

Structural genomic studies of lipoproteins from *Mycobacterium smegmatis* for drug design

By

Feras Moheisen. M Almourfi

B.Sc (Hons) Veterinary Medicine



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

Department of Molecular Biology & Biotechnology

April 2014



In the Name Of Allah (God), The Most Gracious, The Most Merciful

**This Thesis is dedicated to my beloved parents and my
Loving wife and daughter**

Abstract

Tuberculosis (TB) is considered as an old infectious disease that leads to many fatalities in Man. *Mycobacterium tuberculosis* was discovered as the causative agent of tuberculosis by Robert Koch in 1882. Since then, scientists started the first move in order to develop such tools to prevent the disease. According to the WHO 2012 TB report, about one-third of the population is infected with *M.tuberculosis*; TB causes nearly 1.8 million deaths every year. Perhaps most worrying, new strains of *M.tuberculosis* resistant to most or even all-standard anti-TB drugs are spreading throughout the world, making treatment more costly and often impossible. Therefore, we urgently need to discover new drugs to overcome TB. The genomic sequence of *M.tuberculosis* has been completed in 1998 and has helped to shed light on new pathways as drug targets [1].

As part of a drug discovery programme, a structural genomics study of lipoproteins has been launched using the *Mycobacterium smegmatis* as a model organism for *M.tuberculosis*. A lipid-anchored protein is a class of protein that is produced in the cytoplasm as a pre-lipoprotein and attached to the cell membrane following post-translational modification (lipidation). Such proteins represent about 3% of bacterial genomes [1]. Furthermore, all bacteria apparently allocate particular proteins to the cell envelope by a process called post-translational lipid modification in order to produce membrane-anchored lipoproteins that are able to work in the aqueous environment at the membrane interface [2]. Therefore, this project aims to identify new targets suitable for drug discovery and shed light on their role in the cell. Eight targets were identified and put into a pipeline of cloning, over-expression, purification and crystallization for structure determination. Five target proteins of different putative functions were successfully purified with one (Msmeg_0515) annotated as an ABC sugar transporter protein, leading to structure determination.

The structure of Msmeg_0515 (AgaE) has been determined to a high resolution of 1.22 Å. Structure comparison of AgaE with other sugar binding proteins revealed that AgaE shares a similar fold with the maltose / maltodextrin binding protein (MalE) from *E.coli*. Previous bioinformatics studies on the sugar transporters of *M.smegmatis* and *M.tuberculosis* suggested that AgaE is an α -galactoside sugar binding protein [3], however, structural analysis of the binding site of AgaE protein

revealed that it's more similar to malto-oligosaccharide binding proteins. Also, binding assay by Circular dichroism (CD) has revealed a significant affinity of AgaE for maltose, glycerol 3-phosphate and acarbose but not α -galactoside sugars.

Acknowledgements

At this point, it's my pleasure to extend my special gratefulness and gratitude to my supervisor Dr. Patrick Baker for his kind continuous support, encouragements, patient and respectful supervision throughout my PhD studies.

Also, I would like to express my thanks to Prof. David Hornby for giving me the opportunity to do my PhD studies in the department and for his encouragement and fruitful discussions. Also, I owe special thanks for my committee advisors Dr. Rosie Staniforth and Dr. John Rafferty, for their advice and discussions, thanks a lot Rosie for helping me in my CD experiment analysis. I don't forget to thanks Dr. Qaiser Sheikh for his help and support in the first year of my PhD. Many thanks to Fiona for her kindness, birthday cards for every one and excellent lab control, which have a positive impact for every one's research. My thanks also to all my colleagues in the X-ray crystallography group for their contribution, especially Dr. Claudine Bisson.

My deep thanks are also to all my friends and colleagues in the department Dr. Yahya Aldawood, Dr. Heba Alhamal, Dr. Naji Alibrahim, Dr. Ibarhiem Alshubaith, Abdulhameed, Mona Alharbi and Faiza ALmalki. Thanks must also go to my best friends Mishal Alhajmouhammed and Hassan Ayad who have all been supporting me in the past four years in Sheffield.

I would like also to show all my gratefulness and gratitude to my parents, for their love and sacrifices for me all the time. Also, I don't forget my brothers and sisters, specially my sister Khulud.

Finally, I would like to express my appreciation and thanks to my lovely wife Thikra Alamri for her love and support all the time through the last two years at Sheffield, without you the life is impossible and to my lovely and beautiful daughter Noor, who giving me the strength and inspiration to reach my goals.

Table of contents

1. Chapter one: Introduction	23
1.1 Introduction.....	24
1.1.1 Tuberculosis.....	24
1.1.2 Infection process of tuberculosis (TB) disease.....	26
1.1.3 Treatment and diagnosis of TB.....	29
1.1.4 Vaccine development against <i>M.tuberculosis</i>	30
1.2 The bacterium.....	33
1.2.1 <i>Mycobacterium tuberculosis</i>	33
1.2.2 Structure of mycobacterium cell envelope.....	34
1.3 Membrane proteins.....	35
1.3.1 Lipid anchored proteins.....	37
1.3.2 The lipoprotein biosynthetic pathway.....	38
1.3.3 Examples of Lipoprotein in <i>Mycobacterium</i>	41
1.4 Fundamental nutrient pathways of mycobacterium cell envelope.....	41
1.5 <i>M.tuberculosis</i> and structural genomics.....	42
1.6 Identification of drug target.....	42
1.7 Aim of project.....	44
1.7.1 Target selection.....	46
2 Chapter two: Materials and Methods	49
2.1 DNA cloning	50
2.1.1 Recombinant DNA technology and protein production.....	50
2.1.2 Genomic DNA extraction of <i>M.smegmatis</i>	50
2.1.3 Polymerase Chain Reaction (PCR).....	51
2.1.4 pET Vectors.....	52
2.1.5 Cloning with pET28a.....	52
2.1.6 pET Expression hosts.....	52
2.1.7 Purification of PCR product.....	52
2.1.8 Restriction digestion for vector pET28a and PCR product.....	53
2.1.9 DNA ligation and transformation into Dh5 α	55
2.1.10 Confirmation of successful cloning.....	55
2.2 Protein Overexpression.....	56
2.2.1 Transformation.....	56

2.2.2	Small-scale overexpression trials.....	56
2.2.3	Large-scale overexpression.....	57
2.2.4	Production of Seleno-L-methionine incorporated protein.....	57
2.3	Protein purification techniques.....	58
2.3.1	Cell disruption.....	58
2.3.2	Nickel – NTA chromatography.....	59
2.3.3	Size exclusion chromatography (Gel filtration).....	58
2.3.4	Sodium dedecyl sulfate polyacrylamide gel electrophoresis (SDS PAGE).....	59
2.3.5	Protein concentration.....	60
2.4	Protein crystallization.....	61
2.4.1	Introduction.....	61
2.4.2	Protein crystals.....	61
2.4.3	Growing protein crystal.....	62
2.4.4	Vapour diffusion methods.....	64
2.4.4.1	Sitting drop method.....	64
2.4.4.2	Hanging drop method.....	64
2.5	Crystal mounting and cryoprotectants.....	65
2.5.1	Data collection apparatus.....	65
2.5.1.1	Copper-rotating anode system.....	65
2.5.1.2	Diamond light source (Synchrotron).....	66
2.5.1.3	Detectors.....	67
2.5.2	Data collection strategy.....	67
2.5.3	Data collection.....	68
2.6	Processing Data from diffraction images.....	68
2.7	Structure building and refinement.....	69
3	Chapter three: Cloning, over expression, purification and crystallization of protein targets from <i>M.smegmatis</i>.....	71
3.1	Target selection.....	72
3.1.1	Hydrophobic plots analysis.....	72
3.1.2	Generating a truncated protein for selected targets.....	72
3.2	From DNA to protein.....	73
3.2.1	Target amplification.....	73

3.2.2	Gene cloning.....	74
3.2.3	Small and large-scale protein expression.....	76
3.3	Studies on the Msmeg_3621 (NDH2) protein.....	77
3.3.1	Introduction.....	77
3.3.2	Cloning and expression of NDH2 (full length).....	77
3.3.3	NDH2 structure and function.....	79
3.3.4	Protein Purification of NDH2 enzyme.....	82
3.3.4.1	Ni-NTA chromatography of NDH2.....	82
3.3.4.2	Gel filtration chromatography of NDH2.....	82
3.3.5	Sample preparation and crystallization of NDH2.....	85
3.3.6	Discussion.....	85
3.4	Studies on the Msmeg_5007 protein.....	90
3.4.1	Introduction.....	90
3.4.2	Bioinformatics study on LprB.....	91
3.4.3	Protein purification of LprB.....	94
3.4.3.1	Ni-NTA chromatography purification.....	94
3.4.3.2	Gel filtration purification.....	94
3.4.4	LprB protein crystallization.....	96
3.4.5	Identification of more crystal hits.....	97
3.5	Studies on the Msmeg_6050 protein.....	98
3.5.1	Introduction.....	98
3.5.2	Bioinformatics study on Msmeg_6050.....	101
3.5.3	Purification of Msmeg_6050.....	105
3.5.4	Crystallization of Msmeg_6050.....	107
3.6	Studies on the Msmeg_5456 protein.....	108
3.6.1	Introduction.....	108
3.6.2	Bioinformatics study on LpqN protein.....	110
3.6.3	Protein purification of LpqN.....	113
3.6.4	Crystallization of LpqN.....	114
4	Chapter four: AgaE structure determination.....	115
4.1	Bioinformatics study on Msmeg_0515 protein.....	116
4.1.1	Target selection.....	116
4.1.2	Further analysis of Msmeg_0515.....	119
4.1.3	ABC transport system.....	122

4.1.4	Structure of ABC transporters.....	122
4.1.5	The substrate binding domain (SBD).....	124
4.1.6	Carbohydrate binding protein dependent ABC transporters.....	124
4.1.7	Genetic analysis of Msmeg_0515.....	126
4.2	Cloning of Msmeg_0515 (<i>agaE</i>).....	128
4.2.1	Genomic DNA preparation.....	128
4.2.2	Primers design and gene amplification of <i>agaE</i>	128
4.2.3	Restriction digestion of the PCR product and pET28a plasmid.....	131
4.2.4	Sticky end ligation into pET28a plasmid.....	131
4.2.5	Sub-cloning into Dh5 α	131
4.2.6	Identification of successful cloning.....	132
4.3	Protein overexpression and purification of AgaE.....	133
4.3.1	Transformation of pET28 <i>agaE</i> to B121 (<i>DE3</i>) strain.....	133
4.3.2	Optimization of AgaE overexpression.....	133
4.3.3	Large-scale overexpression of AgaE.....	135
4.3.4	Ni-NTA affinity chromatography purification of AgaE.....	135
4.3.5	Gel filtration of AgaE.....	135
4.4	Initial automated crystallization screening of AgaE.....	138
4.4.1	Preparation of AgaE protein sample.....	138
4.4.2	Crystallization of AgaE.....	138
4.4.3	Identifying successful crystallization.....	138
4.4.4	Crystal diffraction test.....	140
4.4.5	X-ray data collection of AgaE crystal.....	141
4.4.6	Native data processing of AgaE crystal.....	143
4.4.7	Structure determination.....	144
4.4.7.1	Molecular replacement.....	144
4.5	Expression and purification of AgaE incorporated with Seleno-methionine.....	144
4.6	Crystallization of Seleno-methionine incorporated AgaE.....	145
4.6.1	X-ray data collection of Seleno-methionine AgaE crystal.....	146
4.6.2	Obtaining experimental phase for AgaE Se-MET data.....	150
4.7	Structure refinement.....	155
4.7.1	Structure building and refinement of AgaE.....	155
4.7.2	Phasing of native data by molecular replacement.....	155
4.7.3	The final model of AgaE.....	156

4.7.4	Alternative conformation of residues in the AgaE structure.....	160
5	Chapter five: Structure of AgaE.....	161
5.1	AgaE structure description.....	162
5.1.1	Detailed features of AgaE structure.....	162
5.1.2	Molecular surface	165
5.2	Crystal contact.....	168
5.3	Structure comparison.....	172
5.3.1	Functional prediction.....	172
5.3.2	Structure classification of solute binding proteins.....	176
5.3.3	Binding site of carbohydrate binding proteins.....	178
5.3.3.1	Hydrophobic surface.....	182
5.3.3.2	Hydrogen bond interaction.....	183
5.3.4	Sugar binding site in other proteins.....	185
5.4	Sugar binding proteins specificity.....	189
6	Chapter six: AgaE – sugar complex.....	195
6.1	Introduction	196
6.2	Binding assays	196
6.2.1	Tryptophan Fluorescence Spectroscopy.....	196
6.2.2	Methodology.....	196
6.2.3	Results.....	197
6.3	Circular dichroism (CD).....	199
6.3.1	Methodology	199
6.3.2	CD data Analysis.....	200
6.4	AgaE – sugar complex.....	203
6.4.1	Co-crystallization of AgaE with Sugar.....	203
6.4.2	Crystallization results.....	203
6.4.3	Data collection and processing of co-crystallized AgaE.....	203
6.4.4	Structure determination and analysis.....	204
6.5	Determination of more crystal structures of AgaE potential sugar complex.....	209
6.6	Structure analysis.....	212
6.6.1	AgaE structure with potential binding of Acarbose.....	212
6.6.2	AgaE structure with potential binding of G3P & Maltose.....	217
6.7	Final discussion.....	217

7 Chapter Seven: Structural studies on phosphoglucose isomerase (PGI) mutants from <i>P.furiosus</i>	221
7.1 Introduction.....	222
7.2 <i>P.furiosus</i> Phosphoglucose isomerase (PGI) structure.....	219
7.3 Active and site function of <i>PfPGI</i>	220
7.4 Identification of possible PGI mutants.....	224
7.5 Overexpression of <i>PfPGI</i> mutants.....	225
7.6 Purification of <i>PfPGI</i> mutants.....	226
7.7 Crystallization of <i>PfPGI</i> mutants.....	227
7.8 Structure analysis of the mutation – carrying loop.....	231
8 Appendix: X-ray crystallography theory, symbols and abbreviation	239
8.1 Abbreviations.....	240
8.1.1 Crystallographic symbols and definition.....	240
8.1.2 Chemicals and Biological symbols.....	240
8.1.3 Miscellaneous.....	241
8.2 References	242

List of Figures

1.1 A map showing the highest rates of global TB infection.....	25
1.2 The process of human host TB and the development of granuloma transfer.....	28
1.3 The chemical structures of compounds used for tuberculosis treatment.....	31
1.4 <i>M.tuberculosis</i> cell wall composition.....	32
1.5 Schematic diagram of the cell wall structure of both Gram positive.....	33
1.6 Schematic structure of the cell wall in <i>M.tuberculosis</i>	34
1.7 Schematic diagram of the possibilities attachments of membrane protein to lipid bilayer.....	36
1.8 The signal sequence composition of bacterial lipoproteins.....	37
1.9 Bacterial lipoprotein biosynthesis pathway.....	40
1.10 The iterative process of structure based drug design discovery.....	43
1.11 A schematic diagram of the predicted way of the lipoprotein membrane attachment.....	46
2.1 The pET28a plasmid.....	54
2.2 A diagram illustrating the protein crystallization phases.....	63
2.3 A diagram illustrates the main techniques used in protein crystallization.....	64
2.4 A schematic diagram of the Diamond Light Source (<i>synchrotron</i>).....	66
3.1 Electrophoresis gels showing the PCR products of gene amplification for all selected targets.....	75
3.2 Hydropathy (a) and transmembrane helix prediction (b) plots of NDH2 from <i>M.smegmatis</i>	78
3.3 The pathway of aerobic electron flow in mycobacterium.....	80
3.4 The predicted 3D structure fold of Msmeg_3621 (NDH2).....	81
3.5 SDS gel showing protein purification steps of truncated NDH2.....	83
3.6 Chromatogram analysis of truncated NDH2 purification.....	84
3.7 Protein sequence alignment based on secondary structures of bacterial NDH2 protein from <i>C.thermarum</i> , <i>M.tb</i> and <i>M.smegmatis</i> and the Ndi1 protein from <i>S.cerevisiae</i>	86
3.8 The 3D structures of NDH2 enzyme from <i>C.thermarum</i> (top,) and Ndi1 enzyme from <i>S.cerevisiae</i> (bottom).....	88
3.9 Structural superposition of the NDH2 enzyme in complex with FAD ⁺ and Ndi1 in complex with FAD ⁺ and NAD ⁺	89

3.10	Hydropathy (a) and transmembrane helix prediction (b) plots of LprB from <i>M.smegmatis</i>	90
3.11	Protein sequence alignment based on secondary structures of LrpB from <i>M.tuberculosis</i> and <i>M.smegmatis</i>	91
3.12	A Blast search against the non-redundant protein sequence.....	92
3.13	The predicted 3D structure fold of LprB.....	93
3.14	SDS gel showing protein purification steps of LprB (full length)	95
3.15	Photographs of LprB native crystals.....	96
3.16	Photographs of LprB native crystals.....	97
3.17	Hydropathy plot of Msmeg_6050 and protein sequence.....	99
3.18	Transmembrane (a) and signal peptides (b) prediction plots of Msmeg_6050.....	100
3.19	Protein sequence alignment based on secondary structures of Msmeg_6050 from <i>M.smegmatis</i> and other metal binding proteins, such as SoxB from <i>T.thermophilus</i> and TroA from <i>Treponema pallidum</i>	101
3.20	A Blast search against the PDB protein structures.....	102
3.21	A Blast search of Msmeg_6050 against the non-redundant protein database.....	103
3.22	The predicted 3D structure fold of Msmeg_6050.....	104
3.23	SDS gel showing protein purification steps of truncated Msmeg_6050.....	105
3.24	Chromatogram analysis of Msmeg_6050 purification.....	106
3.25	Photographs of Msmeg_6050 crystals that were observed after one-year incubation.....	107
3.26	Hydropathy (a) and signal peptides (b) prediction plots of Msmeg_5456 (LpqN).....	109
3.27	A Blast search of LpqN against the non-redundant protein database.....	111
3.28	The predicted 3D structure fold of Msmeg_5456 (LpqN).....	112
3.29	Chromatogram analysis of LpqN purification.....	113
3.30	A photograph of LpqN crystal that was observed after one-year incubation	114
4.1	Hydropathy plot of Msmeg_0515 and protein sequence.....	117
4.2	The predicted signal peptide and transmembrane helix within the first 30 amino acids of Msmeg_0515 (AgaE) protein sequence.....	118

4.3	Sequence alignments of Msmeg_0515 against all non-redundant proteins in the NCBI database.....	120
4.4	A schematic diagram of ABC transporter structure and the transport process...	123
4.5	Carbohydrate binding proteins encoded by different transport systems with their putative predicted substrates in both (a) <i>Mycobacterium smegmatis</i> and (b) <i>mycobacterium tuberculosis</i>	125
4.6	The operon that encodes the Msmeg_0515 gene and the neighboring genes...	127
4.7	A schematic diagram of the putative ABC components, structure and organization of AgaEFKG.....	127
4.8	A schematic representation for the DNA cloning of the <i>agaE</i> (full length) gene to pET28a plasmid	130
4.9	Restriction digestion of pET28a that contains AgaE.....	132
4.10	SDS gel showing the small-scale over-expression of AgaE.....	134
4.11	SDS gel viewing the protein purification steps of Msmeg_0515 (AgaE)...	136
4.12	Chromatogram analysis of AgaE purification.....	137
4.13	Photographs of AgaE native crystals.....	139
4.14	Photographs of selected native AgaE crystals for data collection.....	140
4.15	Photograph of AgaE Se-MET crystals.....	145
4.16	The Se-K edge absorption spectrum of AgaE SAD experiment.....	147
4.17	Example of data collection for the Se-MET AgaE crystal.....	148
4.18	The output results of the SHELX C&D of AgaE SAD experiment.....	151
4.19	Electron density model of the substructure atoms of both original and inverted hands enantiomorphs of AgaE selenium SAD experiment.....	153
4.20	An example of the progress made with respect to the model building AgaE structure by electron density.....	154
4.21	The overall fold of the final structure model of AgaE.....	156
4.22	The properties of the main chain and the side chain for the final refined AgaE structure.....	158
4.23	The results statistics of Molprobit and Ramachandran plot of the final AgaE structure.	159
4.24	An alternative conformation of some AgaE residues.....	160
5.1	Schematic representation of AgaE structure.....	163
5.2	Cartoon representation of the overall fold structure of AgaE.....	164

5.3 Surface electrostatic representation of the AgaE structure.....	166
5.4 Cartoon representation of the putative active site between the two domains interface with complex with PEG.....	167
5.5 Cartoon representations of AgaE molecules packing in the crystal.....	169
5.6 Carton representations of the residues involved in AgaE crystal interface.....	170
5.7 Structure based alignment of agaE and other ABC binding proteins with different bound substrates.....	174
5.8 Highly conserved residues of different solute binding transporter mapped onto AgaE.....	175
5.9 Structural alignment of AgaE structure and other sugar binding proteins.....	177
5.10 Surface electrostatic representation of the carbohydrate binding protein structures with different sugar bound.....	179
5.11 Structural alignment of GacH sugar binding protein structure and AgaE structure.....	181
5.12 Superposition of AgaE (raspberry) and GacH (orange) binding sites.....	182
5.13 Superposition of AgaE (raspberry) and GacH (orange) binding sites.....	184
5.14 Superposition of AgaE (raspberry) and MalE (cyan) binding sites.....	186
5.15 Ligplot diagram representation of the binding site interaction of GacH and MalE with the bond acarbose.....	188
5.16 Sequence alignments of AgaE and other carbohydrate binding proteins....	191
5.17 Superposition of AgaE and TvuCMBP binding sites.....	192
5.18 Superposition of AgaE and EcoUgpB binding sites.....	193
5.19 Superposition of AgaE and ttMtBP binding sites.	194
6.1 Traces of tryptophan fluorescent assay.....	198
6.2 The CD spectrum of AgaE only and in complex with 5mM maltose, acarbose and glycerol 3-phosphate.	201
6.3 The CD spectrum representation of the AgaE complex 5mM added sugar at 222 nm intensity.....	202
6.4 The 3D structure of AgaE (green) in complex with acarbose (wheat) and PEG molecule with electron densities.....	213
6.5 The binding site of AgaE with the acarbose bound.....	213
6.6 Acarbose structure with electron density encountered to 0.6σ	214
6.7 Superposition of 3D structures of AgaE (green), GacH (cyan) and MalE (pink) in complex with acarbose.	215

6.8 Superposition of bound acarbose from the binding sites of AgaE, GacH and MalE.	216
7.1 Cartoon representation of the dimer (white and green subunits) of the wild type PfPGI Mn ⁺² /5PAA 3D-structure.	223
7.2 Specific activity of wild type PfPGI (PY) compared to selected high occurrence/activity mutants RG and AG, and low occurrence/activity mutants AD and VY.....	225
7.3 SDS-PAGE gel analysis of the purification of PfPGI. SDS represents samples taken through the purification process of PfPGI (AG) mutant.....	226
7.4 Photographs of PfPGI mutants crystal.....	228
7.5 A stereo representation of the mF _O – DF _C electron density (grey mesh), contoured at 1.5 σ , for the AG mutant PfPGI structure.....	232
7.6 Mass spectrometry experiment result of the crystallization buffers of AG and RG mutant's crystals.....	233
7.7 Mn ⁺² coordination in the mutant PfPGI structures.	234
7.8 Electron densities of substrate and inhibitor of PfPGI mutants.	235
7.9 Comparison of the structure of the loops adjacent to the mutation position in the four mutant structures.	236

List of tables

1.1 The main drugs used for TB treatment and their mode of action and inhibited target.....	31
1.2 The number of all proteins encoded in the genomic DNA of <i>M.tuberculosis</i>	45
1.3 All targets were selected to run this project.....	47
2.1 The composition of agar and lysogeny broth media.....	50
2.2 Gradient PCR condition of the <i>M.smegmatis</i> genes amplification.....	51
2.3 The composition of a standard PCR reaction mixture.....	51
2.4 The composition of a typical double restriction digestion reaction.....	53
2.5 The composition of a typical ligation reaction.....	55
2.6 Recipes of 12 and 6 % SDS Page ingredients.....	59
2.7 The seven protein crystal system lattices.....	62
3.1 The best PCR annealing temperature used for gene amplification of the target full-length and truncated proteins.....	73
3.2 The primer sequences used in this thesis for gene amplification.....	74
3.3 Over-expression conditions for all truncated target proteins from <i>M.smegmatis</i>	76
3.4 A Blast search against the non-redundant protein sequence.....	91
4.1 Results from Blast search of Msmeg_0515 against all PDB structures.....	121
4.2 Primers of the full length and truncated genes of AgaE.....	129
4.3 The data collection strategies of several crystals of native AgaE.....	141
4.4 Data statistic of three native AgaE crystals.....	142
4.5 Asymmetric unit contents and Matthews's coefficient for AgaE crystals.....	143
4.6 Data collection statistic of Se-Met AgaE crystal.....	149
4.7 The output result of SHELXE calculating of heavy atoms density and pseudo-free CC for the enantiomorph of the determined electron density map.....	152
4.8 The final refinement statistics and validation for the different native crystal structures of AgaE as well as Se-MET incorporated in the crystal structure.....	149
5.1 Accessibility and Buried surface areas of residues on the monomer –monomer interface.....	171
5.2 The unique hydrogen bonds formed between the two monomers in the crystal.....	171

5.3 Results of Dali search for the model putative sugar binding protein Mseg_0515 (AgaE).	173
5.4 Residues involved in the sugar binding of the GacH and MalE binding proteins and their equivalents in AgaE protein structure.....	183
6.1 Photographs of the AgaE-sugar complex crystals.....	205
6.2 Data collection statistics for all AgaE-sugar complex crystals.....	208
6.3 Data collection statistics for all AgaE in complex with acarbose and maltose crystals.....	210
6.4 The refinement statistics for three-crystal structures with potential sugar bound.....	211
7.1 Data collection statistics of PfPGI mutants.....	229
7.2 Final refinement Statistics of PfPGI mutants.....	230
7.3 Mn ²⁺ ligands and coordination distances (Å).....	231

Chapter 1

Introduction

Introduction of the pathogenic bacteria of *Mycobacterium tuberculosis* and an overview of the small-scale project of mycobacterial structural genomics.

1.1 Introduction

1.1.1 Tuberculosis

Tuberculosis (TB) is an infectious disease that is caused by the soil born pathogenic organism *Mycobacterium tuberculosis*. The disease infects both adults and children and mainly has an effect on the lung tissue and leads to Pulmonary TB [4, 5]. However, other parts of the body also can be infected, such as the skin, bones and intestine [4, 6]. Since the twenty-first century, TB has been considered as the second major global infectious disease after HIV. It is estimated that about one third of the world's population is infected with TB, however, the vast majority of this population will not present the disease [4]. The latest World Health Organization (WHO) figures estimate that, in 2012, tuberculosis caused approximately 1.3 million deaths and there were \approx 8.6 million new cases [4]. Although the introduction of antibacterial drug therapies have helped to decrease the rate of mortality by about 45% since 1990, the mortality rate remains high, due to the emergence of drug resistant forms of the disease [4]. Despite this widespread use of antibacterial therapies, developing countries like Africa have had the highest rate of TB because of poverty and the lack of access to treatment (Figure 1.1). Also, immuno-compromised patients, such as those with HIV, are more likely to be infected with TB. [4]. Of the 8.6 million new cases of TB reported in 2012, 1.1 million were of HIV positive patients (75% of these cases were reported in Africa) and of the 1.3 million deaths of TB patients in the same year 0.3 million were also of HIV positive patients [4]. In fact, people with AIDS have a 50-fold increase in TB susceptibility over HIV negative patients [4].

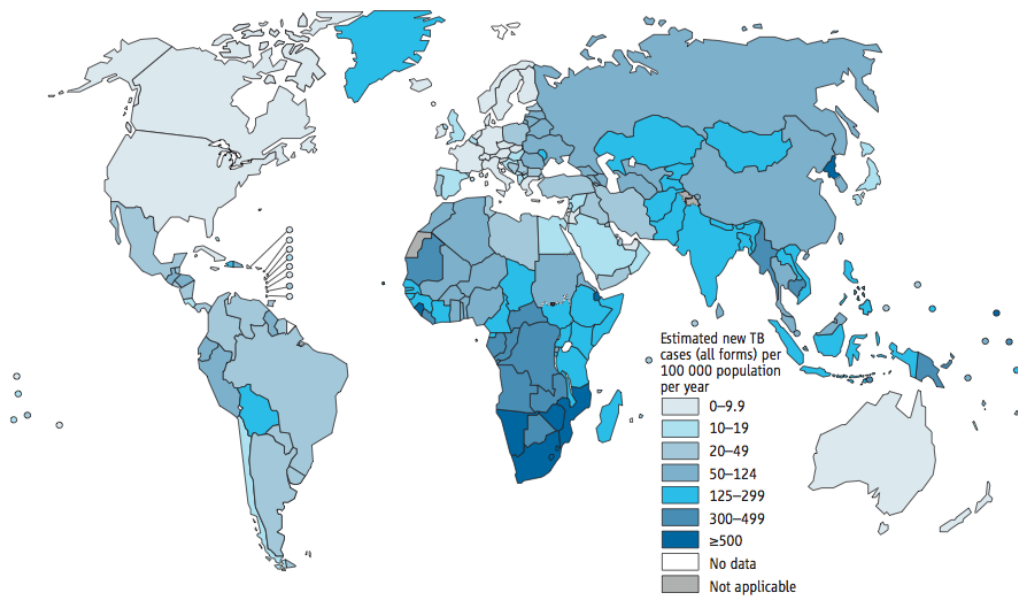


Figure 1.1 A map showing the highest rates of global TB infection. Areas are coloured as shown in the key, based on the estimated number of TB cases per 100,000-population/year. The highest numbers of cases were accounted in Swaziland and South Africa by about 1000 per 100,000 people compared to 10 per 100,000 people in some area of America and other developed countries, such as Australia, Japan, New Zealand and Western Europe. The map has been adapted from the Global Tuberculosis Report 2013 [4].

1.1.2 Infection process of tuberculosis (TB) disease

The main clinical symptoms that can be observed in a patient with TB are coughing with blood, sweating at night and loss in weight [7, 8]. The most common way of transfer of *M.tuberculosis* between people is by inhalation. TB transmission begins with a human source, most often a person with pulmonary TB. When an infected patient sneezes or coughs, aerosols are formed in the lungs and are expelled by coughing [9]. These aerosols contain thousands of micro-particles that carry the bacilli, and can be inhaled by others [9]. The disease affects the lungs in approximately two thirds of cases, but almost all other organs can be the site of TB infection. Although approximately one third of the world population is infected by *M.tuberculosis*, the infection is contained by the immune system in about 90 % of these cases [4, 9]. The TB bacilli can lie dormant for years, being protected by a thick waxy coat [8]. If the immune system is weakened, for instance by an HIV infection or treatment with immunosuppressive agents, the chance of developing active TB become much higher [9].

The process of TB infection in the lung starts by the inhalation of the droplets carrying *M.tuberculosis* bacillus. The bacterium is then ingested by phagocytosis by alveolar macrophages of the lung and dendritic cells (DCs) of the tissue [10]. DCs normally have some effect against a pathogen as a killing agent; however, *M.tuberculosis* can also disturb the normal killing effect of dendritic cells, to allow growth of the infecting agent [11]. The immune system of the body responds against the bacteria and starts to fight the bacteria by forming fibrosis and scar tissue around the bacteria. If the immune system failed to control the growth of TB bacteria, the bacteria will then be active again and infect other parts of the body [12] (Figure 1.2). Active *M.tuberculosis* starts to spread through out the body by developing pro-inflammatory infected cytokines, which activate macrophages again. This in turn, will require the immune system to produce more DCs and other immune cells, such as blood neutrophils and monocytes, which will also be infected with disease [13]. Thus, the bacterium can be transferred into other parts of the body through both infected dendritic cells that have the ability for migration and also through the blood stream to the area, such as neighboring lymph nodes. The lymph node specific T-cells will be activated and the activity of the natural killer (NK) cells will be induced by the release of interleukin 12 (IL-12) and interleukin18 (IL-18) of the newly infected cell cytokines, which in turn will produce gamma interferon (IFN- γ) that

helps to make TNF- α and microbiocide substances that are released by macrophage activation [14, 15].

In the last stage of the disease where granuloma are formed (the pathological feature of the disease), infected tissue starts to be destroyed due to the development of necrosis and macrophage differentiation action [7]. In the stage of advanced TB, the granuloma structure bursts, due to the presence of more infected immune cells, and in turn, the bacilli start to leak out to the air passages allowing the spread of TB [16, 17] (Figure 1.2).

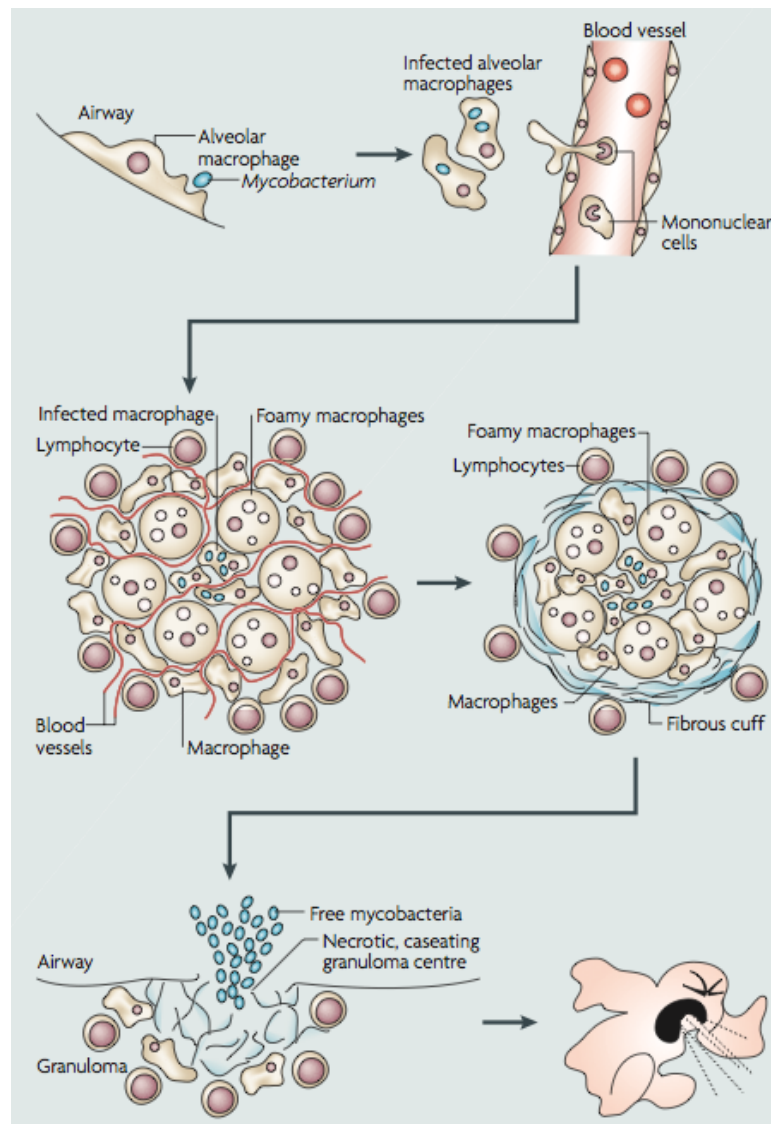


Figure 1.2 The process of human host TB and the development of granuloma transfer. The first stage occurs when the bacterium has been engulfed by phagocytosis by the lung alveolar macrophages. The immune system responds to the infection and leads to the development of a granuloma and transfer of the bacterium into other parts of the body through the blood vessels. The granuloma can lie dormant in the macrophages, however, in the case of advanced tuberculosis; the immune system fails to control the disease, the granuloma ruptures allowing the TB bacteria to spread. This figure has been adapted from Russell, 2007 [16].

1.1.3 Treatment and diagnosis of TB

Since the discovery of *M.tuberculosis* there have been many studies towards finding a proper treatment. For example, in 1943-1947 Selman Waksman, a scientist in America discovered the first antibiotic, streptomycin, that is active against the disease [18]. Streptomycin was generated from the Actinobacteria *Streptomyces griseus*, however, *M.tuberculosis* has become resistant to this antibiotic and other drugs have been developed, such as para-aminosalicylic acids (PAS) in 1948 and isoniazid (INH) which was developed in 1952 [18]. However, these drugs had to be given together to overcome drug resistance in *M.tuberculosis* and have to be administered for a period of two years. Several years later different drugs have appeared for instance, ethambutol (EMB), that took the place of PAS in 1960, mainly because it had to be taken for only 18 months [18]. Later, rifampicin (RIF) and pyrazinamide (PZA) were produced in 1970, which decreased the course of therapeutics even more, to just 6 months [18]. All of these developed drugs are known as first line agents (Figure 1.3). However, the emergence of new *M.tuberculosis* strains resistant to the first line drugs (usually isoniazid and rifampicin) (MDR-TB), resulted in these drugs being replaced by the second line agents, such as capreomycin, amikacin and canamycin, which are less efficient, more expensive and have high pathogenicity [18]. For example, over 450,000 patients have been estimated to have developed resistance for MDR-TB in 2012, which resulted in 170,000 deaths [4]. Recently, strains of *M.tuberculosis* have arisen (XDR-TB), that are also resistant to at least one of the second line antibiotics, such as capreomycin, amikacin and canamycin, which have presented even more problems for treatment.

At present the TB disease is treated by the use of a group of drugs in combination. At first, antibiotics (rifampicin, isoniazid, pyrazinamide, and ethambutol) are normally used for approximately 2 months, based on the activity of TB, and two of them, rifampicin and isoniazid, are used for an additional four months [19]. In case of the emergence of multidrug resistant forms of TB, several choices of other drugs combination are used, including first and second line drugs and the treatment course lasts for 18 months. Treatment also might include surgery and other chemotherapy for cavitary TB cases [20, 21]. Therefore, new drugs are required against this persistent disease [7, 22].

Two methods have been developed and have had a significant impact towards the diagnosis and treatment of the disease, which are chest X-rays and a sample test for a sputum smear under a microscope. However, the development of new drug resistant strains of *M.tuberculosis* have required the development of new methods of diagnosis, such as immunological examination assays and antibiotic resistance tests. The use of these tests have contributed towards the reduction of the percentage of TB deaths by 45% according to WHO, as 65 million of TB cases were diagnosed and treated between 1995 to 2012 [4, 20] (Figure 1.4).

1.1.4 Vaccine development against *M.tuberculosis*

Albert Calmette and Camille Guerin developed a vaccine against TB in 1906 [23]. The vaccine is known as "BCG" (Bacillus of Calmette and Guerin) and is based on the administration of a prophylactic weakened live bacillus to newborn babies [7]. Although the vaccine has several advantages, such as its cheap cost and long lasting activity with the same immunization efficiency, and has been approved to stop children from being infected with meningitis TB, it also has several disadvantages preventing it being used, such as the possibility of developing a new strain of high virulence after long term administration [7, 24]. Also, the vaccine is useless in adults, and doesn't seem to save them from being infected with TB [7, 23].

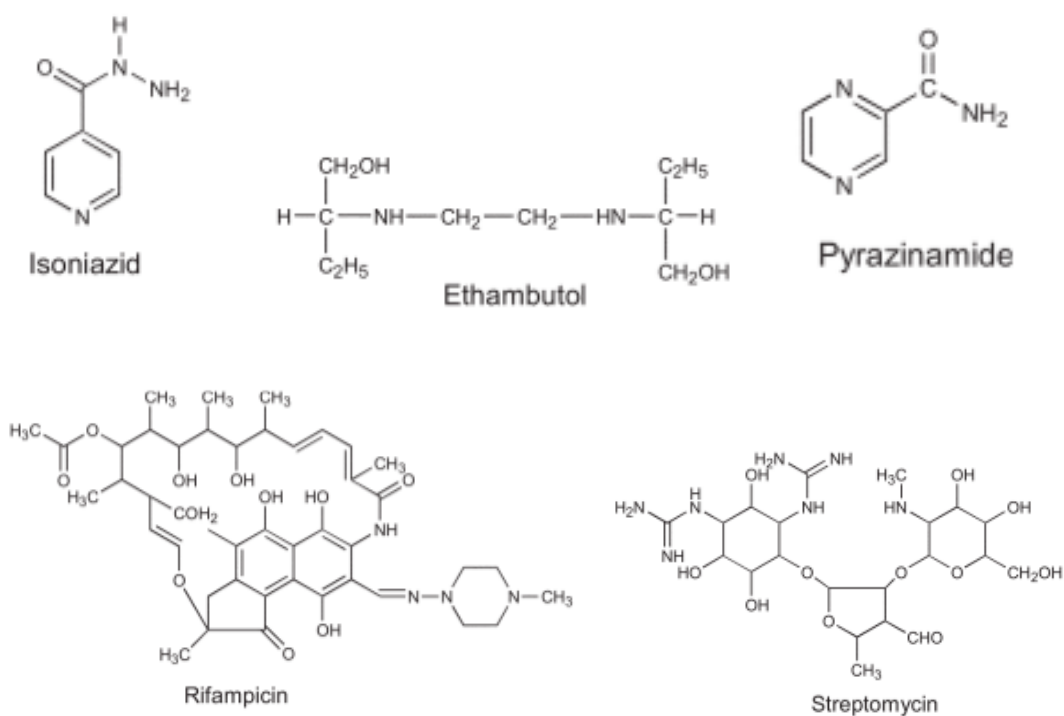


Figure 1.3 The chemical structures of compounds used for tuberculosis treatment.

Antibiotic	Mode of action	Inhibited targets.
Isoniazid (INH)	Inhibit cell wall synthesis.	Mycolic acid synthesis
Ethambutol (EMB)	Inhibit cell wall synthesis	Arabinogalactan.
Pyrazinamide (PZA)	Disrupt plasma membrane and energy metabolism.	Not known.
Rifampicin (RIF)	Inhibit RNA synthesis.	RNA polymerase.
Streptomycin	Inhibit protein synthesis.	S12 rRNA

Table 1. 1 The main drugs used for TB treatment and their mode of action and inhibited target.

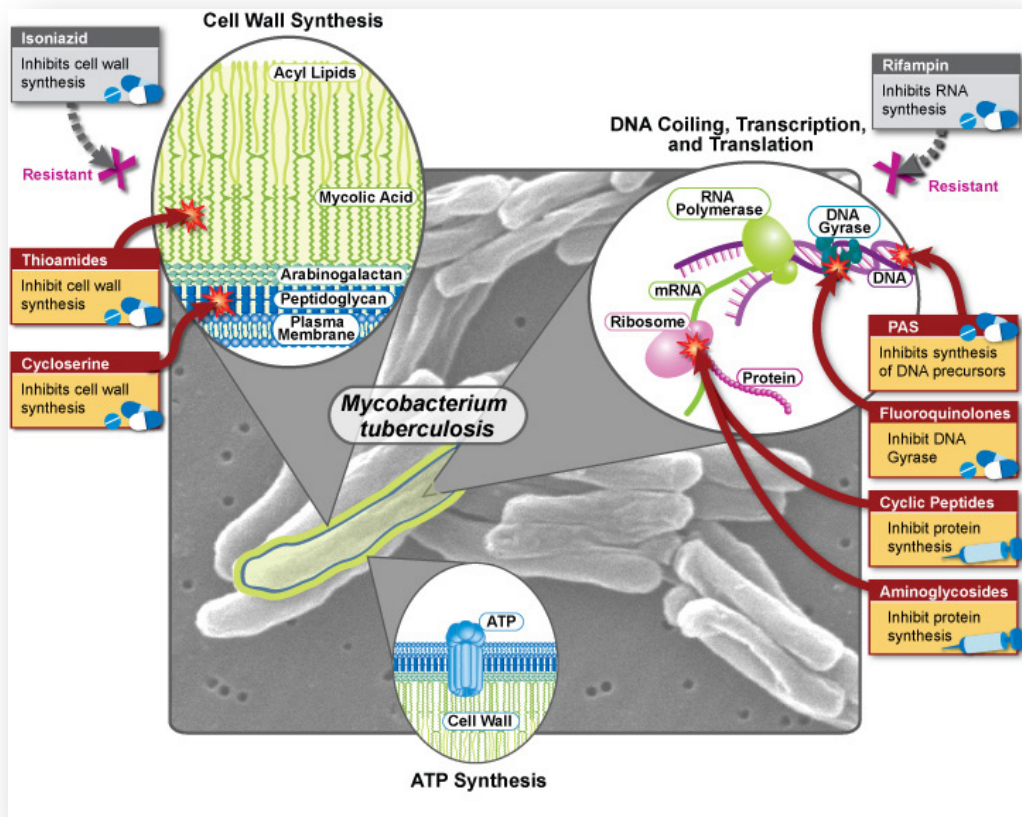


Figure 1.4 *M.tuberculosis* cell wall composition. Drug used in the treatment of patients with multi-drug resistant TB and their modes of actions are shown in the highlighted panels. Figure obtained from the Centers for Disease Control and Prevention website).

1.2 The bacterium

1.2.1 *Mycobacterium tuberculosis*

M.tuberculosis belongs to the Actinobacteria genus, it is a rod shaped aerobic bacteria, and sized 0.5 um in diameter and 1-4 um long [7]. The bacteria are neither motile nor sporulated and the genomic DNA is guanine and cytosine rich (approximately 65% of the total genome) [25, 26]. *M.tuberculosis* does not contain either a capsule or a flagellum. Also, it has a slow replication time of 24hs due to the waxy cell wall. The organism is found in environments, such as soil, water and in host cells, such as alveolar macrophages and has been characterized as an acid-fast Gram-positive bacterium [27]. *M.tuberculosis* was first detected by Robert Koch in 1882, and thus TB is known as Koch's disease [28]. *M.tuberculosis* is resistant to standard Gram staining, since the violet colour of the stain is not retained in the cell envelope [29].

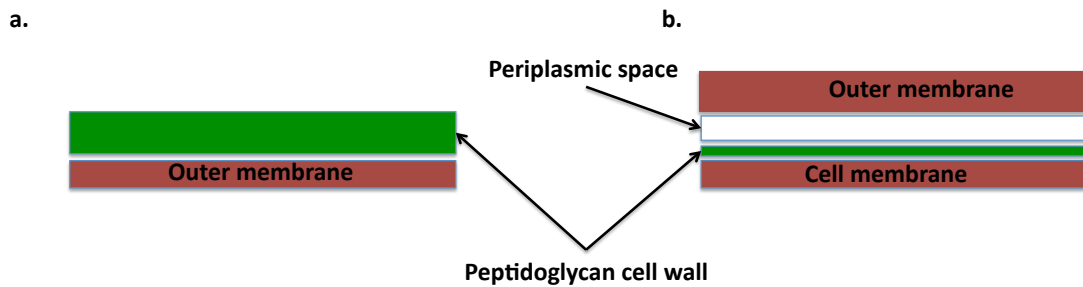


Figure 1.5 Schematic diagram of the cell wall structure of both Gram positive (a) and Gram negative (b). The main difference observed is the lack of the outer cell membrane from the cell wall of gram-positive bacteria.

1.2.2 Structure of mycobacterium cell envelope

The structure of cell wall is unique, compared with other bacteria. It is composed of a long chain of mycolic acid, a fatty acid that contains 60-90 carbon atoms [25]. The mycolic acid is bound to arabinogalactan molecules that make a phosphodiester link with the peptidoglycan [25, 30]. The peptidoglycan of the cell wall of most organisms consists of *N*-acetylmuramic acid, whereas in mycobacterium cell wall, this is replaced by *N*-glycolylmuramic acid [7]. The cell wall of *M.tuberculosis* has mycolic acid forming 60% of its lipid compounds that are present as a mix of two chains of 24 and 64 carbon atoms, respectively, and are attached to the cell wall polysaccharides by covalent bonds [25, 30]. Additionally, the cell wall also consists of two other glycolipid molecules, such as lipoarabinomannan (LAM) and the lipomannan (LM) that are attached to the phosphatidyl-myoinositol mannoside and anchored to the plasma membrane with non-covalent bonds [25, 31] (Figure 1.6).

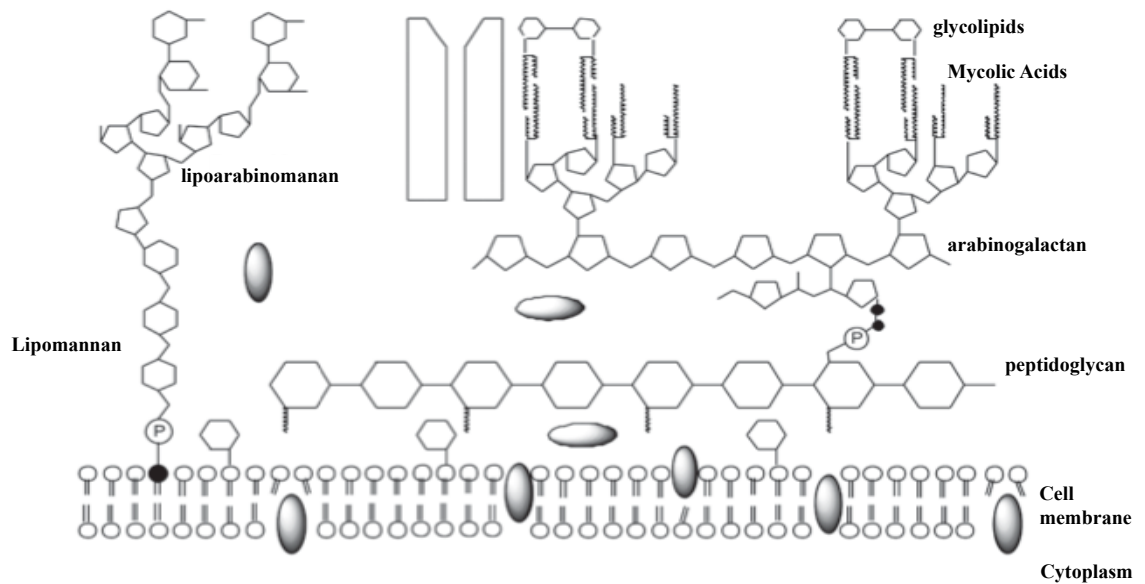


Figure 1.6 Schematic structure of the cell wall in *M.tuberculosis*. The cell membrane is composed of mycolic acids, which are linked to the arabinogalactan that is attached to the peptidoglycan by covalent bonds to form the inner layer of the cell membrane. Also, the cell wall consists of other glycolipid molecules, such as lipoarabinomannan (LAM), and lipomannan (LM) that is anchored to the plasma membrane [7].

1.3 Membrane proteins

Proteins are classified in nature into three main classes, globular, fibrous and membrane proteins. Membrane proteins are responsible for numerous types of function, such as cell signaling, transport of nutrients and cell adhesion. In addition, approximately 30% of the genomes of most organisms code for membrane proteins and over 50% of pharmaceutical drug targets are membrane proteins [32-34].

Membrane proteins can be classified into two main classes; integral membrane proteins and peripheral membrane proteins. Integral membrane proteins (IMPs) are proteins that are embedded in the cell membrane, and their removal from the membrane usually requires the use of detergents. Further, they also can be divided into six groups based on their embedded domain structures and location. The first and the second group are described as a single trans-membrane helix that passes through the cell membrane with the C-terminus either inside or outside the cell. The third and fourth groups contain either a single domain that is composed of several transmembrane helices and can pass through the membrane many times, or by multiple protein domains that can pass through the membrane once but with separated trans-membrane helices or β -pleated sheets, such as MSPA porins [35]. In groups five and six, the protein can either attach to the membrane by covalent bonds, with their lipids as membrane anchors, such as glycosylphosphatidylinositol (GPI) anchors [36] or can pass through the membrane and interact with the membrane phospholipids by an amphipathic helix interaction. However, integral membrane proteins can also consist of multiple β -strands, instead of α -helices, that can cross the phospholipids bilayer of the outer cell membrane of mitochondria and Gram-negative bacteria or the cell wall lipids of Gram-positive bacteria. Therefore, purification process of these types of protein is often very difficult, and requires special detergents, due to the presence of hydrophobic amino acid sequences, which help proteins to interact with the phospholipids bilayer of the cell membrane.

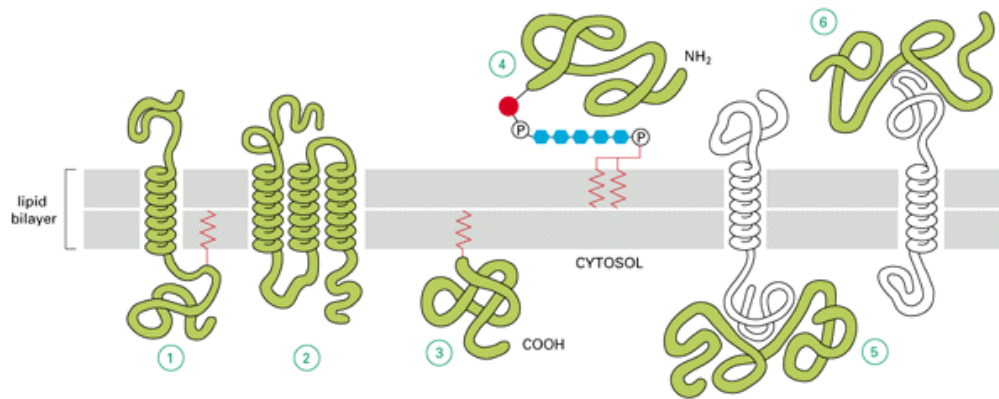


Figure 1.7. Schematic diagram of the possibilities attachments of membrane protein to lipid bilayer. Membrane protein might attach to membrane by covalent interaction of (1) single trans-membrane α -helix or (2) as multiple helices into the cytoplasm. Also, membrane protein might interact covalently by attachment of prenyl molecule or lipid fatty acid as in (3), or (4) by oligosaccharide into the phospholipid bilayer. (5 & 6) represent the non covalent interactions with other integral proteins [2].

The second class of membrane protein is that of peripheral membrane proteins (PMPs). This type of protein is attached to the cell membrane by other molecules that are already bound to the cell membrane, such as integral membrane proteins or membrane lipids. This interaction often occurs by electrostatic interaction, or by polar interaction, forming hydrogen bonds. PMPs thus can be extracted easily from soluble cell fractions by changing pH, or by addition of salt. In addition, PMPs can function as regulator enzymes for different membrane proteins, which work as receptors and channels, such as the Kinase C enzyme that is involved in signal transduction (Figure 1.7) [37].

Finally, lipoproteins, the subject of this thesis, are defined as types of membrane proteins that interact with the cell membrane lipids by a set of hydrophobic residues within their N or C-terminal sequence known as a signal sequence. The signal sequence enables the lipoprotein to be attached into the cell membrane by post-translational modification of a cysteine residue with fatty acyl molecules of the membrane lipid bilayer (N-acyl-S-diacylglyceryl-Cys) [38].

1.3.1 Lipid anchored proteins

A lipid-anchored protein is a class of protein that is produced in the cytoplasm as pre-lipoprotein and attached to the cell membrane by post-translational modification (lipidation). Functionally such proteins are heterogeneous and represent about 3% of bacterial genomes [1]. Furthermore, all bacteria apparently allocate particular proteins to the cell envelope by a process called post-translational lipid modification in order to produce membrane-anchored lipoproteins that are able to work in the aqueous environment of the membrane interface [2].

