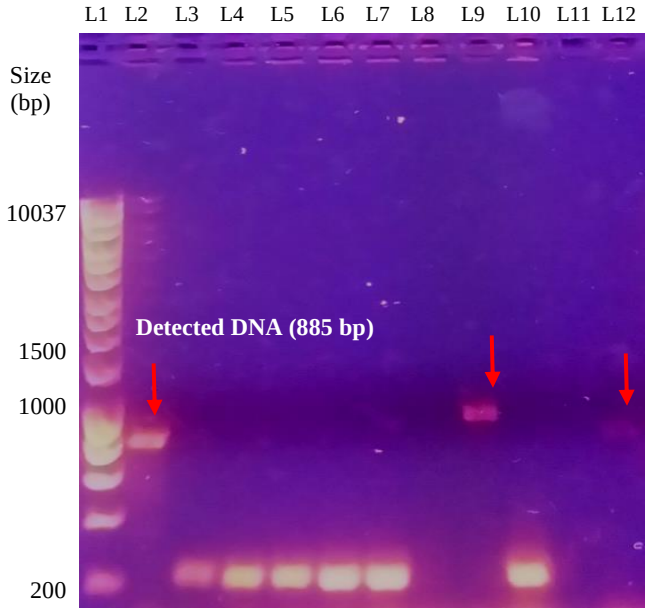


Figure 3.44: Colony PCR of the recombinant plasmid containing KEG15107 on the 1% agarose gel. The colony using T7 primers PCR was performed on 11 colonies obtained from the transformant plate L1: Marker; L3-L7 and L10: nonspecific PCR products; L8 and L11: no PCR product. PCR products from L2, L9, and L12 (red arrows) were sent for sequencing.

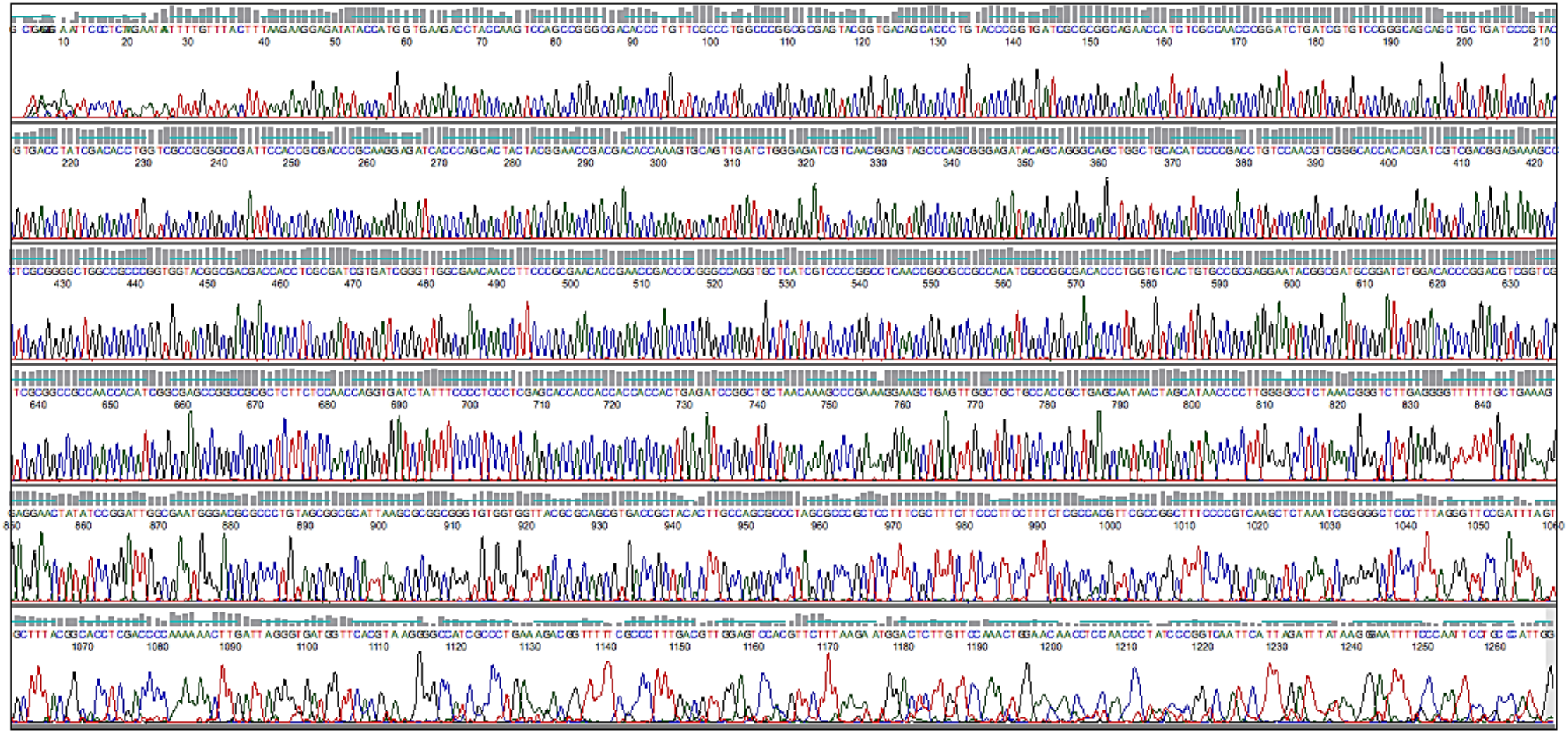


Figure 3.45: Sequencing of the recombinant plasmid pET24d-KEG15107. The KEG15107 sequence was from position 57-732.

```
98CE26          AAGGAGATATACCATGGTGAAGACCTACCAAGTCCAGCCGGGCGACACCC
KEG15107        GT----------------GAAGACCTACCAAGTCCAGCCGGGCGACACCC
                                  ********************************

98CE26          TGTTCGCCCTGGCCCGGCGCGAGTACGGTGACAGCACCCTGTACCCGGTG
KEG15107        TGTTCGCCCTGGCCCGGCGCGAGTACGGTGACAGCACCCTGTACCCGGTG
                **************************************************

98CE26          ATCGCGCGGCAGAACCATCTCGCCAACCCGGATCTGATCGTGTCCGGGCA
KEG15107        ATCGCGCGGCAGAACCATCTCGCCAACCCGGATCTGATCGTGTCCGGGCA
                **************************************************

98CE26          GCAGCTGCTGATCCCGTACGTGACCTATCGACACCTGGTCGCCGCGGCCG
KEG15107        GCAGCTGCTGATCCCGTACGTGACCTATCGACACCTGGTCGCCGCGGCCG
                **************************************************

98CE26          ATTCCACCGCGACCCGCAAGGAGATCACCCAGCACTACTACGGAACCGAC
KEG15107        ATTCCACCGCGACCCGCAAGGAGATCACCCAGCACTACTACGGAACCGAC
                **************************************************

98CE26          GACACCAAAGTGCAGTTGATCTGGGAGATCGTCAACGGAGTAGCCCAGCG
KEG15107        GACACCAAAGTGCAGTTGATCTGGGAGATCGTCAACGGAGTAGCCCAGCG
                **************************************************

98CE26          GGAGATACAGCAGGGCAGCTGGCTGCACATCCCCGACCTGTCCAACGTCG
KEG15107        GGAGATACAGCAGGGCAGCTGGCTGCACATCCCCGACCTGTCCAACGTCG
                **************************************************

98CE26          GGCACCACACGATCGTCGACGGAGAAAGCCTCGCGGGGCTGGCCGCCCGG
KEG15107        GGCACCACACGATCGTCGACGGAGAAAGCCTCGCGGGGCTGGCCGCCCGG
                **************************************************

98CE26          TGGTACGGCGACGACCACCTCGCGATCGTGATCGGGTTGGCGAACAACCT
KEG15107        TGGTACGGCGACGACCACCTCGCGATCGTGATCGGGTTGGCGAACAACCT
                **************************************************

98CE26          TCCCGCGAACACCGAACCGACCCCGGGCCAGGTGCTCATCGTCCCCGGCC
KEG15107        TCCCGCGAACACCGAACCGACCCCGGGCCAGGTGCTCATCGTCCCCGGCC
                **************************************************

98CE26          TCAACCGGCGCCGCCACATCGCCGGCGACACCCTGGTGTCACTGTGCCGC
KEG15107        TCAACCGGCGCCGCCACATCGCCGGCGACACCCTGGTGTCACTGTGCCGC
                **************************************************

98CE26          GAGGAATACGGCGATGCGGATCTGGACACCCGGACGTCGGTCGTCGCGGC
KEG15107        GAGGAATACGGCGATGCGGATCTGGACACCCGGACGTCGGTCGTCGCGGC
                **************************************************

98CE26          CGCCAACCACATCGGCGAGCCGGCCGCGCTCTTCTCCAACCAGGTGATCT
KEG15107        CGCCAACCACATCGGCGAGCCGGCCGCGCTCTTCTCCAACCAGGTGATCT
                **************************************************

98CE26          ATTTCCCCTCCCTCGAGCACCACCACCACCACCACTGA
KEG15107        ATTTCCCCTC------------------------CTAA
                **********                      ** *
```

MVKTYQVQPG DTLFALARRE YGDSTLYPVI ARQNHLANPD
LIVSGQQLLI PYVTYRHLVA AADSTATRKE ITQHYYGTDD
TKVQLIWEIV NGVAQREIQQ GSWLHIPDLS NVGHHTIVDG
ESLAGLAARW YGDDHLAIVI GLANNLPANT EPTPGQVLIV
PGLNRRRHIA GDTLVSLCRE EYGDADLDTR TSVVAAANHI
GEPAALFSNO VIYFPSLEHH HHHH

Figure 3.46: Sequence comparison of the KEG15107 gene construct from the recombinant plasmid with genome sequence and amino acid sequences of His_6-KEG15107. The sequence comparison showed that the recombinant plasmid contained the full length of the KEG15107 gene with the six histag located at the C-terminus of the gene construct. The sequence shows no mutations compared to the original gene sequence. The KEG15107 gene sequence obtained from the sequencing analysis in was translated into an amino acid sequence by the ExPASy translate tool. The amino acid sequence for KEG15107 contains 224 residues.

3.22 Protein over-expression

Small and large-scale protein expression was carried out on the recombinant plasmid containing KEG15107 in BL21 *DE3* cells. The small-scale expression was initially carried out to obtain the optimum conditions for KEG15107 expression then the conditions were used in the larger scale experiment.

3.22.1 Small scale protein expression of KEG15107

Trial experiments of protein expression for the KEG15107 gene were carried out under a range of conditions. The gene was successfully expressed at all the trial conditions but with the different total amount of protein in the soluble fraction. The protein expressed best at either 25 ^{o}C for five hr or 24 hr, and18 ^{o}C for 72 hr (Figure 3.47).

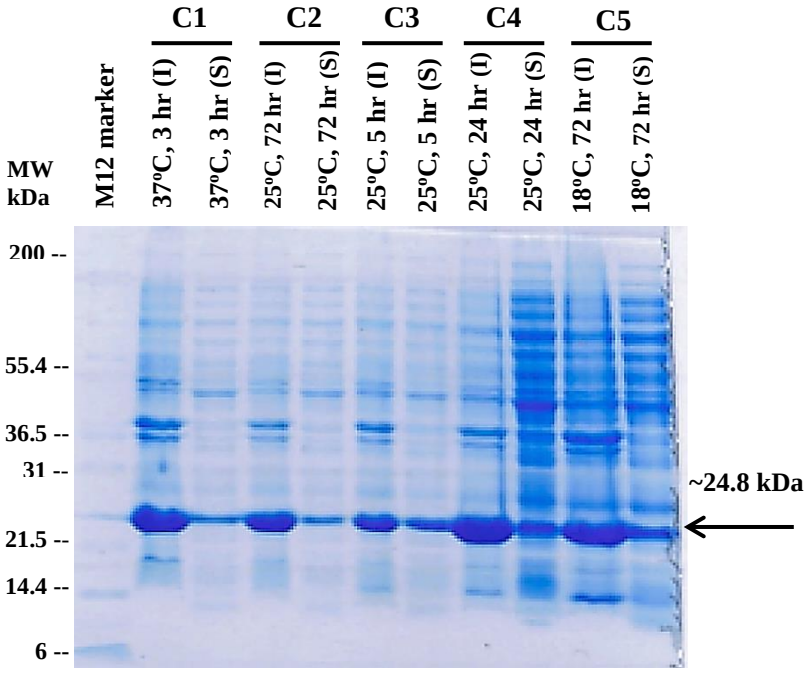


Figure 3.47: SDS-PAGE gel for small-scale expression of KEG15107. The protein expression was performed at five different conditions; (C1): 37^{o}C for 3 hr, (C2): 37^{o}C for 5 hr, (C3): 25^{o}C for 3 hr, (C4): 25^{o}C for 5 hr and (C5): 18^{o}C for 72 hr. SDS-PAGE gel analysis showed that KEG15107 was successfully expressed at all the incubation settings with different amount of protein in the soluble fraction. KEG15107 expressed best under conditions of C3 (25^{o}C, 5 hr), C4 (25^{o}C, 24 hr), and C5 (18^{o}C, 72 hr) respectively. The arrow indicates the position of KEG15107 at the molecular weight (MW) ~24.8 kDa

3.22.2 Large-scale protein expression of KEG15107

Following trial experiments, large-scale expression of KEG15107 was carried out using the optimal condition of 18^{o}C for 72 hr (Figure 3.48). The 2L conical flask containing 500 mL of LB media were grown under this condition, and the cells were harvested by centrifugation at either 10, 000 rpm for 20 min or 5000 rpm for 30 min at 4 ^{o}C. The supernatant was discarded, and the cell paste containing the protein was kept at either -20 ^{o}C for short storage or -80 ^{o}C for long storage.

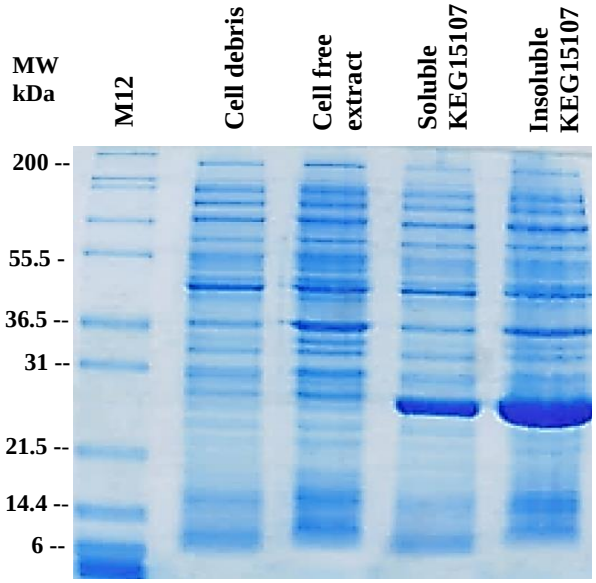


Figure 3.48: SDS-PAGE gel for large-scale expression of KEG15107. The soluble protein was expressed in *E. coli* BL21 *DE3* cells in LB broth induced with a final concentration of 1mM IPTG, at 18 ^{o}C, 250 rpm for 72 hours. L1: M12 Marker, L2: pellet, L3: cell-free extract, L4: soluble protein and L5: Insoluble protein.

3.23 Protein purification

Before protein purification, the cell paste was defrosted and resuspended in Buffer A containing 0.5 M NaCl and 50 mM Tris-HCl buffer (pH 8.0). The cell paste was disrupted by three bursts of sonication on ice, each for 20 seconds. The soluble KEG15107 protein in the cell-free extract (CFE) was separated from the insoluble

fraction by centrifugation at ~45,000g at 4 ^{o}C for 15 min. The protein was purified by affinity chromatography and size exclusion chromatography (SEC) steps.

3.23.1 Purification of KEG15107 by affinity chromatography

The KEG15107 protein was applied to a 5mL HisTrap HP nickel affinity column (GE Healthcare Life Science) and was eluted under a linear gradient 0-70% of 0.5 M imidazole (0-0.35 M) in Buffer A. The protein was fractionated and eluted from the column between a volume ~38 mL to ~50 mL corresponding to approximately 0.2 M imidazole (Figure 3.49 (A)). The KEG15107 protein fractions were subjected to the 12% SDS-PAGE gel to check the purity of the protein which was estimated at ~90% (Figure 3.49 (B)).

3.23.2 Purification of KEG15107 by size exclusion chromatography

The eluted KEG15107 protein from the affinity chromatography column was further purified by size exclusion chromatography (SEC) by using Superdex-200pg (1.6x60 cm HiLoad) in Buffer A. The KEG15107 was successfully fractionated and eluted at an elution volume (Ve) of 69.75 mL as shown in Figure 3.50 (A). Given the calibrated void volume (Vo) and total volume (Vt) of the column (41 mL and 75 mL respectively), gave the Kav value for the KEG15107 protein ~0.38 to give an apparent molecular weight of the KEG15107 protein ~72 kDa (Figure 3.50 (B)). Given the subunit molecular weight of the protein (~24.8 kDa), this suggests that KEG15107 is either a trimer or a tetramer. In this case, KEG15107 could be a tetrameric protein but was eluted through a gel column as a trimeric protein if the protein has binding tendency to the column matrix through hydrophobic interaction. Therefore, the KEG15107 protein maight has a larger retention volume as compared to what it should be. The SDS-PAGE gel (Figure 3.50 (C)) showed that, following SEC, the purity of the product was > 95%.

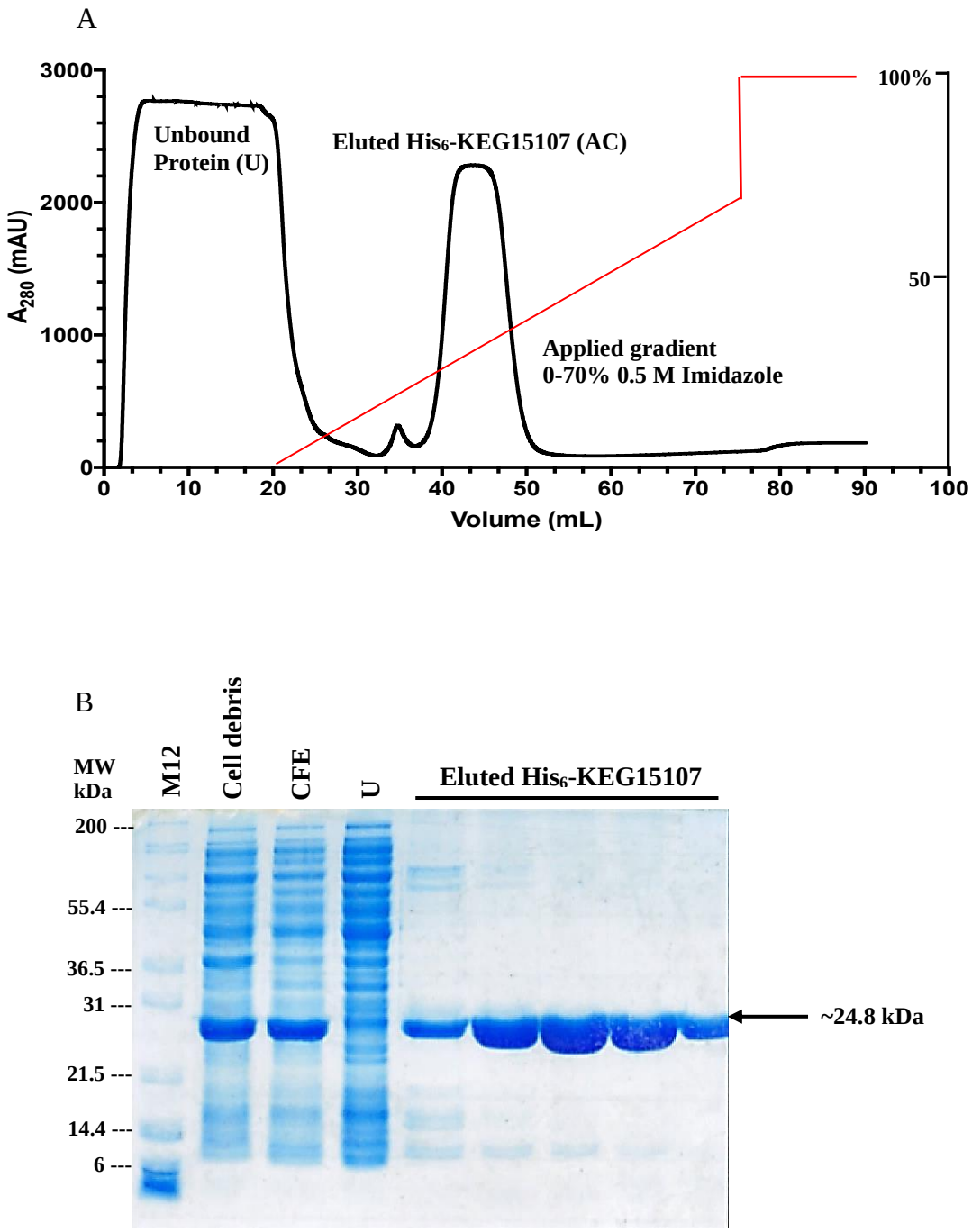


Figure 3.49: The His_6-KEG15107 purification by affinity chromatography. A) Chromatogram profile from the affinity chromatography using a 5mL HisTrap HP column. KEG15107 was fractionated under a linear gradient of 0.5 M imidazole (0-70%) in Buffer A, and the protein was eluted at a single peak at ~0.2 M imidazole. B) SDS-PAGE gel confirmed the presence of His_6-KEG15107 at its expected molecular weight of ~24.8 kDa.

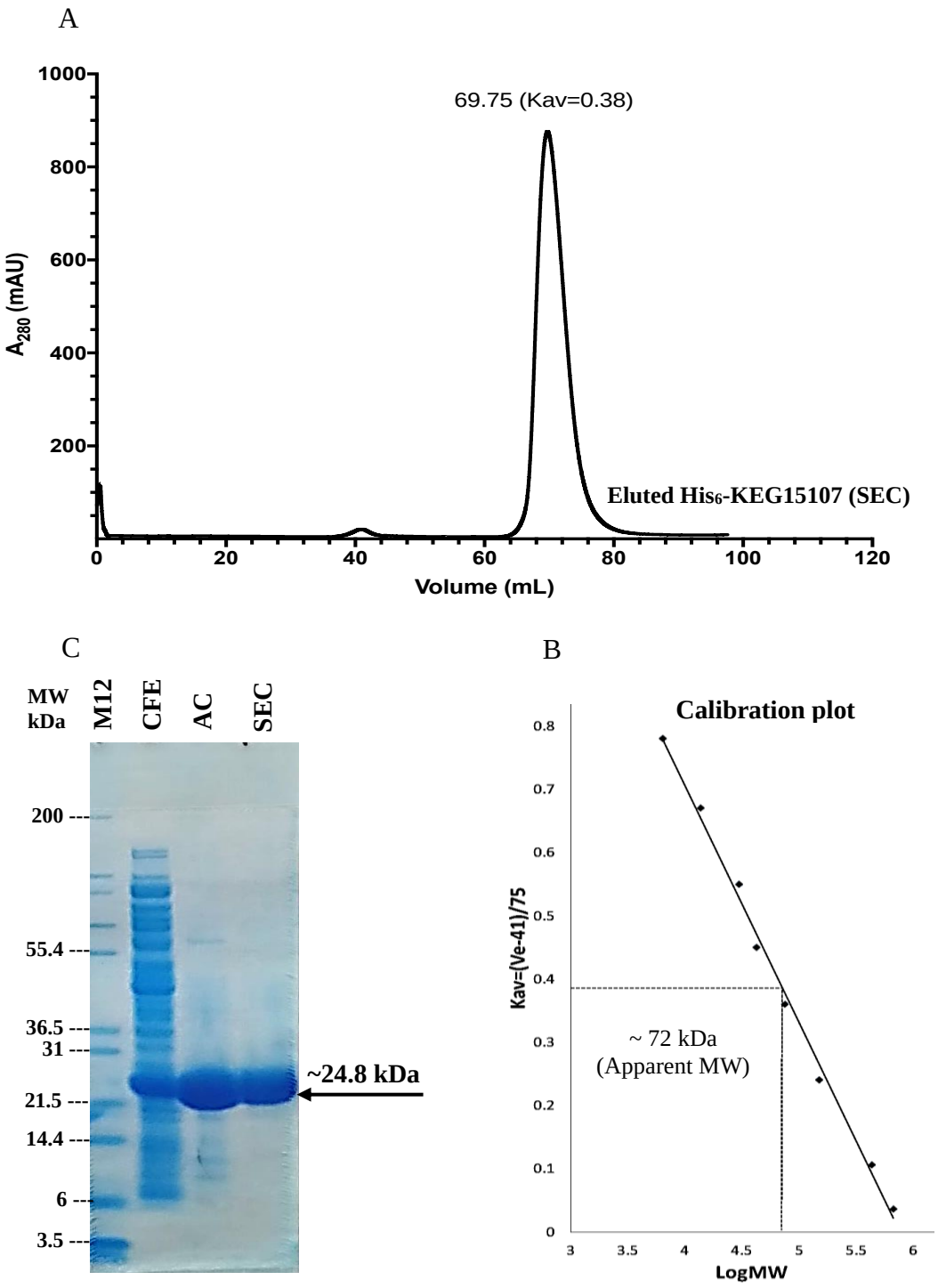


Figure 3.50: The purification profile of His_6-KEG15107 by size exclusion chromatography. A) Chromatogram profile from SEC utilizing a Superdex-200pg column. KEG15107 was fractionated and eluted in Buffer A at elution column volume 69.75 mL, gave the Kav value ~0.38. B) The apparent MW of Hs_6-KEG15017 was ~72 kDa. The Kav (0.38) was plotted against the calibration plot to determine the apparent MW of the protein. C) The SDS-PAGE gel of purified KEG15107 by SEC. The purity of the final product was estimated to be >95%.

3.24 Mass spectrometry analysis on KEG15107 and NAG oligomers

Mass spectrometry analysis was initially performed on KEG15107 to confirm the molecular weight of the corresponding protein. About 2µg/µL of the KEG15107 protein was utilized for mass spectrometry analysis. The mass spectrum results showed that there were two major proton peaks (species A and species B) at molecular mass over charge (m/z) ~24667 and ~24798 Da (Figure 3.51). Species B corresponds to the full-length protein whereas species A reflected the removal of the N-terminal methionine residue. This suggests that the level of methionine amino peptidase activity was not sufficient to fully remove the N-terminal methionine. The other peaks with a larger molecular weight of ~22 Da are thought to be due to the presence of sodium ions present in the buffer during protein purification.

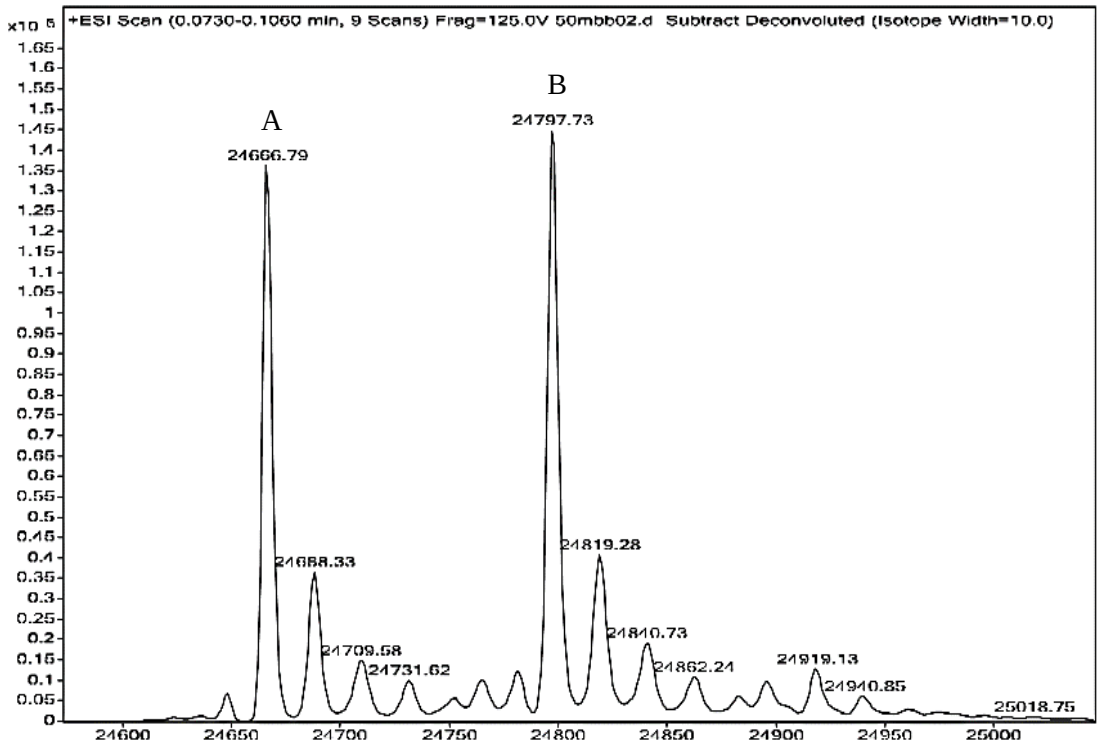
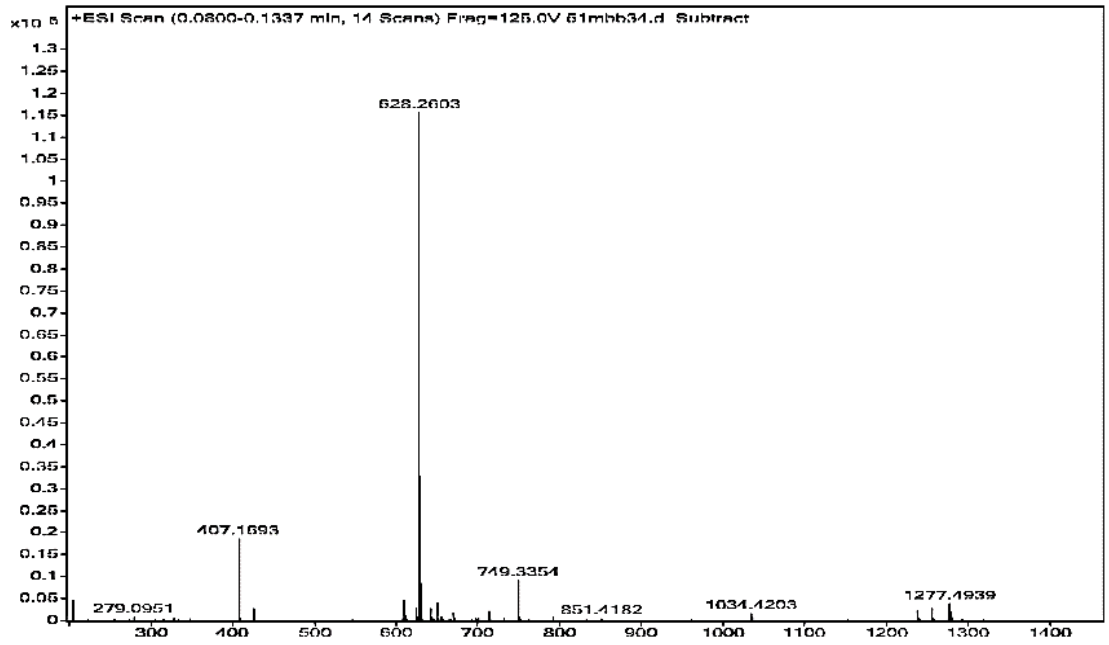


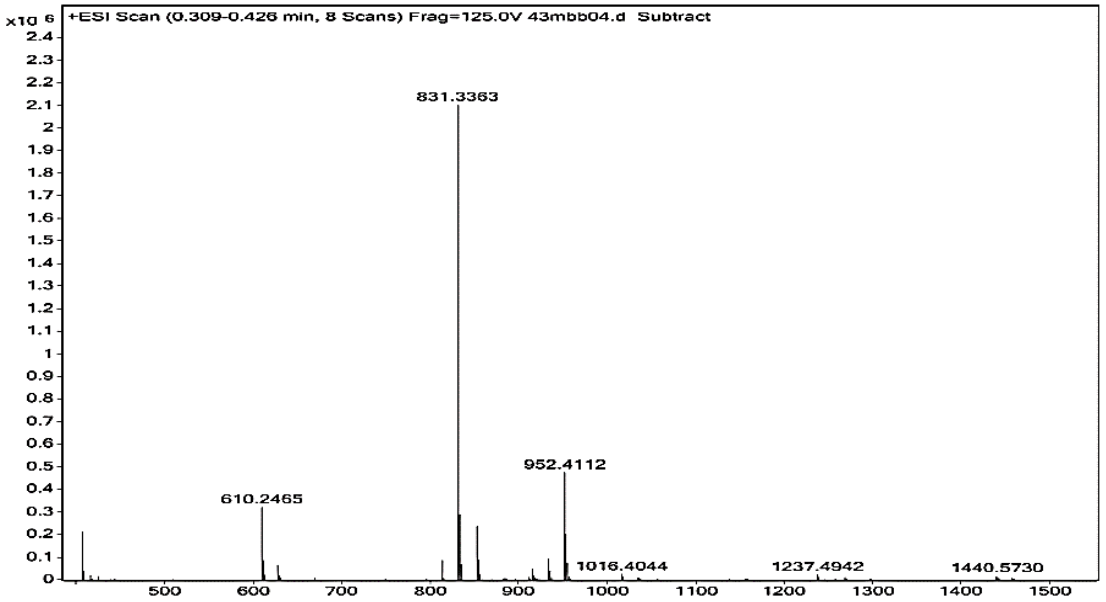
Figure 3.51: The Mass spectrometry profile of His_6-KEG15107. A chromatogram peak obtained from the MS of KEG15107 showed a presence of two species polypeptides; A and B with molecular mass (m/z) 24667 Da and 24798 Da respectively. Species A represents the removal of the N-terminal Met from the full-length protein while species B is the full-length protein.

Mass spectrometry analysis was performed on NAG oligomer samples (NAG_3 to NAG_6) to determine the molecular weight and the purity of the molecules for crystallization purposes. The NAG oligomer samples were prepared at a concentration of 40 mM and mass spectrometry analysis showed that NAG_3 has a mass over charge (m/z) (~628 Da), NAG_4 (~831 Da), NAG_5 (~1034 Da) and NAG_6 (~1237 Da) (Figure 3.52 (A-D)). The purity of NAG_3 was estimated at ~85%, NAG_4 ~80%, NAG_5 ~50% with NAG_6 being less than 20% pure.

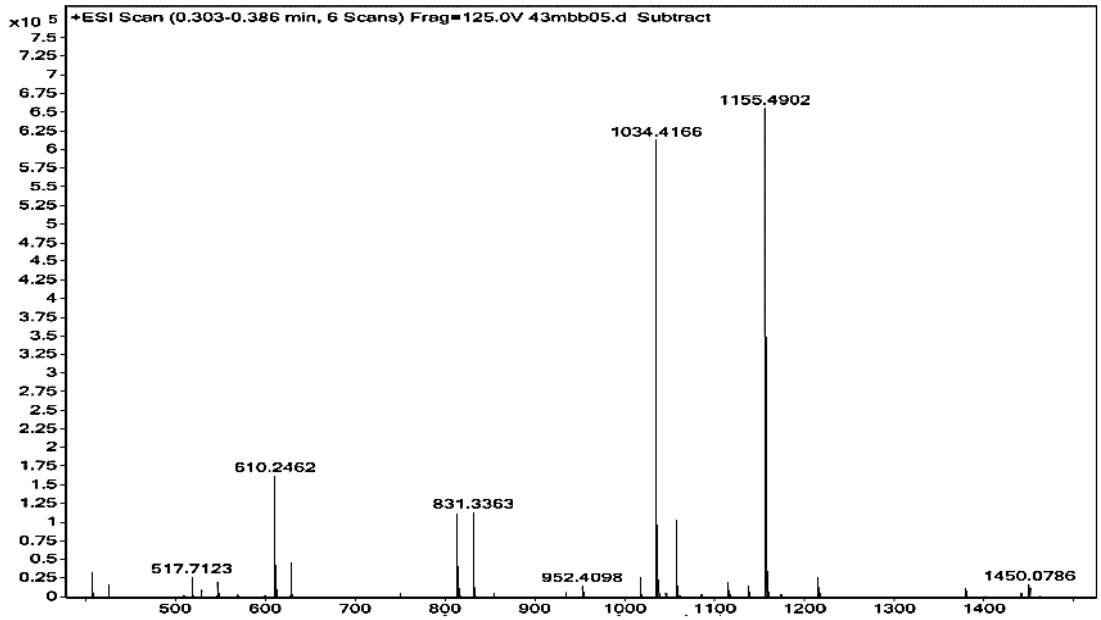
A



B



C



D

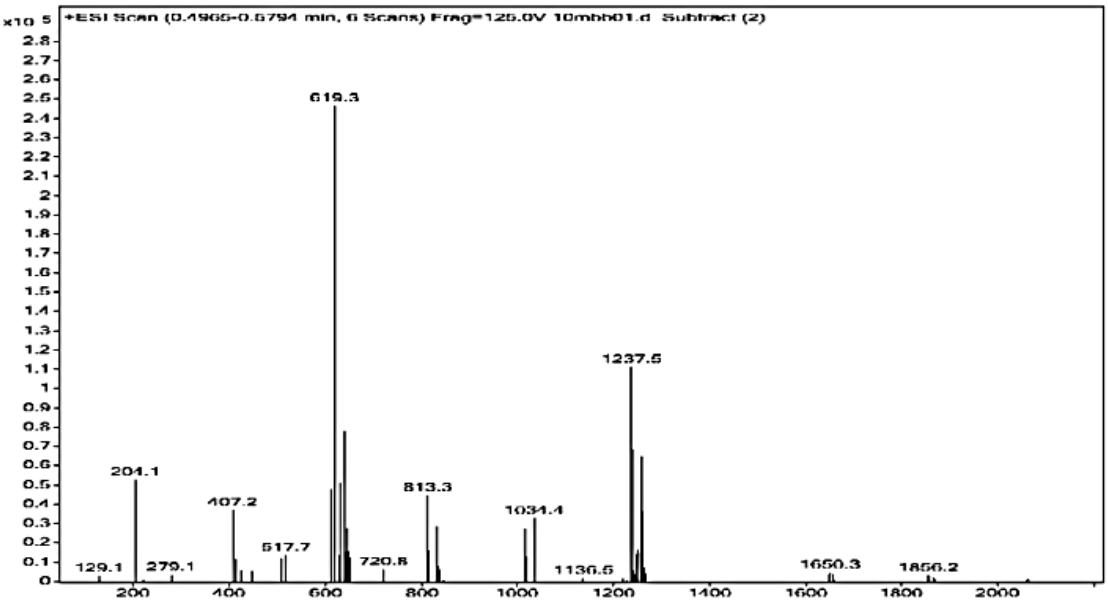


Figure 3.52: Mass spectrometry analysis of NAG oligomers. A) NAG_3 (m/z) was ~628 Da. B) NAG_4 (m/z) was ~831 Da. C) NAG_5 (m/z) was ~1034 Da. D) NAG_6 (m/z) was ~1237 Da. The purity of NAG_3, NAG_4, NAG_5, and NAG_6 were ~85%, ~80%, ~50%, and less than 20% respectively.

3.25 Analysis of KEG15107 in solution by tryptic digest

A tryptic digest was carried out for the KEG15107 protein to determine whether it has a globular fold or whether the four LysM domains are connected by flexible linkers which might be susceptible to be attacked and cleaved by a protease. The amino acid sequence of *KEG15107* was analyzed to identify the position of lysine (K) and arginine (R) residues that might form the target for tryptic cleavage which could indicate they were exposed in the structure. There were 14 possible cleavage sites for trypsin observed in the KEG15107 sequence spread over each of the LysM domains and some of which are close to the inter-domain linkers (Figure 3.53).

The KEG15107 was treated with 10 µg trypsin at 1:1 ratio (w/w). The treated protein was incubated at 37°C for 30 min, 60 min, 120 min and overnight. All the treated samples were kept in ice after the treatment to avoid further reaction between the enzyme and the protein. The digested KEG15107 was subjected to the 12% SDS-PAGE gel (Figure 3.54). None of the treated samples showed evidence of cleavage. These suggest that the LysM domains in the protein are tightly packed and that, the lysine and arginine residues are not accessible to trypsin as in the structure of plant protein containing three LysM domains (AtCERK1) from *Arabidopsis thaliana* (Liu et al., 2012). The three domains of AtCERK1 are tightly packed against each other resulted from some contacts between the residues of the LysM domains with the addition of three disulfide bonds from six cysteine residues of the protein. This is in contrast to the analysis of the LysM domains of AtlA from *Enterococcus faecalis* in which, the six tandem repeats of LysM domains of the protein were suggested to not interact with each other to form a quaternary structure in solution, but instead they behave as 'beads on a string' (Mesnage et al., 2014).

VKTYQVQPGDTLFALARREYGDSTLYPVIARQNHLANPDLIVSGQQLLIPYVTYRHL
VAAADSTATRKEITQHYYGTDDTKVQLIWEIVNGVAQREIQQGSWLHIPDLSNVGHH
TIVDGESLAGLAARWYGDDHLAIVIGLANNLPANTEPTPGQVLIVPGLNRRHIAGDT
LVSLCREEYGDADLDTRTSVVAAANHIGEPAALFSNQVIYFPS

Figure 3.53: Possible cleavage sites for trypsin on the KEG15107 protein. The 14 potential cleavage sites are highlighted in blue boxes. The residues colored in blue are Domain 1, green is Domain 2, orange is Domain 3 and red is Domain 4.

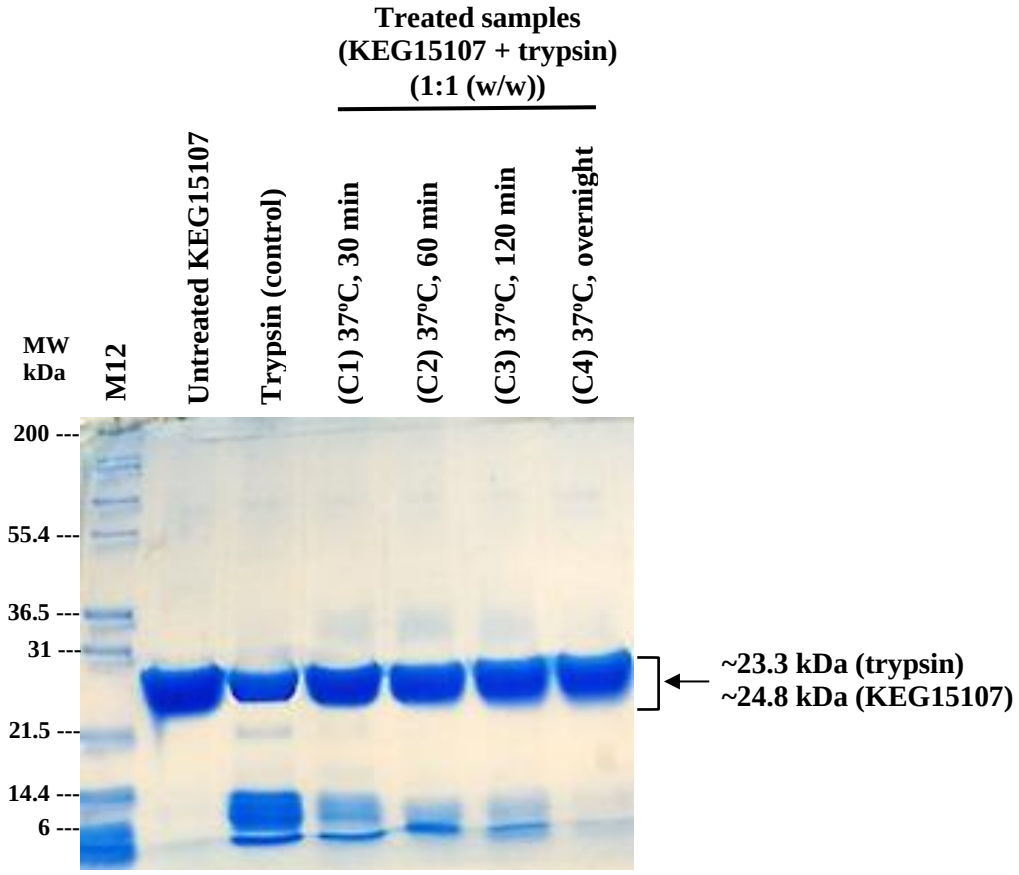


Figure 3.54: SDS-PAGE gel for the trypsin-treated *KEG15107*. The protein samples were treated with ten µg trypsin (final concentration) at 1:1 (w/w). The experiment was carried out at four different incubation settings; C1 (37°C, 30 min), C2 (37°C, 60 min), C3 (37°C, 120 min) and C4 (37°C, overnight). The KEG15107 protein was not digested by trypsin as there are no bands with smaller MW in the treated samples except for the bands (~6-3.5 Da) which present in the trypsin control.

3.26 Crystals of KEG15107

The purified KEG15107 protein was concentrated to 25 mg/mL by using a VIVASPIN device (10, 000 MWCO, Sartorius), and was desalted in freshly prepared 10 mM Tris (pH 8.0) using a Zeba Spin Desalting column (Thermo Scientific). Crystallization trials were carried out on the protein using a Matrix Hydra II PlusOne crystallization robot (BioMATRIX), applying a sitting drop method, and the successful conditions were further optimized using a hanging drop method. The protein was dispensed into 96-well MRC2 sitting-drop crystallization trays in 1:1 ratio of protein: precipitant generating a 200 nL drop, which was allowed to equilibrate through vapor diffusion at 19°C. Commercially available crystallization screens from Molecular Dimension

(Morpheus, $AmSO_4$, JCSG, MPD, PACT, and ProPlex) were used to identify conditions that formed crystals. Further optimization was carried out on the successful conditions in 24-well trays with a reservoir volume of 500 µL and drop volume of 2 µL by using a hanging drop method. The optimization involved varying the protein concentration, the concentration of PEG (%) as the most common precipitant in protein crystallization, the pH of buffers and the salt concentration. The experiments were performed on the apo KEG15107 protein (>95% purity), and the protein in complex with NAG_3, NAG_4, NAG_5, and NAG_6 at different ratios of protein to sugar.

3.26.1 Crystals of apo KEG15107

Crystals of apo KEG15107 grew in various conditions from PACT, PROPLEX, JCSG, MPD, $AmSO_4$, and Morpheus screenings with concentrations of the protein between 10 - 25 mg/mL over pH range of (4.2 to 7.5) with PEG as the most common precipitant. The apo KEG15107 protein crystallized the best at concentration 25 mg/mL with moderate precipitation observed in the drops. The successful crystallization conditions for the apo KEG15107 crystals are listed in Table 3.4. The crystals exhibited a number of different morphologies including plates and rods with sizes up to ~100 µm (Figure 3.55).

3.26.2 Crystals of the KEG15107 complex with various NAG oligomers

Crystallization trials for KEG15107 in complex with a range of NAG oligomers were carried out by co-crystallization as described in Chapter 2 (Section 2.12). Multiple crystal forms of KEG15107 in complex with NAG_4 including rods, plates, and cubes grew in various crystallization condition by co-crystalizing the sugar in the freshly purified KEG15107 protein at 25 mg/mL. The KEG15107–NAG_4 complex was prepared at molar ratio 1:2 (protein: sugar), 1:10 and 1:100. Crystallization trials on the complex at molar ratio 1:100 between the protein and the sugar were carried out with the protein at ~15 mg/mL because the first attempt of crystallization resulted in heavy precipitation observed in the drop. A few crystals of KEG15107-NAG_4 under a re-optimized (1:100) co-crystallization trials grew from ProPlex conditions. Co-crystallization trials of KEG15107-NAG_4 complex with protein to sugar ratios of 1:2 and 1:10 (protein: sugar) were performed at 25 mg/mL (Figure 3.56 (D-J)). The majority of the crystals grew from ProPlex, with additional crystals being formed from

JCSG and PACT conditions (Table 3.5). Similar co-crystallization trials set up of KEG15107-NAG_4 were used for the KEG15107-NAG_3, KEG15107-NAG_5, and KEG15107-NAG_6 complexes, and multiple crystal forms of KEG15107-NAG_3 (Figure 3.56 (A-C)), KEG15107-NAG_5 (Figure 3.56 (L-O)) and KEG15107-NAG_6 (Figure 3.56 (P)) grown in the crystallization drops from various conditions of ProPlex and a few crystals grew from JCSG and PACT conditions (Table 3.5). The complexes crystallized in solutions over a wide pH range of 5.5 to 8.0 with PEG as the most common precipitant.

3.27 Validation of KEG15107 crystals by SDS-PAGE gel and Mass spectrometry analysis

The protein crystals of KEG15107 grown under different conditions from ProPlex and PACT screening solutions were analyzed by 12% SDS-PAGE gel and Mass spectrometry analysis to verify that the crystallized protein was *KEG15107*. The crystals were scooped from condition 1 (C1) (0.2 M lithium sulfate, 0.1 M MES (pH 6), 20% (w/v) PEG 4000), C2 (0.1 M Magnesium chloride, 0.1 M MES (pH6), 8% (w/v) PEG 6000), C3 (0.2 M Ammonium chloride, 0.1 M MES (pH6), 20% (w/v) PEG 6000), C4 (0.2 M sodium acetate trihydrate, 0.1 M Bis-Tris propane (pH6.5), 20% (w/v) PEG3350), C5 (0.2 M lithium chloride, 0.1 M Tris (pH8) , 20% (w/v) PEG 6000) and were washed three times in freshly prepared cryo-protectant containing 25% ethylene glycol to stabilize the crystals and then dissolved in 6 µL of water. The SDS PAGE analysis of the dissolved crystals showed that they run with an identical molecular weight to the protein control of purified KEG15107 as shown in Figure 3.7. To further confirm the identity of the crystallized protein, MS/MS was performed on the sample C3 (Figure 3.57) of the dissolved KEG15107 crystals. The band (C3) on the SDS PAGE gel containing the protein was cut and sent to the Faculty of Science Mass Spectrometry Centre, the University of Sheffield for the MS/MS analysis. The sample was treated with 1% formic acid and 50% acetonitrile to unfold the protein, and it was cleaved into small peptides by trypsin. The amino acid sequence of each of the fragments obtained from the MS/MS analysis was analyzed and compared to the KEG15107 sequence, and the analysis showed that the sequence fragments of the samples were identical to KEG15107. The results of the MS/MS analysis on KEG15107 are presented in Table 3.6.