Analyzing the sequences of potential bacterial lipoproteins to identify possible signal sequences is the first step in identifying their roles and significance. The signal peptide sequence of a lipoprotein consists of different features that can be used to identify lipoprotein from any organism (Figure 1.8). Firstly, they contain an n-region that consists of 5-7 amino acids and usually contain at least two positively charged amino acids. The next part of the sequence contains the h-region (hydrophobic region). This region is composed of around 7-22 amino acids that are mainly uncharged and hydrophobic. Finally, the c-region, called the lipobox, and the first sequence to be identified in bacterial lipoproteins by Hayashi et al, in 1993 [39]. The lipobox consists of a four-amino-acid sequence at the C-terminal end of the signal peptide sequence, including the modified Cys at position +1 and is known as the lipobox [L- (A or S) - (G or A) - C] [38].

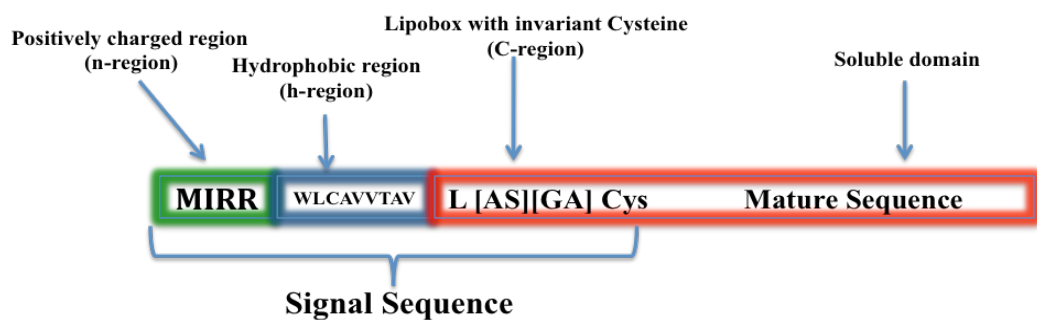


Figure 1.8 The signal sequence composition of bacterial lipoproteins. The sequence composed of three regions [43]. The first region (n-region, green) includes the positive charge residues within the first four residues. The second region (h-region, blue) includes strong hydrophobic residues and the third region (c-region, red) have the lipobox that includes the conserved Cysteine residue [38].

1.3.2 The lipoprotein biosynthetic pathway

There are two pathways that control the translocation of any protein through the inner membrane but do not verify their final site. The first pathway is known as the Secretory translocon pathway (Sec pathway), which functions as a pathway for unfolded proteins through the inner membrane to outside the cell after synthesis in the cytoplasm [40]. The second pathway is known as the Twin-arginine translocation pathway (Tat system), this system requires the presence of two arginine residues within the n-region of the N-terminal signal sequence, and translocates a folded protein through the inner cell membrane into the outside of the cell [41].

Lipid modification then takes place in order to translocate the lipoprotein either within the inner or outer cell membrane. Three enzymes are engaged in the lipid modification of lipoprotein. The signal peptidase II enzyme (LspA) cleaves the pre-lipoprotein in a specific cleavage site sequence [L- (A or S) - (G or A) C] that is found after the H-region of the signal peptides [42, 43]. This cleavage is carried out after the addition of a diacylglyceride unit into the conserved cysteine of the Lipobox (c-region) by a thioether linkage, resulting in the mature protein with a modified cysteine in the N-terminal portion of the protein sequence [44, 45]. This process is called lipidation, and is catalyzed by the enzyme prelipoprotein diacylglyceryl transferase (Lgt) [46, 47] (Figure 1.9).

The process of lipidation is completed by the addition of three fatty acyl chains, that are derived from the phospholipids of the bacterial cell membrane, to the cysteine of the N-terminal protein sequence, by the enzyme phospholipid–apolipoprotein N-acyltransferase (Lnt) [48] (Figure 1.9). An additional system called the localization of lipoprotein system (LoI) is responsible for translocating the mature protein to either the outer membrane or inner membrane. However, it is based on the type of residue following the N-terminal Cys of the signal sequence [49]. If this residue is an Asp, the protein embeds itself in the inner membrane and if it is a Ser, then the protein translocates to the outer membrane [50, 51].

In mycobacterium genomes, lipoproteins have been characterized with features of type II peptide sequence at the N terminal sequence [42]. Furthermore, both enzymes (Lgt and LspA) that are involved in lipoprotein biosynthesis have been found in the

genomic DNA of mycobacteria and other bacteria [52]. This might suggest that their functions are significant for anchoring of lipoprotein in all organisms [52, 53]. However, the third enzyme of phospholipid–apolipoprotein N- acyltransferase (Lnt) was found only in Gram – negative bacteria and is responsible to add a third acyl residue to the amino group of the modified cysteine [54].

Further, membrane lipids in mycobacteria and the conserved cysteine are linked covalently and considered to make lipoproteins attached to the biological membranes by hydrophobic interaction. In Gram-positive bacteria, lipoproteins are attached to the plasma membrane (inner) and in Gram-negative bacteria most of lipoproteins are found in the outer cell membrane [52].

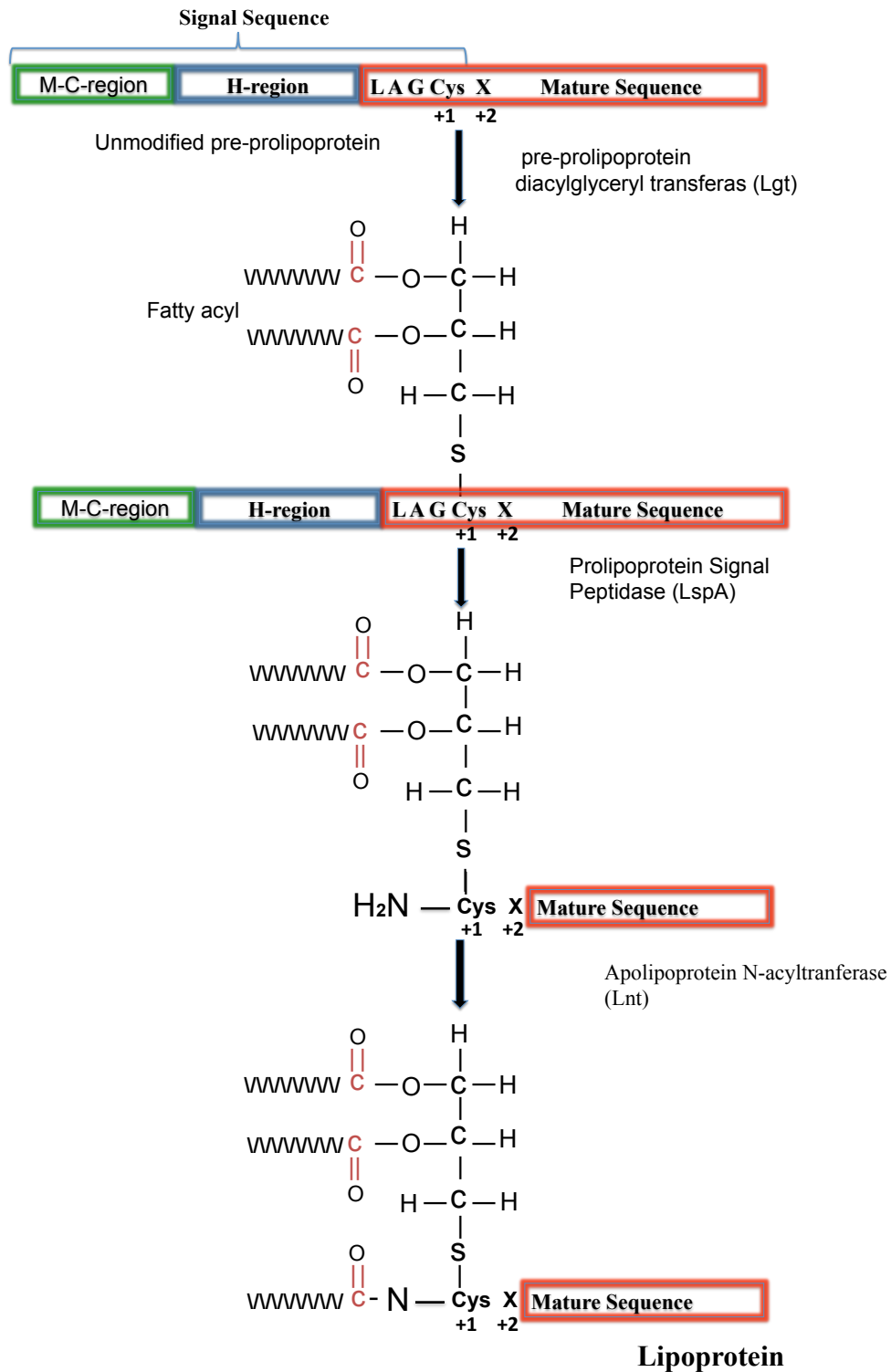


Figure 1.9 Bacterial lipoprotein biosynthesis pathway. Lipoprotein is synthesized in the cytoplasm as a pre-prolipoprotein with N-terminal signal peptides. Lipid modification of the lipoprotein precursor is intermediated by regular activity of three different enzymes (pre-prolipoprotein diacylglyceryl transferase (Lgt), prolipoprotein signal peptidase (LspA) and apolipoprotein N-acyltransferase (Lnt). [55].

1.3.3 Examples of Lipoprotein in Mycobacterium

In mycobacteria lipoproteins have been characterized that have a wide variety of different functions [47, 48]. For example, lipoproteins that are involved in cell adhesion, such as LpqH from *M.tuberculosis*. This lipoprotein has an immunologic role in phagocytosis stimulation, when it binds to the cell receptors of monocytes [56]. Also, lipoprotein functions as receptors of several molecules, such as the phosphate transport receptor (Pst) of *M.tuberculosis*, which causes a significant immune response when used as a vaccine in mice [57]. Lipoproteins can also interact with other proteins and function as ligand binding protein, such as LprF and LprJ lipoproteins of *M.tuberculosis* and *M.smegmatis*, which interact with histidine kinase (KdpD) [58]. Furthermore, lipoproteins are involved in the transport and synthesis of cell envelope and be a lipid essential constituent, such as LpqW and LpqX lipoproteins from *M.tuberculosis* [59]. Thus, lipoproteins can be considered as potential targets for drug design. Finally, lipoproteins also can be designed as a potential vaccine target, such as the Lpp20 lipoprotein, the *Helicobacter pylori* vaccine candidate [60, 61].

1.4 Fundamental nutrient pathways of mycobacterium cell envelope

The mycobacterium cell envelope contains a layer of mycolic acid that is unique among all bacteria. This mycolic acid layer functions as a protector for the organism from toxins and other external agents and is suggested to be resistant against several antibiotics and common disinfectants [25]. The function and structure of this layer are equivalent to the outer cell membrane of the Gram-negative bacteria [62]. Gram-negative bacteria use several pathways to transport solutes in and out of the cells. Examples of these pathways include the lipid pathway, where hydrophobic compounds are dissolved in the lipid bilayer temporarily to enable transport [62]. The porin pathway, where specific hydrophilic compounds pass through a protein channel into the cell and by a process called self – promoted, where polycationic compounds disorder the external membrane in order to pass through [63] [64]. Furthermore, some compounds are transported by integral membrane protein transporters, such as FhuA, which is the protein responsible for the uptake of iron ions and the transport of them into the cell [65]. In addition, in *M.tuberculosis*, essential substances are transported throughout the mycolic acid to the periplasm,

and then transported into the cells by lipoprotein transporters, however, it is still unknown how molecules are transported to the periplasm [62].

1.5 *M.tuberculosis* and structural genomics

The aim of structural genomics is to determine the structure of most of the proteins that are expressed in a particular organism [66]. The *Haemophilus influenza* genome was the first to be completed in 1995, while today, there are more than 100 completed sequences of bacterial genomes [67]. The isolation of *M.tuberculosis* was first conducted in 1905 and H37Rv was the first severe pathogenic and widely spread strain to be isolated and used in research [68]. The sequence of *M.tuberculosis* was first completed in 1998 with 3974 genes identified [53]. In 2002 the genome was reannotated and 82 extra genes were added [53, 68]. The accessibility of the *M.tuberculosis* genome enables the life cycle and virulence of *Mycobacterium* to be studied in greater depth. This also has enabled structural genomic programs to be developed on *M.tuberculosis*. In this procedure all or a certain subset of genes from an organism are overexpressed, proteins purified and their structure determined. In favorable circumstances, the structure obtained can be used to predict the function of unknown proteins, or be used in structure based inhibitor design programs to develop new drugs [69].

For mycobacterium organisms, structural studies have led to the discovery of many targets for drug design, such as the mycobacterial domain of carboxyl transferase [70] and the alternative NADH: ubiquinone oxidoreductase (NDH2) [70]. Also, the completion of mycobacterium genome sequencing in 1998 has a positive effect on mycobacterium research as several genes that are involved in the pathogenicity and life of the organism have been identified. Mycobacterium proteins structures in the protein data bank have risen significantly from less than 100 in 1998 to almost 900 in 2010 [71].

1.6 Identification of drug target

According to McDevitt and Rosenberg [67], there are four different factors to be considered in selecting a drug target in an organism. Firstly, a target should be present in a required variety of organisms. Secondly, a target should not be found in humans or if it is, the protein (s) must be totally different [67, 71]. Thirdly, a target should be essential for the growth of the particular bacterium during infection.

Fourthly, the function of a target should be partially known and understood in order to develop a good high - throughput screen [67]. Most of the available antibiotics which are currently in use, interact with specific proteins that are either involved in processes in the cell, such as transcription or the synthesis of the cell wall [67]. Rifampicin is a transcription inhibitor and Ethambutol (EMB) and isoniazid (INH) are cell wall biosynthesis inhibitors, that are all used as the main drugs for targeting *M.tuberculosis* infection [18].

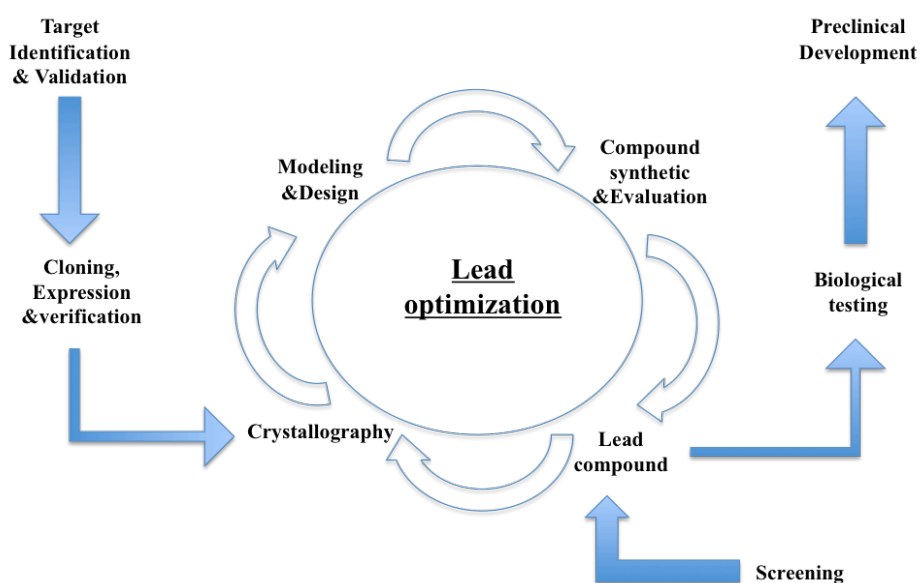


Figure 1.10 The iterative process of structure based drug design discovery.

1.7 Aim of project

The high number of deaths caused by tuberculosis, the emergence of new strains resistant to the available antibiotics, as well as the persistence of the bacteria, have together contributed towards the urgent need for new antibiotics, which have the ability to eliminate the pathogenic strain.

Developing new drugs was a costly (> \$300m) and time consuming (> 20 years) process involving many steps, any of which could fail, leading to a failure in drug development. Therefore, it is very important to find another way to develop new drugs in less time and with less money. Although research into tuberculosis and its causative bacterium has been continuing for a very long time, information about the relationship between the pathogen and the host is still not well understood. Also, aspects of the bacterial life cycle and its interaction mechanisms are also still ambiguous. Thus, the need to understand more about the bacterial life cycle is required in order to be able to identify novel targets for drug design and for use in treatment. Within the proteome of an organism the membrane proteins are still poorly understood biologically compared to globular proteins due to problems in their solubility and isolation. The Mycobacterium genus is composed of about 120 species [72]. *M.tuberculosis* genome encodes about 4000 proteins, and only 8.5 % of these proteins have structures known in the protein data bank, with about 898 *M.tuberculosis* protein structures [68]. Some of which occur in multiple entries, such as shown in Table 1.2. Out of the total number of structures, only 327 structures are unique [68, 71]. Lipoproteins account for about 2.5% of the total number of proteins in the mycobacterium genome [68, 73]. Lipoproteins that are present in mycobacterium species are double the number of lipoproteins that found in other bacteria [52]. Many of lipoproteins have an unknown function [52, 68]. Therefore, the aim of this project was to identify lipoproteins from *M.tuberculosis* based on their attachment to the cell membrane, and then to identify the similar lipoproteins from *M.smegamatis*. The target proteins then enter a process of gene amplification, protein overexpression, protein purification, and crystallization and structure determination, with the aim of identifying their function.

Class	Function	ORFs	Structures in PDB
0	Virulence, detoxification, adaptation	99	20
1	Lipid metabolism	233	31
2	Information pathways	229	25
3	Cell- wall and cell processes	708	31
6	PE and PPE proteins	170	2
7	Intermediate metabolism and respiration	894	143
8	Proteins of unknown function	272	0
9	Regulatory proteins	189	32
10	Conserved hypothetical proteins	1053	43
-	Total protein encoding ORFs	3845	327

Table 1.2 The number of all proteins encoded in the genomic DNA of *M.tuberculosis*. Proteins were classified and arranged according to their functions and solved structures in the PDB [68].

1.7.1 Target selection

All targets were selected for their attachment to the cell membrane and defined as potentially lipid anchored proteins. To determine this, hydrophobicity plots of all the protein sequences encoded within the *M.tuberculosis* genome were generated using the hydropathy scale of Kyte and Doolittle [74]. Each plot was inspected to select the most probable targets that are defined as membrane attached proteins. Targets with a high hydrophobicity in the first 20-50 residues of the protein sequence are considered to be possible lipid anchored proteins. In this way, 55 possible lipid anchored proteins were selected out of 4000 translated proteins. These selected targets were further filtered by checking the availability of their 3D structure. Also, some targets were selected based on their essentiality for *M.tuberculosis* bacterial growth. Essential proteins were identified using Himar1-based transposon mutagenesis [75]. Using this method, three targets were selected as both essential and lipid anchored proteins (Rv1274, Rv2700 and Rv2903c). Further targets (Rv0583c Rv1275 Rv2041c and Msmeg_6050) were identified as lipoproteins and, lastly, Rv1854c, which encodes the NDH2 protein, was added to the list, because of its interest as a drug target [76]. The target genes, which have been studied in this thesis, are shown in table 1.3.

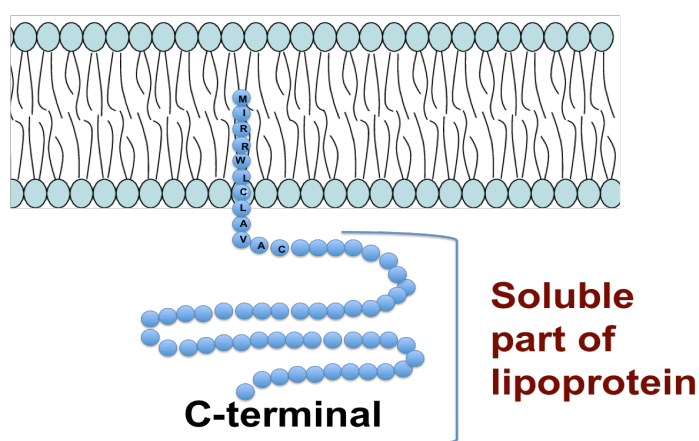


Figure 1.11 A schematic diagram of the predicted way of the lipoprotein membrane attachment. Hydrophobic part that is anchored to the membrane with hydrophobic part of signal sequence while the rest of protein is soluble in the aqueous medium.

No. Of targets	M-Tb targets	<i>M.smegmatis</i> targets	Seq. identity	Number of A. acids (Da)	Gene Length	Predicted function
1	Rv1854c	Msmeg_3621	63%	463	1392	Probable NADH dehydrogenase (NDH2)
2	Rv1274	Msmeg_5007	69%	185	558	Unknown
3	Rv2700	Msmeg_2761	63%	216	651	Unknown
4	Rv2903c	Msmeg_2441	76%	294	885	Probable signal peptidase I (LepB)
5	Rv0583c	Msmeg_5456	48%	228	687	MK35 unknown
6	-	Msmeg_6050	-	300	903	Solute-binding lipoprotein
7	Rv1275	Msmeg_1395	27%	193	582	Unknown
8	Rv2041c	Msmeg_0515	22%	425	1278	Probable sugar transporter sugar binding lipoprotein

Table 1.3 All targets were selected to run this project. Eight targets were selected to work on from *M.smegmatis* with their similar proteins from *M.tuberculosis* and their protein sequence identity and length with their predicted functions.

Chapter 2

Material & Methods

This chapter gives a general description of all the materials and procedures that have been utilized to run this project including the experimental techniques for cloning, over expression, purification, and crystallization of all targets from *M.smegmatis*.

2.1 DNA cloning

2.1.1 Recombinant DNA technology and protein production

The study of protein structure using X-crystallography usually requires a high amount of soluble, stable and well-folded pure protein. As producing enough protein from the native source is difficult and can be dangerous (in case of pathogenic organisms), molecular biology is routinely used to produce high amounts of the required protein target for use in structural studies. Recombinant DNA technology has enabled us to extract and amplify the gene of interest from any organism, by introduction into an expressing plasmid, which is then transformed into a host to be expressed, thus producing the required amount of a target protein.

2.1.2 Genomic DNA Extraction of *M.smegmatis*

An *M.smegmatis* streak plate was obtained from the laboratory of Prof. J.Green in Sheffield. A colony was inoculated using a sterile loop into 5 ml of Luria-Bertani (LB) medium (Table 2.1), and incubated at 37°C for 3 days. 1 ml of this culture was used to extract genomic DNA using a keyPrep bacterial genomic kit (ANACHEM). The manufacturer's protocol was followed to give a final concentration of 20-70 µg in 100 µl of genomic DNA for use in the Polymerase chain reaction (PCR).

Component	g/L
Lysogeny broth media	20
Tryptone	10
Yeast extract	5
Sodium chloride	5
Agar	
Tryptone	10
Yeast extract	5
Sodium chloride	5
Agar	15

Table 2.1 The composition of agar and lysogeny broth media.

2.1.3 Polymerase chain Reaction (PCR)

In order to successfully amplify the gene of interest, PCR is routinely used. The components required are the DNA template, the primers designed to be complementary to the DNA template region, heat stable DNA polymerase, such as Tag polymerase, deoxynucleoside triphosphates (dNTPs), deionized autoclaved water and an appropriate buffer containing Mg^{+2} , to set up the reaction.

The PCR proceeds through a number of cycles for amplification. Every cycle is composed of three stages, completed by varying the reaction temperature. The first stage is denaturation, where the reaction temperature is raised to 95°C in order to break hydrogen bonds between DNA bases to end with single stranded DNA. The second stage is annealing, where the reaction temperature is decreased to less than the DNA primers melting temperature, allowing the primers to bind to their complementary bases on the region of the DNA template. The final stage is elongation, where the DNA polymerase enzyme binds to the DNA and moves forward along the DNA, in order to produce a new duplicate of the preferred region and a temperature suited to the polymerase used.

The extracted *M.smegmatis* genomic DNA was used to amplify all target genes in a PCR reaction of Biomix Red 25 µl as final volume (Bioline) or 50 µl total volume using standard PCR components as shown in (Table 2.2). For each reaction 7 µl of genomic DNA was used and 1µl of 10 pmol forward and reverse primers were added. The reaction was set up using gradient PCR machines as shown in table 2.3.

Initial Denaturation	95°C for 5 minutes
Further Denaturation	95°C for 30 seconds / 35 cycles.
Gradient Annealing	Start from 55°C to 65°C. 5°C < than Tm of primer pair.
Elongation	72°C for 1 minute.
Final Elongation	72°C for 7 minute.

Table 2.2 Gradient PCR condition of the *M.smegmatis* genes amplification.

Reagents	Volume
DNA template	1-7 µl
Forward and reverse primers	1ul each of 10pmol stock.
10mM dNTPs	1µl
10x buffer contain MgCl ₂	5µl -10 µl
ddH ₂ O	To 50 µl total volume.

Table 2.3 The composition of a standard PCR reaction mixture

2.1.4 pET vectors

A large variety of different expression systems have been designed. One example of these systems is the pET system (Novagen). The pET28a vector (Figure 2.1) has been used exclusively in this project. This vector contains multiple restriction sites to be used for cloning and the T7 promoter sequence. The lac operator, ribosomal binding site and a T7 terminator sequence are located after the T7 promoter. The BL21 (DE3) bacterial strain was used as a host for protein over expression.

2.1.5 Cloning with pET28a

The pET28a plasmid was used to clone all gene targets with restriction enzyme digestion. Primers were designed for each gene with restriction enzyme sites *NdeI* and *HindIII* in the forward and reverse primers, respectively, based on their presence in the pET28a plasmid. The primers were also designed to incorporate a His-tag sequence at the N or C-terminal of the protein. As a result of digestion of both plasmid and gene of interest, a sticky end of both plasmid and insert was used for the ligation reaction using T4 DNA ligase. Successful clones were then transformed into BL21 (DE3) *E.coli* cells that lack the T7 polymerase gene, to avoid any disturbance background expression that may possibly be problematic.

2.1.6 pET expression hosts

The *E.coli* DE3 strain is considered as a vehicle that contains the essential genes for protein expression in its chromosomal DNA, including T7 RNA polymerase gene copy and the *lacI* gene. IPTG is used to induce RNA polymerase that is under lacUV5 control. Once the polymerase is expressed, plasmid DNA can be transcribed into mRNA, which is translated and results in a large quantity of the desired protein (Figure 2.1).

2.1.7 Purification of PCR product

In order to visualize the resultant PCR product, a 1% TAE agarose gel was run at 100 V for 50 minutes with Ethidium bromide. The positive bands that appeared at the expected size were cut from gel and transferred to an eppendorf tube. DNA was recovered from these samples using QiaQuick® Gel Extraction Kit (Qiagen) following the manufacturer's protocol. Briefly, The protocol used was; add three volume of buffer QG and incubate at 50°C for 10 minutes. Then add 1 volume of

Isopropanol with mixing. The mixture is then transferred into a QIAquick® column and centrifuged at 14,000 xg for 1 minute. All the liquid was discarded and 500 µl of QG buffer was added to the column and centrifuged again at 14,000 xg for 1 minute. Then the column was washed with 750 µl of PE buffer. The flow through was discarded after centrifugation, of the columns twice, to make sure that all ethanol was removed from the column. The column was transferred into a new Eppendorf tube and DNA was eluted by adding 50 µl of EB buffer and incubating at room temperature for 2 minutes and centrifugation for 1 minute. The final concentration of DNA extracted varied between 20-50 ng µl⁻¹.

2.1.8 Restriction digestion for vector pET28a and PCR product

In order to be able to produce a sticky end for cloning, both vector and PCR product were double digested with 4 units of restriction enzymes NdeI and HindIII (NEB) (Table 2.4). The reactions were incubated at 37°C for 3 hours. Then a 1% TAE agarose gel was run at 100 V for 50 minutes in order to remove unwanted enzymes. DNA was then extracted out of the gel by using QiaQuick® Gel Extraction Kit (Qiagen) and stored at -20°C.

Component	Volume
Vector/PCR product	1 µl (5ng/ul)
10x NEB Buffer 4	2 µl
NdeI enzyme	1 µl
HindIII enzyme	1 µl (4-12 U/µl)
DdH2O	To 20 µl
Total	20 µl

Table 2.4 The composition of a typical double restriction digestion reaction.

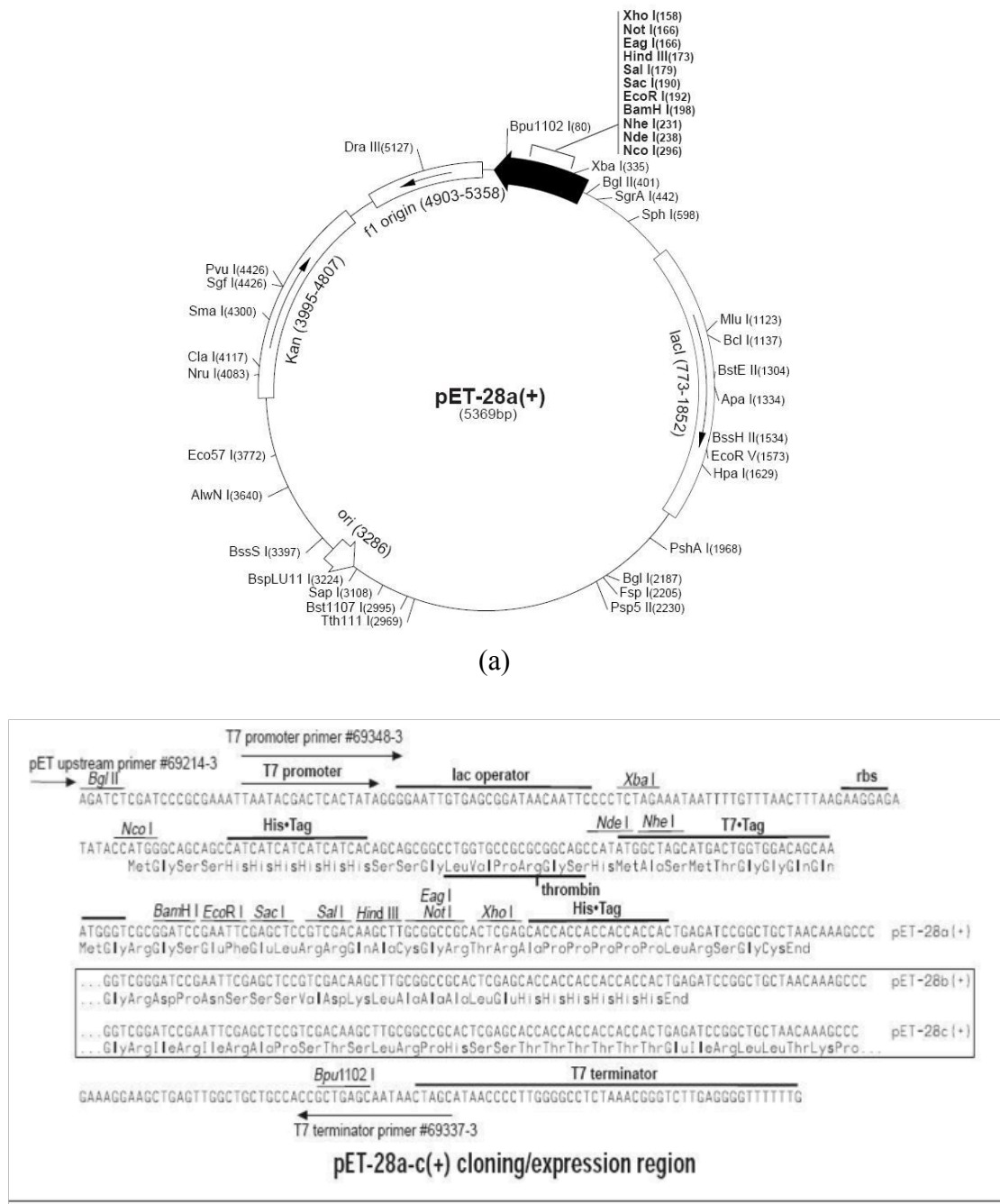


Figure 2.1 The pET28a plasmid. **a.** A circle map of the plasmid shows the essential cloning regions, such as a copy of *lacI* gene and multiple cloning sites (black arrow), replication regions (f1 phage) and pBR322, antibiotic selection gene (Kanamycin) and selection of two three tags (N and C terminal 6xHis tags) and one T7 tag. **b.** The plasmid cloning sites include a T7 promoter sequence and a lac operator sequence and ribosomal-binding site.

2.1.9 DNA Ligation and transformation into DH5 α

The ligation reaction of the target insert into the pET28a vector was performed in an Eppendorf tube containing different concentrations of plasmid and insert (3:1, 1:3 and 3:0). The reaction was also done using T4 ligase enzyme with its appropriate buffer and incubated at room temperature for one hour. After incubation, 1-5 μ l of the ligation mixture is added into 50 μ l of *E.coli* strain DH5 α that had been defrosted gently on ice for 5 minutes. The transformed mixture was incubated on ice for 2-5 minutes and then heat shocked at 42°C in a water bath for 90 seconds. Then, the mixture was placed back on ice for 2 minutes. 450 μ l of LB medium containing 50 μ g ml⁻¹ kanamycin (room Temp) was added to 50 μ l of cells and incubated at 37°C with shaking at 200 RPM. After one hour cells were centrifuged at 3.600 xg for 2 minutes. The LB medium was then discarded and 50 μ l of fresh LB was used to re-suspend cells, which were spread onto 50 μ g ml⁻¹ kanamycin agar plates.

Component	Amount
DNA 50ng/ μ l Msmeg_gene /pET28a	5-10 μ l/5 μ l
10X T4 Ligase	2 μ l
T4 Ligase	1 μ l
ddH ₂ O	To 20 μ l
Total	20 μ l

Table 2.5 The composition of a typical ligation reaction.

2.1.10 Confirmation of successful cloning

In order to confirm successful cloning, several colonies were inoculated into separate 5ml of LB medium containing 50 μ gml⁻¹ kanamycin and incubated at 37°C overnight. The cells were pelleted by centrifugation at 3.600 xg for 5 minutes and the supernatant was discarded. Vector isolation was carried out using a QIAprep Spin MiniPrep Kit (QIAGEN) by following manufacturer's protocol. Cells were re-suspended with 250 μ l of buffer P1, containing RNase. Then 250 μ l of alkaline lysis buffer P2 was added followed by 350 μ l of buffer N3. The samples were centrifuged at 13,000 RPM for 10 minutes to remove broken cells debris. The supernatant was applied onto miniprep spin column and centrifuged again at 13,000 RPM for 1 minute. The column was washed by 740 μ l of buffer PE and centrifuged twice to remove ethanol. Finally, column was transferred into clean 1.5 eppendorf tube and

DNA was eluted by 50 μ l of buffer PE and incubated at room temperature for 2 minutes before being centrifuged for 1 minute. The purified vector was digested using *NdeI* and *HindIII* as described previously, and incubated at 37°C for one hour. The resultant cut vector was checked using 1% TAE agarose electrophoresis gel run at 100 V for 50 minutes.

2.2 Protein over-expression

2.2.1 Transformation

The competent *E.coli* strain BL21 (DE3) is normally used for the over expression of the desired protein (Novagen). The plasmid containing the desired insert was transformed into competent cells by the heat shock method. After removing an eppendorf tube containing 20 μ l of BL21 (DE3) from -80°C and defrosted on ice, a small amount of the desired plasmid DNA (10 ng/ μ l) was added to the cells and incubated on ice for 15-30 minutes. The mixture was then incubated at 42°C for 30 seconds and returned back to ice for 2 minutes. 480 μ l of LB was added to the cells and incubated for one hour at 37°C, before being plated onto LB agar containing 35-50 μ g ml⁻¹ kanamycin.

2.2.2 Small-scale over-expression trials

A small-scale over-expression was carried out to identify the best conditions for each protein. This was done by changing the induction time, temperature and the IPTG concentration. The successful optimized condition was then used to scale the culture up, to produce enough protein for structural studies. A starter culture made with an appropriate antibiotic (50 μ g ml⁻¹ kanamycin) and incubated overnight at 37 °C. This culture was then used to make a secondary culture, (with 1-2% inoculation) with the same selected antibiotic. The secondary culture was incubated at 37 °C and shaker at 250 RPM. Once an OD₆₀₀ of 0.5-0.8 was reached, several concentrations of IPTG (0.1-1 mM), temperature (4-37 °C), shaker speed (150-250 RPM) and time of induction (1-24 hours) were tested to identify the best conditions. A small amount of cells (1.5 ml) were taken before induction to be used for the expression analysis. After expression, all induced cultures were centrifuged at 17,000 xg for 10 minutes. The supernatant was discarded and the pellets stored at -20°C. The cells paste was removed from the freezer and re-suspended in lysis buffer or bug buster (Novagen) and incubated for 30 minutes to break the cell membrane. The soluble fractions of

protein were separated from cell debris and insoluble protein by centrifugation at 70,000 xg for 10 minutes. The supernatant, the cell debris and the un-induced samples were all analyzed for protein expression by running on SDS PAGE.

2.2.3 Large-scale over-expression

After finding the best condition for protein expression, a large-scale over-expression was carried out. The primary culture of 50 ml was grown overnight at 37 °C in a 250 ml conical flask with an appropriate antibiotic and used to inoculate 500 ml LB media in 2l flask to make the secondary culture. Then protein was induced based on the small-scale expression results and centrifuged at 5,000 xg for 30 minutes. The supernatant was discarded and cell paste were stored at -20 °C for protein purification.

2.2.4 Production of Seleno-L-methionine incorporated proteins

A primary culture was grown overnight in LB medium containing 50 µgml⁻¹ kanamycin and used for 2% inoculation of (3 l) secondary culture supplemented with 500 µg per 500 ml LB medium in (2l) flask. After reaching a suitable optical density (OD₆₀₀ of 0.6), cultures were harvested by centrifugation at 5,000 g for 30 minutes. The cell pastes were re-suspended in minimal media containing Potassium dihydrogen phosphate (KH₂PO₄ 4.5 g/l), Dipotassium phosphate (K₂HPO₄ 10.5 g/l), Tri-sodium citrate (Na₃C₆H₅O₇, 0.5 g/l), Ammonium sulfate ((NH₄)₂SO₄ 1 g/l), Adenine (0.5 g/l), Guanosine (0.5 g/l), Thymine (0.5 mg/l), (Uracil 0.5 mg/l), Magnesium sulphate (MgSO₄·7H₂O, 1 g/l), Thiamine (4 g/l), L-lysine (100 mg/l), L-phenylalanine (100 mg/l), L-threonine (100 mg/l), L-valine (50 mg/l), L-isoleucine (50 mg/l), L-leucine (50 mg/l) and Glycerol (5.0 g/l). Then cells were pelleted again for 30 minutes at 5.000 g and resuspended again in minimal media containing 40 mg/L of Seleno-L-methionine. The culture was then divided into several 2 l conical flasks and incubated at 37 °C until OD₆₀₀ reached 0.8. The cultures were induced with IPTG with a doubling of the induction time. The cells were harvested and the cell paste was stored at -20 °C for protein purification.

2.3 Protein purification techniques

In this project two techniques were used to purify the various proteins and are explained as follows.

2.3.1 Cell disruption

Cells were disrupted by sonication (3x-20 seconds) at a volume of 16-micron amplitude on ice in order to purify a protein of interest. Sonication is a technique where the high frequency of sonic pulses is used to break the cell membrane of bacteria. The cell paste was re-suspended first with a proper buffer before cell lysis. The cell debris, insoluble protein and soluble protein were separated by centrifugation at 70.000 xg for 10 minutes and the supernatant used for protein purification.

2.3.2 Nickel- NTA chromatography

The first technique was used in this project is Ni- affinity chromatography. The protein of interest contained six histidine residues at its N or C-terminus, known as a 6-His-tag. This tag has the ability to interact with nickel bound to agarose beads by nitroloacetic acid (NTA) creation inside the column. Molecules that have no affinity or low affinity to the beads are first to be eluted easily from the column either by washing with buffer or with a low concentration of Imidazole. Molecules that have high affinity to the beads are then eluted by increasing the Imidazole concentration. Commonly, a 250 mM concentration of Imidazole is usually able to elute His tagged proteins.

2.3.3 Size exclusion chromatography (Gel filtration)

Further purification of the desired protein can be achieved by applying the sample through a 16x60 Superdex200 gel filtration column (GE Healthcare). This technique is based on the theory of separation of molecules upon their shapes and sizes. The gel filtration column contains porous beads that can be loaded first with small molecules. Larger molecules stay outside the beads, as they are larger than the pores of the beads. Thus they are likely to be eluted out first from the column.

2.3.4 Sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS PAGE)

SDS PAGE was used to evaluate protein purification progress by loading 20 µg of protein sample in 4x NuPAGE buffer (Invitrogen) containing a 10x reducing agent (such as β- mercaptoethanol) and 2 µl of bromophenol blue dye. The protein mixture was boiled before loading to allow the protein to be denatured, at 100 °C for 5 minutes. Two types of SDS PAGE were used, a 12% resolving gel and a 6% stacking gel (Table 2.6). The mark 12™ (Invitrogen) protein standards were used to evaluate the protein bands size. The SDS gel was run at 200 Volts for 50 minutes in 1x SDS buffer. Then, the gel was removed, washed and stained with 0.1% (w/V) Coomassie Brilliant Blue, in 20% (v/v) methanol and 10% (v/v) acetic acid and put on a shaker to clarify the protein bands.

12% resolving gel	Gradients
30% Acrylamide / bisacrylamide (29:1)	2.5 ml
1 M Tris- HCL buffer, pH 8.8	2.35 ml
10% SDS solution	62.5 µl
10% Ammonium persulphate solution	6.25 µl
TEMED	6.25 µl
MilliQ water	1.28 ml
6 % stacking gel	
30% Acrylamide / bisacrylamide (29:1)	0.75 ml
1 M Tris- HCL buffer, pH 6.8	0.47 ml
10% SDS solution	37.5 µl
10% Ammonium persulphate solution	37.5 µl
TEMED	3.75 µl
MilliQ water	2.46 ml

Table 2.6 Recipes of 12 and 6 % SDS Page ingredients.

2.3.5 Protein concentration

The use of X-ray crystallography technique requires crystals, prepared from a high concentration solution of protein. Therefore, in order to increase the concentration of a given protein a Vivaspin sample concentrator (size based on molecular weight of protein 10.000 to 30.000 Da) (GE Healthcare) was used. It is composed of two filters separated by a membrane of polyethersulfone that contains pores of several sizes. The centrifuge is used to force the solution to pass through this membrane. A small molecule together with the solvent will pass thorough the membrane but larger molecules will not be allowed to pass and will remain on the top of the membrane leading to a more concentrated solution. Also, this concentrator can be used to exchange the buffer before crystallization trials.

2.4 Protein Crystallization

2.4.1 Introduction

The determination of the molecular structure of the protein can be carried out using the popular method of X-ray crystallography. In this method, X-rays are diffracted by atoms of the protein in a single crystal. If the intensities of the X-rays are measured and the phase of each reflection can be determined, an electron density map of the crystal can be calculated, leading to a molecular model. However, in order to successfully crystallize a protein, several conditions should be met, including a high amount of concentrated, pure, stable and correctly folded protein and a set of buffer and precipitant conditions, which favor the crystalline form of the protein.

Crystallographic theory is covered in a number of textbooks, including Crystallography Made Crystal Clear [77]. This chapter will cover the methods used in the determination of the protein structures described in this thesis.

2.4.2 Protein crystals

A protein crystal consists of millions of molecules repeated regularly in a specific array of three dimensions, known as a lattice. All molecules inside the protein crystal are interacting with those around them by means of non-covalent interactions, such as hydrogen bonds and the ionic interactions. A protein crystal can contain between 30-90% solvent [78] and thus is fragile and easily broken. The crystal can be defined by its unit cell, which is the simplest and smallest unit that can be repeated by translations along x, y and z. Each unit cell is measured by the length of its three axes, a, b and c and the angles between them, α , β and γ . Furthermore, each unit cell can have internal symmetry, such as rotations. In these cases, the unit cell can be further broken down to the asymmetric unit, which is related to the unit cell by the symmetry operation of the cell. The seven crystal systems with their measurements are described in (Table 2.7). However, along with these crystal systems, there are four different types of lattice: primitive (P) where only one lattice point is found in each corner of the unit cell and body centered (I) where another point in the center of each cell unit is included. The last two lattice types are called all-face centered (F) and c-face centered (C). Because protein molecules are chiral, they can only be crystallized in one of 65 of the total of 230 space groups.

One common method for growing protein crystals is the vapour diffusion method, which was used exclusively in this project. In this method the protein solution is mixed with the precipitation solution in a normally, 1:1 ratio and sealed in the presence of a much larger reservoir of precipitation solution. During the course of the experiment, the protein concentration gradually increases, as water vapour passes from the crystallization drop to the reservoir. As this process occurs, the protein gradually enters the supersaturation zone (Figure 2.2).

Crystal class	Axis system
Cubic	$a=b=c, \alpha = \beta = \gamma = 90^\circ$
Tetragonal	$a=b \neq c, \alpha = \beta = \gamma = 90^\circ$
Trigonal	$a=b, \alpha = \beta = \gamma = 120^\circ$
Hexagonal	$a=b \neq c, \alpha = \beta = 90^\circ, \gamma = 120^\circ$
Orthorhombic	$a \neq b \neq c, \alpha = \beta = \gamma = 90^\circ$
Monoclinic	$a \neq b \neq c, \alpha = \gamma = 90^\circ, \beta \neq 90^\circ$
Triclinic	$a \neq b \neq c, \alpha \neq \beta \neq \gamma \neq 90^\circ$

Table 2.7 The seven protein crystal system lattices.

2.4.3 Growing protein crystals

Producing a crystal is difficult because of several factors that affect protein crystallization, such as the size, stability, mobility and purity of proteins. In order to successfully form protein crystals, several parameters should be considered; the concentration and purity of the protein and the solution used for protein crystallization, such as the buffer, pH, temperature and other factors that are hard to be controlled or manipulated, such as the age and source of purified protein, sound, vibration and the source of the protein as well as the presence of a ligand. Other factors, which are significant for forming crystal, are the type and concentration of the precipitant agents, such as the salts, organic solvents and agents with high molecular weight, such as poly ethylene glycol (PEG). Moreover, many attempts must be performed in order to identify the proper conditions for the protein to form a crystal in combination with all these agents.

As a crystal grows, the solution passes through two different phases until the crystals are formed (Figure 2.2). The first phase is known as the nucleation phase, where

molecules clump together to form nuclei (the nucleation zone). As the experiment continues, the concentration of the precipitant further increases, and the protein concentration decrease, as either crystals (metastable zone), or amorphous precipitate (precipitation zone) are formed.

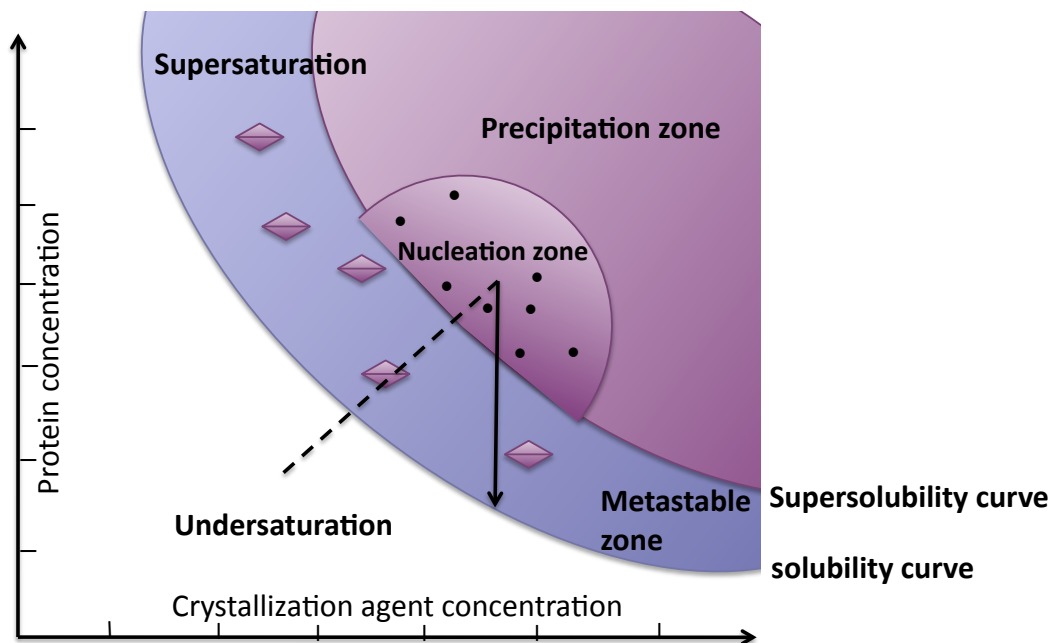


Figure 2.2 A diagram illustrating the protein crystallization phases. As a crystal grows, the solution passes through two different phases (zones) until the crystals are formed. As the experiment continues, the concentration of the precipitant further increases, and the protein concentration decrease, as either crystals (metastable zone), or amorphous precipitate (Precipitation zone) are formed.

2.4.4 Vapour diffusion methods

Two main vapour diffusion methods are commonly used, the sitting drop method and the hanging drop methods (Figure 2.3).

2.4.4.1 Sitting drop method

In this method a small volume of protein solution is mixed with a similar volume of precipitant solution and placed adjacent to a reservoir of the precipitant solution and the system is sealed. This method is easily automated and for this project a Matrix Hydra II PlusOne robot was used to screen against a series of commercially available crystallization conditions in 96 well crystallization plates (Qiagen®). The robot dispensed 200 μ l of protein sample and 200 nl of precipitant solution into the drop. The plates were sealed with tape and stored at 17 °C.

2.4.4.2 Hanging drop method

This method is a variation on the sitting drop, where the protein and precipitant are mixed on a cover slip, which is then inverted over a well containing the precipitant solution and sealed with oil or grease. This method is ideal for optimizing promising conditions seen in initial robot screens and volumes are larger, with typically 1 μ l of protein solution over 1 ml of reservoir.

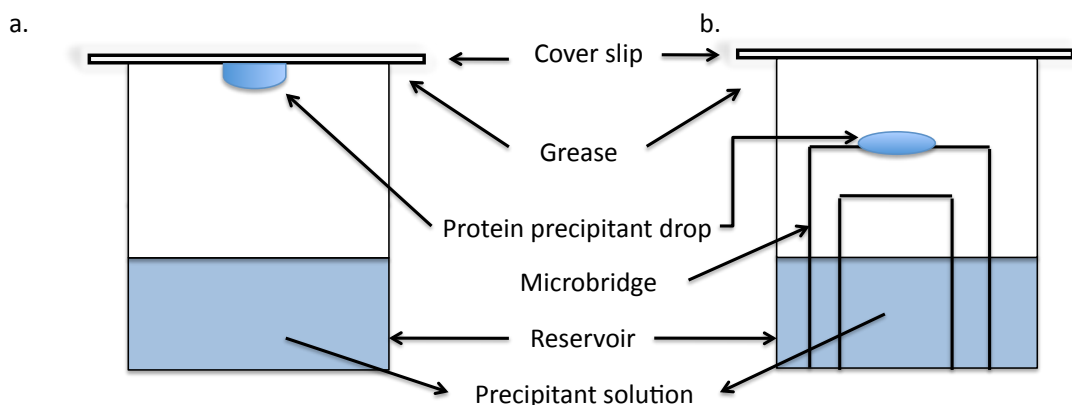


Figure 2.3 A diagram illustrate the main techniques used in protein crystallization. (a) The hanging drops crystallization method. (b) The sitting drops crystallization methods.

2.5 Crystal mounting and cryoprotectants

When X-rays hit protein atoms, free radicals are produced, which react with the protein, breaking bonds and gradually degrading the crystal [77]. This radiation damage can be decreased by keeping the sample at very low temperature, which slows the diffusion of the radicals through the crystal lattice [77]. However, the crystal lattice can be disrupted by the formation of ice around the crystal, which damages the diffraction pattern. Therefore, crystals are placed in a special cryoprotectant solution that contains the crystallization conditions and a small molecule cryoprotectant, such as glycerol, low molecular weight PEG or alcohols, to stop ice crystals forming when the crystal is cooled. The crystal is looped from the drop in a fiber loop, placed for a short time in cryoprotectant solution and then either placed directly in a stream of nitrogen gas at 100 K on the diffractometer, or flash frozen in liquid nitrogen and stored for later use on a synchrotron source.

2.5.1 Data collection apparatus

Different detectors are used for collecting data from a single protein crystal and in this thesis, two types of X-ray source were used, the Rigaku MM007 copper-rotating anode system in Sheffield and the Diamond Light Source near Oxford.

2.5.1.1 Copper-rotating anode system

In a rotating anode, electrons are produced by a heated filament (cathode) and then accelerated in a vacuum through a large potential difference (40 kv) to strike a metal anode, which in the case described here was copper. The impinging electrons have sufficient energy to expel an electron from the K shell of the copper anode. A high-energy electron from a different shell of copper falls back into the vacant K shell, and the excess energy is emitted as an X-ray. For copper, the L-K transition is the most common, resulting in X-rays of a wavelength of 1.54 Å. These are separated from other, less common, energies by the use of a monochromator, Ni filter or coated focusing optics.

2.5.1.2 Diamond light source (Synchrotron)

Additionally, the Diamond light source was used throughout this project for high-resolution data collection. In a synchrotron, electrons are injected by an accelerator into a storage ring (Figure 2.4). In the storage ring the electrons are kept in a continuous circular motion by bending magnets. As the direction of the electron beam is changed by the magnet, electromagnetic radiation is given off tangentially to the storage ring, and the X-ray part of the spectrum is used for the diffraction experiments. Additional magnetic devices called undulators or wigglers can be inserted between the bending magnets; these can increase the intensity of the X-rays. The use of Synchrotron radiation has reduced the time of data collection to minutes compared to > 24 hours using a rotating anode. As different X-ray energies (wavelengths) can be selected, synchrotron radiation is ideal for anomalous scattering experiments, as energy close to the absorption edge of the anomalous element can be chosen.

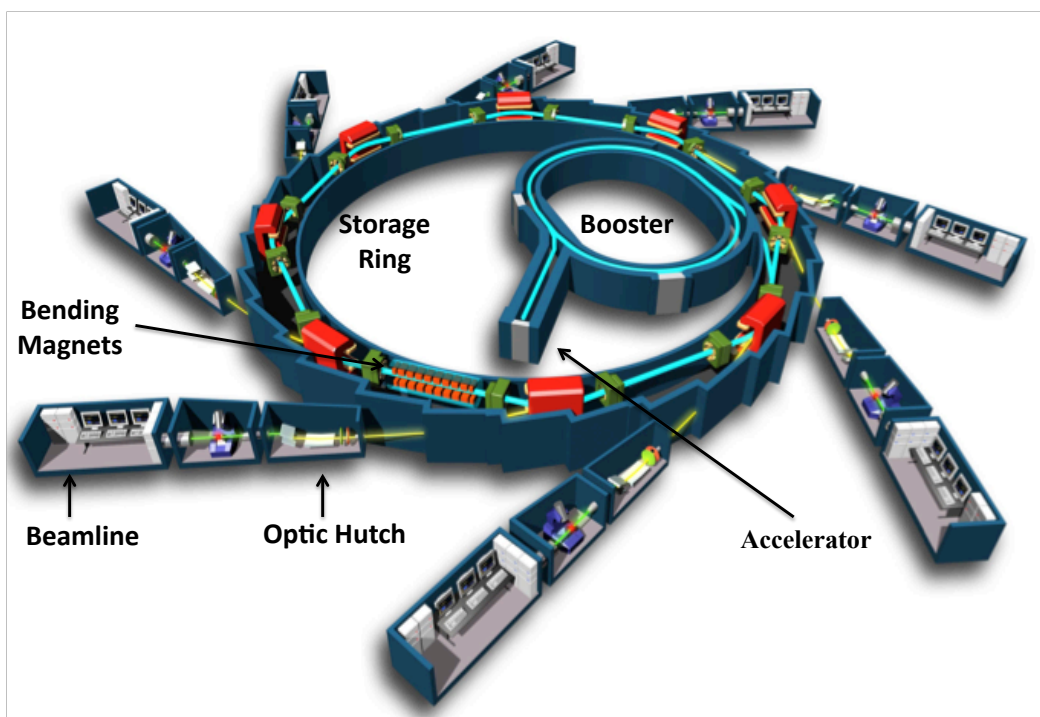


Figure 2.4 A schematic diagram of the Diamond Light Source (*synchrotron*). Figure adapted from the College of Life Sciences website, University of Dundee.

2.5.1.3 Detectors

In this thesis two detectors were used in the diffraction experiments. For the in house data collection a Mar Research image plate was used. In this detector, the diffracted X-rays hit a thin phosphor layer on the detector surface, doped with Eu^{2+} . On irradiation, the Eu^{2+} loses an electron, forming Eu^{3+} . At the end of the exposure, the image is read out, by scanning the detector with a red laser ($\lambda=633$ nm), which returns the Eu^{3+} to Eu^{2+} , releasing radiation of 390 nm, which is counted by a photomultiplier.

For data collected at the Diamond synchrotron, Pilatus pixel detectors were used. These detectors use hybrid pixel technology, where direct detection of the X-rays is achieved using a silicon solid state sensor bonded to a CMOS readout chip. This results in a high dynamic range, zero readout noise, and very short readout times, resulting in high signal/ noise ratio and the ability to use shutterless data collection methods, where the crystal is continuously rotated and the images read out every 0.1 second.

2.5.2 Data collection strategy

A diffraction test is usually carried out to collect information about the crystal and its symmetry. For a data collection experiment, the rotation technique has been used throughout this study. This technique is based on a slight incremental rotation of the crystal during the X-ray beam exposure, so that all the diffracted beams can be measured. Several factors should be considered for data collection to assure that a complete data set is collected. The first factor is the crystal symmetry. According to Friedel's law a complete data set can be collected in a maximum of 180° rotation. This rotation can be reduced by the crystal symmetry, which is described by the space group of the crystal. The second factor is the size of the rotation angle for each exposure called the incremental ϕ angle. This angle used depends on the unit cell dimensions. The length of the unit cell dimensions contributes to the space between each reflection; for instance, the crystal with long dimensions will result in very small separation between the reflections. Therefore, a reduction of the rotational angle for each image can decrease reflection overlap. The last factor that should be considered is the strength of the X-ray beam. Stronger X-rays are preferred for data collection, as they will result in high reflection intensities and improved resolution limit for the collected data.

2.5.3 Data collection

Data collection on single protein crystals was carried out after transferring the crystal from its mother liquor into a cryoprotectant solution, which consisted of the mother liquor plus 20-30% of ethylene glycol. X-ray diffraction images were taken at different ϕ angles of 0° and 90° and then analyzed by the use of the auto-indexing routine in MOSFLM software [79]. MOSFLM estimates the dimensions of unit cell, mosaicity and the predicted space group. Using this information, the strength of beam and exposure time a strategy of data collection was made. For high quality data collection, crystals were sent to the Diamond synchrotron source of beam lines I02, I03, I24 and I04-1.

2.6 Processing data from diffraction images

All data obtained through this project were automatically indexed and integrated by the use of the software MOSFLM and XDS [79, 80]. At this step, the program analyzes the diffraction patterns to determine the distances between all reflections and their symmetry to produce unit cell dimensions and the space groups they belong to. Among the several selections of predicted solutions, that with the highest possible symmetry was routinely selected. Further MOSFLM refinement of the unit cell dimensions was undertaken using more diffraction images. Using these parameters, the data for every image was integrated, using Mosflm [79] or the xia2 pipeline [81] at Diamond, which uses XDS [80] for integration. The data were scaled, merged and analyzed with SCALA [82]. Phase determination was undertaken using either phaser [83] for molecular replacement or SHELX C, D and E for SAD experiments [84, 85]. These techniques are described in the appendix.

2.7 Structure building and refinement

The COOT program was first used to visualize the electron density map and to build a model of the protein structure backbone [86, 87]. Then, the suite of CCP4 programs was used to refine the model to be in agreement with the electron density. The structure refinement was carried out using REFMAC5 in the CCP4 suite [88]. Several cycles of structure refinement and rebuilding were run in order to improve the agreement between the model and the data. In each round of rebuilding the position of the amino acid residues were adjusted to the map, and water molecules and other solvent were added to features in the difference electron density map, where they made sensible interactions with the protein. Progress of refinement was made by comparing the change in R-factor and free R-factor.

Chapter 3

Cloning, over expression, purification and crystallization of protein targets from *M.smegmatis*.

This chapter describes the bioinformatics study on the selected targets of the *M.smegmatis* structural genomics project and their progress of cloning, over expression, purification and crystallization. The target that successfully led to a solved structure will be discussed in chapter 4.

3.1 Target selection

The rationale behind the project described in this thesis was to determine the structure of proteins associated with the membrane by either an N or C terminal helix, or by a lipid anchor. It was hoped that by analysis of the sequences of these types of protein that constructs could be designed that were solely the soluble part of these proteins, and thus might be more favorable towards crystallization than the entire gene product. The initial analysis was thus a hydropathy plot of all of the proteins encoded by the *M.smegmatis* genome.

3.1.1 Hydrophobic plots analysis

The hydropathy scale is a commonly used measure to define hydrophobic sites along protein sequence and is calculated by the relative hydrophobicity of the amino acids [74]. The hydrophobicity is calculated for a window of 19-20 residues, stepping along the sequence. A value > 2.5 for this window often indicates the presence of a transmembrane alpha helix. However, such analyses will also identify N-terminal lipid anchored proteins as they usually contain a hydrophobic portion in the signal sequence, prior to the cysteine residue that is the site of lipidation.

3.1.2 Generating a truncated protein for selected targets

In order to successfully obtain a highly soluble and stable protein that is suitable for crystallization and biochemical studies, attempts were made to truncate all target proteins by cleaving out the hydrophobic part located either at the N-terminal or C-terminal regional of the protein sequence. For this purpose, a wide range of programs was used to help predict the right site for truncation. These included the lipoP 1.0 server that helps to predict and identify a lipoprotein signal peptide, and Signal-3L, SignalP and TMHMM that are designed to predict signal peptides and transmembrane helices for a given protein sequence, respectively. Furthermore, the Phyre server was also used to help predict the secondary and tertiary structure of the unknown protein, based on its homology with known structures in the protein database [89]. Based on these studies, seven targets were selected, namely Msmeg_1395, Msmeg_2441, Msmeg_3621, Msmeg_5007, Msmeg_6050, Msmeg_5456, and Msmeg_0515. Five of these targets are described individually.

3.2 From DNA to protein

3.2.1 Target amplification

The *M.smegmatis* bacterial agar was used to inoculate 5 ml of primary culture which was then incubated at 37°C for 3 days. The culture was checked by eye every day for contamination, as the culture had no antibiotic selection. After 3 days, *M.smegmatis* genomic DNA was extracted using the keyPrep bacterial genomic kit (ANACHEM®).

For each individual target, forward and reverse primers were designed for constructs encoding both the full length and truncated proteins, lacking the hydrophobic residues at the N- or C-terminal region. The selected target gene sequences were obtained from the KEGG database [90]. Both forward and reverse primers were 27-30 base pairs in length and were synthesized by Eurofins MWG Operon Company. All synthesized primers were stored as a stock solution of 100 µM at -20 °C. The working stock was prepared by diluting of the original stock to 10 µM using sterilized water. To amplify the gene of interest, gradient PCR was used testing different annealing temperatures to find the best conditions for gene amplification. The purity of the resultant PCR product was evaluated and checked using 1% agarose gel electrophoresis. The PCR product was extracted from the gel using a gel extraction kit as mentioned detailed in section 2.1.7.

Target	Annealing temperature (°C)
Msmeg_3621	61
Msmeg_5007	58
Msmeg_2441	61
Msmeg_2761	59
Msmeg_6050	61
Msmeg_5456	59
Msmeg_1395	58

Table 3.1 The best PCR annealing temperature used for gene amplification of the target full-length and truncated proteins.

Oligoname (restriction site)	Sequence
NDH2 forward (NdeI) (T)	5' GCGCCATATGAGCCATCCC GGAGCTACGGC3'
NDH2 reverse HindIII	5'-GCGCAAGCTTGGACGCGGCTTTCTCGGTGT-3'
Msmeg_5007 forward NdeI (F)	5'-GCGCCATATGTTGGCGGTTCTGGCG-3'
Msmeg_5007 reverse HindIII	5'-GCGCAAGCTTAAACCTTAAGTGTTA-3'
Msmeg_6050 forward NdeI (F)	5'-GCGCCATATGGTGACGCGGCGGTGC-3'
Msmeg_6050 forward NdeI (T)	5'-GCGCCATATGGGCAGCGACGGCAG-3'
Msmeg_6050 Reverse HindIII	5'-GCGCAAGCTTAATGAGGTTGTTTCGCAATGG-3'
Msmeg_5456 forward NdeI (F)	5'-GCGCCATATGATGATCACGACATTTTC-3'
Msmeg_5456 forward NdeI (T)	5'-GCGCCATATGACGACCGGGGACCAG-3'
Msmeg_5456 Reverse HindIII	5'-GCGCAAGCTTGGGTGTGATGACG-3'
Msmeg_0515 forward NdeI (F)	5'-CCCCATATGGTGATACGACGCTGGTTG-3'
Msmeg_0515 forward NdeI (T)	5'-CCCCATATGTCGTCATCGGGTCC-3'
Msmeg_0515 Reverse HindIII	5'-GGGAAGCTTTTTGCCGGCCAGGG-3'
Msmeg_1395 forward NdeI (F)	5'-GCGCCATATGATGTGCCGGCGTGTGTT-3'
Msmeg_1395 forward NdeI (T)	5'-GCGCCATATGTCGCACACCGTCACCG-3'
Msmeg_1395 Reverse HindIII	5'-GCGCAAGCTTCAGTTCGACTGCAG-3'

Table 3.2 The primer sequences used in this thesis for gene amplification. The sequences written from 5' to 3' with restriction enzyme sites highlighted in bold. Two forward primers with an NdeI restriction site were designed for full-length gene (F) and truncated gene (T), respectively, and reverse primer with HindIII restriction site.

3.2.2 Gene cloning

To successfully clone all the PCR products into the pET28a expression vector, restriction digests using the enzymes NdeI and HindIII were carried out, to cut both the PCR product and the pET28a plasmid. The reaction resulted in a sticky end within both gene and plasmid, which, in turn, could be ligated together using T4 DNA ligase (NEB). The mixture then was transformed into DH5 α competent cells and plated on LB-agar containing 500 μ g/ml kanamycin for selection and incubated overnight at 37 °C. Cells from a single colony were transferred to a primary culture of 5ml and incubated overnight at 37°C in a 250 RPM shaker. Plasmid DNA was extracted using a QIAprep® Miniprep kit (Qiagen). Successful cloning was confirmed by a restriction digestion experiment using the same restriction enzymes (NdeI and HindIII) and agarose gel electrophoresis was run for visualization purpose (Figure 3.1). The plasmid DNA was transformed into a BL21 (DE3) expression strain for gene expression.

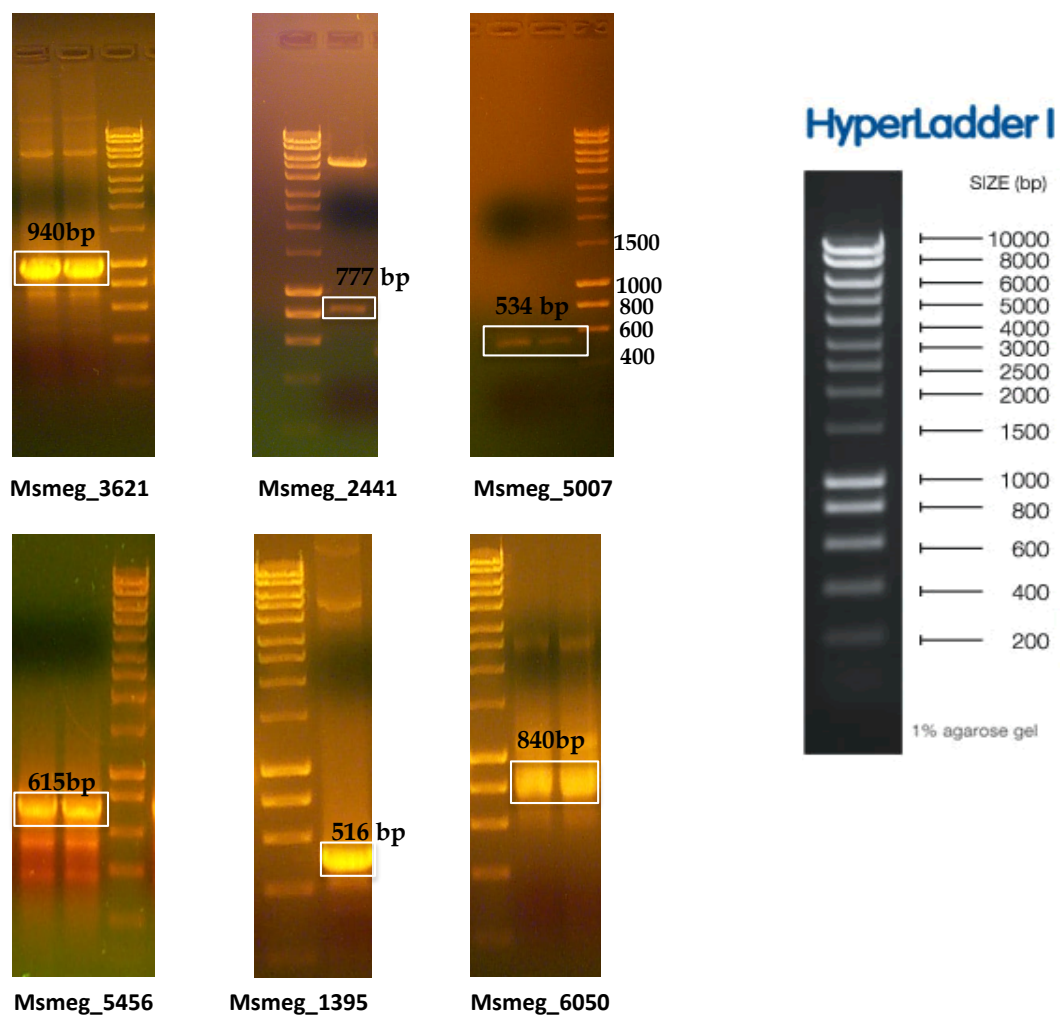


Figure 3.1 Electrophoresis gels showing the PCR products of gene amplification for all selected targets. Bands highlighted in square boxes are the gene targets. Hyper ladder I DNA marker is shown on each gel.

3.2.3 Small and large scale protein expression

Small scale overexpression trials were attempted, by inoculating a single colony of pMsmeg_(gene) into 5ml of LB media containing 50 ugml⁻¹ kanamycin and incubated at 37°C overnight. Each gene was expressed using three 50 ml of LB medium in 250 ml flasks and supplemented with 50 ugml⁻¹ kanamycin, which were then incubated in three different temperatures 37, 25 and 18°C and with variable IPTG concentration 1.0 mM, 0.5 mM and 0.1 mM and for 4 hours induction.

The results revealed that no expression was observed for the full-length protein of any target except for Msmeg_5007, which was expressed in the soluble form only at 18 C° and after 10 hours induction. Thus, small-scale expression trials of the truncated form of the targets were carried out using a similar protocol. However, only three (Msmeg_3621, Msmeg_6050 and Msmeg_5456) could be expressed at low level in the soluble form and thus tested for large-scale expression. The other two targets (Msmeg_1395 and Msmeg_2441) were expressed in the insoluble fraction and attempts to produce soluble proteins by optimizing the expression conditions failed. They were therefore excluded from further studies (data not shown). The expression conditions for each truncated target protein are given in Table 3.3.

Target	Temp (°C)	Induction time (hours)	IPTG concentratio n (mM)	Culture volume (l)	Optical Density OD ₆₀₀	Shaker speed (RPM)
Msmeg_3621 (NDH2)	37	10	0.5	7	1.00	200
Msemg_5007	18	6	0.1	10	0.5	200
Msemg_6050	25	5	1	2	0.5	200
Msemg_5456	25	5	1	2	0.5	200
Msemg_0515	37, 25 & 18	4	1	1	0.6	200

Table 3.3 Over-expression conditions for all truncated target proteins from *M.smegmatis*.

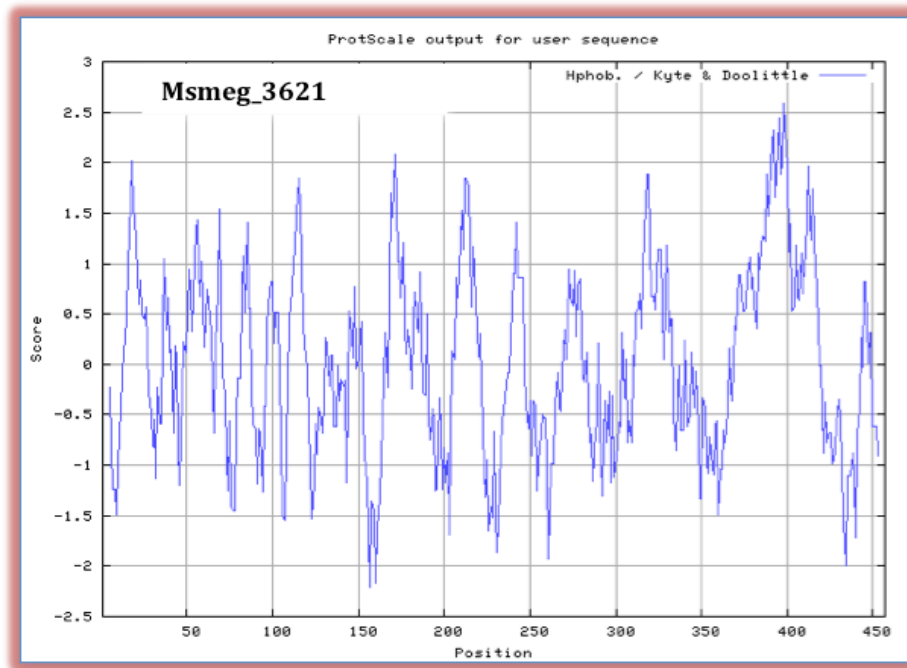
3.3 Studies on the Msmeg 3621 (NDH2) protein

3.3.1 Introduction

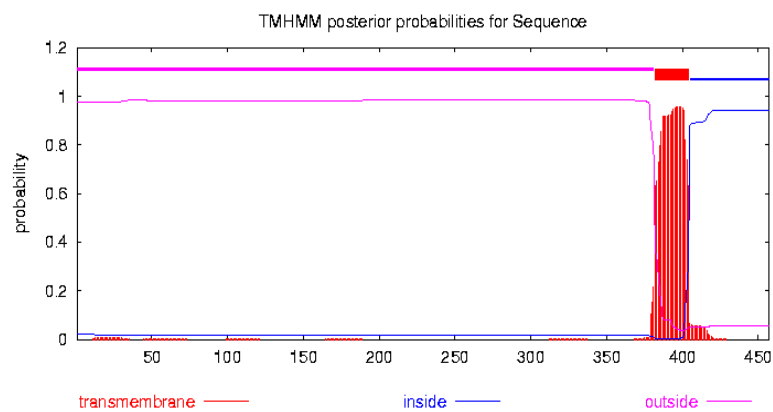
The *M.tb* gene Rv1854c encodes the NDH2 enzyme, which has not been found in mammalian cells, has been shown to be essential in bacteria, including pathogens such as *M.tuberculosis*. NDH2 and is thus a potential drug target [75, 76]. This enzyme is involved in the respiratory chain of mycobacterium and is found in all Mycobacterium species including *M.smegmatis*. Msmeg_3621 where it shares a high protein sequence identity (63%), with NDH2 from *M.tuberculosis*. Although this selected target is not identified as a lipoprotein, it is predicted to attach to the bacterial cell membrane by a transmembrane helix in the C-terminal region (Figure 3.2). The hydropathy plot of NDH2 revealed hydrophobic residues (360-430) at its C-terminal sequence, which may be involved in anchoring the protein to the cell membrane (Figure 3.2a). In addition, these residues are also predicted as a transmembrane helix using the TMHMM server (Figure 3.2b).

3.3.2 Cloning and expression of the full length NDH2

The full-length gene of NDH2 was cloned into a pET28a plasmid with an N-terminal 6xHis-tag as mentioned in section 3.2.2; however, no expression was observed for the full length NDH2. Therefore, it was decided to cut the protein before the C-terminal α -helix in an attempt to produce a soluble protein suitable for crystallization experiment. The truncation site was identified using a combination of the predicted structure from the Phyre server and the hydropathy plot (Figure 3.2). Therefore, NDH2 was truncated at G380 and cloned into pET28a plasmid using the same restriction sites. The truncated NDH2 was overexpressed using the same protocol for the full length as described in section 3.2.3.



a.



b.

Figure 3.2 Hydropathy (a) and transmembrane helix prediction (b) plots of NDH2 from *M. smegmatis*. (a) The hydrophobicity plot was generated by using the Kyte-Doolittle hydrophobicity scale program [74], and indicates a stretch of hydrophobic residues in the C-terminal protein sequence region (380-420). (b) TMHMM plot indicates a potential transmembrane helix at the same place in the sequence [32].

3.3.3 NDH2 Structure and function

NDH2 is a 50-KDa flavoenzyme that plays a significant role in the respiratory chain, where it catalyzes the oxidation of NADH [91]. Significantly, NDH2 appears in the electron transport chain (ETC) of *M.tuberculosis* together with the type 1 NDH (complex1), however, NDH2 is different from complex1, since it is not involved in pumping protons across the membrane and is thus considered as a potential drug target [91]. The electron transport chain of the bacterial cytoplasmic membrane is considered as a factory to generate energy for cellular growth. The respiratory chain usually catalyzes the transportation of electrons, which are produced by oxidizing organic substrates, such as NADH or succinate to an electron acceptor, such as oxygen [92].

The ETC of *M.tuberculosis* is composed of several complexes, which are involved in the power supply. Each of these complexes binds small molecules, such as heme, a flavin prosthetic group or copper atoms [76, 93]. These complexes include four dehydrogenases that work as a complex to transfer electrons from the NADH electron donor in the cytoplasm into the ordinary quinone pool, which is found in the center of the ETC [93]. These dehydrogenases are known as Succinate menaquinone oxidoreductase (SQR), type 1 NDH (NDH-1) that is involved in proton translocation during redox activity and two types of NDH2 enzyme (NDH2 & NDH-2A). NDH2 enzymes both work as a single individual subunit, and are not involved in proton translocation as NDH-1[93]. The oxidation of the quinol is performed either by a reductase bound into the membrane to transfer electrons directly into the cytochrome bd oxidase or by cytochrome bc1 complex that transfers electrons into cytochrome c oxidase (aa3), to be then transferred to oxygen, which is an electron acceptor [93]. In addition, the movement of electrons along the chain contributes to the protein movement across the membrane, which helps to form an electrochemical gradient, which is in turn used by ATP synthase to produce ATP from ADP in the cytoplasm [93].

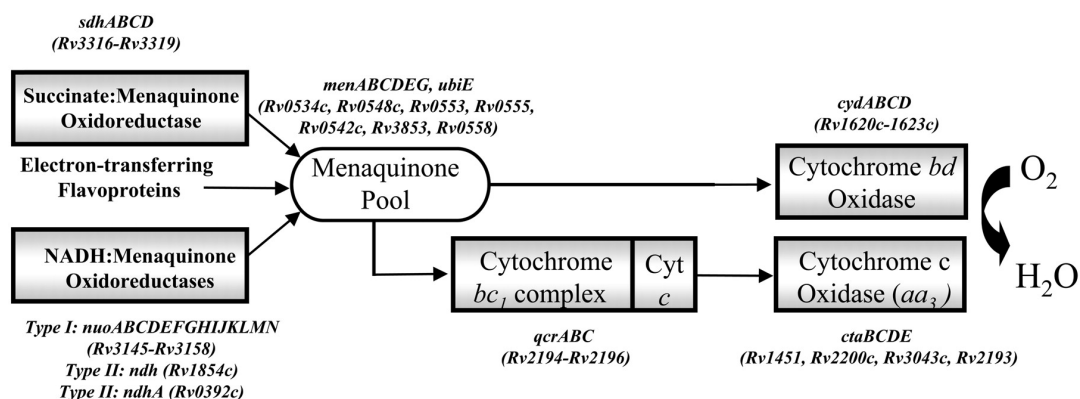


Figure 3.3 The pathway of aerobic electron flow in mycobacterium. Complexes are shown in boxes, with corresponding gene names. The figure was taken from [93].

At present, neither the 3D structure of *M.tuberculosis* nor *M.smegmatis* NDH-2 has been determined and their structure–function relationships have been predicted by the use of sequence similarity to other flavoenzymes, such as the lipoamide dehydrogenases [91, 94, 95]. The protein has been predicted to contain two domains, each with a central β -sheet flanked by helices [96]. Both domains contain the GXGXXG Rossmann fold motif, with one domain binding NADP (H) and the other FMN or FAD [91, 96].

The Phyre server was used to predict a structure for *M.smegmatis*. The best hit (100% confidence) was with the Ndi1 protein that catalyses the NADH oxidation in mitochondria of the *Saccharomyces cerevisiae*. 424 residues (93%) of the *M.smegmatis* NDH2 sequence were covered in the comparison with sequence identity of 28% (Figure 3.4). Structure analysis based sequence alignment and structure prediction to identify the predicted transmembrane α -helix of NDH2 is described below in section 3.3.7.

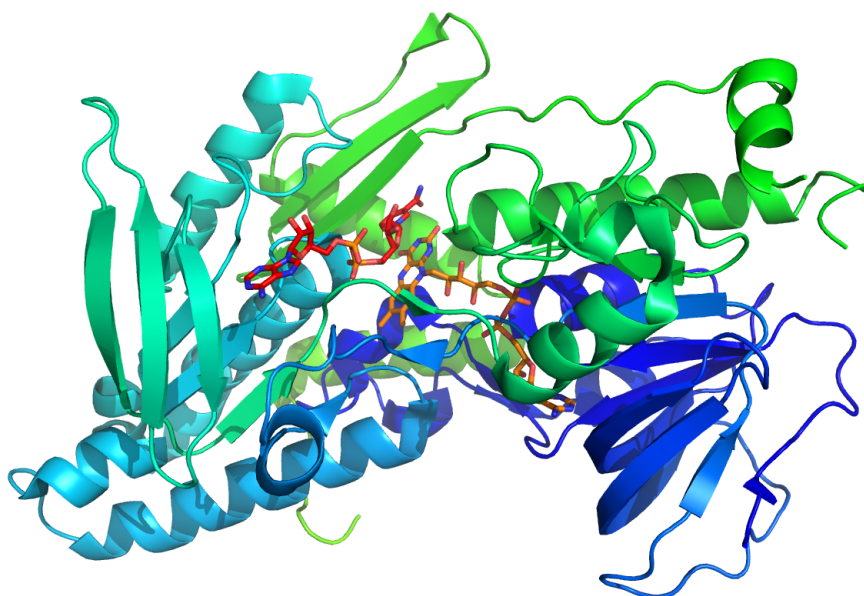


Figure 3.4 The predicted 3D structure fold of Msmeg_3621 (NDH2). The Phyre result revealed that the NDH2 of *M.smegmatis* is predicted to share similar fold with structure of the ndi1 protein from *Saccharomyces cerevisiae* in complex with NAD and FAD (PDB code 4GAP) with 424 residues of NDH2 sequence covered in the comparison (93%) and with low sequence identity 28% and 100% prediction confidence [89]. FAD and NAD are shown as red and orange sticks in the center [97]. Figure was made using Pymol [98].

3.3.4 Protein purification of NDH2 enzyme

3.3.4.1 Ni-NTA chromatography

As the construct of truncated NDH2 contained a N-terminal 6×His tag, a Ni-NTA affinity chromatography column was used as the initial purification step. Thus, 4 g of cell paste was removed from -80°C and resuspended on ice using 10 ml of 50 mM Tris buffer, pH 8.0. Then, cells were sonicated three times on ice at 16-micron amplitude, each for 20 seconds. Cell debris then were removed by centrifugation at 70,000 x g for 10 minutes at 4°C. The soluble protein fractions (~20 ml/ 50 mg) were applied to 5 ml Ni-HP cartridge column after measuring the protein concentration by the Bradford assay. NDH2 enzyme was eluted from the column using 50 ml of 0.35 M imidazole. Fractions of 8 ml were collected. The protein was analysed by SDS-PAGE (Figure 3.5), which showed the presence of lower molecular weight contaminants, thus, a further purification step was used.

3.3.4.2 Gel filtration of pMsmeg_NDH2

Fractions 7 - 14, which had the highest protein concentration (Figure 3.6) were combined for further purification using gel filtration. The sample was reduced to 1 ml using a Viva spin concentrator (MWCO 30000) and injected into a 16x60 Superdex200 column (GE Healthcare), equilibrated with 50 mM Tris, pH.8.0 and 0.5 M NaCl buffer. Gel filtration was performed at a flow rate of 1.5 ml/min and 2ml fractions were collected. The progress of the purification was analysed by SDS-PAGE (Figure 3.5).

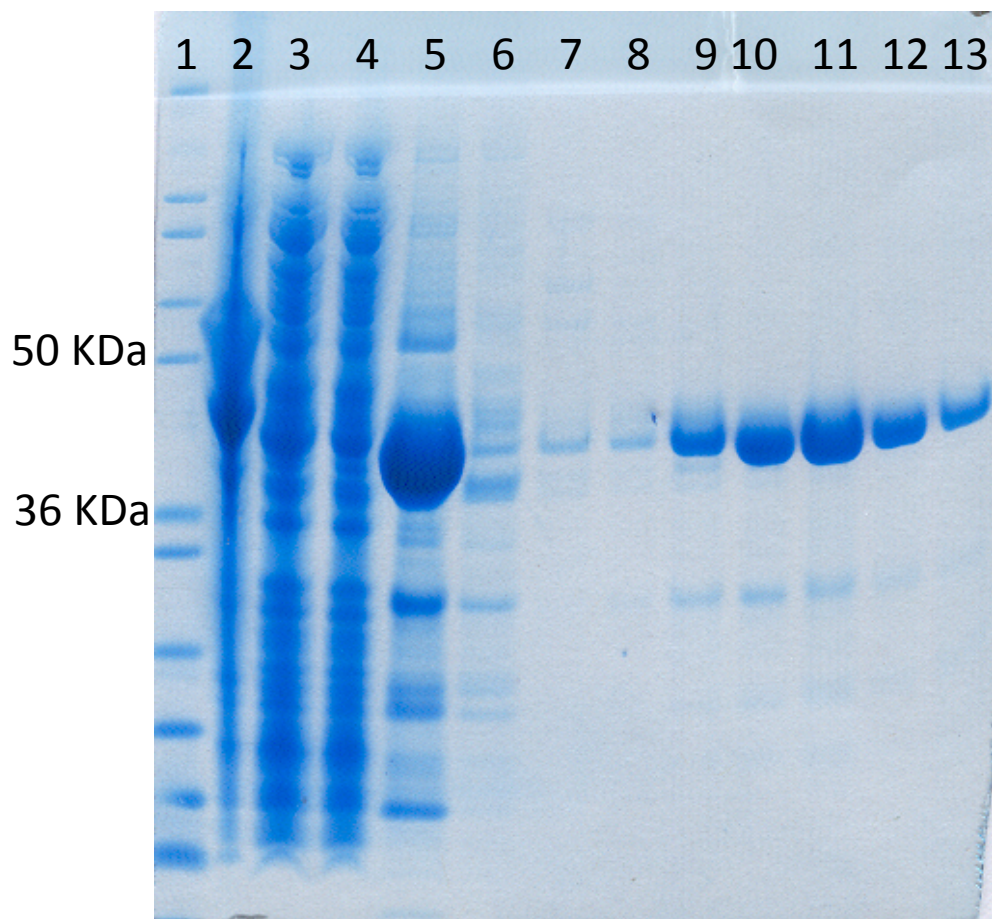
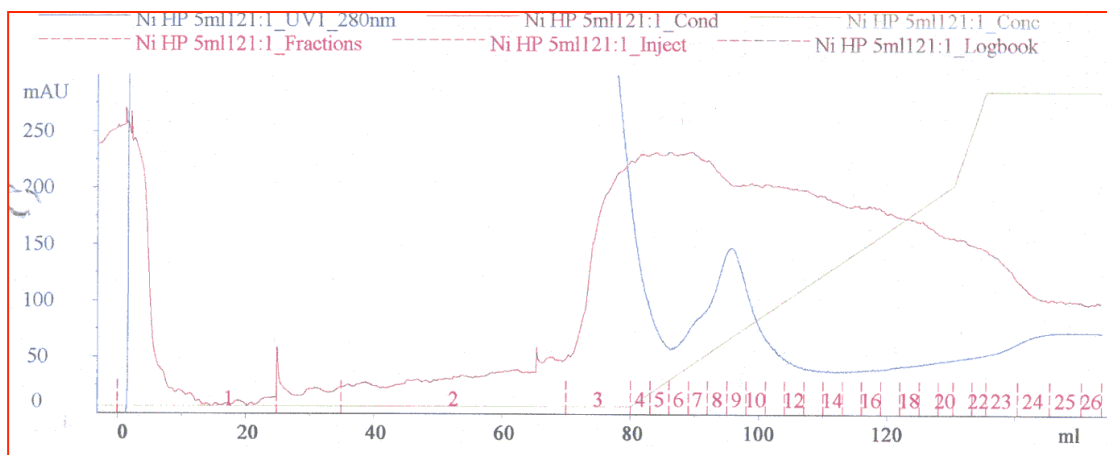


Figure 3.5 SDS gel showing protein purification steps of truncated NDH2. Lane 1; Mark12, lane 2; insoluble fraction of NDH2, lane 3; soluble fraction of NDH2, which indicated a weak expression of soluble NDH2 (~45kDa), lane 4; unbound materials of Ni-NTA affinity column, lane 5; compound fractions 8-11 of Ni-NTA affinity column, lanes 6-13; collected fractions from gel filtration column.

a)



b)

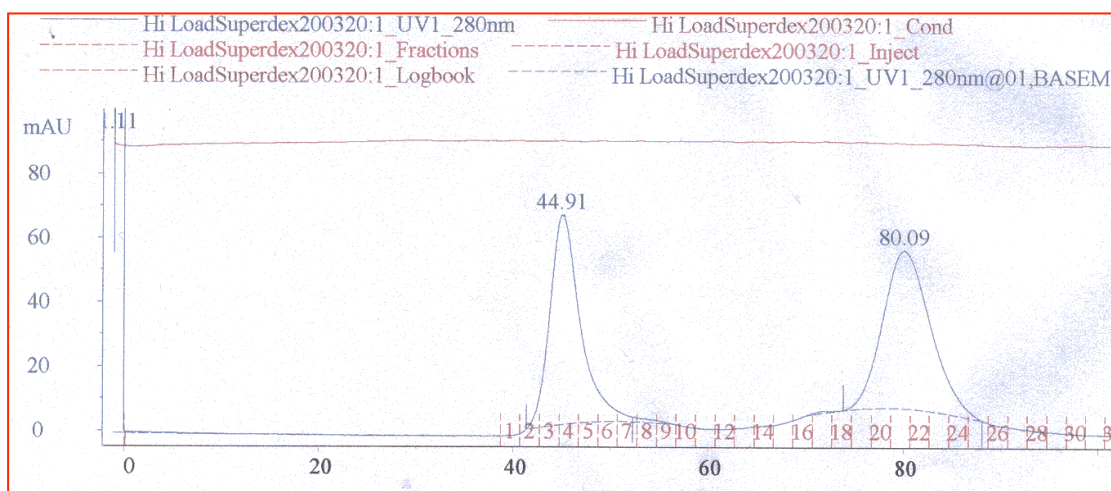


Figure 3.6 Chromatograms obtained during the purification of NDH2. Blue line represents absorption at 280nm, brown line represents conductivity and green line represents gradient of Imidazole concentration: **a.** Chromatography on HisTrap HP column; **b.** Gel filtration on HiLoad Superdex200 column. The gel filtration column step succeeded in removing contaminants left after the initial Ni-NTA purification.

3.3.5 Sample preparation and crystallization of NDH2

To prepare samples for crystallization, the protein from gel filtration was concentrated using a Viva spin column for 10 minutes at 4,500 xg. Subsequently 7ml of 10mM Tris buffer, pH8.0 was added to the 400 μ l of protein solution and spun again for 20-30 minutes at 4500 xg, during this process NDH2 was partially precipitated, but the final sample contained \sim 10 mg/ml NDH2 in 10mM Tris-HCL buffer, pH 8.0.

Crystallization was undertaken using a Hydra II Robot, and the sitting drop vapor diffusion method, with 96 well MRC sitting drop plates. Each reservoir contained 400 μ l of precipitating solution, and the drop was composed of 200 nl of well solution with 200nl of protein solution. Plates were spun at 2000 g prior to incubation at 17°C. Crystallization trials with the JCSG, PACT, PEG and Classic screens (Qiagen Nextal®) were attempted, with both apo NDH2 and NDH2 plus 5 mM FAD⁺ and NAD⁺. The plates were incubated at two temperatures 17°C and 7°C, respectively. However, no crystals were observed in any condition, even after repeating the experiment using the optisalts condition kit (QIAGEN®). The project was thus suspended.

3.3.7 Discussion

During the final stages of writing this thesis, the structure of NDH2 from *Caldalkalibacillus thermarum* was published [99]. This protein shares 23% and 31% sequence identity with NDH2 from *M.smegmatis* and *M.tuberculosis*, respectively. Also, similar enzyme type-II NADH dehydrogenase (Ndi1) is found in the mitochondria of *S.cerevisiae* and has been structurally determined [100]. This protein shares 26%, 28% and 29% sequence identity with NDH2 from *C.thermarum*, *M.tuberculosis* and *M.smegmatis*. A structure based sequence alignment between the four structures revealed a similar secondary structure, but with a number of slight deletions between the yeast Ndi1 and bacterial NDH2 (Figure 3.7). However, the structure of *C.thermarum* NDH2 and *S.cerevisiae* Ndi1 align well with an (r.m.s.d of 1.7Å) [99]. Furthermore, sequence alignment of all these four enzymes revealed that the binding sites of FAD⁺ and quinone are conserved in all of them. Superimposition the two structures of NDH2 and NDi1 also revealed that the FAD⁺ molecules are aligned well (Figure 3.9) [99, 100].

<i>M. tb</i>	1	M-----SPQQEPTAQP RRHRV VI IGSGFGG LNAAK	31
<i>M. smegmatis</i>	1	-----MSHPGATASDRHKV VI IGSGFGG LTAAK	28
<i>C. thermarum</i>	1	-----MSKPS IVIL GAGYGG I VAAL	20
<i>S. cerevisiae</i>	1	MLS KNLYSNKRL L ST NTLVRFASTRSTGVENS GAGPTS FKTMKVIDPQHSD KPNV IL IGSGWGA I SFLK	70
Consensus ss:		eeee hhhhhhhh	
<i>M. tb</i>	32	KLKRA ---D VDIKLI ARTTHHL FQPLL YQVAT GI ISEGEI AP TRV LRKQ-R NVQVLLGNV THIDL AG	96
<i>M. smegmatis</i>	29	TLKRA---D VDV KLIARTTHHL FQPLL YQVAT GI ISEGEI AP TRV LRKQ-K NAQVLLGDV THIDL EN	93
<i>C. thermarum</i>	21	GLQKR L NYNE ADITL V NKNDY H YITTEL HQPA AG TMH DQAR V G IKEL ID E K -- KIKFV KD T VVA ID RE Q	88
<i>S. cerevisiae</i>	71	HID T K --- KYN VS II SPRS Y L F T P L L PS AP V G T VD E K S II E P IV N F AL KK KG N V TY E AE AT S I N P D R	136
Consensus ss:		hhhhh eeeee hhhhh hhhhhh hhhhhhhh eeeeeeeee	
<i>M. tb</i>	97	QCV SELL LG H ----- TY Q TP Y D S L I A A G Q S Y F G N D H F A E F A P G M K S I D D A L E L R G	149
<i>M. smegmatis</i>	94	K T V D S V L L G H ----- T Y S T P Y D S L I A A G Q S Y F G N D H F A E F A P G M K S I D D A L E L R G	146
<i>C. thermarum</i>	89	Q K V T L Q N ----- G E L H Y D Y L V V G L G S E P E T F G I E G L R E H A F S I N S I N S V R I R Q	137
<i>S. cerevisiae</i>	137	N T V T I K S L S A V S Q L Y Q P E N H L G L H Q A E P A E K Y D Y L I S A V G A E P N T F G I P G V T D Y H F L K E I P N S L E I R R	206
Consensus ss:		eeeeee eeee eeee eee hhhhhhhh	
<i>M. tb</i>	150	R I L S A F E Q A E R S S D - P E R R A K L L T F T V V G A G P T G V E M A G Q I A E L A E H T L K G A F R H I D S T K A R V I L L D A A P	218
<i>M. smegmatis</i>	147	R I L G A F E Q A E R S S D - P V R R A K L L T F T V V G A G P T G V E M A G Q I A E L A D Q T L R G S F R H I D P T E A R V I L L D A A P	215
<i>C. thermarum</i>	138	H I E Y Q F A K F A A E P --- E R T D Y L T I V V G A G F T G I E F V G E L A D R M P E L C A E Y - D V D P K L V R I N V E A A P	201
<i>S. cerevisiae</i>	207	T F A A N L E K A N L L P K G D P E R R R L S I V V V G G G P T G V E A A G E L Q D Y V H Q D L R K F L - P A L A E E V Q I H L V E A L P	275
Consensus ss:		hhhhhhhhh h eeeee hhhhhhhhhhhhhhhhhhh eeeee	
<i>M. tb</i>	219	A V L P P M G A K L Q R A A A R L Q K L G V E I Q L G A M V T D V D R N G I T V K D S D --- G T V R R I E S A C K V S A G V S A S R	284
<i>M. smegmatis</i>	216	A V L P P M G E K L G K K A R A R L E K M G V E V Q L G A M V T D V D R N G I T V K D S D --- G T I R R I E S A C K V S A G V S A S P	281
<i>C. thermarum</i>	202	T V L P G F D P A L V N Y A M D V L G G K G V E F K I G T P K R C T P E G V V I E V D G --- E E E I K A A T V V T G G V R G N S	266
<i>S. cerevisiae</i>	276	I V L N M F E K K L S S Y A Q S H L E N T S I K V H L R T A V A K V E E K L L A K T K H E D G K I T E T I P Y G T L I W A T G N K A R P	345
Consensus ss:		hhhhhhhhhhh eeee eeeee eeeee eeee eeee h	
<i>M. tb</i>	285	L G R D L A E Q S R V E L D R A G R V Q V L P D L S I P G Y P N F V V G D M A A V E --- G V P G V A Q G A I Q G A K Y V A S T I K	348
<i>M. smegmatis</i>	282	L G K D L A E Q S G V E L D R A G R V K V Q P D L T L P G H P N F V V G D M A A V E --- G V P G V A Q G A I Q G G R Y A A K I I K	345
<i>C. thermarum</i>	267	I V E K S G F E T --- M R G R I K V D P Y L R A P G H E N I F I V G D C A L I N E E N R P Y P T A Q I A I Q H G E N V A A N L A	331
<i>S. cerevisiae</i>	346	V I T D L F K I P E Q N S K R G L A V N D F L Q V K S N N I F A I G D N A F A G --- L P P T A Q V A H Q E A E Y L A K N F D	408
Consensus ss:		hhhh hhh eee eeeeeee hhhhhhhhhhhhhhhhh	
<i>M. tb</i>	349	A E L A G A N P ----- A E R E P F Q Y F D K G S M A T V S R F S A V A K I G P --- V E F S G F I A W L I W	396
<i>M. smegmatis</i>	346	R E V S G T S P ----- K I R T P F E Y F D K G S M A T V S R F S A V A K V G P --- V E F A G F F A W L C W	393
<i>C. thermarum</i>	332	A L I R G G ----- S M T P F K P H I R G T V A S L G R N D A I G I V G G --- R K V Y G H A A S W L K	376
<i>S. cerevisiae</i>	409	K M A Q I P N F Q K N L S R K D K I D L L F E E N N F K P F K Y N D L G A L A L G S E R A I A I R S G K R T F Y T G G L M T F Y L W	478
Consensus ss:		hhh eee eeeee eeeee eee hhhhhh	
<i>M. tb</i>	397	L V L H L A L I G F K T I T L L S W T V T F L S T R R G L T I T D Q Q A F A R T R L E Q L A E L A E A Q S A A S A K V A S	463
<i>M. smegmatis</i>	394	L V L H L V Y L V G F K T I V T L L S W G V T F L S T K R G Q L T I T E Q Q A Y A R T R I E E L E E I A A A V Q D T E K A A S ---	457
<i>C. thermarum</i>	377	K L I D M R Y L I G G L S L V L K K G R F -----	399
<i>S. cerevisiae</i>	479	R I L Y L S M I L S A R S R L K V F F D W I K L A F F K R D F F K G L -----	513
Consensus ss:		hhhhhhhhhhhhhhhhhhhhhhhh	

Figure 3.7 Protein sequence alignment based on secondary structures of bacterial NDH2 protein from *C.thermarum*, *M.tuberculosis* and *M.smegmatis* and the Ndi1 protein from *S.cerevisiae*. The α -helices residues are highlighted as red colour and β -sheets residues are highlighted as blue colour. The conserved binding motifs of FAD, NAD and quinone are black boxed. The residues involved in binding FAD and located in the C-terminal membrane attachment α -helix are black boxed. The c-terminal membrane attached structure is highlighted as cyan.

Furthermore, structural analysis of NDH2 enzyme from *C.thermarum* and Ndi1enzyme from yeast indicated that both enzymes are attached to the membrane by amphipathic helices that are rich with hydrophobic residues [99, 100], and by comparing their secondary structures with NDH2 from *M.tuberculosis* and *M.smegmatis*, a longer C-terminal domain is observed in Mycobacterial NDH2. Also, it was shown that the α -helix, which is involved in membrane anchoring of NDi1, is longer than its equivalent from NDH2 of *C.thermarum*. These membrane-anchoring domains were also shown to be significant for the stability and activity of both enzymes [99, 100]. Therefore, Truncated NDH2 from *C.thermarum* refused to bind FAD as one of the coordinate moiety (K376) was lost [99, 100]. This residue is replaced by W478, which interacts with the FAD by an adjacent Y476. Also, the equivalent C-terminal helices of the NDH2 and NDi1structures are involved in binding quinone molecules [99, 100]. In Mycobacterial NDH2, the C-terminal domain is longer than both enzymes and contains one more helix. *M.smegmatis* NDH2 was truncated at residue V379 based on the hydropathy plot and Phyre prediction result to obtain a soluble enzyme. However, the truncation included all the equivalent α -helix that is significant for membrane anchor and FAD and quinone binding, and thus affected the stability of the enzyme and may prevent the crystallization of NDH2.

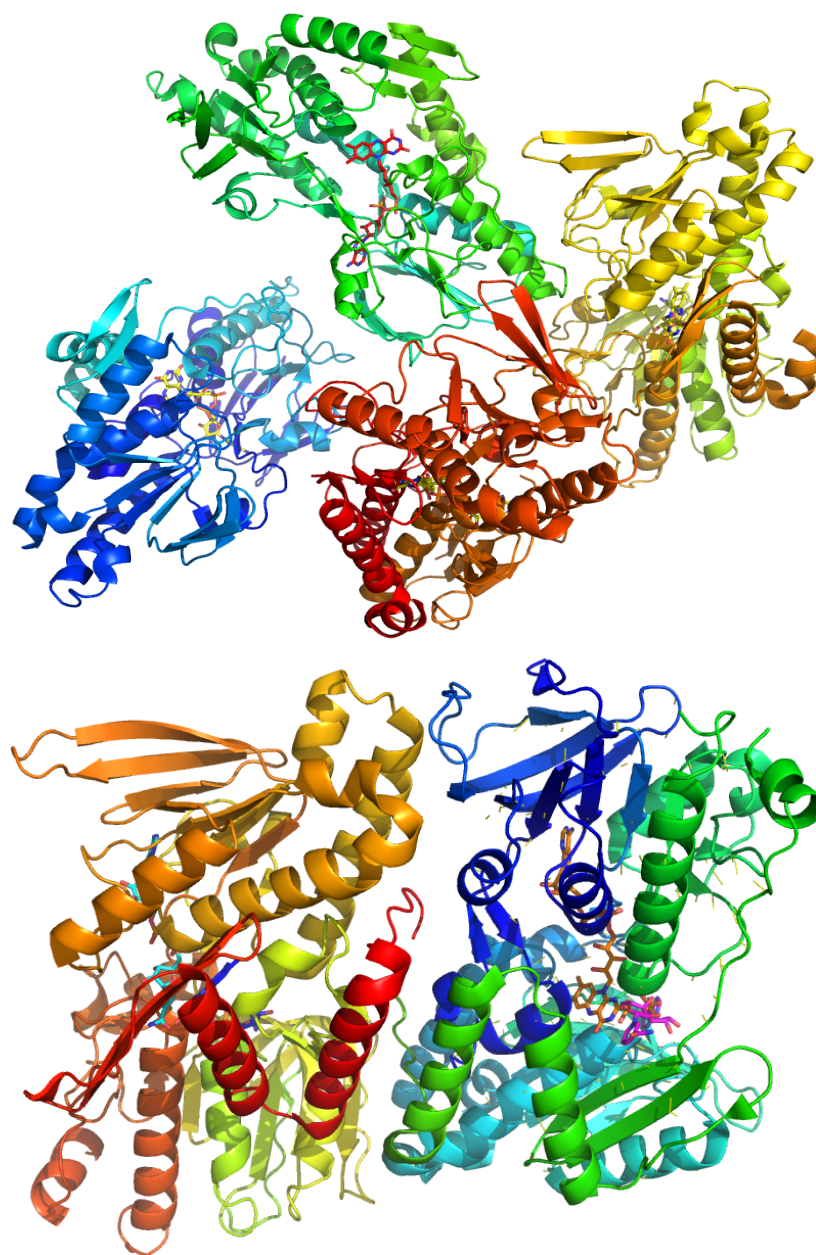


Figure 3.8 The 3D structures of NDH2 enzyme from *C.thermarum* (top,) and Ndi1 enzyme from *S.cerevisiae* (bottom,). The NDH2 is in complex with FAD and Ndi1 is in complex with FAD and NAD. Both structures have been solved recently under (PDB code NDH2; 4NWZ and Ndi1; 4GAP [99, 100]. Figures were produced using Pymol [98].

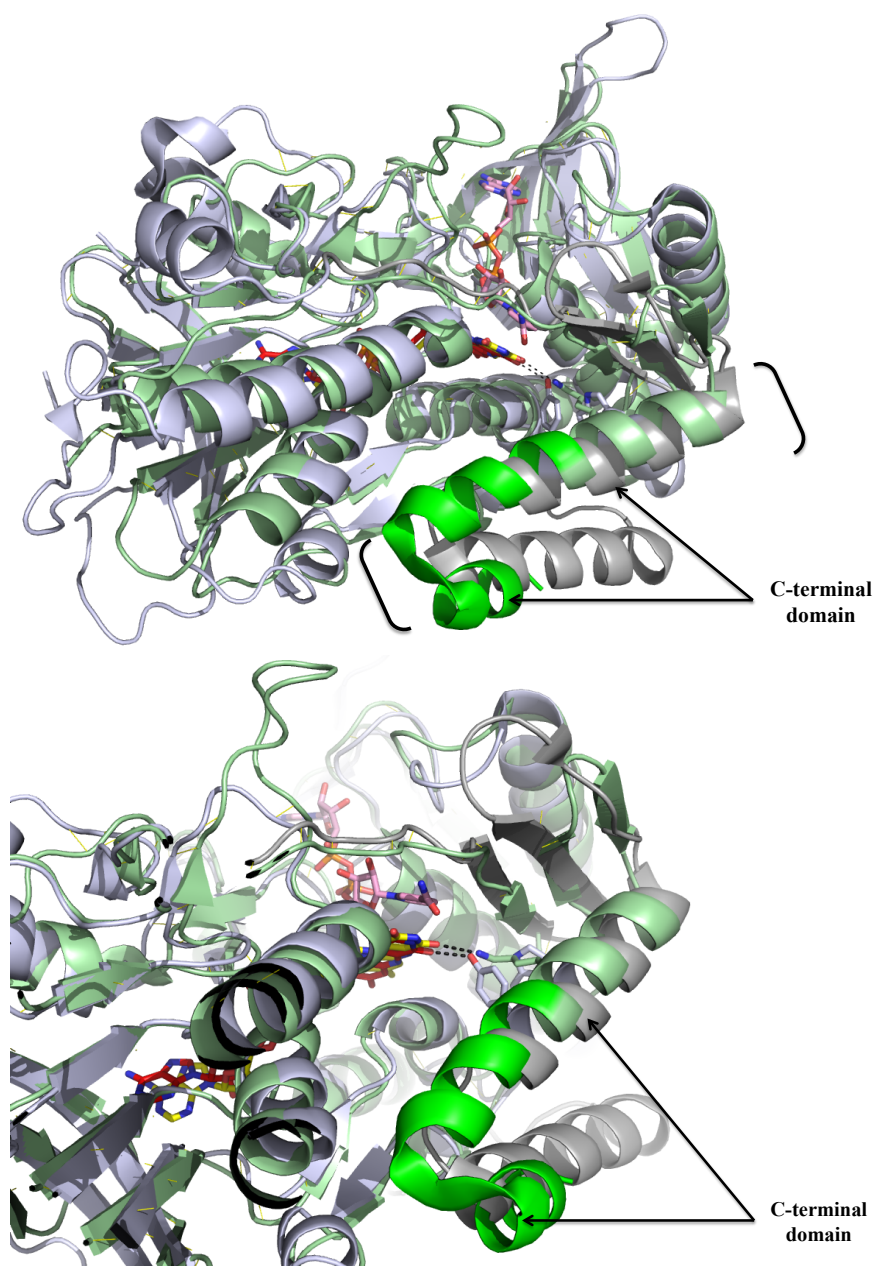


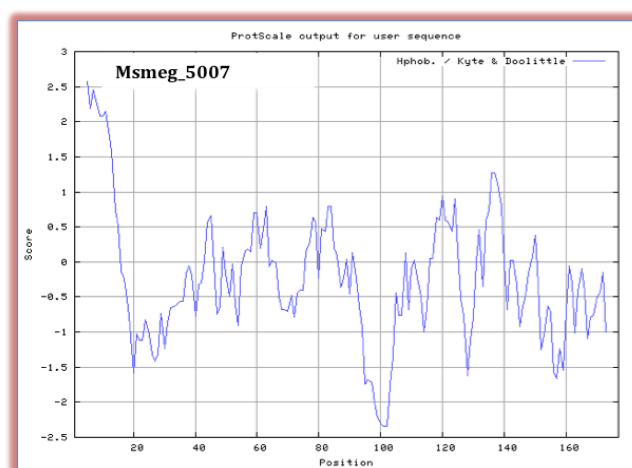
Figure 3.9 Structural superposition of NDH2 enzyme in complex with FAD⁺ and Ndi1 in complex with FAD⁺ and NAD⁺. The two structures overlapped into RMSD of 1.73 Å [99, 100]. The C-terminal membrane attached α -helices are highlighted in NDH2 (green) and Ndi1 (grey). The α -helix residues that are involved in FAD binding are shown as sticks.