Table 3.4: Successful crystallization conditions for apo KEG15107 crystals

Conditions	Components		
	Salt	Buffer (pH)	Precipitant
PACT (A2)	-	0.1 M SPG (5.0)	25 % w/v PEG 1500
A7	0.2 M Sodium chloride	0.1 M Sodium acetate (5.0)	20 % w/v PEG 6000
B7	0.2 M Sodium chloride	0.1 M MES (6.0)	20 % w/v PEG 6000
B8	0.2 M Ammonium chloride	0.1 M MES (6.0)	20 % w/v PEG 6000
F5	0.2 M Sodium nitrate	0.1 M Bis-Tris propane (6.5)	20 % w/v PEG 3350
G10	0.02 M Sodium/potassium phosphate	0.1 M Bis-Tris propane (7.5)	20 % w/v PEG 3350
JCSG (B6)		0.1 M Phosphate/citrate (4.2)	40 % v/v Ethanol 5 % w/v PEG 1000
A9	0.2 M Ammonium chloride	-	20 % w/v PEG 3350
B8	0.2 M Magnesium chloride hexahydrate	0.1 M Tris 7.0	10 % w/v PEG 8000
D5	-	0.1 M HEPES (7.5)	70 % v/v MPD
H3	-	0.1 M BIS-Tris (5.5)	25 % w/v PEG 3350
ProPlex (E11)	-	0.1 M sodium citrate (5.0)	20 % w/v PEG 8000
C7	-	0.1 M sodium citrate (5.6)	20% PEG 4000 20% 2-propanol
E12	0.2 M ammonium sulfate	0.1 M MES (6.5)	20 % w/v PEG 8000
H10	0.05 M magnesium Chloride	0.1 M MES (6.5)	10 % v/v 2-propanol 5 % w/v PEG 4000
H11	0.2 M ammonium acetate	0.1 M sodium HEPES (7.5)	25% 2-propanol
MPD (F8)	0.2 M Magnesium acetate	0.1 M MES sodium salt pH (6.5)	15% (w/v) MPD
F9	0.2 M tri-Sodium citrate	0.1 M HEPES sodium salt (7.5)	15% (w/v) MPD
Morpheus (E4)	Ethylene Glycols 0.12 M (ligand)	0.1 M MES (6.5)	MPD_P1K_P3350 37.5%
H4	Amino acids 0.10 M (ligand)	0.1M MES (6.5)	MPD_P1K_P3350 37.5%
$AmSO_4$ (G2)	2 M sodium chloride	-	2 M ammonium sulfate

Table 3.5: Successful crystallization conditions for KEG15107-$NAG_{(n)}$ crystals

Conditions	Components		
	Salt	Buffer (pH)	Precipitant
ProPlex (B1)	0.2 M lithium sulfate	0.1 M Tris (7.5)	5 % w/v PEG 4000
B12	0.1 M magnesium chloride	0.1 M Na HEPES (7.0)	15 % w/v PEG 4000
C1	0.15 M ammonium sulfate	0.1 M Tris (8.0)	15 % w/v PEG 4000
C6	0.15 ammonium sulfate	0.1 M Na HEPES (7.0)	20 % w/v PEG 4000
C7	-	0.1 M sodium citrate (5.6)	20% PRG 6000 20% 2-propanol
C8	0.16M sodium chloride	0.1 M Tris (8.0)	20 % w/v PEG 4000
C10	0.15 M ammonium sulfate	0.1 M MES (5.5)	25 % w/v PEG 4000
D1	0.17M sodium chloride	0.1 M Na HEPES (7.5)	25 % w/v PEG 4000
E12	0.3 M ammonium sulfate	0.1 M MES (6.5)	20 % w/v PEG 8000
F11	-	0.1 M sodium acetate (5.0)	1.5 M ammonium sulfate
F12	-	0.1 M Na HEPES (7.0)	1.5 M ammonium sulfate
G1	-	0.1 M Tris (8.0)	1.5 M ammonium sulfate
G2	-	0.1 M sodium acetate (5.0)	2 M ammonium sulfate
G3	-	0.1 M Na HEPES (7.0)	2 M ammonium sulfate
G4	-	0.1 M Tris (8.0)	2 M ammonium sulfate
G5	1 M potassium chloride	0.1 M Na HEPES (7.0)	1 M ammonium sulfate
H11	0.2 M ammonium acetate	0.1 M Na HEPES (7.5)	25 % v/v 2-propanol
H12	0.1 M sodium chloride	0.1 M Tris (8.0)	15 % v/v ethanol 5 % v/v MPD
JCSG (H2)	1.0 M Ammonium sulfate	0.1 M BIS-Tris (5.5)	1 % w/v PEG 3350
H4	0.2 M Calcium chloride dihydrate	0.1 M BIS-Tris (5.5)	45 % v/v MPD
H7	0.3 M ammonium sulfate	0.1 M BIS-Tris (5.5)	25 % w/v PEG 3350
PACT (B1)	-	0.1 M MIB (4.0)	25 % w/v PEG 1500
B7	0.2 M Sodium chloride	0.1 M MES (6.0)	20 % w/v PEG 6000

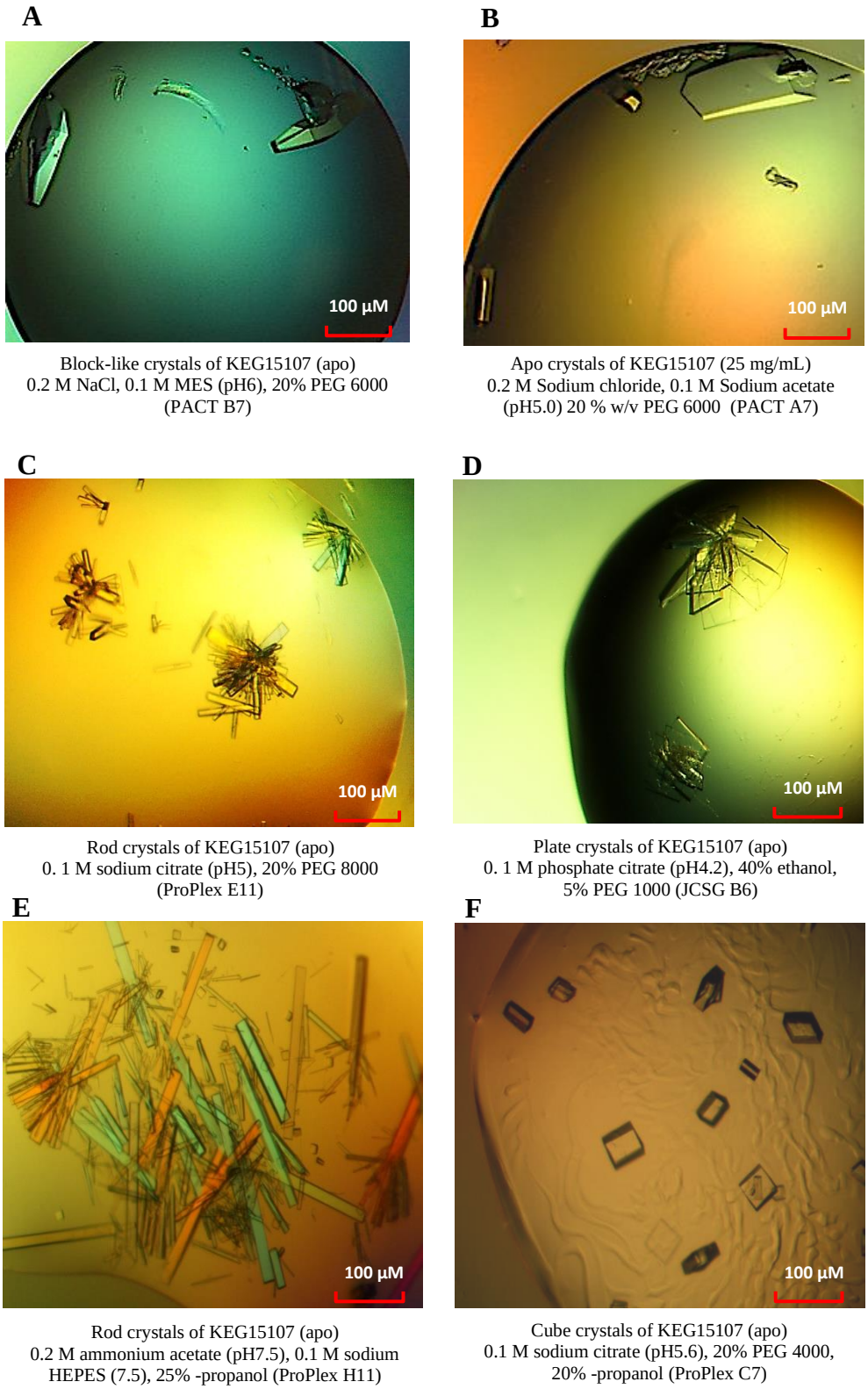
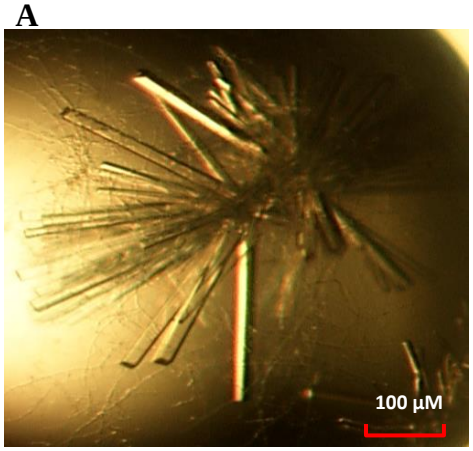


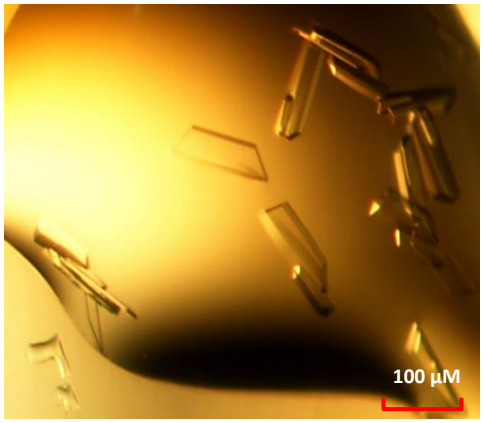
Figure 3.55: Crystals of apo KEG15107 grown in various crystallization conditions. (A-F) Multiple crystal forms of apo KEG15107 including rod, cubic, plate grew in PACT, JCSG and ProPlex conditions and these crystals were sent for data collections.

A



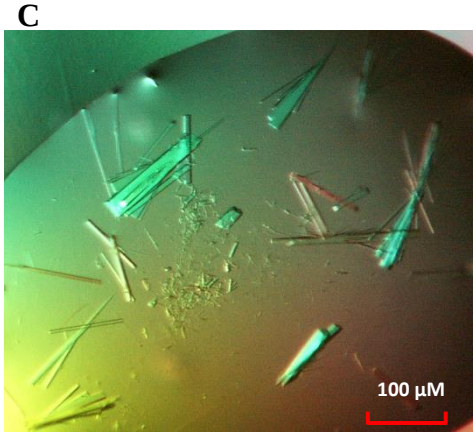
Rod-like crystals of KEG15107-NAG_3
0.1 M sodium citrate (pH 5.6), 20% PEG 4000,
20% 2-propanol
ProPlex C7 (1:10 molar ratio)

B



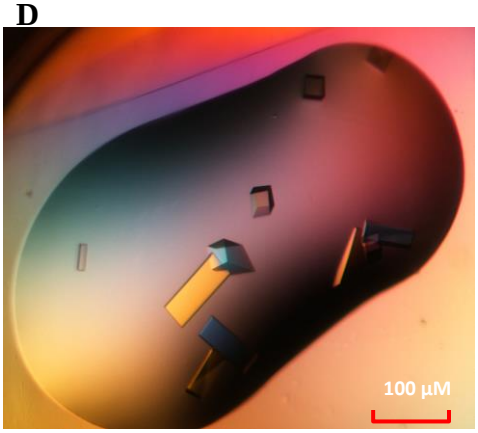
Crystals of KEG15107-NAG_3 (1:10)
0.15 M ammonium sulfate, 0.1 M Na HEPES
(pH7.0), 20 % w/v PEG 4000
(C6 ProPlex)

C



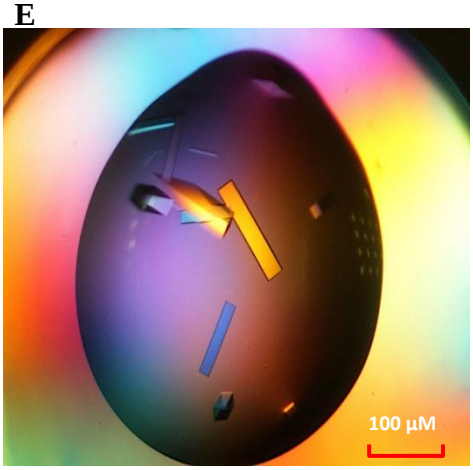
Needle-like crystals of KEG15107-NAG_3
0.2 M sodium chloride, 0.1 M Tris (pH8),
20% PEG 4000
ProPlex C8 (1:100 molar ratio)

D



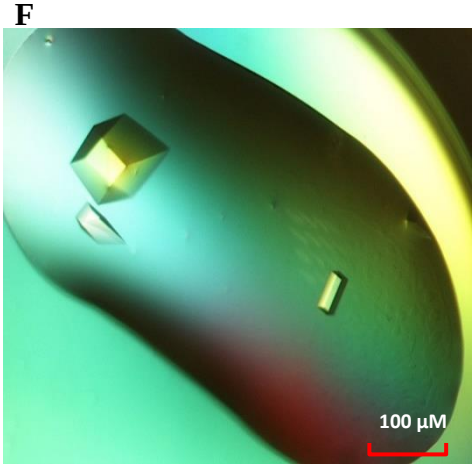
Cube and rod crystals of KEG15107-NAG_4
0.2 M sodium chloride, 0.1 M sodium HEPES
(pH7.5), 25% PEG 4000
ProPlex D1(1:2 molar ratio)

E



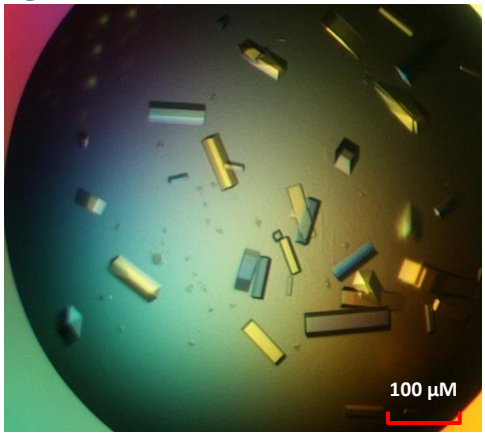
Cube and rod crystals of KEG15107-NAG_4
0.15 M ammonium sulfate, 0.1 M sodium HEPES
(pH7), 20% PEG 4000
ProPlex C6 (1:2 molar ratio)

F



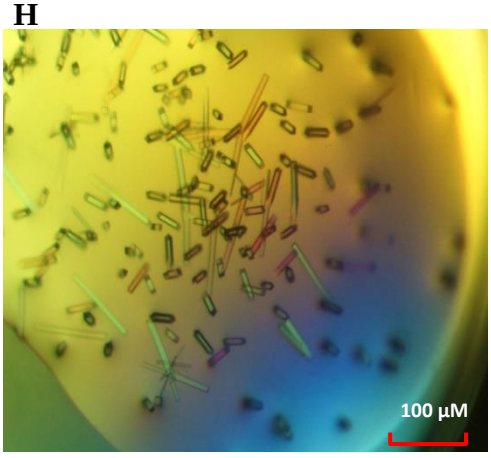
Cube and rod crystals of KEG15107-NAG_4
0.2 M ammonium sulfate, 0.1 M MES (pH6.5),
20% PEG 8000
ProPlex E12 (1:2 molar ratio)

G



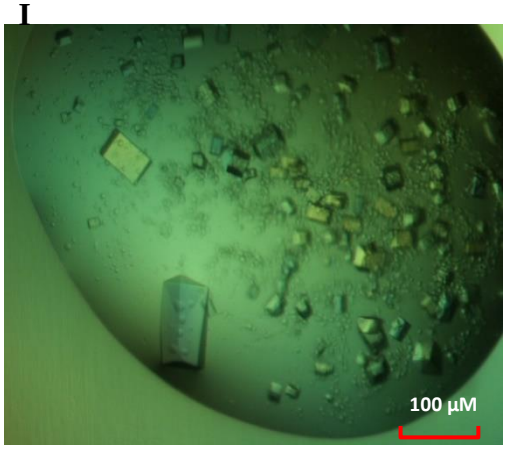
Cube and rod crystals of KEG15107-NAG_4
0.15 M ammonium sulfate, 0.1 M sodium HEPES
(pH7), 20% PEG 4000
ProPlex C6 (1:10 molar ratio)

H



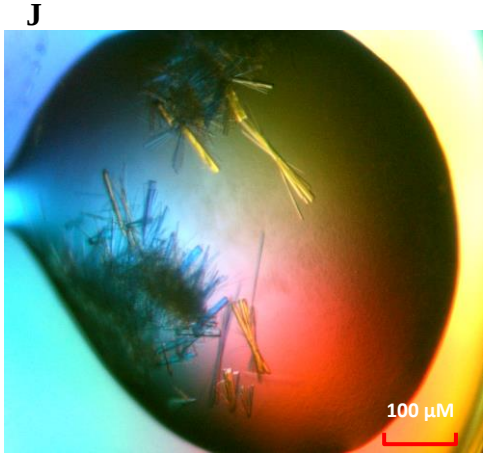
Rod crystals of KEG15107-NAG_4
0.2 M ammonium acetate, 0.1 M sodium HEPES
(pH7.5), 25% 2-propanol
ProPlex H11 (1:10 molar ratio)

I



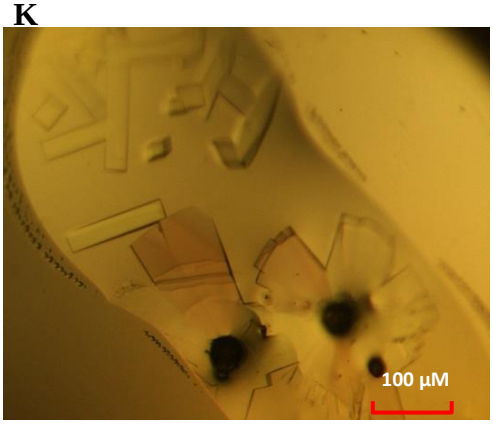
Block-like crystals of KEG15107-NAG_4
0.15 M ammonium sulfate, 0.1 M MES (pH5.5),
25% PEG 4000
ProPlex C10 (1:10 molar ratio)

J



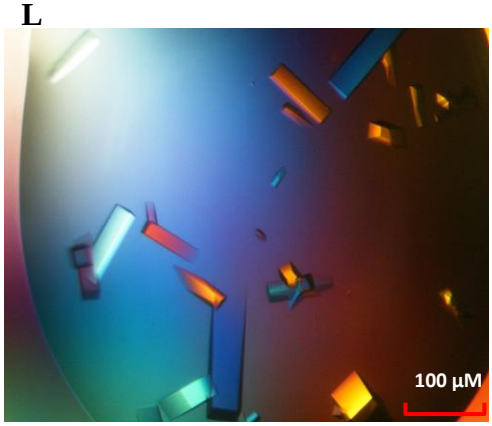
Crystals of KEG15107-NAG_4 (1:10)
0.1 M sodium acetate (pH5.0),
1.5 M ammonium sulfate
(F11 ProPlex)

K



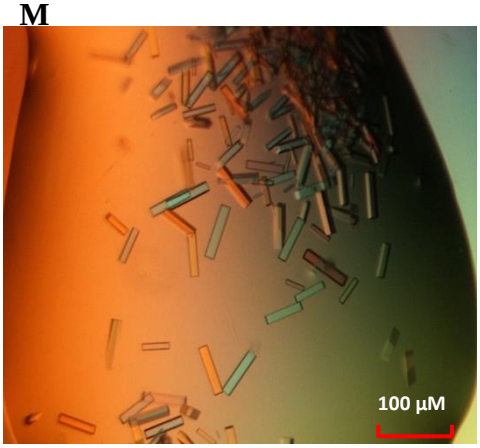
Crystals of KEG15107-NAG_4
0.1 M sodium chloride, 0.1 M Tris (pH8), 15%
ethanol, 5% MPD
ProPlex H12 (1:10 molar ratio)

L



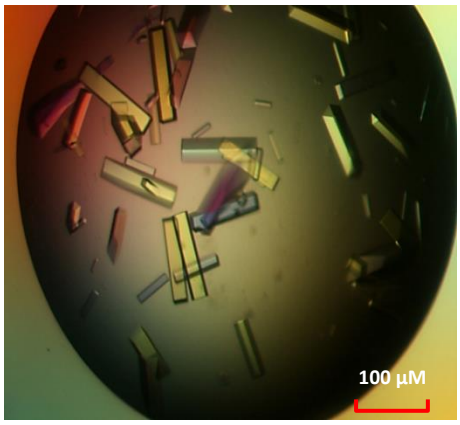
Cube and rod crystals of KEG15107-NAG_5
0.15 M ammonium sulfate, 0.1 M sodium HEPES
(pH7), 20% PEG 4000
ProPlex C6 (1:2 molar ratio)

M



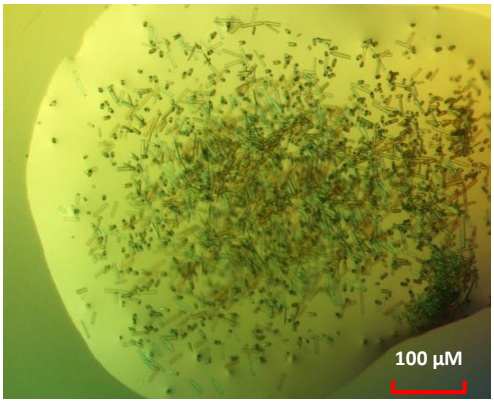
Rod crystals of KEG15107-NAG_5
0.2 M ammonium sulfate, 0.1 M Bis-Tris propane
(pH5.5), 25% PEG 3350
JCSG H7 (1:2 molar ratio)

N



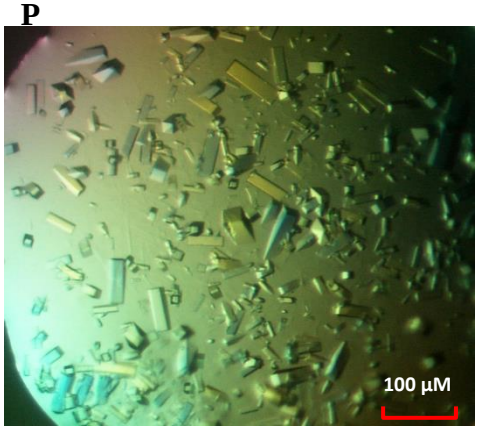
Crystals of KEG15107-NAG_5 (1:10)
0.15 M ammonium sulfate, 0.1 M Na HEPES
(pH7.0), 20 % w/v PEG 4000
(C6 ProPlex)

O



Crystals of KEG15107-NAG_5 (1:10)
0.2 M ammonium acetate,
0.1 M Na HEPES (pH7.5),
25 % v/v 2-propanol (H11 ProPlex)

P



Cube and rod crystals of KEG15107-NAG_6
0.15 M ammonium sulfate, 0.1 M sodium HEPES
(pH7), 20% PEG 4000
ProPlex C6 (1:2 molar ratio)

Figure 3.56: Crystals of the KEG15107 complex in various NAG oligomers grown in various crystallization conditions. (A-P) Multiple crystal forms of KEG15107-$NAG_{(n)}$ complexes including rod, cube, plate mostly grew in ProPlex conditions whereas only a few of them obtained from JCSG conditions.

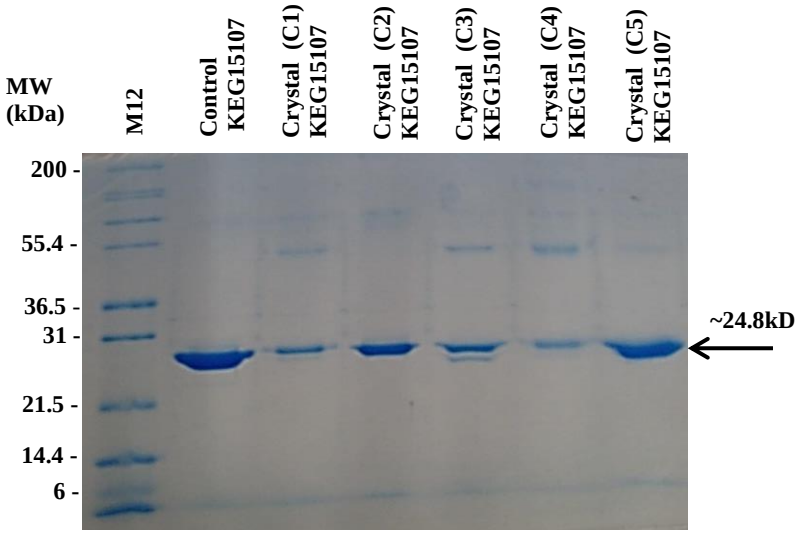


Figure 3.57: Crystal validation for KEG15107 on 12% SDS-PAGE gel. The crystal samples that were loaded onto the gel were crystals dissolved in 6 µL water. The KEG15107 crystals were scoped from different conditions of the crystalline screening solution. All the KEG15107 crystals showed similar in size ~24.8 kDa.

Table 3.6: MS/MS analysis of crystal samples of KEG15107

Peptide fragments identifying by MS/MS	Residues number in KEG15107	MW (Da)	Intensities (I)
MVKTYQVQPGDTLFALAR	1-18	2037.0717	5.90E+06
VKTYQVQPGDTLFALAR	2-18	1906.0312	2.94E+09
TYQVQPGDTLFALAR	4-18	1678.8679	5.83E+08
TYQVQPGDTLFALARR	4-19	1834.969	8.37E+06
REYGDSTLYPVIAR	19-32	1638.8366	5.57E+09
EYGDSTLYPVIAR	20-32	1482.7355	2.95E+09
QNHLANPDLIVSGQQLLIPYVTYR	33-56	2751.4708	9.41E+08
HLVAAADSTATR	57-68	1211.6258	4.35E+06
HLVAAADSTATRK	57-69	1339.7208	7.16E+07
KEITQHYYGTDDTK	69-82	1697.7897	6.00E+04
KEITQHYYGTDDTKVQLIWEIVNGVAQR	69-96	3303.6888	1.01E+07
EITQHYYGTDDTK	70-82	1569.6947	3.01E+07
EITQHYYGTDDTKVQLIWEIVNGVAQR	70-96	3175.5938	6.15E+07
EIQQGSWLHIPDLSNVGHHTIVDGESLAGLAAR	97-129	3519.7859	2.58E+08
WYGDDHLAIVIGLANNLPANTEPTPGQVLIVPGLNR	130-165	3837.0214	3.48E+07
WYGDDHLAIVIGLANNLPANTEPTPGQVLIVPGLNRR	130-166	3993.1225	9.53E+06
RRHIAGDTLVSLCR	166-179	1652.8893	3.57E+07
RHIAGDTLVSLCR	167-179	1496.7882	1.35E+08
RHIAGDTLVSLCREEYGDADLDTR	167-190	2761.309	6.30E+08
HIAGDTLVSLCR	168-179	1340.6871	2.48E+08
HIAGDTLVSLCREEYGDADLDTR	168-190	2605.2078	8.53E+08
EEYGDADLDTR	180-190	1282.5313	1.66E+09
TSVVAAANHIGEPAALFSNQVIYFPSLEHHHHHH	191-224	3766.8505	4.95E+05

[1]MVKTYQVQPG DTLFALARRE YGDSTLYPVI ARQNHLANPD LIVSGQQLLI PYVTYRHLVA[60]
[61]AADSTATRKE ITQHYYGTDD TKVQLIWEIV NGVAQREIQQ GSWLHIPDLS NVGHHTIVDG[120]
[121]ESLAGLAARW YGDDHLAIVI GLANNLPANT EPTPGQVLIV PGLNRRRHIA GDTLVSLCRE[180]
[181]EYGDADLDTR TSVVAAANHI GEPAALFSNQ VIYFPSLEHH HHHH[224]

The protein sequence of KEG15107 from (Figure 3.1.8.4)

CHAPTER FOUR

STRUCTURE DETERMINATION AND ANALYSIS OF KEG15107

4.0 Introduction

In this chapter, data collection, structure determination and structure analysis of KEG15107 and its complexes with either NAG_3, NAG_4 or NAG_5 oligomers are described.

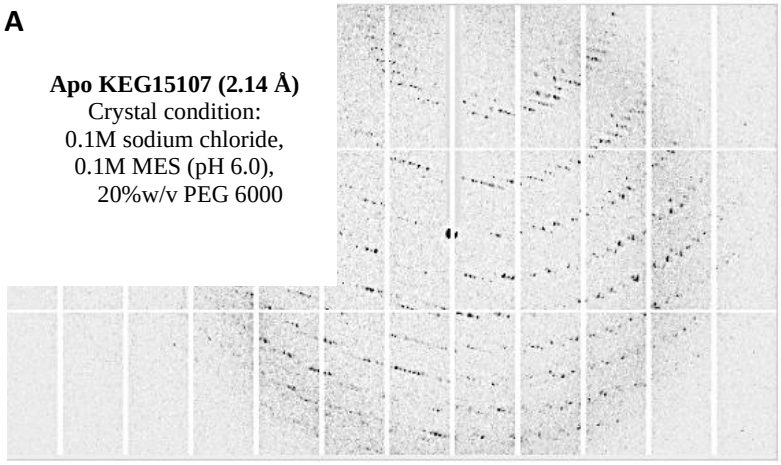
4.1 X-ray data collection and structure determination of KEG15107 and its complexes with polyNAG

All the mounted KEG15107 crystals, including those of the apo protein and its complexes with NAG oligomers, were stored in liquid nitrogen and were exposed to X-rays on beamlines at the Diamond synchrotron, UK. The diffraction data were collected and subsequently processed either by Fast DP, xia2 DIALS or xia2 3dii. The quality of the processed data was initially determined by the statistical analysis of indicators including R-merge, R-meas, completeness, multiplicity, CC half, I/sigma (I) and the resolution of the collected data set. All the data sets of KEG15107, both for the apo protein and its complexes with polyNAG, are high quality with resolutions ranging from 1.25 Å to 2.14 Å (Figure 4.0). A summary of the statistics for the diffraction data is shown in Table 4.0.

The crystals of apo KEG15107 belong to space group $P2_1$ with cell dimensions of a=90.0 Å, b=63.1 Å, c=169.1 Å, β=98.5^o and diffracted to 2.14 Å. Consideration of the possible values of Vm, suggest that the crystals most likely contain eight or nine subunits in the asymmetric unit (AU). The structure of the apo KEG15107 was initially determined in collaboration with Dr. Bisson (University Sheffield) using PhaserMR in the CCP4 suite with a search model of a single subunit of MSMEG3288 from *M. smegmatis*, a homolog that shares about 80% sequence similarity to KEG15107 (Chapter 2: 2.14.1). Structure determination shows there were eight subunits (Chains A-H) in the AU.

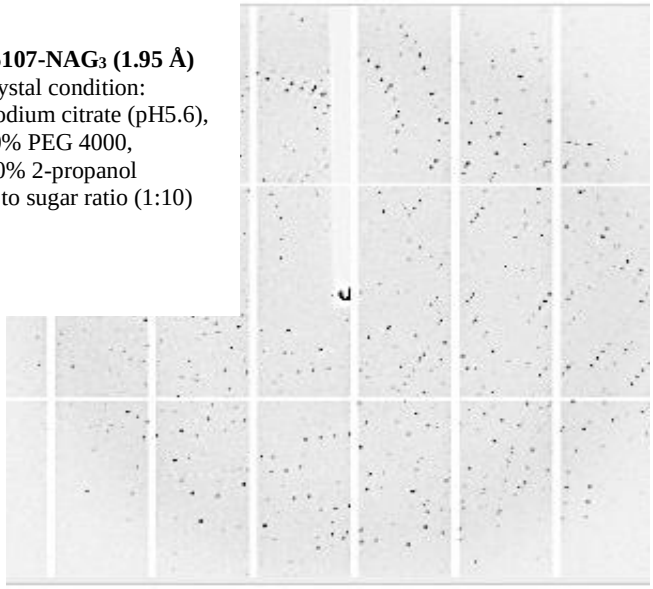
Crystals of a KEG15107-NAG_4 complex were produced from crystallization trials with a 1:2 protein to sugar ratio and belong to space group $I4_122$ with cell dimensions a=121.7 Å, b=121.7 Å, c=202.7 Å and diffracted to 1.25 Å. Consideration of possible value of Vm, suggest that the crystals contain two or four subunits in the AU. The structure of the complex was initially determined by molecular replacement using a monomer of Chain A of the apo KEG15107 structure as the search model. Crystals of a KEG15107-NAG_5 complex, again produced from solutions containing the 1:2 ratio of protein to sugar, also belong to space group $I4_122$ with cell dimensions of a=123.3 Å, b=123.3 Å, c=205.5 Å and diffracted to 1.38 Å. The structure of the complex was determined by molecular replacement (PhaserMR) using a monomer of the KEG15107-NAG_4 structure as a search model.

Subsequently, co-crystallization of KEG15107 with polyNAG substrates was carried out by increasing the sugar concentration to a 1:10 ratio of protein to sugar. Crystals of a KEG15107-NAG_3 complex belong to space group of $I4_122$ with cell dimensions a=121.0 Å, b=121.0 Å, c=201.3 Å and diffracted to 1.95 Å. The structure of the complex was determined by PhaserMR using the monomer of the KEG15107-NAG_4 structure as the search model. Crystals of the complexes from the 1:10 crystallization trials were isomorphous to those produced from the crystallization trials at a 1:2 ratio of protein to sugar. The crystals of these KEG15107-NAG_4 and KEG15107-NAG_5 complexes diffracted to 1.76 Å and 1.60 Å, respectively, and were determined directly by the refinement of the KEG15107-NAG_4 and KEG15107-NAG_5 complexes from the 1:2 crystallization trials.



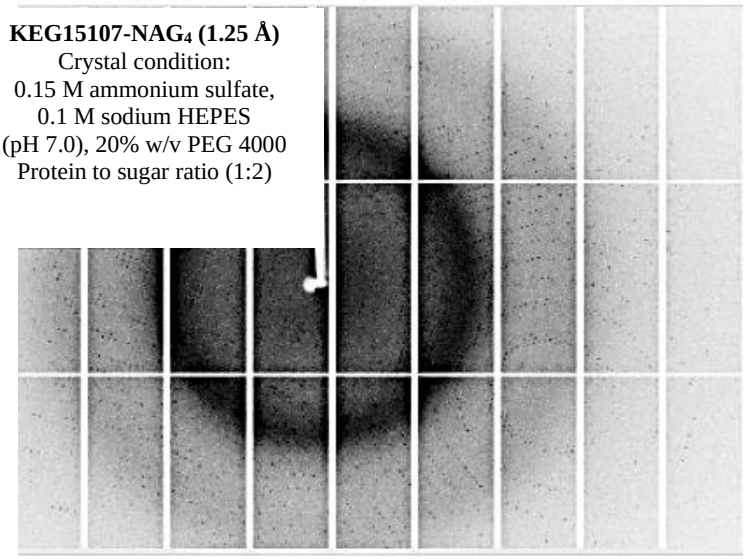
B

KEG15107-NAG$_3$ (1.95 Å)
Crystal condition:
0.1 M sodium citrate (pH5.6),
20% PEG 4000,
20% 2-propanol
Protein to sugar ratio (1:10)



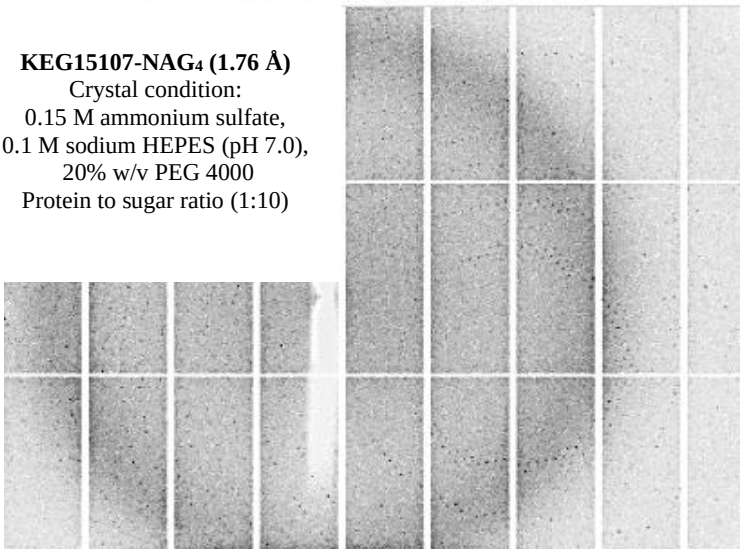
C

KEG15107-NAG$_4$ (1.25 Å)
Crystal condition:
0.15 M ammonium sulfate,
0.1 M sodium HEPES
(pH 7.0), 20% w/v PEG 4000
Protein to sugar ratio (1:2)



D

KEG15107-NAG$_4$ (1.76 Å)
Crystal condition:
0.15 M ammonium sulfate,
0.1 M sodium HEPES (pH 7.0),
20% w/v PEG 4000
Protein to sugar ratio (1:10)



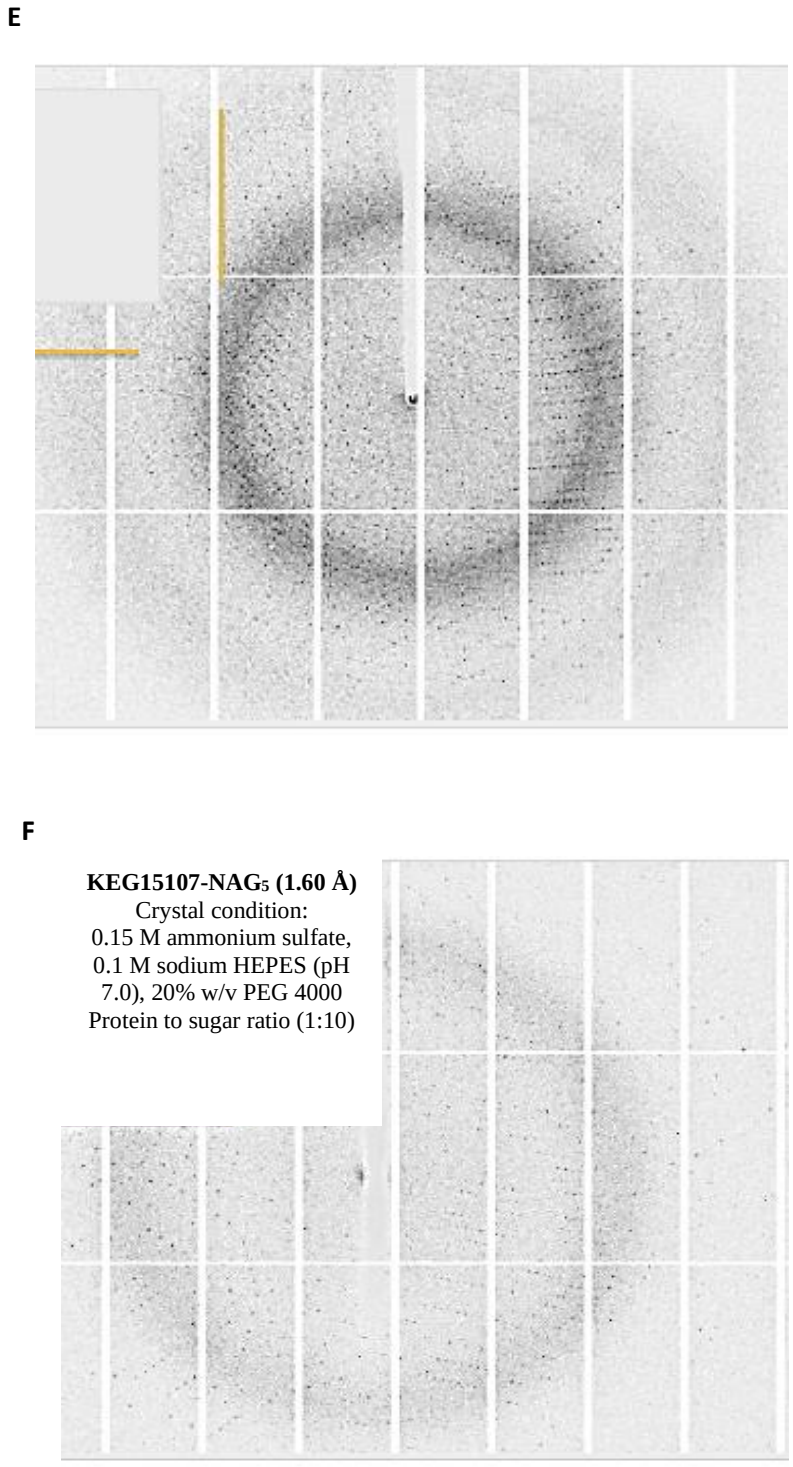


Figure 4.0: Diffraction patterns of the protein crystals of KEG15107 and its complexes. A) Apo KEG15107 (2.14 Å). B) KEG15107-NAG_3 complex (1.95 Å) from the 1:10 ratio of protein to sugar. C-D) KEG15107-NAG_4 (1.25 Å) and KEG15107-NAG_4 (1.76 Å) complexes from the 1:2 and 1:10 ratios of protein to sugar, respectively. E) KEG15107-NAG_5 (1.38 Å) and KEG15107-NAG_5 (1.60 Å) complexes from the 1:2 and 1:10 ratios of protein to sugar, respectively.

Table 4.0 (A): Processing statistics of the apo KEG15107 data

Data set	KEG15107 (apo)
Detector	Pilatus 6M-F
Beamline	I04
Wavelength (Å)	0.92819
No. of images (0.1^o)	3600
Exposure time per image (s)	0.05
Space group	$P2_1$
Cell dimensions	
a, b, c (Å)	90.0, 63.1, 169.1
α, β, γ (^{o})	90.0, 98.5, 90.0
Resolution range (Å)	167.32-2.14 (2.20-2.14)[#]
R_{merge} (I)	0.086 (0.529)[#]
R_{meas} (I)	0.099 (0.670)[#]
CC-half	0.998 (0.838)[#]
$I/\sigma(I)$	9.5 (2.5)[#]
Completeness (%)	98.7 (99.3)[#]
Multiplicity	6.2 (5.0)[#]
No. of reflections	641251 (38475)[#]
Unique reflections	102632 (7657)[#]
Wilson B factor (A^2)	34
Cell volume (A^3)	949959

Values for the highest resolution shell are in parentheses

Table 4.0 (B): Processing statistics of data from the crystals of the KEG15107-NAG complexes

Data sets	KEG15107-NAG_3 protein to sugar ratio (1:10)	KEG15107-NAG_4 protein to sugar ratio (1:2)	KEG15107-NAG_4 protein to sugar ratio (1.10)	KEG15107-NAG_5 protein to sugar ratio (1.2)	KEG15107-NAG_5 protein to sugar ratio (1.10)
Detector	Pilatus 6M-F (25Hz)	Pilatus3 6M	Pilatus 6M-F	Pilatus 6M-F (25Hz)	Pilatus 6M-F
Beamline	I04-i	I03	I04	I04-i	I04
Wavelength (Å)	0.9159	0.9763	0.9795	0.9282	0.9795
No. of images (0.1^o)	2900	1800	3600	2000	3600
Exposure time per image (s)	0.04	0.05	0.05	0.05	0.05
Space group	$I4_122$	$I4_122$	$I4_122$	$I4_122$	$I4_122$
Cell dimensions					
a, b, c (Å)	121.0, 121.0, 201.3	121.7, 121.7, 202.7	121.7, 121.7, 203.4	123.3, 123.3, 205.5	121.7, 121.7, 203.1
α, β, γ (^{o})	90.0, 90.0, 90.0	90.0, 90.0, 90.0	90.0, 90.0, 90.0	90.0, 90.0, 90.0	90.0, 90.0, 90.0
Resolution range (Å)	65.19-1.95 (1.98-1.95)#	65.6-1.25 (1.27-1.25)#	65.64-1.76 (1.79-1.76)#	61.66-1.38 (1.40-1.38)#	65.63-1.60 (1.63-1.60)#
R_{merge} (I)	0.2275 (0.9676)#	0.062 (1.568)#	0.193 (2.645)#	0.1 (2.120)#	0.152 (2.776)#
R_{meas} (I)	0.233 (0.991)#	0.065 (1.665)#	0.197 (2.695)#	0.104 (2.195)#	0.155 (2.851)#
CC-half	0.9977 (0.5285)#	1 (0.489)#	0.999 (0.581)#	0.999 (0.509)#	1 (0.504)#
$I/\sigma(I)$	9.60 (2.96)#	20 (1.3)#	14.5 (1.1)#	14.8 (1.1)#	13.3 (1.1)#
Completeness (%)	100.0 (99.63)#	99.6 (96.8)#	100.0 (100.0)#	100.0 (100.0)#	100.0 (99.2)#
Multiplicity	21.0 (21.5)#	12.9 (8.9)#	26.7 (27.0)#	14.9 (15.0)#	25.3 (19.2)#
No. of reflections	1143909 (57914)#	2.6606e+06 (88402)#	2.0130e+06 (101028)#	2.3909e+06 (119854)#	2.5271e+06 (93176)#
Unique reflections	54592 (2688)#	206633 (9947)#	75376(3744)#	160965 (8006)#	99838 (4850)#
Wilson B factor (A^2)	22	13	19	15	15
Cell volume (A^3)	2947087	3001822	3006214	3125208	3005428

Values for the highest resolution shell are in parentheses

4.2 Refinement of the structure of KEG15107 and its complexes with NAG oligomers

The structure of apo KEG15107 was determined by PhaserMR in the CCP4 suite, and eight protein molecules (monomers) were found to lie in the AU of the crystal. These were designated as Chains A, B, C, D, E, F, G, and H, and were rebuilt in *Coot* and refined in REFMAC5. Examination of crystal packing indicated that the monomers A, B, C, and D form one tetramer in the AU of the crystal while the other four monomers E, F, G, and H formed another tetramer (Figure 4.1 A). These two tetramers, Tetramers 1 and 2 are identical. Comparison of the apo KEG15107 tetramer to the MSMEG3288 tetramer revealed that both structures share similar quaternary structure with rmsd 0.808 Å (Figure 4.1 B). The structure refinement converged at 2.14 Å to a R_{free} value of 28.8% and a conventional R_{factor} of 22.20%. The final model consists of 1624 residues and 1062 water molecules. For most of the subunits, the electron density included virtually all the expected residues of the protein together with one residue from the linker.

The final refined structure of the Chain A of apo KEG15107 was used as a search model for structure determination of the KEG15107-NAG_4 complex from the 1:2 crystallization trials by PhaserMR. Two monomers, designated as Chains Y and Z were found in the AU of the crystal. Examination of crystal packing revealed that in this structure, two monomers formed a dimer (Figure 4.1 C) that was identical to one of the dimer interfaces in the apo KEG15107 tetramer but crystal packing showed no evidence for the formation of a tetramer. The structure of the dimer of KEG15107-NAG_4 was rebuilt and refined in *Coot* and REFMAC5, respectively. The structure refinement converged at 1.25 Å to an R_{free} value of 19.27% and a conventional R_{factor} of 17.902%. The final model consists of 438 amino acids, residues 1-215, each from molecules Y and Z together with five and three residues of the LEHHHHHH tag, respectively, 449 water molecules, four SO_4 molecules, seven PEG molecules, 2 EPE molecules and 2 NAG_4 molecules.

For the KEG15107-NAG_5 complex from the 1:2 crystallization trials and all of the other complexes, structure determination followed an equivalent process to that for the complex of NAG_4. The refinement of the KEG15107-NAG_5 complex converged at

1.38 Å to an R_{free} value of 19.61% and a conventional R_{factor} of 18.00%. The final model consists of 436 amino acids, residues 1-215 from each of molecules Y and Z together with two and four residues of the LEHHHHHH tag, respectively, 600 water molecules, five SO_4 molecules, 11 PEG molecules, 3 EPE molecules, and 2 NAG_5 molecules.

The structure of the KEG15107-NAG_3 complex from the 1:10 ratio of protein to sugar was determined at 1.95 Å to an R_{free} value of 27.0% and a conventional R_{factor} of 22.4%. The final model consists of 432 amino acids, residues 1-215 from each of molecules Y and Z together with one residue of the LEHHHHHH tag, 396 water molecules, and 2 NAG_3 molecules.

For the KEG15107-NAG_4 and KEG15107-NAG_5 complexes, as the crystals from the 1:10 ratio of protein to sugar crystallization were isomorphous to the crystals of KEG15107-NAG_4 complex from the 1:2 crystallization trial, their initial structures were determined directly by the refinement of the corresponding structures of the KEG15107-NAG_4 or KEG15107-NAG_5 complexes as a starting structure. The structure refinement of the complexes converged at 1.76 Å and 1.60 Å to R_{factor}/R_{free} values of 20.43%/21.44% and 19.47%/21.07%, respectively. The final model of the KEG15107-NAG_4 complex consists of 434 amino acids, residues 1-215 each from molecules Y and Z together with one and three residues of the LEHHHHHH tag, respectively, 548 water molecules, two SO_4 molecules, 10 PEG molecules, two EPE molecules and 2 NAG_4 molecules. The final model of the KEG15107-NAG_5 complex consists of 434 amino acids, residues 1-215 from each of the molecules Y and Z with one and three residues of the tag, respectively, four SO_4 molecules, four PEG molecules, one EPE molecules, 524 water molecules and 2 NAG_5 molecules.

Structure validation of the complexes was performed by PROCHECK CCP4 suite and the refinement statistics of all the complexes are shown in Figure 4.2 (B-F). In all the KEG15107-NAG structures, the electron density for the bound ligands was of high quality. Since the structures of complexes with NAG_4 and NAG_5 were essentially identical for the 1:10 compared to the 1:2 protein to sugar ratios, only the latter two structures will be discussed given the higher resolution of the data from the crystals grown in the 1:2 crystallization trials (Figure 4.1 D). There was no evidence of sugar binding on Domains 2, 3 and 4 in the KEG15107-NAG_3, KEG15107-NAG_4 and

KEG15107-NAG_5 complexes from the crystallization trials conducted at a higher sugar concentration.

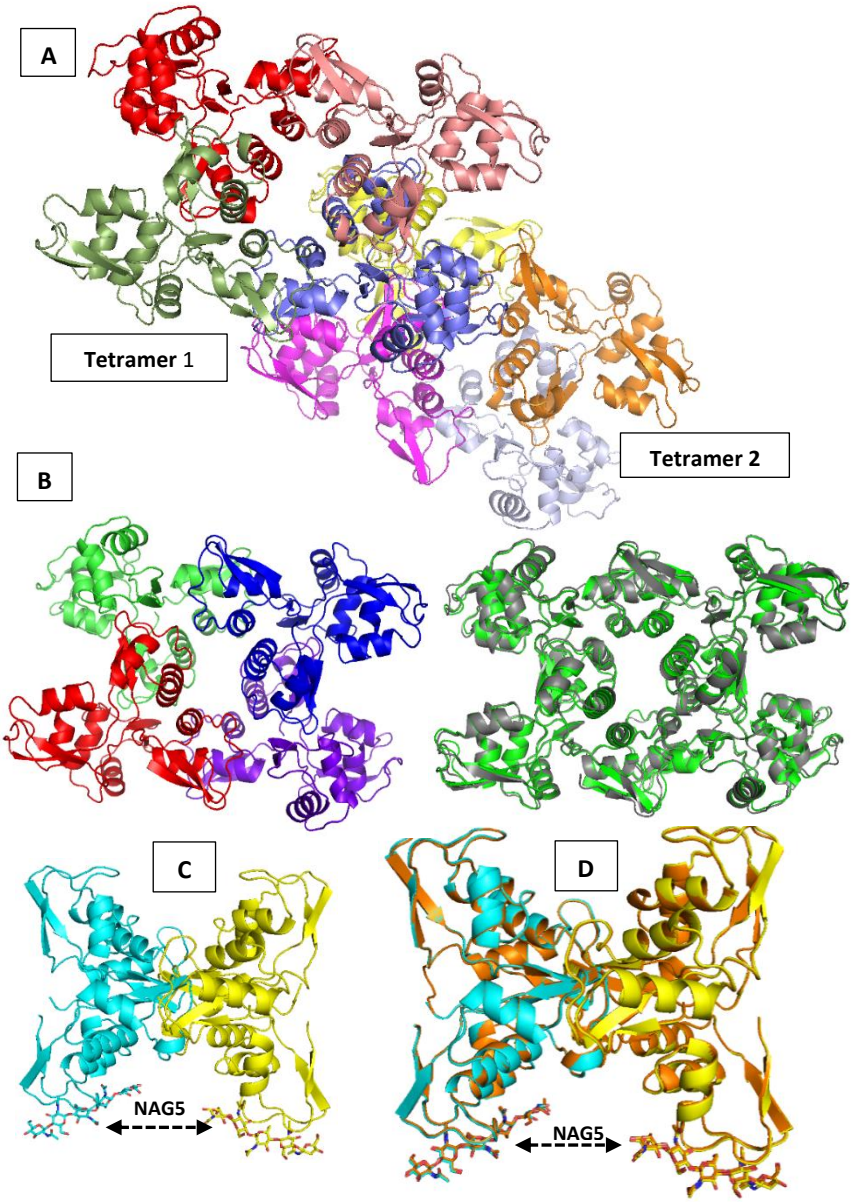
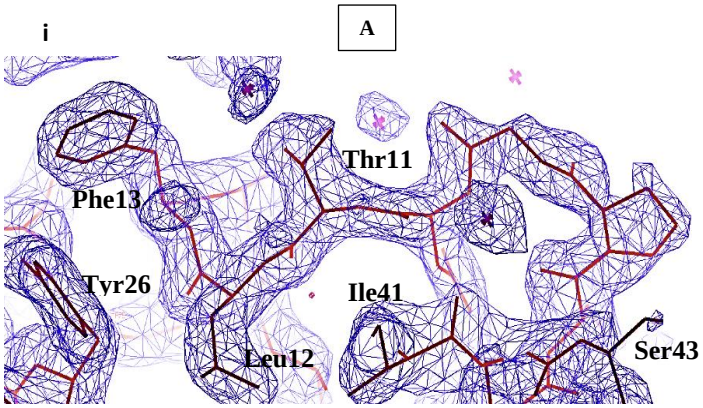


Figure 4.1: Three-dimensional structures of the apo KEG15107 tetramer and its dimer in a complex with NAG_5 oligomer A) Eight monomers of apo KEG15107 form two tetramers (Tetramers 1 and 2) in the AU of the apo crystal. **Key to colors:** (Tetramer 1 of KEG15107: red, dark green, brown and purple, Tetramer 2: yellow, pink orange and grey). B) Tetramer of apo MSMEG3288 (**left**). **Key to colors:** (Tetramer of MSMEG3288: green, red, blue and purple). Superposition of apo tetramer of MSMEG3288 (green) and apo tetramer of KEG15107 (grey) (**right**). C) Two monomers of KEG15107 form a dimer in the AU of the complex with NAG oligomers. D) Superposition of the KEG15107-NAG_5 structures from the 1:2 and 1:10 ratios of protein to sugar revealed that the complexes are closely related. **Key to colors:** (Dimer of KEG15107-NAG_5 from 1: 2 ratio of crystallization trials: light blue and yellow) and (Dimer of KEG15107-NAG_5 from 1:10 ratio of crystallization trials: orange).

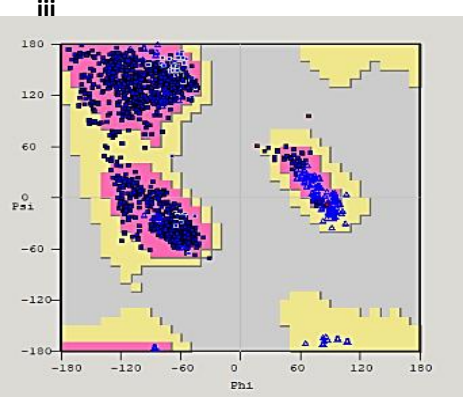
i



ii

Refinement statistics		
Resolution (Å)		2.14
R_{work} (%)		22.2
R_{free} (%)		28.8
Root mean square deviations		
Bond length (Å)		0.017
Bond angles (deg)		1.78
Ramachandran plot		
Non-glycine residues in most favored regions (%)		90.7
Non-glycine residues in additional allowed regions (%)		9
Non-glycine residues in generously allowed regions (%)		0.3
Chains	Residues	Number of Non-H atoms (average B (A^2) total/main chain/side chain)
A	2-216	1670 (30/33/31)
B	1-216	1676 (34/38/36)
C	1-216	1677 (38/41/40)
D	2-216	1670 (40/43/41)
E	2-216	1662 (35/39/37)
F	1-216	1677 (31/35/33)
G	2-216	1670 (34/38/36)
H	1-215	1669 (33/37/35)
Water		1062 (39)

iii



iv

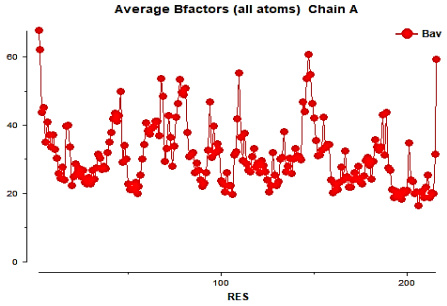
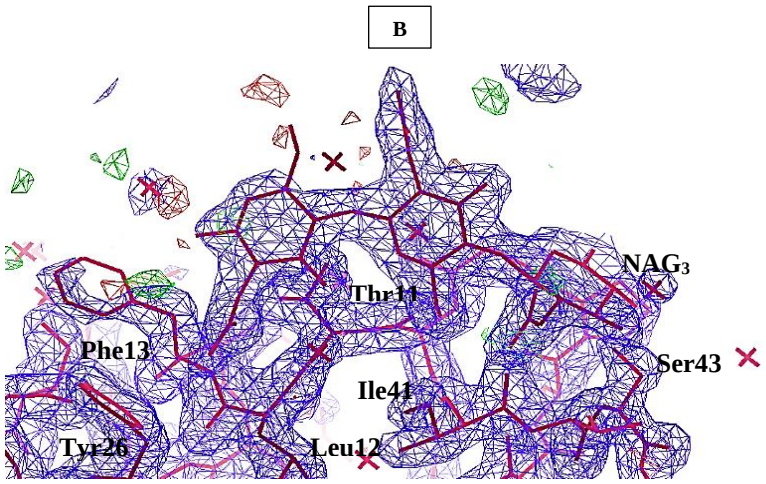


Figure 4.2: The refined structures of apo KEG15107 and its complexes with NAG oligomers. A) A representative portion of electron density map (i) following the final round of refinement of apo KEG15107 together with the refinement statistics (ii) and the Ramachandran plot (iii). The refinement statistics for the respective structure is shown in the table. The B factor plot (iv) for Chain A is also shown in the figure.