3.4 Studies on the Msmeg_5007 protein

3.4.1 Introduction

Msmeg_5007 (LprB) is a ≈ 19 KDa, putative uncharacterized lipoprotein, with 178 amino acids and 534 bp DNA size. It is only found in Mycobacterium species and no homolog has been detected in other organisms. The *lprB* gene has been shown to be essential for mycobacterium survival [75]. The hydropathy plot of LprB revealed 20 hydrophobic residues at its N-terminal, which may be involved in anchoring the protein to the cell membrane (Figure 3.10.a). In addition, no transmembrane helices are predicted using the TMHMM server, which might suggest that these hydrophobic residues refer to a short signal sequence (Figure 3.11.b). Therefore, the full-length protein sequence and a truncated protein at Cys 15 were attempted to be expressed.

a)



b)

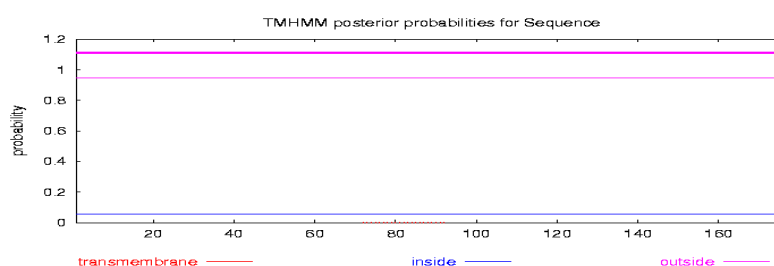


Figure 3.10 Hydropathy (a) and transmembrane helix prediction (b) plots of LprB from *M.smegmatis*. a) The hydrophobicity plot was generated by using the Kyte-Doolittle hydropathy scale program [74], and indicates a stretch of hydrophobic residues in the N-terminal protein sequence region (1-20 residue). b) TMHMM plot indicates no transmembrane helix at the same place in the sequence [32].

RNA-binding S4 domain-containing protein [Gordonia bronchialis DSM 43247]

Sequence ID: [ref|YP_003274952.1](#) Length: 293 Number of Matches: 1

[▶ See 2 more title\(s\)](#)

Range 1: 121 to 289 [GenPept](#) [Graphics](#) ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
143 bits(361)	2e-38	Compositional matrix adjust.	83/181(46%)	108/181(59%)	20/181(11%)
Query 1	MAVLAAAMVPVFAACSTDE----	PASPEVPOSES	SAGAPTHG	PPFPQCGGISDQTVSEL	56
Sbjct 121	LIALIAALMLIVAGCSSDDTPSDPAKPPTPQA----	PGRAGD	PPFGECGGLTTEEVSSI	176	
Query 57	TQVPLVNTATNSSGCQWLQGGSI	LGP	HFSFTWFRGSP	IGRERKTEELSRASVEDINIEG	116
Sbjct 177	TRLGVLNTNTIKNPSVCEWDSTGTRTGPVASF	NWYRGS	PIGRERATEQLSRAS	TTDIEIKG	236
Query 117	HGGFIAVGEDPLKPGDVT-LCEIGIQFDDDF	IEWSVSY-SQK	PPF--DPCEVAKELTRQS	172	
Sbjct 237	HRGFIA-----HDATAICEV	GIEFGADFF	EWSVSAGS	NSPLTIEQVCDATRELSRLS	288
Query 173	I			173	
Sbjct 289	I			289	

putative RNA-binding protein [Gordonia soli]

Sequence ID: [ref|WP_007619464.1](#) Length: 296 Number of Matches: 1

[▶ See 1 more title\(s\)](#)

Range 1: 158 to 292 [GenPept](#) [Graphics](#) ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
131 bits(329)	1e-33	Compositional matrix adjust.	68/143(48%)	83/143(58%)	16/143(11%)
Query 39	GPFFPQCGGISDQTVSEL	TQVPLVNTATNSSGCQWLQGGSI	LGP	HFSFTWFRGSP	IGRE 98
Sbjct 158	GPFFGECGGVTIDDVARLTKFPLSATVNN	NPSACEWSSNDRTGPVASF	NWYRGS	PIGRE 217	
Query 99	RKTEELSRASVEDINIEGHGGFIAVGEDPLKPGDVT	LCEIGIQFDDDFIEWSVSY	SQK--R	TE+LSR S +DI I GH GFIA	156
Sbjct 218	RATEQLSRESTKDIEINGHKGFIA-----	SDVGICEV	GIDFGDFF	EWSVSAGAAASV 269	
Query 157	-----PFPDPCEVAKELTRQSI			173	
Sbjct 270	TGGEVPPTEEICDATRELSRLSI			292	

Figure 3.12 A Blast search against the non-redundant protein sequence. The highest similarity is with putative RNA binding domains from *Gordonia species*.

Additional Genetic analysis revealed that the *lprB* gene is located within an operon that contains pesticide degrading monooxygenase (Msmeg_5001), O-methyltransferase (Msmeg_5003), DNA repair Exonuclease (Msmeg_5004), and phosphohistidine phosphatase (Msmeg_5006). Furthermore, the structure of LprB was predicted using the Phyre server. This suggested that LprB shares a similar structure of beta hairpin fold with RNA binding domain (mog1p) and a protein involved in oxygen enhancement (PsbP), with 33.5 % prediction confidence and 14% sequence identity.

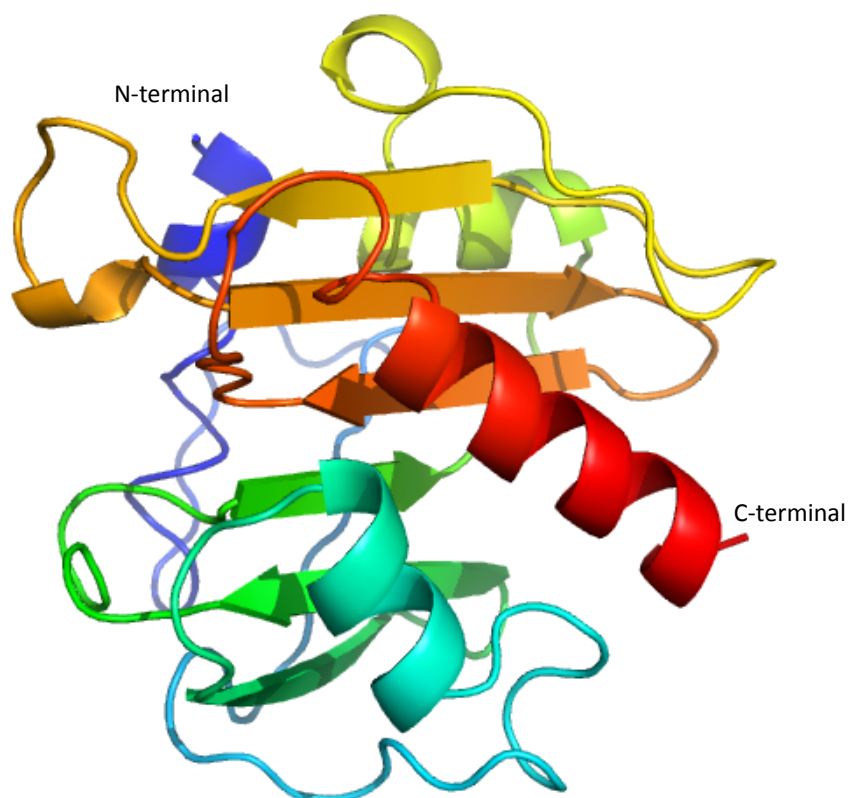


Figure 3.13 The predicted 3D structure of LprB. Phyre results revealed that LprB is predicted to share similar hairpin fold with RNA binding protein from corynebacterium diphtheriae (PDB code 2I8G) with 133 residues of LprB sequence were covered in the comparison (86.0 %) and low sequence identity of 14 % and 33.84 % prediction confidence.

3.4.3 Protein Purification of Msmeg_5007 (LprB)

3.4.3.1 Ni-NTA chromatography

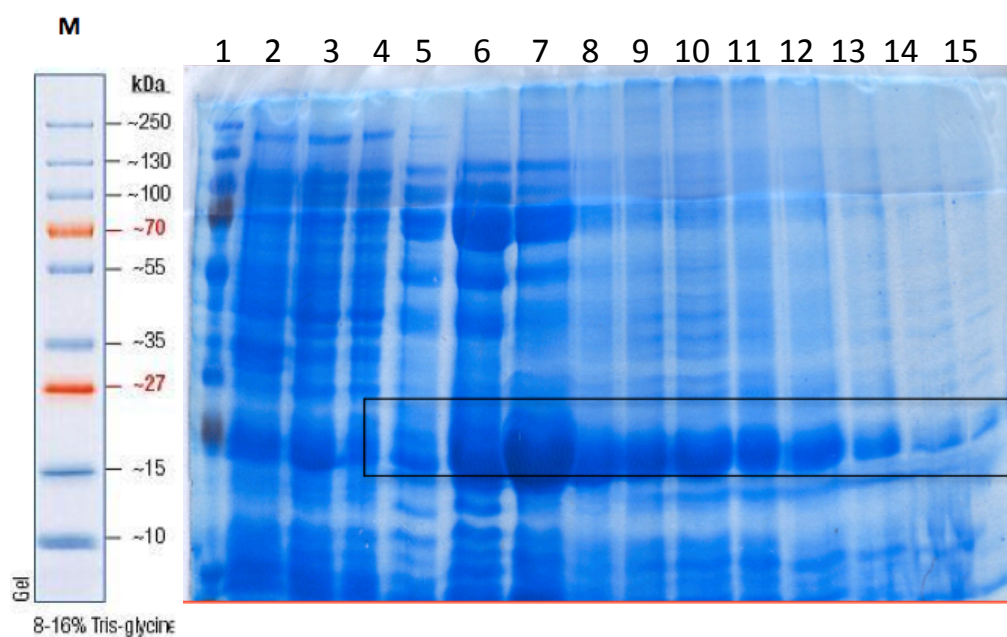
As the expression level of soluble LprB (full length) was very low, the amount of cells was increased to produce more protein suitable for a crystallization experiment. Therefore, 10L of LB medium was used and cells were induced by 0.1mM IPTG at 18 C° and for 10 hours to produce 3g of cell paste. Cells were then transferred into 1 l Beckman centrifuge tubes in order to spin down all cells at 3,600 ×g for 20 minutes. Then, all LB media were discarded, except a little to be used for the resuspension of the cells. Cells were centrifuged again using table top centrifuge (Sigma 3-16K) at 4,500 ×g for 20 minutes and were stored at -80°C.

Cells of LprB were defrosted and resuspended in a buffer containing 50mM Tris-HCl, pH 8.0, and 0.5M NaCl. Then, cells were disrupted on ice by sonication (3x-20 seconds) at a volume of 16-micron amplitude. Cell debris was removed at 4°C by centrifugation at 70,000 xg for 10 minutes. The cell free extract were measured with a Bradford assay with a total protein concentration of 8.2 mg/ml. Then, the supernatant of LprB was applied onto a 5ml Ni-HP cartridge affinity column, which was equilibrated with the same buffer. LprB was eluted from the column by 50ml gradient Imidazole (0 - 0.5M) Tris-HCL buffer, pH 8.0, 3ml fractions were collected (7-11) and were pooled together, with a total protein concentration estimated at 12 mg/ml by Bradford assay.

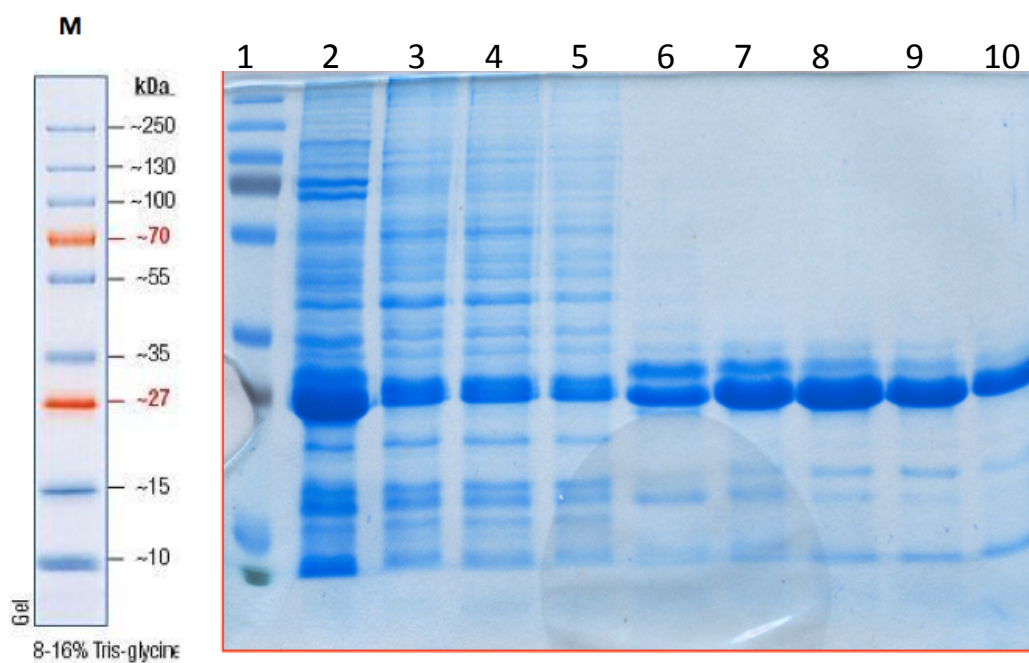
3.4.3.2 Gel filtration purification

The Volume of the LprB was reduced to 2ml using a Vivaspin concentrator with MWCO 30000 size and applied on a 16x60 Superdex200 gel filtration column (GE Healthcare), equilibrated with 50 mM Tris-HCl, pH 8.0, and 0.5M NaCl buffer. Gel filtration was performed at flow rate 1.5ml/min with the same buffer and 2ml fractions were collected. The fractions 22-24 with the highest peaks of protein concentration were combined and further concentrated using VivaSpin concentrator.

To prepare sample for crystallisation, the buffer was exchanged in the sample with 10mM tris-HCl pH8.0 using diafiltration cup with Viva Spin concentrator. Finally, the LprB protein was concentrated to 12mg/ml. The purification progress was analysed by 12% SDS-gel (Figure 3.14).



(a)



(b)

Figure 3.14 SDS gel showing protein purification steps of LprB (full length). The molecular weight of LprB is ~19 (kDa). **a.** An SDS gel showing the steps of Ni-NTA affinity column purification, which indicates LprB expressed in both fractions soluble and insoluble form (lanes 2 and 3, respectively), Lane 4; unbound materials, lanes 5-15 collected fractions. **b.** An SDS gel showing the gel filtration purification. Lane 1; Protein marker SM#811, lane 2; fractions of Ni-NTA column, lanes 3-10; collected fractions.

3.4.4 LprB protein crystallization

LprB crystallizations were undertaken using the same protocol as described for NDH2 (Section 3.3.6). Several hits with flexible plate like morphology and micro crystals were identified with different conditions within the PEG screen A2 (0.1 M Sodium acetate pH 4.6, 30% PEG300), A4 (0.1 M Sodium Acetate pH4.6, 25%PEG1000), and F5 (0.2 M Magnesium chloride, 0.1M Tris pH 8.5, 50% ethylene glycol) and the PACT screen A7 (0.2M Sodium chloride, 0.1M Sodium Acetate pH 5.0, 20% PEG 6000) and the JCSG screen A9 (0.2 M Ammonium chloride + 20% PEG3350) (Figure 3.15). Hanging drop methods were used to optimize several crystal conditions by varying the pH between 4 and 7 and the PEG concentration from 10 to 40%, however, the same crystal morphology was obtained. These crystals failed to diffract in initial X-ray diffraction experiments.

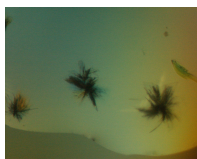
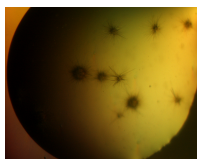
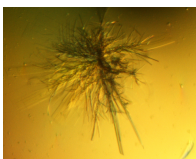
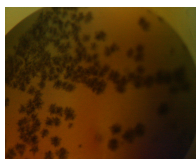
JCSG-A9 (0.2M Ammonium chloride+ 20% PEG3350).	PACT-A7 (0.2M Na chloride, 0.1M Na acetate pH 5, 20% PEG 6000).	PEG- F5 (0.2M Magnesium chloride, 0.1M Tris pH8.5, 50% ethylene glycol)	PEG-A2 (0.1 M Sodium acetate pH 4.6, 30 %(v/v) PEG 300).
			

Figure 3.15 Photographs of LprB native crystals. Crystals were grown over a three months period in different conditions and showed different morphology; however, no diffraction could be recorded from these crystals. Crystals were grown within three months.

3.4.5 Identification of more crystal hits

More crystals were observed after one year and two years incubation time at 17°C. Crystals were mounted in a loop within their mother solution and 25% ethylene glycol. Then, they were flash cooled and stored in liquid nitrogen to be sent to Diamond for X-ray diffraction testing (Figure 3.16). However, the X-ray diffraction tests revealed that all crystals were actually salt crystals and thus, this project also has been suspended.

<p>PACT-H1 (0.2M Na fluoride, 0.1M Bis Tris propane pH 8.5, 20% PEG 3350).</p> 	<p>PEG-F8 (0.2 M Magnesium formate, 20 %(w/v) PEG 3350).</p> 	<p>PEG-A4 (0.1 M Sodium acetate pH 4.6, 25 %(v/v) PEG 550 MME).</p> 	<p>PEG-A2 (0.1 M Sodium acetate pH 4.6, 30 %(v/v) PEG 300).</p> 	<p>Classic-E11 (0.1 M HEPES sodium salt pH 7.5, 1.5 M Lithium sulfate).</p> 
<p>MPD-F6 (0.1 M BICINE pH 9.0, 40 %(v/v) MPD).</p> 	<p>MPD-E9 (0.1 M MES pH 6.0, 20 %(v/v) MPD).</p> 	<p>MPD -E6 (0.1 M BICINE pH 9.0, 10 %(v/v) MPD).</p> 	<p>MPD-F6 (0.1 M BICINE pH 9.0, 40 %(v/v) MPD).</p> 	<p>MPD-G4 (0.1 M Imidazole. HCl pH 8.0, 15 %(w/v) MPD, 5 %(w/v) PEG 4000).</p> 
<p>PEG-A2 (0.1 M Sodium acetate pH 4.6, 30 %(v/v) PEG 300).</p> 	<p>PEG-A2 (0.1 M Sodium acetate pH 4.6, 30 %(v/v) PEG 300).</p> 	<p>JCSG-E8 (1 M di-Ammonium phosphate, 0.1 M Na acetate pH 4.5).</p> 		

Figure 3.16 Photographs of LprB native crystals. Crystals were grown within different conditions and with different crystal forms.

3.5 Studies on the MsmeG_6050 protein

3.5.1 introduction

MsmeG_6050 is \approx 32.0 (kDa), 300 amino acids and 903bp DNA size putative solute binding lipoprotein that belongs to the ABC transport system, MsmeG_6050 has been annotated as non-essential gene of mycobacterium [75]. MsmeG_6050 was assigned as a potential target due to the presence of highly hydrophobic N-terminal 20 residues, (Figure 3.14). The analysis of these hydrophobic residues using the TMHMM server revealed no potential transmembrane helix and thus, these residues are highly likely to form an N-terminal signal sequence [32, 101]. Analysis using the signal peptides prediction SignalP server, showed that the first 20 residues were probably a signal sequence, and that the signal would be cleaved after residue A20 [102]. However, analysis of the sequence of MsmeG_6050 using the DOLOP server, which predicts lipid-anchored proteins, suggested that Cys 18 would be the site of lipid attachment and that the signal would be cleaved between Ala17 and Cys18 [1, 38]. In addition, DOLOP clearly showed the presence of the n-region, h-region and C-region of the lipoprotein sequence (Figure 3.17).

In the light of these analyses, it was decided to produce a construct that started at Ala20 to the C-terminus in an attempt to obtain a soluble protein for crystallization. However, this was not the case for this construct as a truncated protein was produced with the protein sequence starts from the amino acid Thr 33.

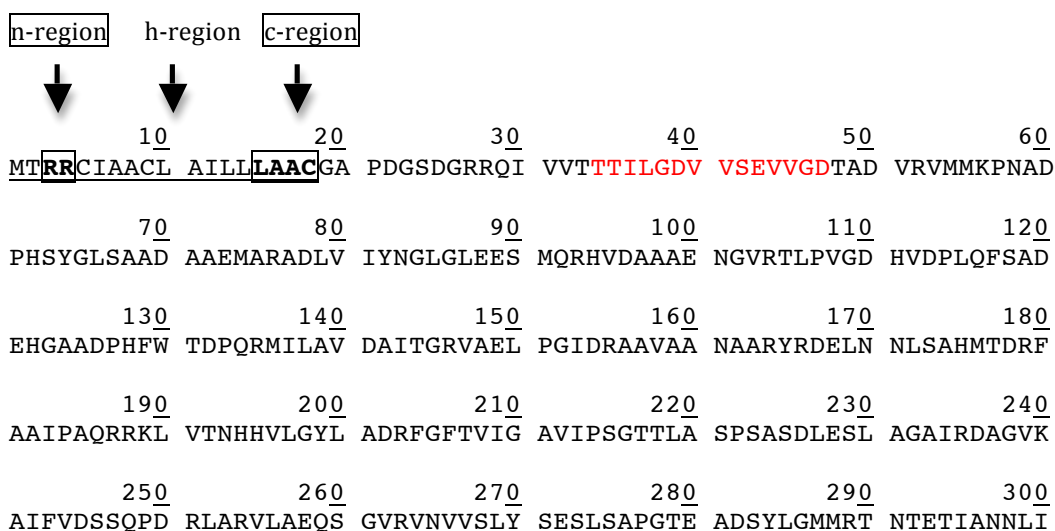
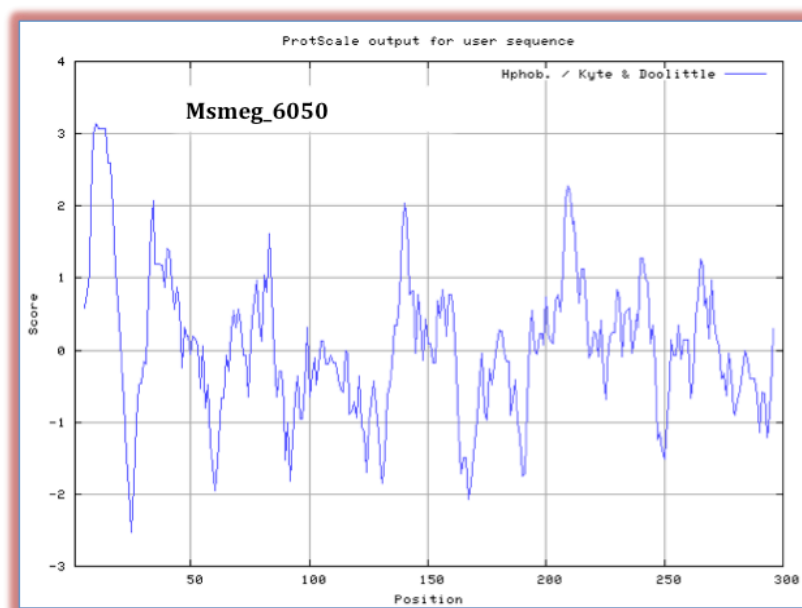
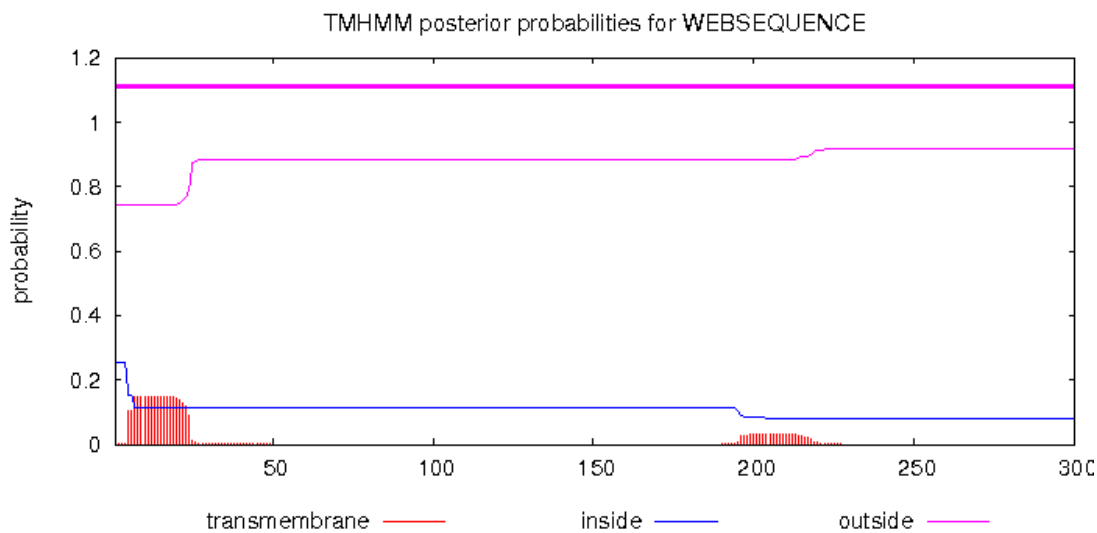
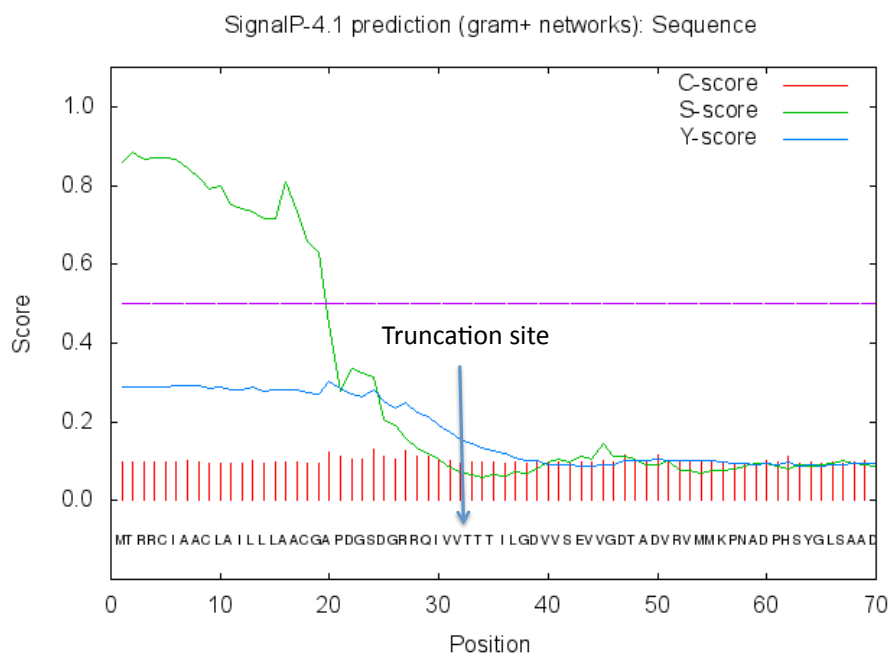


Figure 3.17 Hydropathy plot of Msmeg_6050 and protein sequence. The hydrophobicity plot was generated using Kyte-Doolittle hydropathy scale program [74], and indicated a high value of hydrophobic residues within the first 20 residues. Protein sequence shows the signal peptides (underlined) within the first 20 residues. The lipoprotein features are highlighted and boxed based on DOLOP server. Red colour residues indicated the start of truncated construct and predicted cleavage site.



(a)



(b)

Figure 3.18 Transmembrane (a) and signal peptides (b) prediction plots of *Msmeg_6050*. A. The transmembrane plot was generated using the TMHMM server [32], which indicated no significant sign for the presence of transmembrane helix within the first 20 residues of the N-terminal protein sequence. b) The prediction of signal peptides and cleavage site of truncated protein using SignalP server [102].

In addition, Msmeg_6050 shares considerable sequence identity with the surface antigen (PSAA) from *S.pneumoniae* (PDB code 2ZK7) and the DNA repair protein from *Bacillus caldotenax* (UvrB) (PDB code 1D9X) with 27% sequence identity (Figure 3.20).

Chain A, Crystal Structure Of Pneumococcal Surface Antigen Psaa E205q In The Metal-free, Open State
Sequence ID: [pdb|3ZK8|A](#) Length: 278 Number of Matches: 1
[▶ See 1 more title\(s\)](#)

Range 1: 1 to 276		GenPept	Graphics	▼ Next Match	▲ Previous Match
Score	Expect	Method	Identities	Positives	Gaps
130 bits(326)	6e-35	Compositional matrix adjust.	74/279(27%)	132/279(47%)	9/279(3%)
Query 27	RRQIVVTTTILGDVVSEVVGDTADVRRVMMKPNADPHSYGLSAADAEMARADLVIYNGLG				86
	+ ++V T +I+ D+ + GD D+ ++ DPH Y D + + ADL+ YNG+				
Sbjct 1	KLKVVATNSIADITKNIAGDKIDLHSIVPIGQDPHEYEPLPEDVKKTSEADLIFYNGIN				60
Query 87	LEESMQRHVDAAAENGVRT-----LPVGDHVDPLQFSA--DEHGAADPHFWTDPQRMILAV				140
	LE EN +T V D VD + +E G DPH W + + I+				
Sbjct 61	LETGGNAWFTKLVENAKKTENKDYFAVSDGVDVIYLEGQNEKGKEDPHAWLNLENGIIFA				120
Query 141	DAITGRVAELPGIDRAAVAANAARYRDELNNLSAHMTDRFAAIPAQRRLVTNHHVLYL				200
	I +++ ++ N Y D+L+ L D+F IPA+++ +VT+ Y				
Sbjct 121	KNIAKQLSAKDPNNKEFYEKNLKEYTDKLDKLDKESKDKFNKIPAEKKLIVTSQGAFKYF				180
Query 201	ADRFGFTVIGAVIPSGTTLASPSASDLESLAGAIRDAGVKAIFVDSSQPDRRLARVLAEQS				260
	+ +G V A I T + +++L +R V ++FV+SS DR + +++ +				
Sbjct 181	SKAYG--VPSAYIWEINTEEEGTPEQIKTLVEKLRQTKVPSLFVESSVDDRRPMKTVSQDT				238
Query 261	GVRVNVVSLYSELSAPGTEADSYLGMRTNTETIANNL		299		
	+ + ++++S++ G E DSY MM+ N + IA L				
Sbjct 239	NIPI-YAQIFTDSIAEQGKEGDSYSSMMKYNLDKIAEGL		276		

Figure 3.20 A Blast search against the PDB protein structures. Additional sequence similarity (27%) is detected with a surface antigen protein structure from *S.pneumoniae* (PDB code 2ZK7).

Furthermore, a Blast search against the non-redundant protein database revealed that Msmeg-6050 has homology with several ABC metal transporters from mycobacterium species and other organisms, such as *Streptomyces species* and *Amycolatopsis azurea* with high sequence identity of between 54-59%, respectively (Figure 3.21).

The structure of Msmeg_6050 protein was predicted using the Phyre server. This suggested that Msmeg_6050 shares a similar structure fold with various metal binding proteins, such as the zinc binding protein from *Treponema pallidum* (PDB code 1TOA) with 100% prediction confidence and 27% sequence identity [89, 103] (Figure 3.22).

zinc ABC transporter periplasmic-binding protein ZnuA [Streptomyces sp. PAMC26508]

Sequence ID: [ref|YP_007857411.1](#) Length: 328 Number of Matches: 1[▶ See 2 more title\(s\)](#)

Range 1: 23 to 325		GenPept	Graphics			▼ Next Match	▲ Previous Ma
Score	Expect	Method	Identities	Positives	Gaps		
315 bits(807)	8e-103	Compositional matrix adjust.	165/303(54%)	209/303(68%)	8/303(2%)		
Query	5	CIAACLAILLLAACGAPDGS	DGRRQIVVTTTILGDVVSEVVD	TADVRVMMKPNADPHSY	64		
Sbjct	23	LLMGLIALTTVAAGTACTTGGDQPRIVVTTN	ILGDI	TRQIVGDEAEVTVLMKPDADPHSF	82		
Query	65	GLSAADAEMARADLVIY	NGLGLEESMQRHVDA	AAAENGVRTLPVGDHVDPLQFSA-----	119		
Sbjct	83	GLSAVQAAELERADLVVFNGLGLEENVLRHVDA	AARESGVATFEAGKAVDPLTFHAGDGG	142			
Query	120	--DEHGAADPHFWTDPQRMILAVDAITGRVAE	-LPGIDRAAVAANAARYRDELNLSAHM	176			
Sbjct	143	PEEEAGQDPDFWTDPRVRKASGLIADQVVEHVGGVDEQAIRANAARYEKQLADLTW	M	202			
Query	177	TDRFAAIPAQRRLV	TNHHVLYGLADRFGFTVIGAVIPSGTTLASPSASDLES	LAGAIRD	236		
Sbjct	203	EKSPALIPEDERALV	TNHHVLYGLADRFGFRVIGAVVPSGTTLASPSSDLRALTRAMEE	262			
Query	237	AGVKAIFVDSQPDR	LARVLAEQSGVRVNVVSLYSELSAPGTEADSYLGM	MRTNTETIA	296		
Sbjct	263	AGVRTVFADSSQPDKLAQVLRTELGGQVSVV	LYSESLTRKDAGAGTYLQMMRANTTAIT	322			
Query	297	NNL	299				
Sbjct	323	DSL	325				

Zinc ABC transporter, periplasmic-binding protein ZnuA [Amycolatopsis azurea]

Sequence ID: [ref|WP_005159882.1](#) Length: 299 Number of Matches: 1[▶ See 1 more title\(s\)](#)

Range 1: 21 to 296		GenPept	Graphics			▼ Next Match	▲ Previous Match
Score	Expect	Method	Identities	Positives	Gaps		
327 bits(838)	7e-108	Compositional matrix adjust.	165/279(59%)	202/279(72%)	5/279(1%)		
Query	23	GSDGRR-QIVVTTTILGDVVSEVVD	TADVRVMMKPNADPHSYGLSAADAEMARADLVI	81			
Sbjct	21	GSGGKSASVVVTTN	ILGDI	TRAVVGDQAEVTVLMKPNADPHSPGISAQQAQVERAGLIV	80		
Query	82	YNGLGLEESMQRHVDA	AAAENGVRTLPVGDHVDPLQFSADEHGAADPHFWTDPQRMILAVD	141			
Sbjct	81	YNGLGLEEGMLRTVHTA	ENGVPALPAGDRANPISFAGNK---	PDPFWTDFARVDRVVK	137		
Query	142	AITGRV-AELPGIDRAAVAANAARYRDELNLSAHM	TDRFAAIPAQRRLV	TNHHVLYGL	200		
Sbjct	138	AI	V A + G+D A + ANA RYR E++ L MT++F IP +RRKLV	TNHHV GYL	197		
Query	201	ADRFGFTVIGAVIPSGTTLASPSASDLES	LAGAIRDAGVKAIFVDSQPDR	LARVLAEQS	260		
Sbjct	198	AQRVGFVVGAVIPGGTTLASPSSDLKALADTVRAAGVPVVPADSSQPDR	LARVLAEQA	257			
Query	261	GVRVNVVSLYSELSAPGTEADSYLGM	MRTNTETIANNL	299			
Sbjct	258	GLHVAVTPLFSELS	EPGEGAATYLEMMRANTESIT	TGL	296		

Figure 3.21 A Blast search of Msmeg_6050 against the non-redundant protein database. The highest similarity is with the putative zinc binding proteins from several organisms, such as *Streptomyces species* and *Amycolatopsis azurea* with sequence identity of between 54 and 59 %, respectively.

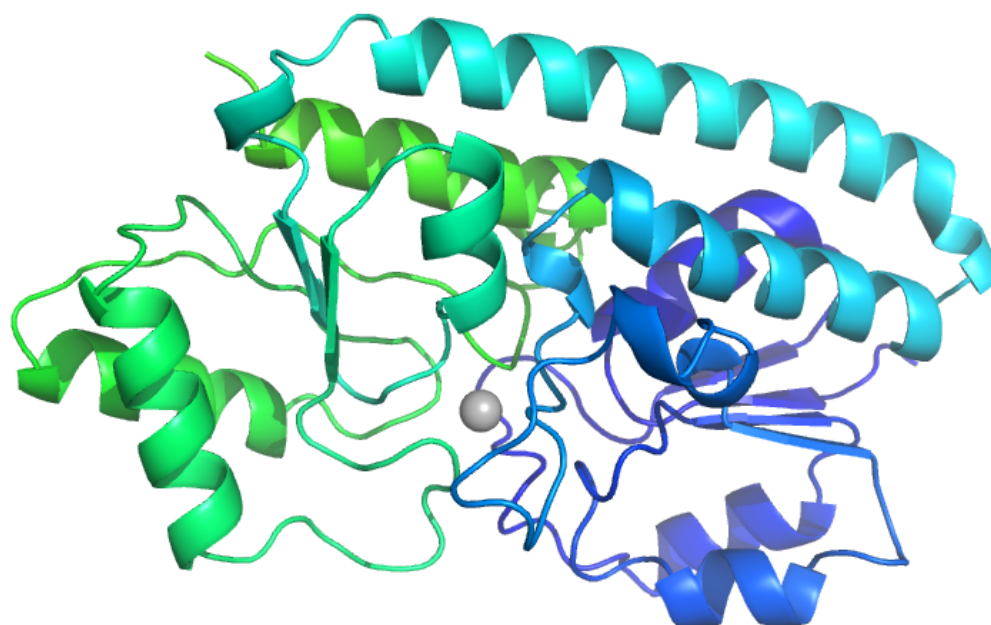


Figure 3.22 The predicted 3D structure fold of Msmeg_6050. Phyre results revealed that Msmeg_6050 is predicted to share similar fold with metal receptor from *Treponema pallidum* (PDB code 1TOA) with 271 residues of Msmeg_6050 sequence were covered in the comparison (90%) and low sequence identity of 27% and 100% prediction confidence [89]. Zinc is shown as a gray sphere in the center. Figure was made by using Pymol[98].

3.5.3 Purification of Msmeg_6050

The truncated Msmeg_6050 construct (20-300 a.a) has been purified using two steps of Ni-NTA affinity column and gel filtration. The same protocols were followed as given in section 3.2.2. 0.5 ml of protein at a final concentration of 10.5 mg/ml was produced using 1g of cell paste and prepared for crystallization experiment. SDS PAGE was used to evaluate the purification progress.

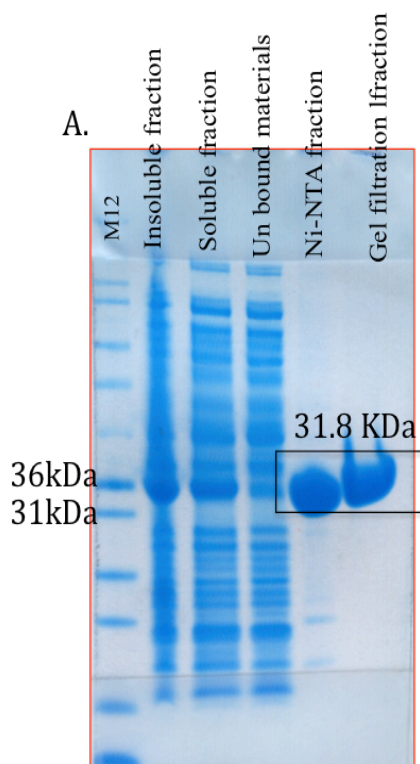


Figure 3.23 SDS gel showing protein purification steps of truncated Msmeg_6050. The molecular weight of Msmeg_6050 is ~32 (kDa). Gel-representing Ni-NTA affinity column purification, indicating Msmeg_6050 expressed in both soluble and insoluble fractions. The gel filtration purification step indicated high-purified protein fraction was obtained.

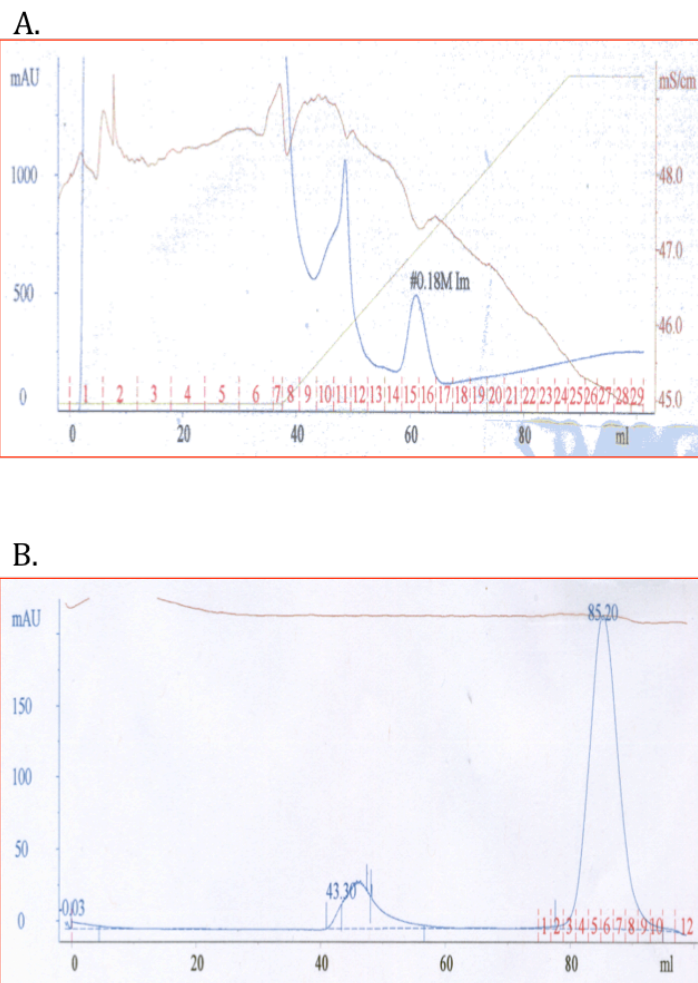


Figure 3.24 Chromatograms obtained during the purification of Msmeg_6050. Blue line represents absorption at 280nm, brown line represents conductivity and green line represents gradient of Imidazole concentration: **a.** Chromatography on HisTrap HP column; **b.** Gel filtration on HiLoad Superdex200 column. The gel filtration column step succeeded in removing low molecular weight contaminants left after the initial Ni-NTA purification.

3.5.4 Crystallization of Msmeg_6050

The truncated Msmeg_6050 (20-300 amino acids) was put into crystallization screens following the same protocols as given in section 3.3.3. Crystals were observed after one year incubation in two different crystallization condition of JCSG screen (Figure 3.25). However, X-ray diffraction test experiments showed that they were salt crystals. Thus, this project was suspended.

JCSG-A	JCSG-A6
(0.2 M Lithium sulfate, 0.1 M Na acetate pH 4.5, 50% PEG 400).	(0.2 M Lithium sulfate, 0.1 M Phosphate-citrate pH 4.2, 20% PEG 1000).

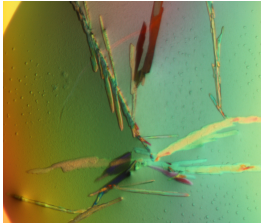
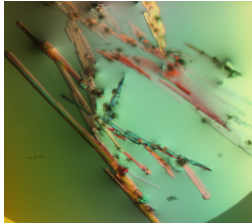
	
--	---

Figure 3.25 Photographs of Msmeg_6050 crystals that were observed after one-year incubation. However, X-ray diffraction test revealed that these are salt crystals.

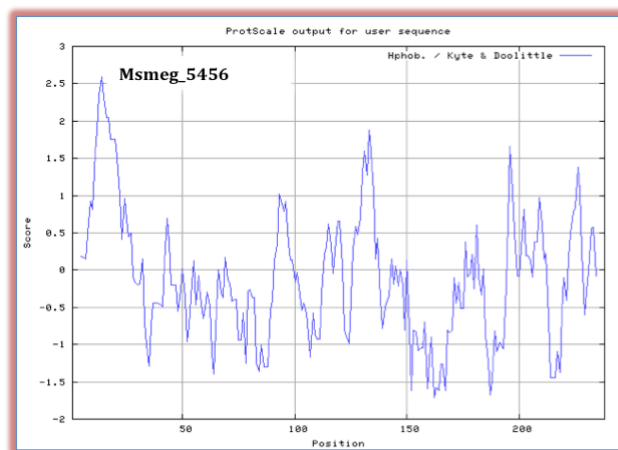
3.6 Studies on the Msmeg_5456 protein

3.6.1 Introduction

Msmeg_5456 (LpqN) is ≈ 24.7 (kDa), putative uncharacterized lipoprotein, with 238 amino acids and 717bp DNA length. LpqN has been annotated as a non-essential gene of *M.tuberculosis* [75]. LpqN was assigned as a potential target due to the presence of highly hydrophobic N-terminal 35 residues, (Figure 3.26). These hydrophobic residues were analyzed using the TMHMM server and revealed no potential transmembrane helices and thus, these residues are highly likely to form an N-terminal signal sequence [32, 101]. Analysis using the signal peptide prediction server SignalP showed that the first 32 residues were probably a signal sequence, and that the signal would be cleaved after residue Q32 [102].

However, analysis of the sequence of LpqN using the DOLOP server, revealed that LpqN does not meet the same criteria of a lipoprotein and thus a truncated LpqN was made based on the predicted cleavage site by the SignalP server [1, 38]. This construct (A33-P238) was made in an attempt to obtain a soluble protein for crystallization.

a)



b)

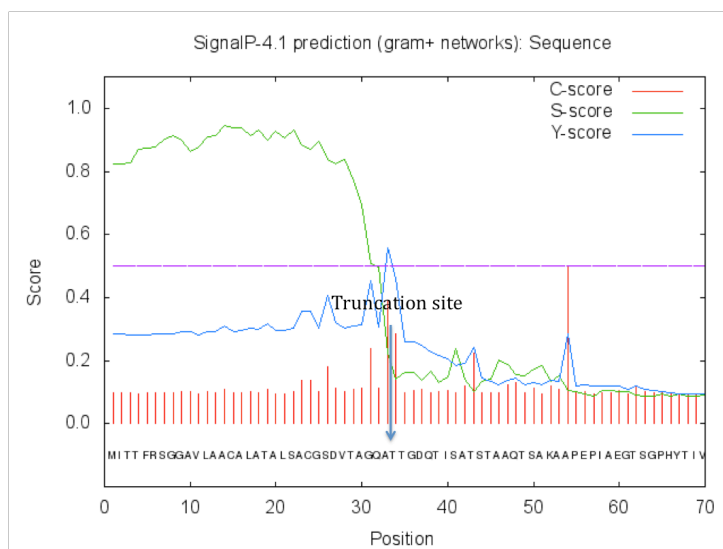


Figure 3.26 Hydropathy (a) and signal peptides (b) prediction plots of Msmeg_5456 (LpqN). a. The hydrophobicity plot was generated using the Kyte-Doolittle hydrophobicity scale program (Kyte and Doolittle 1982[74], which indicated hydrophobic residues within the first 32 residues of the N-terminal protein sequence. b) The prediction of signal peptides and the cleavage site required for the truncated protein [102]

3.6.2 Bioinformatics study on LpqN

A Blast search of LpqN against proteins of known structure and function in the PDB revealed only five hits to different crystal structures with sequence identity between 38 and 43%. The crystal structure of the putative regulator from *E.coli* CFT073 (PDB code 3HFI), and crystal structures of the interferon inhibitory domain and RNA domain from Reston Ebola virus (VP35) (PDB code 3L2A and 3KS4 respectively) share the highest identity of protein sequence of 43%, 39% and 39% respectively. In addition, LpqN shares a lower sequence identity with other RNA binding domains from human *Homo sapiens* (PdB code 4KRE) and the membrane rotor of the V-type ATPase from *Enterococcus hirae* (PdB code 2BL2) with 31% and 28% sequence identity, respectively. However, only a small part of the LpqN sequence (< 80/238) was covered in the alignments.

Furthermore, a Blast search against the non-redundant proteins revealed that LpqN is closely related to proteins among several mycobacterium species with sequence identity between 40-67%. These proteins include Mk35 antigen protein of *Mycobacterium vulneris* (67% sequence identity for 161 residues), the conserved LpqN protein from *M.tuberculosis* (48% sequence identity, 109 residues) and serine-threonine protein kinase of *M.smegmatis* (47% sequence identity, for residues 107/238) (Figure 3.27).

MK35 lipoprotein [Mycobacterium vulneris]

Sequence ID: [emb|CDO32200.1](#) Length: 233 Number of Matches: 1

Range 1: 1 to 233 [GenPept](#) [Graphics](#) ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
308 bits(790)	1e-102	Compositional matrix adjust.	161/240(67%)	187/240(77%)	9/240(3%)
Query 1		MITTFRSGGAVLAACALATALSACGSDVDTAGQATTGDTISATSTAAQTSAKAAPEPIAE			60
Sbjct 1		M T+ R+GG V+AA AL L+ACG +G AT + T S+ ++ T KAA EP			54
Query 61		GT--SGPHYTIVDYIRDNGITEMPVHRGDPGTPVLDIPIPGWADAGADTPEWAWSAMVS			118
Sbjct 55		GT+G YTIVDYIRDN ITE PVHRGDPGTP LD+PIP GW DA + PEWAW A+VS			114
Query 119		TDPAFADDPPIALMSRLTGDVDPKILEYAPNEIKNLPYDGESEGTAEDELSGFDAY			178
Sbjct 115		T P F+ DPP IIALMS+LTGDVDPKILEYAPNEI+NLPGY +G +G+AD+LSGFDAY			173
Query 179		QIGGMYVRDGGQRLIAQKTVVIPGQDGLYVLQNLNADGTEDQLGTLDDATSVIDEKTVITP			238
Sbjct 174		QIGG YV+DG RLIAQKTVVIPG DGL+VLQ NADGTEDQ+G L+DAT+ IDE+TVITP			233

serine/threonine protein kinase [Mycobacterium smegmatis str. MC2 155]

Sequence ID: [ref|YP_889750.1](#) Length: 618 Number of Matches: 1[▶ See 2 more title\(s\)](#)

Range 1: 403 to 618 [GenPept](#) [Graphics](#) ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
197 bits(502)	2e-55	Compositional matrix adjust.	107/228(47%)	145/228(63%)	12/228(5%)
Query 11		VLAACALATALSACGSDVDTAGQATTGDTISATSTAAQTSAKAAPEPIAEGTSGPHYTIV			70
Sbjct 403		VL + A+++ G + +AG A T + TA++T P+ E YTI			452
Query 71		DYIRDNGITEMPVHRGDPGTPVLDIPIPGWADAGADTPEWAWSAMVSTDPAFADDPPI			130
Sbjct 453		DYIRDN I+E PVHRGDP TP L +P PPGWADAG TP WA+SA+V+ DPA DPP +			511
Query 131		IALMSRLTGDVDPKILEYAPNEIKNLPYDGESEGTAEDELSGFDAYQIGGMYVRDGGQ			190
Sbjct 512		I+L+S+L G+VDP K+LE+APNE++NL ++ G T LSGF + Q+GG Y++DG++			570
Query 191		RLIAQKTVVIPGQDGLYVLQNLNADGTEDQLGTLDDATSVIDEKTVITP		238	
Sbjct 571		R I QKTVV+ G+YVLQ+NAD G L+D ID++ I P			618

Figure 3.27 A Blast search of LpqN against the non-redundant protein database. The highest sequence similarity is with putative Mk35 antigen lipoprotein from *M.vulneris*, and serine / threonine protein kinase from *M.smegmatis* with sequence identity of 67 and 47 %, respectively.

The structure of LpqN protein was predicted using the Phyre server. The result suggested that LpqN shares a similar structure fold with the uncharacterized protein from *Jonesia denitrificans* (PDB code 3LDY) and with the PA94 putative regulator from *Pseudomonas aeruginosa* (PDB code 3LDY) with 100% prediction confidence and 26% sequence identity, respectively (Figure 3.28) [89] [104].

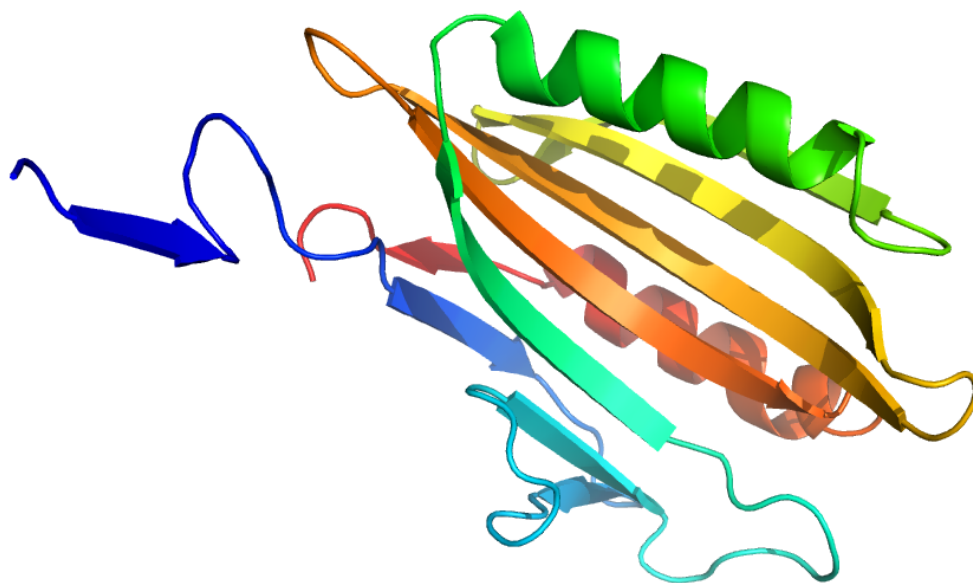


Figure 3.28 The predicted 3D structure fold of Msmeg_5456 (LpqN). Phyre results revealed that LpqN is predicted to share similar fold with uncharacterized protein from *Jonesia denitrificans* (PDB code 3LDY) (not published) with 153 residues of LpqN sequence were covered in the comparison (66%) and low sequence identity of 21% and 100% prediction confidence [89]. Figure was made using Pymol.

3.6.3 Protein purification of LpqN

The truncated LpqN (residues Q33-P238) was purified using a one step of Ni-NTA affinity column. The same protocol has been followed as described in section 3.2.2. LpqN protein of high concentration fractions were combined and transfer into 10 mM Tris-HCL buffer pH 0.8, and concentrated to 26 mg/ml for crystallization experiment. SDS PAGE was used to evaluate the purification progress (Figure 3.29).

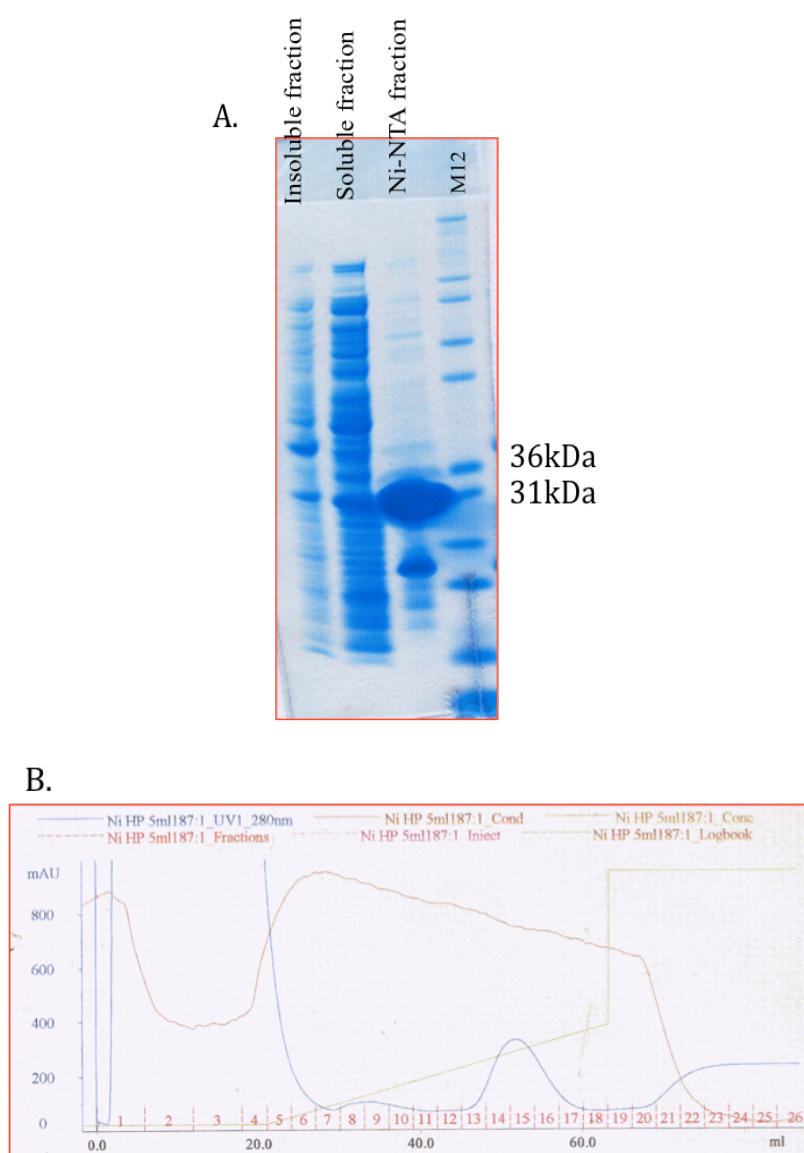


Figure 3.29 Chromatogram analysis of LpqN purification. a. SDS gel showing protein purification of truncated LpqN, the gel was run 4 days after the purification, an additional band with lower molecular weight appear, which might suggested some degradation has occurred. b. Chromatography on HisTrap HP column. Blue line represents absorption at 280nm, brown line represents conductivity and green line represents gradient of Imidazole concentration.

3.6.4 Crystallization of LpqN

For crystallization the truncated construct of LpqN (26 mg/ml protein in 10 mM Tris-HcL, pH 8.0) was put into crystallization screens following the same protocols as given in section 3.3.3. A single crystal was observed after one year incubation in one crystallization condition of PEG screen (Figure 3.30). However, X-ray diffraction test experiment showed that they were salt crystals. Thus, this project was suspended.

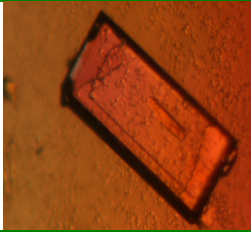
Crystal	Condition PEG-G10
	(0.2 M Potassium sulfate, 20 %(w/v) PEG 3350).

Figure 3.30 A photograph of LpqN crystal that was observed after one-year incubation. However, X-ray diffraction test revealed that this is a salt crystal.

Chapter 4

AgaE structure determination

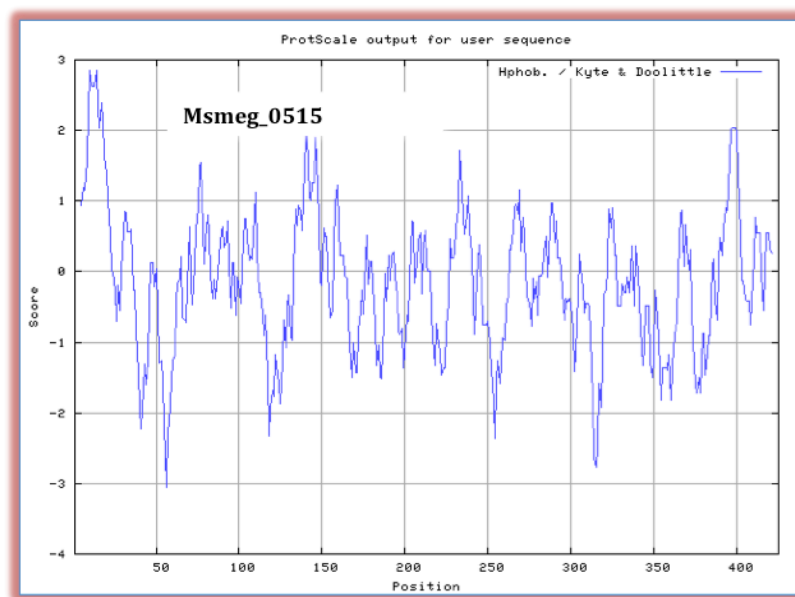
This chapter explains the materials, procedures and results of the cloning, over-expression, purification and crystallization of the Putative sugar binding protein Msmeg_0515 (AgaE).

4.1 Bioinformatics studies on Msmeg_0515 protein

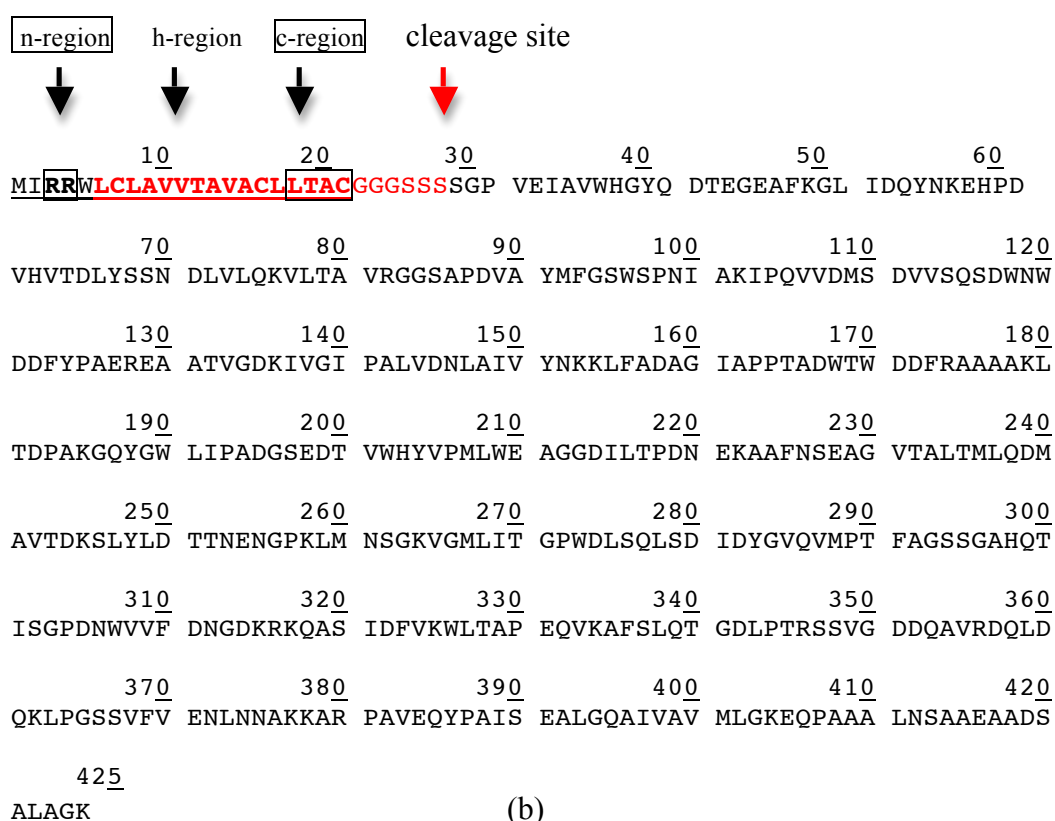
4.1.1 Target selection

Msmeg_0515 protein is one of the eight selected targets for structural studies. The original criteria for target selection was based on hydropathy plots of the entire gene produced in *M.smegmatis*. Msmeg_0515 was assigned as a potential target due to the presence of highly hydrophobic N-terminal 30 residues (Figure 4.1). This suggested that the N-terminal residues might be a trans-membrane helix or perhaps a signal sequence. Analysis using the TMHMM server indicated (not surprisingly) that the 30 N-terminal residues had a high potential of being a trans membrane helix (Figure 4.2a) [32, 101]. Analysis using the signal peptides prediction SignalP server [102] showed that the first 27 amino acid residues were probably a signal sequence, and that the signal would be cleaved after S27 (Figure 4.2b). Analysis of the sequence of Msmeg_0515 using the DOLOP server, which predicts lipid-anchored proteins, suggested that cysteine 21 would be site of lipid attachment and that the signal would be cleaved between Ala 20 and Cys 21 [1, 38]. In addition, DOLOP clearly showed the presence of the n-region, h-region and C-region of the lipoprotein features (Figure 4.1).

In the light of these analyses, it was decided to produce a construct that started from Ser 28 to the C-terminus in an attempt to obtain a soluble protein for crystallization.



(a)



(b)

Figure 4.1 Hydropathy plot of Msmeg_0515 and protein sequence. a) The hydropobicity plot was generated by using Kyte-Doolittle hydropathy scale program and indicated high value of hydrophobic residues within the first 30 residues. b) The protein sequence shows the signal peptide (underlined) within the first 20 residues. The lipoprotein features are highlighted and boxed based on DOLOP server. Red colour residues indicate the predicted transmembrane helix and the cleavage site is assigned by a red coloured arrow.

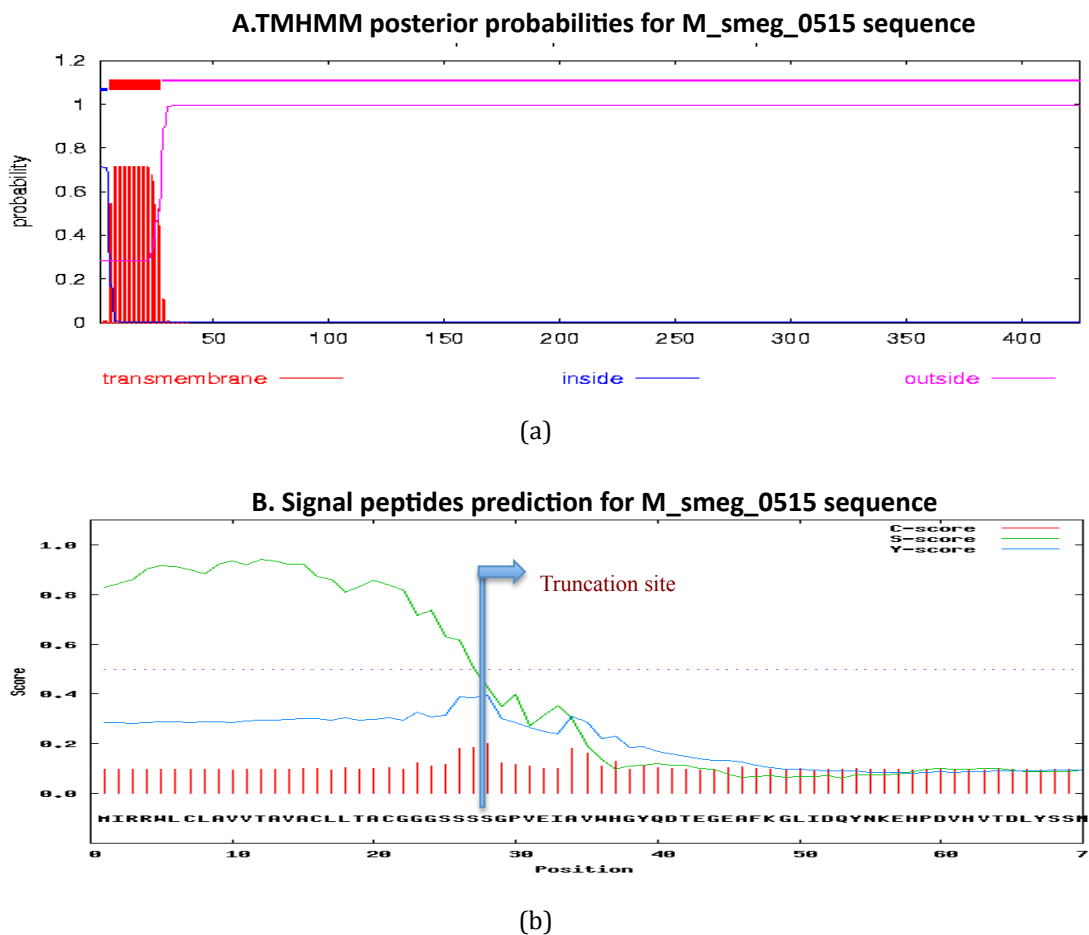


Figure 4.2 Predicted signal peptide and transmembrane helix within the first 30 amino acids of Msmeg_0515 (AgaE) protein sequence. a) A trans-membrane helix is predicted using TMHMM server, described by red lines at the N-terminal of the other protein sequence with pink lines showing the soluble part of protein. b) Prediction of signal peptides located within the first 27 residues of the sequence. The truncation of Msmeg_0515 (AgaE) was made based on this result.

4.1.2 Further analysis of Msmeg_0515

The primary sequence of Msmeg_0515 was compared against all non-redundant proteins in the NCBI database. The results revealed that Msmeg_0515 shares sequence homology to a putative extracellular solute binding protein of different Gram-positive bacteria species that belong to Actinobacteria genus, such as *Frankia* sp, *Conexibacter woesei*, *Catenulispora acidiphila*, *Salinibacterium* sp, with high sequence identity of 52%, 46%, 46% and 41%, respectively (Figure 4.3). Moreover, the Msmeg_0515 protein sequence showed homology with several putative sugar binding proteins and ABC type sugar dependent ABC transporters of the Gram positive *Streptomyces* sp and the gram variable –staining *Gardnerella vaginalis*, with sequence identity 41% and 29%, respectively.

Furthermore, in order to investigate the possible function and structure of Msmeg_0515, the primary sequence was also applied against all proteins with structures known in the protein data bank (PDB), which in turn, may lead to a predicted function or substrate specificity [105]. The results indicated several hits of numerous crystal structures that have similar structure but low sequence identity of between 31-21 % (Table 4.1). The most significant homology structure was with the crystal structure of ABC maltooligosaccharide / Acarbose binding protein (GacH) from *S.glaucescens* with 27% sequence identity (PDB code. 3K01). Additionally, the result of the Blast searches also indicated other crystal structures that are predicted to have a similar structure with Msmeg_0515, which are bound to different sugars, such as maltodextrin, acarbose, and D-glucose.

extracellular solute-binding protein [Frankia sp. Eu1c]
 Sequence ID: [ref|YP_004019447.1](#) Length: 439 Number of Matches: 1
[▶ See 2 more titles\(s\)](#)

Score	Expect	Method	Identities	Positives	Gaps
407 bits(1046)	3e-135	Compositional matrix adjust.	207/395(52%)	270/395(68%)	6/395(1%)
Query 33	I A V H G V Q D T E G E A F K G L I D Q Y N K E H P D V H V D L Y S S - N D L V L Q K V L T A V R G G S A P D V A Y				
Sbjct 48	I + H O Y D E L Q + N + H P V + + N D L Q K + G + D D + A Y				
Query 92	M F S S H S N I A K I P Q V D M S D V S G S D W H D D F P A E R E A A T V G K I V G I P A L V D N L A I V Y				
Sbjct 108	+ G S + + A K + P + V D + + V + + H + D F Y + E R + A T V K I V G I P A L V D N L S L V Y				
Query 152	N K K L F A D A G I A P P T A D W T M D F R A A A K L T O P A K G Q G W L I P A D S E D T V H V V P M L M E A				
Sbjct 168	N K K L F D A G V A P P T D S W Q D F R A A A K L T A G N G T G Y H S V D G S E D V V R Y L A M L W Q A				
Query 212	G D D I L T P D N E K A F N S E A G V A L T M L Q D M A V T D K S L Y D T T N E N G K L M S G K V M L I T G				
Sbjct 228	G D D I L T D N K A A F S A G A L A A L T Q L R D M T V D S V L T D T G W Y L L P S G K I A M L I T G				
Query 272	P W D L S Q L - S D I D Y V Q V M P F A G S S G A G Q T I S G F D M V F D M G R K R Q A S I D F P K W L D A P				
Sbjct 288	P W D S I S T S D V S V G Y T L D P - - - G Y N G H E T I S G P D I Y M L F D H S A P R Q A A I D F I T W L T S P				
Query 331	E Q V A F A L Q G D L P R S V G D D Q A V R D L D Q L P G S E V E N L N N A K K A R P A V E Q P A I S +				
Sbjct 345	+ P + + T G D L P R S + + L Q R P G V E V E N L M K R P + Y + S				
Query 391	E A L G Q A I V A M L G K E P A A L N S A E A A S A L A G K 425				
Sbjct 404	A + G Q + + V + I G + Q A A L + A + + + A L A G				

extracellular solute-binding protein [Conexibacter wosei DSM 14684]
 Sequence ID: [ref|YP_003397578.1](#) Length: 444 Number of Matches: 1
[▶ See 2 more titles\(s\)](#)

Score	Expect	Method	Identities	Positives	Gaps
361 bits(926)	3e-117	Compositional matrix adjust.	182/400(46%)	256/400(64%)	10/400(2%)
Query 30	P V E I A V H G V Q D T E G E A F K G L I D Q Y N K E H P D V H V - T D L Y S S N D L V L Q K V L T A V R G G S A P D				
Sbjct 49	P V E I F H G Q N T A Q T I E G L V D R F N A S P D V K V A E G A L A D S L Y Q K T T A A L A G G K P D				
Query 89	V A Y M F G S W S P N I A K I P Q V D M S D V S G S D W H D D F P A E R E A A T V G K I V G I P A L V D N L A				
Sbjct 109	V Y F C + + A + P + + D + D V + + M N D D F Y R E A T V K + + P A L + D + L A				
Query 149	I V Y N K L F A D A G I A P P T A D W T M D F R A A A K L T O P A K G Q G W L I P A D S E D T V H V V P M L M E A				
Sbjct 169	V Y N R L F R E A G I P A P R A G W T W D D Y A I A R Q L T D S S Q G F G A W P G D E D T V R L W F M V				
Query 209	W E A G D I L T P D N E K A F N S E A G V A L T M L Q D M A V T D K S L Y D T T - N E N G K L M S G K V G				
Sbjct 229	M + G G D + P D E + A F E + G + T + T + D H A V T D + S L Y D T + E + + M + G + G				
Query 267	M L I T G P W L S Q L - S D I D Y V Q V M P F A G S S G A G Q T I S G F D M V F D M G R K R Q A S I D F P K W L D A P				
Sbjct 289	M + T G W + + + + D Y G M D + + S T I S G D M + + P D N G R + A + + P				
Query 325	K W L T A P E Q V A F S L Q G D L P R S V G D D Q A V R D L D Q L P G S E V E N L N N A K K A R P A V E Q P A I S +				
Sbjct 346	Q W L L P E Q D A V M D V D A G S L F L R R S T A - Q P I W R H A G V E V G L D V P T A A L E Q A - R V R P T I				
Query 385	Q Y P A I S E A L G Q A I V A M L G K E P A A L N S A E A A S A L A G K 424				
Sbjct 404	Y P + S E A + G I V V + I G P A L + A + + A L A G				

sugar transporter sugar binding lipoprotein [Streptomyces cattleya NRRL 8057 = DSM 46488]
 Sequence ID: [ref|YP_006051438.1](#) Length: 429 Number of Matches: 1
[▶ See 2 more titles\(s\)](#)

Score	Expect	Method	Identities	Positives	Gaps
304 bits(779)	2e-95	Compositional matrix adjust.	175/432(41%)	248/432(57%)	23/432(5%)
Query 6	L C L A V V A V A C L L T A C G G G S S S - - - - - G P V E I A V H G V Q D T E G E A F K G L I D Q Y N K E				
Sbjct 2	L L G A S A V A A G L T A T V S G S A A M D G V G D R V T I E L H G G V G S S K V A E A M R E F R N R				
Query 58	H P D V H V - - T D L Y S S N D L V L Q K V L T A V R G G S A P D V A Y M F G S W S P N I A K I P Q V D M S D V S G				
Sbjct 62	H P K I R V D A G G G A V A D M L Q K V T A A L A A G A P V A Y I G S L P N I A R S P Q V D L T S W T F G				
Query 116	S D N W D D F P A E R E A A T V G K I V G I P A L V D N L A I V Y N K L F A D A G I A P P T A D W T M D F R A				
Sbjct 122	G A T P M Q V P A E R A V T V G S V G A L P A L I D L A V Y N K L F A D A G I A P P T A D W T M D F R A				
Query 176	A A A L T O P A K G Q G W L I P A D S E D T V H V V P M L M E A G D I L T P D N E K A F N S E A G V A L T				
Sbjct 182	A + L T P D + G + G P G E D T M W L W P M I D L G G E I V G P D G R S I G F - A D G V R A L Q				
Query 236	M L Q D M A V T D K S L Y D T T - - N E N G K L M S G K V M L I T G P W L S Q L S D - - I D Y G V Q V M P F				
Sbjct 241	L + D + D + + D + + G V + E + + G + + M + T G W L + + D + D G V + S T +				
Query 292	A G S S G A G Q T I S G F D M V F D M G R K R Q A S I D F P K W L T A P E Q V A F S L Q G D L P - T R S V G				
Sbjct 300	- - - S G R L I T S G P D M V F D M G R S A K E F V S L I Q P D G V D L L M L P L R S A				
Query 351	D O A V R Q D L Q K L P G S S V E N L N N A K K A R P A V E Q P A I S E A L G Q A I V A M L G K E P A A				
Sbjct 357	+ R + + G V F + L + A + R P + Y P I S + A L Q A I + + V + G P A A				
Query 411	I N S A E A A S A L 422				
Sbjct 414	+ + A + A A L				

ABC transporter substrate-binding protein [Gardnerella vaginalis]
 Sequence ID: [ref|WP_004123085.1](#) Length: 423 Number of Matches: 1
[▶ See 1 more title\(s\)](#)

Score	Expect	Method	Identities	Positives	Gaps
175 bits(443)	3e-46	Compositional matrix adjust.	125/427(29%)	196/427(45%)	21/427(4%)
Query 6	L C L A V V A V A C L L T - A C G G S S S - S G P V E I A V H G V Q D T E G E A F K G L I D Q Y N K E H P D V				
Sbjct 7	I C - A L I G A A M I I S V S A C S A K S D A N G A T S I I W Y N D G A N A T F D A M V D P N A S H K M I				
Query 62	H V T D L Y S S N D L V L Q K V L T A V R G G S A P D V A Y M F G S W S P N I A K I P Q V D M S D V S G S D W H D				
Sbjct 66	K I K T A S V N S D P M T L R A S A S K S L P I S I D S L W V P Q I A R K M L D L S K V I S S K - T L D 123				
Query 122	D P P A E R E A A T V G K I V G I P A L V D N L A I V Y N K L F A D A G I A P P T A D W T M D F R A A A R L T				
Sbjct 124	D P A + + + K V + P + N L A + Y N K + + A G + P T W D + A +				
Query 182	D P - A G Q Y G W L I P A - D G S E D T V H V V P M L M E A G D I L T P D N E K A F N S E A G V A L T M L Q D				
Sbjct 184	E K T G K P G Y L L T Q A G D N S E G L T W N F Q V N L N Q A G S E L T K D N S K A A P N P E G K K A M F W M D				
Query 240	M A V T D K S L Y D T T N E N G K L M S G K V M L I T G P W L S Q L S D - - I D Y G V Q V M P F A G S S G				
Sbjct 244	L I K S G V P T A R W G E - - - - - F E K G G S A G S G S W I N A D P P F P P V A R A P H - - P K D G				
Query 297	A H Q I T S G F D M V F D M G R K R Q A S I D F P K W L T A P E Q V A F S L Q G D L P - T R S V G				
Sbjct 296	T G + + V E N D R + A + + M + P E Q V + S + G L D S V				
Query 357	D O L D Q K L P G S S V E N L N N A K K A R P A V E Q P A I S E A L G Q A I V A M L G K E P A A L N S A E				
Sbjct 356	D + + + P F V E + + A R P Y P I S + A + G + + A L + A +				
Query 417	A A D S A L A 423				
Sbjct 415	A + V A 421				

Figure 4.3 Sequence alignments of Msmeg_0515 against all non-redundant proteins in the NCBI database. The results suggested that Msmeg_0515 is a sugar binding protein.

Protein	Organism	Ligand	Sequence identity %	PDB code
GachH Receptor	<i>S.glaucescens</i>	Malto-oligosaccharide-Acarbose	27% (105/396)	3k01
Maltose binding protein (PfMBP)	<i>P.furiosus</i>	Maltotriose	22% (74/333)	1ELJ
XOS BINDING PROTEIN	<i>Bifidobacterium animalis subsp. lactis BI-04</i>	Xylotetraose	26% (110/428)	3ZKK
UgpB binding protein	<i>E.coli</i>	Sn-glycerol-3-phosphate	24%(98/412)	4AQ4
Maltose binding protein (TvuCMBP)	<i>Thermoactinomyces vulgaris</i>	Gamma-Cyclodextrin	26%(79/299)	2ZYK
TMBP	<i>T.litoralis</i>	Trehalose	24% (97/402)	1EU8
Maltose binding protein (Male)	<i>Alicyclobacillus acidocaldarius</i>	MALTOSE	27%(105/36)	1URG
Maltose binding protein (Male)	<i>The Phytopathogen X.citri</i>	Unknown	22% (88/399)	3UOR
Extracellular Solute-binding Protein	<i>S.aureus</i>	PEG	23%(92/407)	4HS7

Table 4.1 Results from Blast search of Msmeg_0515 against all PDB structures. The first 10 hits revealed that Msmeg_0515 might well share similar structure with other carbohydrate binding proteins from different organisms.

4.1.3 ABC transport system

As the sequence alignment suggested that Msmeg_0515 might possibly be part of an ABC transport system, a brief overview of these proteins is given below:

The ATP binding cassette is a very large super-family of proteins that are expressed in all Kingdoms of life, including bacterial and eukaryotic species [106]. They function as transporters of different macromolecules, such as carbohydrates, amino acids, ions and antibiotics. They are divided into two subfamilies based on their function, substrate direction and translocation as importers and exporters [107]. In the *E. coli* and *Bacillus subtilis* genomes, ~5% of the expressed gene encode ABC transporters [108, 109] whereas, in *M. tuberculosis*, 2.5% of its genome encodes putative ABC transporters, only one of which has been recently characterized structurally and biochemically [110, 111]. Most ABC transporters in bacteria are importers and they function to uptake essential molecules into cells, such as the maltose permease of *E. coli*. In contrast, ABC transporters in eukaryotic systems only function as exporters of essential nutrients [107, 112]. In addition, ABC exporters have been found in all organisms, whereas, importers are only present in prokaryotes [113].

4.1.4 Structure of ABC transporters

In general, the structure of the ABC transporters system is composed of four domains joined together (figure 4.4); the first two domains are called the membrane spanning domain (MSDs) and they function as a translocation pathway. The second two domains are called nucleotide binding domains (NBDs) or ATP hydrolyzing domains and they function as energy providers to transport specific substrates through ATP binding and hydrolysis. An additional domain is found in the bacterial ABC transporter called the periplasmic binding domain (PBD) or the substrate binding domain (SBD). This domain is the most studied and divergent one among the ABC transporters in bacteria. It is found in the periplasm of Gram negative bacteria as a soluble protein, and as a cell membrane lipid-anchored protein in Gram positive bacteria and *Mycobacterium* [114, 115].

In bacteria, the domains of MSDs and NBDs are either organized as two domains encoded within the same gene of a specific DNA cluster or as individual domains within two separate genes of the same DNA cluster. However, in eukaryotes both MSDs and NBDs are encoded in one gene as multi-domains [110, 116, 117].

Furthermore, the membrane-spanning domain MSDs is less well conserved than the nucleotide-binding domain NBDs among the ABC transporters in eukaryotes [116, 117].

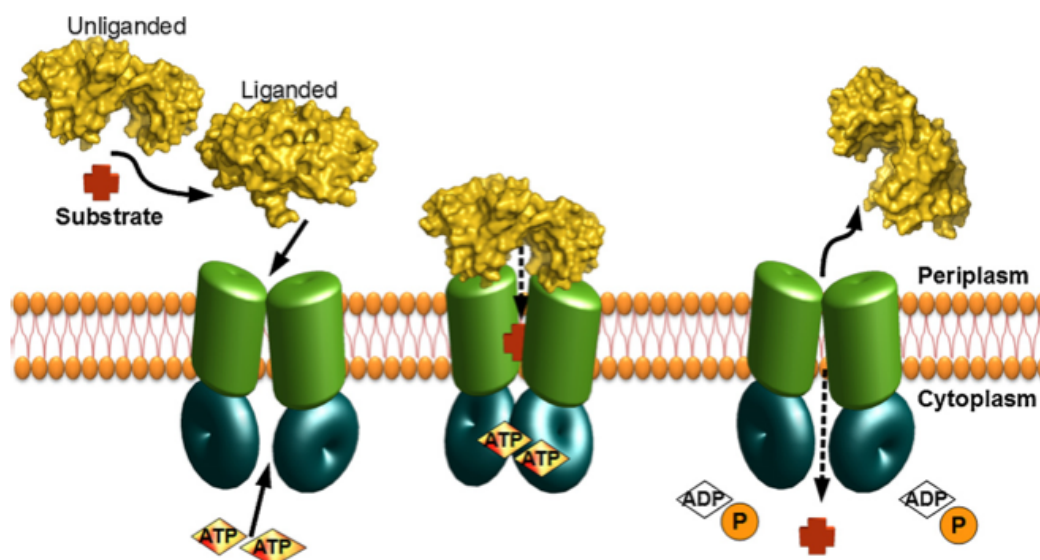


Figure 4.4 A schematic diagram of ABC transporter structure and the transport process. The first step of transportation of a specific substrate is the binding of the substrate into its extracellular receptor (yellow). Then, the receptor binds to its second component of membrane spanning domains (green) to be passed through the cell membrane. The third component of ABC transporters is the Nucleotide binding domains (cyan) or ATPase, which receives the substrate and transports it into the cell cytoplasm by ATP hydrolysis ATPs hydrolyse action; it also contributes to the dissociation of the extracellular receptor with unliganded form. The figure was adapted from [118, 119].

In *E.coli* and *B.subtilis*, the ABC transporters have been divided based on the type of their substrates into 10 and 12 subgroups, respectively; however, in *Mycobacterium*, only nine subgroups are found [108-110]. Five families of mycobacterium ABC transporters are involved in the import of the wide range of nutrients, such as peptides, amino acids, sugars, iron and anions, whereas the other four families are thought to be responsible for the export of antibiotics, drugs and other unknown substrates [110].

4.1.5 The Substrate binding domain (SBD)

The substrate binding protein is a significant domain of the ABC transporter system that imports and exports essential nutrients through the permeable barrier of the cell membrane [120, 121]. The binding proteins exist in both the Gram negative and Gram-positive bacteria; however, they are diverse in their location in both organisms. They have been found in the periplasmic space of Gram-negative bacteria and anchored to the inner cell membrane of Gram positive bacteria as a lipoprotein [115]. These types of lipoprotein are attached to the membrane by the conserved cysteine that is found in the N-terminal of the protein as part of the signal peptide that directs and translocates the protein. A lipid is covalently attached to the cysteine and this lipid embeds itself in the membrane [122]. It has been suggested that each binding protein has its own affinity that is specific to the particular substrate or to several substrates of the same class [123].

4.1.6 Carbohydrate binding protein dependent ABC transporters

The ABC Carbohydrate transporters are divided into two subfamilies: CUT 1 and CUT2, based on their substrate specificity [124]. The CUT 1 family is responsible for the transportation of di and oligosaccharides and many other nutrients, such as polyols, whereas, the CUT 2 family is responsible for the transport of monosaccharide only [124, 125].

M.smegmatis possesses 28 putative systems of carbohydrate transporters that belong to five different families, such as the ABC family (figure 4.5), the phosphotransferase system (PTS), the major intrinsic protein family (MIP), the major facilitator superfamily (MFS), and the sodium solute superfamily (SSS) [3, 126-129]. Nineteen of these systems belong to the ABC family. All ABC carbohydrate protein components are encoded within the same operon including the substrate binding protein [110, 126]. The genomic DNA of *M.smegmatis* possesses 18 genes that encode sugar-binding proteins belonging to the ABC permease family, based on sequence homology with other ABC transporters from different organisms, such as *E.coli* and *Streptomyces species* [3]. The substrate specificity for several binding proteins encoded in the *M.smegmatis* genome have been determined on the basis of the sequence and genetic similarity with other carbohydrate binding proteins. These ABC permease proteins are predicted to be responsible for the transport of sugars,

such as α -glucosides, sugar alcohol, xylose, β -galactosides and arabinose in *M. smegmatis* [3].

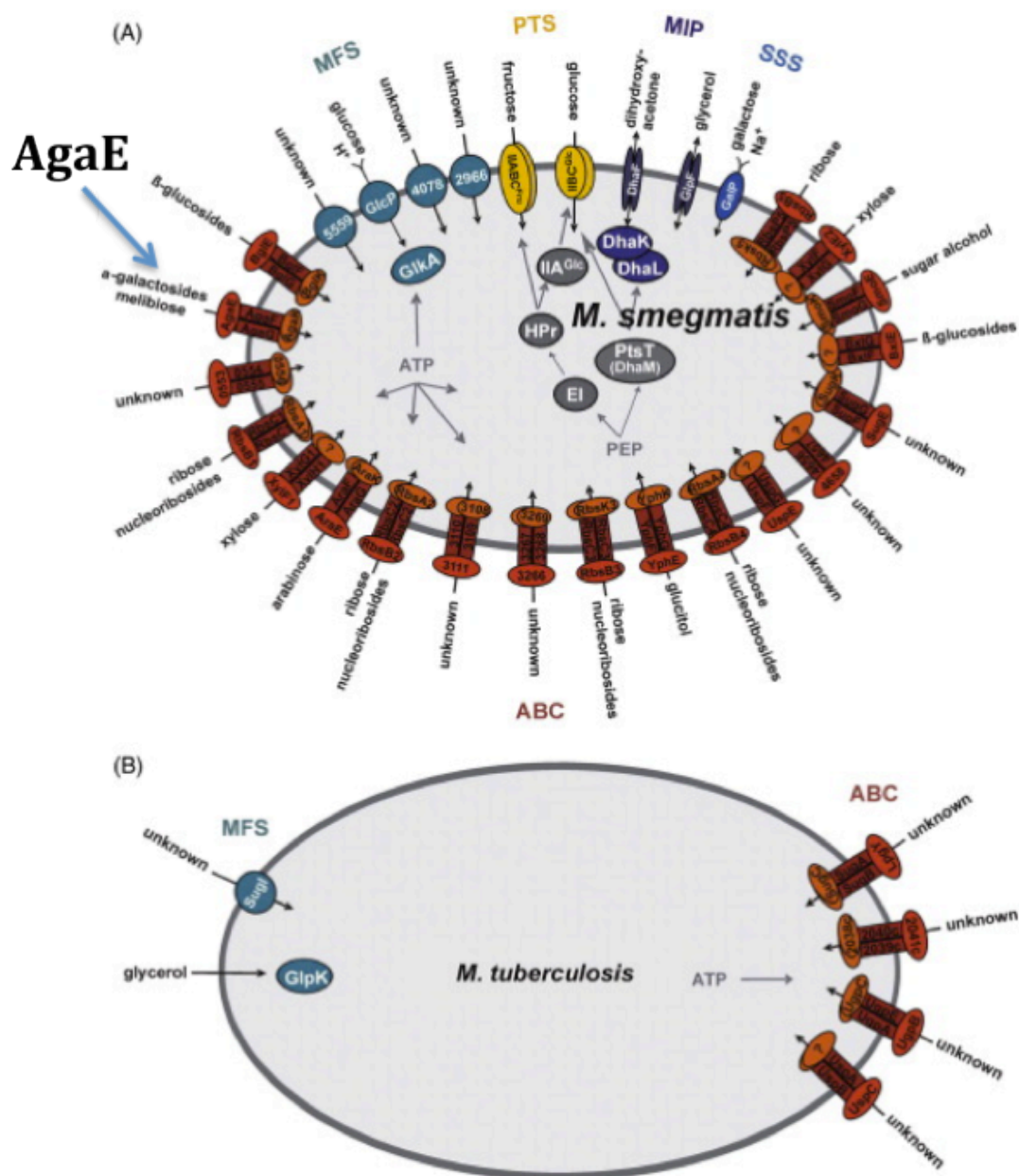


Figure 4.5 Carbohydrate binding proteins encoded by different transport systems with their putative predicted substrates in both (a) *M. smegmatis* and (b) *M. tuberculosis*. *M. smegmatis* possesses 28 putative systems involved in carbohydrate transport, 19 of which are ABC transporter (dark red). *M. tuberculosis* has only four ABC carbohydrate transporters. Figures were adopted from [3].

4.1.7 Genetic analysis of Msmeg_0515

Msmeg_0515 is one of the predicted ABC permease proteins located in cluster two of the sugar transporters of the *M.smegmatis* genome; it is called *agaE* [3]. The operon that contains the Msmeg_0515 (*agaE*) gene was predicted to be a α -galactosides uptake operon, located downstream of the first cluster that is responsible for β -glucoside uptake genes [3]. The operon contains several genes *agaR*, *Z*, *S*, *X*, *P*, *A*, *E*, *F*, *G*, *K* and *B* (Figure 4.6). The first six genes of the operon that are located upstream of Msmeg_0515 (*agaE*) gene are predicted to have several functions as follows: Msmeg_0509 or *agaR* is predicted to function as a transcriptional regulator for the whole operon [3]. The second gene in the operon known as Msmeg_0510 or *agaZ* was predicted to be a D-tagatose –bisphosphate aldolase binding protein that is followed by Msmeg_0511 or *agaS* gene that encodes isomerase enzyme [3]. Both Msmeg_0512 that is predicted to belong to the ATPase family and Msmeg_0513 that is also, predicted to be integral to the membrane protein are of unknown function. The sixth gene in the operon is called Msmeg_0514 or *agaA*; it encodes an enzyme that functions as hydrolyzing enzyme for α -galactoside sugars [3, 130]. Downstream of the Msmeg_0515 (*agaE*) gene are further genes that belong to the components of the ABC transporter family, such as Msmeg_0516 and Msmeg_0517 (*agaF* and *agaK*), respectively (Figure 4.7). These proteins are predicted to encode the membrane spanning domains (MSDs), and Msmeg_0518 (*agaG*) is predicted to encode the nucleotide binding domains (NBDs), with predicted substrate specificity for glycerol 3-phosphate transport. The last gene in the operon located downstream of Msmeg_0515 (*agaE*) is a gene that encodes a porin; it has been suggested that this protein works as a gate for the sugar to get in and to be transported by the ABC permeases, including Msmeg_0515 (AgaE) [3, 131].

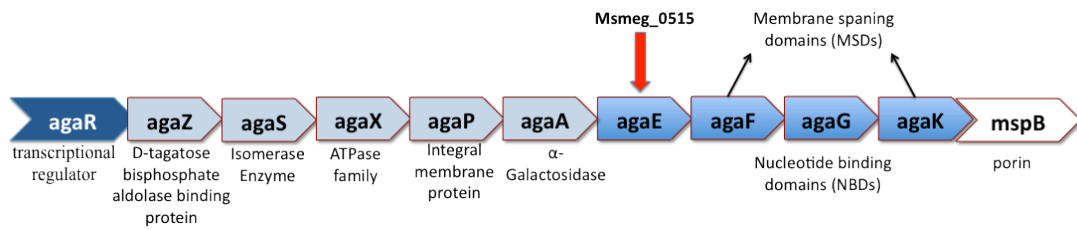


Figure 4.6 The operon that encodes the *Msmeg_0515* gene and the neighboring genes. The dark arrow color indicates the regulatory gene (*agaR*), the light blue indicates the carbohydrate metabolic genes and the dark blue indicates the genes involved in the transport system ([3]).

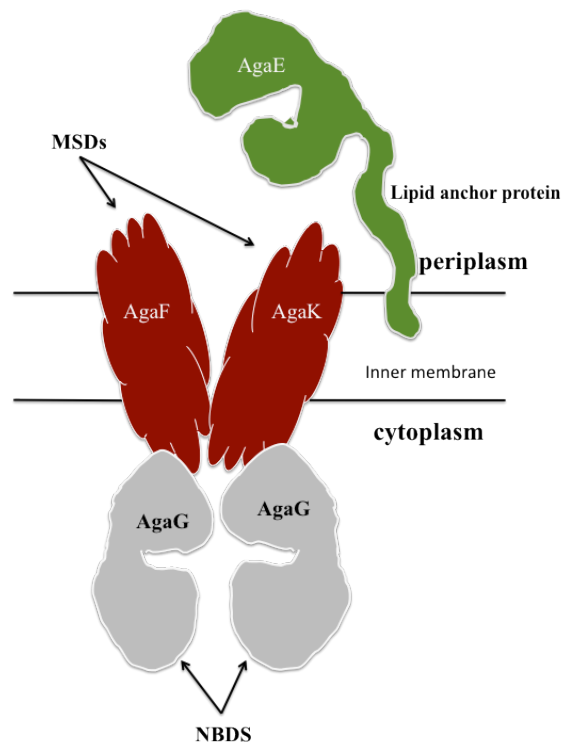


Figure 4.7 A schematic diagram of the putative ABC components, structure and organization of AgaEFGK. The first component is the extracellular binding protein (AgaE, green), which is predicted to be membrane anchored by the fatty acyl chain linked to the conserved cysteine in the N-terminal protein sequence. The second and third components are the membrane spanning domains (AgaF & K, red) and the nucleotide binding domains (AgaG, grey).

4.2 Cloning of Msmeg_0515 (*agaE*) gene

4.2.1 Genomic DNA preparation

The fast growing *M.smegmatis* MC2 155 strain was kindly provided by Prof. Jeff Green (Sheffield University) in an agar plate, which was used to grow a primary culture in 5ml LB medium. One colony was inoculated into a 5 ml LB medium and incubated at 37°C for 3 days. As the culture did not have an antibiotic selection gene, the culture was daily checked by the eye to see if there was any contamination. After 3 days, genomic DNA from the cells was extracted using keyPrep bacterial genomic kits (ANACHEM®) as pointed out in section 2.1.2. The genomic DNA was stored at -20°C.

4.2.2 primers design and gene amplification of *agaE*

To aid in the purification of the AgaE protein, it was decided to clone the AgaE 28-425 truncated protein into a pET28a vector, this plasmid contains NdeI and HindIII restriction sites and so primers were designed with an NdeI site in the forward primer and an HindIII site in the reverse primer (Table 4.2). These restriction enzymes were selected, as the gene itself did not contain these cut sites. The designed construct was missing the N-terminal 1- 27 amino acid residues and the primers were designed lacking the first 81 bases of the *agaE* gene, which lacks its own methionine as a start codon (Figure 4.1). The restriction site of NdeI enzyme was used to provide an alternative start codon for gene expression. All primers were synthesized by Eurofins MWG Operon Company in 100 µM stock and stored at -20°C.

To amplify the *agaE* gene, various gradient PCR runs were undertaken with different annealing temperatures between 55 and 65°C, in order to optimize the quality of the PCR product. The *agaE* gene was amplified best at an annealing temperature of 58°C (Figure 4.8). 1% agarose gel electrophoresis was used to evaluate and check that the resulting PCR product was pure enough to be used for an efficient ligation reaction with the pET28a vector. The PCR product then was extracted using gel extraction kit as described in section 2.1.7.

Primers	Sequence
Msmeg_0515 forward NdeI (F)	5'-CCCCATATGGTGATACGACGCTGGTTG-3'
Msmeg_0515 forward NdeI (T)	5'-CCCCATATGTCGTCATCGGGTCC-3'
Msmeg_0515 Reverse HindIII	5'-GGGAAGCTTTTTGCCGCCAGGG-3'

Table 4.2 Primers of the full length and truncated genes of AgaE. The forward primers contain the NdeI restriction site (red letters), and the reverse primer contains the HindIII restriction site.

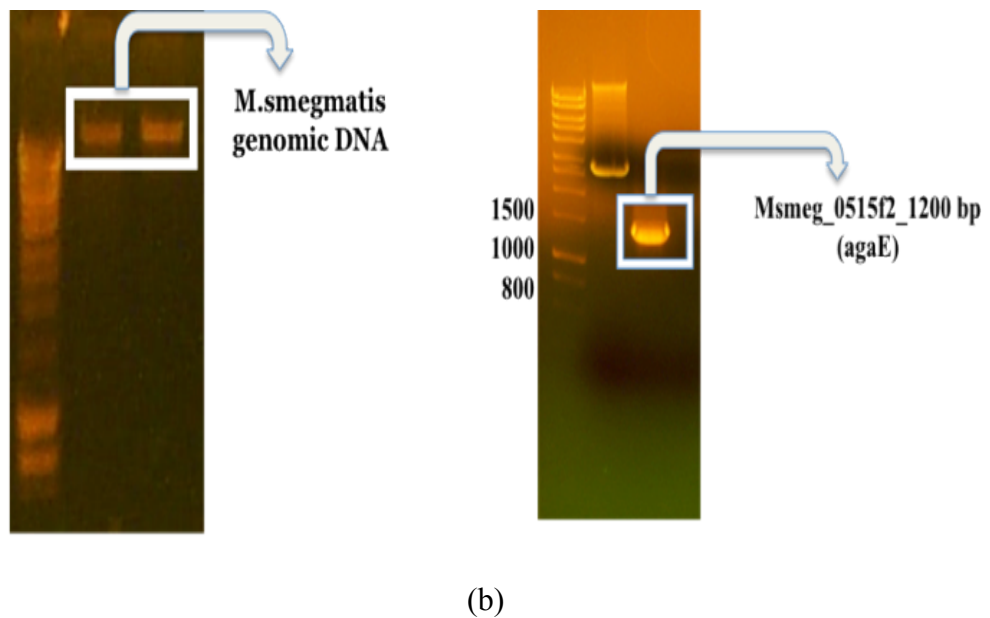
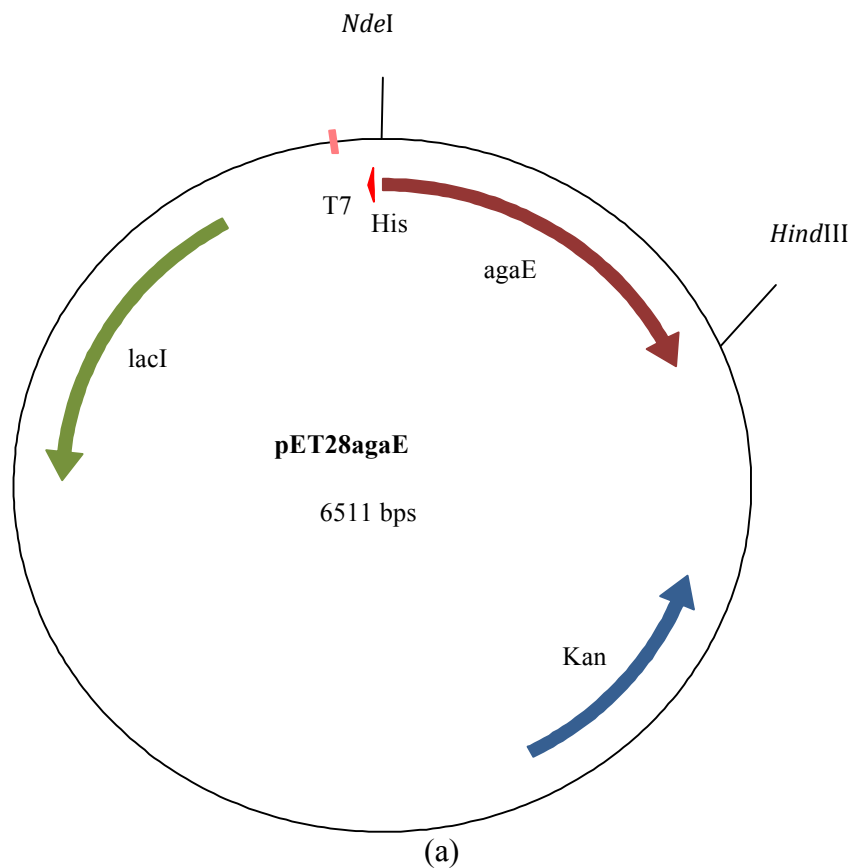


Figure 4.8 A schematic representation for the DNA cloning of the *agaE* (full length) gene to pET28a plasmid . A. displays the circle pET28agaE (6511 bp) that contains the *agaE* gene with 6×His tag at C-terminal and kanamycin resistant gene. B. the genomic DNA extraction of the *M.smegmatis* (left) and the PCR purification of the *Msmeg_0515* gene using 1% agarose electrophoresis gel stained with ethidium bromide, and analyzed under UV light.

4.2.3 Restriction digestion of the PCR product and pET28a plasmid

Both the PCR product and plasmid were double digested by NdeI and HindIII in order to provide a sticky end suitable for cloning purposes. The PCR product and the plasmid were incubated with enzymes and the required buffer at 37°C for 2 hours. 1% of agarose gel electrophoresis was used to recover the DNA fragments and to remove the enzymes. The DNA was extracted from a gel by using the Qiagen DNA gel extraction kit. The restriction digestion reaction was set up based on the New England Biolabs protocol as described in section 2.1.8.

4.2.4 Sticky end ligation into pET28a plasmid

In two separate eppendorf tubes, a ligation mixture of 3:1 and 1:3 of the digested pET28a plasmid and PCR product, respectively were added in order to provide sticky end ligation. A T4 ligase (New England Biolabs) was added to the reaction and the mixture was incubated at 37°C for 1 hour as described in section 2.1.9.

4.2.5 Sub-cloning into Dh5 α

The *E.coli* DH5 α strain was used for sub-cloning the ligation mixture of pET28a and Msmeg_0515 gene. One of the 50 μ l pre-aliquoted tubes stored at -80°C was used to transform the 1 μ l of the ligation mixture after 5 minutes of gentle resuspension on ice. The cells were incubated on ice for 30 minutes, and heat shocked at 42°C for 30 seconds. The cells were then incubated back on ice for 2 minutes to complete the transformation. Then, 450 μ l of LB media stored at room temperature was added to 50 μ l of cells to reach a final volume of 500 μ l. The cells were incubated at 37°C for one hour. The transformed culture was then centrifuged for one minute to collect the cells. The media was discarded and the cells were re-suspended in 50 μ l of LB media. Finally, LB media plates containing 50 μ g of kanamycin were used to grow the cells overnight at 37°C.

4.2.6 Identification of Successful cloning

Successful clones were identified by a restriction digestion experiment. Several colonies of the transformed cells were inoculated into 5 ml of LB media containing 50 μ g of kanamycin and incubated at 37°C overnight. Then, the culture was centrifuged at 4000 \times g for 5 minutes. The LB media were discarded and the plasmid in the cells was recovered using a Qiagen mini prep kit (section 2.22). The resulting DNA plasmid was digested using NdeI and HindIII enzymes and incubated at 37°C for 2 hours. The reaction mixture was loaded onto a 1% agarose gel, which showed two clear bands, of the correct size for plasmid and the insert (Figure 4.9).

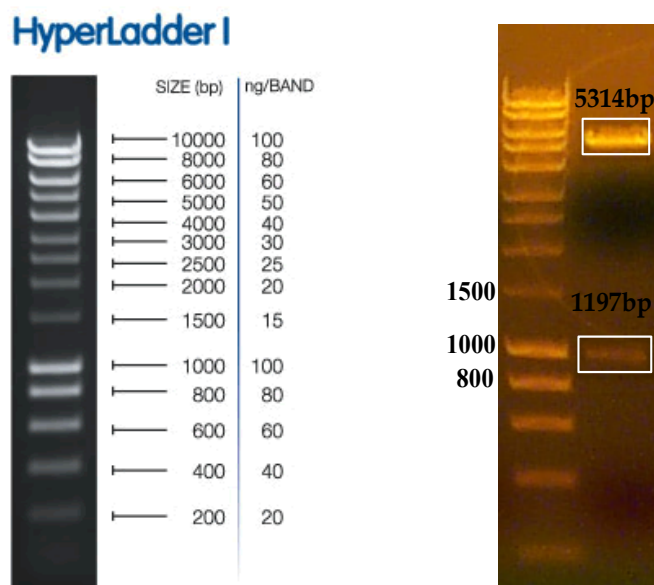


Figure 4.9 Restriction digestion of pET28a that contains AgaE. The top band corresponds to the pET28a vector (~ 5314 bp), and the lower band corresponds to the AgaE (1197 bp).

4.3. Protein expression and purification of AgaE

4.3.1 Transformation of pET28agaE to Bl21 (DE3) strain

In order to be able to test the expression of the cloned AgaE truncated protein, the successful cloned gene was transformed to a competent bacterial strain of BL21 (DE3) using the protocol described in 4.2.3. After transformation, LB media containing 50 μgml^{-1} kanamycin was added to 80 μl of cells, to reach a final volume of 100 μl and incubated at 37 °C for further cell growth. After 1 hour of incubation, the cells were centrifuged and discarded. All cells were re-suspended by 50 μl of the LB media and spread onto 50 μgml^{-1} kanamycin LB plates. Finally, the plates were incubated at 37°C overnight.

4.3.2 Optimization of AgaE overexpression

One colony of the transformed AgaE in Bl21 (DE3) strain was inoculated in 5 ml of LB media supplemented by 50 μgml^{-1} kanamycin; it was incubated overnight at 37 °C as a primary culture. This culture was used to make a secondary culture using 250 ml flasks. 500 μl of the primary culture were inoculated into 50 ml of fresh and sterile LB media supplemented by 50 μgml^{-1} of kanamycin. The secondary culture was then incubated for 2 hours until an optical density of 0.6 (OD_{600}) was reached. Before induction, 1ml of the culture was taken and the cells were stored at -20°C to be used in an evaluation expression as a pre induction fraction. In order to find out an optimal temperature for the AgaE protein expression, the cultures were induced with 50 μl of 1mM IPTG and incubated at different temperatures of 18 °C, 25 °C and 37 °C. After 4 hours of induction, the cultures were transferred into Falcon tubes and centrifuged at 3600 xg for 20 minutes at 4°C by using sigma 3-16k centrifuge. The LB media were discarded and the cells were stored at -80°C. The cells were thawed on ice in the next day and re-suspended with 2ml of Tris-HCL buffer, pH 8.0. The cells were sonicated on ice 2x at 16 microns for 20 seconds. In addition, the pre-induction fraction was resuspended in 100 μl of Tris buffer to be run on SDS gel. 5 μl of SDS loading buffer and 1 μl of DTT were added to 15 μl of (15 μg) all fractions; they then were boiled at 95°C for 5 minutes. Protein concentration was assayed using the method of Bradford. Finally, a 12 % of SDS-PAGE gel was run at 200 for 45 minutes to monitor the protein expression. As demonstrated by the SDS-gel, the

protein was highly expressed at all temperatures (figure 4.10); thus, 37°C was selected for large-scale protein purification.