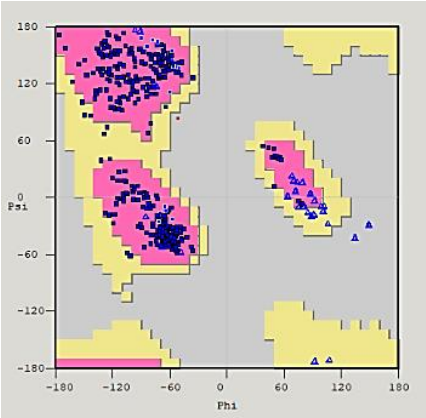
i



ii

Refinement statistics		
Resolution (Å)		1.95
R_{work} (%)		22.4
R_{free} (%)		27.0
Root mean square deviations		
Bond length (Å)		0.015
Bond angles (deg)		1.95
Ramachandran plot		
Non-glycine residues in most favored regions (%)		90.3
Non-glycine residues in additional allowed regions (%)		9.7
Chains	Residues	Number of Non-H atoms (Average B (Å²)total/ main chain/side chain)
Y	1-216	1683 (29/29/29)
Z	1-216	1716 (30/31/30)
Water		396 (40)
NAG	6	88 (29)/(35,27,27,30,27,26)
PEG	8	56 (57)

iii



iv

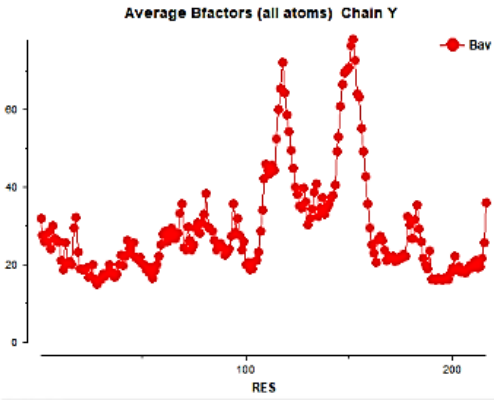
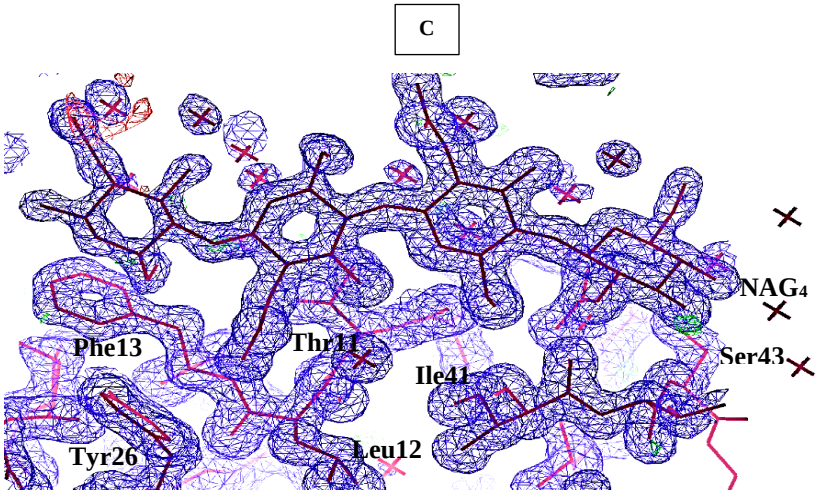


Figure 4.2: B) A representative portion of the electron density map (i) following the final round of refinement of KEG15107-NAG_3 from the 1:10 crystallization trials together with the refinement statistics (ii) and the Ramachandran plot (iii). The refinement statistics for the respective structure is shown in the table. The B factor plot (iv) for Chain Y is also shown in the figure. Protein residues from position 115-118 and 146-156 that are located at the cleft area of the sugar binding possessed high B-factor.

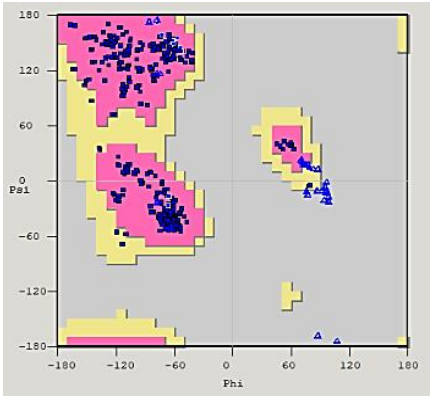
i



ii

Refinement statistics		
Resolution (Å)		1.25
R_{work} (%)		17.90
R_{free} (%)		19.27
Root mean square deviations		
Bond length (Å)		0.013
Bond angles (deg)		1.59
Ramachandran plot		
Non-glycine residues in most favored regions (%)		91.5
Non-glycine residues in additional allowed regions (%)		8.5
Chains	Residues	Number of Non-H atoms (average B (Å²) total/ main chain/side chain)
Y	1-220	1696 (18/21/20)
Z	1-218	1705 (17/20/19)
Water		449 (28)
NAG	8	114 (20)/ (19,15,18,25,28,17,16,21)
EPE	2	30 (30)
PEG	7	49 (48)
SO_4	4	20 (28)

iii



iv

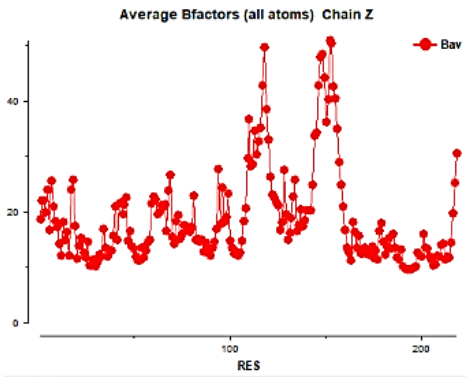
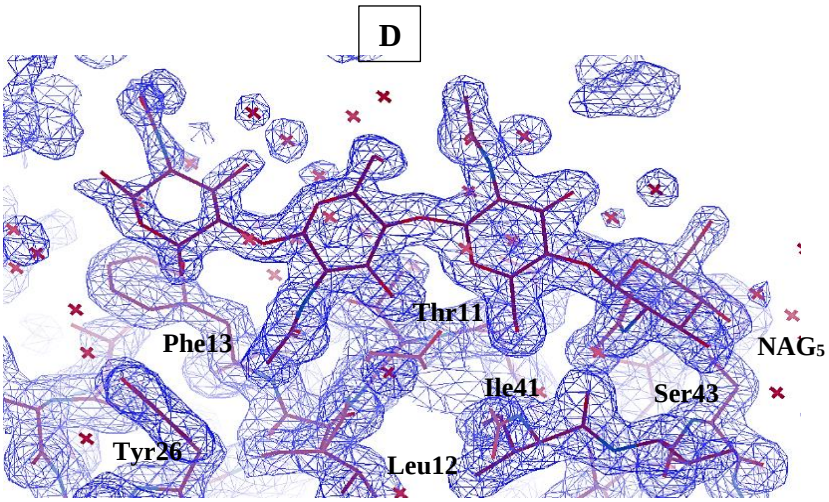


Figure 4.2: C) A representative portion of the electron density map (i) following the final round of refinement of KEG15107-NAG_4 from the 1:2 crystallization trials together with the refinement statistics (ii) and the Ramachandran plot (iii). The refinement statistics for the respective structure is shown in the table. The B factor plot (iv) for Chain Z is also shown in the figure. Protein residues from position 115-118 and 146-156 that are located at the cleft area of the sugar binding possessed high B-factor.

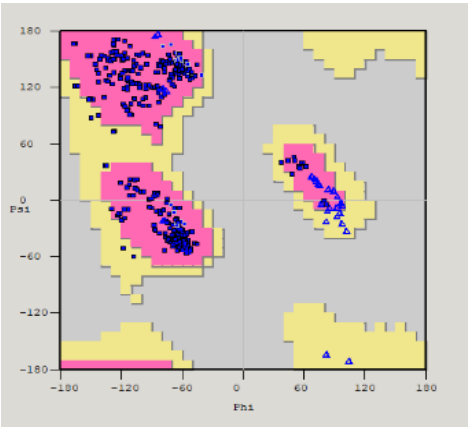
i



ii

Refinement statistics		
Resolution (Å)		1.76
R_{work} (%)		20.43
R_{free} (%)		21.44
Root mean square deviations		
Bond length (Å)		0.013
Bond angles (deg)		1.64
Ramachandran plot		
Non-glycine residues in most favoured regions (%)		91.4
Non-glycine residues in additional allowed regions (%)		8.6
Chains	**Residues**	**Number of Non-H atoms (Average B (A^2)total/ main chain/side chain)**
Y	1-216	1686 (24/28/26)
Z	1-218	1705 (24/28/26)
Water		548 (39)
NAG	8	114 (30)/(45,30,23,24,42,28,22,26)
SO_4	2	10 (59)
PEG	10	70 (30)
EPE	2	30 (32)

iii



iv

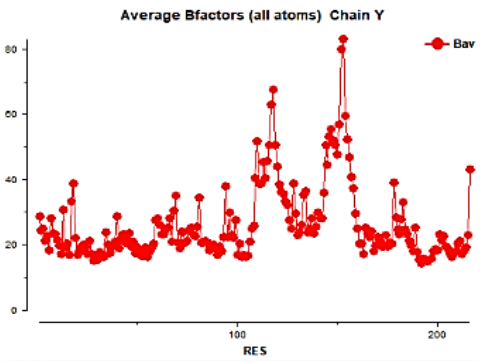
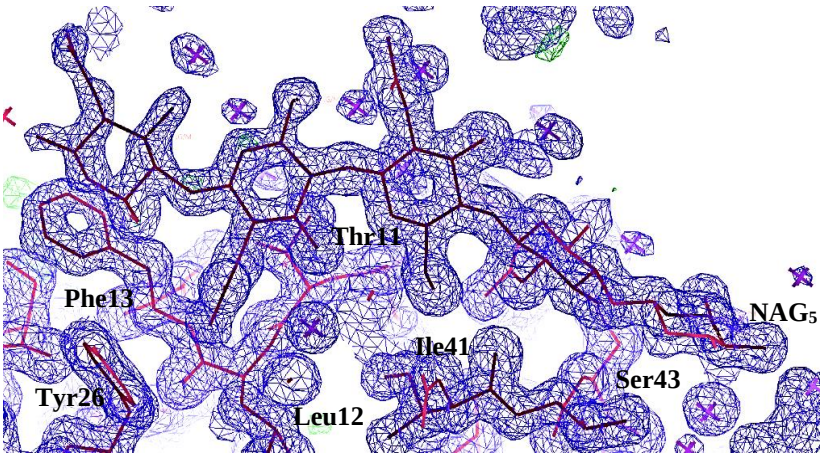


Figure 4.2: D) A representative portion of the electron density map (i) following the final round of refinement of KEG15107-NAG_4 from the 1:10 crystallization trials together with the refinement statistics (ii) and the Ramachandran plot (iii). The refinement statistics for the respective structure is shown in the table. The B factor plot (iv) for Chain Y is also shown in the figure. Protein residues from position 115-118 and 146-156 that are located at the cleft area of the sugar binding possessed high B-factor.

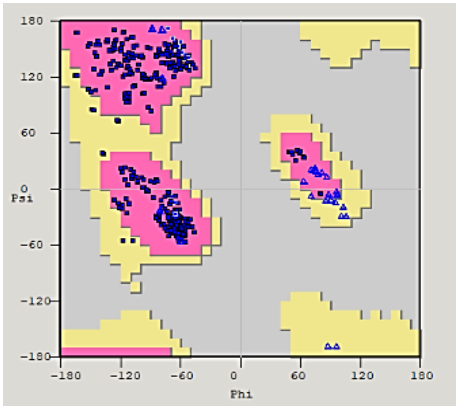
E



ii

Refinement statistics		
Resolution (Å)		1.38
R_{work} (%)		18.00
R_{free} (%)		19.61
Root mean square deviations		
Bond length (Å)		0.013
Bond angles (deg)		1.50
Ramachandran plot		
Non-glycine residues in most favored regions (%)		90.7
Non-glycine residues in additional allowed regions (%)		9.3
Chains	**Residues**	**Number of Non-H atoms (average B (A²)total/ main chain/ side chain)**
Y	1-217	1686 (19/21/20)
Z	1-219	1706 (18/21/20)
Water		600 (35)
NAG	10	143 (25)/(64,21,16,17,27,21,16,15,19,36)
PEG	11	77 (50)
EPE	3	45 (51)
SO_4	5	25 (79)

iii



iv

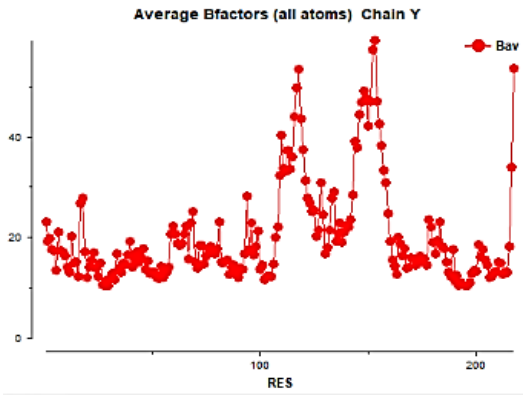
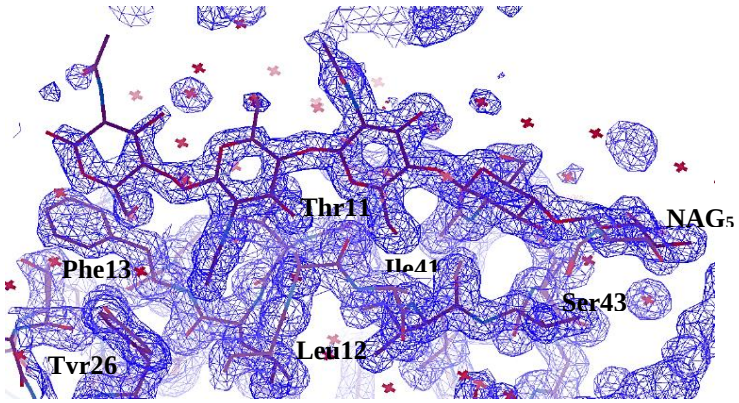


Figure 4.2: E) A representative portion of the electron density map (i) following the final round of refinement of KEG15107-NAG_5 from the 1:2 crystallization trials together with the refinement statistics (ii) and the Ramachandran plot (iii). The refinement statistics for the respective structure is shown in the table. The B factor plot (iv) for Chain Y is also shown in the figure. Protein residues from position 115-118 and 146-156 that are located at the cleft area of the sugar binding possessed high B-factor.

F

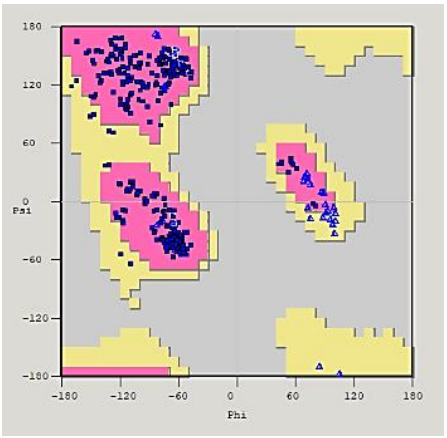
i



ii

Refinement statistics		
Resolution (Å)		1.60
R_{work} (%)		19.47
R_{free} (%)		21.07
Root mean square deviations		
Bond length (Å)		0.014
Bond angles (deg)		1.79
Ramachandran plot		
Non-glycine residues in most favored regions (%)		91.1
Non-glycine residues in additional allowed regions (%)		8.9
Chains	Residues	Number of Non-H atoms (average B (Å²) total/ main chain/side chain)
Y	1-216	1686 (20/22/21)
Z	1-218	1705 (19/22/21)
Water		524 (34)
NAG	10	143 (28)/(48,20,19,20,26,52,23,18,20,32)
SO_4	4	20 (48)
PEG	4	28 (39)
EPE	1	15 (42)

iii



iv

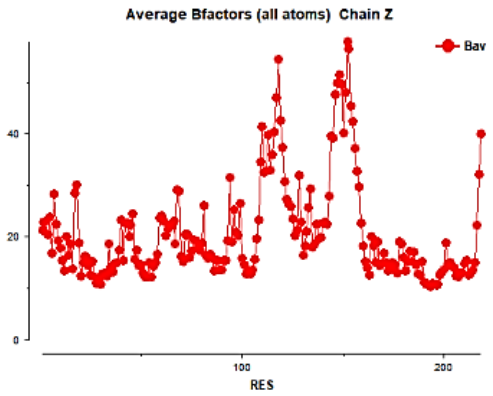


Figure 4.2: F) A representative portion of the electron density map (i) following the final round of refinement of KEG15107-NAG_5 from the 1:10 crystallization trials together with the refinement statistics (ii) and the Ramachandran plot (iii). The refinement statistics for the respective structure is shown in the table. The B factor plot (iv) for Chain Z is also shown in the figure. Protein residues from position 115-118 and 146-156 that are located at the cleft area of the sugar binding possessed high B-factor.

4.3 Three-dimensional X-ray structure of apo KEG15107

The three-dimensional structure of apo KEG15107 shows that the protein contains four domains each approximately 50 residues and similar in the fold to the classical LysM domain with two α-helices (H1 and H2) packed onto each other, and facing two strands of an antiparallel β-sheet (S1 and S2) (Figure 4.3 A-B). These four domains were designated as LysM 1, 2, 3 and 4 and are linked to each other by small loops (Figure 4.3 C). Two loops in the LysM fold, designated as L1 and L2 lie between S1 and H1 and between H2 and S2, respectively, and are adjacent in the 3-D structure with a shallow cleft between them (Figure 4.4 and 4.5). Structure-based sequence analysis showed that the four domains only share low sequence identity (~10%) (Figure 4.3 D).

The sequence conservation is concentrated in four regions corresponding to Motifs 1, 2, 3 and 4 (Figures 4.3 D and 4.4). Motif 1 (GDTL) in Domains 1 and 4 resides within the L1 loop. In Domains 2 and 3, the equivalent residues of this motif are ADST and GESL, respectively. The conserved residues of Motif 2 (YGD/T) lie on a loop connecting the H1 and H2 helices of the domains. Motif 3 corresponds to the conservation of an asparagine residue at the start of the L2 loop (N33, N90, N143, and N197). Motif 4 is located on the L2 loops and includes the strong conservation of a glycine residue (G44, G100, G154) in Domains 1, 2 and 3 but with this glycine not being conserved in Domain 4 (N208).

Superposition of Domains 2, 3 and 4 onto Domain 1 of KEG15107 confirmed that these domains possess a very similar fold with rmsd values of 1.030 Å, 0.312 Å, and 0.865 Å, respectively (Figure 4.5). The main differences between the domains lie in the conformation of the L1 loop which is similar in Domains 1, 2 and 4 but different in Domain 2 (Figure 4.5 A-B). These results show that the cleft surface is similar in Domains 1, 3 and 4, while in Domain 2 it is less open as a result of the different conformation (Figure 4.5 C-D).

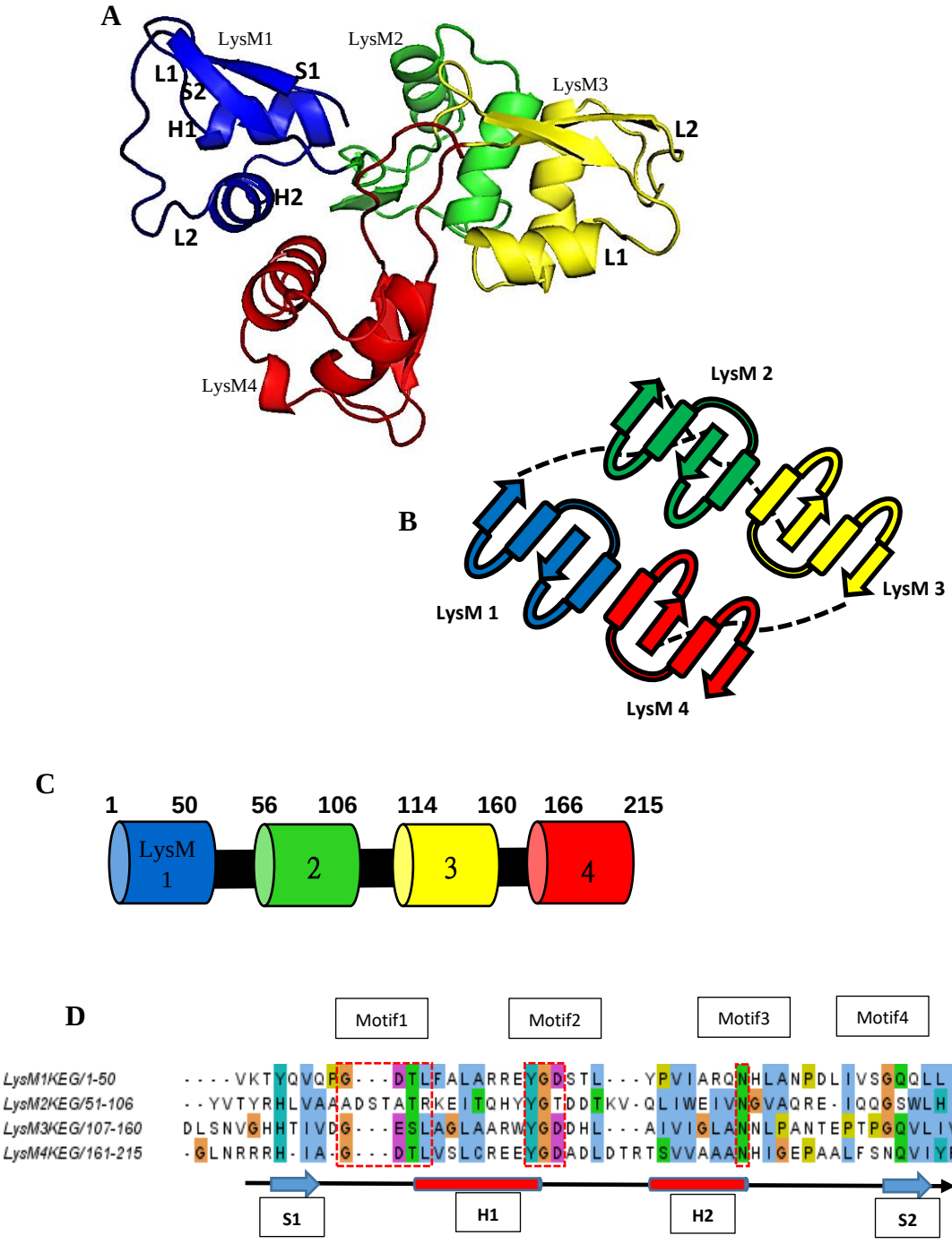


Figure 4.3: A cartoon representation of the three-dimensional structure of KEG15107 from *M. avium*. A) The four LysM domains are colored in blue (Domain 1), green (Domain 2), yellow (Domain 3) and red (Domain 4). B) A schematic illustration of the domain arrangement showed that the four domains are packed against each other. C and D) The architecture of the domain arrangement of KEG15107 showing Domains 1, 2, 3 and 4 are linked to each other by small loops that are similar in length together with the structure-based sequence alignment of the four LysM domains of KEG15107 to show that the sequence conservation is concentrated in four regions.

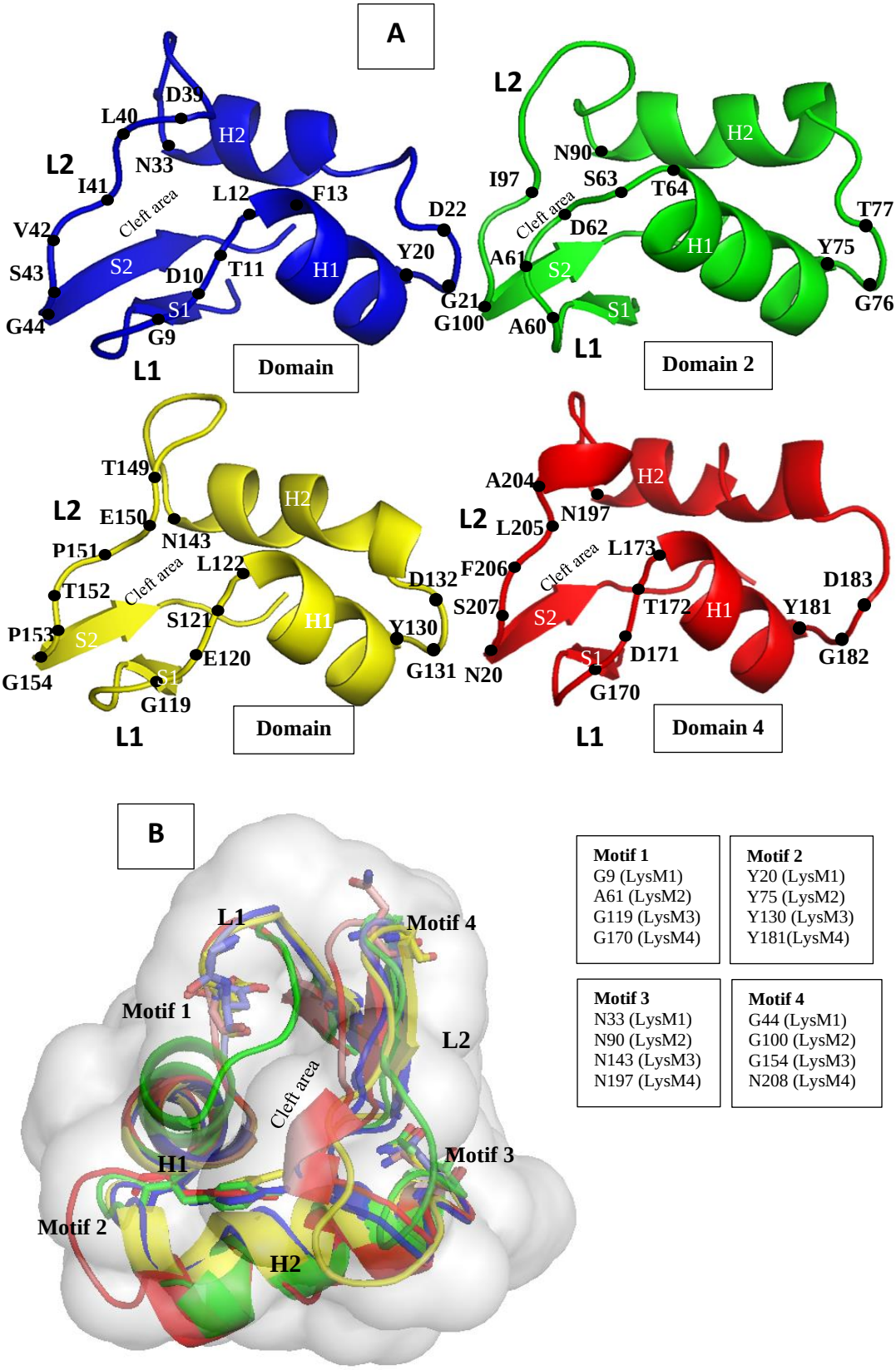


Figure 4.4: A schematic diagram of the structures of the four LysM domains of KEG15107. A) The diagram shows the position of the α-carbon atoms of residues forming the four regions which correspond to the four conserved motifs. B) Superposition of the four LysM domains reveals four highly conserved regions with identified residues displayed in the table.

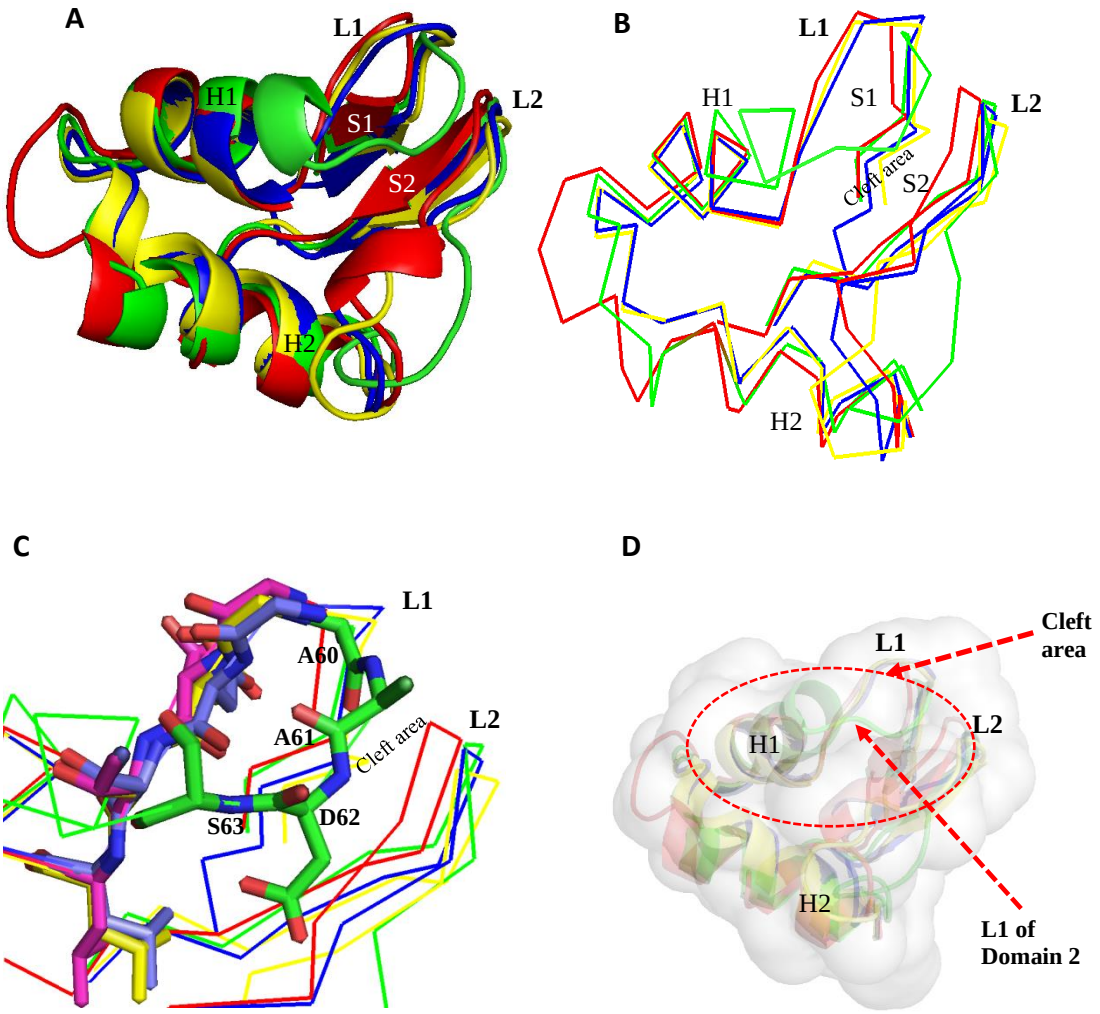


Figure: 4.5: Analysis of the domains of KEG15107. A) and B): The superposition of Domains 1, 2, 3 and 4 of KEG15107 revealed that the domains share a similar LysM fold (rmsd values for the similarity given associated in the table). C) The superposition of the four LysM domains of KEG15107 showed the L1 of Domain 2 was the most different compared to the other domains. This difference in conformation blocks the cleft area between L1 and L2 for Domain 2. D) The superposed structures of the four domains in a surface representation clearly showed the cleft area in between the L1 and the L2 motifs. Key color: Blue (Domain 1), Green (Domain 2), Yellow (Domain 3), Red (Domain 4).

4.3.1 KEG15107 is a globular protein

Structure analysis of apo KEG15107 revealed that it is a globular protein with each of its four LysM domains packed tightly against each other. The structure shows that there are multiple interactions occur between the various domains. The interactions between Domains 1 and 4 are mediated by contacts involving the main chain and side chains of residues from helix 2 of the respective domains (Figure 4.6 A). These contacts involved a mix of hydrogen bonds mediated by water molecules together with hydrophobic and van der Waals interactions. Domains 2 and 3 interact through contacts principally mediated by side chains of hydrophobic residues from helix 2 in both domains (Figure 4.6 B). Protein contacts between Domains 1 and 2 involve the YGD motifs located on the loop between the two helices of Domain 1, and residues located at the very beginning of strand 1 of Domain 2. These contacts involve direct and water-mediated hydrogen bonds (Figure 4.6 C). Domains 3 and 4 also interact with each other through the YGD motif in Domain 3 and the residues located at the very beginning of strand 1 of Domain 4 (Figure 4.6 D).

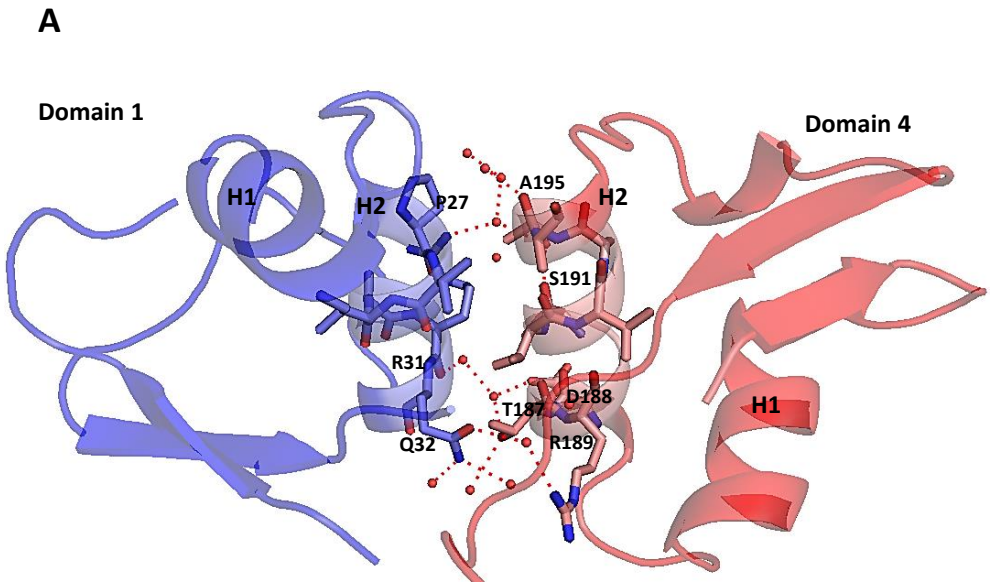


Figure 4.6: Crystal contacts of four LysM domains of KEG15107. A) The diagram illustrates the contacts between Domains 1 and 4 through residues on H2 of both domains.

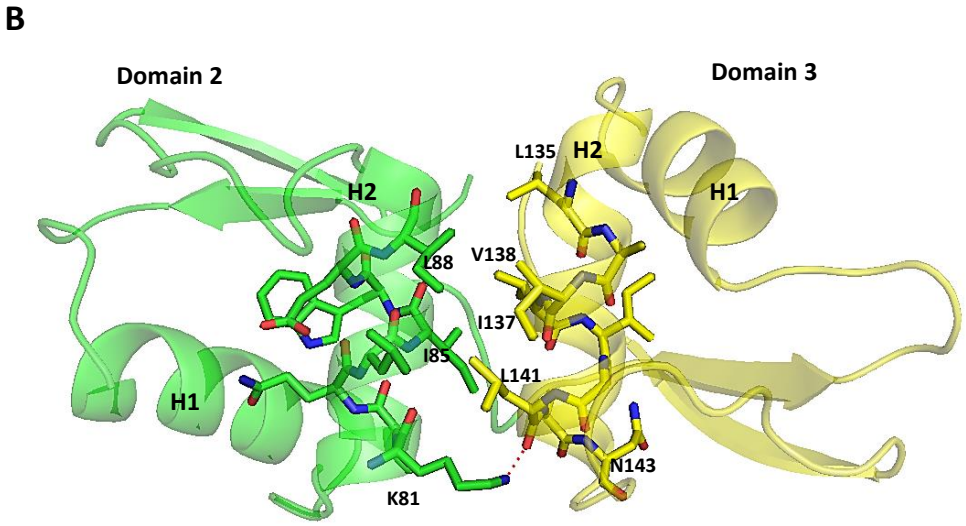


Figure 4.6: B) The diagram illustrates the contacts between Domains 2 and 3 involving residues from H2 of both domains.

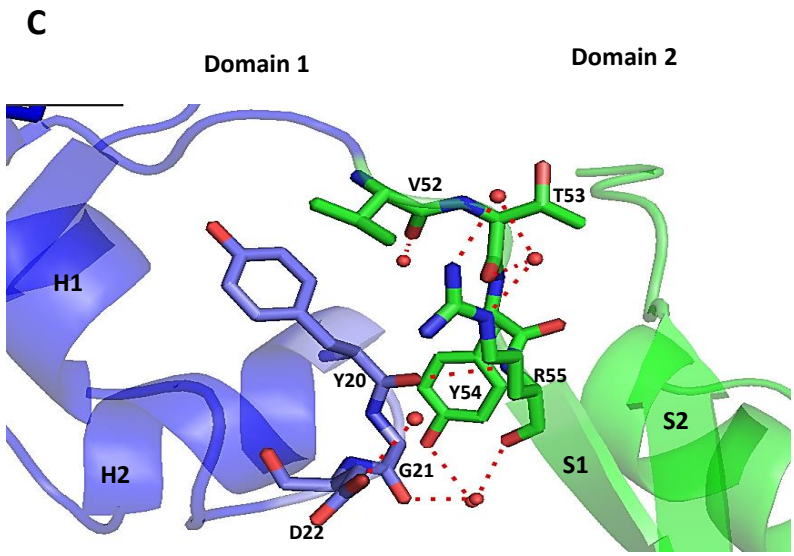


Figure 4.6: C) The diagram illustrates the contacts between Domains 1 and 2 involving the YGD motif located on the loop between H1 and H2 of Domain 1 and residues at the very beginning of S1 of Domain 2.

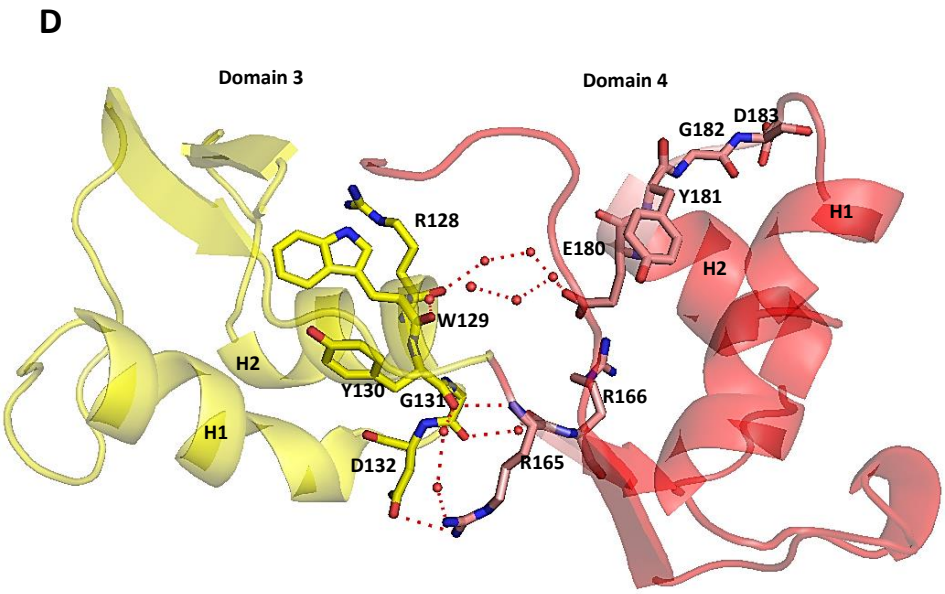


Figure 4.6: D) The diagram illustrates the contacts between Domains 3 and 4 involving the residues from the YGD motif on Domain 3 and residues at the very beginning of Domain 4.

4.4 Analysis of the KEG15107 tetramer

The structure of the apo KEG15107 protein revealed that it forms a tetramer and the asymmetric unit of the crystal consist of eight monomers, designated as subunits A, B, C and D, and four equivalent subunits of a second tetramer, E, F, G and H (Figure 4.1 A) and (Figure 4.7 A). Superposition of the four monomers reveals that they share almost an identical conformation (Figure 4.7 B). The formation of the KEG15107 tetramer is mediated by contacts between Domain 1 of one monomer with its symmetry-related partner, and more extensive contacts between Domain 4 of one monomer with Domains 2 and 4 of a 2-fold related partner. There are two examples of each of these interactions in the tetramer indicated by circles (Figure 4.7 A). Overall, each of the monomers of the tetramer has an accessible surface area of approximately 11800 A^2 (value calculated by PDBePISA and rounded to the nearest 50 A^2 averaged over four subunits). Interactions between Domain 4 and Domains 2 and 4 of asymmetry-related partner involve approximately 1450 A^2 of buried area. Interactions between Domain 1 and its symmetry-related partner involve approximately 450 A^2 of the buried area of the monomers. Thus, the interactions involving Domain 4 dominate the subunit assembly forming the dimer with the interactions between Domains 1 which give rise

to the tetramer being far weaker. This provides one possible explanation as to why the tetramer breaks down to a dimer on the binding of polyNAG to Domain 1.

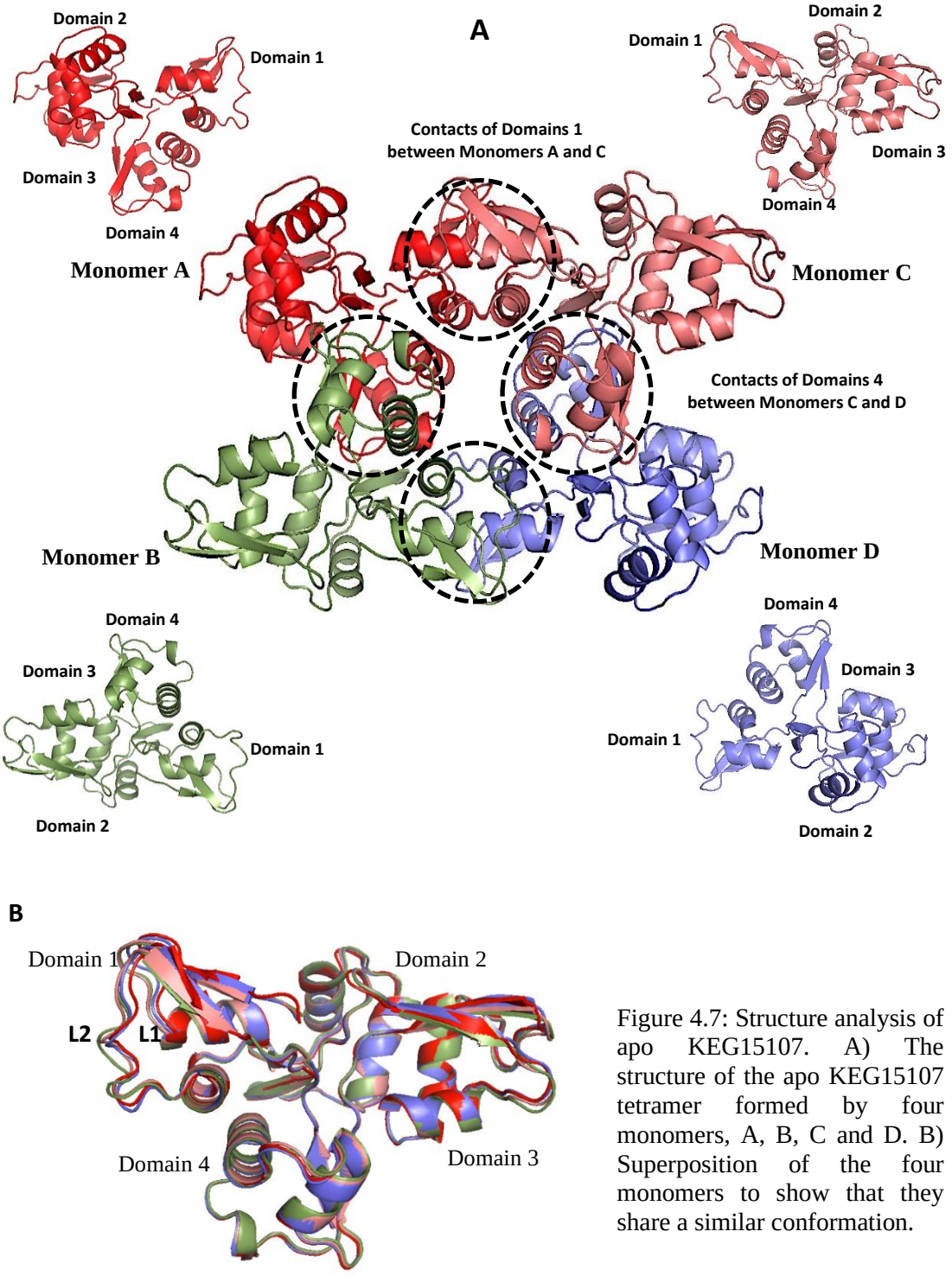


Figure 4.7: Structure analysis of apo KEG15107. A) The structure of the apo KEG15107 tetramer formed by four monomers, A, B, C and D. B) Superposition of the four monomers to show that they share a similar conformation.

4.5 Molecular recognition of oligosaccharides by KEG15107

The crystallization trials were subsequently carried out on KEG15107 in complex with poly$NAG_{(n)}$ substrates to further investigate the biological function of the protein. The KEG15107 protein was co-crystallized with three different NAG oligomers (NAG_3, NAG_4, and NAG_5) using either 1:2 and/or 1:10 protein to sugar ratios.

Structure analysis revealed the KEG15107-NAG_3, KEG15107-NAG_4 and KEG15107-NAG_5 complexes were consistently observed as a dimer formed by two subunits (designated as Y and Z). The accessible surface areas of the monomers are 11450 A^2 (calculated by PDBePISA again rounded to 50 A^2 and averaged between two subunits). Contacts between the dimer involve interactions between Domain 4 of one monomer in Domains 2 and 4 of the 2-fold symmetry-related partner as seen previously in the structure of apo KEG15107. The buried area between the Monomers Y and Z is 1450 A^2. A NAG oligomer was observed bound to Domain 1 of each subunit of the dimer at the cleft region in between the L1 and L2 loops (Figure 4.8 A). The NAG_5, NAG_4 and NAG_3 molecules were observed to bind in similar conformations (Figure 4.12). There was no evidence for binding of an oligosaccharide to any of the other domains of the protein. Analysis of the crystal packing showed that there were no tetramers in the crystals. Superposition of the three structures of the complexes showed that the dimers were very closely related (Figure 4.8 B and C).

Structural comparison of the apo tetramer and the dimers of KEG15107 in complex with NAG molecules showed that the dimer in the latter is essentially identical to one of the dimer interfaces in the tetramer (Figure 4.9 A). The detailed comparison revealed that the interactions in the tetramer of the apo protein between the 2-fold related copies of Domain 1 are abolished in the sugar complexes due to the adverse steric interactions between atoms in the sugar and residues that in the apo protein form contacts occurring the dimer interface involving Domain 1 (See Chapter 5: Section 5.3.3: Figure 5.9). Thus oligosaccharide binding to Domain 1 blocks association of the dimers preventing tetramer formation. Moreover, since these binding sites for the sugar are partially buried from the solvent in the tetramer, this implies that the tetramer must dissociate to dimers prior to sugar binding. Comparison of the conformation of Domain 1 in the apo protein and in the complex with sugars shows that the side chain of Phe13 moves significantly to allow access to the S4 sugar binding site in Domain 1 (Figure 4.10).

Examination of binding of different NAG oligomers to Domain 1 of KEG15107 showed that five binding sites could be identified, designated as S0, S1, S2, S3, and S4 (Figure 4.11). The sugar molecules bound to Domain 1 were designated as N0, N1, N2, N3, and N4. The binding sites together with bound sugars are numbered from the non-reducing end to the reducing end of the oligosaccharides (Numbering, 1983). For NAG_5, all five sites are occupied for both subunits in the AU (Figure 4.11 B). However, in the complex of NAG_4, in Chain Y the binding of sugar is observed in S0, S1, S2, and S3 whereas in Chain Z sugars occupy sites S1, S2, S3 and S4 (Figure 4.11 C). In the complex of KEG15107-NAG_3, the binding sites were observed in S1, S2, and S3 (Figure 4.11 A). Superposition of the NAG_3, NAG_4, and NAG_5 complexes revealed that these molecules share similar conformations (Figure 4.11 D). Given that binding sites S1, S2, and S3 are occupied in all the complexes, these are assumed to have a higher binding affinity while the S0 and S4 sites are presumed to be of lower affinity towards sugar molecules.

Detailed analysis of the binding pockets of Domain 1 revealed three prominent cavities in the shallow groove between the L1 and L2 loops corresponding to binding sites S1, S2, and S3 (Figure 4.12 A). The N-acetyl substituents at C-2 of the NAG molecules N1 and N3 fit well to cavities of these binding pockets. Numerous hydrogen bonds and van der Waals interactions can be identified between the protein and the NAG molecules (Figure 4.12 B and C). The NAG molecules mainly interact with the main chain of residues of the protein from the L1 and the L2 regions (G9, L12, F13, D39, I41, and S43). At the S1 site, three residues, G9, I41, and S43 interact with the C-2 N-acetyl group and hydroxyl group on the pyranose ring of the N1 sugar molecule through their main chain atoms, while the side chain of S43 also interacts with the C-3 hydroxyl group of the N1 sugar ring. The C-6 hydroxyl of the N2 sugar makes hydrogen bonds with both main chain atoms of I41. In the S3 binding pocket, interactions involving the C-2 N-acetyl group of the N3 sugar with the main chain of the F13 and the C-3 hydroxyl group of the main chain of D39 can be identified. The side chain of D39 makes direct hydrogen bonds to the C-2 N-acetyl group of the N3 sugar. A water-mediated hydrogen bond is also found between the main chain of L12 and the C-3 hydroxyl of the N3 sugar. The pyranose ring of the N4 sugar makes a π-stacking interaction to the aromatic ring of F13, while the N0 sugar was observed to make a single hydrogen bond to the side chain of S43 at the S1 binding site. The findings of oligosaccharide binding by Domain

1 of KEG15107 is similar to the AtCERK1 protein in chitin recognition. The binding sites S0 to S3 of Domain 1 of KEG15107 is equivalent to the chitin binding sites of NAG1 to NAG4 of AtCERK1 from A. thaliana (Liu et al., 2012).

The C-3 hydroxyl of the NAG molecules of N1 and N3 are bound in sterically restricted pockets. The replacement of the hydroxyls by larger lactic acid substituents in NAM molecules would prevent NAM residues from binding at these sites. Thus, S1 and S3 can only bind NAG residues while sites S0, S2 and S4 can bind NAG or NAM as the conformation of the oligosaccharide is such that the C-3 hydroxyl faces the solvent where the larger lactyl substituent to which the peptide linker that covalently connects different peptidoglycan chains in the cell wall can be accommodated. Detailed information about the interactions between the KEG15107 protein and NAG molecules is shown in Table 4.1. Domain 1 interactions are established mainly through hydrogen bonds between the bottom part of the sugar and the main chain of KEG15107. Specific recognition of the N-acetyl groups in the binding pockets of S1 and S3 allows KEG15107 to discriminate NAG and NAM molecules in the binding of peptidoglycan. However, given the similarity of the four LysM domains, this raises questions as to why no binding of sugars has been observed to Domains 2, 3 and 4.