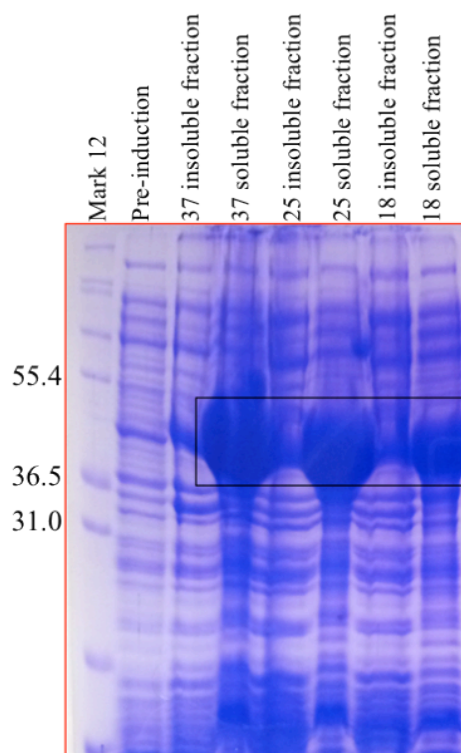


Figure 4.10 SDS gel showing the small-scale over-expression of AgaE. The highlighted boxed bands indicate the expression of AgaE in soluble fractions at three different temperatures 37, 25 and 18°C. The molecular weight of AgaE is 45.5 kDa.

4.3.3 Large scale overexpression of AgaE

For large scale overexpression, 2 l of LB media were prepared using 4 flasks of 500 ml. The primary culture was prepared by the inoculation of one colony of the transformed clone in BL21 (DE3) to 50 ml of the LB media supplemented by 50 μ l of 50 μ gml⁻¹ kanamycin in 250 ml flask and incubated overnight (15-18 hours) at 37°C at 250 RPM. 5 ml of this culture was used to inoculate 500 ml of LB media supplemented with 50 μ gml⁻¹ kanamycin. The secondary culture was incubated at 37°C and 200 RPM for approximately 2 hours to reach an optical density (OD₆₀₀) of 0.6. The cells were induced by the addition of 1mM of IPTG and incubated at 37 °C and 200 RPM for an additional 4 hours. 1ml of non-induced cells was taken for expression analysis. After 4 hours induction, the cells were transferred into 1 l Beckman centrifuge tubes and spun at 3,600 \times g for 20 minutes. Most of the LB media was discarded, the cells resuspended and centrifuged again using Sigma 3-16K centrifuge at 3,600 \times g for 20 minutes and stored at -80°C.

4.3.4 Ni-NTA Affinity Chromatography purification of AgaE

As the construct of AgaE contained a N-terminal 6 \times His tag, Ni-NTA affinity chromatography was used as the initial purification step. 1 g of the cell paste was removed from -80°C and re-suspended on ice using 10 ml of 50 mM Tris buffer pH 8.0. Then, the cells were disrupted on ice by sonication (3 \times -20 seconds) at a volume of 16-micron amplitude. The cell debris was then removed at 4°C by centrifugation at 70,000 \times g for 10 minutes. The soluble protein fractions (~20 ml / 50mg) were applied to the 5 ml Ni-HP cartridge column after measuring the protein concentration using the Bradford assay. The AgaE protein was eluted from the column with 50 ml of 0.35M imidazole. 3 ml fractions were collected and pooled together. The protein was analysed by SDS-PAGE gel, which displayed lower molecular weight contaminants; thus, a further purification step was used.

4.3.5 Gel filtration of AgaE

The fractions from the Ni-NTA affinity chromatography step were measured using the Bradford assay and fractions 13 and 14, which had the highest protein concentration (Figure 4.12) were combined for further purification by gel filtration. The sample was reduced to 1 ml using a Viva spin concentrator (MWCO 30000); it was then injected into a 16 \times 60 Superdex200 gel filtration column (GE Healthcare),

which was equilibrated with 50 mM Tris pH.8.0 and 0.5 M NaCl buffer. The Gel filtration was performed at a flow rate of 1.5 ml/min and 2 ml fractions were collected. Fractions 21-24 were combined based on the recording fraction peak of the protein predicted size. The progress of the purification was analysed by SDS-PAGE using 12% Bis-Tris NOVAX gel (Invitrogen) (Figure 4.11).

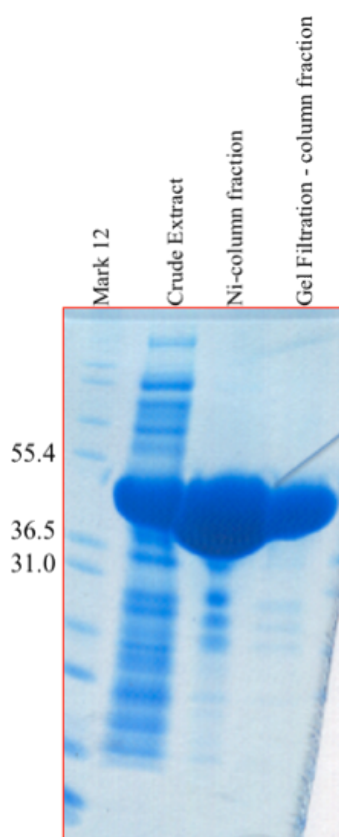


Figure 4.11 SDS gel viewing the protein purification steps of *Msmeg_0515* (AgaE). The molecular weight of AgaE is 45.5 (kDa). Samples were taken from the crude extract, Ni-column and gel filtration respectively to be analyzed.

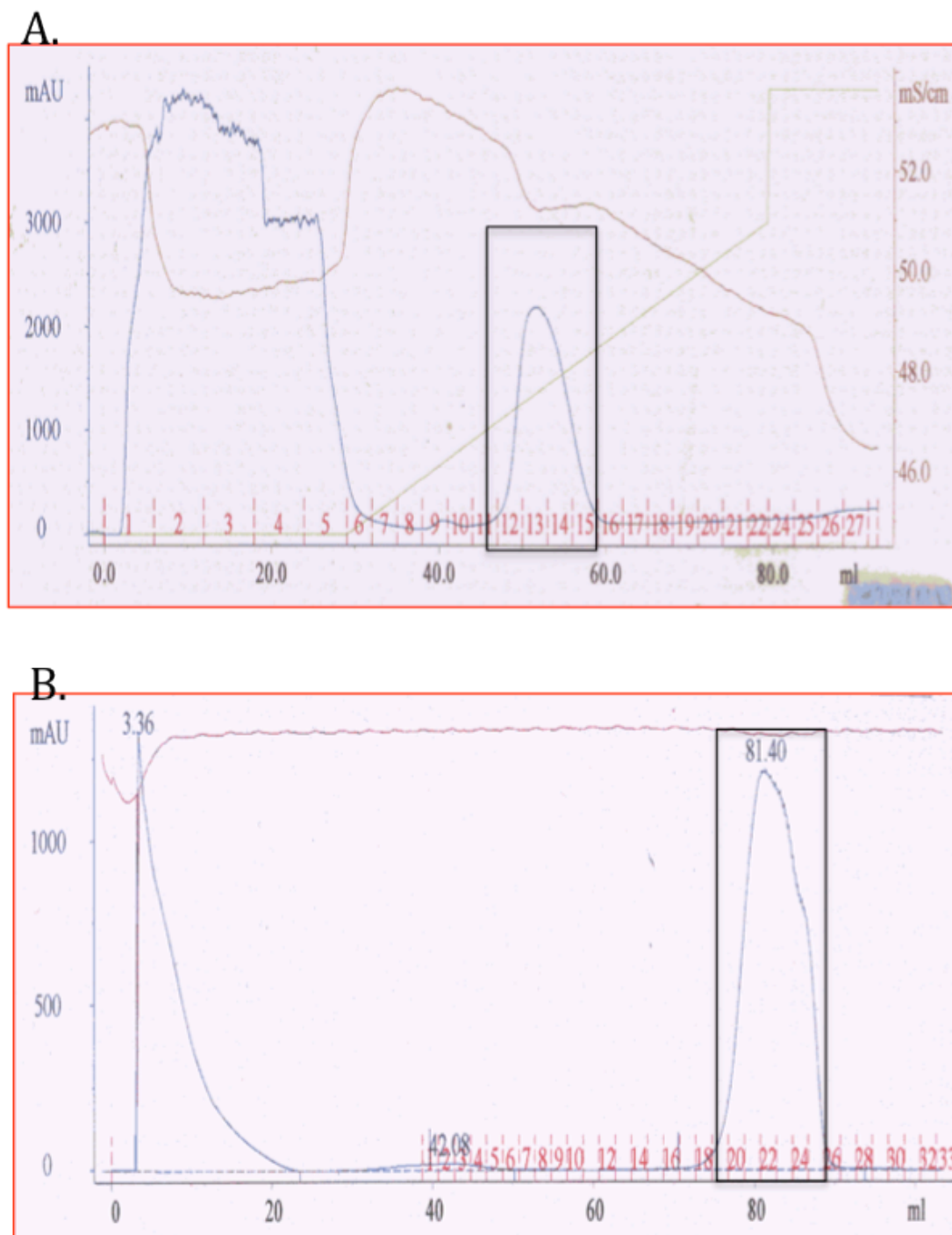


Figure 4.12 Chromatograms obtained during the purification of AgaE. Blue line represents absorption at 280nm, brown line represents conductivity and green line represents gradient of Imidazole concentration: **a.** Chromatography on HisTrap HP column; **b.** Gel filtration on HiLoad Superdex200 column. The gel filtration column step succeeded in removing low molecular weight contaminants left after the initial Ni-NTA purification.

4.4 Initial automated crystallization screening of truncated AgaE

4.4.1 Preparation of the protein sample

To obtain protein for crystallization experiments, fractions 21-24 from the gel filtration were pooled and concentrated to 11.5 mgml⁻¹ using a Viva spin concentrator. The purification buffer was exchanged to one with low salt by adding 7ml of 10mM Tris buffer pH 8.0 to the 400 µl of protein solution and centrifuging again for 20-30 minutes at 4500 rpm.

4.4.2 Crystallization of AgaE

Sitting drop vapour diffusion trials were carried out in order to identify an optimal condition for the protein to be crystallized. The Hydra II crystallization robot system was used together with several commercial screen conditions, such as JCSG+, PACT, PEG and classic (Qiagen Nextal®) using 96 well MRC plates a drop sized 200 nl and reservoir size of 200 nl. The plates were sealed using a special sheet (Molecular Dimension) in order to prevent the drop from being dried and to view it clearly. All the plates were centrifuged for 2 minutes and at a speed of 2000 rpm to make sure that the two drops of buffer and protein were mixed well using a Grant-Bio LMC-3000 R-2 rotor. Finally, all the plates were placed in a crystallization room at 17°C.

4.4.3 Identifying of successful crystallization

The crystallization plates were viewed after 24 hours, 3 days and 7 days for crystal growth. After 3 days of incubation, the recombinant AgaE protein started to form crystals under different conditions of the JCSG screen (A9, B9, H6, 7,8). The trays were left to equilibrate further, and the crystal grew bigger. In addition, after 1 month, crystals were also observed in a number of further conditions in the JCSG, PACT and PEG screens (Figure 4.13).

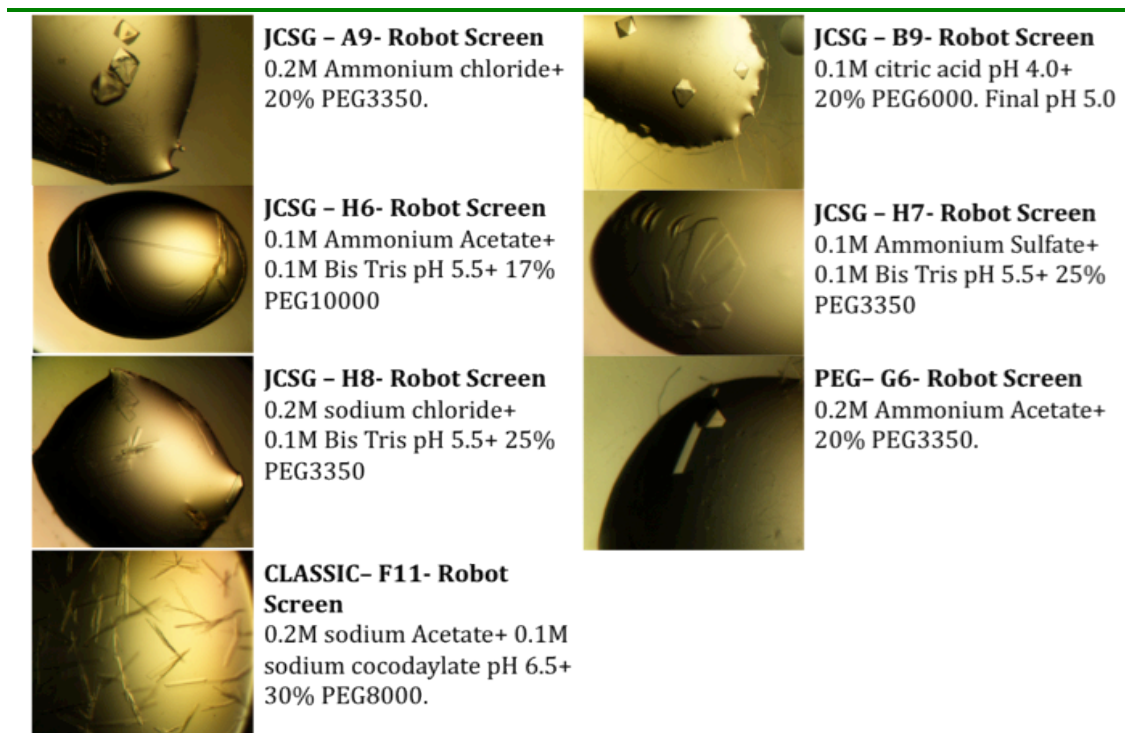


Figure 4.13 Photographs of AggE native crystals; the crystals were grown within different conditions and with different crystal forms.

4.4.4 Crystal diffraction test

Crystals were selected for data collection according to their size and quality. Single octahedral shaped crystals were obtained from wells A9 and B9 of the JCSG screen with dimensions 0.3mm×0.2mm×0.2mm. These crystallization conditions contained 0.2 M Ammonium chloride pH 6.3, 20% PEG 3350 and 0.1M citric acid pH5.0, 20% PEG600, respectively.

In order to test the diffraction, crystals were washed with a cryoprotectant solution (crystallization buffer and 25% ethylene glycol), mounted onto the diffractometer using a fiber loop and flash cooled to 100K with an Oxford Cryo-systems Cryostream 700.

The X-ray generator of Sheffield University (Rigaku MM007 copper rotating anode generator and MAR345 Research image plate) was used for initial X-ray diffraction experiments. Two 1° rotation images at 90° to each other and 5 minutes exposure were obtained. The resulting diffraction was auto-indexed using the Mosflm software (Leslie & Powell, 2007). Using this procedure, the crystal of AgaE was determined to be in the orthorhombic class, in point group P222 with cell dimensions of $a=63.88$, $b=69.15$, $c=100.71$ Å and $\alpha=\beta=\gamma=90^\circ$. Data were collected 2.14Å and processed using the Mosflm software (Leslie & Powell, 2007)[79]. As only the even $h00$ $0k0$ and $00l$ reflections were present; the space group was assigned to $P2_12_12_1$ (Table 4.4). All crystals were then saved in liquid Nitrogen to be sent to the Diamond synchrotron for high resolution data collection.

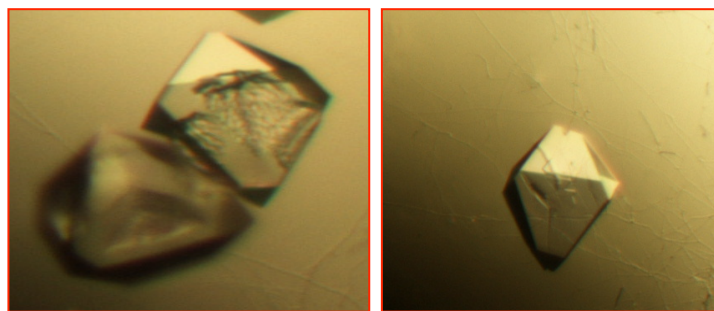


Figure 4.14 Photographs of selected native AgaE crystals for data collection. Crystal (1-left) was taken from JCSG – A9 and diffracted to 1.49 Å. Crystal (2) was taken from JCSG-B9 and diffracted to 1.35Å.

4.4.5 X-ray data collection of AgaE crystals

In order to collect a high-resolution data, several native AgaE crystals were selected from the robot trials of the JCSG screen wells A9, B9 and H12. All the crystals were mounted and washed with a cryoprotectant solution made of a crystallization condition buffer with the addition of 25% ethylene glycol and flash cooled to 100 K with a gaseous nitrogen stream. The crystals were then stored in liquid nitrogen and sent to beam line I03 at the Oxford Diamond light source for data collection. Two diffraction images 90° apart and 1° oscillation were collected using a ADSC Q315r detector and auto-indexed using the Mosflm software to predict the best strategy for data collection [79]. Data collection strategies and statistics for all the three native crystals are shown in Tables 4.3 and 4.4 respectively.

Data strategies	Crystal I	Crystal II	Crystal III
Detector	MAR345 image plate	ADSC Q315r	ADSC Q315r
Phi start	273.0°	335.0°	335.0°
Phi Oscillation	1.0°	0.5°	0.2°
No of images	140	230	900
Resolution	2.14 Å	1.48 Å	1.35 Å

Table 4.3 The data collection strategies of several crystals of native AgaE

DATA SET	Crystal I Native	Crystal II Native	Crystal III Native
Wavelength (Å)	0.97630	0.96860	0.97620
Energy (KeV)	12.7	12.8	12.7
Space group	P 2 ₁ 2 ₁ 2 ₁	P 2 ₁ 2 ₁ 2 ₁	P 2 ₁ 2 ₁ 2 ₁
Unit cell parameters			
a (Å)	63.88	64.06	64.03
b (Å)	69.15	69.26	69.22
c (Å)	101.71	100.74	100.47
$\alpha = \beta = \gamma (^{\circ}) =$	90.000	90.000	90.000
Resolution range (Å)	53.94-2.14 (2.03-2.14)	34.6-1.48 (1.52-1.48)	64.0-1.35 (1.39-1.35)
Unique observation	28899 (3694)	75131 (5493)	91428 (4448)
R_{merge}	0.103 (0.403)	0.047 (0.586)	0.097 (0.382)
R_{pim}	0.059 (0.235)	0.029 (0.352)	0.048 (0.43)
Completeness (%)	98.1 (87.4)	99.7 (99.9)	92.8 (62.1)
Anomalous completeness (%)	95.7 (84.1)	97.1 (99.3)	82.5 (38.7)
Multiplicity	3.9 (3.8)	4.6 (4.6)	5.3 (2.3)
Anomalous multiplicity	2.0 (2.0)	2.3 (2.3)	2.6 (0.9)
Mean (I)/ σ (I)	8.4 (3.1)	15.8 (2.3)	11.0 (2.0)

Table 4.4 Data statistic of three native AgaE crystals.

Matthews	Crystal I	Crystal II	Crystal III
Molecules in the AU	1	1	1
Probability (based on data resolution)	1.00	1.00	1.00
Probability (all proteins in the PDB)	1.00	1.00	1.00
V _m (Å ³ / Da)	2.4	2.5	2.5
Solvent content (%)	49.7	49.8	50.4
Molecular weight (Da)	45500	42.734	45500

Table 4.5 Asymmetric unit contents and Matthews coefficient for AgaE crystals. The results revealed one molecule only per asymmetric unit.

4.4.6 Native data processing of AgaE crystals

The native data from crystals I and II were processed automatically using the Xia2 pipeline implemented at Diamond light source [81]. The processed data used in the structure solution of AgaE, came from the 3D mode of Xia2; XDS and XSCALE were used to process and merge data, respectively [80, 132]. The space group was checked using Pointless [82], which revealed that an orthorhombic $P2_12_12_1$ system, with cell dimensions of $a = 64.0 \text{ \AA}$, $b = 69.0 \text{ \AA}$, $c = 100.0 \text{ \AA}$, and all angles $\alpha=\beta=\gamma=90^\circ$ (Table 4.4) [79]. The asymmetric unit contents were estimated using the method of Matthews's [78], which gave a V_m of ~ 2.5 for one AgaE in the A.U. (Table 4.5).

4.4.7 Structure determination

4.4.7.1 Molecular Replacement:

As the sequence comparisons of AgaE revealed that the most likely structure would be of a maltose- maltodextrin binding protein, attempts were made to determine the AgaE truncated structure by molecular replacement using a search model of a maltose binding protein from the thermoacidophilic bacterium *Alicyclobacillus acidocaldarius* (PDB code, 1URD). The search model had a sequence identity of 27% to AgaE and it was predicted to share the same 3D structure as AgaE by the Phyre server [89]. The Chainsaw program of the Phaser was used to replace all the amino acids of the search model by a poly alanine chain [133]. Phaser was used to search for one copy of the molecule using the data of AgaE as an input [83]. Unfortunately, no convincing solution to the molecular replacement could be obtained.

The resulting map of the best solution was poor and discontinuous; additionally, no improvement was seen after the refinement with an R-factor in Refmac5 of 0.52 [134]. As there is some variation in the position of the two domains of this family of proteins, depending upon substrate binding, an alternative MR strategy was employed. The search model was split into two domains and attempts were made using the Phaser to find solutions for the individual domains. Again, no convincing solution could be found. It was thus decided to try to express the AgaE protein as a Seleno-methionine derivative, in order to obtain initial phases by exploiting the anomalous diffraction of the selenium atoms.

4.5 Expression and purification of AgaE incorporated with Seleno-methionine

The AgaE protein containing Seleno-L-methionine was overexpressed using the same protocol as that for the native protein over-expression. However, after reaching the suitable optical density of $OD_{600} \sim 0.6$, the cultures were harvested by centrifugation at 5.000 xg for 30 minutes. The LB media was discarded and the cells were re-suspended in minimal media supplemented with Seleno-methionine as detailed in (section 2.2.5). The expression was induced using 1mM IPTG for a further 8 hours. Using this protocol, 1g of Se-Met AgaE cell paste was obtained. The purification followed that for the native protein, with a Ni-NTA column followed by

gel filtration and buffer exchange to give a solution of Se-Met AgaE at 11 mg/ml in 10mM Tris pH8.0.

4.6 Crystallization of Seleno-methionine incorporated AgaE

Several robot screens were used in order to identify suitable condition for Se-Met AgaE protein to be crystallized. The JCSG, PACT, PEG and classic screens were used with purified AgaE at 11.5 mg/ml in 10mM Tris pH8.0. 200 nl of the well solution were mixed with 200nl of the protein for each well. All trays were then incubated as the native protein at 17°C. Crystals were obtained under numerous conditions in the JCSG, PACT and classic screens. However, only a single crystal from JCSG – B2 (0.1M citric acid pH 4.0 and 20% PEG6000), was sent to the Diamond synchrotron for data collection and phase determination experiment.

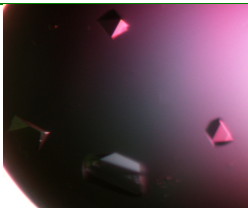
Crystal	Condition
	JCSG – B9- Robot Screen 0.1M citric acid pH 4.0 20% PEG6000.

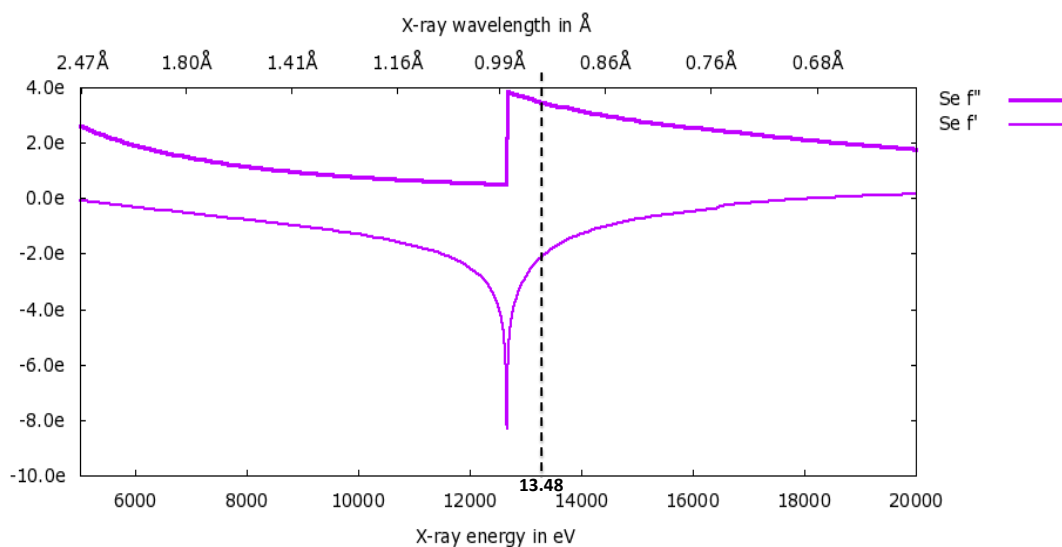
Figure 4.15 Photograph of AgaE Se-MET crystals. Crystals were found under the same conditions as native crystal.

4.6.1 X-ray data collection of Seleno-Methionine AgaE crystals

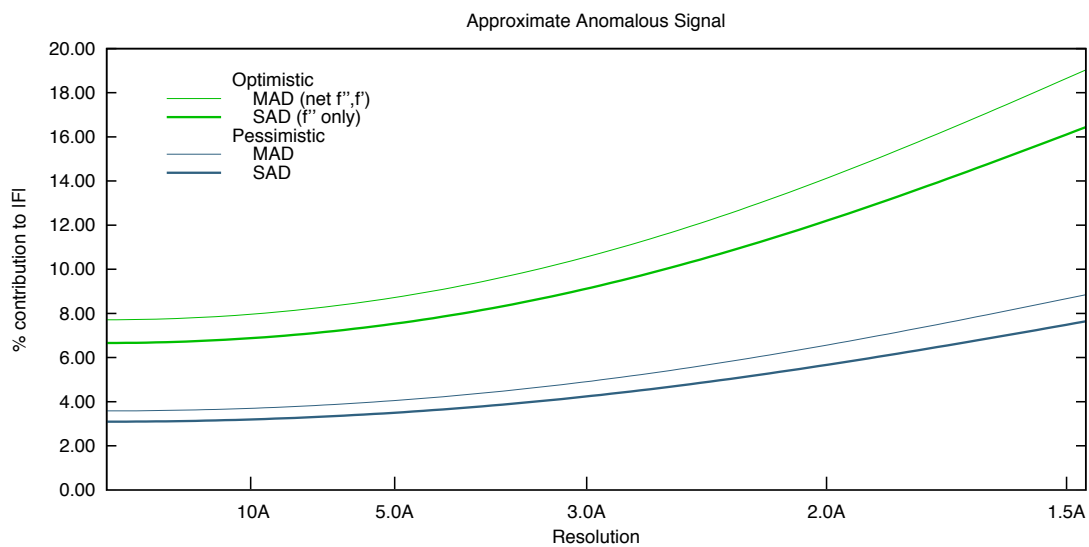
As crystals of the Se-MET protein grew under the same conditions as those seen for the native protein, single crystal was selected from JCSG-B9 for data collection. The crystal was mounted and saved in liquid nitrogen for data collection. The I04-1 beam line of the Diamond light source was used to collect data using a SAD experiment.

The theoretical Se-K edge absorption spectrum is shown in figure 4.16-a. It can be seen that at energies above 12.66 KeV, a significant anomalous diffraction signal should be present in a protein crystal containing selenium, however, the exact extent of the contribution of the selenium anomalous diffraction depends on many factors including the number of selenium atoms compared to the other atoms in the unit cell, the quality of diffraction and the precise X-ray energy chosen. The amino acid sequence of AgaE contains ten methionine residues. The predicted signal was estimated for AgaE using the BMSC web server (Figure 4.16-b). It can be seen that at $E=13.48$ eV (the energy of beamline I04-1), the size of the anomalous signal should be approximately 4-10% at 2\AA .

Three initial images of 0.5° rotation and at 45° to each other were taken on beamline I04-1 to test the diffraction and crystal quality and to estimate a good strategy for data collection. These images showed that the crystal diffracted well (Figure 4.17), and so single wavelength (SAD) data at energy of 13.48 KeV (0.91997\AA) were collected on this crystal. To ensure that good quality anomalous data were collected, a total of 7200 images of 0.2° rotation were collected, so that each reflection was measured an average 30 times. The data was processed using the Xia2 Diamond system in its 3dii mode. Data were indexed in space group $P2_12_12_1$ with similar unit cell dimensions to the native crystals (Table 4.4). Due to rotation damage of the crystal only the first 4268 images were processed, however this still gave an anomalous multiplicity of 15.8 (Table 4.6).



(a)



(b)

Figure 4.16 The Se-K edge absorption spectrum of AgaE SAD experiment. a) The Se-K edge absorption spectrum. b) The predicted anomalous signal for AgaE. Figures were created using the BMSC web server.

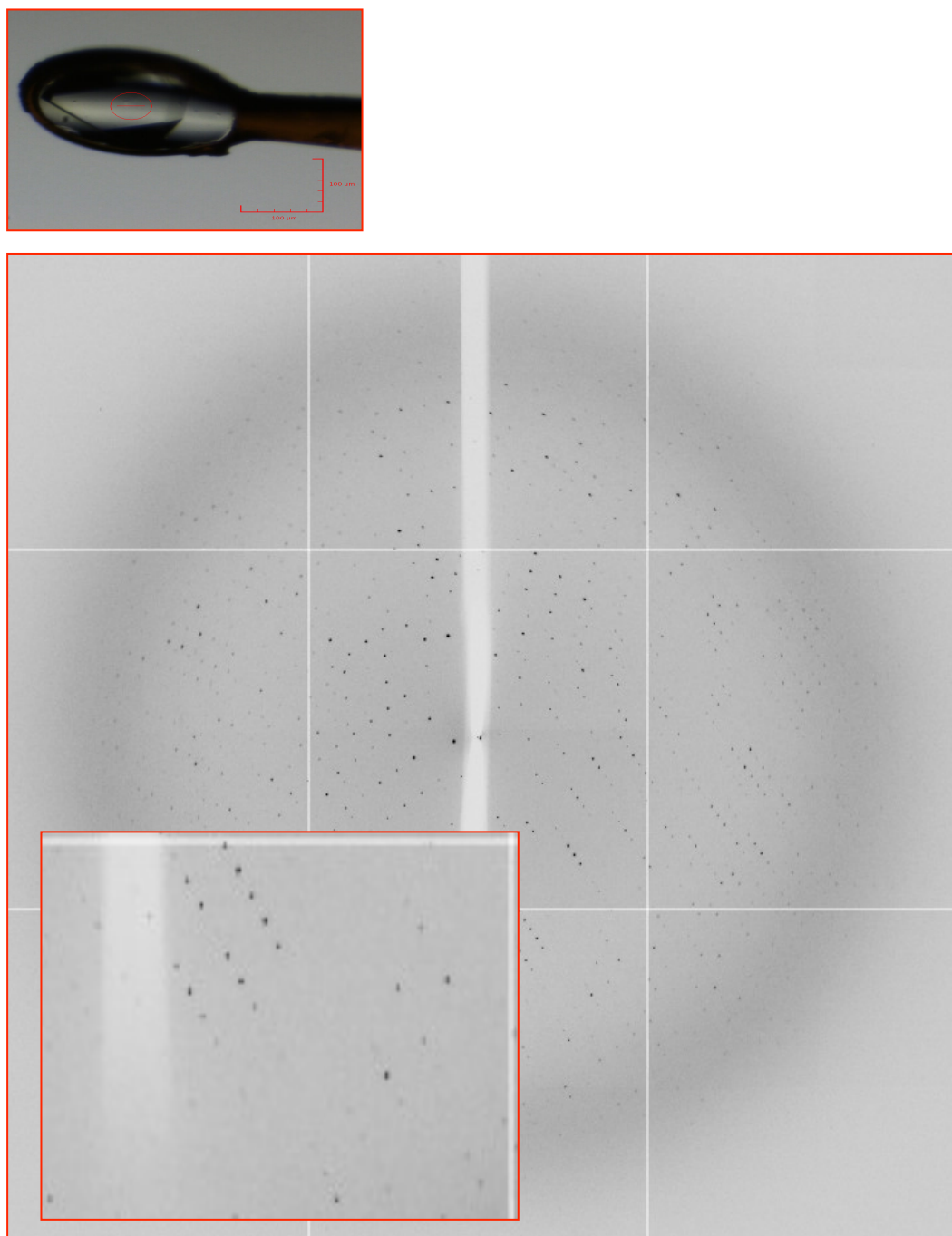


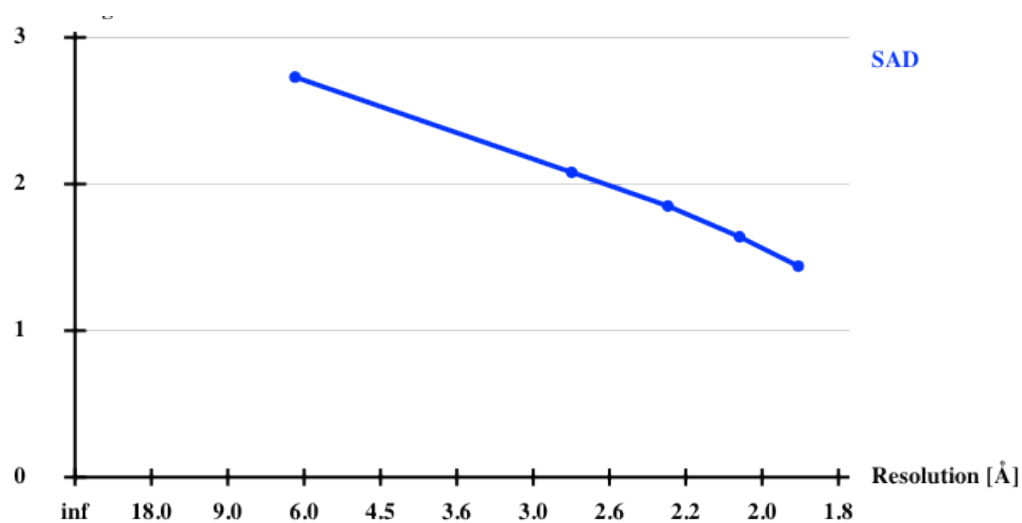
Figure 4.17 Example of data collection for the Se-MET AgaE crystal. A. Crystal was mounted onto a fiber loop and exposed to 100 μm X-ray beam. B. Diffraction image of the AgaE crystal; diffraction data were collected to a resolution of 1.77 \AA for of Se-MET crystals using ADSC Q315 CCD detector at the beam line I03 of the Diamond synchrotron source.

DATA SET	Se-MET Crystal
Energy (KeV)	13.48
Wavelength (Å)	0.91997
Space group	P 2 ₁ 2 ₁ 2 ₁
Unit cell parameters	
a (Å)	63.9
b (Å)	69.6
c (Å)	101.4
$\alpha = \beta = \gamma$ (°) =	90.0
Resolution range (Å)	(57.4-1.77) (1.82-1.77)
Observations	1348574 (86932)
Unique reflections	44560 (3206)
Rmerge	0.086 (0.877)
Rpim	0.016 (0.167)
Completeness (%)	99.4 (97.6)
Anomalous completeness (%)	99.4 (97.5)
Multiplicity	30.3 (27.1)
Anomalous multiplicity	15.8 (13.8)
Mean (I)/ σ (I)	30.5 (5.6)

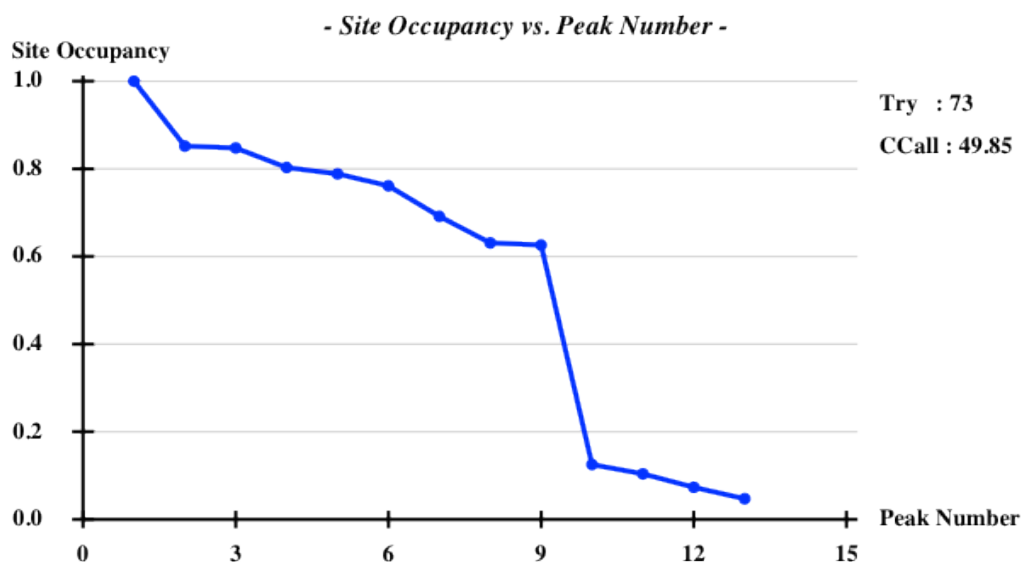
Table 4.6 Data collection statistics of the Se-Met AgaE crystal data set (values in brackets refer to the high resolution shell).

4.6.2 Obtaining experimental phase for AgaE S-MET data

In order to solve the phase problem, the three SHELX programs C, D and E were used in the graphical user interface of HKL2MAP [84, 85]. In the first step SHELXC was used to prepare the data from the scaled, unmerged file from Xia2 and to calculate the extent of the anomalous signal. This shows a strong anomalous signal with $\text{Dano} / \text{sig} (\Delta_{\text{ano}}) > 1.2$ to 1.9\AA and a correlation coefficient of $>30\%$ to 2.2\AA (Tables 4.7a). Then, SHELXD was used to calculate the positions of the 10 potential selenium atoms in the structure, with 100 tries and to a maximum resolution of 2.3\AA . The SHELXD results (Figure 4.18), gave the position of 9 Se atoms with an occupancy > 0.6 , with a steep fall off in occupancy to 0.1 for the next site. There was a high correlation of 49.8 % between the observed and calculated patterns. It thus seems likely that the crystals of AgaE contained nine selenium atoms, in good agreement with the 10 Met residues in the sequence, assuming the N-terminal methionine had been cleaved during experiment. The protein phases were then calculated for both hands of the selenium substructure, using SHELXE. 20 cycles of structure phasing and density refinement using a 50% fractional solvent content were run. It can be seen that the phases calculated from the inverted hand enantiomorph of the substrate gave substantially better mean figure of merit and correlation coefficient and the resulting map had a much higher connectivity and contrast than the map from the original hand substructure (Table 4.7b). The two SHELXE maps were inspected in coot to check for continuous good connectivity displaying a clear map for the protein structure (Figure 4.19). This clearly showed that the inverted hand map was correct.

Dano / sig (Δ_{ano}) vs. Resolution

(a)



(b)

Figure 4.18 The output results of the SHELX C & D of AgaE SAD experiment. **a.** Shows the anomalous signal from the AgaE S_MET crystal datasets. A graph of Dano / sig (Δ_{ano}) was plotted against the resolution which indicated a value of above 1.3 with a good anomalous signal. **b.** Peaks show the best solution of 9 sites of heavy atoms with occupancies between 0.6 and 1.0 [84].

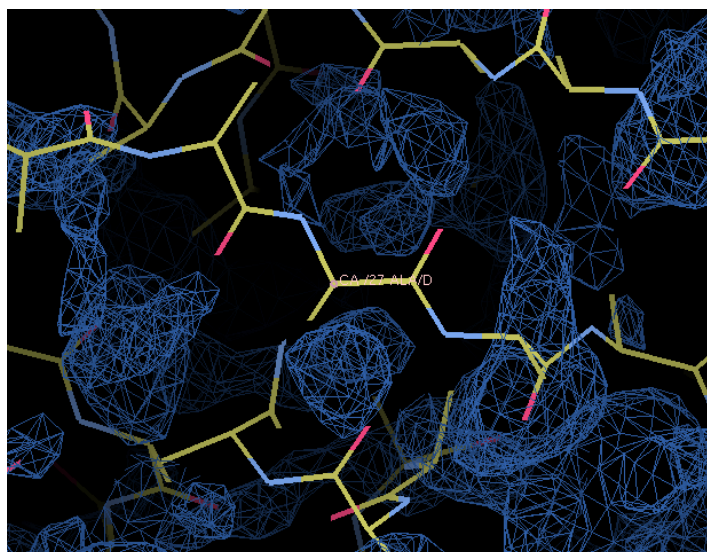
Resl.	Inf	8.0	6.0	5.0	4.0	3.5	3.0	2.6	2.4	2.2	2.0	1.77
N(data)	560	695	866	1952	1942	3421	4929	3809	5290	7603	13433	
Chi-sq	0.65	0.80	1.01	0.98	1.16	1.07	0.98	0.97	1.10	1.19	1.36	
<I/sig>	70.1	67.0	66.9	73.7	71.0	56.9	43.5	33.7	26.6	18.8	9.0	
%Complete	97.6	96.0	98.2	99.4	99.8	99.1	99.6	99.6	99.3	99.5	98.6	
<d"/sig>	4.16	3.52	2.96	2.27	2.23	2.00	1.81	1.57	1.43	1.21	1.02	
CC(anom)	94.1	86.1	82.9	76.3	73.5	64.2	56.8	45.4	35.7	20.8	6.7	

(a)

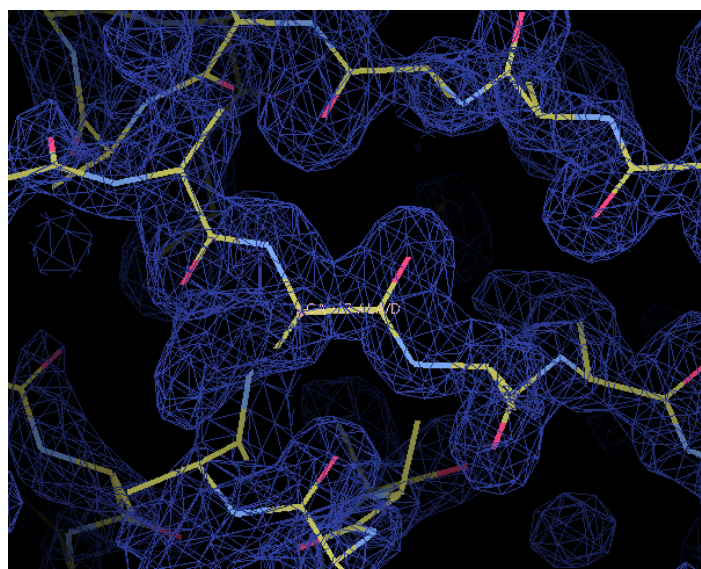
Hand	Original	Inverted
Contrast	0.22	0.7
Connectivity	0.6	0.8
Mean FOM	0.31	0.7
Correlation coefficient	34.7	72.0

(b)

Tables 4.7 The output result of SHELXE calculating of heavy atoms density and pseudo-free CC for the enantiomorphs of the determined electron density map. The original enantiomorph (Top) and the inverted enantiomorph (Bottom).



(a)



(b)

Figure 4.19 Electron density model of the substructure atoms of both original and inverted hands enantiomorphs of AgaE selenium SAD experiment. Maps are contoured at 1.6σ and viewing the same regions. **a)** Shows the original substructure hand from SHELXE. **b)** Shows the inverted substructure hand.

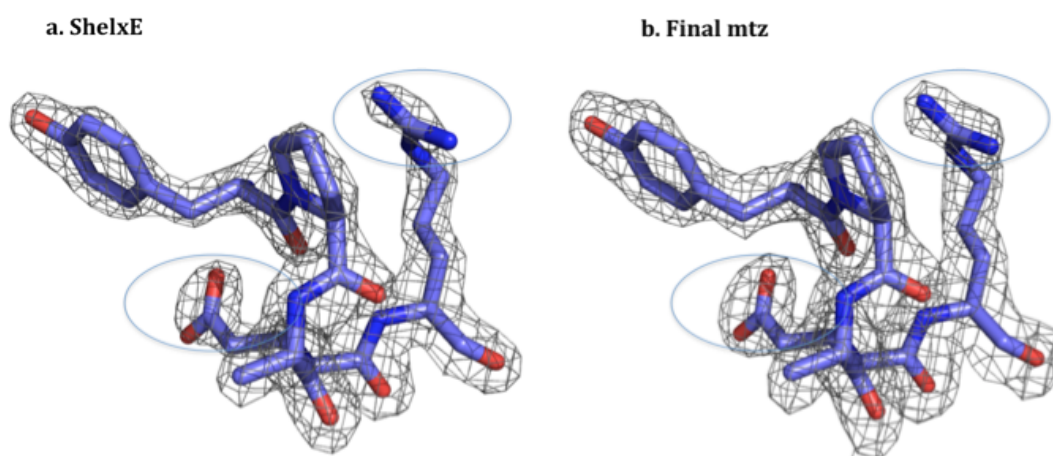


Figure 4.20 An example of the progress made with respect to the model building AgaE structure by electron density. Maps are shown for the same residues Tyr124, Pro125, Ala126, Glu127 and Arg128 for the same model after SHELXE and for the final mtz, both are contoured at sigma level of 1.5. a. The electron density of AgaE after SHELXE and b. the electron density of AgaE for the final mtz model [84]. Differences in electron density quality are highlighted by blue rings.

4.7 Structure refinement

4.7.1 Structure building and refinement of AgaE

SHELXE was run again on the inverted hand including autobuild of the main chain of the protein. The polyalanine model produced had 347 out of the expected 398 residues, built in 8 chains. The model was completed in coot by manually adding side chains and missing residues using the position of the Selenium atoms as a guide.

4.7.2 Phasing of native data by molecular replacement

The complete refined Se-Met AgaE structure was used as a search model in molecular replacement to determine the structure of the three sulphur methionine crystals, for which data had been collected (Table 4.4). In each case clear solutions for the rotation and translation functions could be seen, with each structure built using rounds of refinement in Refmac5 and rebuilding in Coot [87, 134]. Refinement statistics for only the highest resolution structure was shown in table 4.8.

4.7.3 The final model of (agaE)

The structure was built and refined further using Refmac5 in CCP4 with an initial R-factor of 0.35 and R-free 0.35 [134]. 14 cycles of refinement were carried out in order to reduce the R-factor for best structure using coot for structure evaluation [87]. The final structure model has been built to an R-factor of 0.20 and R-free 0.23. Waters were added to the model with selected sigma level of 1.0; however, all waters were later checked manually and unsuitable water with poor density was removed.

There was very little difference between the native structures, apart from the resolution. Each model of AgaE has 398 residues with well-defined electron density; however the two N-terminal residues (28 and 29) in crystal II have poor density. The final models have an RMSD deviation of bond lengths and angles of about 0.02 Å and 1.9°, respectively (Table 4.8). The water molecules for all crystal structures were added to the models with discrete electron density over 3.4 sigma in the difference maps and checked that they made reasonable hydrogen bonds with the protein atoms. Finally, the models were validated using the programs PROCHECK and MOLPROBITY [135, 136]. The final model for crystal III at 1.35Å had 277 water molecules, one 6 carbon PEG molecule and 3 ethylene glycol molecules. This is the structure that is described in the following chapter.

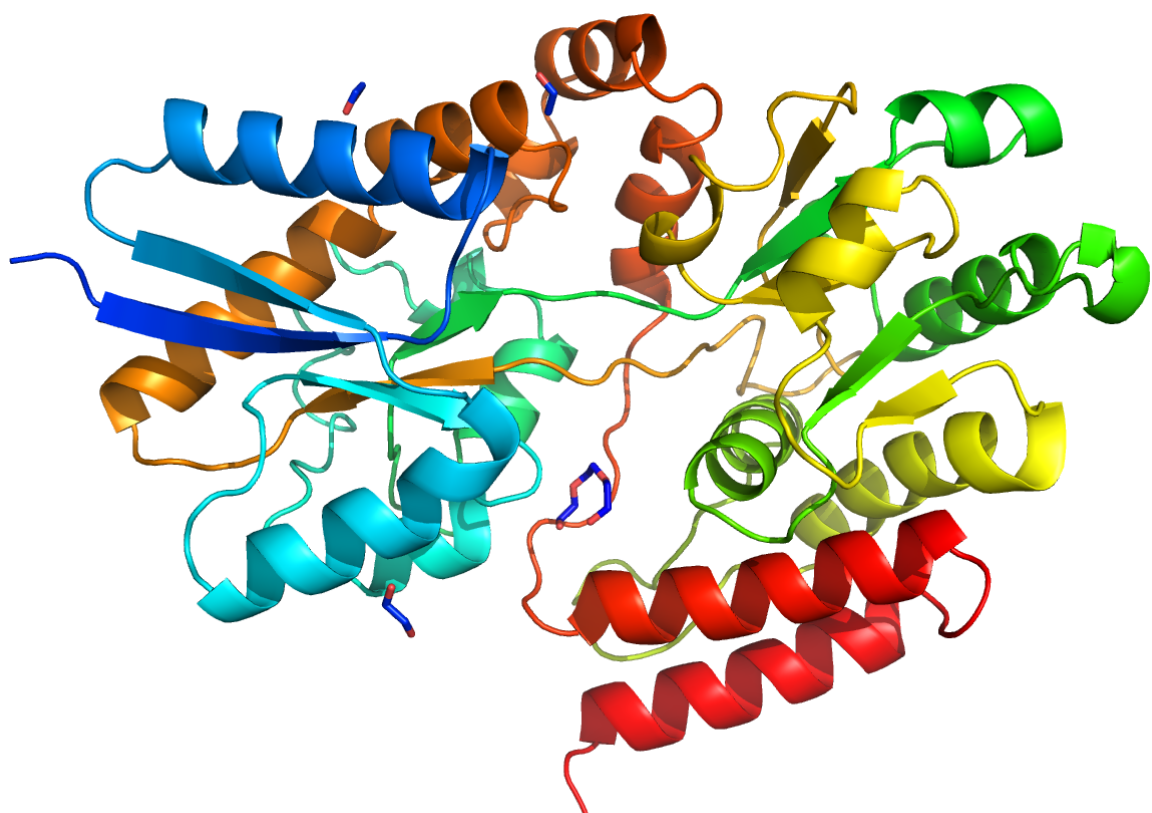
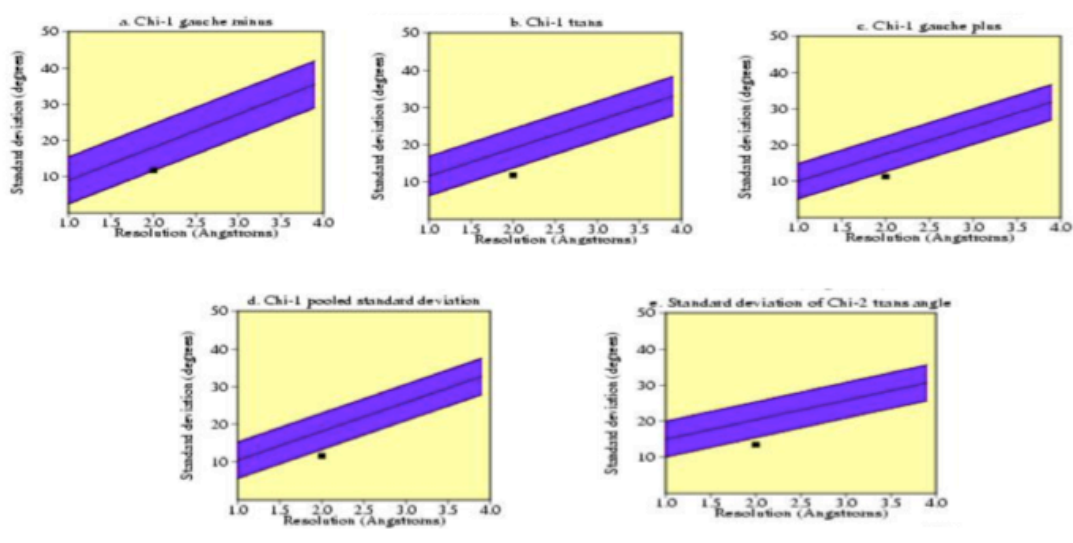


Figure 4.21 The overall fold of the final structure model of AgaE. The position of PEG and ethylene glycol are shown as blue sticks. Cartoon image was made using Pymol [98].

Model	Crystal III	Crystal S-MET
Resolution (Å)	1.35Å	1.77Å
Number of reflections	86449	42260
Protein molecules per asymmetric unit	1	1
Number of atoms	3143	3296
Number of waters	277	187
Number of EDO	3	7
Number of PEG	1	1
Ramachandran favoured (%)	98.0	97.3
Ramachandran outliers (%)	0	0.25
Poor rotamers (%)	0.94	1.84
RMSD bond (Å)	0.02	0.012
RMSD angle (°)	1.9	1.3
Average B-factors (Å²)		
Main chain (Å²)	18	29
Side chain (Å²)	22	30
Waters	31	51
EDO	25	36
PEG	32	32
R-factor	0.16	0.18
R-Free	0.20	0.24
Molprobit score	1.1	1.9
	98 th percentile*	90 th percentile

Table 4.8 The final refinement statistics and validation for the different native crystal structures of AgaE as well as S_MET incorporated in the crystal structure.

a. Side-chain parameters



b. Main-chain parameters

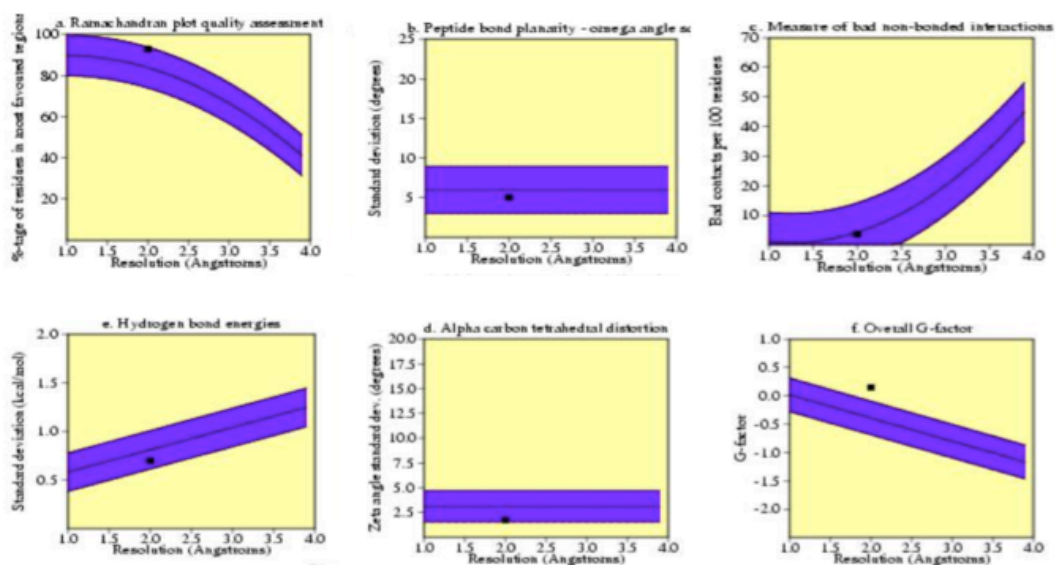
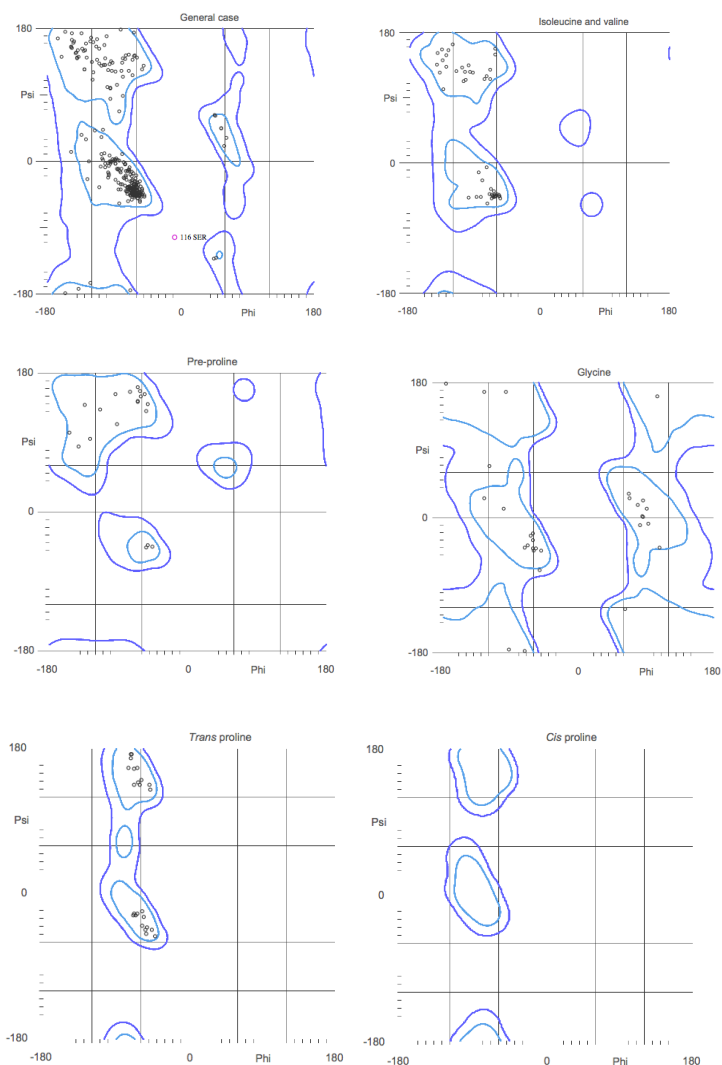


Figure 4.22 The properties of the main chain and the side chain for the final refined **AgaE** structure. All residues were shown to be within the expected site. The figure was produced using PROCHECK program [136].



All-Atom Contacts	Clashscore, all atoms:	4.73	97 th percentile* (N=841, 1.77Å ± 0.25Å)	
	Clashscore is the number of serious steric overlaps (> 0.4 Å) per 1000 atoms.			
Protein Geometry	Poor rotamers	6	1.84%	Goal: <1%
	Ramachandran outliers	1	0.25%	Goal: <0.05%
	Ramachandran favored	392	97.27%	Goal: >98%
	MolProbity score [^]	1.58	90 th percentile* (N=11232, 1.77Å ± 0.25Å)	
	Cβ deviations >0.25Å	0	0.00%	Goal: 0

(a)

All-Atom Contacts	Clashscore, all atoms:	3.01	98 th percentile* (N=466, 1.35Å ± 0.25Å)	
	Clashscore is the number of serious steric overlaps (> 0.4 Å) per 1000 atoms.			
Protein Geometry	Poor rotamers	3	0.94%	Goal: <1%
	Ramachandran outliers	0	0.00%	Goal: <0.05%
	Ramachandran favored	387	97.97%	Goal: >98%
	MolProbity score [^]	1.10	98 th percentile* (N=3057, 1.35Å ± 0.25Å)	
	Cβ deviations >0.25Å	1	0.27%	Goal: 0

Figure 4.23 Results of Molprobity and Ramachandran plot of the final native and Se-met AgaE structures. The results indicated excellent score of Molprobity and all residues are within the favored regions. The figures were produced using Molprobity server [135].

4.7.4 Alternative conformation of residues in the AgaE structure

The Se-MET structure of AgaE has a number of residues that were built with two conformations, as there was clear electron density (at a contour level of 1.5σ) indicating this. The residues Leu-66, Trp-190 and Asp-119 were built to have two conformations, each of 50% occupancy (Figure 4.24).

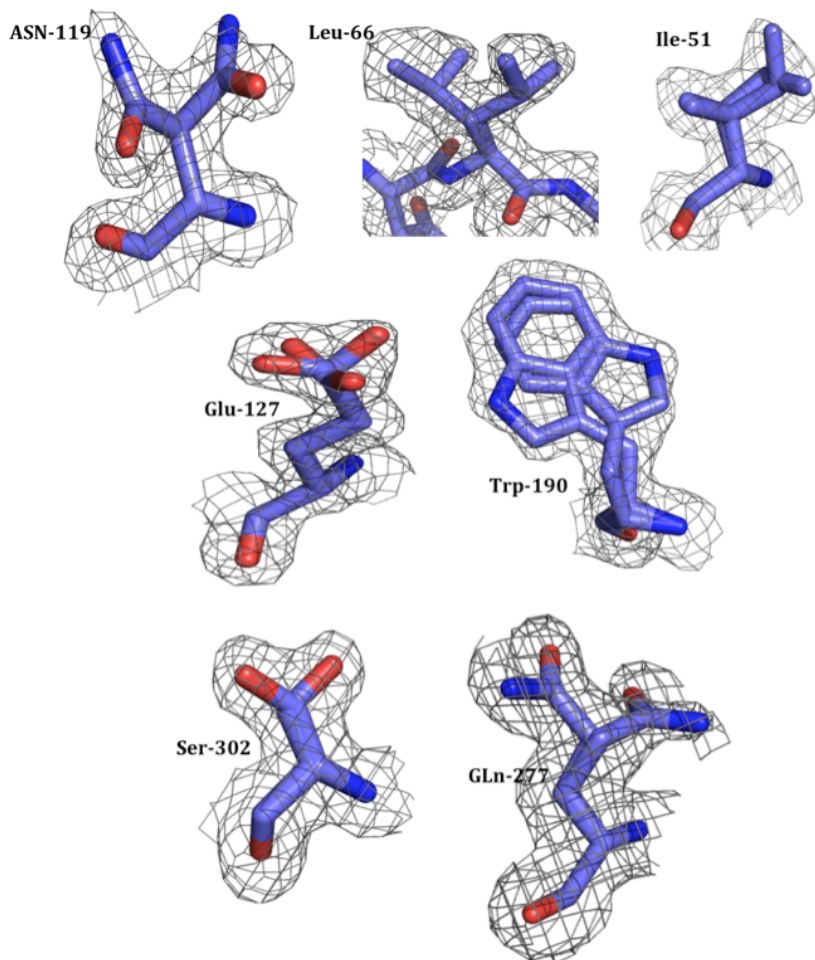


Figure 4.24 An alternative conformation of some AgaE residues. The Map was contoured at 1.5σ .

Chapter 5

Structure of AgaE

In this chapter, a description of the truncated AgaE protein structure will be detailed, together with a comparison to other similar structures. Also, the analysis will include a discussion of the putative active site, and the protein possible function.

5.1 AgaE structure description

5.1.1 Detailed features of AgaE Structure

The AgaE Crystal structure (residues 28-425) was solved in space group $P2_12_12_1$ with one AgaE molecule in the asymmetric unit and with electron density present for all residues. The overall model structure of the truncated AgaE revealed that AgaE folds into two separate domains called as the N and C terminal domains.

A schematic of the folding pattern observed in AgaE is shown in Figure 5.1. It can be seen that the N-terminal domain (residues 28-143 and 305-375) is constructed from a central five stranded β -sheet, surrounded by α -helices. Four of the β -strands are folded in the contiguous N-terminal domain, with the fifth strand interdigitating into the sheet after the residues in the C-terminal domain. This last strand runs antiparallel to the others, and a mixed parallel and antiparallel β -sheet is formed.

In a similar fashion, the central β -sheet in the C-terminal domain (residues 148-286 & 385-425) has one strand running in an opposite direction to the others, and this is the first strand after the linking loop (6 residues long) from the N-terminal domain. This arrangement again gives a mixed parallel and antiparallel β -sheet, with the same architecture as the sheet in the N-terminal domain. The strands in both domains are linked by α -helices and often more than one helix is present between strands.

The connectivity between the two domains is rather complicated. From the N-terminus, four β -strands and associated linking α -helices are folded in domain 1. An extended loop connects domain 1 with domain 2. The β -sheet of domain 2 and its linking helices is folded and then another extended loop folds the polypeptide chain back into domain 1, to form the extra strand of the sheet in this domain. The chain then folds five α -helices (α -13 to α -17), before a third loop runs back into domain 2, to finally fold two further α -helices (α -18 & α -19), and then to the C-terminus.

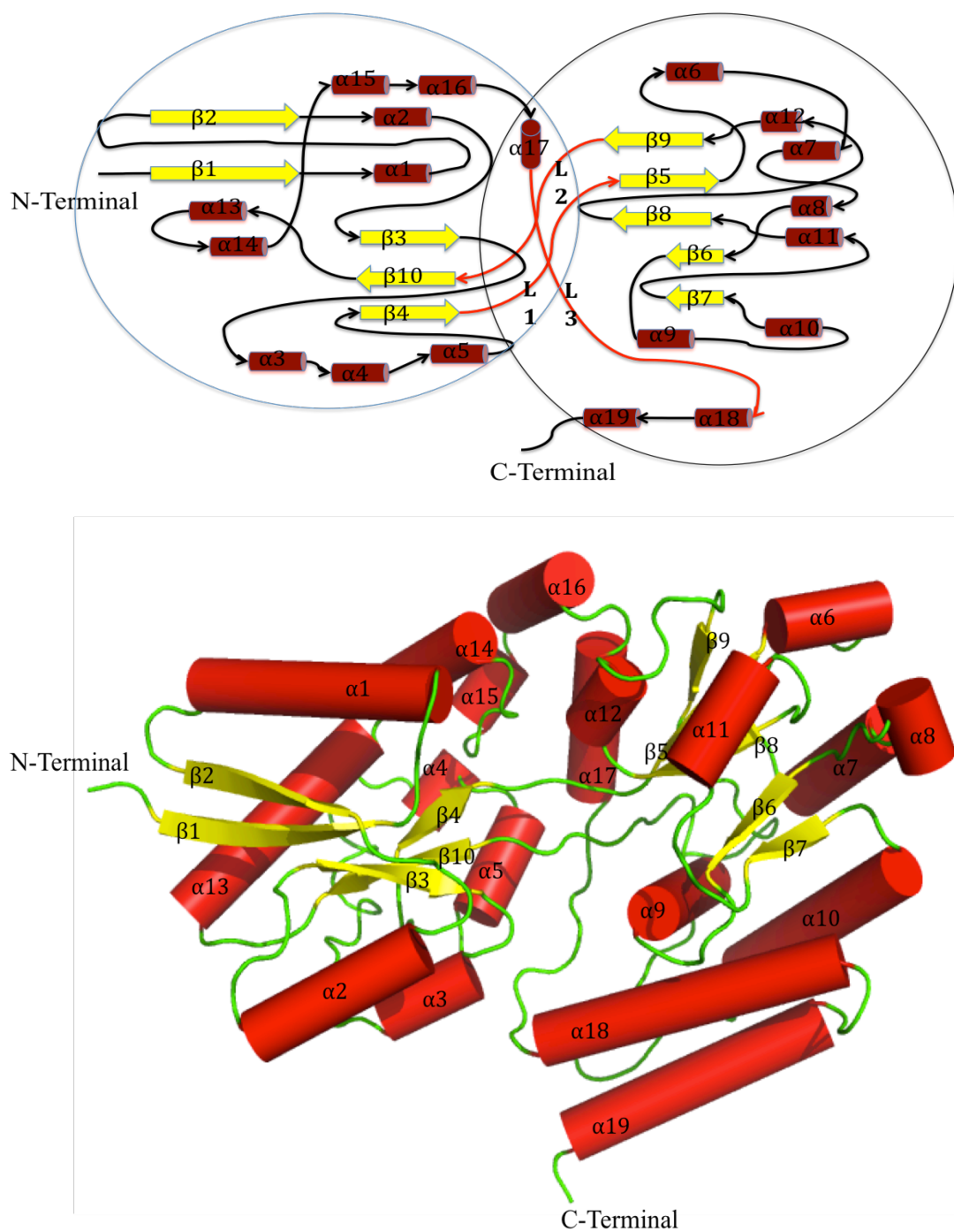


Figure 5.1 Schematic representation of AgaE structure. The top view is a schematic diagram, two domains are circled, β -stands colored yellow and α -helices dark red, hinge region connected both domains are colored light red as L1, L2 and L3. The bottom view is the number of helices and beta strands based on the primary structure of AgaE. Figure was created using Pymol [98].

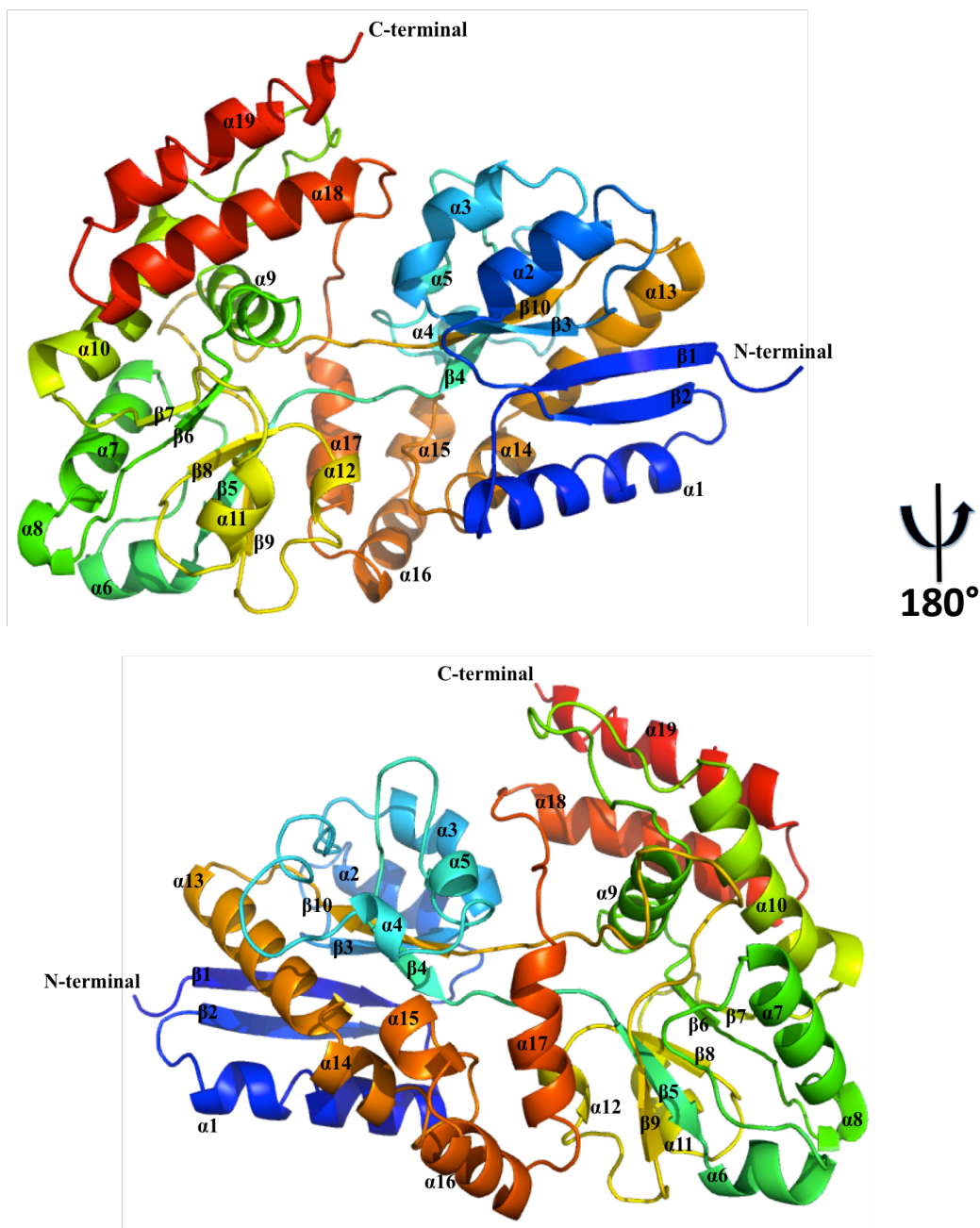


Figure 5.2 Cartoon representation of the overall fold structure of AgaE. The individual strands and helices are numbered. The molecule in the lower view has been rotated 180° with respect to the upper view.

5.1.2 Molecular surface

The two domains of AgaE fold into a compact, ellipsoidal, structure of dimensions (70.7 Å, 53.7 Å), with a clear deep, cleft running between them (Figure 5.2). The electrostatic surface of AgaE was calculated using Pymol [98] and is shown in figure 5.3. The surface has a mixed positively and negatively charged appearance, but at the base of cleft there is a deep negatively charged depression, formed by the side chains of residues (D41, E43, D71, D121, D122, D154, E198, D199, D274, D305, D342 and D351), possibly indicating a binding site. Indeed, in the structure clear difference electron density could be seen in this area, which could be interpreted as a molecule of polyethylene glycol (6-carbons) (Figure 5.4). It can be seen that the PEG molecule makes hydrogen bond interaction with Ser 95 and packs against R380 and P381 in the cleft.

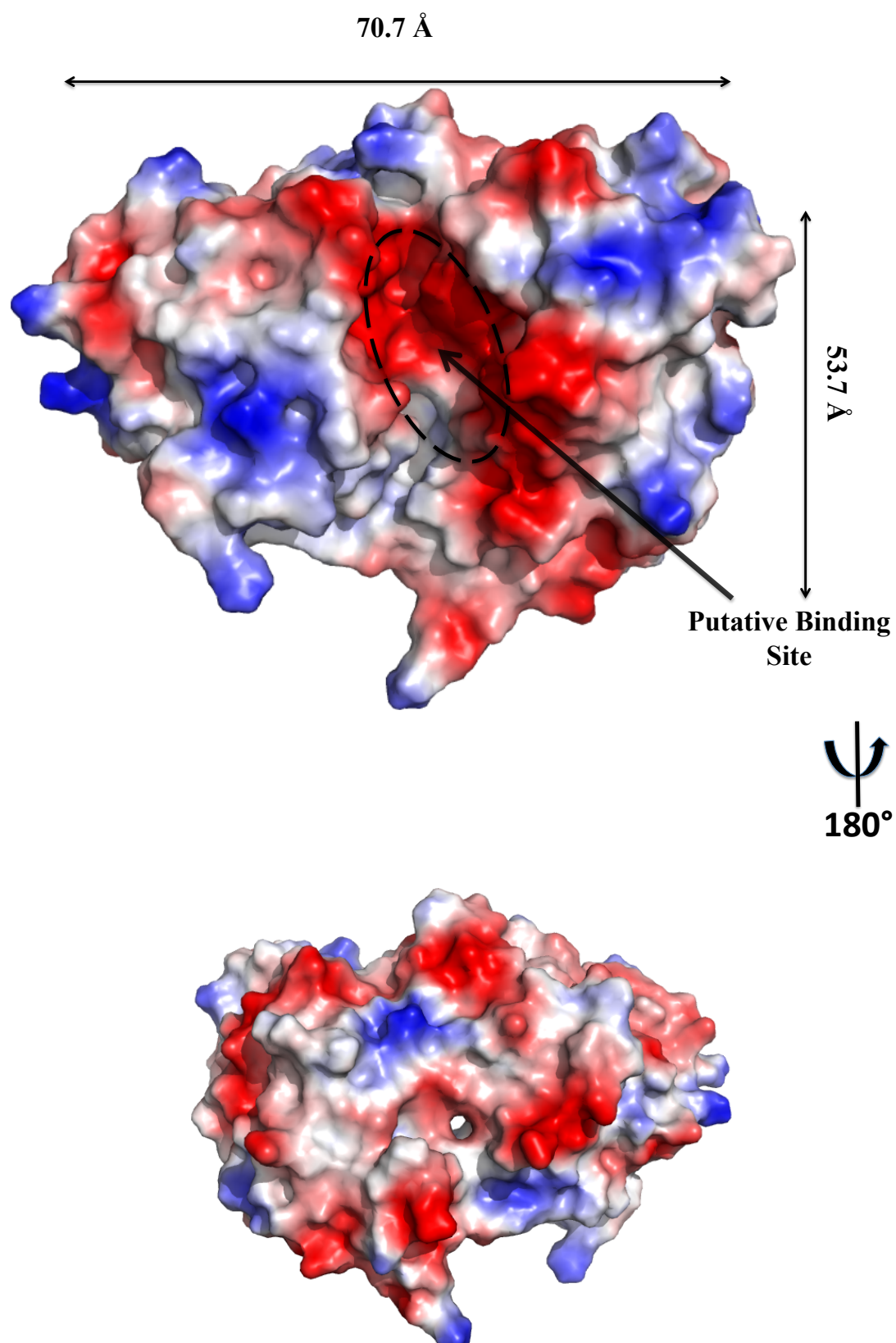


Figure 5.3 Surface electrostatic representation of the AgaE structure. The view in (a) is the front view, (b) is the back view of a. Positively charged residues are colored blue and negatively charged residues are colored red, respectively. The putative binding site is shown between the two domains in (a) with highly negatively charged site. The figure was produced using Pymol [98].

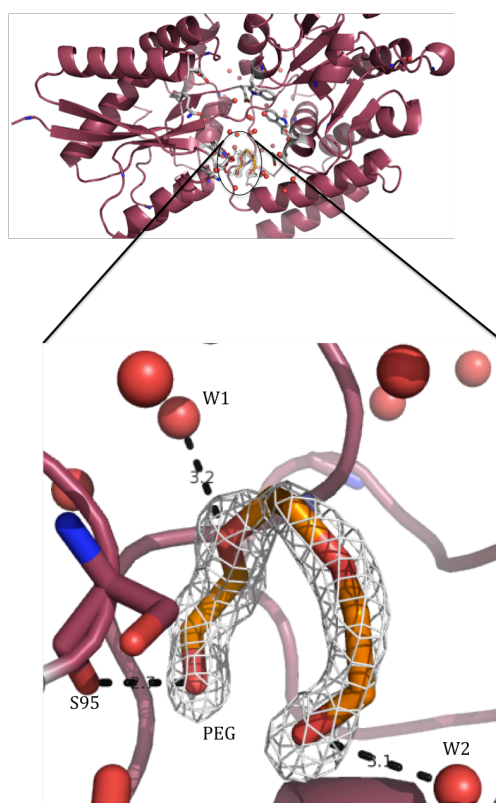
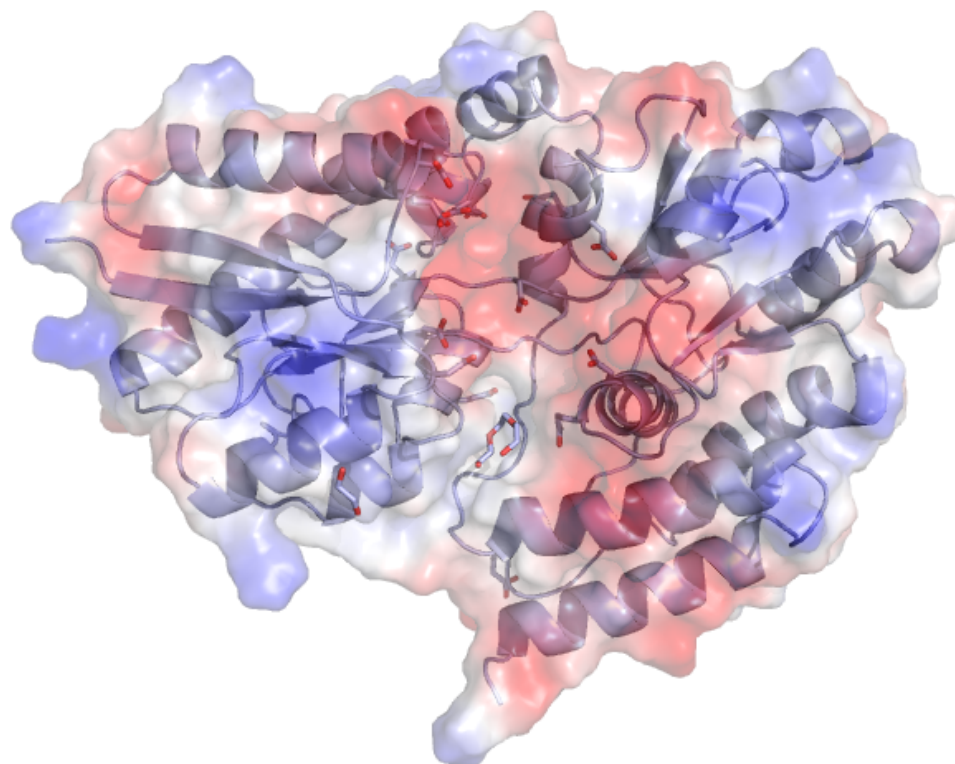


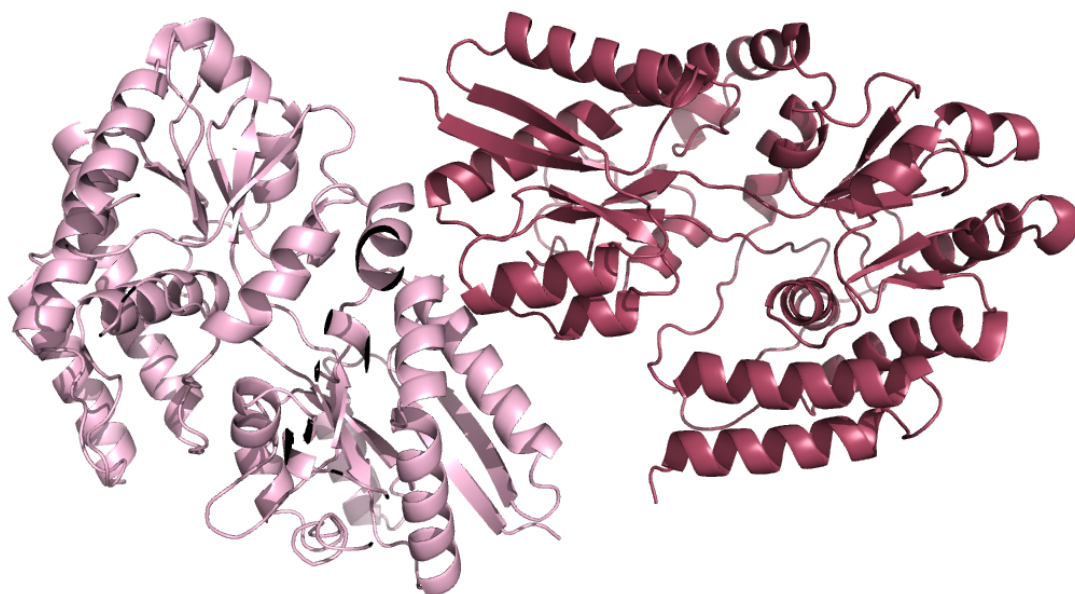
Figure 5.4 Cartoon representation of the putative active site between the two domains interface in complex with PEG. The PEG molecule is coordinated by two water molecules and makes hydrogen bond with Serine 95.

5.2 Crystal contact

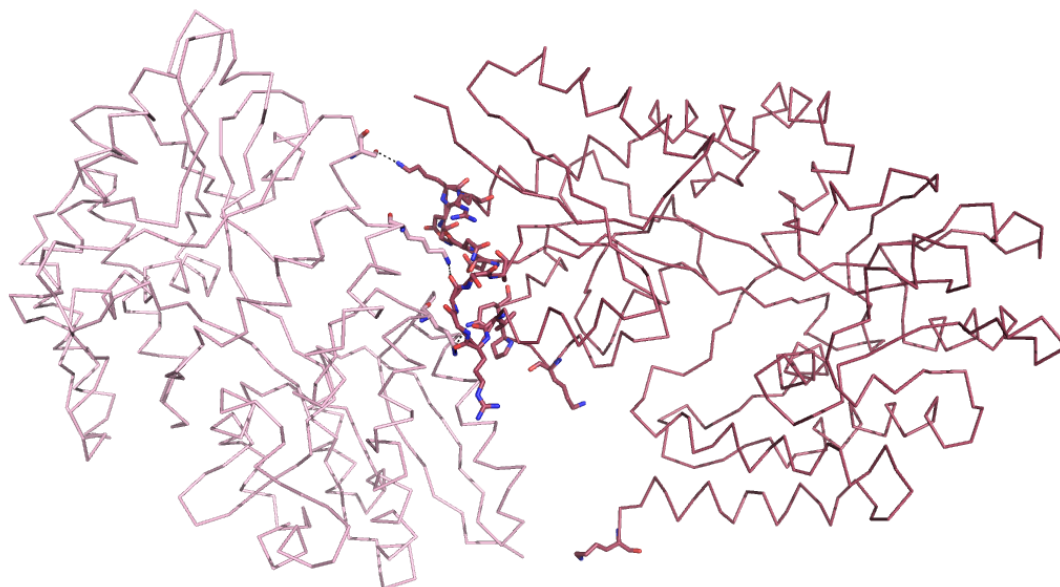
AgaE crystallizes with a monomer in the asymmetric unit, and so the crystal packing was analyzed to see if AgaE forms oligomers using the PISA server (http://www.ebi.ac.uk/msd-srv/prot_int/pistart.html) [137]. For chain A, which represents the AgaE protein structure, there are total of 398 residues in the protein chain and 354 of the residues contain some atoms that are exposed to the surface. The solvent accessible area for this protein is 16650 Å², with solvation energy for folding (ΔG) of -382.4 Kcal/mol.

The PISA server revealed that the most extensive crystal packing occurs between AgaE molecules related to each other by a 2-fold screw axis of symmetry (Figure 5.5). This interface is formed by residues in $\alpha 2$, the loop between $\alpha 2$ and $\beta 3$, $\alpha 3$ and the loop between $\beta 10$ and $\alpha 13$ of the N-terminal domain, packing against residues from $\alpha 1$, $\alpha 14$ and $\alpha 16$ from the N-terminal domain and $\alpha 12$ from the C-terminal domain of the symmetry related molecule.

A total of 15 residues from one molecule and 19 residues from the other are involved in the interface (Figure 5.6 a). The average interface area was calculated by PISA to be approximately 500 Å², which represents 3% of the total surface area. This lies outside the normal buried surface area of protein oligomers (5-25%) [138] and therefore, AgaE most likely occurs as a monomer. This is in good agreement with results from gel filtration (section 4.3.5), where AgaE elutes as a monomeric species.

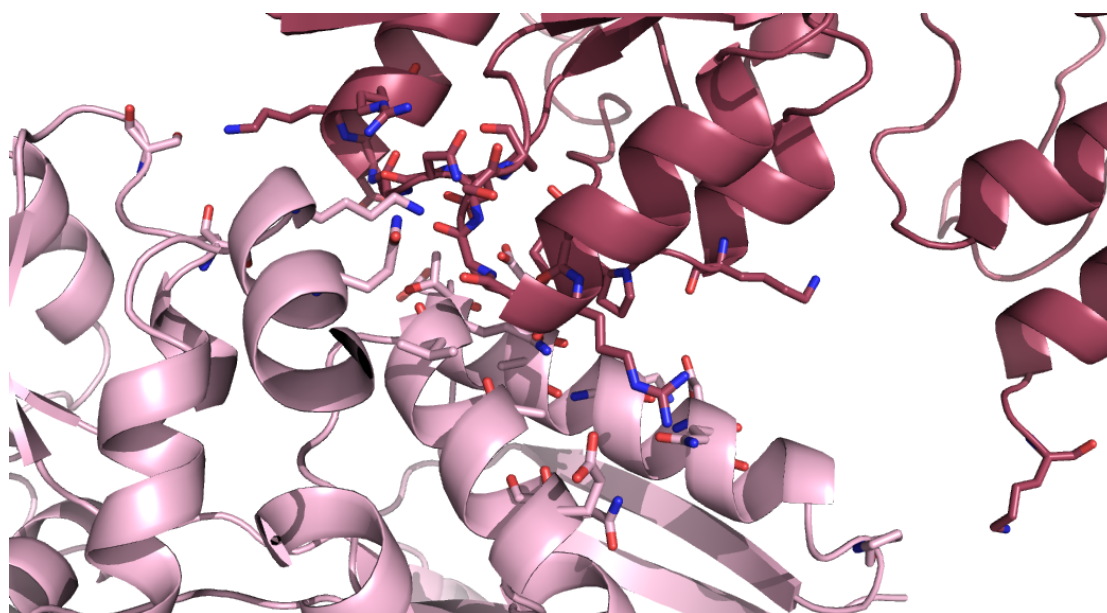


(a)

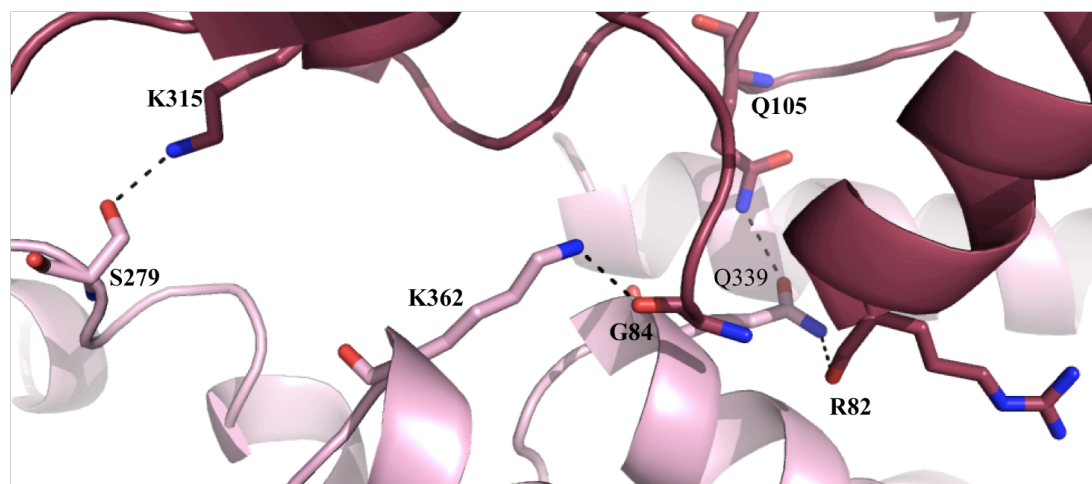


(b)

Figure 5.5 Cartoon representations of AgaE molecules packing in the crystal. **a)** A cartoon representation of the most extensive interface seen in the crystal packing of AgaE. **b)** A ribbon representation of both molecules with residues involved in the interface [137]. Figures were produced using Pymol [98].



(a)



(b)

Figure 5.6 Cartoon representations of the residues involved in AgaE crystal interface. **a)** Residues that are involved in the interface between two subunits in the AgaE structure (raspberry) and its symmetry related molecule (pink). **b)** Residues that were predicted to form potential hydrogen bonds between the two monomers are shown as sticks. Results were predicted using PISA server [137]. Figures were produced using Pymol [98].

Residue	Accessible surface area (Å ²)	Buried surface area (Å ²)	Accessible surface area (Å ²)	Buried surface area (Å ²)
	Monomer I		Monomer II	
ASP 41	118	-	118	14.5
THR 42	83	-	83	47
GLU 45	117	-	117	85.8
ALA 46	4	-	4	3.4
LYS 48	96	-	96	12.8
GLY 49	39	-	39	16.9
ASP 52	67	-	67	15.3
GLN 53	63	-	63	25.3
PRO 59	109	-	109	4.6
VAL 81	26	9.5	26	-
ARG 82	230	103	230	-
GLY 83	63	41.5	63	-
GLY 84	68	58.2	68	-
SER 85	72	20.5	72	-
ALA 86	9	0.37	9	-
LYS 102	123	10.6	123	-
ILE 103	7	-	7	-
PRO 104	119.3	87.2	119.3	-
GLN 105	77.4	62.5	77.4	-
SER 276	80.5	-	80.5	6.3
SER 279	96.1	-	96.1	27
ASP 311	57.1	45.3	57.1	-
ASN 312	44.4	20.3	44.4	-
GLY 313	38	13.2	38	-
ASP 314	119.2	12.2	119.2	-
LYS 315	155.4	40.1	155.4	-
ARG 316	86.3	27.2	86.3	-
GLU 331	119.5	-	119.5	27
GLN 332	15	-	15	2.5
VAL 333	11.4	-	11.4	-
LYS 334	76	-	76	-
ALA 335	27	-	27	15
LEU 338	57	-	57	38.3
GLN 339	100	-	100	70
GLN 358	102	-	102	34.0
LYS 362	176	-	176	90.0
LYS 425	234.5	5	234.5	5

Table 5.1 Accessibility and buried surface areas of residues on the monomer –monomer interface.

Residues in monomer 1	Residues in monomer 2	Bond distance (Å)
GLN 105[NE2]	GLN 339[OE1]	3.06
LYS 315[NZ]	SER 279[OG]	2.79
ARG 82[O]	GLN 339[NE2]	3.00
GLY 84[O]	LYS 362[NZ]	2.98

Table 5.2 The unique hydrogen bonds formed between the two monomers in the crystal.

5.3 Structure comparisons

5.3.1 Functional prediction

As the structure appeared to indicate that AgaE may well bind a small molecule in the cleft between the two domains, the structure was compared to others in the database to gain clues on possible function. Therefore, the refined model of AgaE was submitted to the Dali server [139] to identify proteins with a similar 3D structure and also with known function. The Dali search revealed several hits with high Z-scores, but each had low sequence identity (Table 5.3). These hits revealed that AgaE shares a similar structure with ABC sugar transporter proteins from several organisms, with the highest agreement with the Maltose binding protein from the phytopathogen *Xanthomonas citri* (PDB_code 3UOR; Z score 42.6; r.m.s.d. \approx 2.1 Å for 356 of C α -atoms), the Acarbose/Maltose binding protein from *Streptomyces glaucescens* (PDB_code 3K01; Z score 39.8; r.m.s.d. \approx 1.9 Å for 336 of C α -atoms), the cyclo/maltodextrin binding protein from *Thermoactinomyces vulgaris* (PDB_code 2ZYO; Z score 39.8; r.m.s.d. \approx 3.0 Å for 319 of C α -atoms), the trehalose/maltose-binding protein from *Thermococcus litoralis* (PDB_code 1EU8; Z score 39.2; r.m.s.d. \approx 2.9 Å for 334 of C α -atoms) and the maltose-binding protein (MBP) (PDB_code 3MBP; Z score 39.8; r.m.s.d. \approx 2.3 Å for 339 of C α -atoms) [140-143]. The AgaE structure is also similar to other solute binding proteins, such as sperimidine and putrescine binding domains from *Streptococcus pneumonia* (PDB_code 4EQB; Z score 23.7; r.m.s.d. \approx 3.1 Å for 284 of C α -atoms) (unpublished). In addition, other enzymes share a similar fold, such as thiaminase I from *Bacillus thiaminolyticus* (PDB_code 2THI; Z score 24.5; r.m.s.d. \approx 3.4 Å for 283 of C α -atoms) [144].

Taken together, these structural similarities indicated that AgaE is most likely a solute binding protein. Therefore, The sequences of the protein that formed the top hits with AgaE in the Dali search were aligned with that of AgaE based on the structural similarity of these proteins (Figure 5.7). Each of these proteins contains approximately 420 residues, yet only 20 are strongly conserved across the family. These residues were plotted on the structure of AgaE (Figure 5.8), where it can be seen they are dispersed across the structure and indicate that these residues probably play a structural role in defining the fold of the protein family.

Protein	PDB	Z-Score	RMSD	Seq. identity %	Substrate	Organism
MalE	3uor	42.6	2.1	21	Maltose	<i>Xanthomonas citri</i>
GacH	3k01	39.8	1.9	27	Acarbose & Maltose	<i>Streptomyces glaucescens</i>
TvuCMBP	2ZYO	39.8	3.0	21	Cyclo & maltodextrin	<i>Thermoactinomyces vulgaris</i>
TMBP	1EU8	39.2	2.9	23	Trehalose	<i>Thermococcus litoralis</i>
MalE	4HW8	39.1	2.6	23	PEG	<i>Staphylococcus aureus</i>
MalE	2GHB	38.6	2.8	22	Maltose	<i>Thermotoga maritima</i>
MalE	3PUY	36.0	2.5	20	Maltose	<i>Escherichia coli K-12</i>
MalX	2XD2	36.0	2.4	20	Maltopentaose	<i>Streptococcus pneumoniae</i>
GL-BP	2Z8F	31.9	3.3	18	Galacto-N-biose-/lacto-N-biose.	<i>Bacillus subtilis</i>
ttGBP	2B3B	29.1	4.2	16	Glucose	<i>Bacillus subtilis</i>
MalE	3MBP	33.0	2.3	21	Maltose and maltotriose	<i>Escherichia coli K-12</i>
MalE	3KJT	35.9	2.4	21	Maltose	<i>Escherichia coli K-12</i>
MalE	1URS	33.6	3.5	25	Maltose	<i>Alicyclobacillus acidocaldarius</i>
PotD	4EQB	24.2	3.1	12	Spermidine & Putrescine.	<i>Streptococcus pneumoniae</i>
Thiaminase	2THI	24.9	3.4	12	2,5-Dimethyl-pyrimidin-4-ylamine	<i>Bacillus thiaminolyticus</i>

Table 5.3 Results of Dali search using the model of the putative sugar binding protein Mseg_0515 (AgaE). The first 15 hits are listed based on the highest Z-score, RMSD score and sequence identity respectively. Although the AgaE structure shares similar structure to these proteins, the sequence identity is very low.

residues are highlighted and boxed. Consensus residues (aa) represents the highly conserved amino acid residues, which are shown as bold and uppercase letters as following; **aliphatic** residues (I, V, L): *l*, **aromatic** residues (Y, H, W, F): *@*, **hydrophobic** residues (W, F, Y, M, L, I, V, A, C, T, H): *h*, **alcohol** residues (S, T): *o*, **polar** residues (D, E, H, K, N, Q, R, S, T): *p*, **tiny** residues (A, G, C, S): *t*, **small** residues (A, G, C, S, V, N, D, T, P): *s*, **bulky** residues (E, F, I, K, L, M, Q, R, W, Y): *b*, and **charged** (D, E, K, R, H): *c*. Consensus secondary structure prediction (ss) are shown underneath the alignment as **alpha-helix**: *h* and **beta-strand**: *e*. This figure was created using PROMALS3D [145].

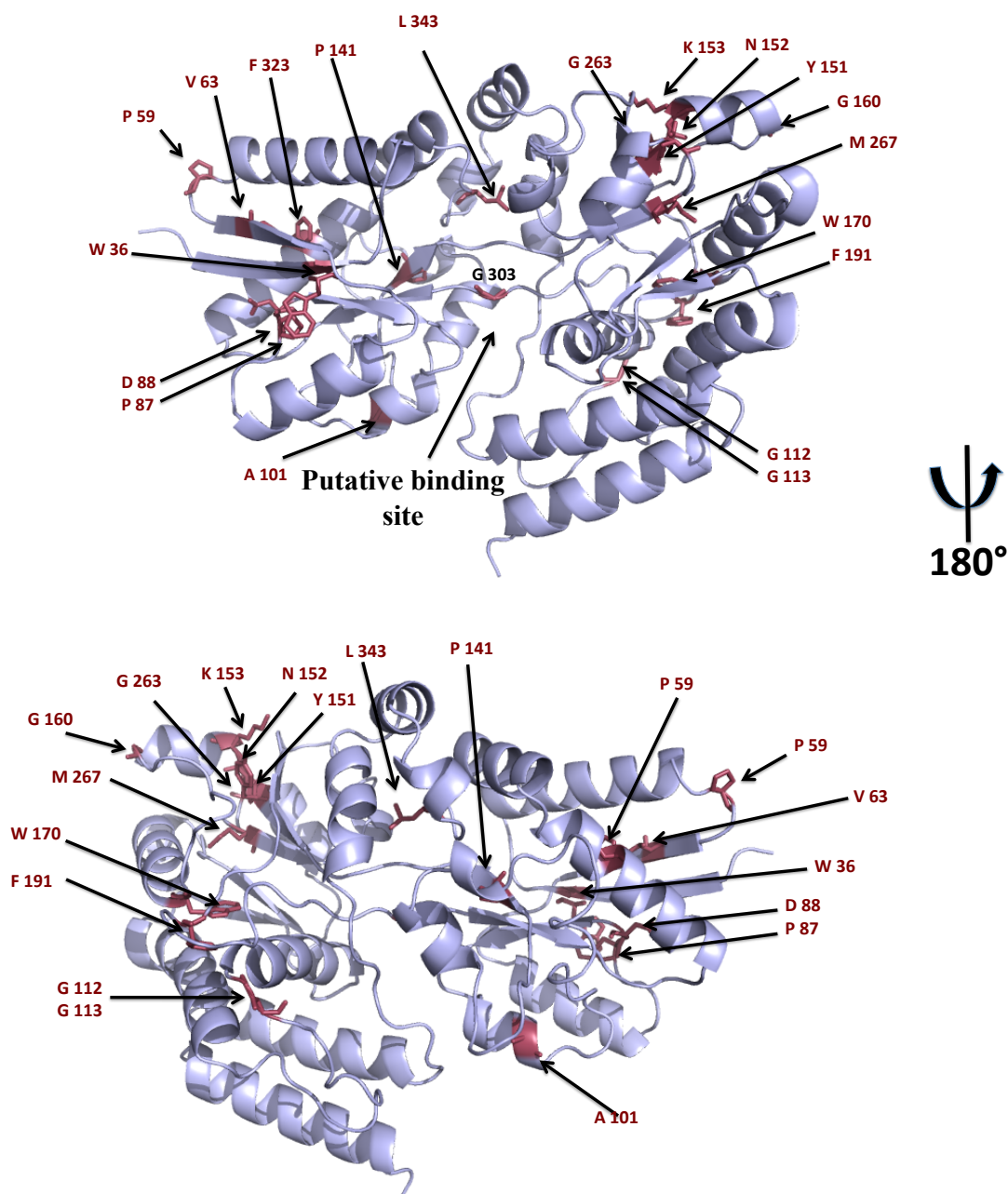


Figure 5.8 Highly conserved residues of different solute binding transporter mapped onto AgaE. **b)** The back view of (a) with a vertical rotation of $Y=180^\circ$ between the two views. The highly Conserved residues were identified from sequence alignment of different ABC transporter shown in figure 5.7, which were boxed and shaded. The structure is colored blue and highly conserved residues are in raspberry color.

5.3.2 Structure classification of solute binding proteins

The structure of AgaE shares overall structural similarity with other solute binding proteins. Each domain of this family contains a five stranded β -sheet surrounded by α -helices. Both domains are linked to each other by a hinge region, composed of two loops and an α -helix from both domains that controls the movement of the two domains, upon substrate binding in the cleft between them. The structures of Solute Binding Proteins have been classified into two groups [146]. The classification is based on the variation on the number of linkages (hinge) between the two domains, and the β -sheet topology of both domains. The first class includes structures that have the β -sheet topology of $\beta 2\beta 1\beta 3\beta 4\beta 5$, with the hinge region formed by three loops connecting both domains together. This class includes structures that bind monosaccharides, such as the *Escherichia coli* glucose/ galactose binding protein [146, 147]. The second class includes structures that have the β -sheet topology $\beta 2\beta 1\beta 3\beta n\beta 4$ and the hinge region between the two domains is formed by two loops only. The beta strand βn of the second class corresponds to the interdigitating strand that occurs after the other domain has folded [148]. This class includes structures that bind di and oligosaccharides, such as the maltose binding protein of *Escherichia coli* [124, 149]. An additional domain was identified as class three of a solute binding protein, and in this class, the domains are joined by only one connecting linker, of an α -helix between both domains, such as the staphyloferrin binding protein (HtsA) from *Staphylococcus aureus* [146, 150].

Solute binding proteins have also been classified into six groups (A-F) based on structural alignments of all crystal structures submitted into the protein data bank (PDB) and their substrate specificities [148]. The group A SBPs are ABC transporters belonging to class III of the Fukami-Kobayashi classification and are specific for metal binding, while group B members belong to class I SBPs and mainly bind carbohydrates. However, this group has also been found to bind other molecules such as amino acids [146]. The group C, D and F are all class II SBPs and bind different types of substrate, such as carbohydrates, iron and thiamine, however, both C and D groups have an extra domain and they are larger in size (class C > 55 kDa and class D > 40 kDa) [148]. Also, the lengths of the hinge regions between the two domains are shorter in group D (4-5 amino acids) than in group F SBPs (8-10 amino acids). Furthermore, group D SBPs was also subdivided into further three

subgroups based on the substrate binding specificity; subgroup I SBPs are carbohydrate specific, such as MBP from *E.coli*; subgroup II are thiamine and polyamine specific binding proteins, subgroup III are SBPs that bind phosphate, sulfate and oxyanion molecules and finally, subgroup IV includes those SBPs that are specific for ferric iron binding [148]. The last group of the SBPs classification is group E, which belongs to the TRAP transporters protein family.

As the overall structure of AgaE resembles that of maltose binding protein structure from *E.coli*, which was categorized as a class II SBP belonging to group D-1 in the substrate specificity based classification, it can be suggested that AgaE is a carbohydrate binding protein belonging to class II substrate binding protein.

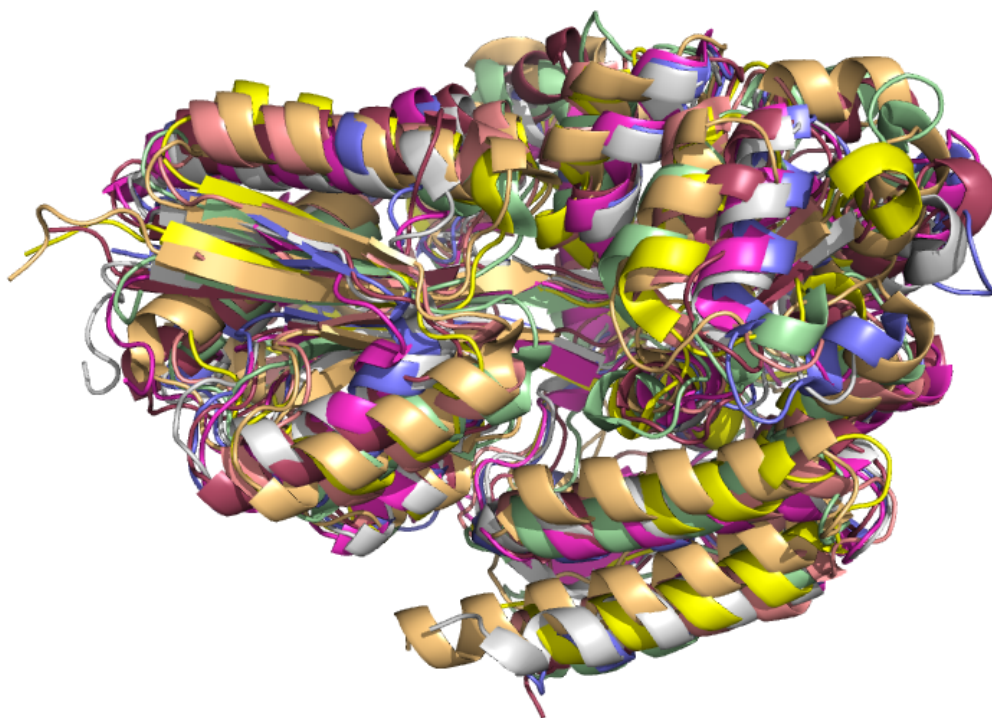


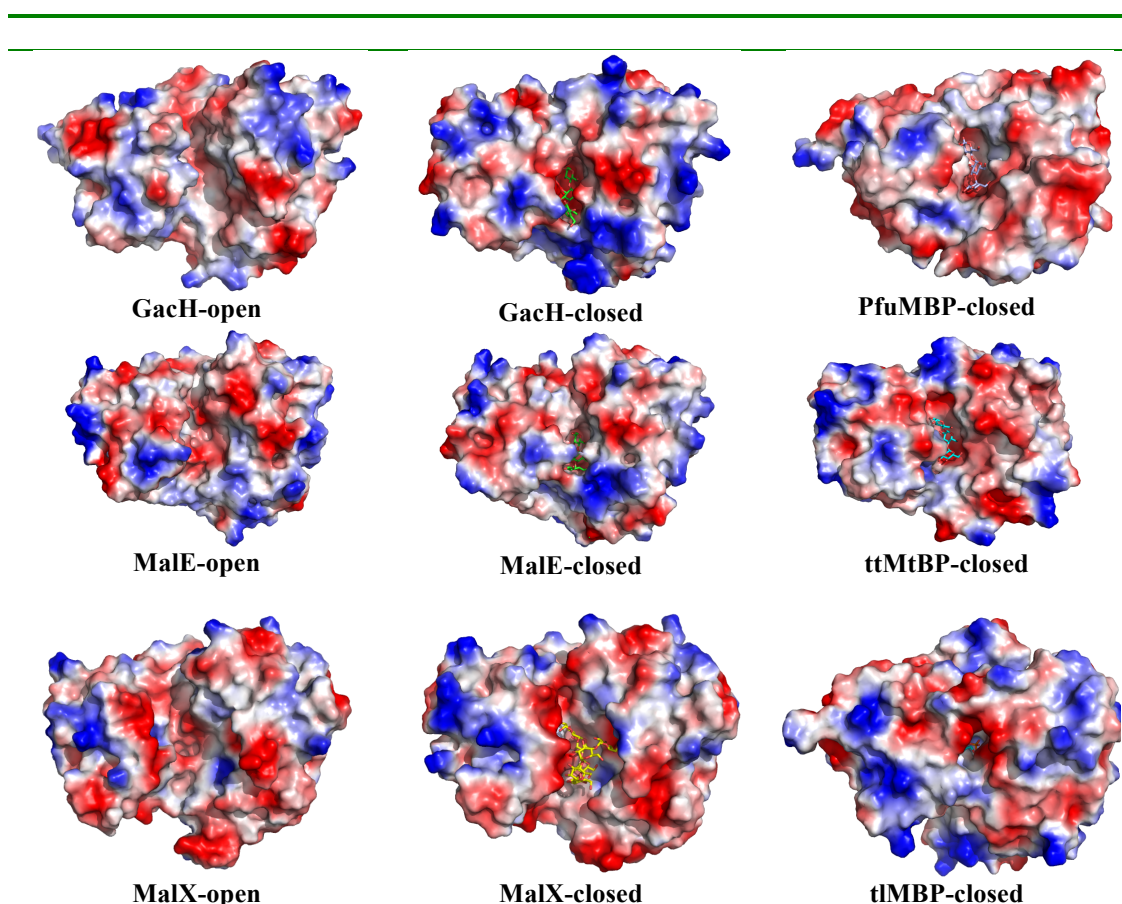
Figure 5.9 structural alignment of AgaE structure and other sugar binding proteins.