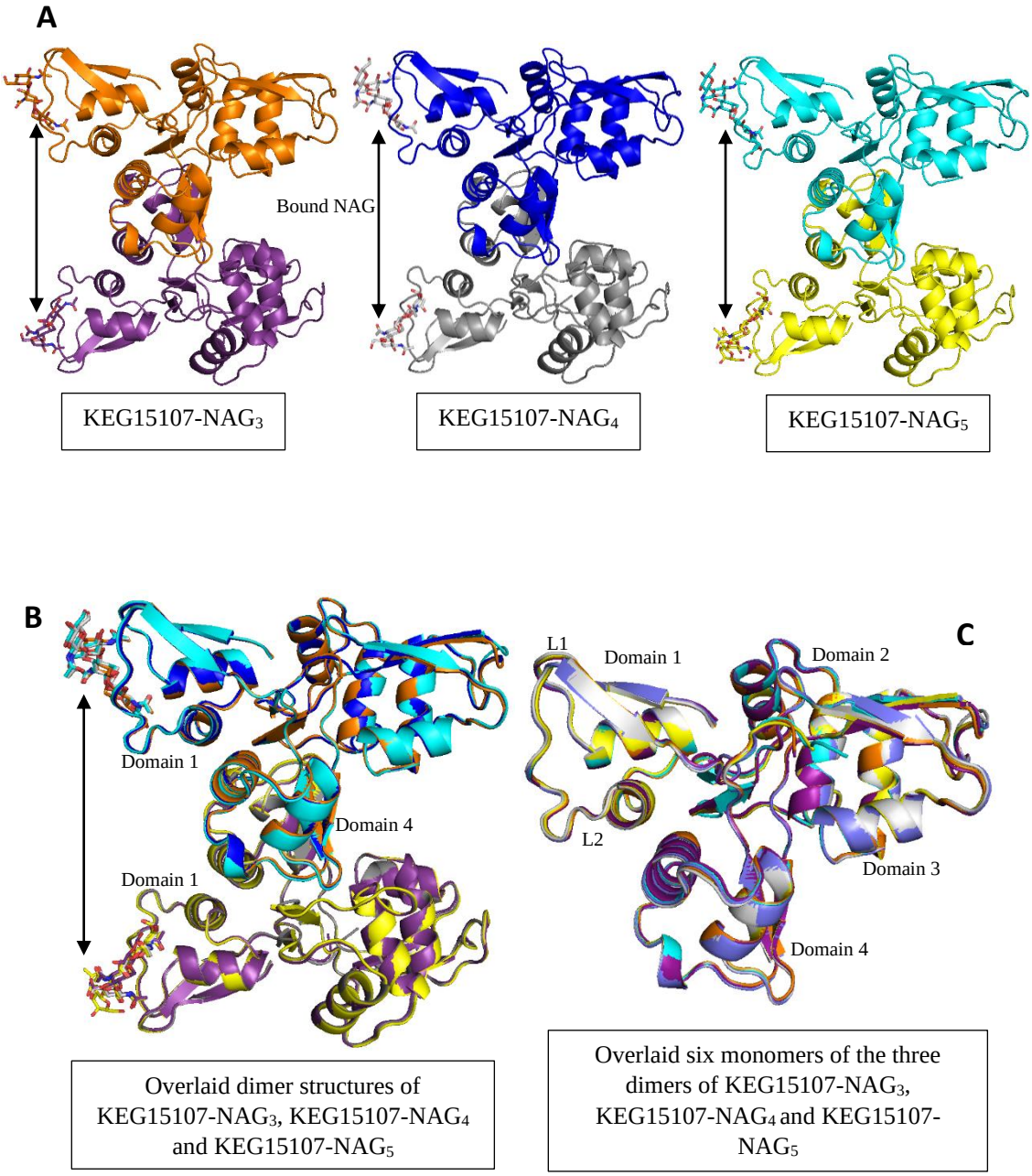
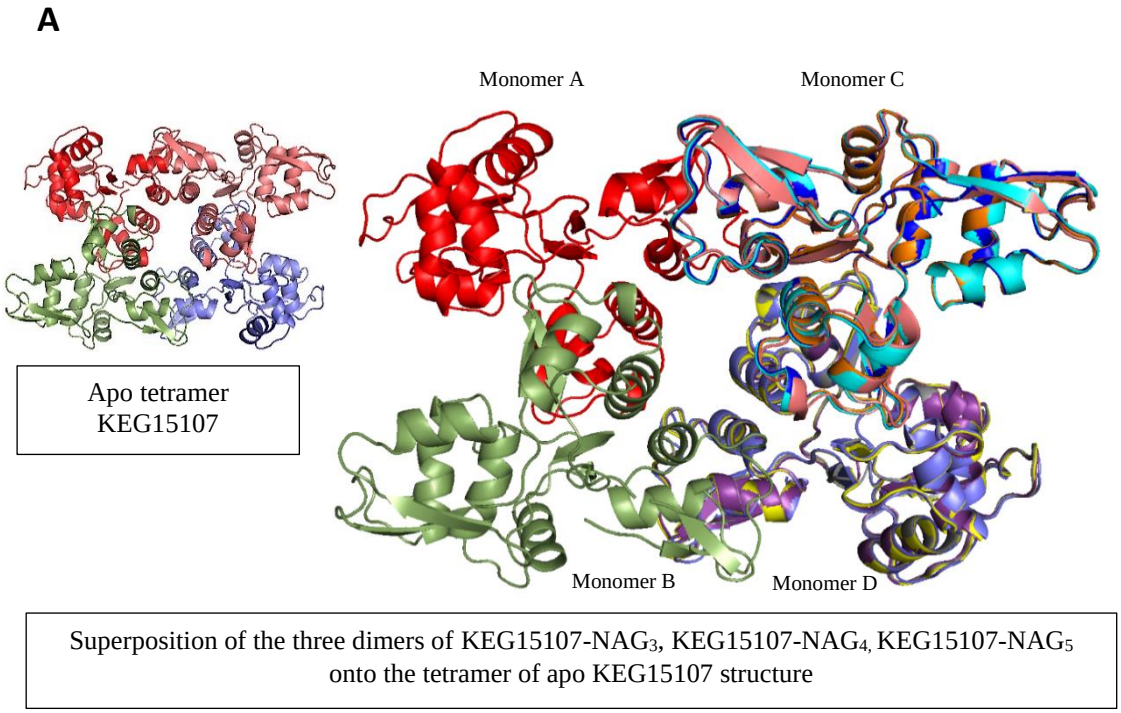


Figure 4.8: Comparison of the three-dimensional structures of KEG15107 in complex with NAG_3 (1:10), NAG_4 (1:2) and NAG_5 (1:2) oligomers. A) The KEG15107 protein was observed as a dimer in the NAG complexes. B) Superposition of the three KEG15107-$NAG_{(n)}$ complexes show the complexes share an identical dimeric quaternary structure. NAG molecules bound on Domain 1 are indicated by arrows. C) All the six monomers of the three dimers of the complexes were superposed onto each other and the analysis shows that the conformation of the subunits was essentially identical. (Key colors: the color scheme above (KEG15107-NAG_3 (orange and purple), (KEG15107-NAG_4 (blue and grey), (KEG15107-NAG_5 (light blue and yellow)).



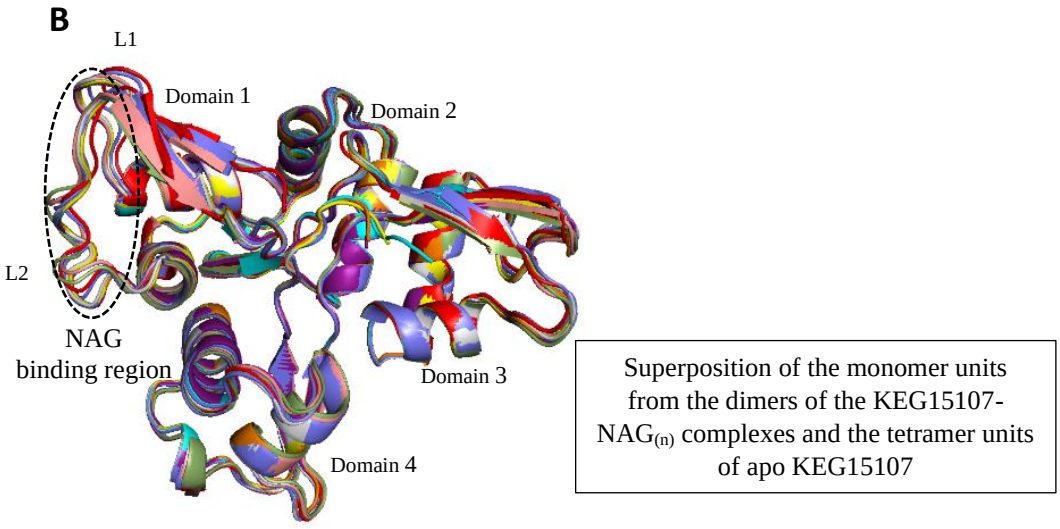


Figure 4.9: Structural analysis between the apo tetramer KEG15107 and the dimeric structures of KEG15107–$NAG_{(n)}$ complexes. A) Superposition of the dimeric structure of the KEG15107-NAG complexes onto the apo tetramer KEG15107 revealed that the dimer of the protein was identical to the half of the tetramer B) Superposition of 10 monomers from the apo KEG15107 tetramer and dimer shows that the structures were highly similar except the L1 and L2 loops (circled area) of Domains 1 of the monomers. The color scheme is similar to Figure 4.8.

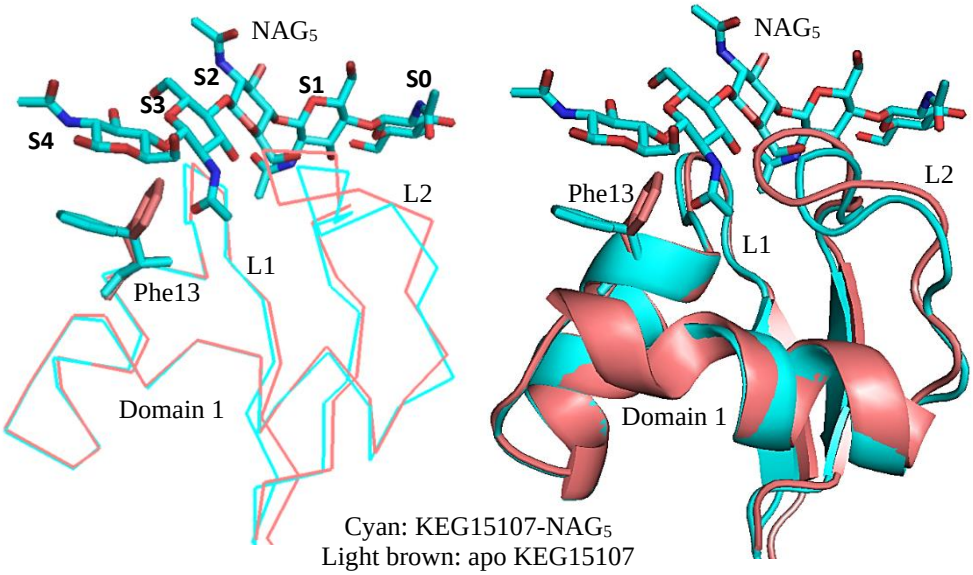


Figure 4.10: Conformational change in the binding region of Domain 1 of KEG15107 triggered by sugar binding. Detailed comparison of the three-dimensional structure of the apo tetramer and the dimer of the KEG15107-NAG complex revealed that Phe13 of Domain 1 of the apo KEG15107 flips to allow sugar molecules in site S4 to bind.

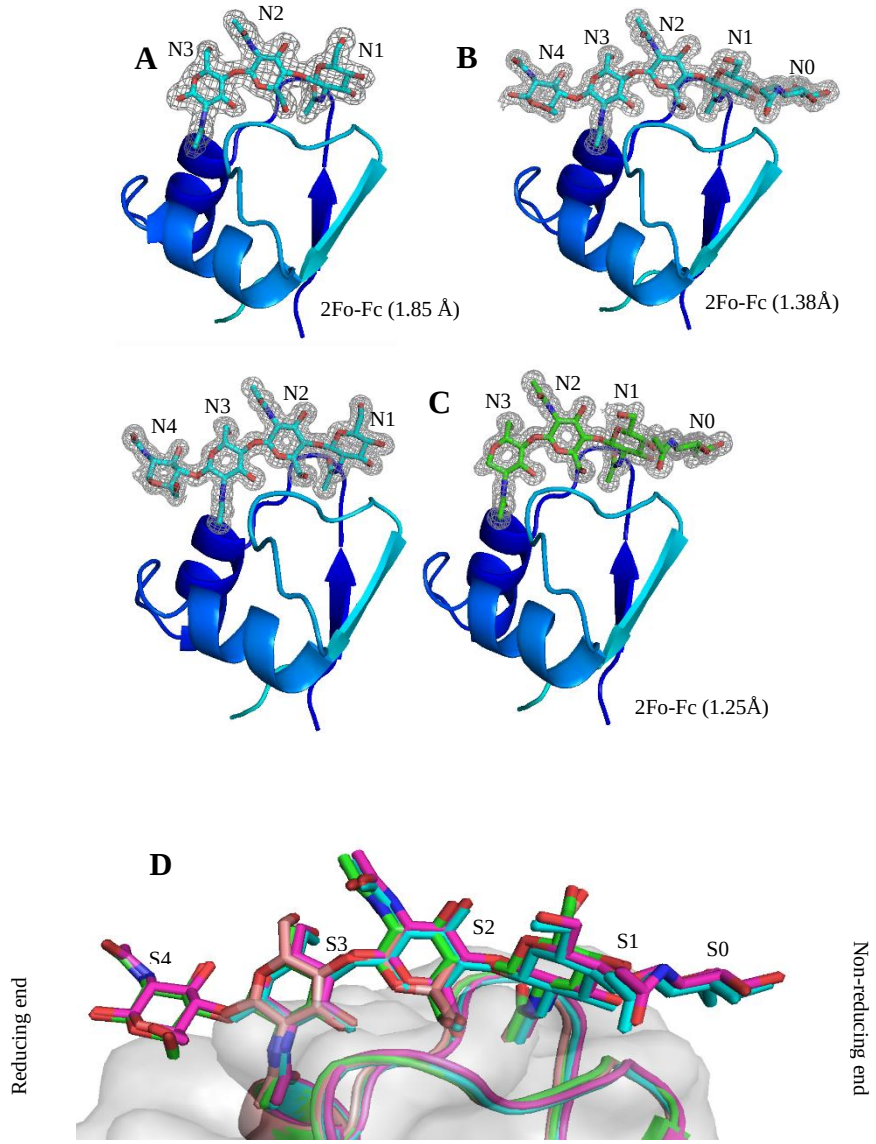


Figure 4.11: Three-dimensional structures of Domain 1 of KEG15107 either with bound NAG_3 or NAG_4 or NAG_5 oligomers. A) Domain 1 in a complex with NAG_3 oligomer. In the NAG_3 complex, both subunits in the AU are occupied by the NAG molecules at similar binding sites (S0-S2). B) Domain 1 in a complex with NAG_5 oligomer. In the NAG_5 complex, both subunits in the AU are occupied by the NAG molecules at similar binding sites (S0-S4). C) Domain 1 in a complex with NAG_4 oligomer. In the NAG_4 complex, the two subunits in the AU show different patterns of sugar binding either to sites S0-S3 or S1-S4. D) Superposition of bound NAG_3 (Tint), NAG_4 (cyan and green) and NAG_5 (pink) revealed five binding sites, S0, S1, S2, S3, and S4 on the shallow cleft of Domain 1.

Table 4.1: Molecular interactions between oligosaccharide residues and protein residues of KEG15107

Sugar binding pockets on Domain 1 of KEG15107																				
S0				S1				S2				S3				S4				
Protein		Sugar	H-Bond length (Å)	Protein		Sugar	Bond length (Å)	Protein		Sugar	H-Bond length (Å)	Protein		Sugar	H-Bond length (Å)	Protein		Sugar	H-Bond length (Å)
MC	SC	PR/SC		MC	SC	PR/SC		MC	SC	PR/SC		MC	SC	PR/SC		MC	SC	PR/SC	
S43		SC(M)	2.9	G9		SC(A)	3.0	D39		SC(M)	2.8	L12 +H_2O		PR(H)	2.9		F13	PR	Π-stacking
				I41		SC(A)	2.9	I41		SC(M)	3.0	F13		SC(A)	2.9				
				S43		PR(H)	3.1	I41		SC(M)	3.2	D39		PR(H)	3.2				
					S43	PR(H)	3.5						D39	SC(A)	2.9				

MC: main chain, SC: side chain, PR: pyranose ring, H: hydroxyl group, M: methyl group, A: N-acetyl group, H-bond: hydrogen bond

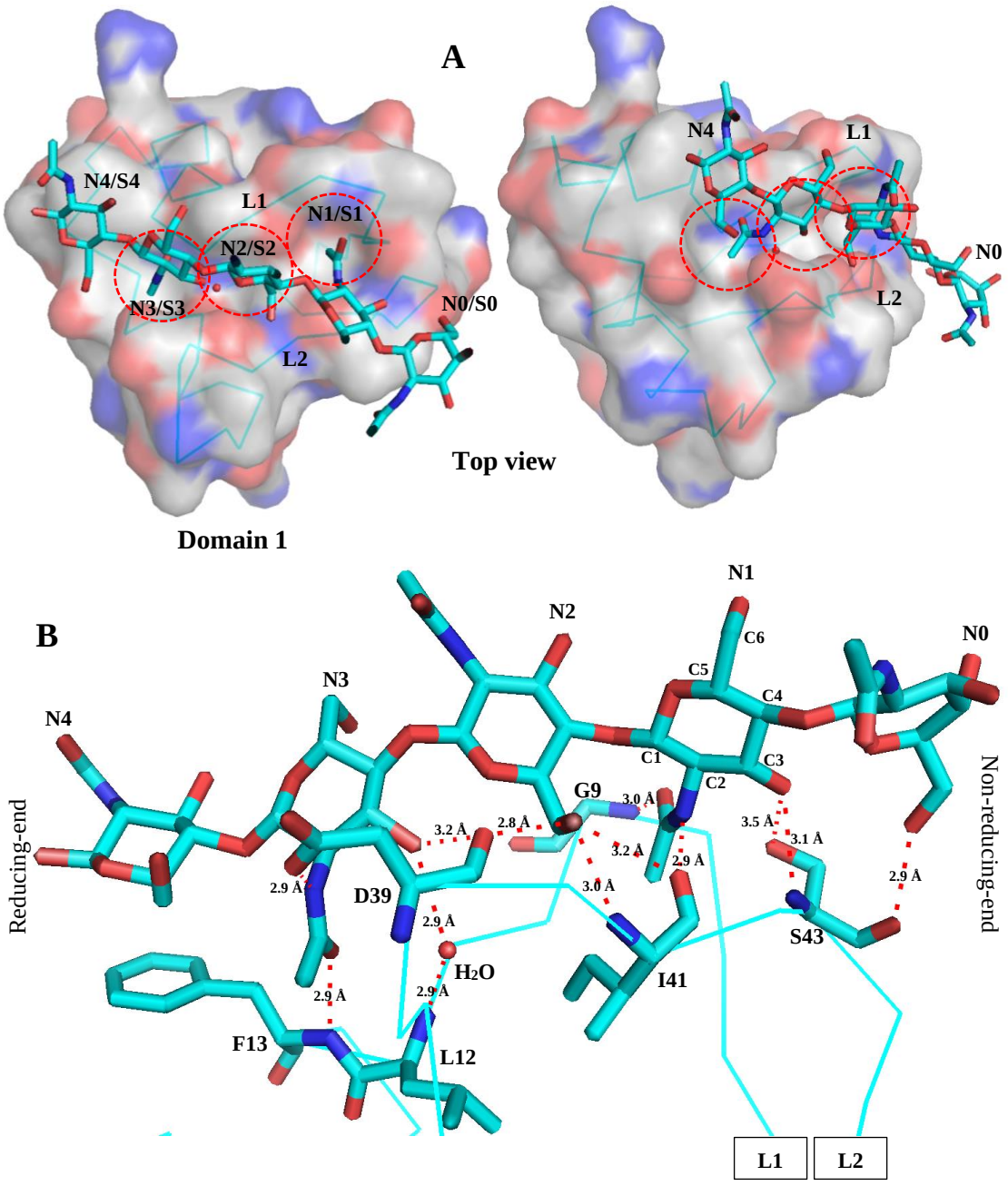


Figure 4.12: Oligosaccharide recognition by Domain 1 of KEG15107. A) Analysis of the atomic interactions between NAG_5 molecules and protein residues from Domain 1 of KEG15107, revealed that the protein has three extensive and two minor sugar binding pockets. The three higher affinity pockets, S1, S2 and S3 are circled in red. The sugar-binding pockets lie between the L1 and the L2 loops of Domain 1 (S0, S1, S2, S3, and S4). B) Detailed analysis of the oligosaccharide binding showed interactions occurred between the NAG molecules and the protein residues. A water molecule trapped in the shallow groove of the protein mediates hydrogen bonds between the N3 NAG molecule and the protein residues at the S3 binding pocket.

4.6 LysM domains of species variants possess very similar oligosaccharide binding pockets

Seven LysM modules found in different proteins whose structures had been determined were analyzed against the LysM Domain 1 of KEG15107, from *M. avium*. These were from NlpC-endopeptidase from *Thermus thermophilus* (4UZ3), SleL-spore cortex from *Bacillus cereus* (4S3J), ykuD-peptidase from *Bacillus subtilis* (4A1I), Ecp6-secreted protein from *Cladosporium fulvum* (4B8V), CVNH, a sugar-binding protein from *Magnaporte oryzae* (5C8Q), AtCERK1-protein kinase from *Arabidopsis thaliana* 4EBZ), and MSMEG3288 from *Mycobacterium smegmatis* (Dr Bisson and DWR, Sheffield University; personal communication). Comparison of the three-dimensional structures of these LysM domains revealed that the modules possess an identical LysM fold (Figure 4.13). The L1 and L2 loops were all broadly similar. Sequence alignment of the respective LysM modules shows that they share a highly conserved GDTL motif (Figure 4.14 C). In LysM 2 of Ecp6, its groove area of the module is shaped by amino acids ^{20}GDTL^{24} and ^{47}NLIE^{50} (Sánchez *et al*., 2013) located on the L1 and L2 loops, respectively, and this finding is highly similar to Domain 1 of KEG15107 involving residues ^{9}GDTL^{12} and ^{40}LIVS^{43}.

Detailed analysis of the three-dimensional X-ray structures of the LysM domains from various organisms, Ecp6, CVHH, AtCERK1, NlpC, and MSMEG3288, in a complex with NAG_4, NAG_4, NAG_4, NAG_6, and NAG_3, respectively, showed that they shared similar binding pockets for sugar molecules as seen in Domain 1 of KEG15107. The binding pockets, assigned as S1, S2 and S3 were clearly observed on each of the cleft areas of the LysM domains (Figure 4.14 A). Common to each of the binding pockets of these LysM domains is the observation that three sugar molecules occupied Sites S1, S2, and S3. Structure comparison to AtCERK1 (Liu et al., 2012) and CVNH (Koharudin et al., 2015) reveals that these proteins recognize NAG molecules in a similar manner to KEG15107 protein. Residues on the L1 and L2 loops of the proteins make hydrogen bonds to the NAG molecules through their main chain to the hydroxyl and N-acetyl groups of the sugar. The respective residues of these proteins specifically recognize N-acetyl groups of NAG molecules at binding pockets S1 and S3 as the molecules fit well to the pocket cavities and make numerous hydrogen bonds which is also seen in Domain 1 of KEG15107. A similar observation of the N-acetyl recognition

by KEG15107 was also seen in the Ecp6, NlpC and AtlA proteins (Sánchez *et al.*, 2013, Mesnage *et al.*, 2014, Wong *et al.*, 2015). The aromatic ring side chain of F13 of Domain 1 of KEG15107 contributes to the oligosaccharide binding and this aromatic ring is also a key determinant of protein specificity (Boraston et al., 2004). A similar result is seen in MSMEG3288 whilst the other equivalent residues to this F13 in NlpC/P60, AtCERK1, Ecp6, and CVNH are Y28, E114, T24 and R67, respectively (Figure 4.14 A).

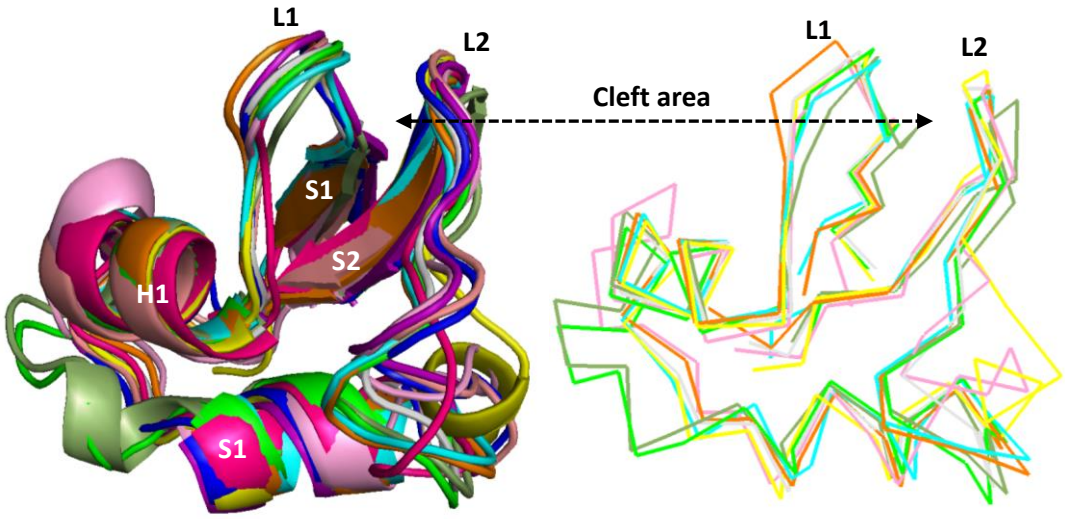
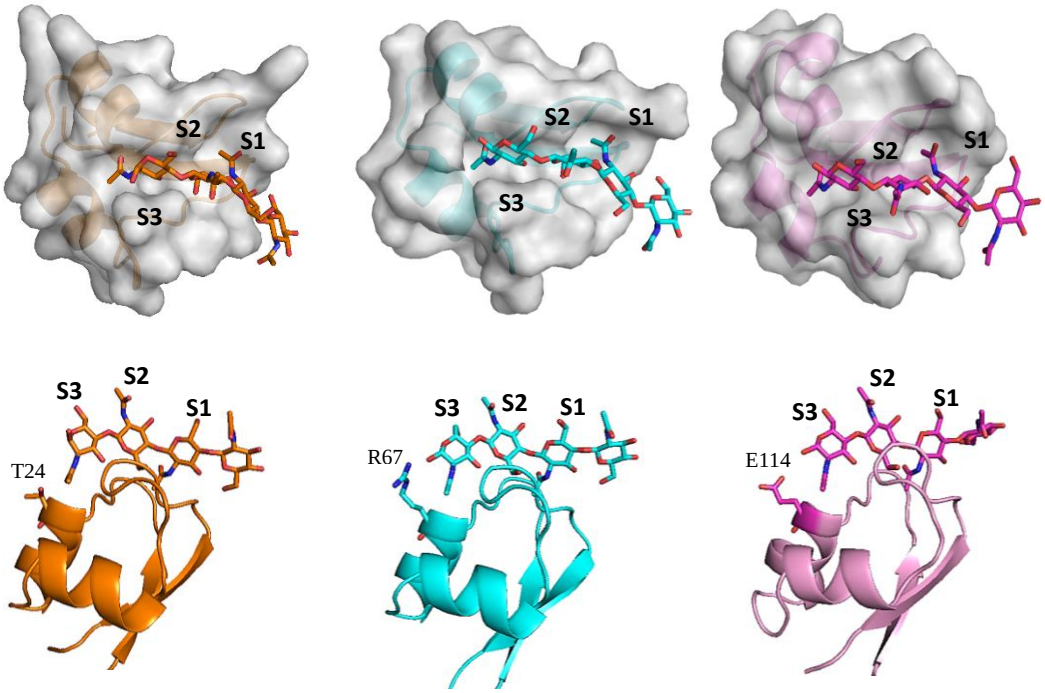


Figure 4.13: Structure alignment of LysM Domain 1 of KEG15107 against other LysM domains from various proteins of species variants. A total of eight structures of LysM domains found in bacterial, fungal and plant proteins were superposed onto each other. Two adjacent loops designated as L1 and L2 which were seen in Domain 1 of KEG15107 were also observed on the other LysM domains. All the aligned LysM domains were obtained from PDB except for KEG15107 and MSMEG3288. Indicator: KEG15107 (LysM1-*M. avium*, green), **NlpC/P60** (LysM1-endopeptidase of *Thermus thermophilus*,**4UZ3,** grey**)**, **SleL** (LysM1-spore cortex of *Bacillus cereus*, **4S3J,** red), **ykuD** (LysM1-peptidase of *B. subtilis*, **4A1I,** yellow), **Ecp6** (LysM1-secreted protein of *Cladosporium fulvum*, **4B8V**, orange), **CVNH** (LysM1sugar-binding protein of *Magnaporte oryzae*, **5C8Q**, cyan), **AtCERK1** (LysM2-protein kinase of *Arabidopsis thaliana*, **4EBZ**, pink), MSMEG3288 (LysM1-*M. smegmatis*, dark green).

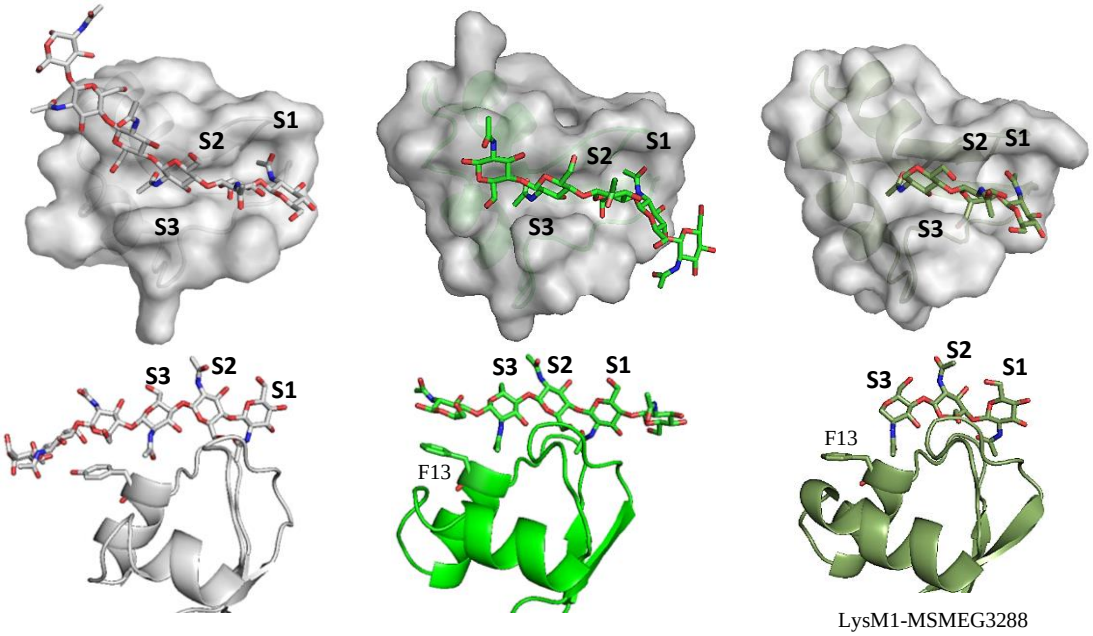
A



LysM1-Ecp6 (pdb:4b8v)
C. fulvum
(Sánchez-Vallet et al., 2013)

LysM-CVNH (pdb:5c8p)
M. oryzae
(Koharudin, Debiec, & Gronenborn, 2015)

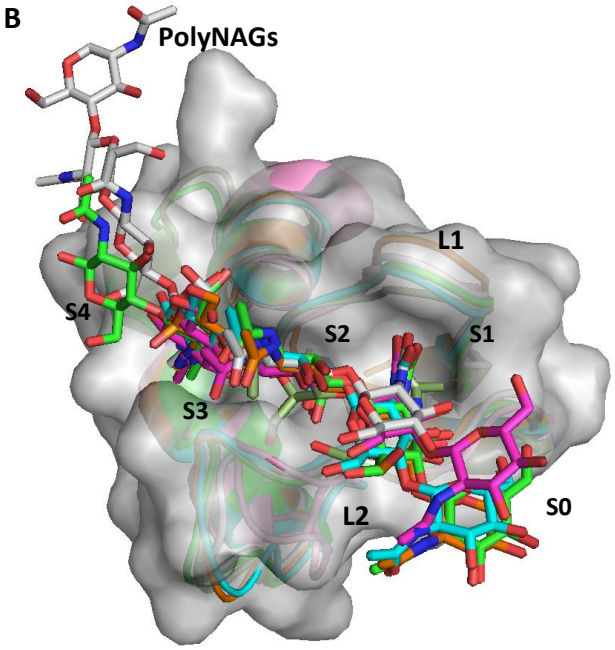
LysM2-AtCERK (pdb:4ebz)
A. thaliana
(Liu et al., 2012)



LysM1-NlpC
T. thermophilus
(Wong et al., 2015)

LysM1-KEG15107
M. avium

LysM1-MSMEG3288
M. smegmatis
(C. Bisson)



LysM domains

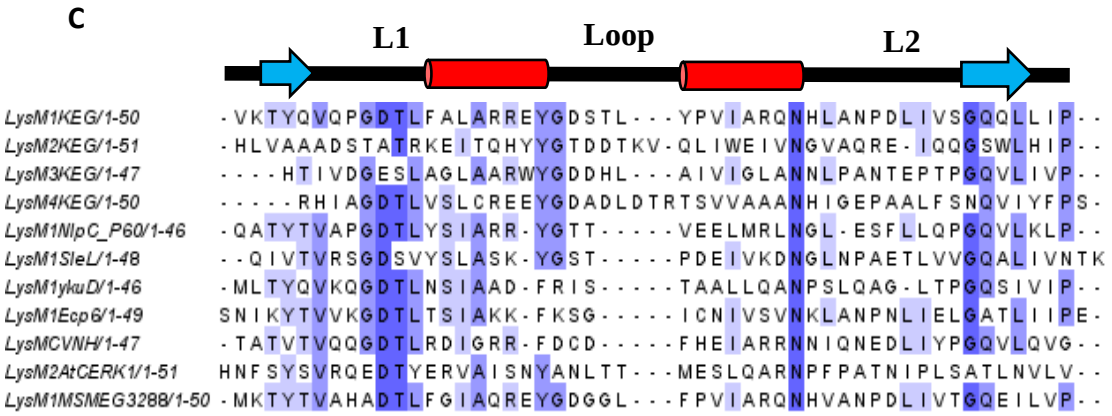


Figure 4.14: Analysis of LysM domains from various proteins of species variants. A) Three-dimensional structure of LysM domains of Ecp6, CVNH, AtCERK1, NlpC, KEG15107 and MSMEG3288 in complex with NAG oligomers. The LysM domains from different proteins of species variants exhibit highly similar binding pockets to NAG molecules and equivalent protein residues at the binding sites which suggested contribute for oligosaccharide binding. B) Structure comparison of LysM domains from various proteins shows the module bind NAG oligomers in a similar manner. Five binding sites designated as S0, S1, S2, S3, and S4 could be identified in the shallow groove of these LysM domains, in which, S1, S2 and S3 are prominent pockets, whereas the S0 and S4 are a minor pocket for a LysM domain. C) Sequence alignment of LysM domains from various proteins.

CHAPTER FIVE

MASS SPECTROMETRY ANALYSIS OF OLIGOSACCHARIDE BINDING TO A LysM DOMAIN

5.0 Introduction

The mass spectrometry analysis was initially performed on KEG15107 to explore the different quaternary structures implied by the X-ray structures of the apo KEG15107 protein and its polyNAG complexes which had revealed a tetrameric and a dimeric structure, respectively. The MS analysis was also initiated as, in the three-dimensional structures of the KEG15107-$NAG_{(n)}$ complexes, the sugar bound only to Domain 1 with no evidence of sugar-binding to Domains 2, 3 or 4, a situation that was exactly the same as had been previously observed for the binding of polyNAG to the KEG15107 homolog, MSMEG3288 (C.Bisson and DWR; personal communication). Given the sequence differences between these two proteins, this observation might have suggested that the other three LysM domains were not functional in the respect of peptidoglycan recognition. This raised a question as to whether the other LysM domains of KEG15107 and equivalent LysM domains in the other proteins are capable of binding to sugar polymers. In addition, if all these domains are sugar-binding modules, do they have an equivalent binding affinity towards the sugar polymer as that of Domain 1 of KEG15107? Subsequently, the analysis was extended to explore the binding of a range of NAG oligomers to three LysM domain-containing protein, Rv1288 and a single LysM domain-containing protein, Ygau, to examine the function of the LysM domain further.

5.1 Analysis of the quaternary structure of KEG15107 in solution by mass spectrometry

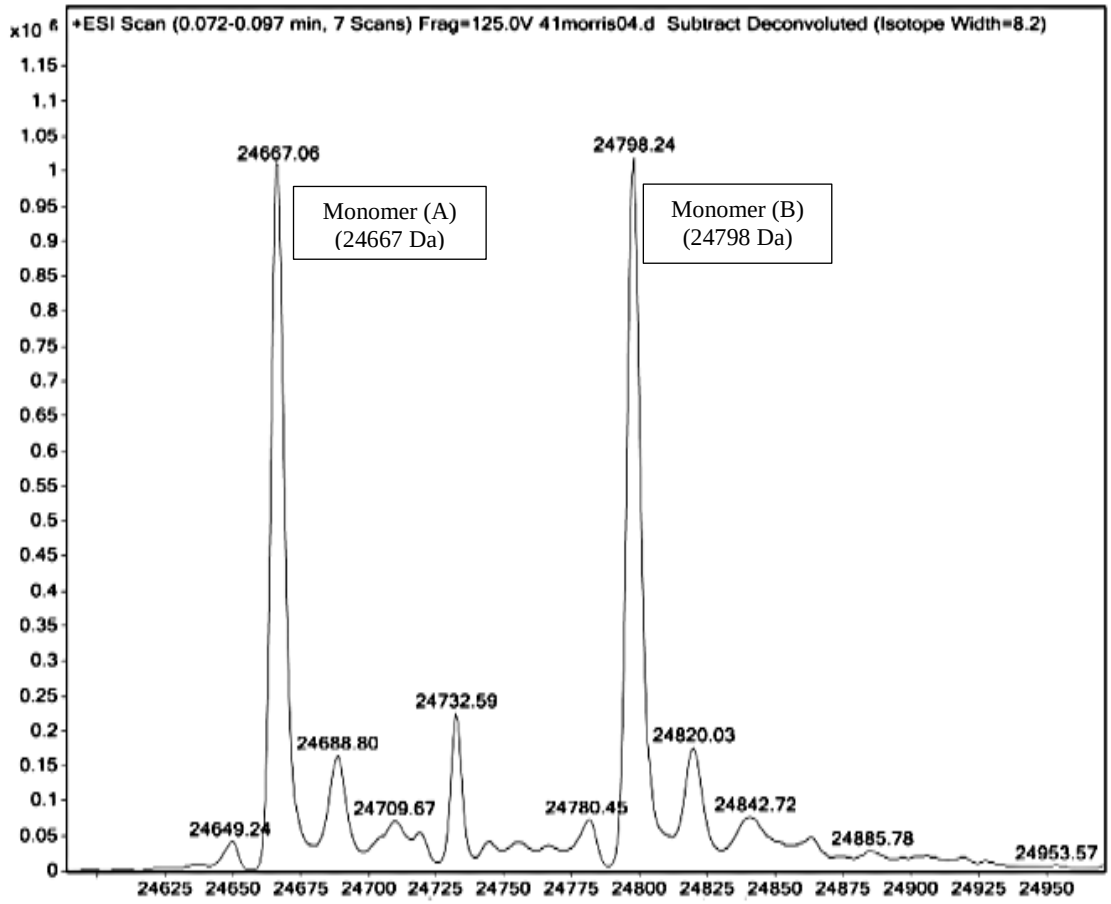
The observation that NAG binding would disrupt the KEG15107 tetramer interface, and that dissociation into a dimer is necessary to permit binding of polyNAG to Domain 1, suggests that multiple quaternary structures of KEG15107 exist. To examine the

quaternary structure of KEG15107 in solution, mass spectrometry (MS) analysis was performed on the purified KEG15107 protein.

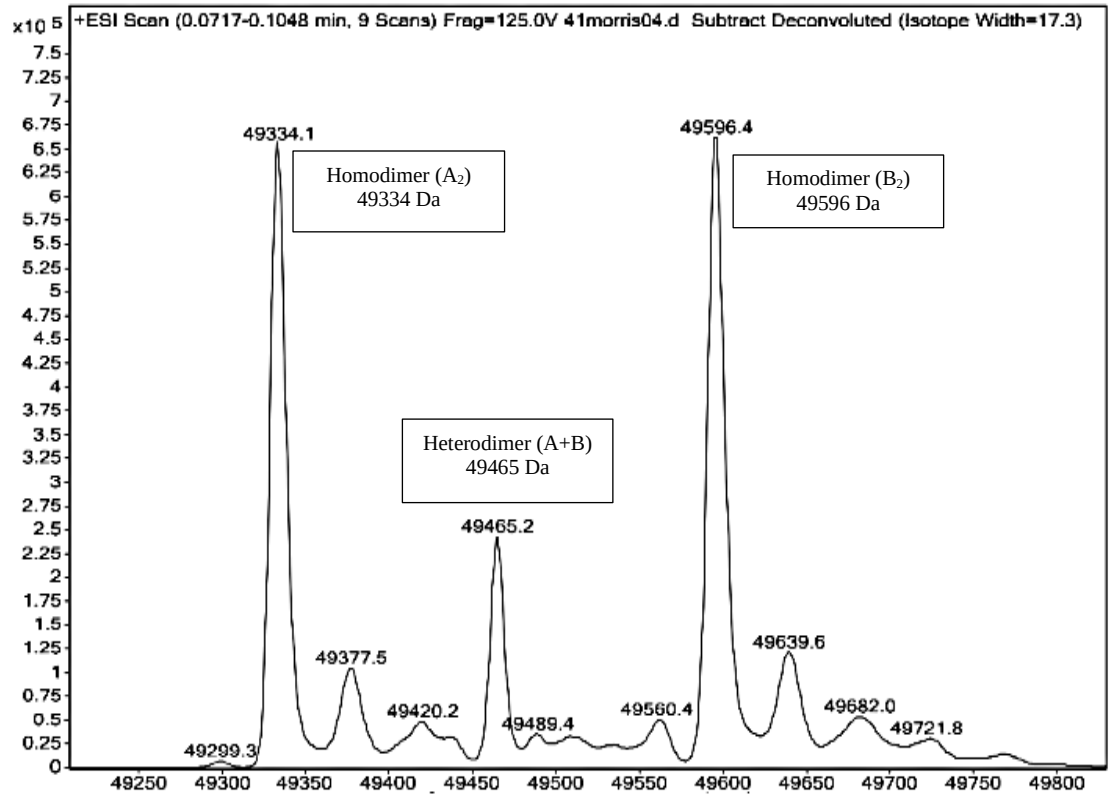
One microliter of a solution of the KEG15107 (~ 0.1 mM) in Tris buffer (pH 8.0) was analyzed by MS in which the sample was directly injected into the MS chamber. Detailed analysis of the deconvoluted spectra showed the protein consisted of two monomeric species, form A (without an N-terminal methionine) and form B (with an N-terminal methionine) in approximately equal quantity with masses of 24667 Da and 24798 Da, respectively (Figure 5.0 A). However, other significant peaks in the mass spectrum with masses 49334 Da, 49596 Da, and 49465 Da indicate the presence of homodimers of forms A and B and a heterodimer of A and B, respectively (Figure 5.0 B). At a much lower level of significance, the presence of homotetrameric and heterotetrameric species of KEG15107 was observed in the mass spectrum with masses 98667 Da (A_4), 99191 Da (B_4) and 98929 Da (A_2+B_2), respectively (Figure 5.0 C). The identity of other very minor peaks in this region of the spectrum is unclear. The presence of sodium ions which were probably carried over from buffers during purification of the sample was detected by the peaks in the spectrum which differ by multiple increments of 22 Da. The mass spectrum did show minor peaks corresponding to the positions of trimeric and pentameric species (Figure 5.0 D). The significance of this is unclear and they may not be biologically relevant.

Considering the results of the X-ray analysis, overall, these results most likely suggest that KEG15107 exists as an equilibrium between monomeric, dimeric and tetrameric species. Comparison of the relative abundance of the peaks suggests that the monomer is the most abundant species although there is a significant proportion of a dimer (~30%) but with very little tetramer (~1%) (Figure 5.0 D). However, the significance that can be attached to the relative amounts of each peak is questionable and may depend on the relative stability of the different complexes in the mass spectrometer.

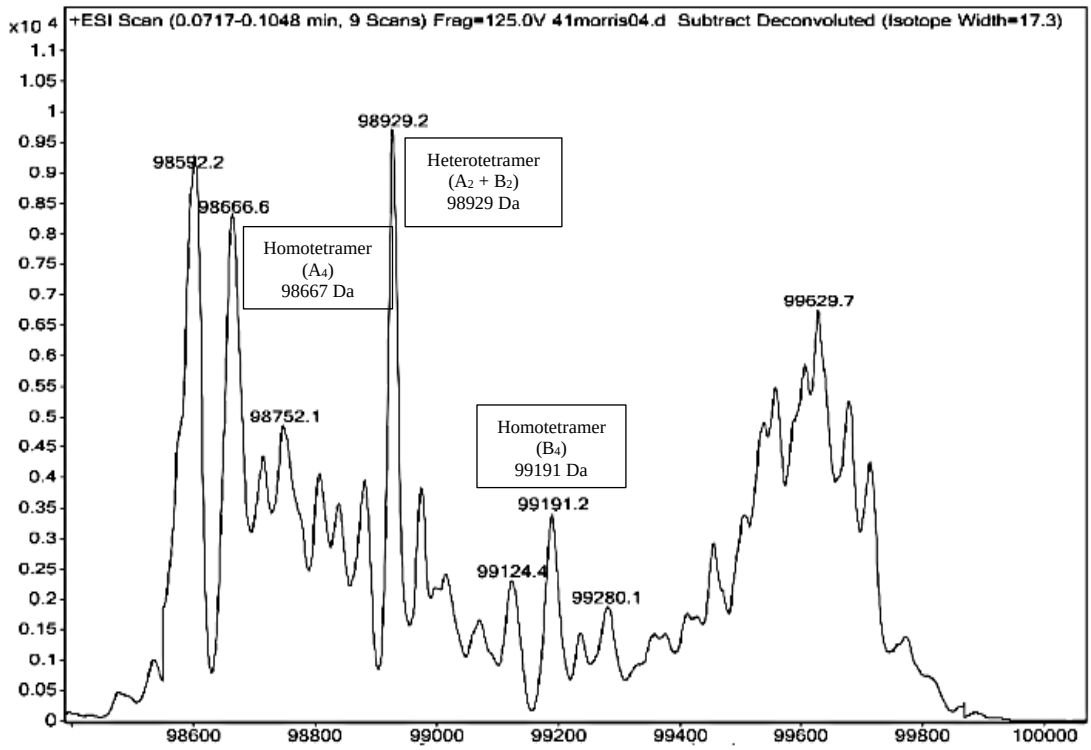
A


+ESI Scan (0.072-0.097 min, 7 Scans) Frag=125.0V 41morris04.d Subtract Deconvoluted (Isotope Width=8.2)
x10 6
24667.06
24798.24
Monomer (A)
(24667 Da)
Monomer (B)
(24798 Da)
24732.59
24688.80
24820.03
24709.67
24780.45
24842.72
24649.24
24885.78
24953.57

B


+ESI Scan (0.0717-0.1048 min, 9 Scans) Frag=125.0V 41morris04.d Subtract Deconvoluted (Isotope Width=17.3)
x10 5
49334.1
49596.4
Homodimer (A2)
49334 Da
Homodimer (B2)
49596 Da
Heterodimer (A+B)
49465 Da
49465.2
49377.5
49639.6
49420.2
49560.4
49682.0
49489.4
49721.8
49299.3

C



D

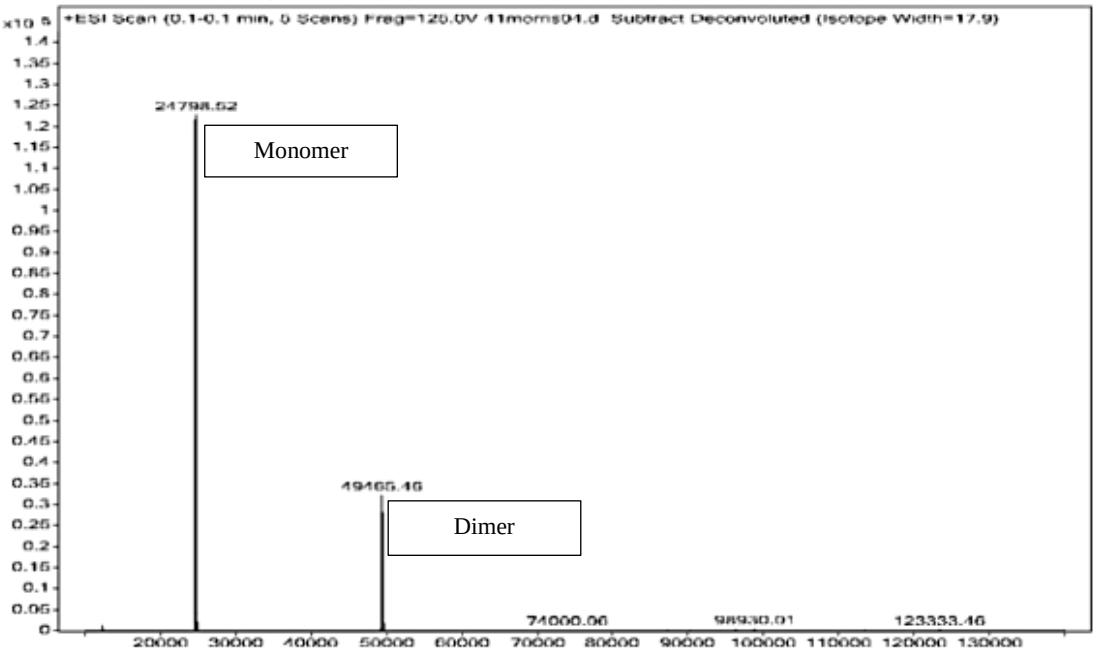


Figure 5.0: The deconvoluted mass spectra profiles of apo KEG15107. A) KEG15107 was detected as monomeric species A and species B with molecular masses 24667 and 24798 Da, respectively, corresponding to the samples without and with the N-terminal methionine. B) KEG15107 was observed as a homodimeric species with molecular masses 49334 (A_2) and 49596 Da (B_2), and as a heterodimer with a molecular mass 49465 Da. C) KEG15107 was detected as a homotetrameric species (A_4) and (B_4) with

molecular masses 98667 Da and 99191 Da, respectively, while a heterotetramer A_2+B_2 (99829 Da) species was also detected at a much lower intensity. Selected peaks in the mass spectra are labeled with their expected molecular masses indicated in the boxes. The presence of sodium ions bound to the protein can be identified by peaks in the mass spectrum with molecular mass differences of multiple of 22 Da. D) The deconvoluted spectrum up to the mass range of 130 kDa.

5.2 Binding analysis of oligosaccharides by LysM domains

All the X-ray structures of KEG15107 in complex with NAG_3, NAG_4, and NAG_5 showed binding only to Domain 1 and, despite co-crystallizing the protein with high concentrations of sugar (1:10 molar ratio), there was no evidence of binding to the other three LysM domains. This situation is similar to that seen in structural studies of the homolog of *KEG15107* from *M. smegmatis*, *MSMEG3288* (C.Bisson and DWR, personal communication). Given the similarity in the structure of all four of the LysM binding domains in KEG15107 and their sequence similarity, as indicated by the structure-based sequence alignment (Chapter 4: Figure 4.18), all the four LysM domains of the protein might have been expected to bind the oligosaccharide. Therefore, to further investigate whether Domains 2, 3 and 4 of KEG15107 bind to oligosaccharides, the KEG15107 protein, in complex with different polyNAG substrates, were analyzed by mass spectrometry.

5.2.1 MS analysis of the KEG15107-NAG_5 at a 1:2 protein to sugar ratio

A sample of KEG15107 complexed with NAG_5 was prepared in 10 mM Tris buffer at 1:2 of protein to sugar ratio to make a final concentration of the protein, 0.1 mM and 0.2 mM for the NAG_5. One microliter of the sample was injected into the MS chamber for analysis. The m/z values from the mass spectrum of the sample detected a peak for polyNAG_5 (1034 Da) (Figure 5.1). The deconvoluted mass spectrum of this complex showed the presence of a monomer, a dimer and a tetramer species of KEG15107 (Figure 5.2).

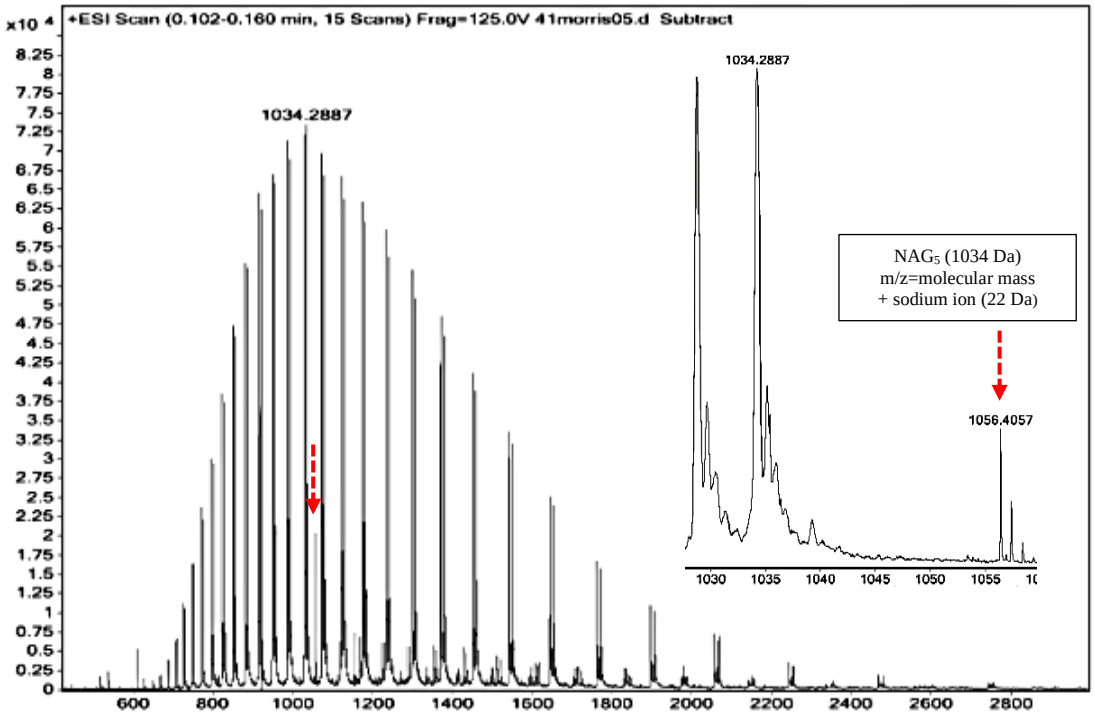


Figure 5.1: The peak in the array of multiple ions of the mass spectrum indicated by a red arrow with a m/z value of 1056 indicates the presence of a NAG_5 species (1034 Da) bound to a sodium ion (22 Da). This peak appears surrounded by peaks corresponding to the protein.

Peaks for the monomeric species A and B of the KEG15107 protein were identified at molecular masses of 24667 Da and 24798 Da, respectively (Figure 5.2 A). Further analysis of the mass spectrum revealed the existence of peaks providing strong evidence for oligosaccharide binding corresponding to A+1NAG_5 (25701 Da) and B+1NAG_5 (25832 Da), and at a lower level of significance, two oligosaccharides bound to species A and species B with masses 26735 Da, and 26866 Da, respectively. The protein was also detected in a homodimeric form at 49334 Da (species A) and 49596 Da (species B), and with two oligosaccharide chains bound corresponding to A_2+2NAG_5 (51403 Da) and B_2+2NAG_5 (51664 Da) (Figure 5.2 B).

Figure 5.2 (A)

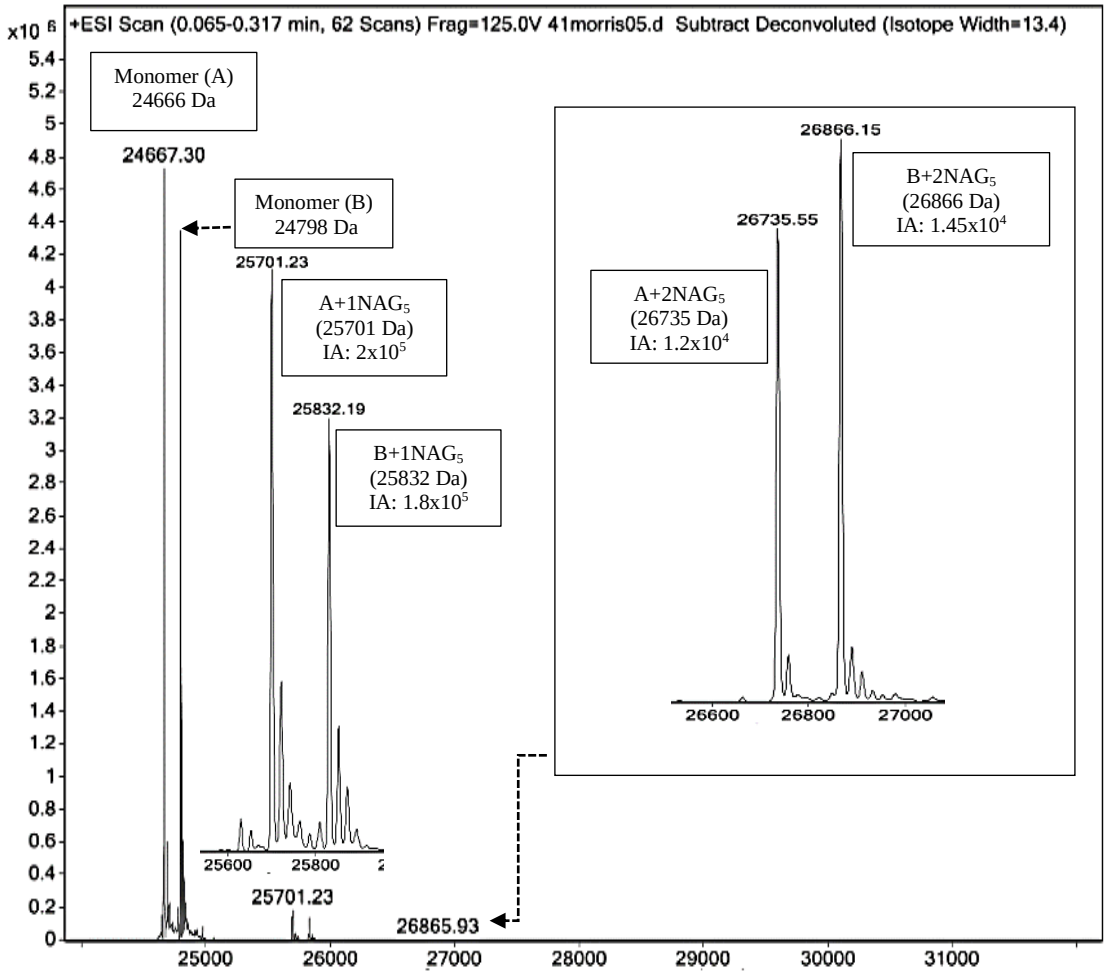
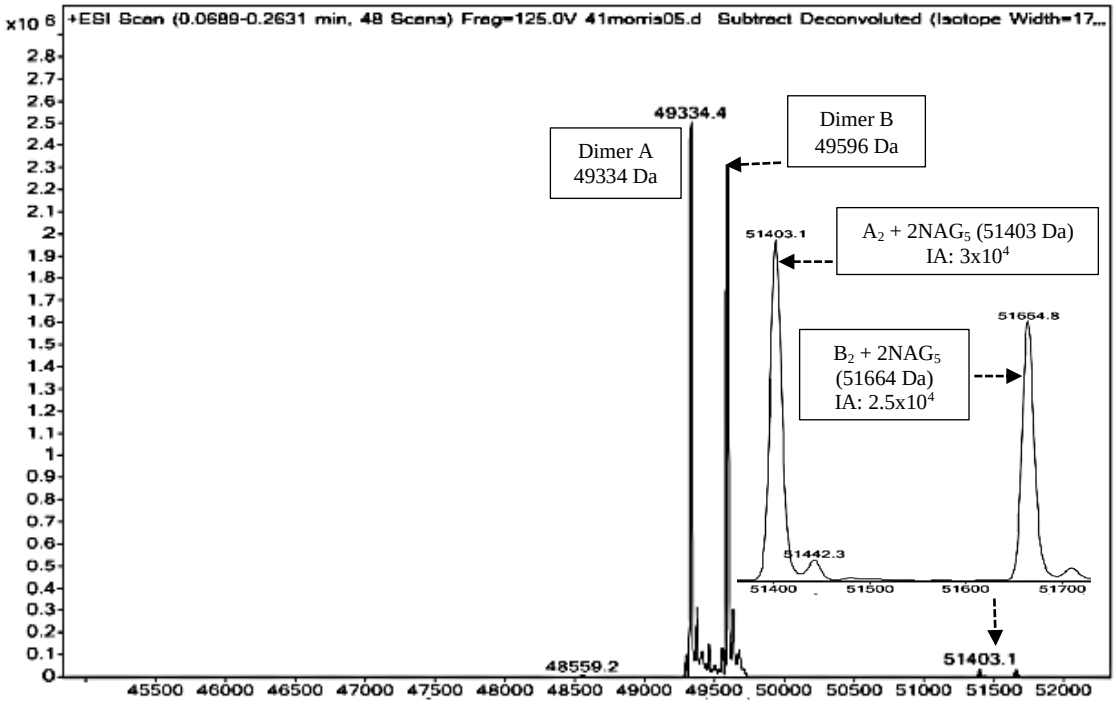


Figure 5.2 (B)



Although the data are much noisier, the presence of peaks corresponding of two homotetrameric KEG15107 for species A and B with four oligosaccharides bound was detected in the mass spectrum corresponding to the molecular mass of A_4+4NAG_5 (102805 Da) and B_4+4NAG_5 (103327 Da) (Figure 5.2 C). However, the other minor species in this region of the mass spectrum could not be interpreted but it is worth noting that impurities in the sample of NAG_5 might also interfere with the nature of the molecules that are bound to the protein. Therefore, since NAG_3 and NAG_4 have a higher purity than NAG_5, these oligosaccharides were subsequently used to simplify the MS analysis (Chapter 3: Figure 3.52). Analysis of the deconvolution of the full mass spectrum suggested that the proportion of the dimer was ~5% (Figure 5.2 D) lower than that seen for the apo protein (~30% (Figure 5.0 D)). Minor peaks for a trimer and a pentamer were also present but the biological function of these is unclear.

Figure 5.2 (C)

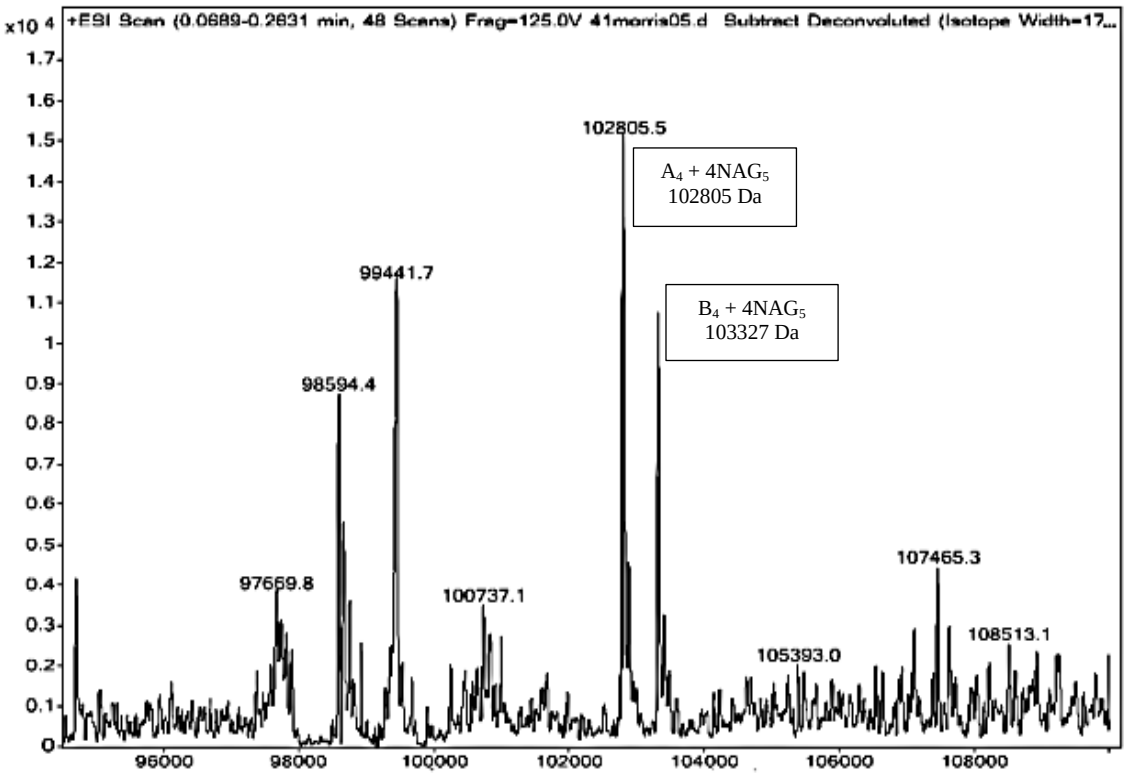


Figure 5.2 (D)

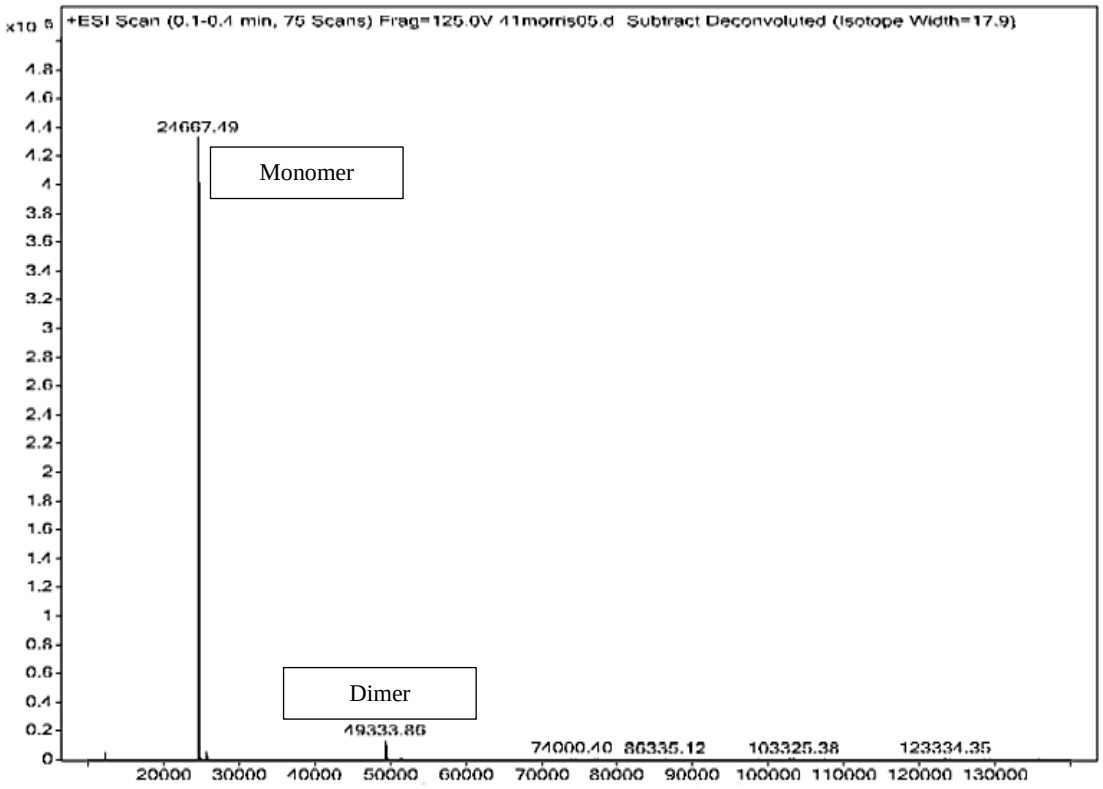


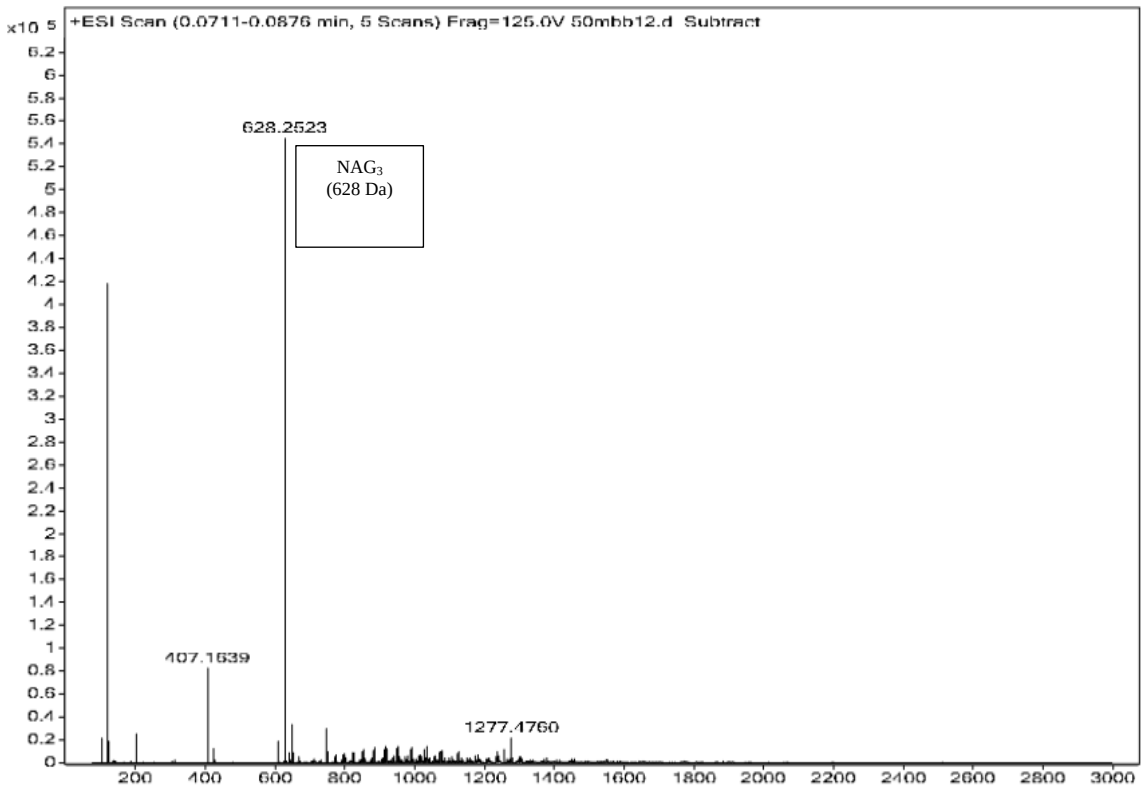
Figure 5.2: Mass spectrometry analysis of the KEG15107-NAG_5 complex at 1:2 of a protein to sugar ratio. A) The monomer region of the spectrum. The spectrum showed the presence of KEG15107 in a monomeric form of A (24667 Da) and B (24798 Da), each bound to either one or two NAG_5 molecules. The expected masses of specific complexes are shown in boxes. B) The dimer region of the spectrum. The presence of the dimeric species of KEG15107, A, and B, each with two molecules of NAG_5 bound, (51403 Da) and (51664 Da), respectively, were indicated by the peaks at the expected masses. C) The tetramer region of the spectrum. The mass spectrum showed the presence of the homotetrameric form of the KEG15107, A_4, and B_4 bound to four molecules of NAG_5 with masses of 102805 Da and 103327 Da, respectively. This region includes a high background, and the significance of the peaks is lower. The ion abundance (IA) for selected peaks is indicated. D) The deconvoluted spectrum up to the mass range of 130 kDa.

5.2.2 MS analysis of the KEG15107-NAG_3 at a 1:200 protein to sugar ratio

To further investigate oligosaccharide binding by the LysM domains, MS analysis was carried out on the KEG15107 protein in a complex with NAG_3 at much higher concentration of the sugar using a 1:200 ratio of protein to sugar, in the hope of observing binding to all four LysM domains of KEG15107 with oligosaccharides.

The sample that was injected to the mass spectrometry contained 0.1 mM protein and 20 mM of NAG_3 in 10 mM Tris buffer. The presence of NAG_3 in the analyzed sample was detected in the mass spectrum with a mass of 628 Da (Figure 5.3 A). Detailed analysis of the deconvoluted spectrum provided clear evidence that led to the identification of the monomeric and the dimeric species of KEG15107 with the latter being ~8% of the former along with other minor peaks corresponding to a trimer, a tetramer and a pentamer could be also identified (Figure 5.3 E).

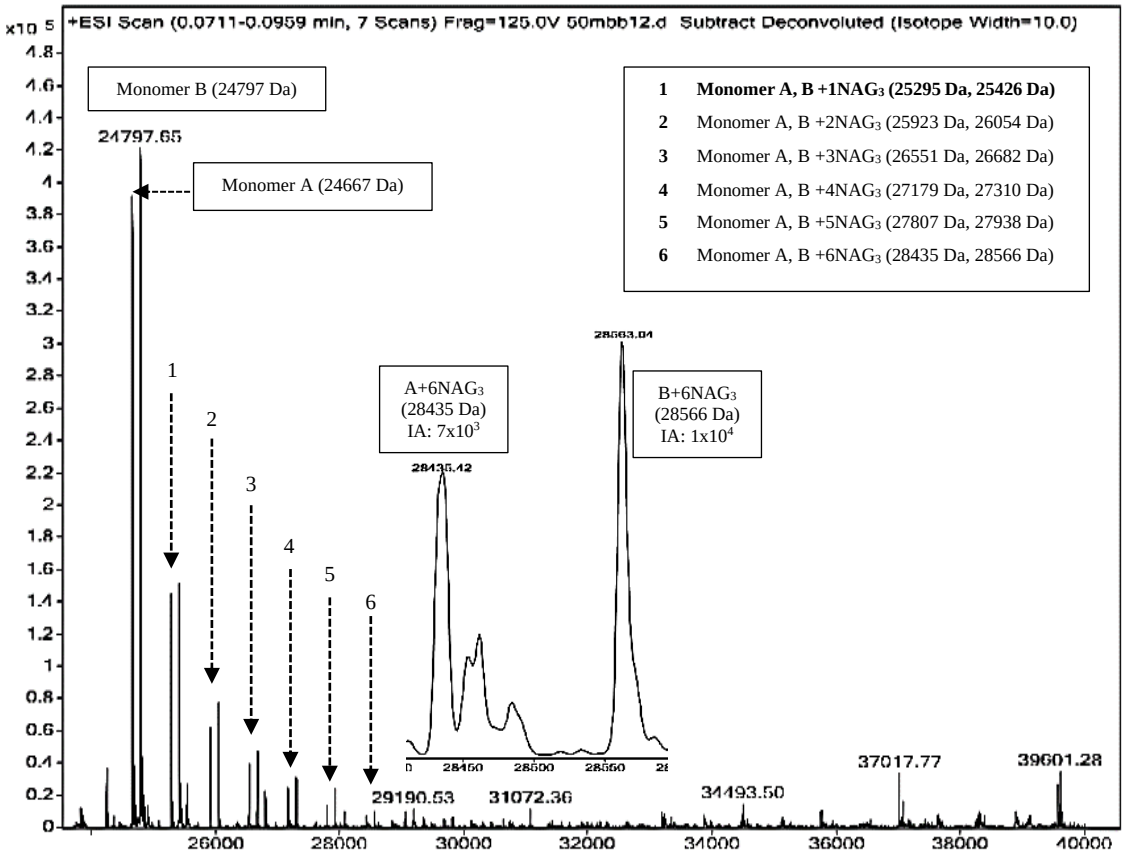
Figure 5.3 (A)



Peaks in the monomer region of the mass spectrum, corresponding to the binding of one, two, three or four NAG_3 molecules per monomer to each sequence variant of KEG15107 (+/- the N-terminal Met) with masses (25295 and 25426 Da), (25923 and 26054 Da), (26551 and 26682 Da), and (27179 and 27310 Da), respectively, could clearly be seen (Figure 5.3 B). In addition, other peaks in the monomer region, whilst being at a lower level is consistent with the binding of five and six NAG_3 molecules.

One interpretation of this is that each of the binding grooves can be occupied by more than one sugar chain.

Figure 5.3 (B)



Analysis of the mass spectrum in the region of the dimer showed binding of two or four or six NAG_3 molecules (Figure 5.3 C). It was noticeable that the peaks corresponding to the binding six NAG_3 molecules were much lower than the peaks corresponding to the binding of two or four NAG_3 molecules. However, despite these peaks being smaller, they were detected with the expected masses (A_2+6NAG_3 (53102 Da)) and (B_2+6NAG_3 (53360 Da)). In the tetramer region, the significance of the peaks is unclear given the low level of these peaks compared to the noise in the mass spectrum (Figure 5.3 D). In contrast, the presence of the trimeric species of monomer A with three NAG_3 molecules can be clearly identified (75886 Da) albeit at a low level of significance.

Figure 5.3 (C)

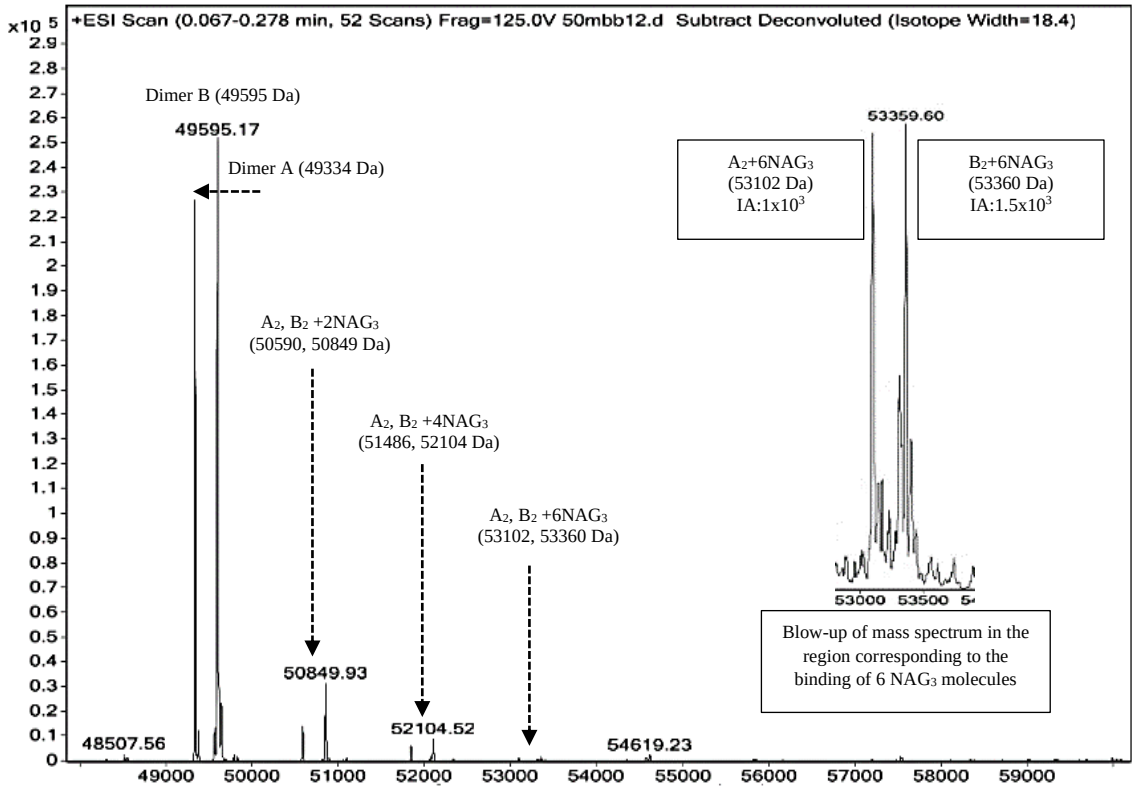


Figure 5.3 (D)

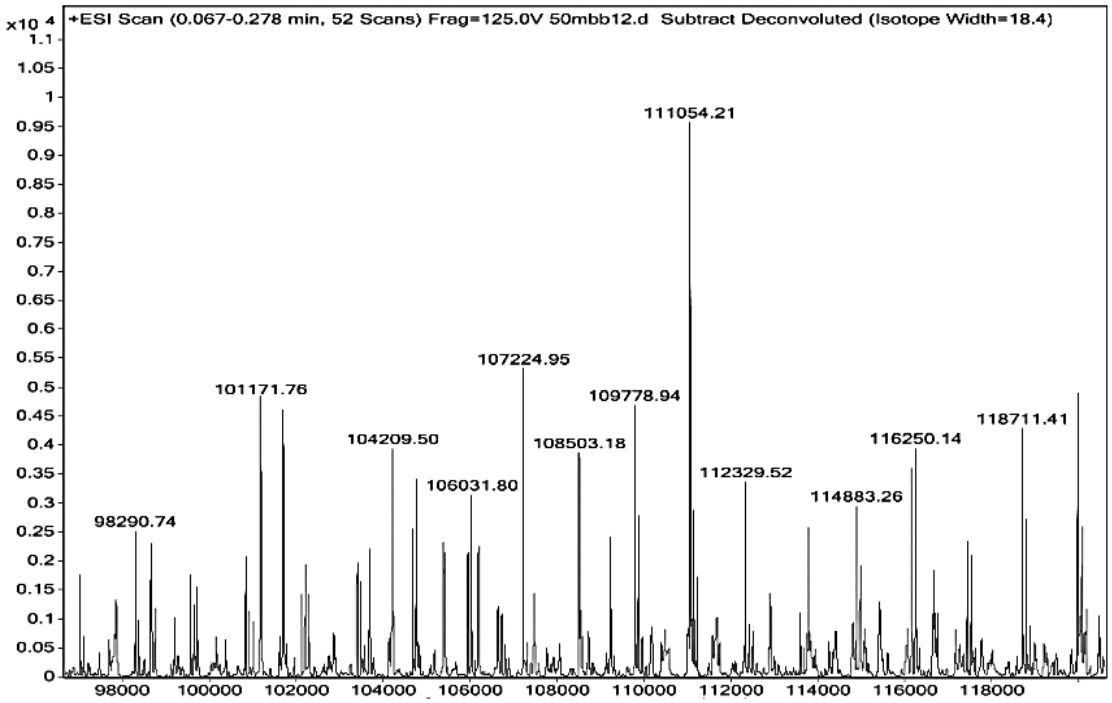


Figure 5.3 (E)

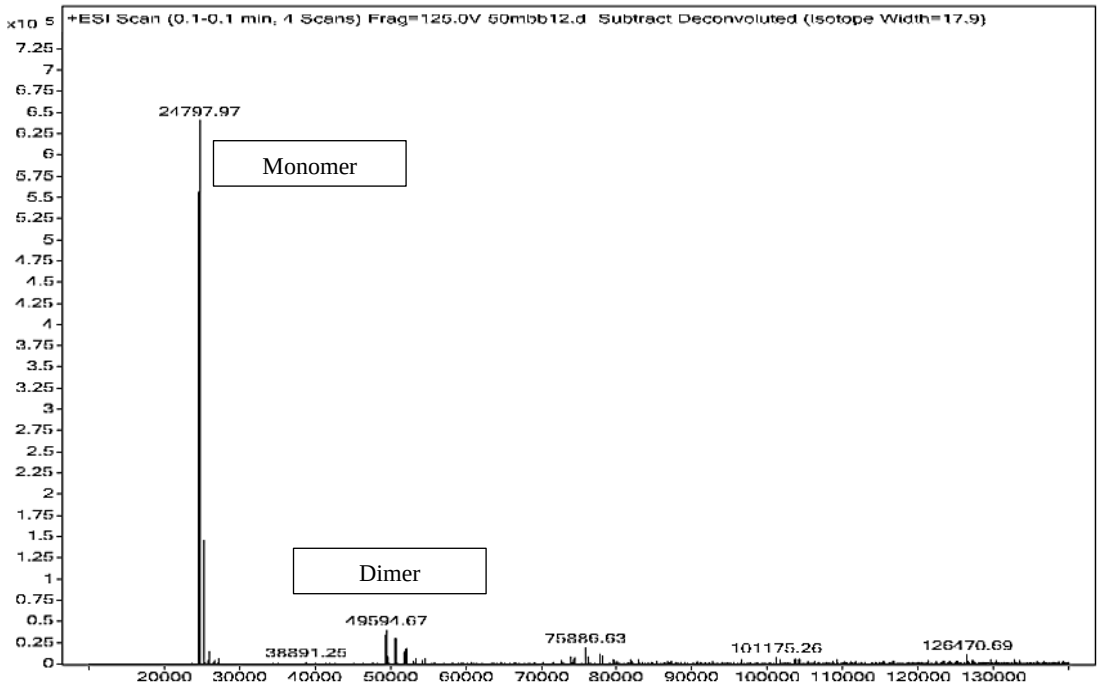


Figure 5.3 (F)

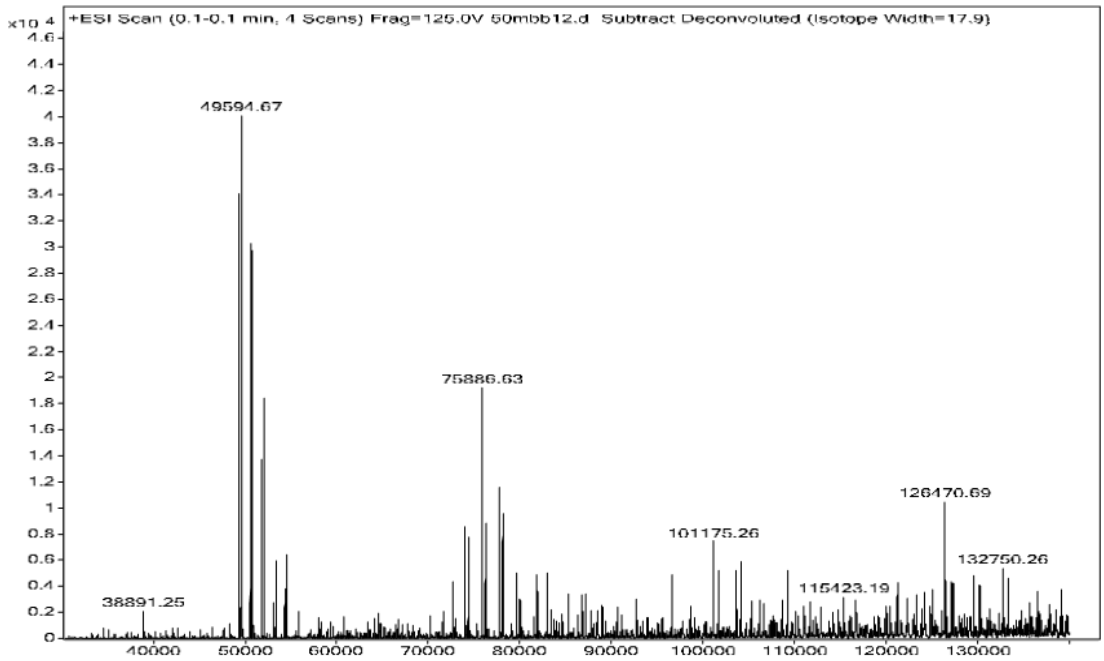


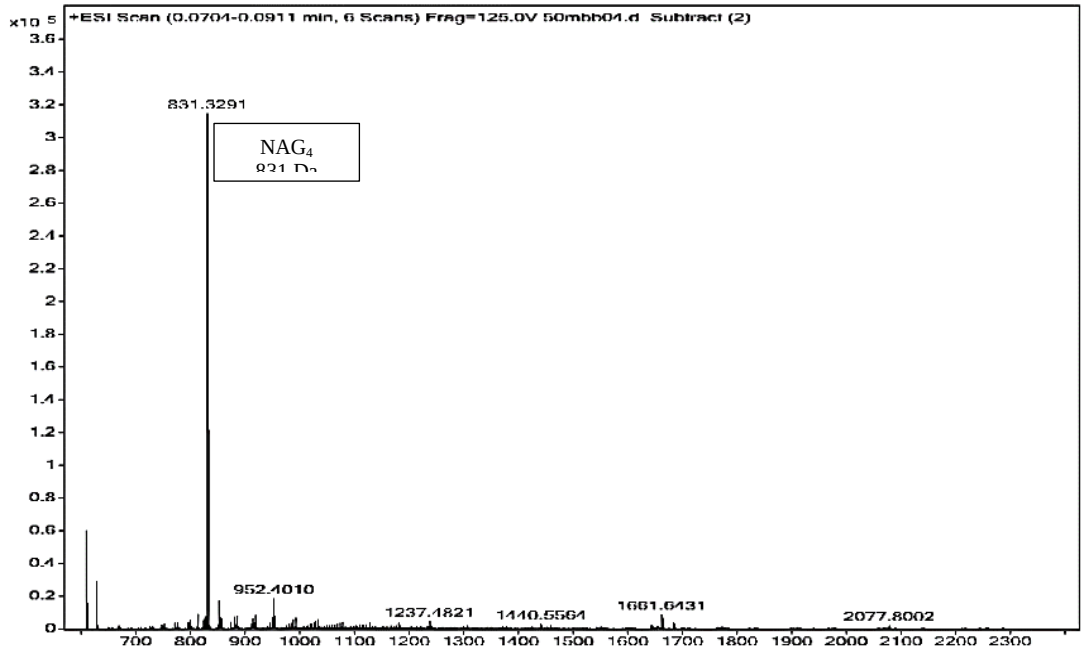
Figure 5.3: Mass spectrometry analysis of the KEG15107-NAG_3 complexes at a 1:200 protein to sugar ratio. A) The array of multiple ions in the raw experimental data shows a peak with molecular mass of 628 Da corresponding to a NAG_3 molecule. B) The monomer region of the deconvoluted mass spectrum shows the presence of KEG15107 in a monomeric form of A (24667 Da) and B (24798 Da), each bound to either one, two, three, four, five or six NAG_3 molecules. The expected masses of specific complexes are shown in boxes. C) The dimer region of the spectrum. The presence of the dimeric species of KEG15107, A and B, each bound to either two or four or six

NAG_3 molecules, are indicated by the peaks at the expected masses. The expected masses of selected complexes are shown above the peaks highlighted with arrows. The ion abundance (IA) for selected peaks is indicated. D) There were no significant peaks in the mass spectrum in the region corresponding to a tetramer of KEG15107. E) The deconvoluted spectrum up to the mass range of 130 kDa. F) The deconvoluted mass spectrum from 40 to 130kDa.

5.2.3 MS analysis of the KEG15107-NAG_4 at a 1:200 protein to sugar ratio

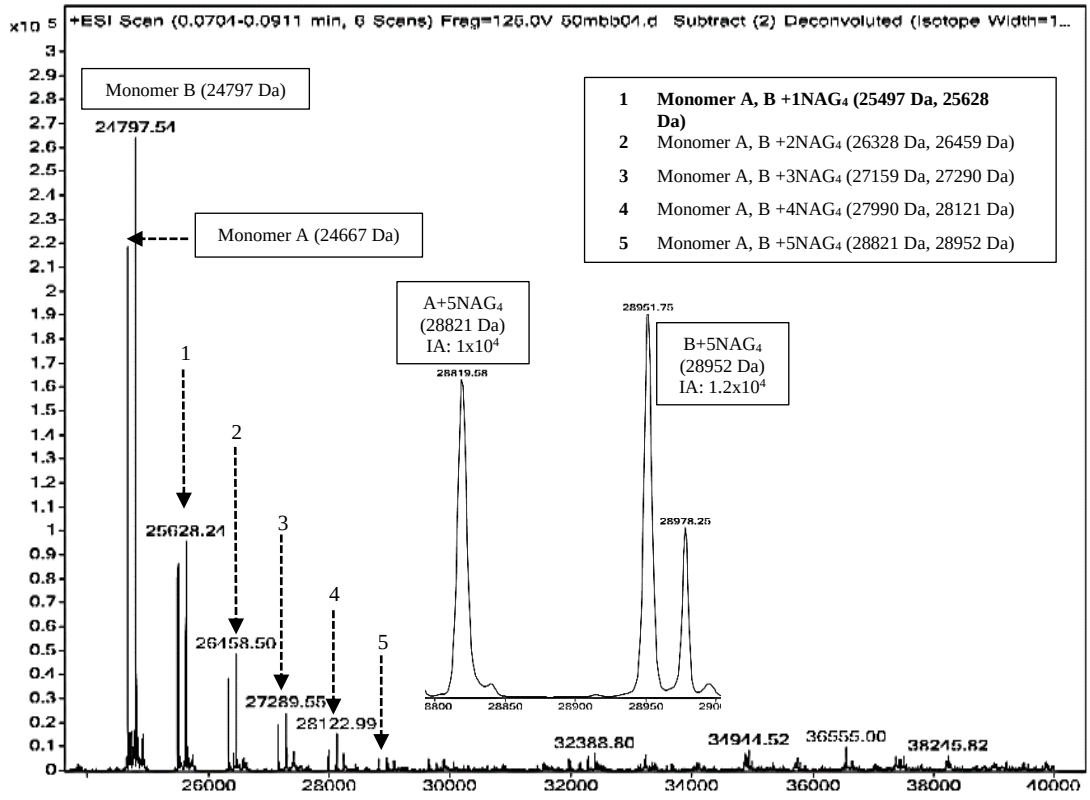
One microliter of a complex of the KEG15107 protein (0.1 mM) with NAG_4 (20 mM) at 1:200 protein to sugar ratio, prepared in 10 mM Tris buffer, was injected into the mass spectrometer. The presence of a NAG_4 molecule was detected in the raw mass spectra at the expected molecular mass of 831 Da (Figure 5.4 A). The presence of a strong monomeric peak for forms A and B of KEG15107 in the sample of the KEG15107-NAG_4 complex could be identified by peaks of 24666 Da and 24797 Da (Figure 5.4 E) with other peaks indicating of sugar binding (Figure 5.4 B). At the higher sugar concentration, the proportion of the dimer and the other species was much reduced probably less than one percent abundance (for example peaks for dimer at 49332 Da and 49594 Da for species A and B, respectively (Figure 5.4 C). The proportion of protein in the trimeric, tetrameric and pentameric forms is significantly reduced (Figure 5.4 D).

Figure 5.4 (A)



The oligosaccharide binding pattern observed in the KEG15107-NAG_4 and KEG15107-NAG_3 at a higher protein to sugar ratio was similar. Peaks corresponding to the binding of either one, two, three, four or five molecules of NAG_4 were observed for forms A and B of KEG15107 monomer at the expected masses (25497 and 25628 Da), (26328 and 26459 Da), (27159 and 27290 Da), (27990 and 28121 Da), (28821 Da, 28952 Da) and (29652 Da, 29783 Da), respectively (Figure 5.4 B). The consistent findings that up to five or six oligosaccharides bind to the monomer of KEG15107 at high concentration of NAG_3 or NAG_4, respectively, indicates that all four LysM domains are capable of binding one or two oligosaccharides.

Figure 5.4 (B)



Other peaks corresponding to the binding of two, four or five NAG_4 molecules bound to the dimer of forms A and B of KEG15107 with masses (50994 and 51257 Da), (52657 and 52917 Da), and (54317 and 54580 Da), respectively, were identified in the mass spectrum (Figure 5.4 C). As previously observed in the analysis of the complexes with NAG_3, the peaks corresponding to the binding of six NAG_4 molecules were detected at lower ion abundance. Again, no peaks corresponding to a tetramer were detected in the mass spectrum (Figure 5.4 D).

Figure 5.4 (C)

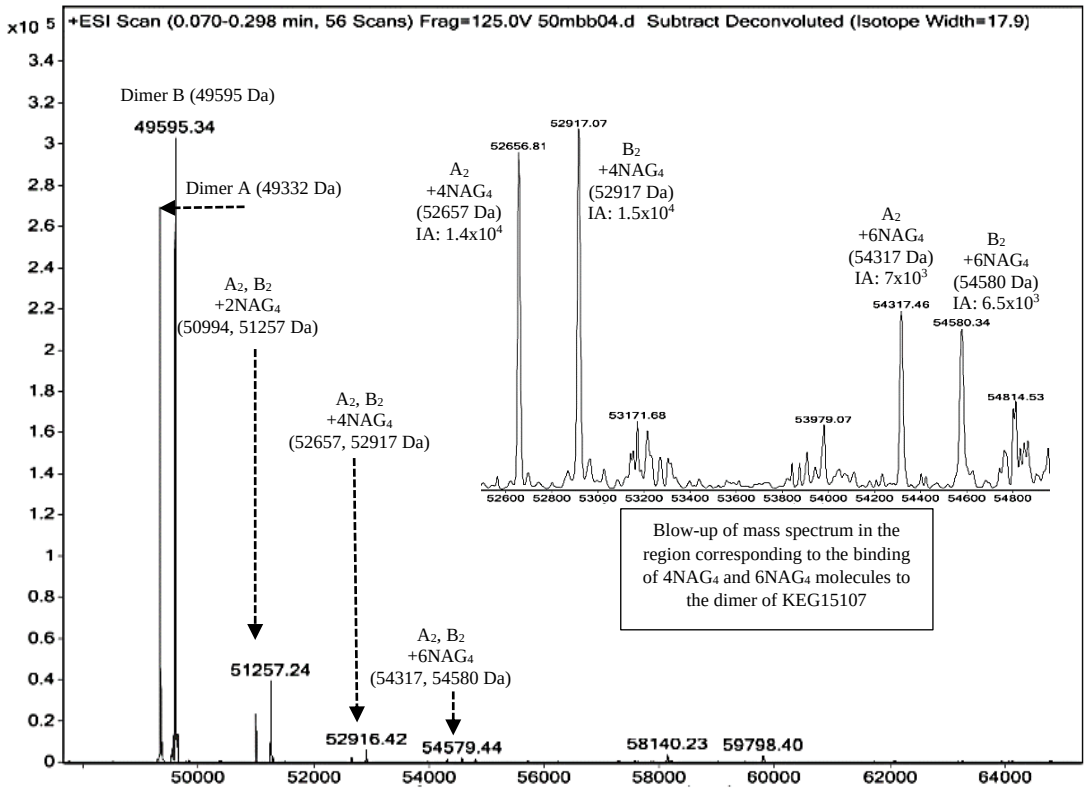


Figure 5.4 (D)

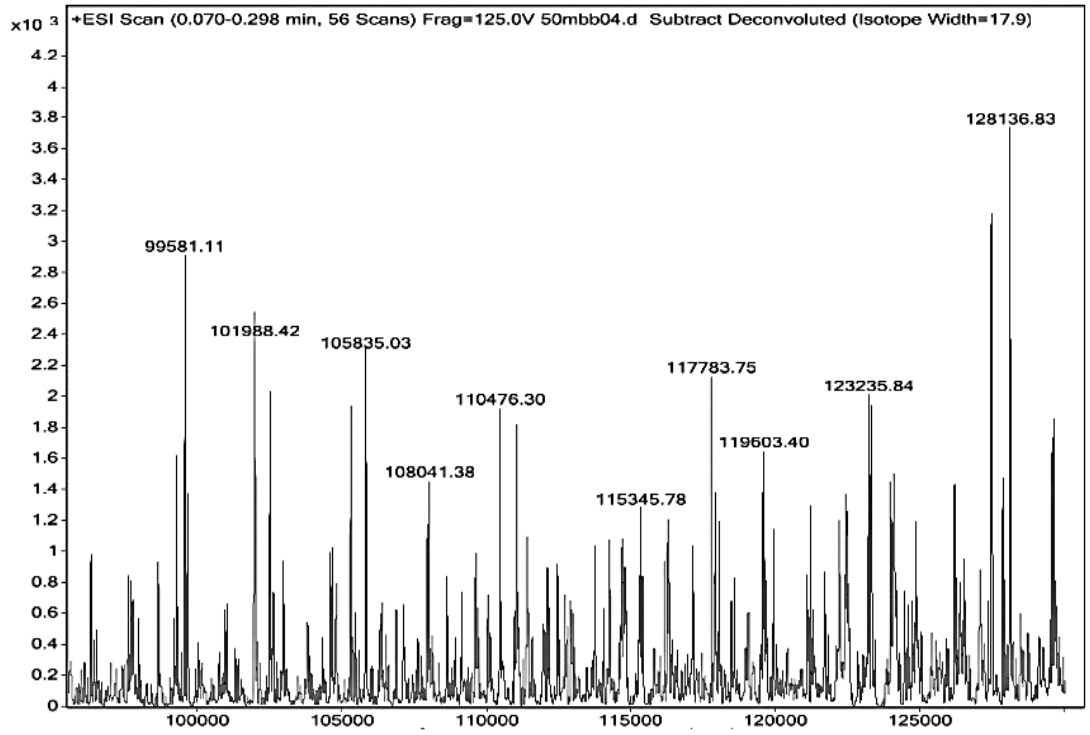


Figure 5.4 (E)

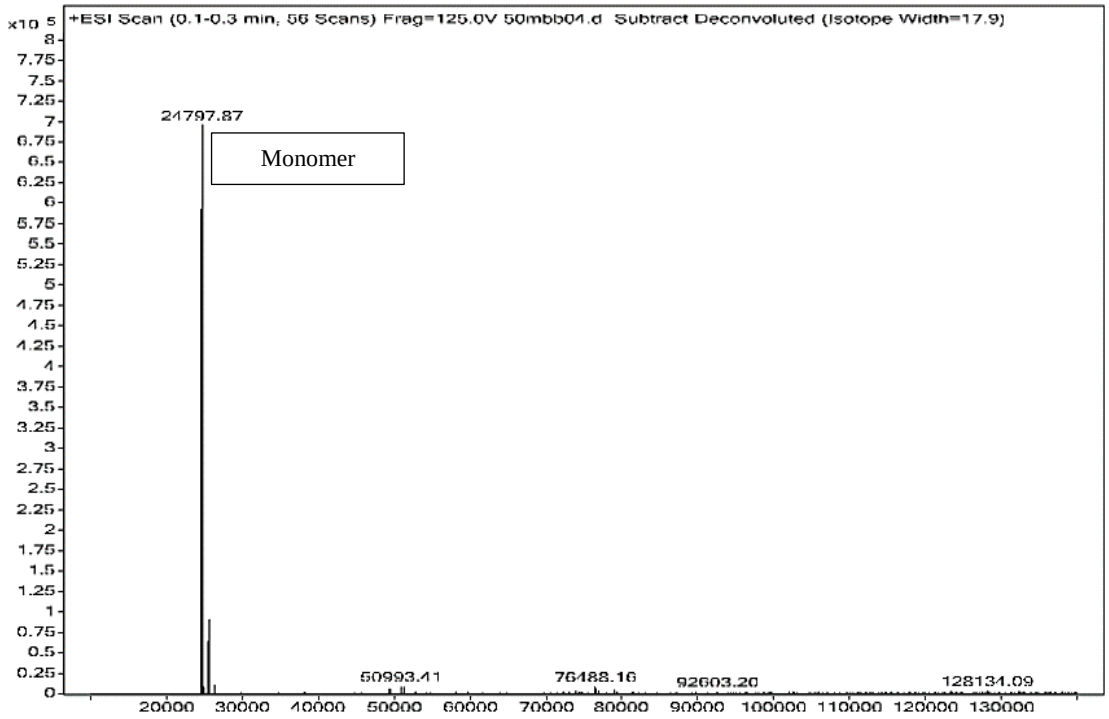


Figure 5.4: Mass spectrometry analysis on the KEG15107-NAG_4 complex at a 1:200 protein to sugar ratio. A) The array of multiple ions in raw experimental data shows a peak with a molecular mass of 831 Da, indicating the presence of NAG_4 in the analyzed sample. B) The monomer region of the spectrum. The spectrum shows the presence of KEG15107 in a monomeric form of A (24667 Da) and B (24798 Da), each bound to either one or two or three or four or five NAG_4 molecules. The expected masses of specific complexes are shown in boxes. C) The dimer region of the spectrum. The presence of the dimeric species of KEG15107, A and B, each bound to either two or four or six oligosaccharides, are indicated by the peaks at the expected masses. The expected masses of the specific complexes are shown in the boxes. Ion abundance (IA) is indicated for the selected peaks. D) The peaks in the mass spectrum in the region corresponding to a tetramer of KEG15107 were at a very low level of significance. E) The deconvoluted spectrum up to the mass range of 130 kDa.

5.3 Mass spectrometry analysis of oligosaccharide binding to LysM domains from other proteins.

5.3.1 MS analysis on Trc1

Mass spectrometry analysis of oligosaccharide binding to the Trc1 protein which contains three LysM domains was subsequently performed. Trc1 is a truncated protein derived from Rv1288 from *M. tuberculosis* where the full-length protein contains three

LysM domains in its C-terminal region and a hypothetical esterase catalytic domain in the N-terminal region of the protein. One microliter of a complex of the Trc1 protein (0.1 mM) with NAG_4 (20 mM) at 1:200 of a protein to sugar ratio, prepared in 10 mM Tris buffer was analyzed.

The raw mass spectrum data showed the presence of the NAG_4 molecule with a molecular mass of 831 Da (Figure 5.5 A). Detailed MS analysis of the deconvoluted mass spectra provided clear evidence that led to the identification of a monomeric species of Trc1 (17328 Da) to which no NAG_4 molecules had been bound (Figure 5.5 B).

Figure 5.5 (A)

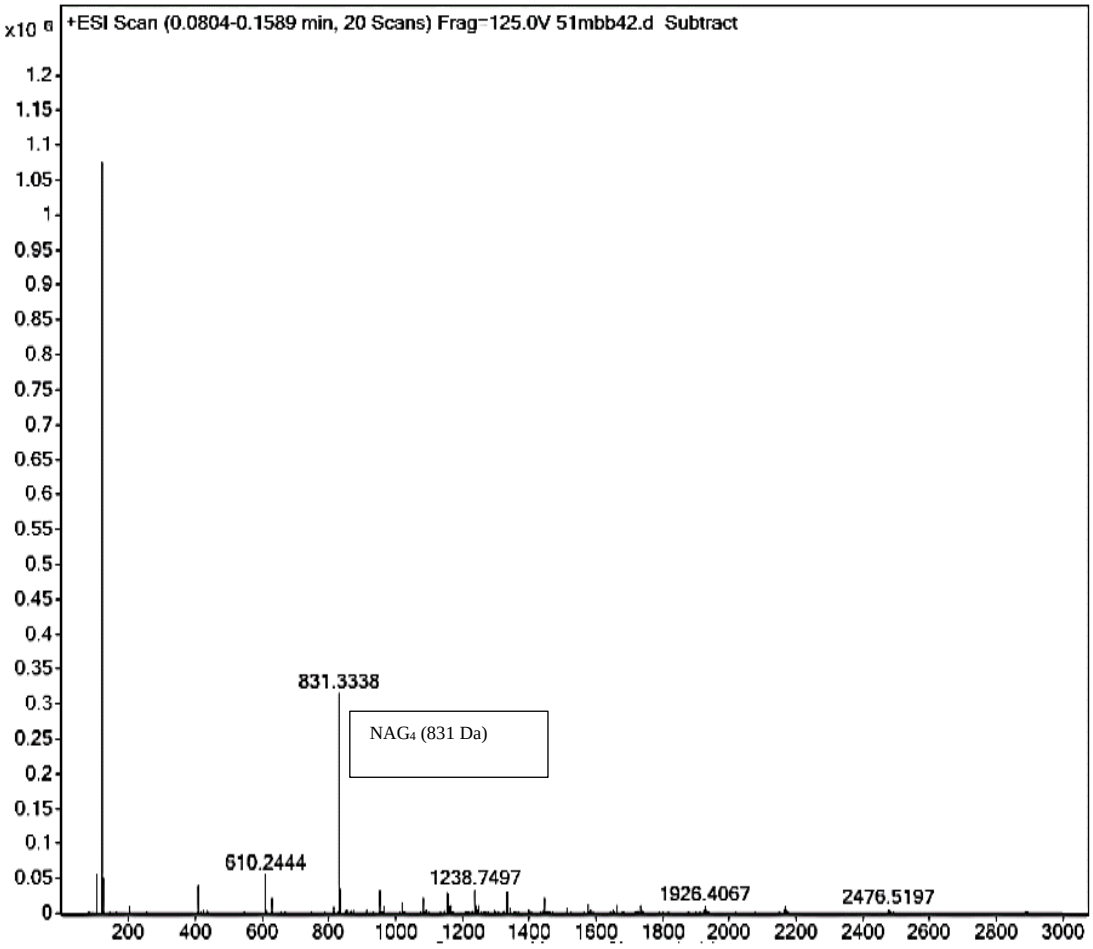
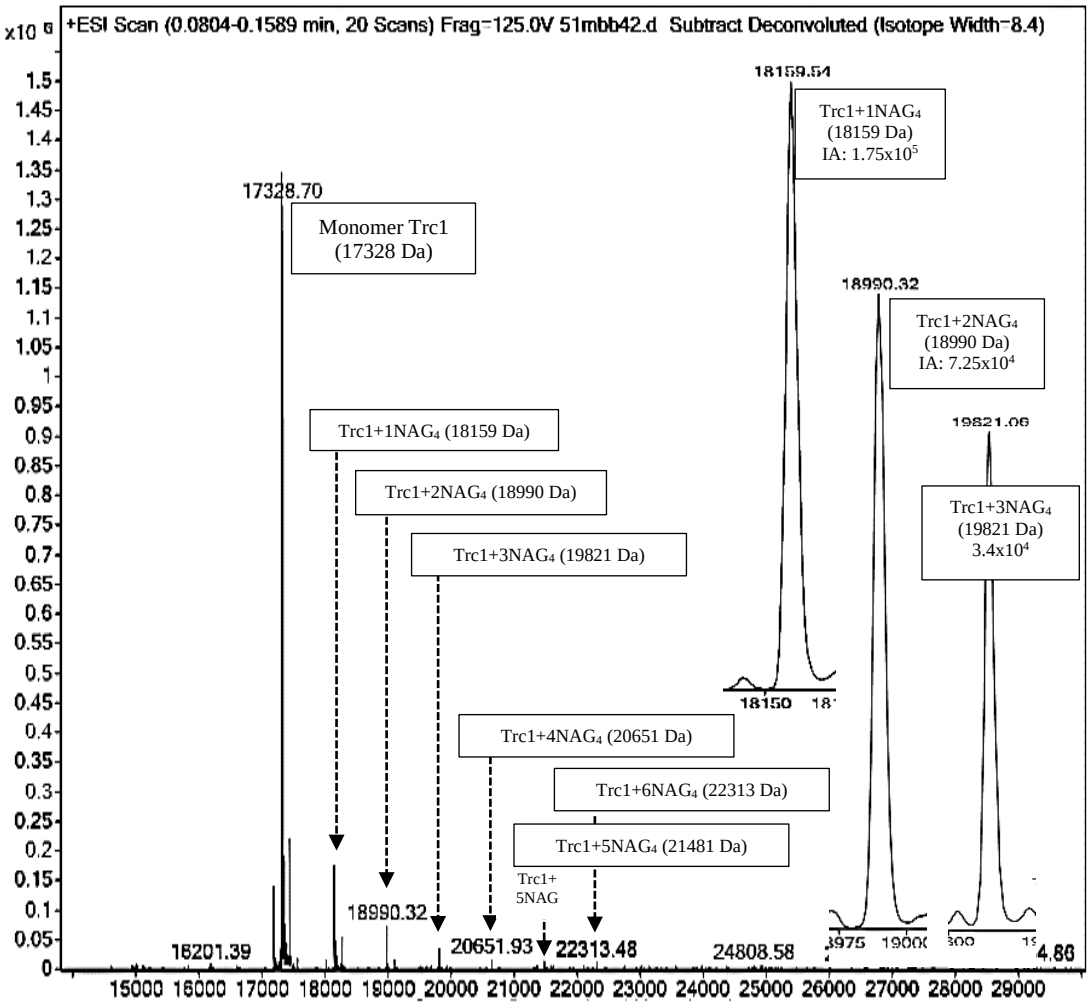


Figure 5.5 (B)



Other peaks in the mass spectrum led to the identification of Trc1 with one, two or three NAG_4 molecules bound to the monomer. The selected peaks of the complexes were detected at the expected molecular masses which corresponded to Trc1+1NAG_4 (18159 Da), Trc1+2NAG_4 (18990 Da) and Trc1+3NAG_4 (19821 Da), respectively. The observation of three NAG_4 molecules per monomer of Trc1 by the MS analysis suggests that all the domains can bind oligosaccharides as seen in KEG15107. Minor peaks in the mass spectrum appear to correspond to the binding of four (20651 Da), five (21481 Da) and six (22313 Da) molecules of NAG_4. In the region of the mass spectrum corresponding to possible dimers, trimers or tetramers of Trc1, no peaks could be seen indicating a monomeric quaternary structure which is consistent with the gel filtration (Figure 5.5 C).

Figure 5.5 (C)

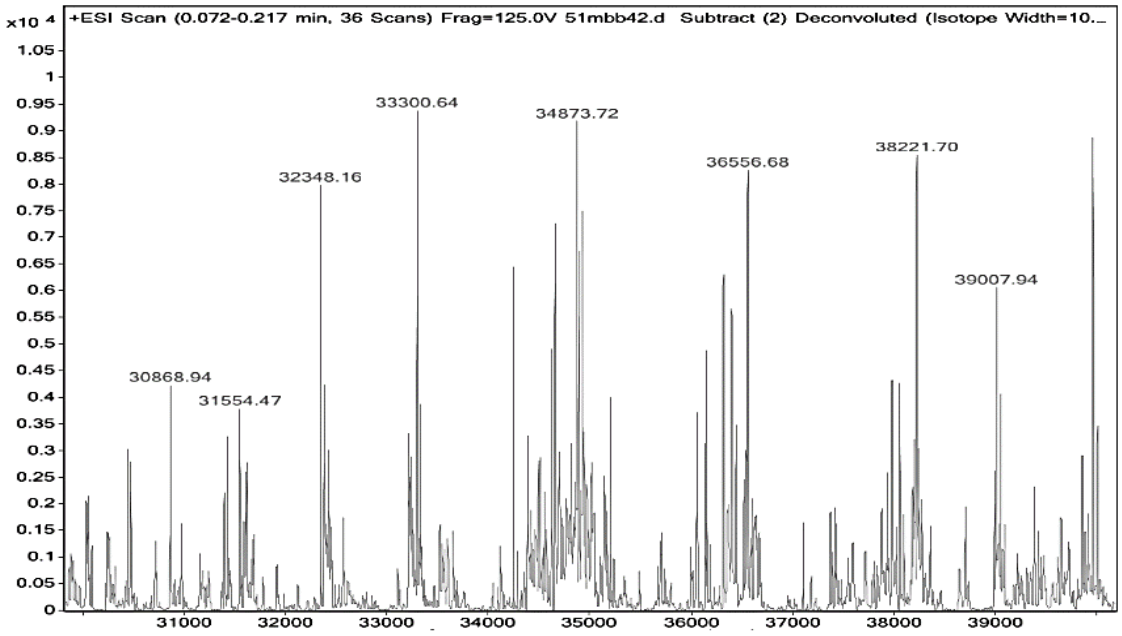


Figure 5.5: Mass spectrometry analysis on the Trc1-NAG_4 complex at a 1:200 protein to sugar ratio. A) The array of multiple ions in raw experimental data shows a peak with a molecular mass of 831 Da, indicating the presence of NAG_4 in the analyzed sample. B) Trc1 was detected as a monomer with a molecular mass 17328 Da. Analysis of the monomer region of the mass spectrum showed clear peaks for one, two and three NAG_4 molecules and smaller peaks for four, five and six NAG_4 molecules bound to Trc1 with expected molecular masses shown in the boxes with the arrows above the selected peaks. The IA for the complexes was indicated for the selected peaks in the blow-up mass spectrum. C) There were no significant peaks in the mass spectrum in the regions corresponding to a dimer or a trimer of Trc1.

5.3.2 MS analysis on YgaU

B. pseudomallei YgaU, a protein-containing a single LysM domain, was included in the study to further investigate the oligosaccharide binding by LysM domains. A stock solution of YgaU at a final concentration ~10 mg/mL in 10 mM of Tris buffer (pH8.0), possibly containing a low level of sodium chloride from the purification was provided by Dr. Sedelnikova (Sheffield University). For the MS analysis, the protein and sugar samples were prepared in two buffers containing either potassium chloride or sodium chloride at different concentrations (50, 100 and 200 mM) either without or with NAG_5 at a final ratio 1:200 of protein to sugar (final concentration of the YgaU protein (0.1

mM) and final concentration of NAG_5 (20 mM). Despite having a lower purity, NAG_5 was chosen for the analysis as a comparison to the KEG15107-NAG_5 complex as it was suspected that affinity of NAG_5 might be higher than that of the shorter oligosaccharides (Chapter 5; Section 5.5). Since the YgaU protein had been suggested to be responsible as a potassium sensor in *E. coli* (Ashraf et al., 2016), therefore, to investigate the effect of this on oligosaccharide binding analysis, the YgaU-NAG_5 complexes were analyzed in the buffers containing potassium and sodium.

The presence of NAG_5 (1034 Da) in a sample of YgaU in a buffer with 50 mM sodium chloride was detected in the mass spectrum as an additional peak compared to the protein alone (16322 Da) corresponding to the binding of NAG_5 (17356 Da) (Figure 5.6 A). In general, whether in the presence or absence of NAG_5, in either buffer containing sodium or potassium ions, the YgaU protein was observed principally as a monomer with the mass spectrum only showing a very low level of the dimer (approximately 1%) (Figure 5.6 B (i-iv)). Moreover, there were no significant peaks for any other quaternary structures of YgaU in the mass spectrum and at this stage, it is not clear whether the dimer peak is of any biological significance.

Figure 5.6 (A)

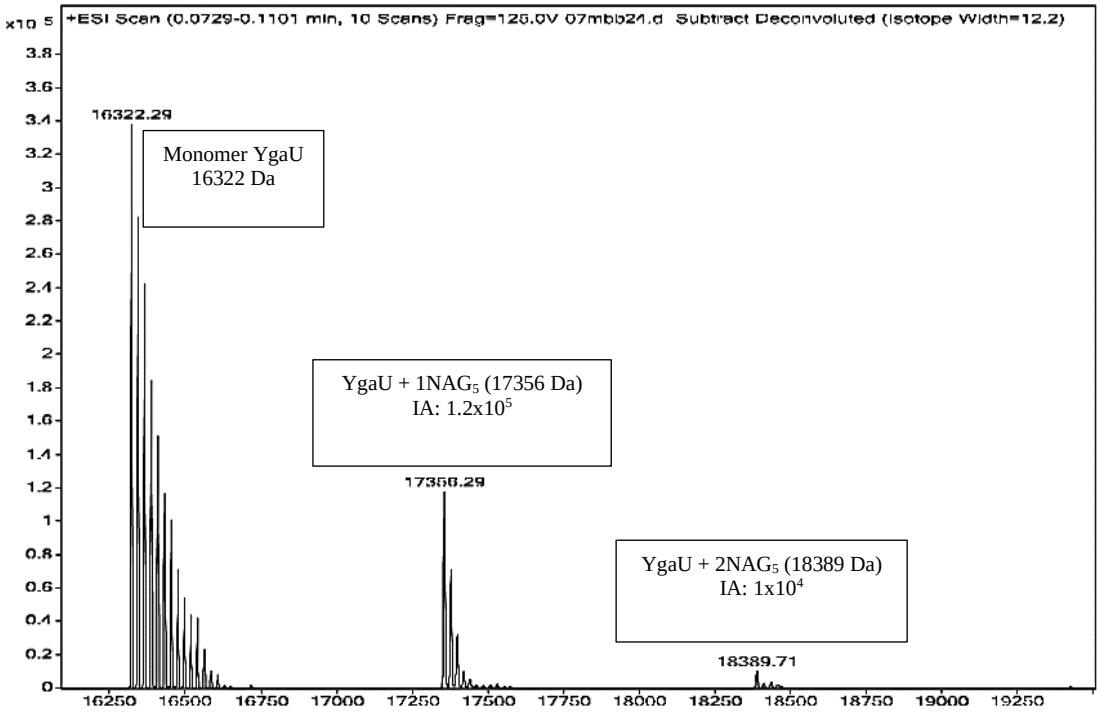


Figure 5.6 (B-i)

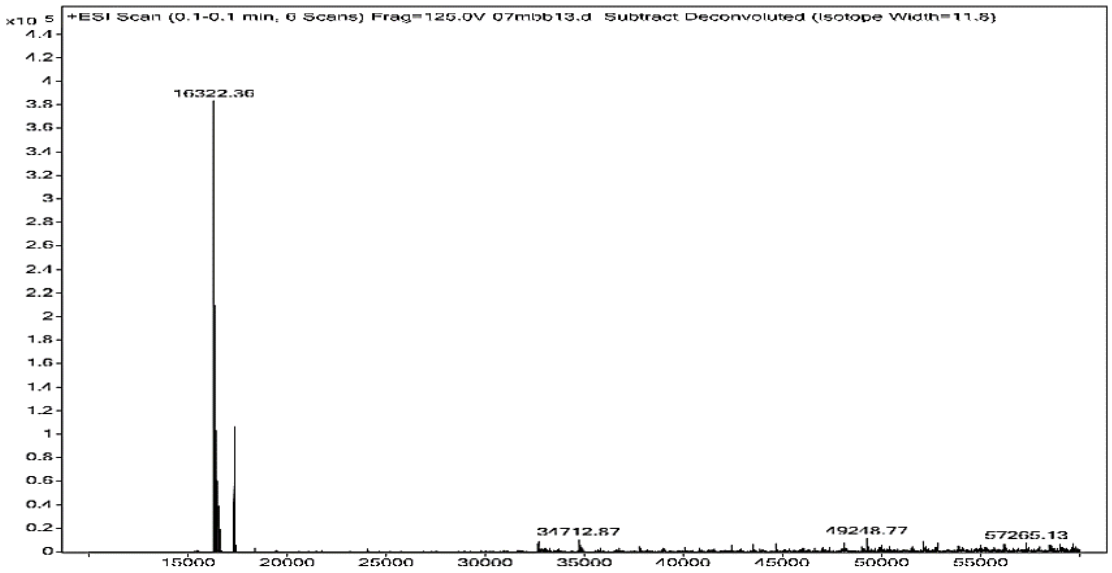


Figure 5.6 (B-ii)

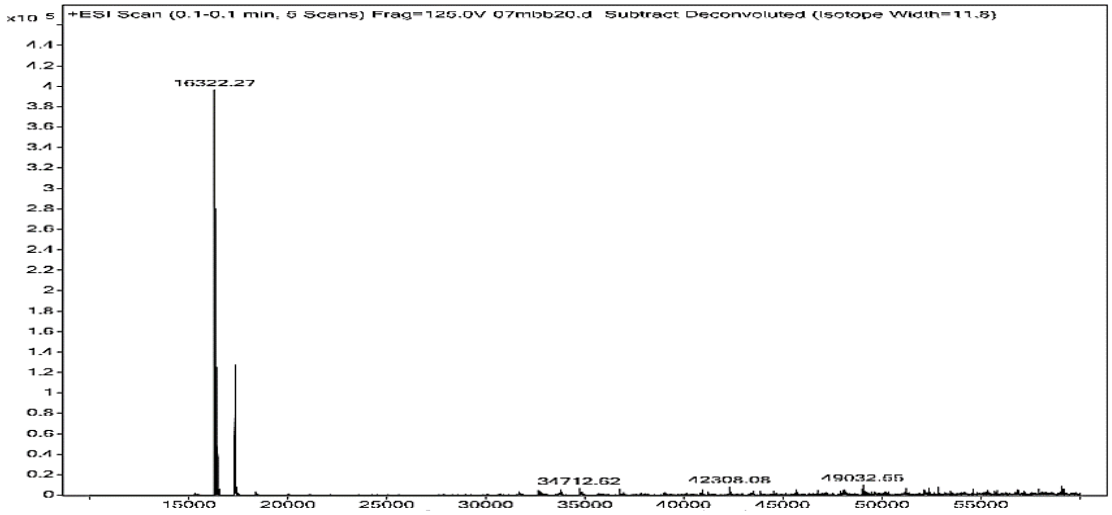


Figure 5.6 (B-iii)

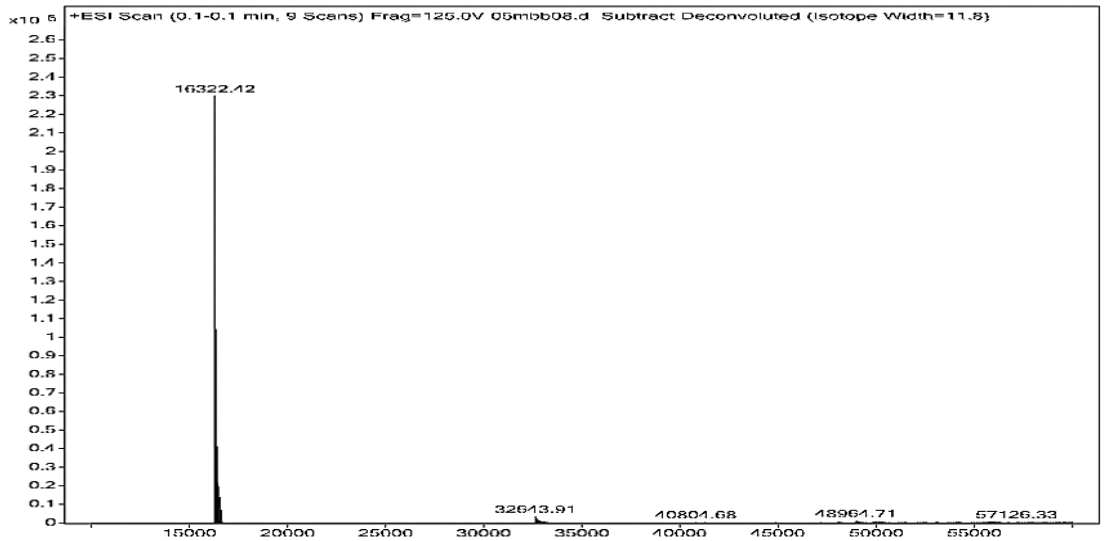
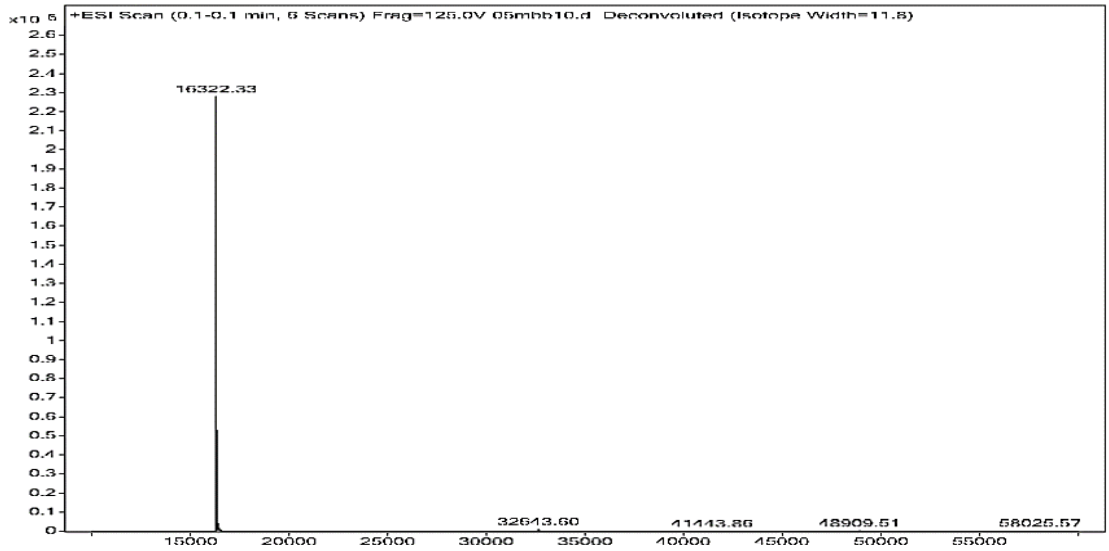


Figure 5.6 (B-iv)



In the monomeric region of the mass spectrum, peaks corresponding to the binding of one NAG_5 molecule to the monomer of YgaU could be identified (17356 Da). A second, the much smaller peak corresponding to binding of two NAG_5 molecules to the monomer of YgaU (18389 Da) could also be seen at a lower ion abundance (Figure 5.6 A). The two NAG_5 molecules bound to YgaU suggests that the LysM domain of the protein can be occupied by two oligosaccharide chains as seen in KEG15107 and Trc1. Comparing the protein peaks in the samples with either sodium or potassium, multiple peaks could be observed indicating the major buffer cation in the solution (Figures 5.6 B-v and B-vi). In the case of the sample with potassium, minor peaks associated with the presence of sodium from the purification can also be seen.

Figure 5.6 (B-v)

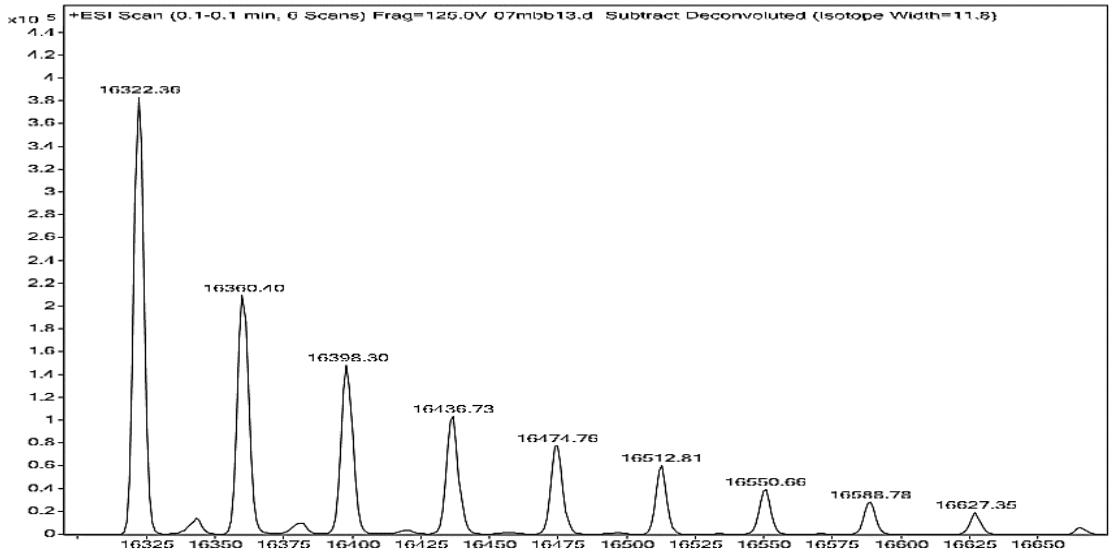
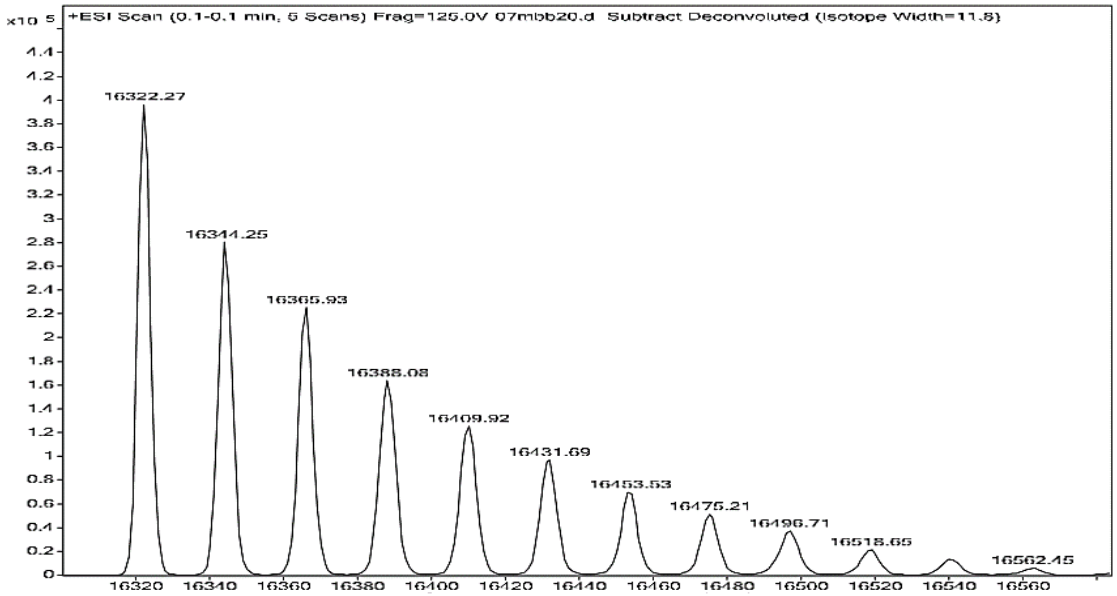


Figure 5.6 (B-vi)



Peaks corresponding to the binding of one or two NAG_5 molecules to the dimer of YgaU with molecular masses of 33780 and 34713 Da, respectively, were detected in the dimer region of the mass spectrum (Figure 5.6 C). The peak for the binding of one NAG_5 molecule to the dimer was at a much lower level of significance than for the binding of two NAG_5 molecules to the dimer. Again, the biological significance of the dimer is not clear and it is possible that these peaks represent molecules non-specifically aggregating to each other. The crystal structure of the apo *Burkholderia pseudomallei* YgaU is monomeric (Claudine Bisson and DWR, personal communication) as is the structure of *E. coli* YgaU (Ashraf *et al.*, 2016). However, whether this is an artefact of crystallization or whether apo YgaU dimerizes or whether it dimerizes in the presence of polyNAG remains to be determined.

Figure 5.6 (C)

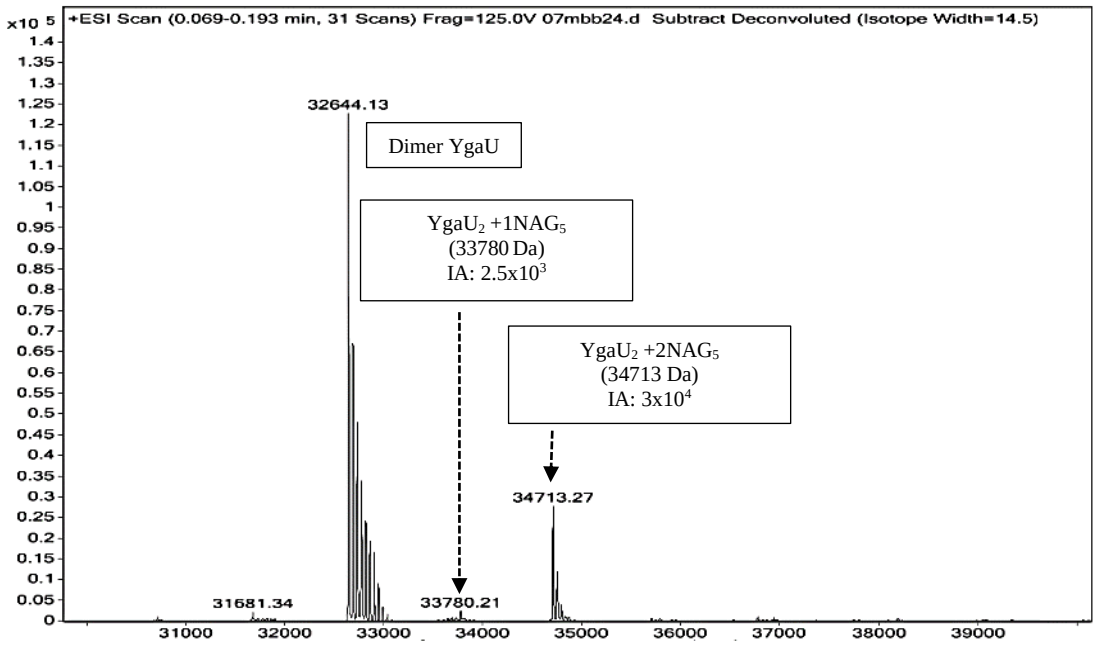


Figure 5.6: Mass spectrometry analysis of the YgaU-NAG_5 complex at a 1:200 protein to sugar ratio in buffers containing potassium or sodium chloride. A) In the sample of YgaU in 50 mM sodium chloride the monomer with a molecular mass 16322 Da was observed. Analysis of the monomer region of the mass spectrum showed peaks for the binding of one and two NAG_5 molecules to YgaU with expected molecular masses 17356 Da and 18389 Da, respectively, indicated in the boxes. B) Deconvoluted mass spectra of YgaU in samples with (i) 200 mM potassium chloride and 20 mM NAG_5 (ii) 200 mM sodium chloride and 20 mM NAG_5 (iii) 50 mM potassium chloride (iv) 10 mM Tris (pH 8). B-v and B-vi) Expanded views of the deconvoluted spectra corresponding to samples B-iii and B-iv, respectively, to show the presence of the potassium ions (38 Da) and sodium ions (22 Da) in the buffers. Note the presence of minor sodium peaks in the buffer containing potassium ions (v). C) An expanded view of the deconvoluted mass spectra of the sample of YgaU as in (A) in the region of the dimer. Peaks corresponding to the expected molecular mass of the dimer (32644 Da) and to two NAG_5 molecules bound to the dimer of YgaU (34713 Da) can be seen. A much smaller peak corresponding to one NAG_5 molecule bound to the dimer was also observed (33780 Da).

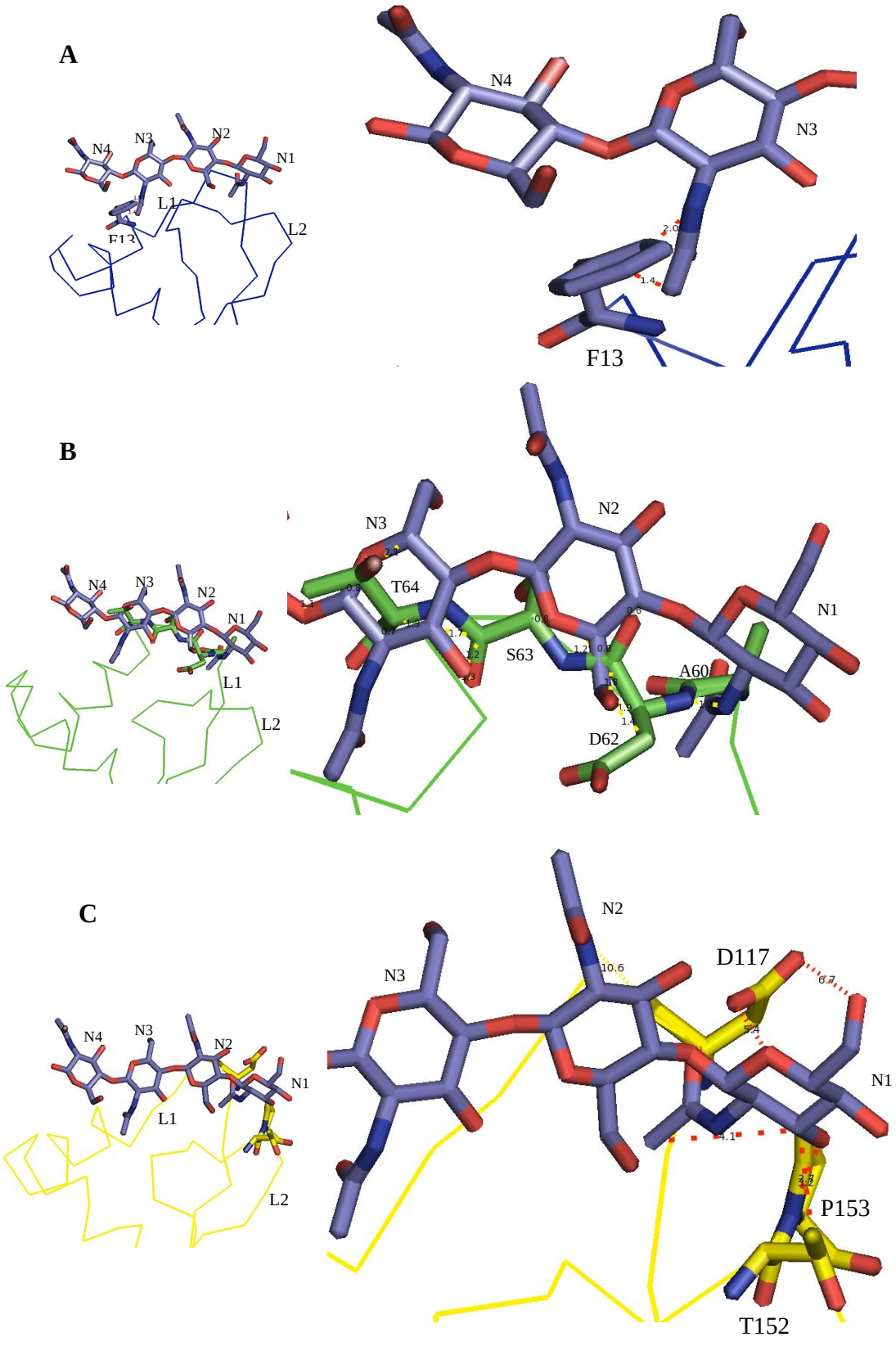
5.3.3 Implications of the mass spectrometry analysis for the binding of sugar chains to LysM domains of KEG15107

The findings from MS analysis on the apo KEG15107 protein revealed that this protein has multiple quaternary structures in which a monomer and a dimer and a tetramer are

the predominant species where the relative abundance of the monomer is greater than the dimer or the tetramer.

The identification of one, two, three or four oligosaccharides bound to the monomeric species of KEG15107 by NAG_3 and NAG_4 under conditions of high sugar concentration indicates that as well as the LysM Domain 1 binding oligosaccharide, as indicated by the crystallography, Domains 2, 3 and 4 can also bind sugar chains. The findings are further validated by the observation of either two or four oligosaccharides bound to the dimeric form of KEG15107-NAG_3, or KEG15107-NAG_4 complexes. The observation that the dimer can only bind six rather than eight oligosaccharides suggests that at least one of the binding sites is sterically hindered in the dimer. However, the binding of six oligosaccharides to the dimer species remains somewhat uncertain as the ion abundance of the mass peak for this particular species was very small possibly suggesting that in the dimer binding to two of the four LysM domains is also hindered. The observation that the tetramer can only bind four sugar chains suggests that in this quaternary structure, three of the LysM binding sites are sterically hindered (Figure 5.2).

Detailed analysis of the structure of the apo monomer of KEG15107 was carried out to evaluate the accessibility of the oligosaccharide binding sites to validate the mass spectrometry analysis. Superposition of the LysM Domain 1 from the KEG15107-NAG structure with a NAG chain bound on the domain onto Domains 2, 3 and 4 of the apo monomer KEG15107 structure shows that all the domains can bind to oligosaccharides as their sugar-binding sites are fully exposed to solvent but conformational changes may be required to permit sugar binding (Figure 5.7 and Table 5.0). For example, a conformational change is required to alter the position of the aromatic ring of Phe13 on Domain 1 as in the apo protein its position blocks sugar binding at sites S3 and S4 of the domain. In addition, in Domain 2, the L1 motif that must border the sugar-binding site must also move to allow the binding of the oligosaccharide. The conformation of Domain 3 appears to be compatible with sugar-binding but that of Domain 4 also requires a conformational change of its L2 motif to permit sugar-binding.


A
N4
N3
N2
N1
L1
L2
F13
N4
N3
F13
B
N4
N3
N2
N1
L1
L2
N3
N2
T64
S63
A60
D62
N1
C
N4
N3
N2
N1
L1
L2
N2
N3
D117
10.6
6.7
5.4
N1
4.1
P153
T152

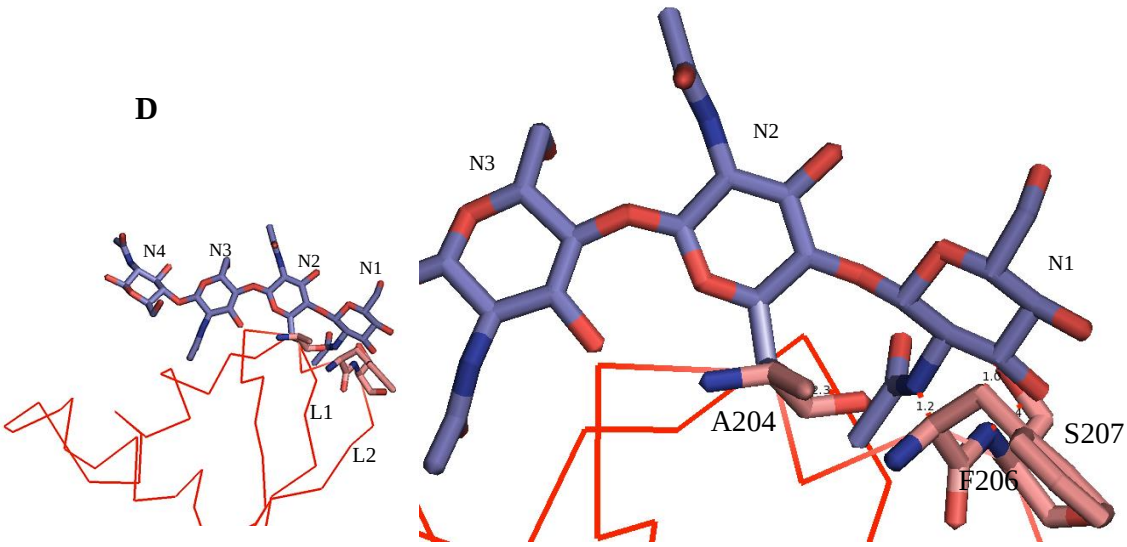


Figure 5.7: Sugar binding analysis of the four domains of the apo KEG15107 monomer. A) The binding site on Domain 1 of the KEG15107 monomer is fully exposed to solvent but the aromatic ring of F13 must move to allow the sugar to bind. B) The binding site on Domain 2 of the KEG15107 monomer is fully exposed to solvent but the residues on L1 of Domain 2 cause steric hindrance. C) The binding site on Domain 3 of the KEG15107 monomer is totally exposed to solvent. D) The binding site on Domain 4 of the KEG15107 monomer is fully exposed to solvent but conformational changes to residues on the L2 region are required for sugar-binding.

Table 5.0: Oligosaccharide binding analysis to the monomer of apo KEG15107

LysM domains	Monomer	
	Exposure to solvent	Conformation for sugar-binding
Domain 1	Fully exposed	The aromatic ring of F13 blocks access to sugar-binding site but moves on sugar recognition
Domain 2	Fully exposed	The L1 loop causes steric hindrance blocking sugar binding
Domain 3	Fully exposed	Conformation is compatible with NAG binding
Domain 4	Fully exposed	The L2 loop causes steric hindrance blocking sugar binding

In the dimer form of KEG15107, Domains 1, 2 and 3 appeared to be fully exposed to solvent and therefore, binding to oligosaccharides is possible. However, the binding site on Domain 4 is partially buried from the solvent by its interaction with a two-fold related partner. Steric hindrance would, therefore, prevent the oligosaccharide binding to this domain in the dimer. The observation that two or four oligosaccharides can be bound to the dimer of KEG15107 is consistent with the analysis of the exposure of the different binding sites in the dimer (Figure 5.8 and Table 5.1). The presence of small peaks for the binding of six oligosaccharides is again consistent with the MS results.

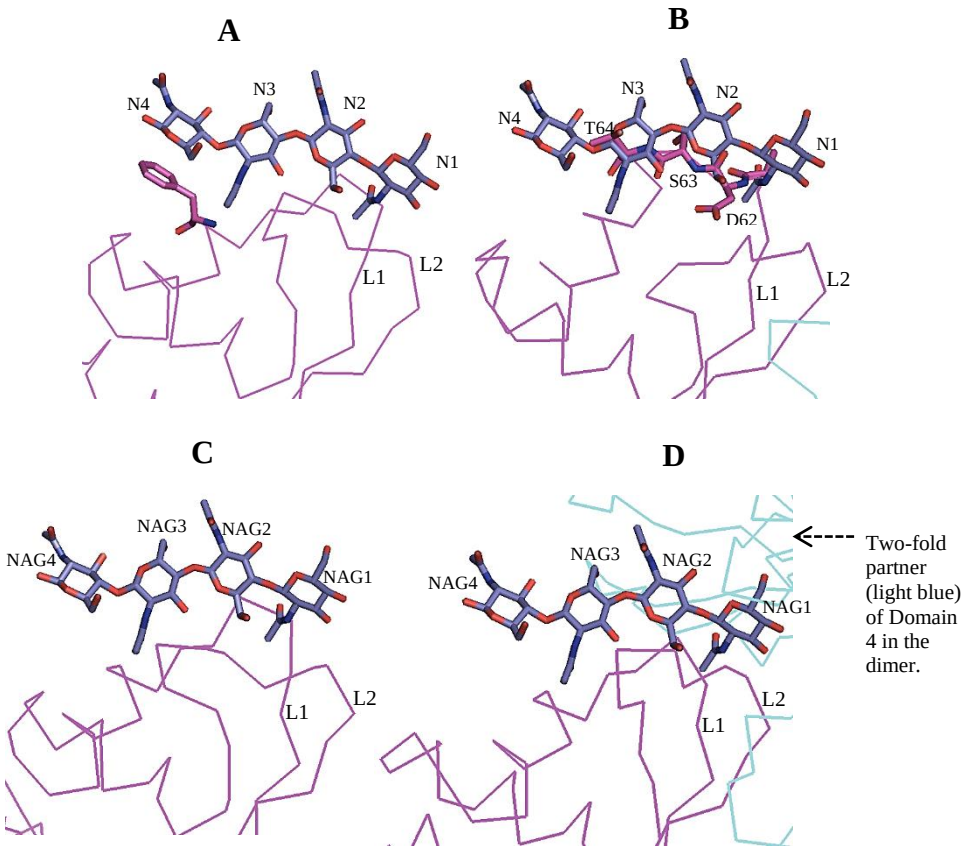


Figure 5.8: Sugar binding analysis of the four domains of the KEG15107 dimer. A) The binding site on Domain 1 of the KEG15107 monomer is fully exposed to solvent and the aromatic ring of F13 moves to allow the sugar to bind. B) The binding site on Domain 2 of the KEG15107 dimer is fully exposed to solvent but the position of L1 blocks the binding pocket. C) The binding site on Domain 3 of the KEG15107 monomer is totally exposed to solvent. D) The binding site of Domain 4 in the KEG15107 dimer is partially buried from the solvent by interactions across the dimer interface preventing sugar-binding.

Table 5.1: Oligosaccharide binding analysis to the dimer of the oligosaccharide complex of KEG15107

LysM domains	Dimer	
	Exposure to solvent	Conformation for sugar-binding
Domain 1	Fully exposed	Conformation is compatible with NAG binding
Domain 2	Fully exposed	The L1 motif causes steric hindrance blocking sugar binding
Domain 3	Fully exposed	Conformation is compatible with NAG binding
Domain4	Partially buried from solvent	The L1 and L2 motifs cause steric hindrance blocking sugar binding

In the tetramer of KEG15107, Domains 1 and 4 are buried from the solvent by their interaction with two-fold related partners and therefore, steric hindrance would prevent the oligosaccharide binding to these domains in the tetramer. Domain 3 appears to be fully exposed to the solvent and therefore oligosaccharide binding is possible whilst Domain 2 is partially exposed to solvent and steric hindrance could prevent the binding to this domain (Figure 5.9 and Table 5.2). In general, for the tetramer, the evidence for the binding of polyNAG is less convincing but the data for the binding to NAG_5 seemed clear (Figure 5.2 C).

Table 5.2: Oligosaccharide binding analysis to the tetramer of apo KEG15107

LysM domains	Tetramer	
	Exposure to solvent	Conformation for sugar-binding
Domain 1	Partially buried	Conformation causes steric hindrance
Domain 2	Fully exposed	The L1 motif causes steric hindrance to sugar-binding
Domain 3	Fully exposed	Conformation is compatible with NAG binding
Domain4	Fully buried	Conformation causes steric hindrance

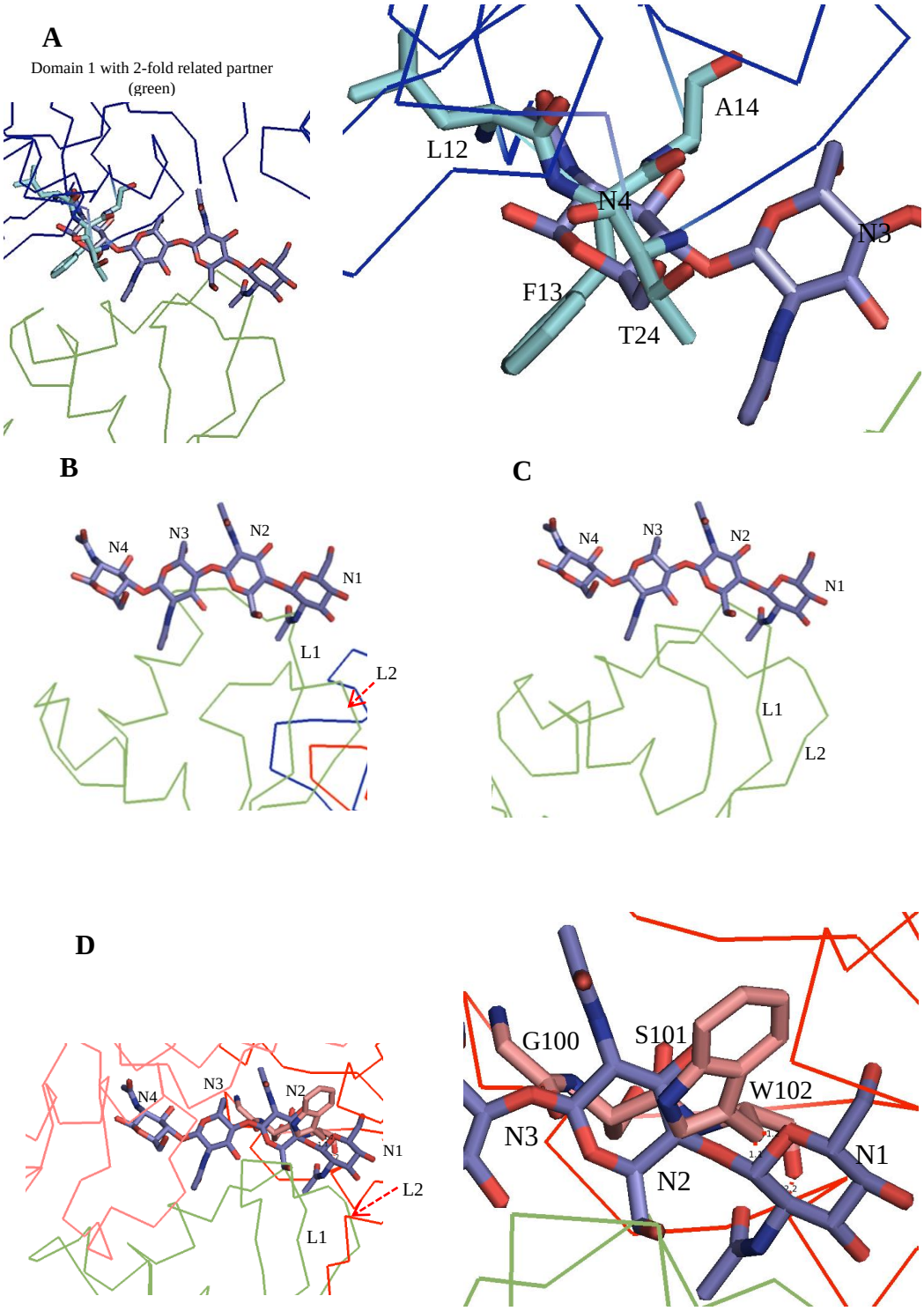


Figure 5.9: Sugar binding analysis of the four domains of the apo KEG15107 tetramer. A) Domain 1 on the KEG15107 tetramer was partially buried from solvent. Sugar binding to the binding sites of the domain in the tetramer causes steric hindrance to the two-fold partner of Domain 1 in the tetramer. B) The binding site on Domain 2 of the KEG15107 tetramer is fully exposed to solvent but the L1 motif causes steric hindrance to block sugar-binding. C) The binding site on Domain 3 of the KEG15107 tetramer is fully exposed to solvent. D) The binding site on Domain 4 of the KEG15107 tetramer is fully buried from solvent and binding to sugar causes steric hindrance.

5.4 Analysis of binding affinity between domains of KEG15107 and NAG_4

The observation of a monomeric species of KEG15107 binding principally to one or two or three or four molecules of polyNAG suggested that each of the LysM Domains 1, 2, 3 and 4, bind to oligosaccharides but probably with different affinities. Therefore, to further investigate the binding affinity of the four LysM domains of KEG15107, MS analysis was further carried out on the KEG15107-NAG_4 complex at a 1:100, 1:50 and 1:2 ratio of protein to sugar.

Analysis of the complex of KEG15107-NAG_4 by mass spectrometry at a 1:100 ratio of protein to sugar is similar to the equivalent analysis at a 1:200 ratio (Refer Figure 5.4). The presence of NAG_4 molecule in the complex is shownin Figure 5.10 A. However, the fraction of the KEG15107-NAG_4 complexes from the 1:100 ratio sample compared to that of unbound KEG15107 reduces significantly for the monomeric and dimeric species of the protein as the concentration of sugar decreases (Figure 5.10 B-C and Figure 5.13). For the monomeric and the dimeric species of KEG15107, the binding of up to five or six oligosaccharides could be identified, respectively, in both samples (Figure 5.4 and Figure 5.10).

Figure 5.10 (A)

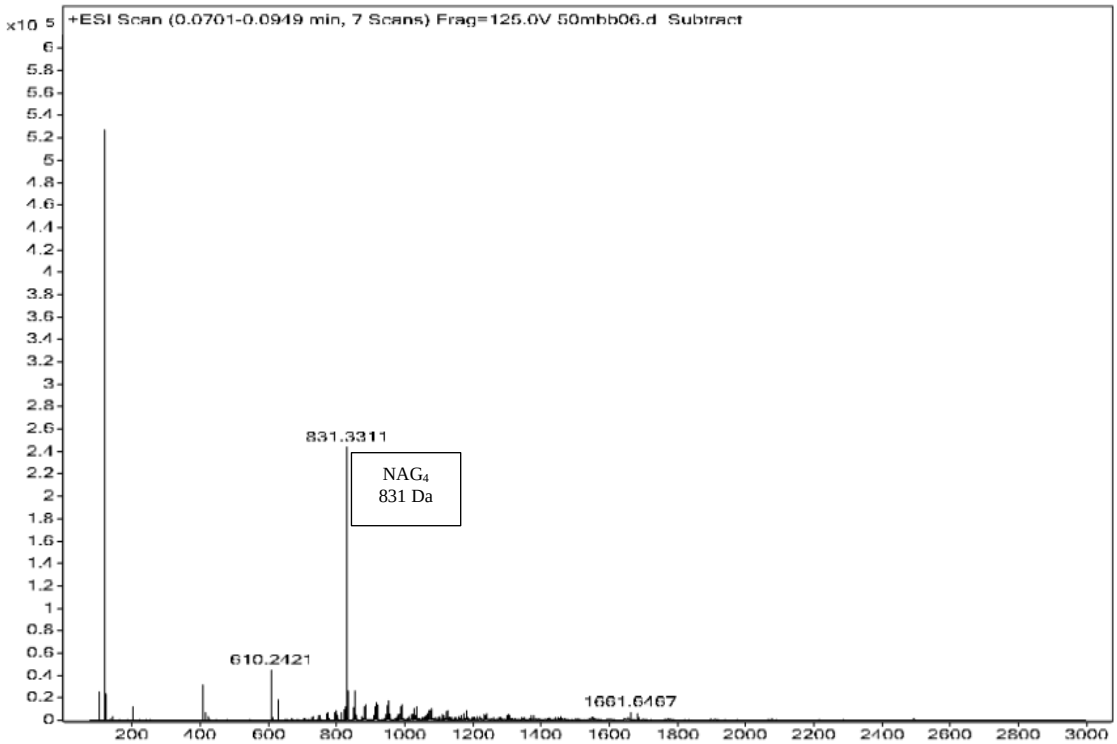


Figure 5.10 (B)

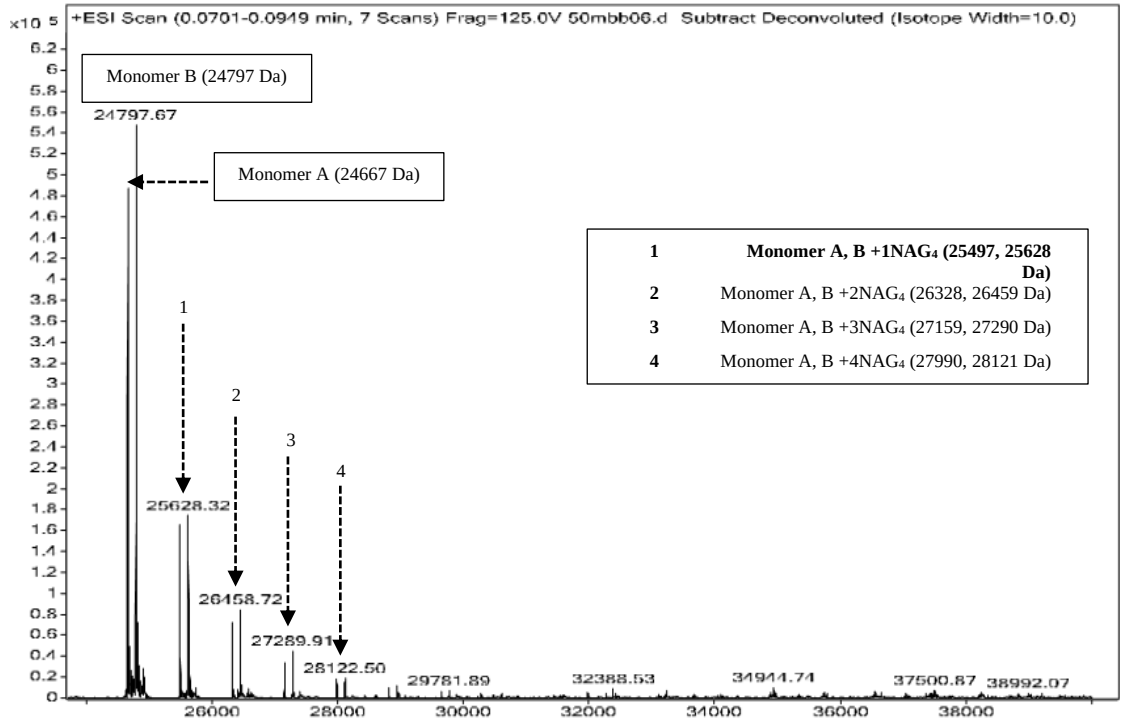
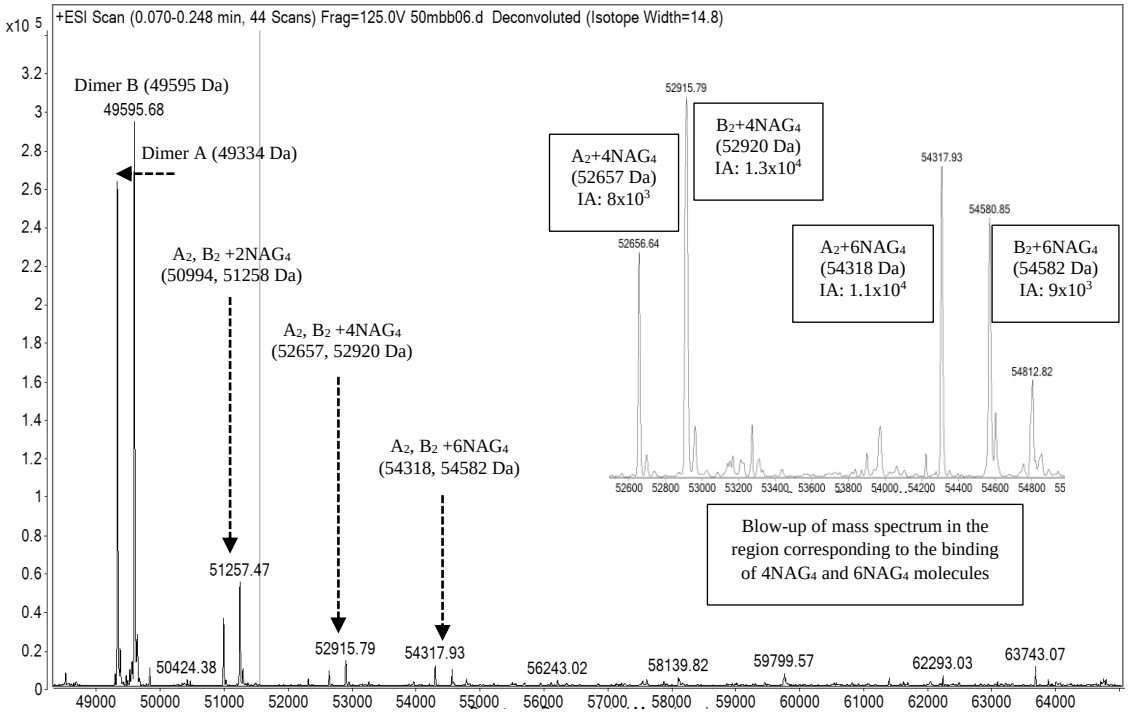


Figure 5.10 (C)



The analysis of the results of oligosaccharide binding to the tetramer with NAG_4 provided no convincing evidence for significant interaction unlike the earlier studies with NAG_5 (Figure 5.2 C) and (Figure 5.10 D).

Figure 5.10 (D)

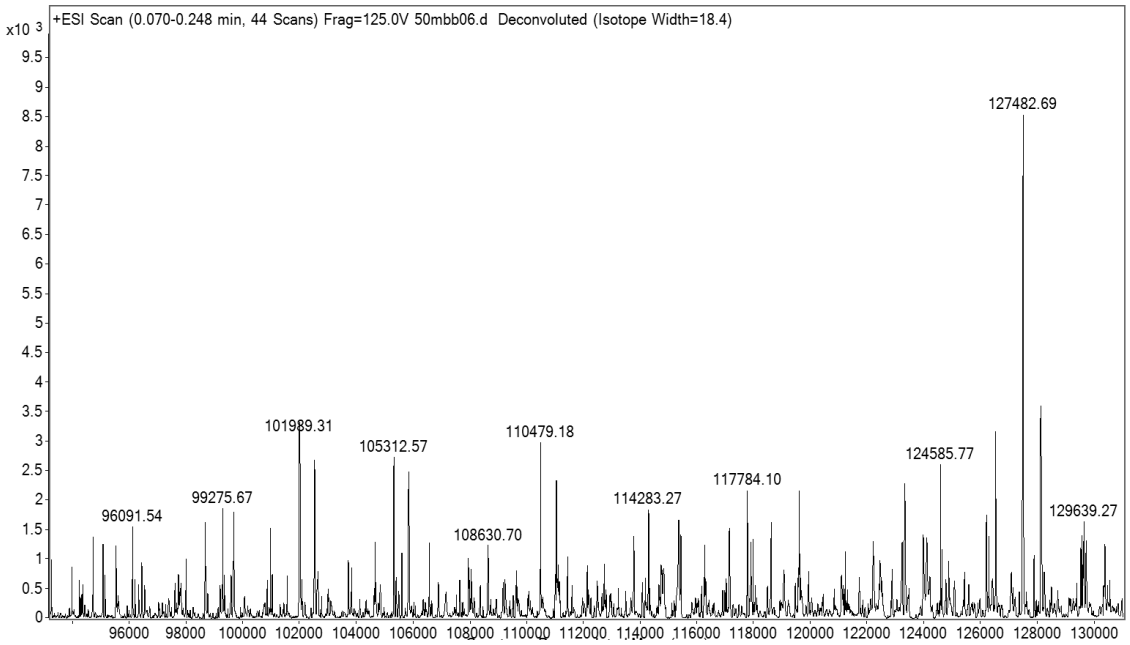


Figure 5.10: Mass spectrometry analysis on the KEG15107-NAG_4 complex at a 1:100 of protein to sugar ratio. A) The convoluted mass spectrum shows a peak with a molecular mass of 831 Da corresponds to a NAG_4 molecule. B) The monomer region of the mass spectrum. One or two or three or four NAG_4 molecules to the KEG15107 protein were identified at a significant level of abundance. C) The dimer region of the mass spectrum. The peaks with expected molecular masses for two, four or six NAG_4 molecules bound to the KEG15107 dimer were detected. The molecular masses of the complexes are shown in boxes. Ion abundance for the selected peaks is also indicated in the boxes. D) There was no significant peak identified in the tetramer region of the mass spectrum.

One possible explanation of this could be that NAG_5 might bind with slightly higher affinity to KEG15107 compared to NAG_4. If this is the case then the domain could be regarded as having five and not four sugar-binding sites. This might explain the observation in the crystal structure of KEG15107 with NAG_4 where the two molecules in the AU bind NAG_4 differently with sites S1 to S4 occupied by NAG in one monomer of the asymmetric unit and S0-S3 occupied in the other. Clearly, the binding affinity of different NAG oligomers to KEG15107 should be the subject of further work.

Interestingly, estimates of the affinity of the binding of NAG oligomers to a single LysM domain in AtlA, determined in solution by surface plasmon resonance (SPR) analysis, show that they increase with oligosaccharide chain length, with apparent Kd values of approximately 400 μM, 40 μM, 12 μM and 6 μM for NAG_3, NAG_4, NAG_5 and NAG_6, respectively. This too suggests that the interactions of the LysM domains involve more than four NAG residues. Whilst these affinity values are modest they do not reflect the binding strength in the cell wall where clustering of the ligands in different peptidoglycan strands would result in a marked increase in apparent affinity due to the high local concentration of the substrate.

A similar trend of protein-sugar complex formation was seen in the mass spectrum profiles of the KEG15107-NAG_4 complexes from the 1:50 ratio (Figure 5.11). The presence of NAG_4 molecule in the complex is shown Figure 5.11 A. Binding of one or two or three or four NAG_4 molecules to KEG15107 could be detected in the monomer region of the mass spectrum (Figure 5.11 B). For the dimeric species of the protein, the fraction of the bound complexes reduced significantly as expected when the concentration of sugar is lowered (Figure 5.11 C and Figure 5.14). Only insignificant peaks were observed in the tetramer region of the mass spectrum and there was no evidence from this spectrum to support the existence of the KEG15107 tetramer complex with NAG molecules (Figure 5.11 D).

Figure 5.11 (A)

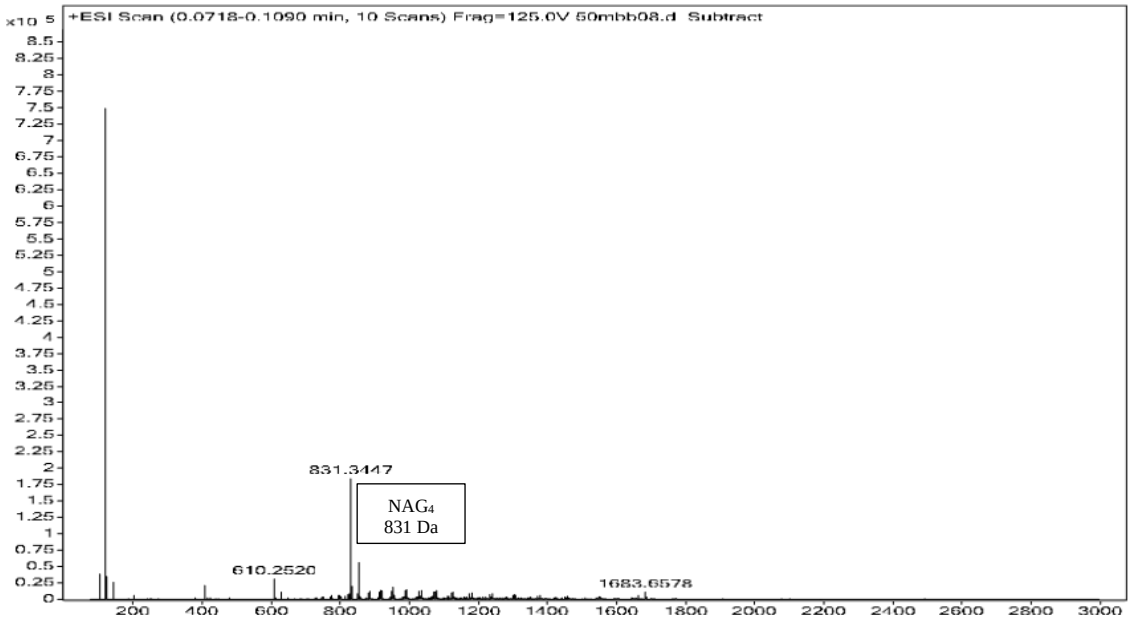


Figure 5.11 (B)

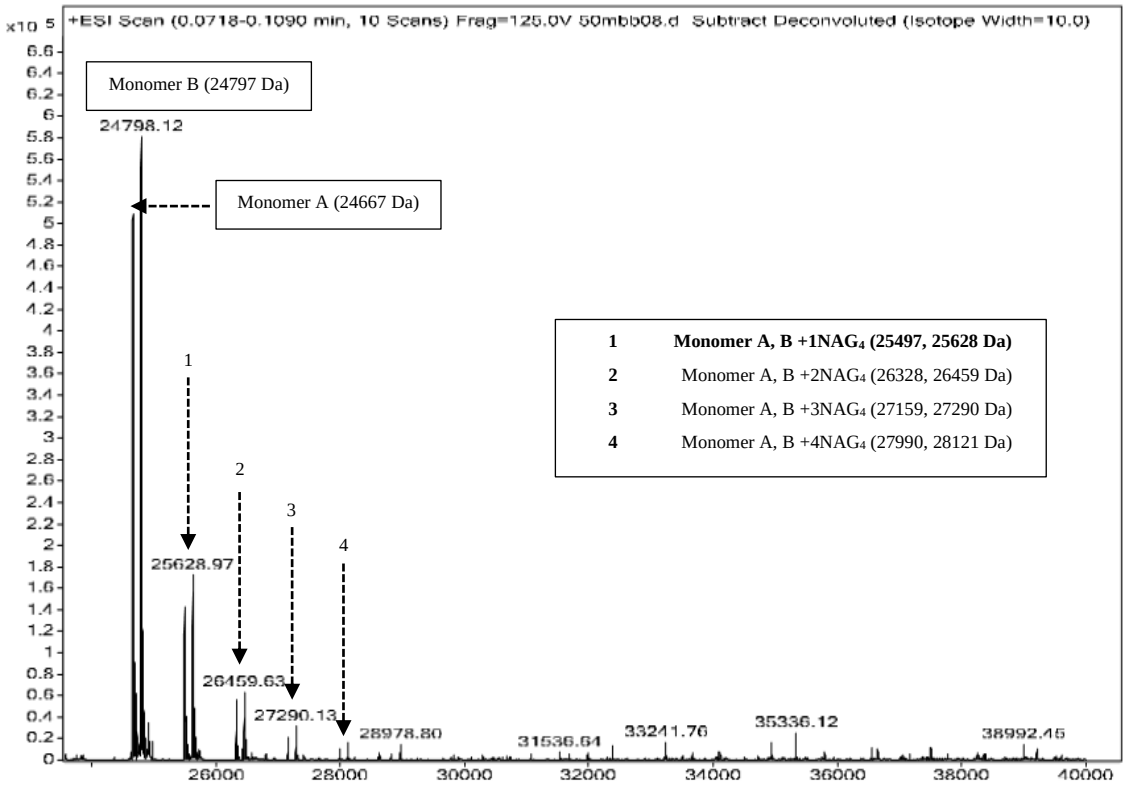


Figure 5.11 (C)

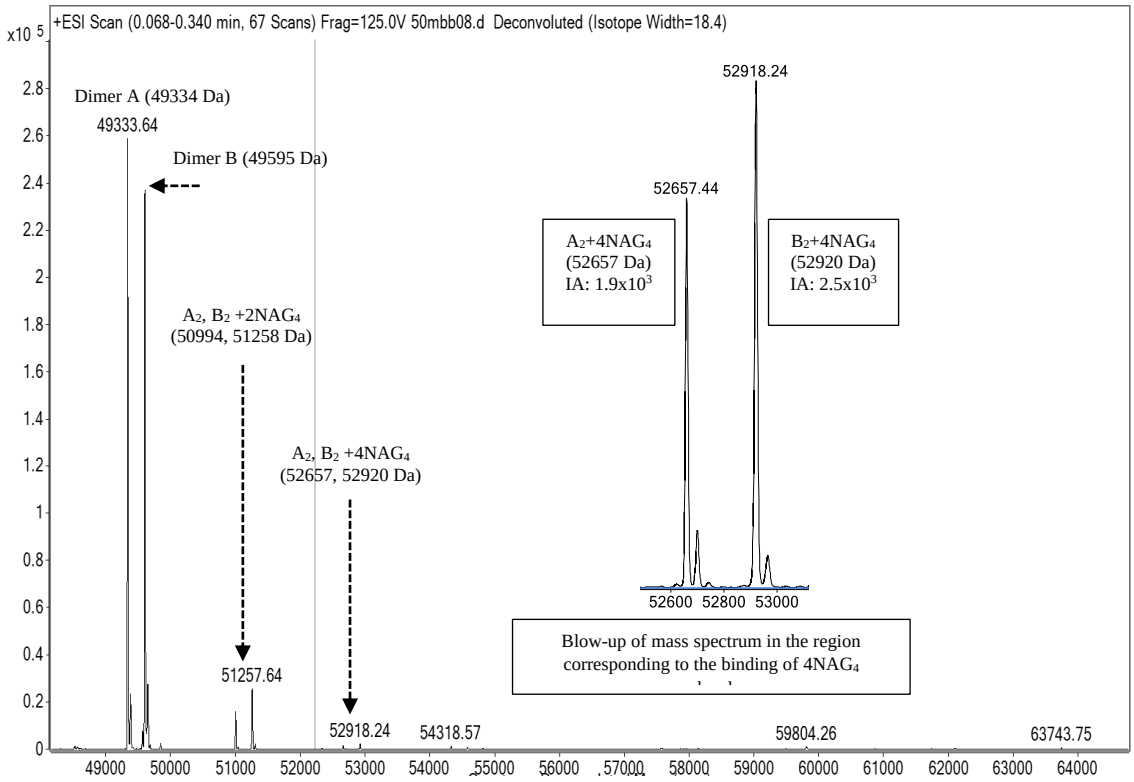


Figure 6.11 (D)

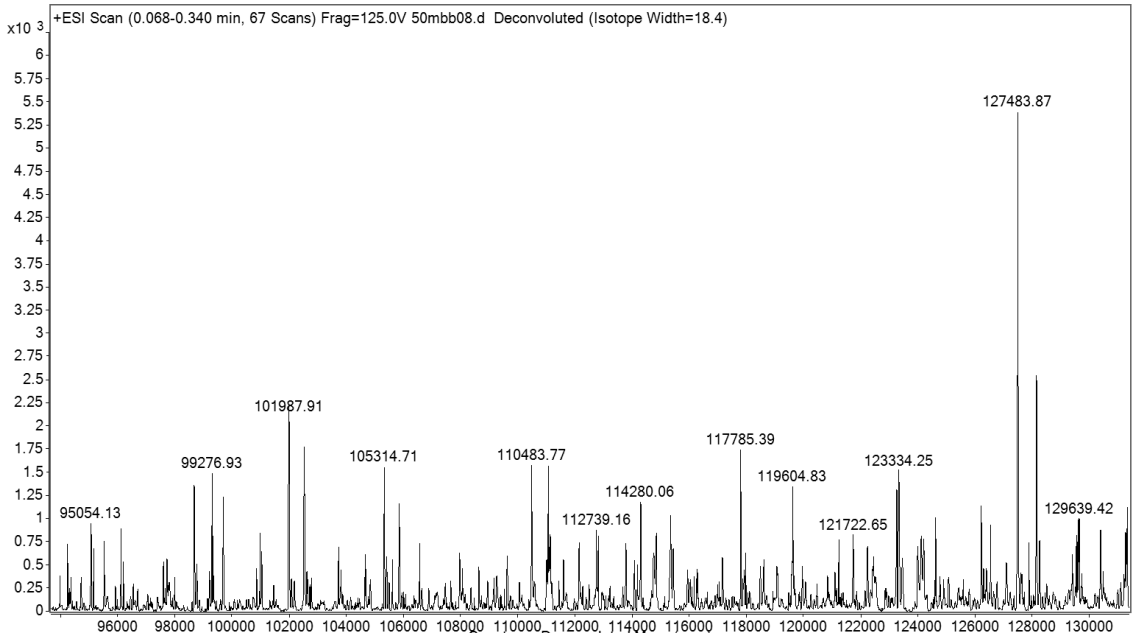


Figure 5.11: Mass spectrometry analysis on the KEG15107-NAG_4 complex at a 1:50 of protein to sugar ratio. A) The convoluted mass spectrum shows a peak with a molecular mass of 831 Da corresponding to a NAG_4 molecule. B) The monomer region of the mass spectrum. One, two, three or four NAG_4 molecules bound to the KEG15107 protein were identified at a significant level of abundance. The molecular masses of the complexes are shown in boxes. C) The dimer region of the mass spectrum. The peaks with expected molecular masses for two or four NAG_4 molecules bound to the KEG15107 dimer were detected and the masses are shown in boxes. Ion abundance for the selected peaks is also indicated in the boxes. D) There was no significant peak corresponding to a tetramer complex of the KEG15107 protein with polyNAG chain identified in the tetramer region of the mass spectrum.

Using a lower concentration of NAG_4 (1:2 ratio) in the monomer region only complexes with one or two NAG_4 molecules could be identified and again the fraction of molecules with NAG_4 bound decreased (Figure 5.13). A similar situation could be seen in the dimer region of the spectrum (Figure 5.12).

Given that the different complexes of oligosaccharide bound to KEG15107 can be identified by mass spectrometry could this technology be used to determine the binding affinity between the protein and NAG oligomers? Two problems with this approach can be identified. Firstly, it is unclear if mass spectrometry can be used in this instance

in a quantitative manner as the different species might not fly in the mass spectrometer with equal efficiency. Secondly, given the various equilibria that can be identified, including the different quaternary structures, and the different numbers of NAG molecules bound to each subunit (including multiple, non-overlapping, occupancy of each of the sites), a full kinetic analysis is premature given the limited data available. Nevertheless, as shown in Table 5.3, the fraction of molecules binding NAG_4 does decrease as the sugar concentration is lowered.

Figure 5.12 (A)

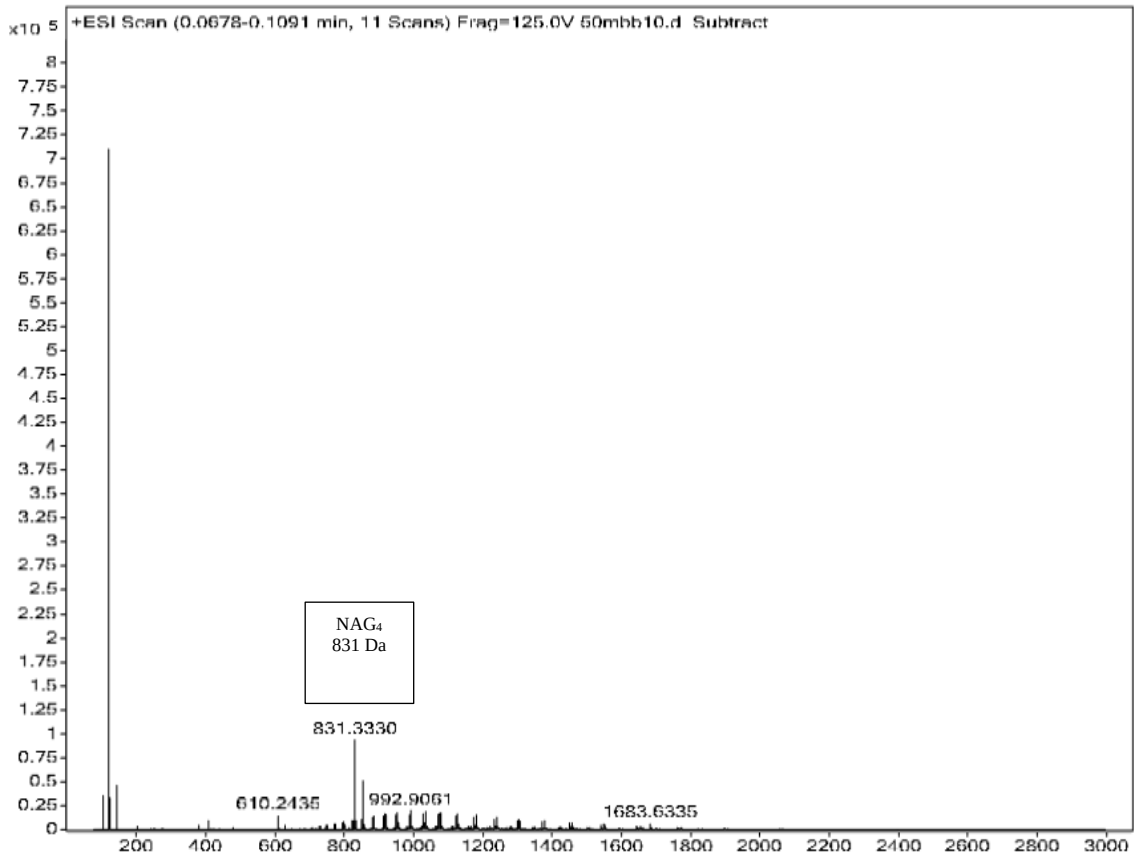


Figure 5.12 (B)

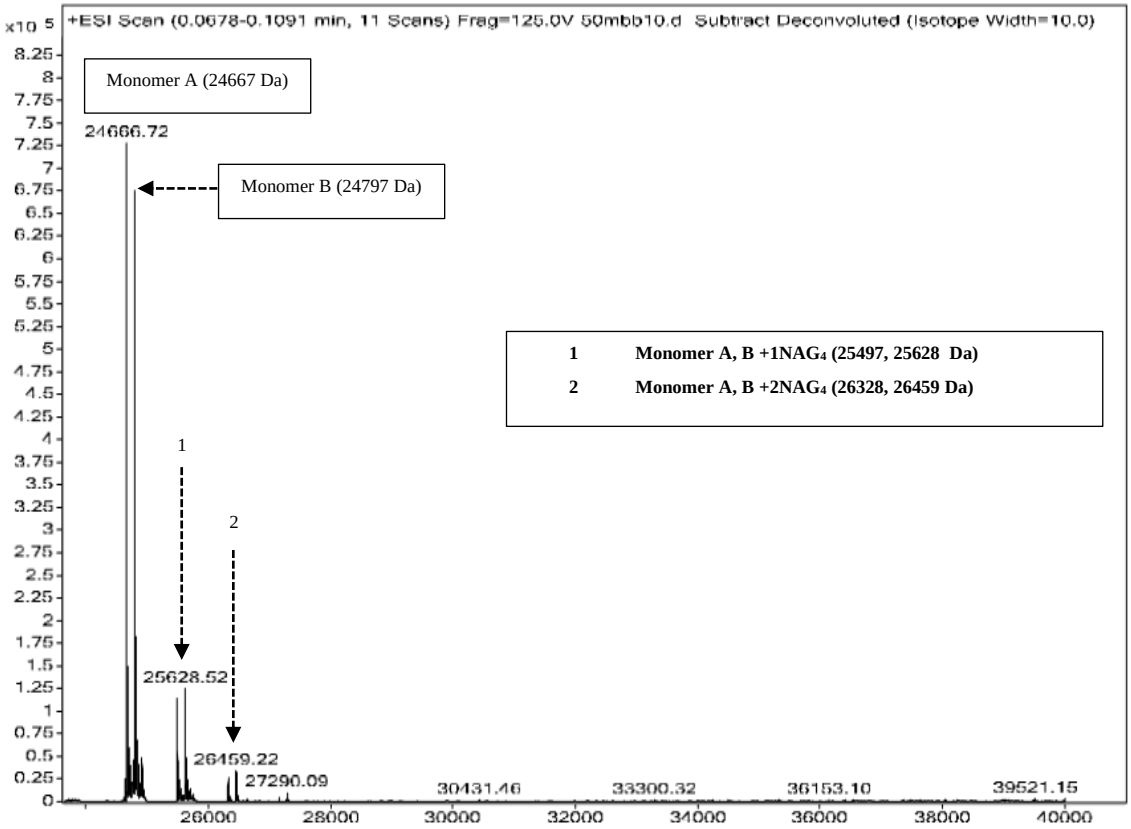


Figure 5.12 (C)

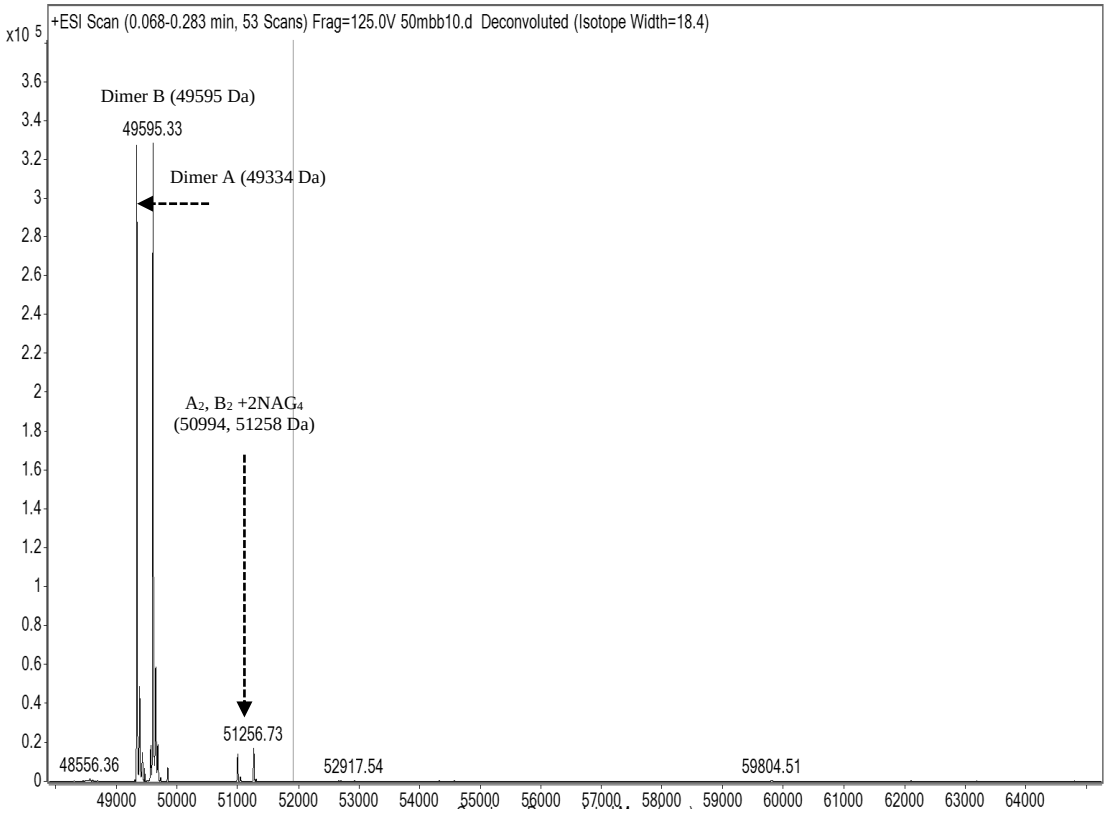


Figure 5.12 (D)

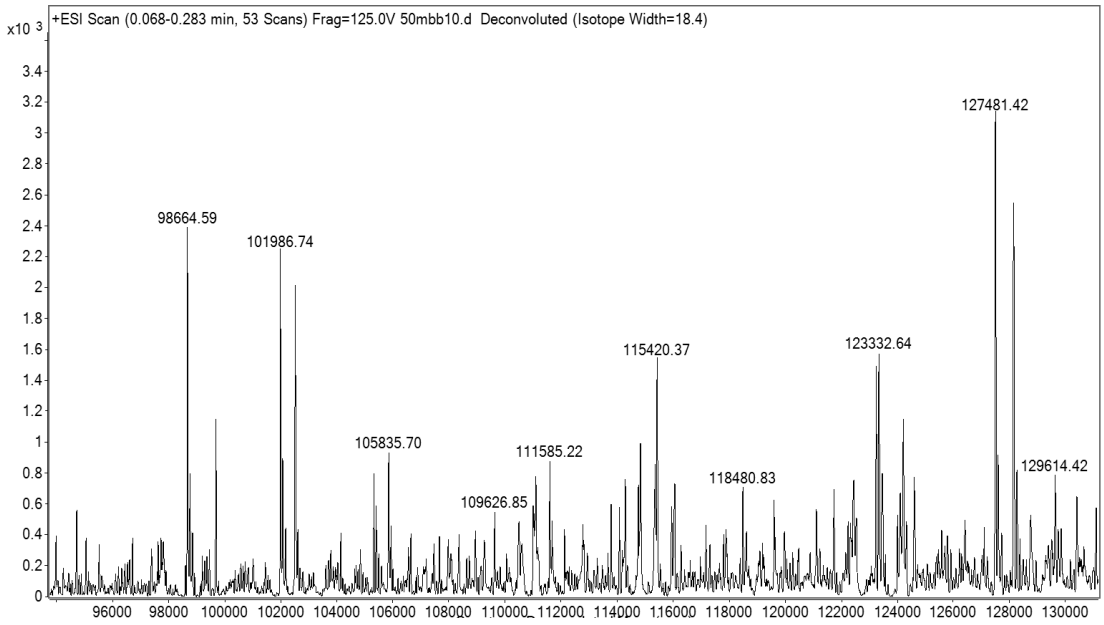


Figure 5.13: Mass spectrometry analysis on the KEG15107-NAG_4 complex at a 1:2 of protein to sugar ratio. A) The convoluted mass spectrum shows a peak with a molecular mass of 831 Da corresponds to a NAG_4 molecule. B) The monomer region of the mass spectrum. One or two NAG_4 molecules bound to the KEG15107 protein were identified at a significant level of ion abundance. The molecular masses of the complexes are shown in boxes. C) The dimer region of the mass spectrum. Two NAG_4 molecules bound to the KEG15107 dimer were detected in the region of the mass spectrum. D) There was no significant peak corresponding to a tetramer complex of the KEG15107 with polyNAG chain identified in the tetramer region of the mass spectrum.

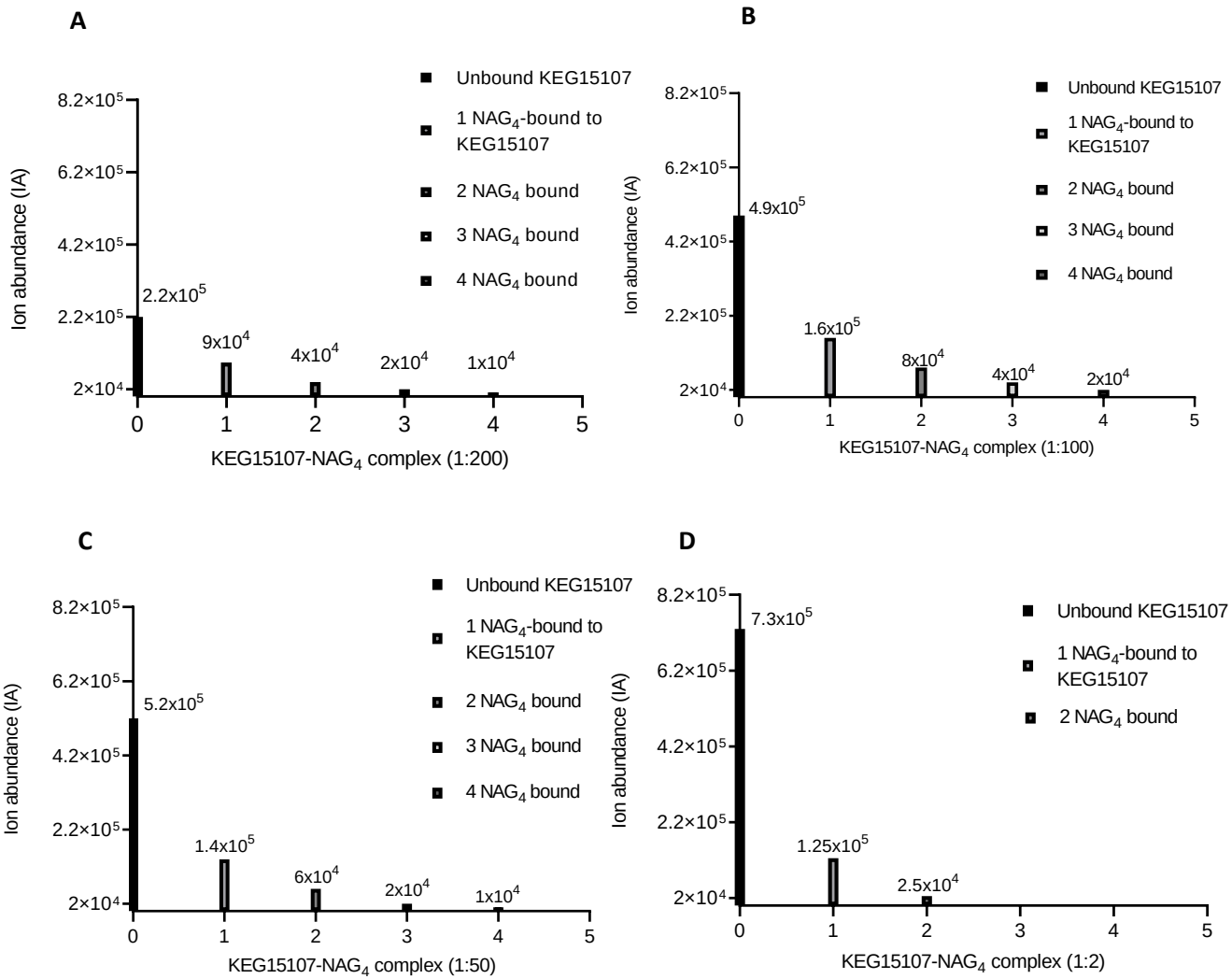


Figure 5.13: The KEG15107-NAG_4 complexes detected in the monomeric region of the mass spectrum for species A of the protein at four different ratios of protein to sugar. A) The ion abundance of unbound KEG15107 and the principal NAG_4 complexes (with one, two, three or four NAG_4 molecules) from the sample with a 1:200 protein to sugar ratio. B) As (A) but with the sample at a 1:100 protein to ratio. C) As (A) but with the sample at a 1:50 ratio. D) As (A) but with the sample at a 1:2 ratio.

Table 5.3: Analysis of the interaction between species A of KEG15107 and NAG_4

Protein to sugar ratio (concentration)	Total amount of KEG15107 and its complexes with NAG_4 (IA)	1 NAG_4 bound (IA)	2 NAG_4 bound (IA)	3 NAG_4 bound (IA)	4 NAG_4 bound (IA)	Total fraction bound [£]
1:200 P (0.1 mM) S (20 mM)	$3.8x10^5$	$9x10^4$ (0.24)#	$4x10^4$ (0.11)#	$2x10^4$ (0.05)#	$1x10^4$ (0.03)#	0.43
1:100 P (0.1 mM) S (10 mM)	$7.9x10^5$	$1.6x10^5$ (0.20)#	$8x10^4$ (0.10)#	$4x10^4$ (0.05)#	$2x10^4$ (0.03)#	0.38
1:50 P (0.1 mM) S (5 mM)	$7.5x10^5$	$1.4x10^5$ (0.19)#	$6x10^4$ (0.089)#	$2x10^4$ (0.03)#	$1x10^4$ (0.01)#	0.31
1:2 P (0.1 mM) S (0.2 mM)	$8.8x10^5$	$1.25x10^5$ (0.14)#	$2.5x10^4$ (0.03)#	ND	ND	0.16

#Values in parentheses are the fraction of the KEG15107-NAG_4 complexes in the total of expected analytes
P: concentration of the protein, S: concentration of NAG_4.
The ion abundance values given above are those for species A of KEG15107 (the species with the N-terminal methionine)
ND: not detected
[£]The total fraction bound is the sum of the ion abundance of molecules of species A binding NAG_4 compared to the sum of the ion abundance of all peaks related to species A

5.5 Interpretation of multiple oligosaccharide binding by KEG15107

Three-dimensional structures of KEG15107-NAG_3, KEG15107-NAG_4, and KEG15107-NAG_5 consistently suggested that only Domain 1 binds to the oligosaccharides. However, mass spectrometry analysis suggested that all four LysM domains of KEG15107 bind oligosaccharides. Moreover, the observation of minor peaks in the mass spectrum consistent with the binding of two oligosaccharides to each LysM domain of KEG15107, Trc1, and YgaU can be explained if two sugar chains simultaneously bind to different parts of the five sugar-binding sites (S0, S1, S2, S3 and S4). However, it would be expected that occupancy of all five sites by one oligosaccharide molecule would have higher affinity than the partial occupancy of the sites by multiple sugar chains.

Illustrative models for the possible occupancy of the different sites are provided in Figure 5.14. Note that the models showing multiple occupancy (Figure 5.14 B-D) necessarily leave at least one site unoccupied so as to avoid the steric clash that would occur between the C1 hydroxyl at the end of one sugar chain and the C4 hydroxyl of the second chain which normally are linked following the formation of the β-1-4

glycosidic bond if two different chains occupied adjacent sites. Interestingly, the observation in the crystal structure that only binding to one LysM domain of KEG15107 could be seen is similar to the crystal structure analysis of AtCERK1 in which the polyNAG bound to the only one of its three LysM domains following a crystal soaking experiment (Liu et al., 2012). In the structure of Ecp6, the chitin scavenger of *C. fulvum*, one molecule of polyNAG was seen to be bound simultaneously by two of its three LysM domains (LysM1 and LysM3) following co-crystallization (Sánchez-Vallet et al., 2013). This type of cooperative binding is possible for polyNAG where the sites of the oligosaccharide chains have a similar structure but not for peptidoglycan where the larger lactyl group on the C3 of NAM and any associated peptide linkers would block binding (Figure 5.17).

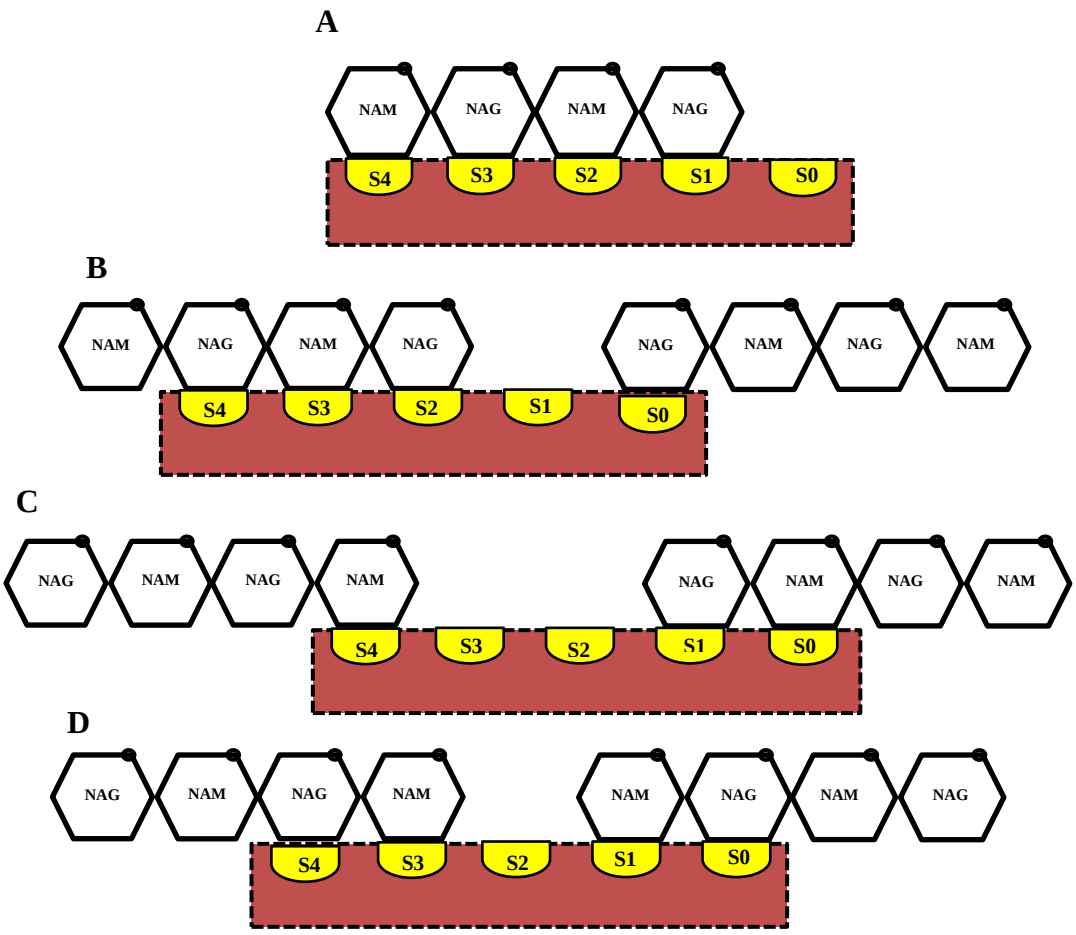
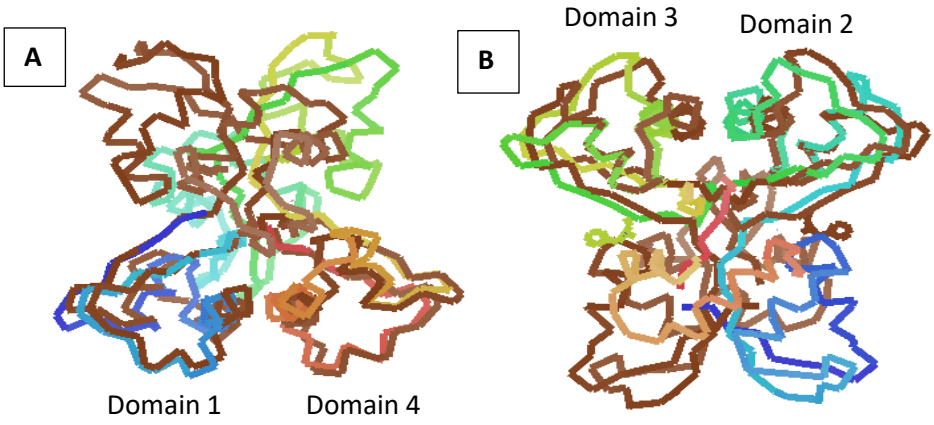


Figure 5.14: Possible models for dual oligosaccharide recognition by the LysM domains of KEG15107, YgaU, and Trc1. A) The five NAG binding pockets of KEG15107 are proposed to be occupied by one oligosaccharide oligomer. B-D) The five NAG binding pockets on a single LysM domain are occupied by two oligosaccharide molecules in different patterns with at least one site being unoccupied.

5.6 Implication of oligosaccharide recognition by KEG15107 for the architecture of peptidoglycan chains

Analysis of the monomer of KEG15107 shows that Domains 1 and 4 are related by a local pseudo-2-fold axis of symmetry (Figure 5.15 A) with an rmsd of 2.10 Å for superposition of equivalent residues (Table in Figure 5.15). Equally, Domains 2 and 3 are also related by a local pseudo-2-fold axis symmetry (Figure 5.15 B and Table). Moreover, the interface between Domains 2 and 3 is very similar to that between Domains 1 and 4. However, superposition of Domains 2 and 3 does not superimpose Domains 1 and 4 and thus the two local 2-fold axes are independent. The consequences of this local symmetry is that the equivalent oligosaccharide binding sites are also symmetrically related.

Based on the crystal structures, and in conjunction with the mass spectrometry analysis, the observation that each of the four domains of KEG15107 could bind sugar chains led to an attempt to model the relative position of the four oligosaccharide chains should they be bound at the same time to a monomer. The model was constructed by superimposing the coordinates for the complex of Domain 1 with polyNAG to each of Domains 2, 3 and 4 in turn to locate the approximate position of the oligosaccharide chain on the other domains. Given that there are four LysM domains there are six possible combinations of domains where the oligosaccharide chains would be in different orientations and with different separations (LysM 1-LysM 2, LysM 1-LysM 3, LysM 1-LysM 4, LysM 2-LysM 3, LysM 2- LysM 4 and LysM 3-LysM 4). Each of these are now considered in turn.



LSQKAB statistics for the structure superposition of Domains 1, 2, 3 and 4 of KEG15107		
	Domain 1 (residues compared)	R.m.s.d XYZ displacement values (Å)
Domain 4 (Equivalent residues)	(12-21) vs (173-182) (25-36) vs (189-200) (173-182) vs (12-21) (189-200) vs (25-36)	2.10
	Domain 2 (residues compared)	R.m.s.d XYZ displacement values (Å)
Domain 3 (Equivalent residues)	(66-74) vs (121-129) (121-129) vs (66-74)	0.64

Figure 5.15: Structure alignment of four LysM domains of KEG15107. A) Structure alignment between Domains 1 and 4 shows these two domains are related by local pseudo-2-fold symmetry. B) Structure alignment between Domains 2 and 3 shows these two domains are related by local pseudo-2-fold symmetry. The rmsd values for the LSQKB superposition between the four domains are shown in the table.

5.6.1 The orientation of LysM 1-LysM 4 and LysM 2-LysM 3

Examination of the model reveals that the bound sugar chains on Domains 1 and 4 lie approximately in the same plane and as a result of local 2-fold symmetry, they are in an antiparallel orientation, being separated by ~37 Å (measured from the oxygen of the glycosidic bonds between the sugars bound in the S2 and S3 sites) (Figure 5.16 A (i)). Equally, a similar situation is observed for the orientation of the oligosaccharides bound to Domains 2 and 3 which are separated by ~38Å (Figure 5.16 A (ii)).

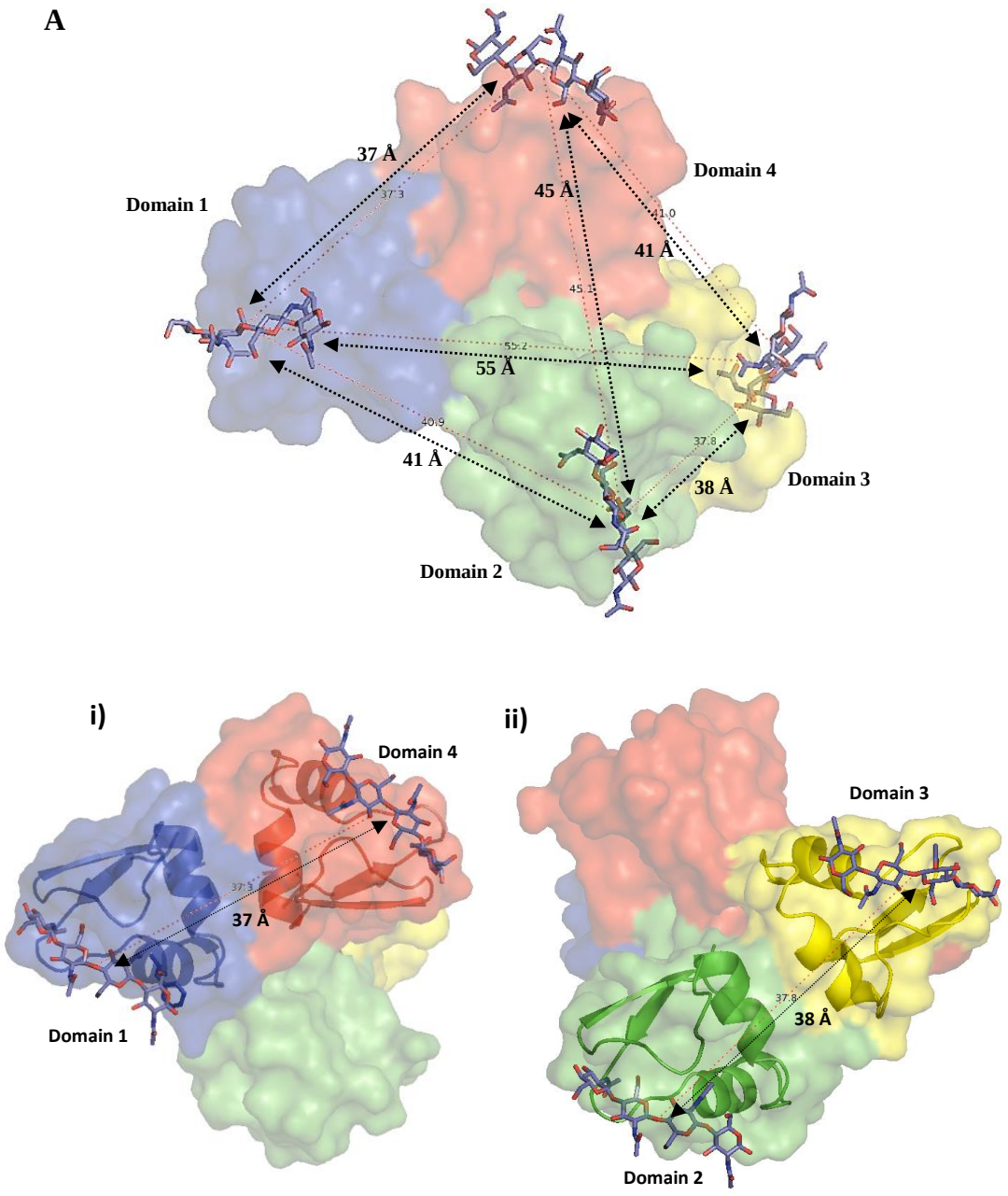


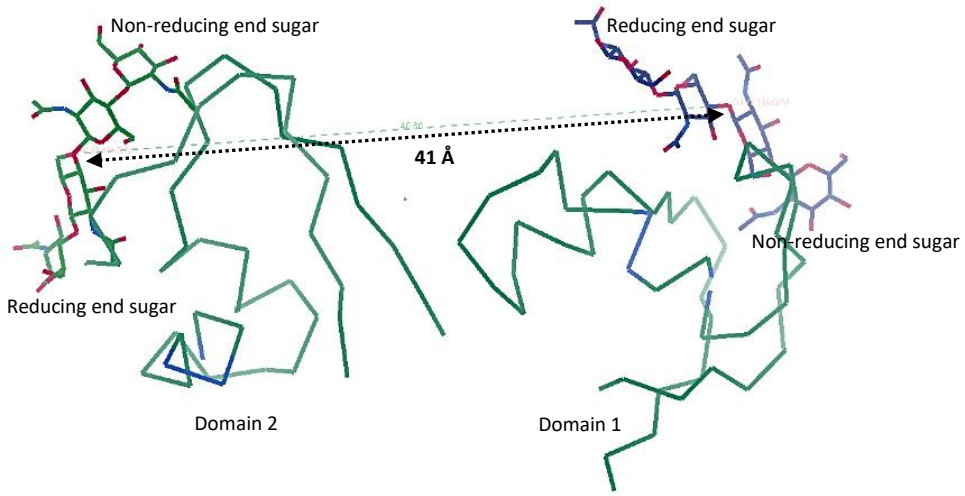
Figure 5.16 (A)

5.6.2 The orientation of LysM 1-LysM 2 and LysM 3-LysM 4

Given the local symmetry, the relationship between LysM 1 and LysM 2 is similar to that between LysM 3 and LysM 4. The oligosaccharide binding sites again lie approximately in the same plane and are separated by ~41 Å. In this case, the two oligosaccharides lie approximately at 90^{o} to one another with the reducing end of one

chain (in either Domain 1 or 3) closest to the non-reducing end of the other chain (in either Domain 2 or 4, respectively) (Figure 5.16 B (i and ii)).

B **i)**



ii)

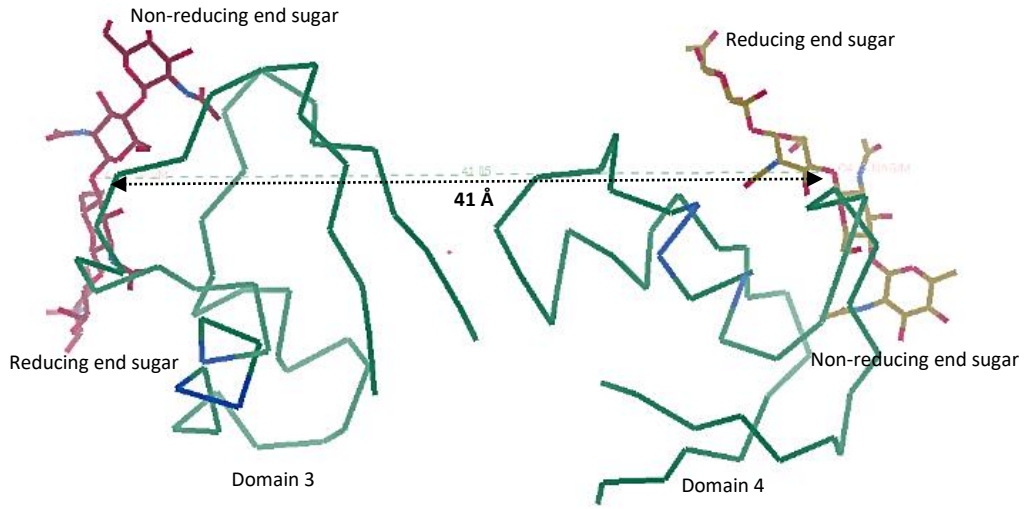


Figure 5.16 (B)

5.6.3 The orientation of LysM 2-LysM 4 and LysM 1-LysM 3

The two oligosaccharides associated with Domains 2 and 4 lie on opposite faces of the molecule and are inclined at an angle of ~120^{o} to each other being separated by ~45 Å. For these chains, the closest approach is between sugars at their non-reducing ends (Figure 5.16 B (iii)). In the case of the orientation of Domains 1 and 3, the two chains again lie on opposite faces of the monomer but in approximately the same plane being separated by ~55 Å in parallel orientation (Figure 5.16 B (iv)).

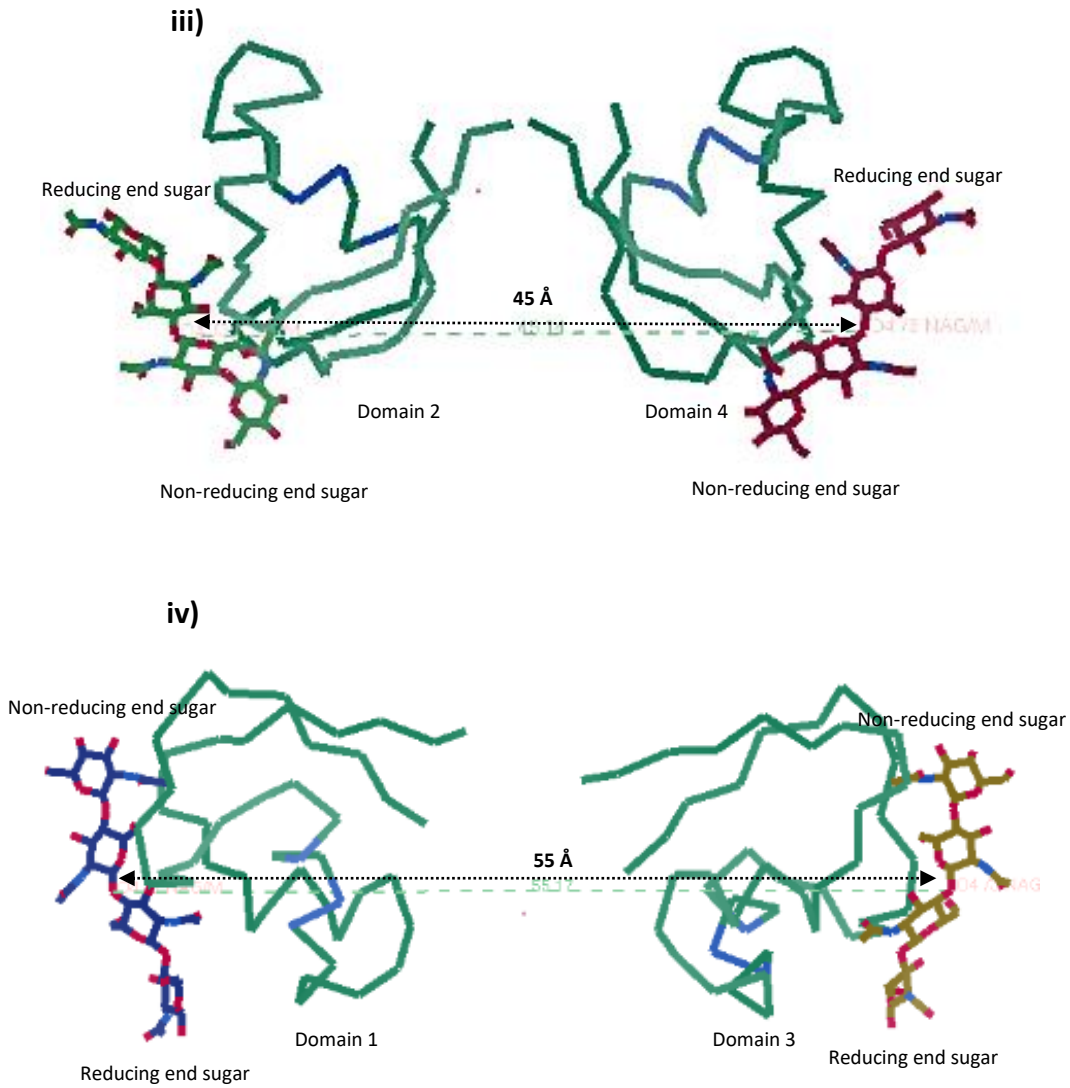


Figure 5.16 (B)

5.6.4 Insights into the separation of peptidoglycan chains in the cell wall

If some or all of the different LysM domains interact with the peptidoglycan layers in the cell wall at the same time, then the relative orientations of these domains on the KEG15107 monomer might reflect the actual distances and orientations separating different peptidoglycan strands in the cell wall. Thus, one possibility is that the chains might be separated by ~37 Å in an antiparallel orientation (consistent with the separation of Domains 1 and 4, or Domains 2 and 3) (Figure 5.16 C). Alternatively, some chains might be parallel and separated by a greater distance of ~55 Å (consistent with the separation of Domains 1 and 3). In other orientations, the chains would be inclined with respect to each other with different separations again. Currently, our understanding of the separation and orientation of peptidoglycan strands in the cell wall is incomplete. Data from electron microscopic analysis on the cell envelopes of *E. coli* by Braun and colleagues and others (see for example Braun *et al*., 1973) suggest that covalently linked glycan chains are probably separated by ~10 - 40 Å following an analysis of the length of short peptides that cross-link the peptidoglycan chains. However, whether these linked chains are simultaneously recognized by two LysM domains is not clear. However, this distance is compatible with the LysM domain separation seen in KEG15107. No restriction is placed on the separation of the glycan chains that are not covalently linked. Direct visualization of glycan chains in the peptidoglycan polymer of *E.coli* by AFM analysis suggests that the chains are spaced by ~27 Å (R. D. Turner, Mesnage, Hobbs, & Foster, 2018), a distance compatible with the EM analysis and with the separation of the difference LysM domains on KEG15107. However, whether these chains are recognized by multiple LysM domains remains an unresolved question.

The arrangement of the four LysM domains of KEG15107 combined with the pseudo-symmetry might suggest that different layers of covalently linked glycan chains could be bridged such that antiparallel covalently linked peptidoglycan chains in one layer could be recognized by two-fold related LysM domains (e.g: Domains 1 and 4) whilst the other two-fold related domains (Domain 2 and 3) could recognize antiparallel covalently liked peptidoglycan chains in a second adjacent layer. Clearly, as more data emerges from experiments elsewhere, it may become possible to support the findings of others on the separation and orientation of glycan chains using measurements of the

separation of the LysM domains from this and other structures combined with a modeling approach as described above.

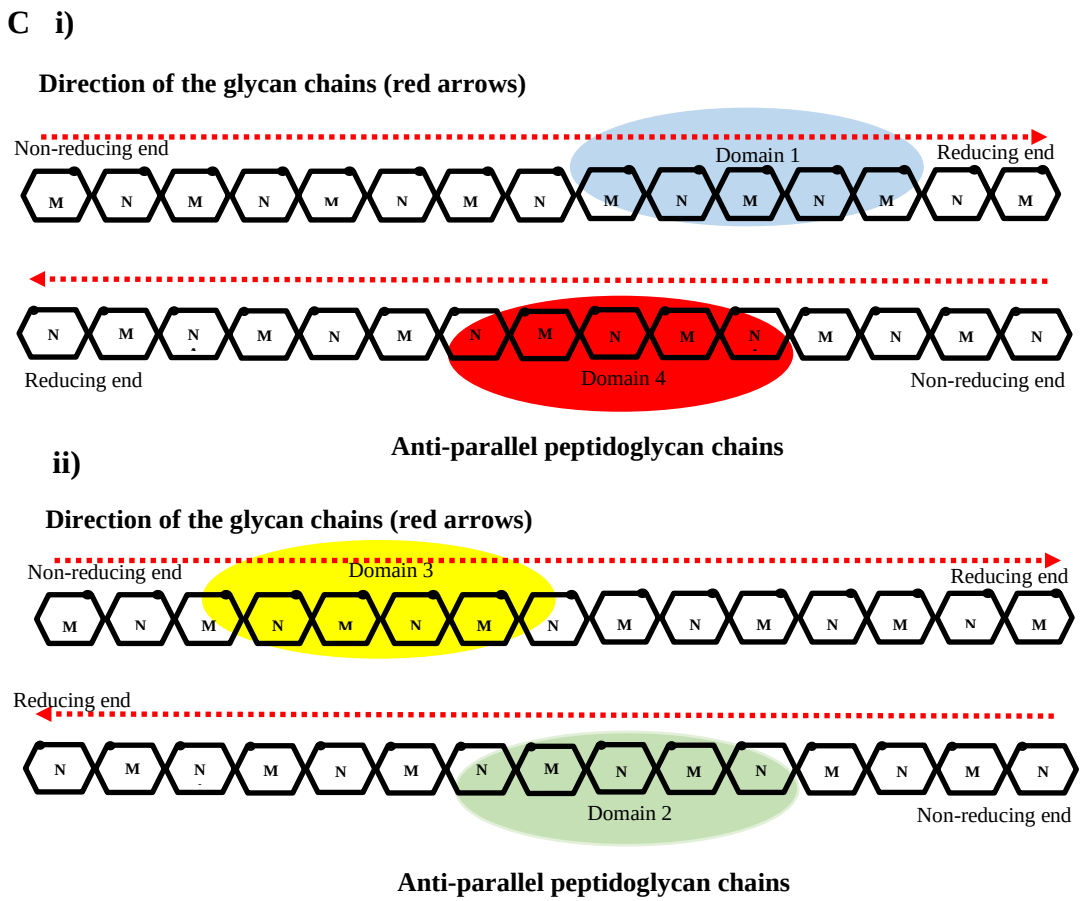


Figure 5.16: A proposed model of peptidoglycan layers based on the observation of oligosaccharide recognition by LysM domains of KEG15107 from *M. avium.* A) Center to center distance between the oligosaccharide binding sites on the different LysM domains. The distance is measured from the oxygen of glycosidic bonds between the sugar bound in the S2 and S3 sites. i) The oligosaccharide chains which bind to Domains 1 and 4 are in an antiparallel orientation as the two domains are related by local 2-fold symmetry. ii) The oligosaccharide chains which bind to Domains 2 and 3 again, are in an antiparallel orientation as the two domains are related by local 2-fold symmetry. B) The separation and orientation of oligosaccharide chains bound to different combinantion of the LysM domains. i and ii) Oligosaccharide chains bound to Domains 1 and 2, and Domains 3 and 4 are antiparallel. iii and iv) Oligosaccharide chains bound to Domains 2 and 4 are inclined at 120^{o}, and Domains 1 and 3 are approximately parallel. v) The separation between disaccharide units is 10.4 Å. C) The arrangement of multiple LysM domains of KEG15107 and the way each of the four domains recognize oligosaccharide suggests that antiparallel covalently linked peptidoglycan chains in one layer could be recognized by two-fold related LysM domains (e.g: Domains 1 and 4) whilst the other two-fold related domains (e.g:Domain 2 and 3) could recognize antiparallel covalently liked peptidoglycan chains in a second adjacent layer.

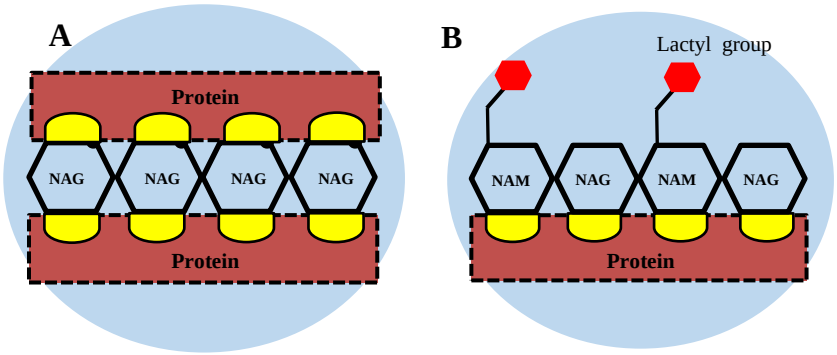


Figure 5.17: A comparison between the binding mode of polyNAG and polyNAG-NAM by LysM domains. A) LysM domains can bind to both sides of the polyNAG polymer as they present equivalent face to the protein. B) LysM domains cannot bind to both sides of polyNAG-NAM polymers as the NAM residues contain larger C3 lactyl groups on one face of the molecules and therefore binding on that side causes steric hindrance.

CHAPTER SIX

DISCUSSION AND FUTURE WORK

LysM domains are widely distributed

The LysM domain was initially found in a lysozyme-like enzyme from *Bacillus* phage phi 29 (Garvey et al., 1986). The LysM domain principally binds N-acetyl glucosamine (NAG) and N-acetyl muramic acid (NAM) sugars, major constituents of molecules such as peptidoglycan, chitin and other related molecules including lipochito-oligosaccharides such as the Nod-factor. This module is widely distributed in all domains of life (Steen et al., 2003, Radutoiu et al., 2003, Visweswaran, Dijkstra, & Kok, 2011, Visweswaran, Leenhouts, Van Roosmalen, Kok, & Buist, 2014, Kitaoku, Fukamizo, Numata, & Ohnuma, 2017). LysM domains frequently occur as part of multifunctional polypeptides that are commonly involved in aspects of bacterial cell wall synthesis and breakdown including hydrolases, amidases, peptidases and esterases as well as being associated with other domains such as protein kinases of plants and fungi (Vollmer et al., 2005, Buist, Steen, Kok, & Kuipers, 2008, Radutoiu et al., 2003). The role of the LysM module has been suggested to position the neighboring domains by anchoring the substrate thereby facilitating the interaction of the catalytic domains of the associated enzymes (Visweswaran et al., 2014). Through studies of the interactions between the LysM domains and its polysaccharide substrate, it might be possible to learn more about the molecular architecture of peptidoglycan itself, an understanding that is far from complete at the moment.

The four LysM domains of KEG15107 form a tightly packed structure

KEG15107 from *M. avium* is a LysM containing protein with four tandem LysM domains that, unusually, lacks any associated catalytic domains. As part of this thesis, the structure and function of this protein had been investigated to reveal that its four LysM domains form a protease-resistant globular structure. This

arrangement is in stark contrast to the beads on a string model suggested for the multiple LysM domains of AtlA which, like the LysM domains of KEG15107 do not contain disulphide bridges between the multiple domains (Mesnage et al., 2014). The formation of tightly packed LysM domains of KEG15107 is therefore similar to AtCERK1 protein, in which its three LysM domains closely interact with each other being linked through disulfide bonds (Liu et al., 2012) and to *C. fulvum* Ecp6, a protein with three LysM domains that are also closely packed together and stabilized by disulfide bonds (Sánchez-Vallet et al., 2013). The findings from this thesis therefore suggest that the association of the LysM domains into tightly packed structures does not require disulphide bonds suggesting that the beads on a string model of AtlA is not universal and indeed, in the absence of the structure of AtlA more data is needed to clarify how the domains in the latter protein are connected.

KEG15107 forms different quaternary structures in equilibrium with each other

The structures of the apo proteins of both KEG15107 and MSMEG3288 have a tetrameric quaternary structure with a total of 16 LysM domains whereas the complexes with oligosaccharides are dimeric as a result of adverse steric interactions with the oligosaccharide bound to Domain 1 interfering with the association of these domains in the tetramer. Moreover, consideration of the structure of the tetramer suggested that the two dimers would have to separate to allow access to the sugar binding pockets on Domain 1. This suggests that the dimer is probably in equilibrium with the tetramer in solution. Initial investigation of this possible equilibria by mass spectrometry on the apo protein surprisingly revealed the presence of monomers, dimers, and tetramers suggesting the assembly of the protein is highly dynamic and prompting further investigation of KEG15107 and its complexes with oligosaccharides.

All LysM domains of KEG15107 binds oligosaccharide

The observation from the crystallographic analysis that only one of the three LysM domains of AtCERK1 binds oligosaccharide is similar to the behavior of KEG15107 and its homolog of MSMEG3288. At first, this suggested that the

additional LysM domains might not be involved in oligosaccharide recognition. However, whilst in the presence of low polyNAG substrate concentration (0.1mM) the protein appears to be mainly monomeric or dimeric, there was clear evidence for the binding of multiple oligosaccharide chains to the protein. Further studies showed as the concentration of sugar increased the protein was commonly seen as a monomer with an increasing proportion of the molecules having sugar bound and mass spectrometry analysis showed that major species corresponding to the binding of one, two, three or four sugar chains to the monomer of KEG15107 could be identified. Moreover, minor species could be seen which suggested that each of the sugar binding sites on the individual LysM domains could be occupied by two sugar chains. Taken together, these results strongly suggest that all of the LysM domains can bind oligosaccharides and the observation from some crystallographic studies that only one of the multiple domains might bind oligosaccharide is a misleading artefact presumably reflecting the difficulties of obtaining crystals with multiple oligosaccharides bound as a result of the heterogenous mixture of species in solution (mixtures of different domains with or without sugar bound) combined with issues relating to the formation of favourable crystal contacts. Equally, when soaking substrates to pre-existing crystals, problems might arise where conformational changes might not be possible in the crystal lattice or where the binding sites might be blocked by neighboring molecules in the crystals. In both these cases, whether co-crystallising or soaking to obtain substrate complexes, the different affinity of the domains for their oligosaccharides could mean that the concentration of the sugar necessary to saturate the sites might be difficult to achieve. More generally, the experiments undertaken using mass spectrometry to analyze the binding of oligosaccharides to other LysM containing proteins including to Trc1 (the three LysM domain segment of the Rv1288 sequence) and to full-length YgaU, which contains a single LysM domain, provide a similar picture to that shown with regards to oligosaccharide binding to KEG15107.

In general, the binding affinity of sugar residues to proteins is generally believed to be low (in the millimolar range), but the affinity increases when the binding involves several sugar residues to the protein (Rudd, Wormald, & Dwek, 2004).

Clearly, the possibility of multiple interactions between different oligosaccharides chains and proteins could well be a factor for triggering biological function. The presence of multiple LysM domains in KEG15107 and other proteins containing multiple LysM domains might enhance oligosaccharide recognition and thus increase the binding affinity of the proteins towards the sugar molecules facilitating activities such as peptidoglycan synthesis, hydrolysis or modification and alterations to the number of LysM modules in a protein might reduce the efficiency of the enzymatic activities of the protein as it can be seen in the N-acetylglucosaminidase in *Lactococcus lactis.* As the number of the LysM domains of the protein was trunccated, the enzyme activities was significantly regressed (Steen et al., 2005).

Molecular recognition of oligosaccharides and protein specificity

Five sugar-binding sites (S0, S1, S2, S3, and S4) could be identified on Domain 1 of KEG15107. Extensive interactions between the N-acetyl groups of the sugars N1 and N3 with the main chains and side chains of protein residues at sites S1 and S3 on the L1 and L2 loops of Domain 1 suggest that these sites exclusively recognize N-acetyl glucosamine molecules and thus discriminate the NAG and NAM residues which are components of peptidoglycan. These observations are similar to those noted in the oligosaccharide binding pockets in the plant LysM 2 of AtCERK1 (Liu et al., 2012) as well as fungal LysMs of Ecp6 (Sánchez-Vallet et al., 2013) and bacterial LysM of AtlA (Mesnage et al., 2014) and NlpC/p60 (Wong et al., 2015).

In the case of KEG15107, the aromatic ring of Phe13 moves significantly from the position it occupies in the apo protein to allow the sugar N4 to be accommodated in the binding site of Domain 1. This finding is also observed in MSMEG3288 (Personal communication with C. Bisson and DWR). In the case of the NlpC/P60 endopeptidase of *T.thermophilus* and PrChi-A from *Pteris ryukyuensi,* an aromatic side chain in an equivalent position (Y28 and Y72, respectively) is also involved in a conformational change on oligosaccharide binding (Wong et al., 2015) (Ohnuma, Onaga, Murata, Taira, & Katoh, 2008a).

Insights into peptidoglycan architecture from studies of LysM domains

Unlike structural studies on protein or nucleic acid our understanding of the structure and architecture of oligosaccharide including those in the cell wall is far from complete. The confirmation in studies presented here that in proteins containing multiple LysM domains each of the domains can bind oligosaccharide chains to suggest that they might act together and that level the separation and relative orientation of their oligosaccharide binding sites might relate the underline separation of oligosaccharide chains in the cell wall. Clearly, the most interesting finding from the structure is that binding sites of some of the domains in KEG15107 are separated by ~40 Å and in an antiparallel orientation. It remains to be seen if this geometry exists for peptidoglycan chains in the cell wall and whether studies such as those described here, could lead to a deeper understanding of the cell wall and to the design of inhibitors that might be used as novel antimicrobial agents.

Future work

Site directed mutagenesis on the KEG15107 protein is worth to try in order to further investigate its binding affinity towards oligosaccharide. The key residues of the protein that make direct hydrogen bonds through their side chains to the hydroxyl groups or to the pyranose ring of the oligosaccharide could be attempted by replacing the residues with alanine. The mutated protein with and without oligosaccharide then could be subjected to Isothermal titration calorimetry (ITC) or Microscale thermophoresis (MST) to determine dissosiation and association activities of the complex (Kd).

Investigating the localization of the KEG15107 protein could be attempted too as it will provide information of the distribution of the protein in the bacterial cells. The distribution of the protein provides insight about the biological function of the protein and how important the protein is for the cells. A protein fusion technique with GFP would be the best attemption in which the expressed

KEG15107 protein is tagged with green fluorescent that be simply observed under fluorescent microscope.

References

Andersson, K. & Hovmöller, S. (n.d.). The average atomic volume and density of proteins. *Crystalline Materials 1998*. Retrieved from doi:10.1524/zkri.1998.213.7-8.369.

Ashraf, K. U., Josts, I., Mosbahi, K., Kelly, S. M., Byron, O., Smith, B. O., & Walker, D. (2016). The Potassium Binding Protein Kbp Is a Cytoplasmic Potassium Sensor. *Structure*, *24*(5), 741–749. https://doi.org/10.1016/j.str.2016.03.017Andersson, K. & Hovmöller, S. (n.d.). The average atomic volume and density of proteins. *Crystalline Materials 1998*. Retrieved from doi:10.1524/zkri.1998.213.7-8.369

Ashraf, K. U., Josts, I., Mosbahi, K., Kelly, S. M., Byron, O., Smith, B. O., & Walker, D. (2016). The Potassium Binding Protein Kbp Is a Cytoplasmic Potassium Sensor. *Structure*, *24*(5), 741–749. https://doi.org/10.1016/j.str.2016.03.017

Bateman, A., & Bycroft, M. (2000a). The structure of a LysM domain from E. coli membrane-bound lytic murein transglycosylase D (MltD) 1 1Edited by P. E. Wight. *Journal of Molecular Biology*, *299*(4), 1113–1119. https://doi.org/10.1006/jmbi.2000.3778

Bateman, A., & Bycroft, M. (2000b). The structure of a LysM domain from E. coli membrane-bound lytic murein transglycosylase D (MltD)1. *Journal of Molecular Biology*, *299*(4), 1113–1119. https://doi.org/http://dx.doi.org/10.1006/jmbi.2000.3778

Battye, T. G. G., Kontogiannis, L., Johnson, O., Powell, H. R., & Leslie, A. G. W. (2011). iMOSFLM: A new graphical interface for diffraction-image processing with MOSFLM. *Acta Crystallographica Section D: Biological Crystallography*, *67*(4), 271–281. https://doi.org/10.1107/S0907444910048675

Baum, E. Z., Montenegro, D. A., Licata, L., Turchi, I., Webb, G. C., Foleno, B. D., & Bush, K. (2001). Identification and characterization of new inhibitors of the Escherichia coli MurA enzyme. *Antimicrobial Agents and Chemotherapy*, *45*(11), 3182–3188. https://doi.org/10.1128/AAC.45.11.3182-3188.2001

Béliveau, C., Potvin, C., Trudel, J., Asselin, a, & Bellemare, G. (1991). Cloning, sequencing, and expression in Escherichia coli of a Streptococcus faecalis autolysin. *Journal of Bacteriology*, *173*(18), 5619–5623.

Berrazeg, M., Diene, S., Parola, P., Drissi, M., Raoult, D., & JM Rol. (2012). *New Dehli*

metallo-beta-lactamase around the world: An eReview using Google Maps (Poster). (3), 7278.

Bolton, M. D., Van Esse, H. P., Vossen, J. H., De Jonge, R., Stergiopoulos, I., Stulemeijer, I. J. E., … Thomma, B. P. H. J. (2008). The novel Cladosporium fulvum lysin motif effector Ecp6 is a virulence factor with orthologues in other fungal species. *Molecular Microbiology*, *69*(1), 119–136. https://doi.org/10.1111/j.1365-2958.2008.06270.x

Boraston, A. B., Bolam, D. N., & Gilbert, Harry J.Davies, G. J. (2004). Carbohydrate-binding modules: fine-tuning polysaccharide recognition. *Biochemical Journal*, *382*(3), 769–781. https://doi.org/10.1042/bj20040892

Buist, G., Steen, A., Kok, J., & Kuipers, O. P. (2008a). LysM, a widely distributed protein motif for binding to (peptido)glycans. *Molecular Microbiology*, *68*(4), 838–847. https://doi.org/10.1111/j.1365-2958.2008.06211.x

Buist, G., Steen, A., Kok, J., & Kuipers, O. P. (2008b). LysM, a widely distributed protein motif for binding to (peptido)glycans. *Molecular Microbiology*, *68*(4), 838–847. https://doi.org/10.1111/j.1365-2958.2008.06211.x

Cardona, P. J. (2018). Pathogenesis of tuberculosis and other mycobacteriosis. *Enfermedades Infecciosas y Microbiologia Clinica*, *36*(1), 38–46. https://doi.org/10.1016/j.eimc.2017.10.015

Carotenuto, G., Chabaud, M., Miyata, K., Capozzi, M., Takeda, N., Kaku, H., … Genre, A. (2017). The rice LysM receptor-like kinase OsCERK1 is required for the perception of short-chain chitin oligomers in arbuscular mycorrhizal signaling. *New Phytologist*, *214*(4), 1440–1446. https://doi.org/10.1111/nph.14539

Chan, A. P., Choi, Y., Brinkac, L. M., Krishnakumar, R., DePew, J., Kim, M., … Fouts, D. E. (2018). Multidrug resistant pathogens respond differently to the presence of co-pathogen, commensal, probiotic and host cells. *Scientific Reports*, *8*(1), 1–12. https://doi.org/10.1038/s41598-018-26738-1

Chen, L., Walker, D., Sun, B., Hu, Y., Walker, S., & Kahne, D. (2003). Vancomycin analogues active against vanA-resistant strains inhibit bacterial transglycosylase without binding substrate. *Proceedings of the National Academy of Sciences*, *100*(10), 5658–5663. https://doi.org/10.1073/pnas.0931492100

Christopher Walsh, & Timothy, and W. (n.d.). *Antibiotics: Challenges, Mechanisms, Opportunities* (2016th ed.). American Society for Microbiology Press.

Crick, D. C., Mahapatra, S., & Brennan, P. J. (2001). Biosynthesis of the

arabinogalactan-peptidoglycan complex of Mycobacterium tuberculosis. *Glycobiology*, *11*(9), 107–118. https://doi.org/10.1093/glycob/11.9.107R

Dauter, Z. (1999). Data collection strategies. *Acta Crysatallography*, *59*, 1703–1717. https://doi.org/10.1007/978-3-319-02078-5_4

Davis, I. W., Leaver-Fay, A., Chen, V. B., Block, J. N., Kapral, G. J., Wang, X., … Richardson, D. C. (2007). MolProbity: All-atom contacts and structure validation for proteins and nucleic acids. *Nucleic Acids Research*, *35*(SUPPL.2), 375–383. https://doi.org/10.1093/nar/gkm216

de Jonge, R., Peter van Esse, H., Kombrink, A., Shinya, T., Desaki, Y., Bours, R., … Thomma, B. P. H. J. (2010). Conserved Fungal LysM Effector Ecp6 Prevents Chitin-Triggered Immunity in Plants. *Science*, *329*(5994), 953–955. https://doi.org/10.1126/science.1190859

Desvaux, M., Dumas, E., Chafsey, I., & Hébraud, M. (2006a). Protein cell surface display in Gram-positive bacteria: From single protein to macromolecular protein structure. *FEMS Microbiology Letters*, *256*(1), 1–15. https://doi.org/10.1111/j.1574-6968.2006.00122.x

Desvaux, M., Dumas, E., Chafsey, I., & Hébraud, M. (2006b). Protein cell surface display in Gram-positive bacteria: From single protein to macromolecular protein structure. *FEMS Microbiology Letters*. https://doi.org/10.1111/j.1574-6968.2006.00122.x

Dirk-Jan Scheffers, M. G. P. (2005). Bacterial Cell Wall Synthesis: New Insights from Localization Studies. *ASM*, *69*(4), 585–607.

Dodson, E. J., Winn, M., & Ralph, A. (1997). [32] Collaborative computational project, number 4: Providing programs for protein crystallography. *Methods in Enzymology*, *277*, 620–633. https://doi.org/10.1016/S0076-6879(97)77034-4

Dreisbach, A., Dijl, J. M. Van, & Buist, G. (2011). *The cell surface proteome of Staphylococcus aureus*. 3154–3168. https://doi.org/10.1002/pmic.201000823

Eckert, C., Lecerf, M., Dubost, L., Arthur, M., & Mesnage, S. (2006). Functional analysis of AtlA, the major N-acetylglucosaminidase of Enterococcus faecalis. *Journal of Bacteriology*, *188*(24), 8513–8519. https://doi.org/10.1128/JB.01145-06

Emsley, P., & Cowtan, K. (2004). Coot: Model-building tools for molecular graphics. *Acta Crystallographica Section D: Biological Crystallography*, *60*(12 I), 2126–2132. https://doi.org/10.1107/S0907444904019158

Evans, P. R. (2011). An introduction to data reduction: Space-group determination, scaling and intensity statistics. *Acta Crystallographica Section D: Biological Crystallography*, *67*(4), 282–292. https://doi.org/10.1107/S090744491003982X

Evans, P. R., & Murshudov, G. N. (2013). How good are my data and what is the resolution? *Acta Crystallographica Section D: Biological Crystallography*, *69*(7), 1204–1214. https://doi.org/10.1107/S0907444913000061

Finn, R. D., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Mistry, J., Mitchell, A. L., … Bateman, A. (2016). The Pfam protein families database: Towards a more sustainable future. *Nucleic Acids Research*, *44*(D1), D279–D285. https://doi.org/10.1093/nar/gkv1344

Fisher, S. J., Levik, K. E., Williams, M. A., Ashton, A. W., & McAuley, K. E. (2015). SynchWeb : a modern interface for ISPyB . *Journal of Applied Crystallography*, *48*(3), 927–932. https://doi.org/10.1107/s1600576715004847

Ford, N., Matteelli, A., Shubber, Z., Hermans, S., Meintjes, G., Grinsztejn, B., … Getahun, H. (2016). TB as a cause of hospitalization and in-hospital mortality among people living with HIV worldwide: A systematic review and meta-analysis. *Journal of the International AIDS Society*, *19*(1), 1–5. https://doi.org/10.7448/IAS.19.1.20714

Frederick C. Neidhardt, John L. Ingraham, M. S. (1990). *Physiology of the Bacterial Cell*.

Garvey, K. J., Saedi, M. S., & Ito, J. (1986). Nucleotide sequence of Bacillus phage ??29 genes 14 and 15: Homology of gene 15 with other phage lysozymes. *Nucleic Acids Research*, *14*(24), 10001–10008. https://doi.org/10.1093/nar/14.24.10001

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. *Nucleic Acids Research*, *31*(13), 3784–3788. https://doi.org/10.1093/nar/gkg563

Gautam, A., Rishi, P., & Tewari, R. (2011). UDP-N-acetylglucosamine enolpyruvyl transferase as a potential target for antibacterial chemotherapy: Srecent developments. *Applied Microbiology and Biotechnology*, *92*(2), 211–225. https://doi.org/10.1007/s00253-011-3512-z

Getahun, H., Matteelli, A., Chaisson, R. E., & Raviglione, M. (2015). Latent *Mycobacterium tuberculosis* Infection. *New England Journal of Medicine*, *372*(22), 2127–2135. https://doi.org/10.1056/NEJMra1405427

Gómez, M. I., O'Seaghdha, M., Magargee, M., Foster, T. J., & Prince, A. S. (2006). Staphylococcus aureus protein A activates TNFR1 signaling through conserved IgG binding domains. *Journal of Biological Chemistry*, *281*(29), 20190–20196. https://doi.org/10.1074/jbc.M601956200

Gust, A. A., Willmann, R., Desaki, Y., Grabherr, H. M., & Nürnberger, T. (2012). Plant LysM proteins: Modules mediating symbiosis and immunity. *Trends in Plant Science*, Vol. 17, pp. 495–502. https://doi.org/10.1016/j.tplants.2012.04.003

Hayafune, M., Berisio, R., Marchetti, R., Silipo, A., Kayama, M., Desaki, Y., … Shibuya, N. (2014). Chitin-induced activation of immune signaling by the rice receptor CEBiP relies on a unique sandwich-type dimerization. *Proceedings of the National Academy of Sciences*, *111*(3), E404–E413. https://doi.org/10.1073/pnas.1312099111

Healing, D. E. (2009). Bacteria in Biology, Biotechnology and Medicine, 4th edition. In *Journal of Medical Microbiology* (Vol. 47). https://doi.org/10.1099/00222615-47-4-369b

Heijenoort, J. v. (2001). Formation of the glycan chains in the synthesis of bacterial peptidoglycan. *Glycobiology*, *11*(3), 25R-36R. https://doi.org/10.1093/glycob/11.3.25R

Heikinheimo, P., Goldman, A., Jeffries, C., & Ollis, D. L. (1999). *Of barn owls and bankers: a lush variety of α/β hydrolases*. 141–146.

Hett, E. C., & Rubin, E. J. (2008). Bacterial Growth and Cell Division: a Mycobacterial Perspective. *Microbiology and Molecular Biology Reviews*, *72*(1), 126–156. https://doi.org/10.1128/MMBR.00028-07

Hsu, C.-Y., Wu, C.-W., & Talaat, A. M. (2011). Genome-Wide Sequence Variation among Mycobacterium avium Subspecies paratuberculosis Isolates: A Better Understanding of Johne?s Disease Transmission Dynamics. *Frontiers in Microbiology*, *2*(December), 1–14. https://doi.org/10.3389/fmicb.2011.00236

Igor Tvaroska, T. B. (1989). Anomeric and Exo-Anomeric Effects in Carbohydrate Chemistry. *Advances in Carbohydrate Chemostry and Biochemistry*, 48–119.

Imberty, A., Delage, M. M., Bourne, Y., Cambillau, C., & Pérez, S. (1991). Data bank of three-dimensional structures of disaccharides: Part II, N-acetyllactosaminic type N-glycans. Comparison with the crystal structure of a biantennary octasaccharide. *Glycoconjugate Journal*, *8*(6), 456–483.

https://doi.org/10.1007/BF00769847

Jerse, a E., Yu, J., Tall, B. D., & Kaper, J. B. (1990). A genetic locus of enteropathogenic Escherichia coli necessary for the production of attaching and effacing lesions on tissue culture cells. *Proceedings of the National Academy of Sciences of the United States of America*, *87*(20), 7839–7843. https://doi.org/10.1097/00005176-199208000-00024

Jin, B. S., Han, S. G., Lee, W. K., Ryoo, S. W., Lee, S. J., Suh, S. W., & Yu, Y. G. (2009). Inhibitory mechanism of novel inhibitors of UDP-N-acetylglucosamine enolpyruvyl transferase from Haemophilus influenzae. *Journal of Microbiology and Biotechnology*, *19*(12), 1582–1589. https://doi.org/10.4014/jmb.0905.05036

Kantardjieff, K. A., & Rupp, B. (2003). Matthews coefficient probabilities: Improved estimates for unit cell contents of proteins, DNA, and protein-nucleic acid complex crystals. *Protein Science*, *12*(9), 1865–1871. https://doi.org/10.1110/ps.0350503

Katarina Mikusova, Richard A Slayden, G. S. B. and P. J. B. (1995). *Biogenesis of the Mycobacterial Cell Wall and the Site of Action of Ethambutol AND*. *39*(970), 2484–2489.

Kaufmann, S. H. E., Weiner, J., & von Reyn, C. F. (2017). Novel approaches to tuberculosis vaccine development. *International Journal of Infectious Diseases*, *56*, 263–267. https://doi.org/10.1016/j.ijid.2016.10.018

Kibirige, D., Ssekitoleko, R., Mutebi, E., & Worodria, W. (2013). Overt diabetes mellitus among newly diagnosed Ugandan tuberculosis patients: a cross sectional study. *BMC Infectious Diseases*, *13*(1), 122. https://doi.org/10.1186/1471-2334-13-122

Kitaoku, Y., Fukamizo, T., Numata, T., & Ohnuma, T. (2017). Chitin oligosaccharide binding to the lysin motif of a novel type of chitinase from the multicellular green alga, Volvox carteri. *Plant Molecular Biology*, *93*(1–2), 97–108. https://doi.org/10.1007/s11103-016-0549-5

Koharudin, L. M. I., Debiec, K. T., & Gronenborn, A. M. (2015). Structural Insight into Fungal Cell Wall Recognition by a CVNH Protein with a Single LysM Domain. *Structure*, *23*(11), 2143–2154. https://doi.org/10.1016/j.str.2015.07.023

Kraulis, P. J. (2002). MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. *Journal of Applied Crystallography*, *24*(5), 946–950. https://doi.org/10.1107/s0021889891004399

Lebedev, A. A., Young, P., Isupov, M. N., Moroz, O. V., Vagin, A. A., & Murshudov,

G. N. (2012). JLigand: A graphical tool for the CCP4 template-restraint library. *Acta Crystallographica Section D: Biological Crystallography*, *68*(4), 431–440. https://doi.org/10.1107/S090744491200251X

Lee, M., Hesek, D., Llarrull, L. I., Lastochkin, E., Pi, H., Boggess, B., & Mobashery, S. (2013). Reactions of all escherichia coli lytic transglycosylases with bacterial cell wall. *Journal of the American Chemical Society*, *135*(9), 3311–3314. https://doi.org/10.1021/ja309036q

Lenz, L. L., Mohammadi, S., Geissler, A., & Portnoy, D. A. (2003). SecA2-dependent secretion of autolytic enzymes promotes Listeria monocytogenes pathogenesis. *Proc Natl Acad Sci U S A*, *100*(21), 12432–12437. https://doi.org/10.1073/pnas.2133653100\r2133653100 [pii]

Liu, T., Liu, Z., Song, C., Hu, Y., Han, Z., She, J., … Chai, J. (2012). Chitin-Induced Dimerization Activates a Plant Immune Receptor. *Science*, *336*(6085), 1160–1164. https://doi.org/10.1126/science.1218867

Lovering, A. L., Safadi, S. S., & Strynadka, N. C. J. (2012). Structural Perspective of Peptidoglycan Biosynthesis and Assembly. *Annual Review of Biochemistry*, *81*(1), 451–478. https://doi.org/10.1146/annurev-biochem-061809-112742

Low, L. Y., Yang, C., Perego, M., Osterman, A., & Liddington, R. (2011). Role of net charge on catalytic domain and influence of cell wall binding domain on bactericidal activity, specificity, and host range of phage lysins. *Journal of Biological Chemistry*, *286*(39), 34391–34403. https://doi.org/10.1074/jbc.M111.244160

Maan, P., Kumar, A., Kaur, J., & Kaur, J. (2018). Rv1288, a Two Domain, Cell Wall Anchored, Nutrient Stress Inducible Carboxyl-Esterase of Mycobacterium tuberculosis, Modulates Cell Wall Lipid. *Frontiers in Cellular and Infection Microbiology*, *8*(December), 1–16. https://doi.org/10.3389/fcimb.2018.00421

Mahapatra, S., Yagi, T., Belisle, J. T., Espinosa, B. J., Hill, P. J., McNeil, M. R., … Crick, D. C. (2005). Mycobacterial lipid II is composed of a complex mixture of modified muramyl and peptide moieties linked to decaprenyl phosphate. *Journal of Bacteriology*, *187*(8), 2747–2757. https://doi.org/10.1128/JB.187.8.2747-2757.2005

Malanovic, N., & Lohner, K. (2016). Gram-positive bacterial cell envelopes: The impact on the activity of antimicrobial peptides. *Biochimica et Biophysica Acta - Biomembranes*, *1858*(5), 936–946.

https://doi.org/10.1016/j.bbamem.2015.11.004

Matias, V. R. F., Al-amoudi, A., Dubochet, J., & Beveridge, T. J. (2003). Cryo-Transmission Electron Microscopy of Frozen-Hydrated Sections of Escherichia coli and Pseudomonas aeruginosa Cryo-Transmission Electron Microscopy of Frozen-Hydrated Sections of Escherichia coli and Pseudomonas aeruginosa. *Journal of Bacteriology*, *185*(20), 6112–6118. https://doi.org/10.1128/JB.185.20.6112

Matthews, B. W. (1968). Solvent content of protein crystals. *Journal of Molecular Biology*, *33*(2), 491–497. https://doi.org/10.1016/0022-2836(68)90205-2

McCoy, A. J., Grosse-Kunstleve, R. W., Adams, P. D., Winn, M. D., Storoni, L. C., & Read, R. J. (2007). Phaser crystallographic software . *Journal of Applied Crystallography*, *40*(4), 658–674. https://doi.org/10.1107/s0021889807021206

Mcguffin, L. J., Bryson, K., & Jones, D. T. (2000). *The PSIPRED protein structure prediction server*. *16*(4), 404–405.

McRee, D. E. (1999). XtalView/Xfit - A versatile program for manipulating atomic coordinates and electron density. *Journal of Structural Biology*, *125*(2–3), 156–165. https://doi.org/10.1006/jsbi.1999.4094

Meroueh, S. O., Bencze, K. Z., Hesek, D., Lee, M., Fisher, J. F., Stemmler, T. L., & Mobashery, S. (2006). Three-dimensional structure of the bacterial cell wall peptidoglycan. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(12), 4404–4409. https://doi.org/10.1073/pnas.0510182103

Mesnage, S., Dellarole, M., Baxter, N. J., Rouget, J.-B., Dimitrov, J. D., Wang, N., … Williamson, M. P. (2014). Molecular basis for bacterial peptidoglycan recognition by LysM domains. *Nature Communications*, *5*. https://doi.org/10.1038/ncomms5269

Miyata, K., Kozaki, T., Kouzai, Y., Ozawa, K., Ishii, K., Asamizu, E., … Nakagawa, T. (2014). The bifunctional plant receptor, OsCERK1, regulates both chitin-triggered immunity and arbuscular mycorrhizal symbiosis in rice. *Plant and Cell Physiology*, *55*(11), 1864–1872. https://doi.org/10.1093/pcp/pcu129

Moellering, R. C. (2012). MRSA: the first half century. *Journal of Antimicrobial Chemotherapy*, *67*(1), 4–11. https://doi.org/10.1093/jac/dkr437

Murshudov, G. N., Skubák, P., Lebedev, A. A., Pannu, N. S., Steiner, R. A., Nicholls, R. A., … Vagin, A. A. (2011). REFMAC5 for the refinement of macromolecular crystal structures. *Acta Crystallographica. Section D, Biological Crystallography*,

67(Pt 4), 355–367. https://doi.org/10.1107/S0907444911001314

Neuberger A & Deenen V. (1994). Bacterial Cell Wall. *ELSEVIER*, pp. 263–274. https://doi.org/10.1016/b978-0-12-134350-7.50031-9

Nikolaidis, I., Favini-Stabile, S., & Dessen, A. (2014). Resistance to antibiotics targeted to the bacterial cell wall. *Protein Science*, *23*(3), 243–259. https://doi.org/10.1002/pro.2414

Numbering, A. (1983). Symbols for Specifying the Conformation of Polysaccharide Chains: Recommendations 1981. *European Journal of Biochemistry*, *131*(1), 5–7. https://doi.org/10.1111/j.1432-1033.1983.tb07224.x

Ohnuma, T., Onaga, S., Murata, K., Taira, T., & Katoh, E. (2008a). LysM Domains from Pteris ryukyuensis Chitinase-A . *Journal of Biological Chemistry*, *283*(8), 5178–5187. https://doi.org/10.1074/jbc.m707156200

Ohnuma, T., Onaga, S., Murata, K., Taira, T., & Katoh, E. (2008b). LysM domains from Pteris ryukyuensis chitinase-A: A stability study and characterization of the chitin-binding site. *Journal of Biological Chemistry*, *283*(8), 5178–5187. https://doi.org/10.1074/jbc.M707156200

Pace, J. L., & Yang, G. (2006). Glycopeptides: Update on an old successful antibiotic class. *Biochemical Pharmacology*, *71*(7), 968–980. https://doi.org/10.1016/j.bcp.2005.12.005

Percudani, R., Montanini, B., & Ottonello, S. (2005). The anti-HIV cyanovirin-N domain is evolutionarily conserved and occurs as a protein module in eukaryotes. *Proteins: Structure, Function and Genetics*, *60*(4), 670–678. https://doi.org/10.1002/prot.20543

Perlstein, D. L., Zhang, Y., Wang, T. S., Kahne, D. E., & Walker, S. (2007). The direction of glycan chain elongation by peptidoglycan glycosyltransferases. *Journal of the American Chemical Society*, *129*(42), 12674–12675. https://doi.org/10.1021/ja075965y

Petutschnig, E. K., Jones, A. M. E., Serazetdinova, L., Lipka, U., & Lipka, V. (2010). The Lysin Motif Receptor-like Kinase (LysM-RLK) CERK1 is a major chitin-binding protein in Arabidopsis thaliana and subject to chitin-induced phosphorylation. *Journal of Biological Chemistry*, *285*(37), 28902–28911. https://doi.org/10.1074/jbc.M110.116657

Queiroz, A., & Riley, L. W. (2017). Bacterial immunostat: Mycobacterium tuberculosis lipids and their role in the host immune response. *Revista Da Sociedade Brasileira*

de Medicina Tropical, *50*(1), 9–18. https://doi.org/10.1590/0037-8682-0230-2016

Radutoiu, S., Madsen, L. H., Madsen, E. B., Felle, H. H., Umehara, Y., Grønlund, M., … Stougaard, J. (2003). Plant recognition of symbiotic bacteria requires two LysM receptor-like kinases. *Nature*, *425*, 585–592. https://doi.org/10.1038/nature02039

Rhodes, G. (1993a). An Overview of Protein Crystallography. In *Crystallography Made Crystal Clear* (pp. 5–27). https://doi.org/10.1016/B978-0-12-587075-7.50006-0

Rhodes, G. (1993b). Crystallography Made Crystal Clear. In *Acedemic Press INC*. https://doi.org/10.1016/0307-4412(94)90199-6

Richard J Reece. (2004). *Analysis of Genes and Genomes*. Willey Blackwell.

Rudd, P. M., Wormald, M. R., & Dwek, R. A. (2004). Sugar-mediated ligand-receptor interactions in the immune system. *Trends in Biotechnology*, *22*(10), 524–530. https://doi.org/10.1016/j.tibtech.2004.07.012

Ruhland, G. J., Hellwig, M., Wanner, G., & Fiedler, F. (1993). Cell-surface location of Listeria-specific protein p60--detection of Listeria cells by indirect immunofluorescence. *Journal of General Microbiology*, *139*(3), 609–616. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8473866

Sánchez-Vallet, A., Saleem-Batcha, R., Kombrink, A., Hansen, G., Valkenburg, D. J., Thomma, B. P. H. J., & Mesters, J. R. (2013). Fungal effector Ecp6 outcompetes host immune receptor for chitin binding through intrachain LysM dimerization. *ELife*, *2013*(2). https://doi.org/10.7554/eLife.00790

Schneewind, O., & Missiakas, D. M. (2012). Protein secretion and surface display in Gram-positive bacteria. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1592), 1123–1139. https://doi.org/10.1098/rstb.2011.0210

Seeliger, D., & de Groot, B. L. (2010). Ligand docking and binding site analysis with PyMOL and Autodock/Vina. *Journal of Computer-Aided Molecular Design*, *24*(5), 417–422. https://doi.org/10.1007/s10822-010-9352-6

Shi, X. Z., Zhou, J., Lan, J. F., Jia, Y. P., Zhao, X. F., & Wang, J. X. (2013). A Lysin motif (LysM)-containing protein functions in antibacterial responses of red swamp crayfish, Procambarus clarkii. *Developmental and Comparative Immunology*, *40*(3–4), 311–319. https://doi.org/10.1016/j.dci.2013.03.011

Siew, N., Saini, H. K., & Fischer, D. (2005). A putative novel alpha/beta hydrolase ORFan family in Bacillus. *FEBS Letters*, *579*(14), 3175–3182.

https://doi.org/10.1016/j.febslet.2005.04.030

Singleton, P. (n.d.). *Bacteria in Biology, Biotechnology and Medicine* (6th Editio).

Stagg HR, Lipman MC, M. T. and J. H. (2017). Isoniazid resistant tuberculosis: a cause for concern. *Int J Tuberc Lung Dis*, *21*(2), 129–139. https://doi.org/10.5588/ijtld.16.0716.Isoniazid

Steen, A., Buist, G., Horsburgh, G. J., Venema, G., Kuipers, O. P., Foster, S. J., & Kok, J. (2005). AcmA of Lactococcus lactis is an N-acetylglucosaminidase with an optimal number of LysM domains for proper functioning. *FEBS Journal*, *272*(11), 2854–2868. https://doi.org/10.1111/j.1742-4658.2005.04706.x

Steen, A., Buist, G., Leenhouts, K. J., El Khattabi, M., Grijpstra, F., Zomer, A. L., … Kok, J. (2003a). Cell wall attachment of a widely distributed peptidoglycan binding domain is hindered by cell wall constituents. *Journal of Biological Chemistry*, *278*(26), 23874–23881. https://doi.org/10.1074/jbc.M211055200

Steen, A., Buist, G., Leenhouts, K. J., El Khattabi, M., Grijpstra, F., Zomer, A. L., … Kok, J. (2003b). Cell wall attachment of a widely distributed peptidoglycan binding domain is hindered by cell wall constituents. *Journal of Biological Chemistry*. https://doi.org/10.1074/jbc.M211055200

Takade, A., Umeda, A., Matsuoka, M., Yoshida, S. ichi, Nakamura, M., & Amako, K. (2003). Comparative studies of the cell structures of Mycobacterium leprae and M. tuberculosis using the electron microscopy freeze-substitution technique. *Microbiology and Immunology*, *47*(4), 265–270. https://doi.org/10.1111/j.1348-0421.2003.tb03394.x

Tanaka, K., Nguyen, C. T., Liang, Y., Cao, Y., & Stacey, G. (2012). Role of LysM receptors in chitin-triggered plant innate immunity. *Plant Signaling & Behavior*, *8*(1), 147–153. https://doi.org/10.4161/psb.22598

Tang, J., Yam, W. C., & Chen, Z. (2016). Mycobacterium tuberculosis infection and vaccine development. *Tuberculosis*, *98*, 30–41. https://doi.org/10.1016/j.tube.2016.02.005

Turner, M. S., Hafner, L. M., Walsh, T., & Giffard, P. M. (2004). Identification and characterization of the novel LysM domain-containing surface protein Sep from Lactobacillus fermentum BR11 and its use as a peptide fusion partner in Lactobacillus and Lactococcus. *Applied and Environmental Microbiology*, *70*(6), 3673–3680. https://doi.org/10.1128/AEM.70.6.3673-3680.2004

Turner, R. D., Mesnage, S., Hobbs, J. K., & Foster, S. J. (2018). Molecular imaging of

glycan chains couples cell-wall polysaccharide architecture to bacterial cell morphology. *Nature Communications*, *9*(1). https://doi.org/10.1038/s41467-018-03551-y

Uhlen, Guss, Nilsson, Gatenbeck, P. & L. (1984). Complete sequence of the staphylococcal gene encoding protein A. A gene evolved through multiple duplications. *J. Biol. Chem. 259, 1695-1702*, *64*(12), 1350–1355.

Varki, A., Cummings, R. D., & Esko, J. D. (2017). *Essentials of Glycobiology, 3rd edition* (p. Print). p. Print. https://doi.org/10.1016/S0962-8924(00)01855-9

Vishwanath Venketaraman, D. K. and B. S. (2015). Mycobacterium tuberculosis. *Journal of Immunology Research*, *26*(6), 555–556. https://doi.org/10.1016/j.tim.2018.02.012

Visweswaran, G. R. R., Dijkstra, B. W., & Kok, J. (2011a). Murein and pseudomurein cell wall binding domains of bacteria and archaea-a comparative view. *Applied Microbiology and Biotechnology*, Vol. 92, pp. 921–928. https://doi.org/10.1007/s00253-011-3637-0

Visweswaran, G. R. R., Dijkstra, B. W., & Kok, J. (2011b). Murein and pseudomurein cell wall binding domains of bacteria and archaea-a comparative view. *Applied Microbiology and Biotechnology*, *92*(5), 921–928. https://doi.org/10.1007/s00253-011-3637-0

Visweswaran, G. R. R., Dijkstra, B. W., & Kok, J. (2012). A genetically engineered protein domain binding to bacterial murein, archaeal pseudomurein, and fungal chitin cell wall material. *Applied Microbiology and Biotechnology*, *96*(3), 729–737. https://doi.org/10.1007/s00253-012-3871-0

Visweswaran, G. R. R., Leenhouts, K., Van Roosmalen, M., Kok, J., & Buist, G. (2014). Exploiting the peptidoglycan-binding motif, LysM, for medical and industrial applications. *Applied Microbiology and Biotechnology*, Vol. 98, pp. 4331–4345. https://doi.org/10.1007/s00253-014-5633-7

Volkmar Braun, Helga Gnirke, U. H. and K. R. (1973). Model for the structure of the shape maintaining layer of the Escherichia coli cell envelope. *Journal of Bacteriology*, *114*(3), 1264–1270.

Vollmer, W., Blanot, D., & De Pedro, M. A. (2008). Peptidoglycan structure and architecture. *FEMS Microbiology Reviews*, Vol. 32, pp. 149–167. https://doi.org/10.1111/j.1574-6976.2007.00094.x

Vollmer, W., Blanot, D., De Pedro, M. A., Scheffers, D., Pinho, M., Nikolaidis, I., …

Dessen, A. (2005). Structural and mechanistic basis of penicillin-binding protein inhibition by lactivicins. *FEMS Microbiology Reviews*, *30*(2), 565–569. https://doi.org/10.1128/MMBR.69.4.585

Vollmer, W., Höltje, J., & Ho, J. (2004). *The Architecture of the Murein (Peptidoglycan) in Gram-Negative Bacteria : Vertical Scaffold or Horizontal Layer (s)? MINIREVIEW The Architecture of the Murein (Peptidoglycan) in Gram-Negative Bacteria : Vertical Scaffold or Horizontal Layer (s)?†*. *186*(18), 5978–5987. https://doi.org/10.1128/JB.186.18.5978

Walker, J. M. (2002). Protein Protocols Handbook, The. In *The Protein Protocols Handbook* (Vol. 0). https://doi.org/10.1385/1592591698

Walsh, C. (2017). *Antibiotics: Challenges, Mechanisms, Opportunitiesle*.

WHO. (2018). *Global tuberculosis report 2018*.

Williams, C. J., Headd, J. J., Moriarty, N. W., Prisant, M. G., Videau, L. L., Deis, L. N., … Richardson, D. C. (2018). MolProbity: More and better reference data for improved all-atom structure validation. *Protein Science*, *27*(1), 293–315. https://doi.org/10.1002/pro.3330

Winn, M. D., Ballard, C. C., Cowtan, K. D., Dodson, E. J., Emsley, P., Evans, P. R., … Wilson, K. S. (2011). Overview of the CCP4 suite and current developments. *Acta Crystallographica Section D: Biological Crystallography*, *67*(4), 235–242. https://doi.org/10.1107/S0907444910045749

Winter, G. (2010). Xia2 : an Expert System for Macromolecular Crystallography Data Reduction . *Journal of Applied Crystallography*, *43*(1), 186–190. https://doi.org/10.1107/s0021889809045701

Winter, Graeme, & McAuley, K. E. (2011). Automated data collection for macromolecular crystallography. *Methods*, *55*(1), 81–93. https://doi.org/10.1016/j.ymeth.2011.06.010

Wlodawer, A., Minor, W., Dauter, Z., Jaskolski, M., & Physics, B. (2015). Protein crystallography for non-crystallographers, or how to get the (but not more) from published macromolecular structures. *Febs J.*, *275*(1), 1–21. https://doi.org/10.1111/j.1742-4658.2007.06178.x.Protein

Wong, J. E. M. M., Midtgaard, S. R., Gysel, K., Thygesen, M. B., Sorensen, K. K., Jensen, K. J., … Blaise, M. (2015). An intermolecular binding mechanism involving multiple LysM domains mediates carbohydrate recognition by an endopeptidase. *Acta Crystallographica Section D: Biological Crystallography*,

71, 592–605. https://doi.org/10.1107/S139900471402793X

World Day TB, 2017. (2017). World TB Day 2017: Advances, Challenges and Opportunities in the “End-TB” Era. *International Journal of Infectious Diseases*, *56*, 1–5. https://doi.org/10.1016/j.ijid.2017.02.012

Yao, Y., Barghava, N., Kim, J., Niederweis, M., & Marassi, F. M. (2012). Molecular structure and peptidoglycan recognition of mycobacterium tuberculosis ArfA (Rv0899). *Journal of Molecular Biology*, *416*(2), 208–220. https://doi.org/10.1016/j.jmb.2011.12.030

Ying Zhang, Beate Heym, Bryan Allen, D. Y. S. C. (1992). The Catalase-Peroxidase Gene and Isoniazid Resistance of Myc. *Nature*, 591–593.