5.3.3 Binding site of carbohydrate binding proteins

In order to investigate the possible binding site of AgaE, the binding sites of the solute binding proteins were compared to identify any similar features that were also present in the AgaE structure. The binding pocket of a typical solute binding protein is located in the cleft in the interface between the two globular domains, that are linked to each other by the three-hinge regions, forming a bilobate structure [151]. Comparing the AgaE structure with other solute binding proteins, such as the maltose and acarbose binding protein from *S.glaucescens* (GacH) and the maltose binding protein (MalE) from *E.coli* revealed that all solute binding proteins interact with their different substrates with a similar mode of action. In all the protein / sugar complexes determined, the sugar moiety binds in the cleft between the two domains. Upon binding the sugar, the two domains move closer together, by up to 6.0 Å, to exclude bulk water and to allow residues from both sides of the cleft to interact with the sugar. The molecular surfaces of this family of proteins all exhibit a bilobal structure (Figure 5.10), however the electrostatic surface of the cleft varies between them, reflecting the variety of substrates bound (Figure 5.10).

The binding sites of a number of different members of the solute binding proteins family were compared, and a number of similarities were seen across the family. These include residues that make direct and water mediated hydrogen bonds to the various ligands as well as other residues involved in van der Waals contacts [151]. The number and type of these interactions depends upon the size of the solute bound. The protein with the closest structural similarity to AgaE is the solute binding protein (GacH) from *S.glaucescens*, which is the receptor of the putative oligosaccharide ATP binding cassette transporter (GacFG). GacH is an acarbose / maltose binding protein [141], and its structure in complex with maltotetraose (4 rings of α -D-Glucopyranosyl sugar) and acarbose (a maltose bound to an acarviosin moiety) and 5C (acarviosyl-1,4-maltose-1,4-glucose-1,4-glucose) have been determined by Vahedi-Faridi, Licht et al. 2010). As GacH binds four ring solutes, its structure forms a good template to identify possible sugar binding regions of AgaE. The overall structure of the open form GacH binding protein is superimposed very well with AgaE structure with a root-mean-square deviation (RMSD) of C α position of 1.9Å despite the low sequence identity of 25% (Figure 5.11). Although both structures share the same fold, their secondary structures have slight variations in the number and length of α -helices and β -strands. For example, additional small α -helices are

found within the GacH binding protein structure corresponded to loops in AgaE structure, such as the loop that is located in between α_{12} and β_{10} [141]. Also, the two antiparallel β -strands present in the C-domain in AgaE are not conserved in all MBP structures and replaced by short α -helices in maltopentose binding protein (MalX) from *S.pneumoniae* (PDB_code 2XD2) [152]. Furthermore, the hinge region that connects the two domains of GacH consists of three loops and link two β -strands are longer than those found in AgaE hinge regions (7-10 and 4-5 residues long, respectively) and α -helix [141]. The structure of GacH in complex with acarbose is shown in figure 5.11. The binding is constructed from residues from the N-domain, the C-domain and the loops that joins the two domains. The residues from the N-domain and the joining loop mainly form hydrogen bonds to the hydroxyls of the solute, whereas, those from the C-domain construct a hydrophobic surface that interacts with the sugar rings.



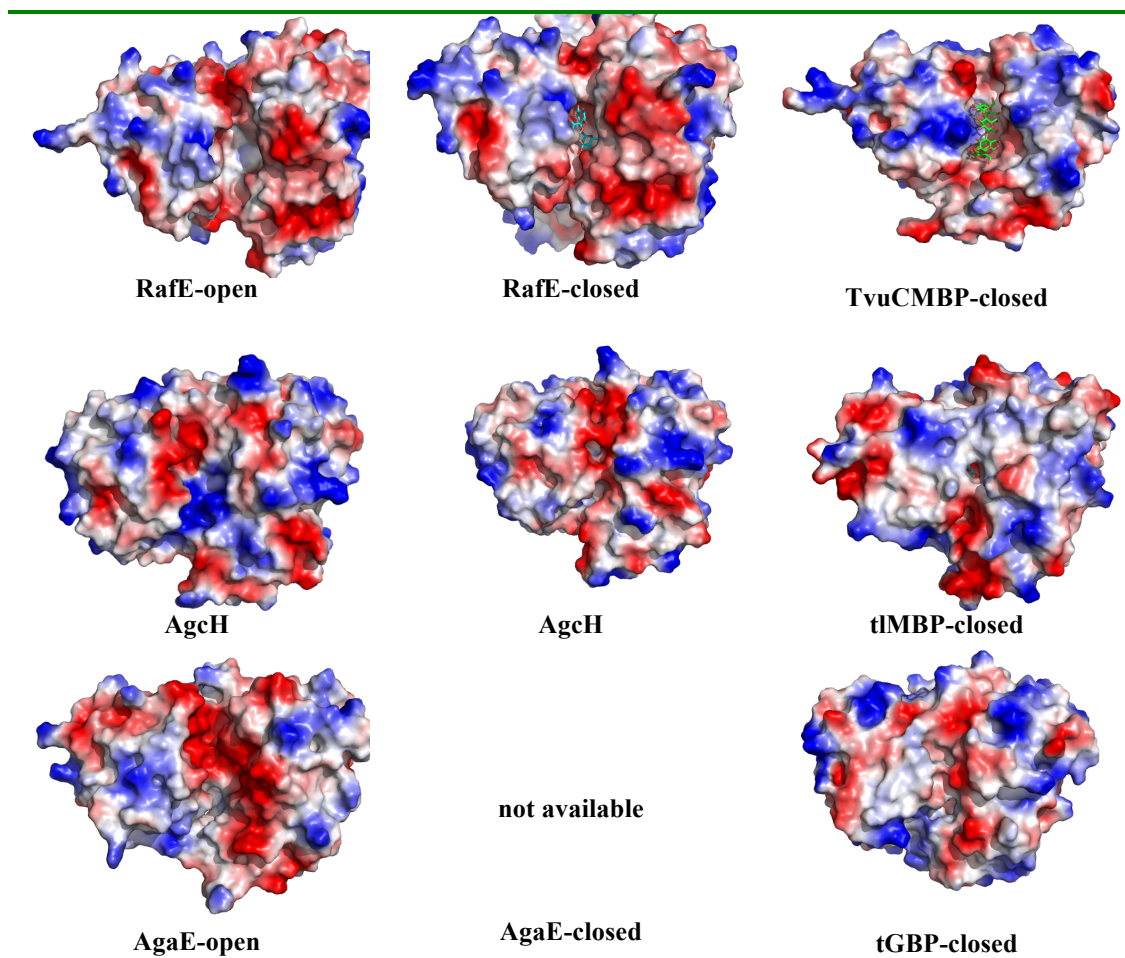


Figure 5.10 Surface electrostatic representation of the carbohydrate binding protein structures with different sugar bound. The surface electrostatic was calculated for both open and closed forms based on the structure available in the protein database. Positive charged residues are colored blue and negative charged residues are colored red, respectively. The putative Binding site is shown between the two domains. The figure was produced using Pymol [98].

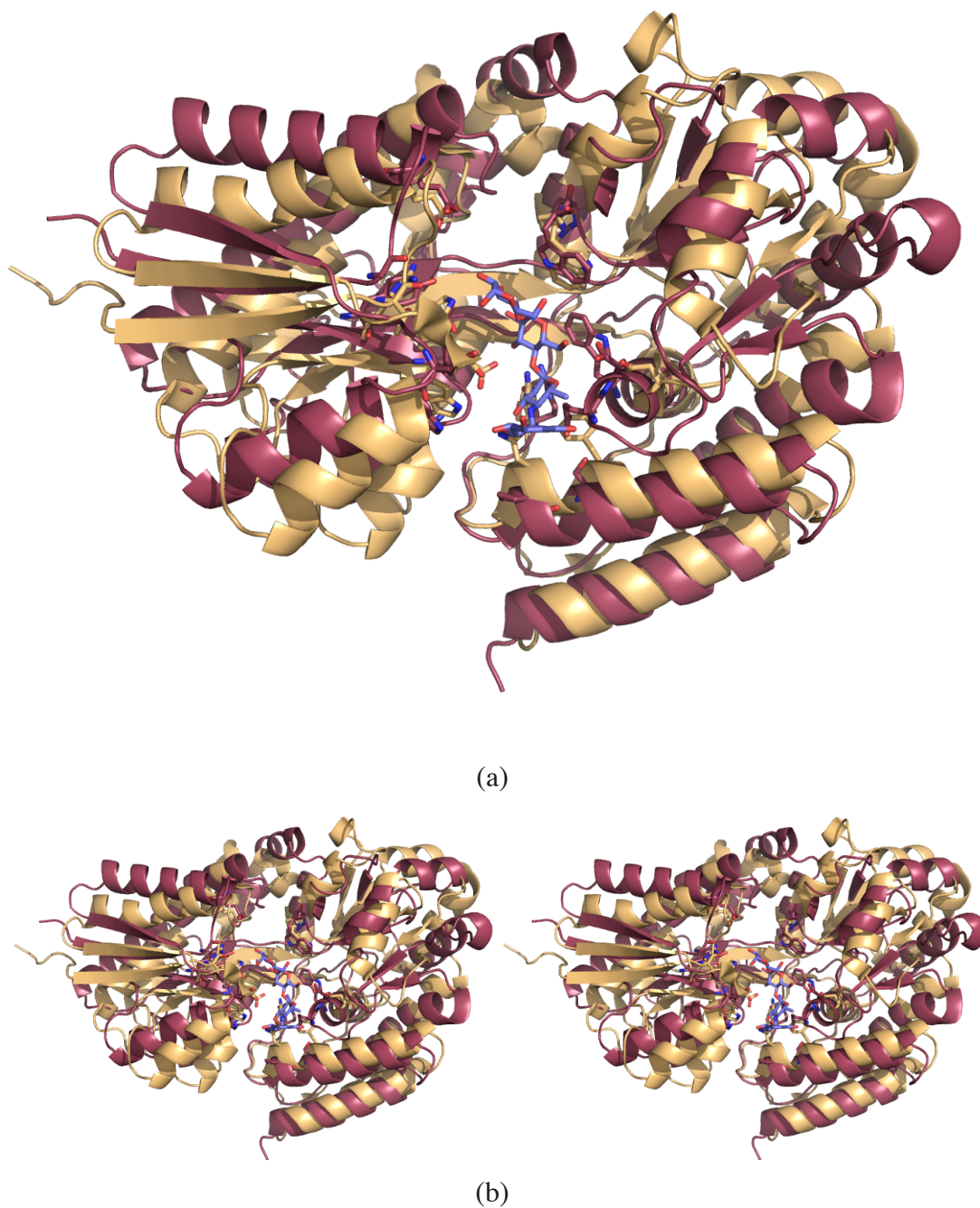


Figure 5.11 Structural alignment of GacH sugar binding protein structure and AgaE structure. a) The overall structure of GacH binding protein (orange) is superimposed very well with AgaE structure (raspberry) with a root-mean-square deviation (RMSD) of $C\alpha$ position of 1.9Å. The bound acarbose of GacH is shown in between the two domains as purple stick and binding residues from each structure as sticks. b) Stereo view of a.

5.3.3.1 Hydrophobic surface

The C-domain residues Y182, W183, W243, W254 and F368 form a hydrophobic surface on one side of the cleft in GacH (Figure 5.12). The equivalent residues in AgaE (W202, H203, W273, Y386 AND P387, table 5.3), are also hydrophobic and aromatic and form a surface with similar properties (Figure 5.12).

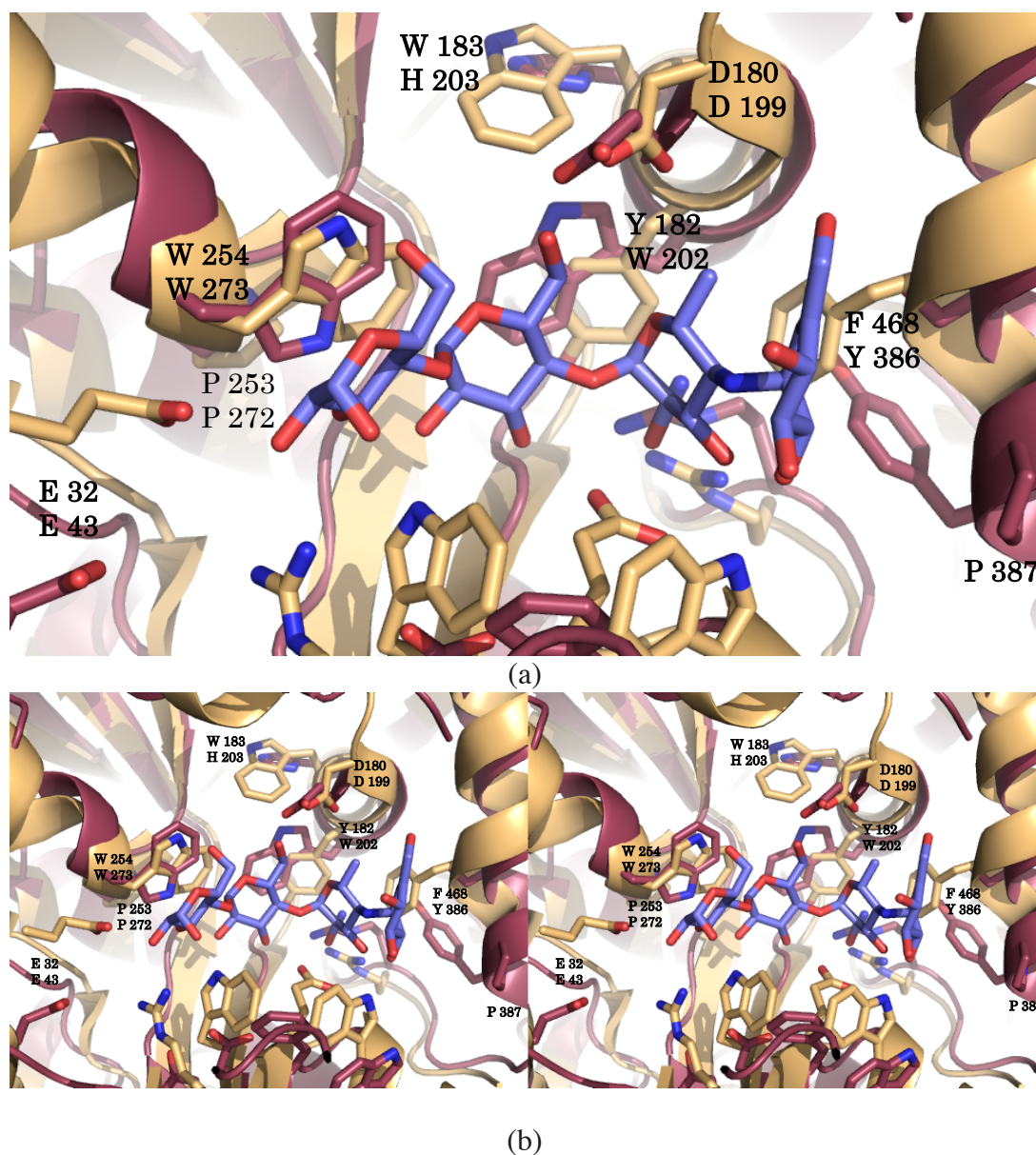


Figure 5.12 Superposition of AgaE (raspberry) and GacH (orange) binding sites. a) The C-domain binding residues alignment of AgaE and GacH. b) Stereo representation of a. the bound sugar (acarbose) of GacH is shown as purple stick.

5.3.3.2 Hydrogen bond interaction

In GacH, residues from the N-domain provide side chains that form hydrogen bonds with the hydroxyl substituents of the four rings acarbose solute. Residues E32 and R81 form hydrogen bonds to the 2-OH of the second ring, with the carboxyl of E83 forming hydrogen bonds to both the 2-OH of the third ring. R358 interacts with the 2-OH of the third ring and the N_ε of W385 interacts with both the 2-OH and the 3-OH of the fourth ring. Both D133 and R358 are found on the loops that join the two domains.

In the AgaE structure there are residues with equivalent position and functional groups to all of these acarbose hydrogen-bonding residues in the GacH structure (Table 5.4). In the AgaE structure the bound PEG molecule partially occupies the position of rings three and four of the acarbose in the GacH / acarbose complex, with S95 forming a hydrogen bond to the PEG. As both the hydrophobic surface and the hydrogen bonding residues are conserved in the AgaE structure and PEG binds in a similar position, it seems likely that a carbohydrate type solute may bind to AgaE.

GacH	MalE	AgaE	GacH	MalE	AgaE
E32-O1	D14-O2	E43	Y182	Y155	W202
R81-NH1	K15-N	Y91	W183	F156	H203
E83-O1 & O2	D65-O1 & O2	S95	W234	Y210	-
W86-NE1	R66-NE & NH2	W96	W254	W230	W273
			F368	W340	Y386
D133-O2	E111-O1	D145	-	Y341	P387
W290-N1	-	D305	-	-	-
R358-NH1 & NH2	M330	R380	-	-	-
			-	-	-

a) Hydrogen binding residues

b) Hydrophobic binding residues

Tables 5.4 Residues involved in the sugar binding of the GacH and MalE binding proteins and their equivalents in AgaE protein structure. A) Hydrogen binding residues. B) Hydrophobic binding residues

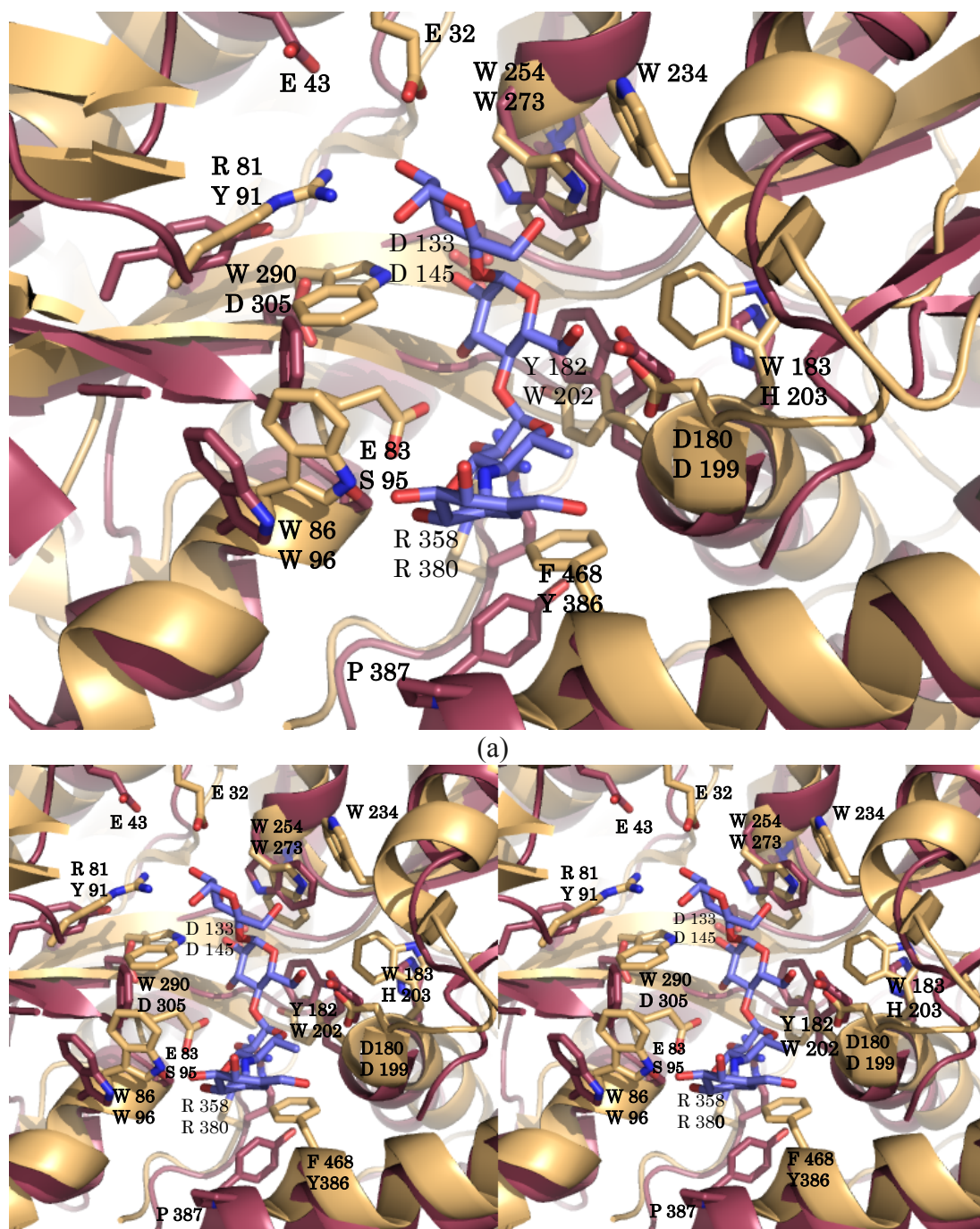


Figure 5.13 Superposition of AgaE (raspberry) and GachH (orange) binding sites. a) N-domain (left) and C-domain (right) binding residues alignment of AgaE and GachH are shown as stick. b) Stereo of (a). Acarbose is shown as purple stick in between the two domains.

5.3.4 Sugar binding site in other proteins

The analysis described above was extended to the protein typical solute binding protein MalE from *E.coli*. This protein binds maltose and maltodextrin and its structure has been determined in complex with acarbose [143] (PDB_code 3JYR). However, when this structure is superimposed on that of the GacH / acarbose complex, the acarbose moieties do not fully superimpose. Rings 2, 3, and 4 of the acarbose in GacH occupy the same space as rings 1, 2 and 3 of the acarbose in the MalE / acarbose complex, showing how this family of proteins could bind sugars with more rings than acarbose (Figures 5.14 and 5.15). In the MalE structure, D14 and K15 occupy the equivalent space as the first ring of the acarbose in the GacH structure. The carboxyl of D14 in MalE coordinate the 1-OH of the first ring of acarbose and the amine of K15 hydrogen bonds to the 2-OH of the same ring. These residues do not have equivalents in GacH. Furthermore, the loop between helix (α 1) in GacH where these residues occur is longer than that in MalE, providing more space for the acarbose to bind. However, in AgaE D41 lies in an equivalent position to D14 of MalE, and upon domain closure in AgaE, may well occupy a similar position. The hydrophobic surface described for GacH and AgaE is also present in MalE, being formed by residues Y155, F156, Y210, W230, W340 and Y341 (Figures 5.14 and 5.15). Also, in MalE, a similar set of residues provide hydrogen bonding to the acarbose, as seen in GacH, but as the acarbose moiety is displaced by one ring, the precise interactions are different. Residues involved are D14, K15, R66, E44 and E111. There is no direct equivalent to R358 in GacH, with MET 330 occupying this position in MalE (Table 5.4).

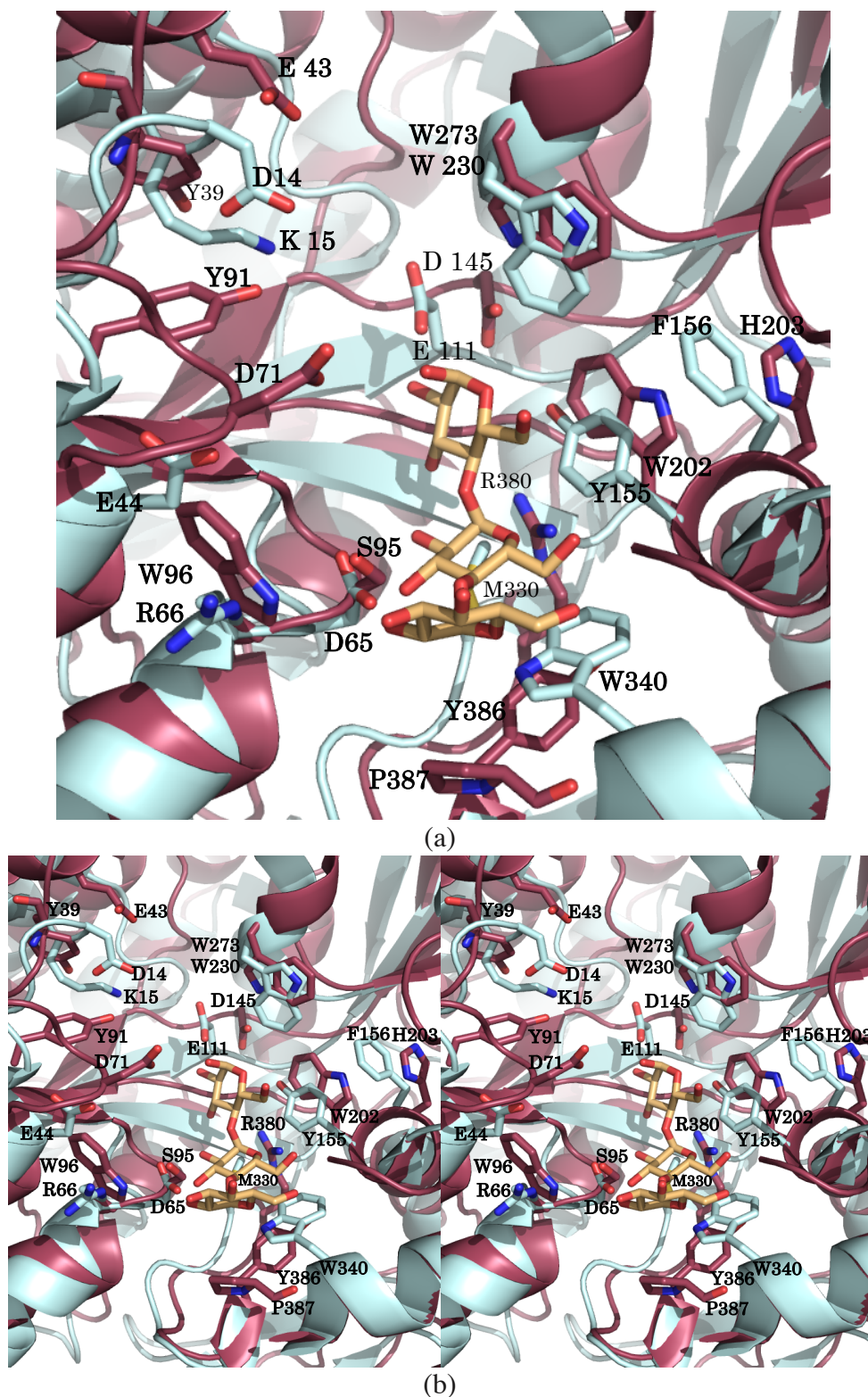
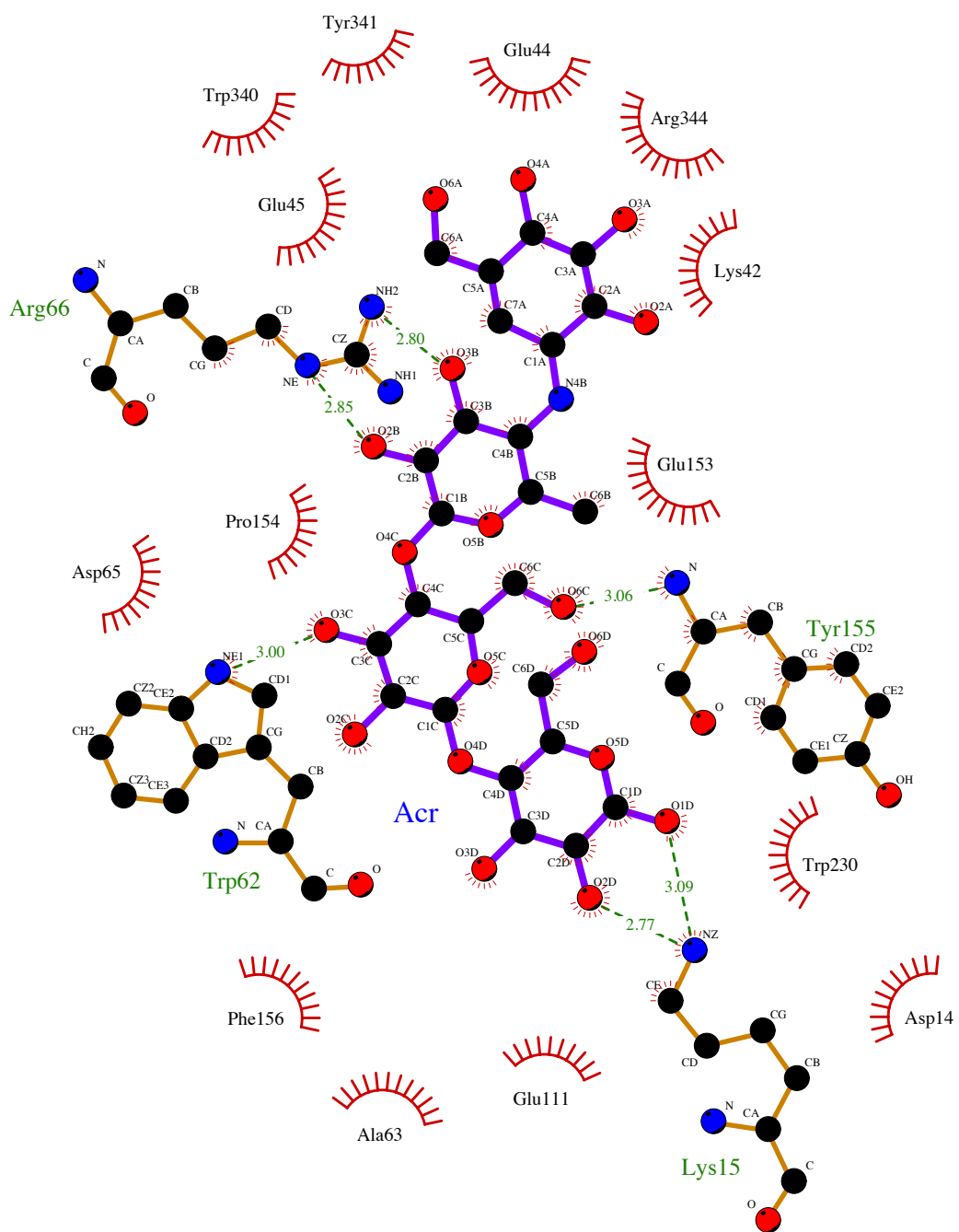
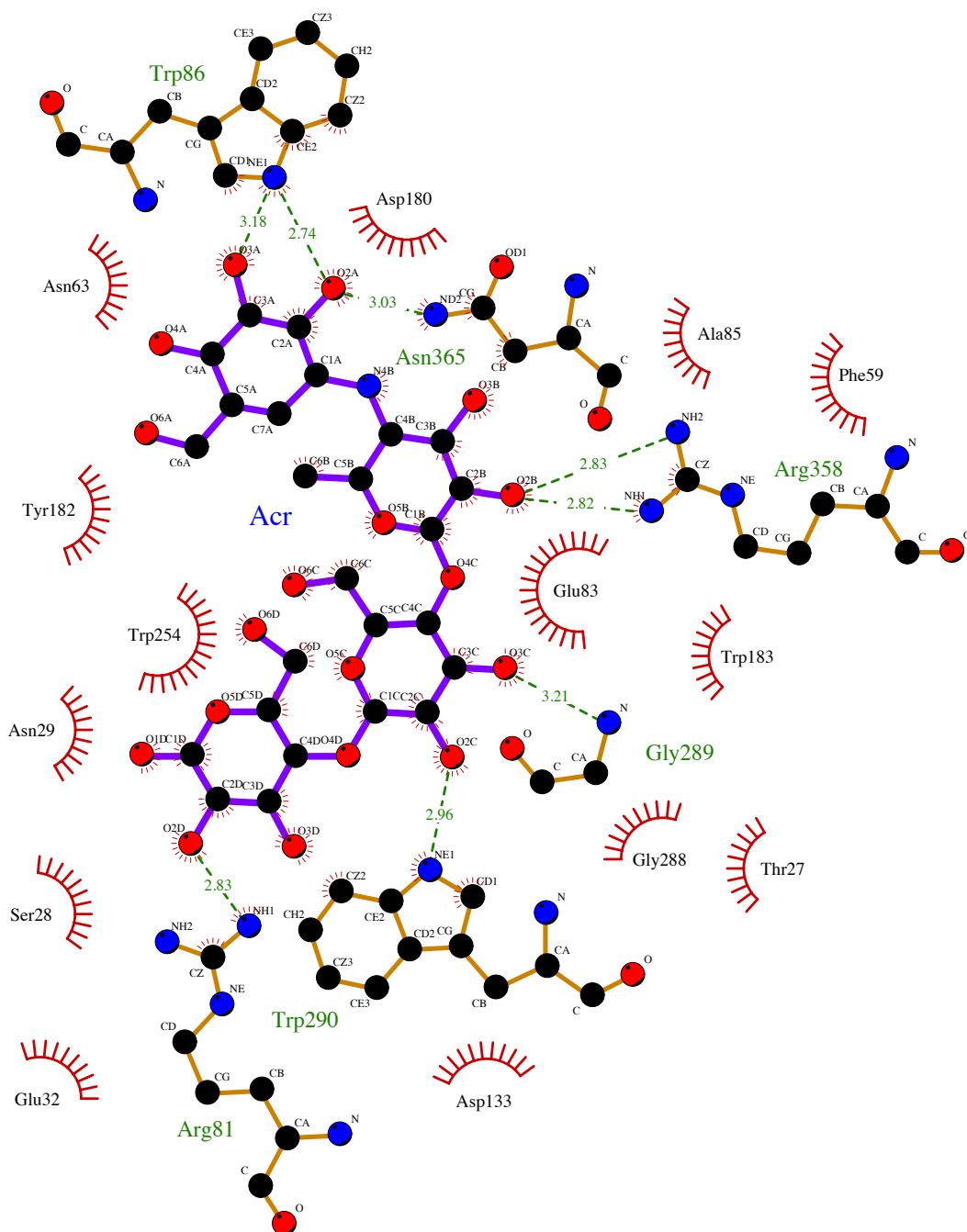


Figure 5.14 Superposition of AgaE (raspberry) and MalE (cyan) binding sites. a) N-domain (left) and C-domain (right) binding residues alignment of AgaE and GacH are shown as stick. b) Stereo of (a). Acarbose is shown as brown stick in between the two domains.



(a) MalE



(b) GacH

Figure 5.15 Ligplot diagram representation of the binding site interaction of GacH and MalE with the bound acarbose. Residues in both sites are involved in direct and waters mediated hydrogen bonds and hydrophobic interaction. Figures were created using LigPlot [153].

5.4 Sugar binding proteins specificity

This analysis extended to mono-di and oligosaccharides binding proteins that have a similar structure to AgaE. These include five different types of solute binding proteins, malto-oligosaccharide binding proteins including GacH, MalE, ttMtBP, *PfuMBP*, *TvuCMBP* and MalX, Trehalose binding protein (tlMBP), raffinose binding protein (RafE), glycerol 3 phosphate binding proteins (UgpB) and monosaccharide binding proteins ttGBP and AgcH. The aligned sequences for these proteins are shown in figure 5.16, with the residues that bind the carbohydrate boxed. In all these structures, the residues that form the binding site come from three areas, site I (N-domain residues), site II (C-domain residues) and site III (loops between the domains). When these structures are compared to that of GacH and MalE, some indications about substrate specificity of this family can be made (Figures 5.17, 5.18 and 5.19). The hydrophobic surface is conserved throughout proteins that bind maltooligosaccharides, but not those that bind other substrates. These residues are lying on $\alpha 9$, $\alpha 11$, $\alpha 12$ and $\alpha 18$ (Site III). Similarly, the residues that form hydrogen bonds are lying on $\beta 1$, $\beta 3$ and $\beta 10$ and helices $\alpha 1$, $\alpha 2$ and $\alpha 3$ (site I) and have similar properties. For ttGBP and AgcH, which bind monosaccharide, the loop between $\alpha 17$ and $\alpha 18$ (L3, site III) is longer and takes a different path to GacH. This occlude the rings two and three of acarbose, indicating that AgaE could bind oligosaccharides. For *MtbUgpB* and *EcoUgpB*, which both bind G3P, R377 from site II, would occlude the position of rings three and four of acarbose in GacH (Figure 5.18). The equivalent residues are on $\alpha 18$ in AgaE S390 and in GacH F368 but R380 from loop 3 in AgaE occupies the same space as R377 of the G3P binding proteins, perhaps indicating that AgaE could bind G3P (Figure 5.13). In the raffinose (trisaccharides) binding protein (RafE), the first ring of the ligand occupies the space of F67 on $\alpha 2$ (site I) in GacH; in addition, the loop between $\beta 2$ and $\alpha 2$ is shorter in RafE, to provide space for the raffinose substrate. AgaE has a structure similar to GacH in this area, and therefore probably does not bind raffinose. In disaccharides trehalose binding protein tlMBP, R363 on L3 occupies the space of ring three of the acarbose in GacH. However, the equivalent residue in GacH and AgaE is also arginine, but this residue occupies a different position in the complex, and thus no definitive argument can be made for the specificity of AgaE for trehalose. Taken together, these analyses suggest that AgaE may well bind G3P or malto-oligosaccharides like

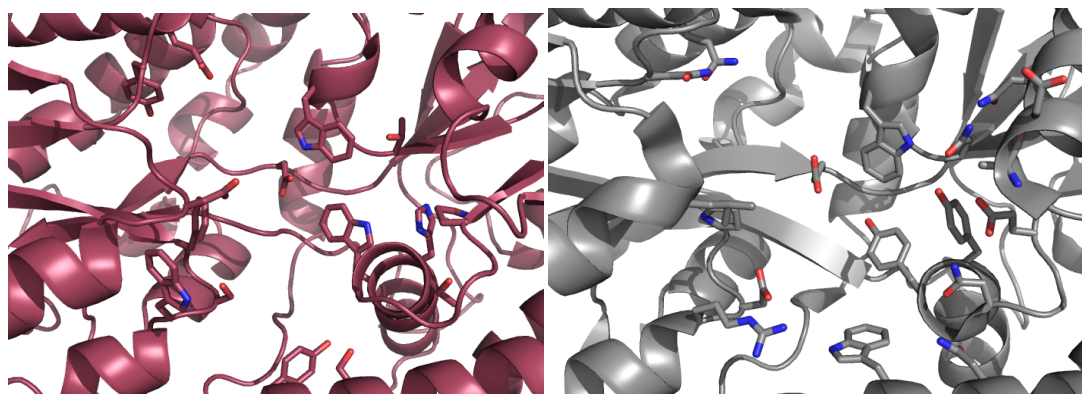
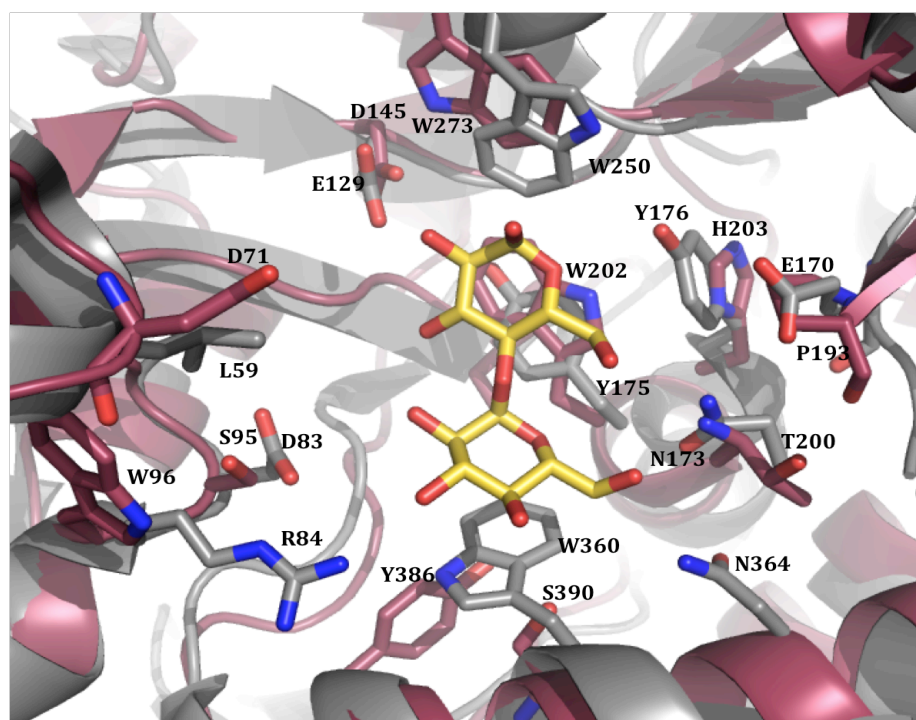
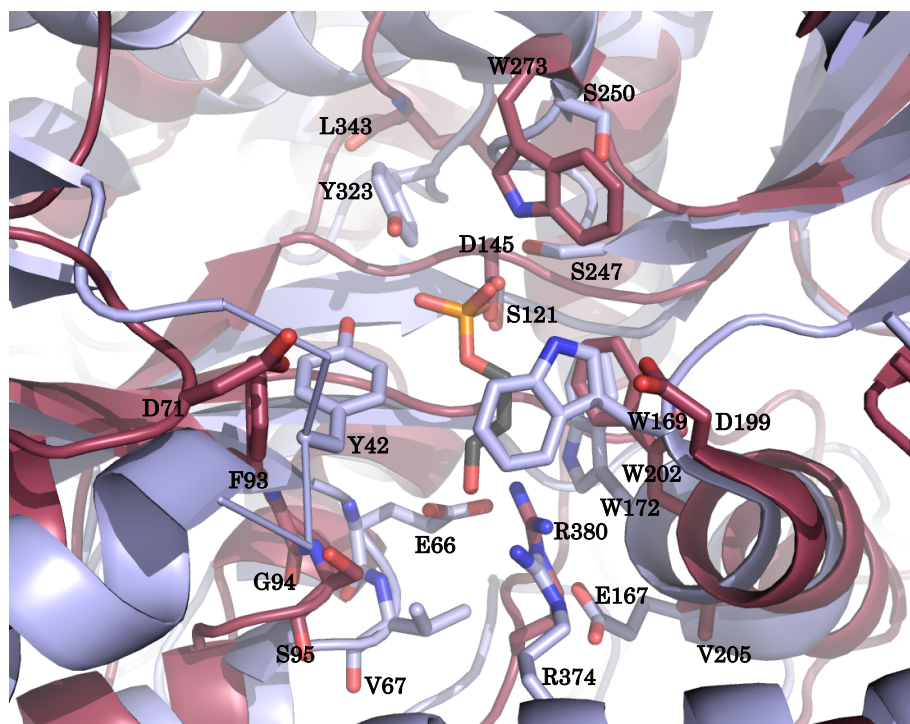
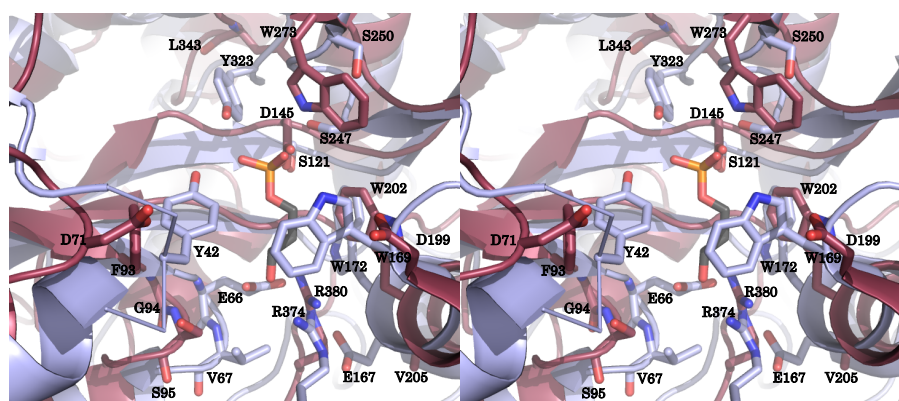


Figure 5.17 Superposition of AgaE and TvuCMBP binding sites. a) An overview picture of the binding residues of AgaE (raspberry) and TvuCMBP (gray) alignment. b) The binding site of *M. smegmatis* AgaE (left) and *Thermoactinomyces vulgaris* TvuCMBP (right) shown in the same orientation.



(a)



(b)

Figure 5.18 Superposition of AgaE and EcoUgpB binding sites. a) An overview picture of the binding residues of AgaE (raspberry) and UgpB closed form (brown). b) The binding site of *M.smegmatis* AgaE (left) and *E.coli* UgpB (right) shown in the same orientation.

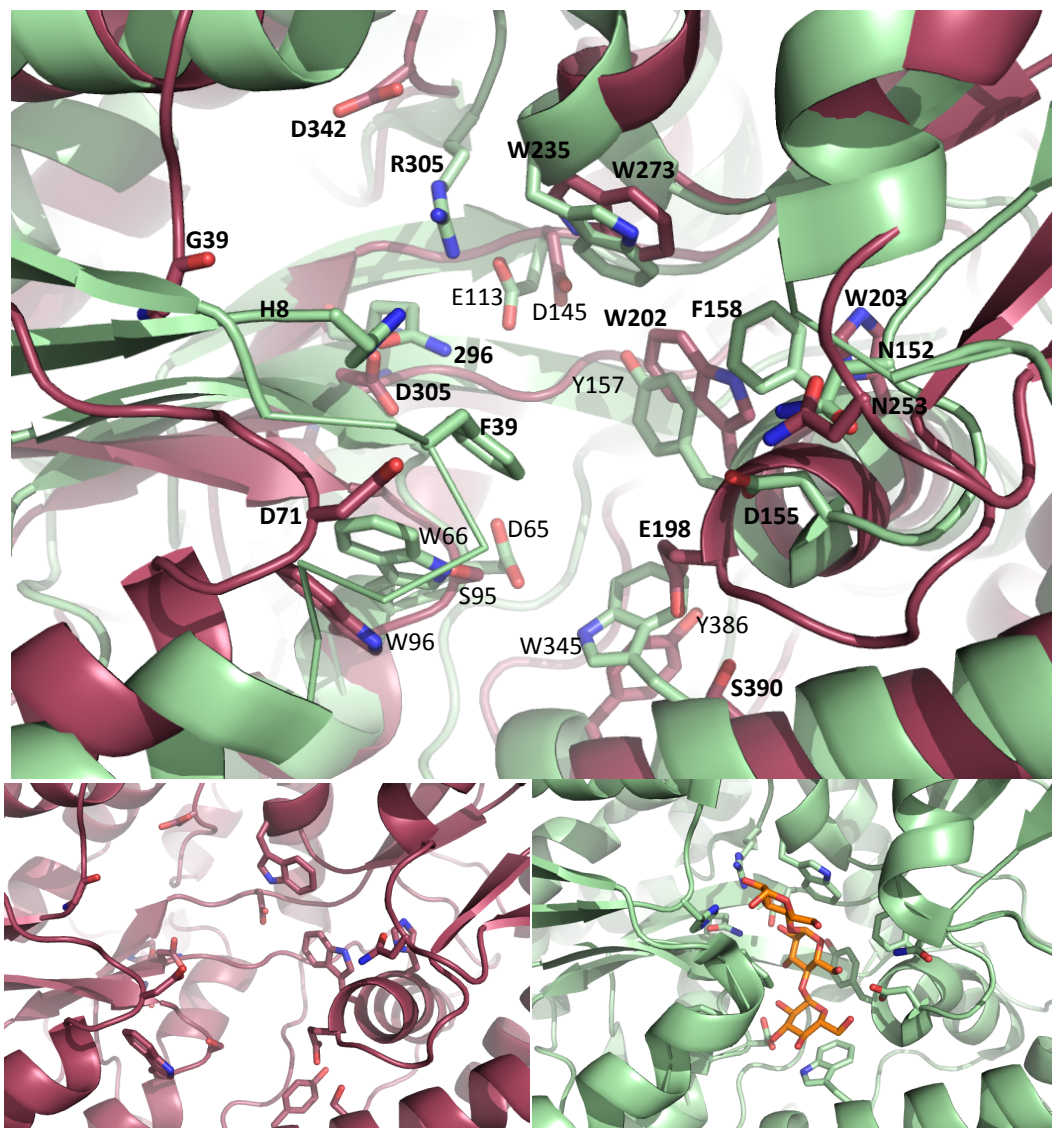


Figure 5.19 Superposition of AgaE and ttMtBP binding sites. a) An overview picture of the binding residues of AgaE (raspberry) and ttMtBP closed form (green). b) The binding site of *M.smegmatis* AgaE (left) and *Thermus thermophilus* ttMtBP (right) shown in the same orientation.

Chapter 6

AgaE – sugar complex

6.1 Introduction

As the structural analysis described in chapter five suggested that AgaE may well bind a mono or oligosaccharide substrate, a number of different experimental techniques were used to try and discover the substrate of AgaE. These included tryptophan fluorescence, circular dichroism (CD) assays and co-crystallization. For each of these experiments a large variety of different possible ligands were investigated, as the structural analysis were by no means certain in identifying possible substrates.

6.2 Binding Assays

6.2.1 Tryptophan Fluorescence Spectroscopy

In order to gain a clear idea of the potential rational substrate that binds to AgaE active site, tryptophan fluorescence was performed using 16 different sugars. This assay will help to detect any interaction that might occur between AgaE and any sugar, by measuring the alteration in the intrinsic fluorescence of the AgaE tryptophan residues, which produce a high signal when excited at a wavelength of 280 nm - 295 nm excitation. The tryptophan emission fluorescence depends on the surrounding solvent. For example, a decrease in the polarity in the solvent surrounding the tryptophan will result in a lower wavelength and an increase in the intensity of the fluorescence emission spectrum. Other aromatic residues also can produce fluorescence signals, such as tyrosine and phenylalanine. However, the emissions of these two residues are measured at lower wavelength 274 nm and 257 nm, respectively.

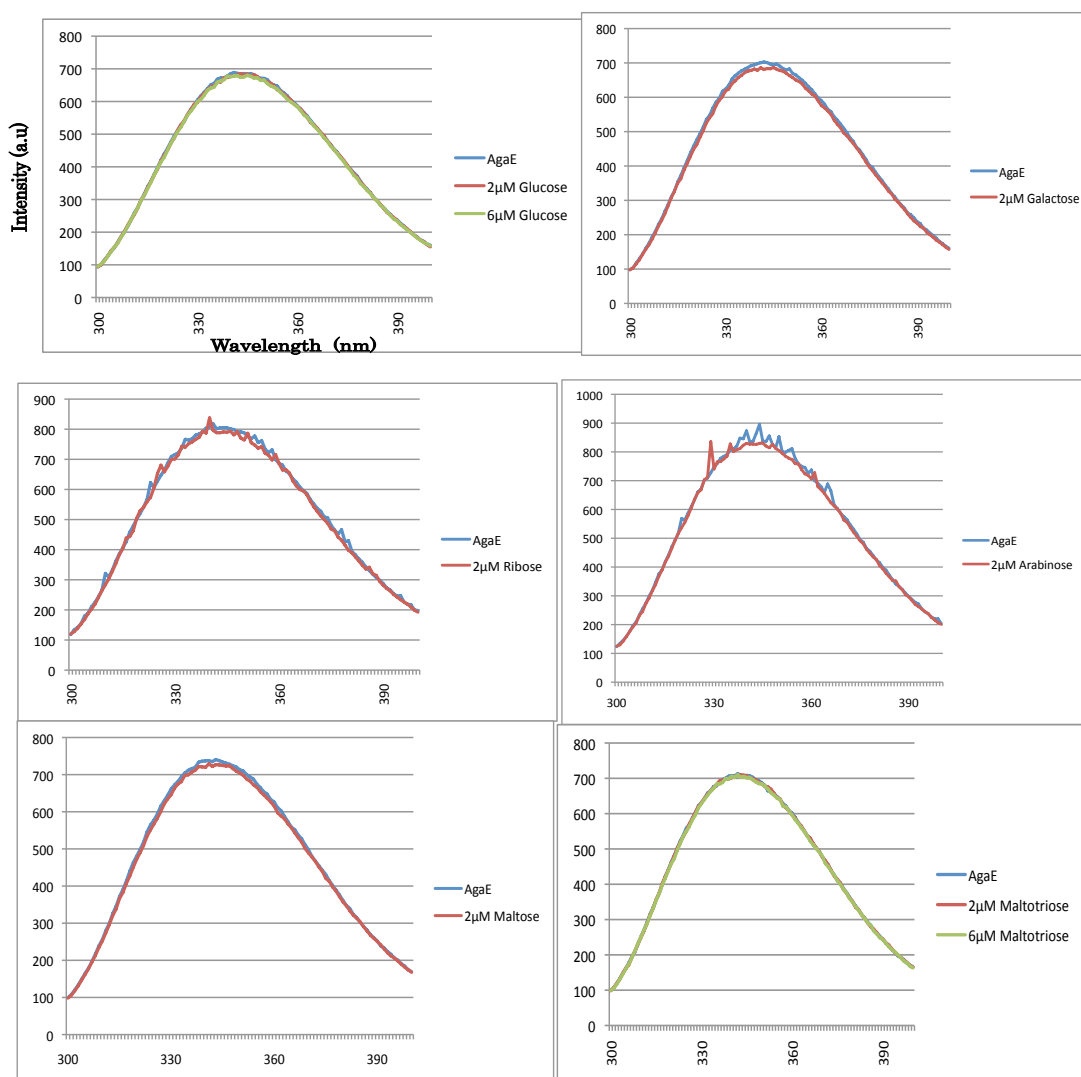
6.2.2 Methodology

The Cary Eclipse fluorimeter (Varian Ltd, UK) was used to measure the change in UV fluorescence of the intrinsic AgaE tryptophan residues. As binding of a sugar in the family of proteins usually results in the two domains moving closer together, it was hoped that this movement may give rise to a change in tryptophan fluorescence. The AgaE protein sequence contains 12 tryptophan and 11 tyrosine residues, respectively. The UV fluorescence of AgaE was measured using a spectrophotometer cuvette containing 0.2 μ M protein in 3 volumes of a solution of 10 mM Tris-HCl buffer pH 7.4 at 30°C. The AgaE excitation was measured at 280nm with a cut of 5 nm width and an emission wavelength of 300-400nm with a slit width of 20 nm.

AgaE had a maximum emission at about 370 nm with a 280 nm excitation. Then, 2 μM and 6 μM of each ligand, respectively, was added to the 3-volume solution containing 0.2 μM protein in similar buffer and the titration was measured with a 5 nm and 20 nm window for excitation and emission, respectively.

6.2.3 Results

The binding affinity for each sugar was tested separately. The sugars tested are shown in table 6.1. However, there was no indication of any ligand-protein interaction with all components that have been examined, as there was no change in the fluorescence emission of any of them.



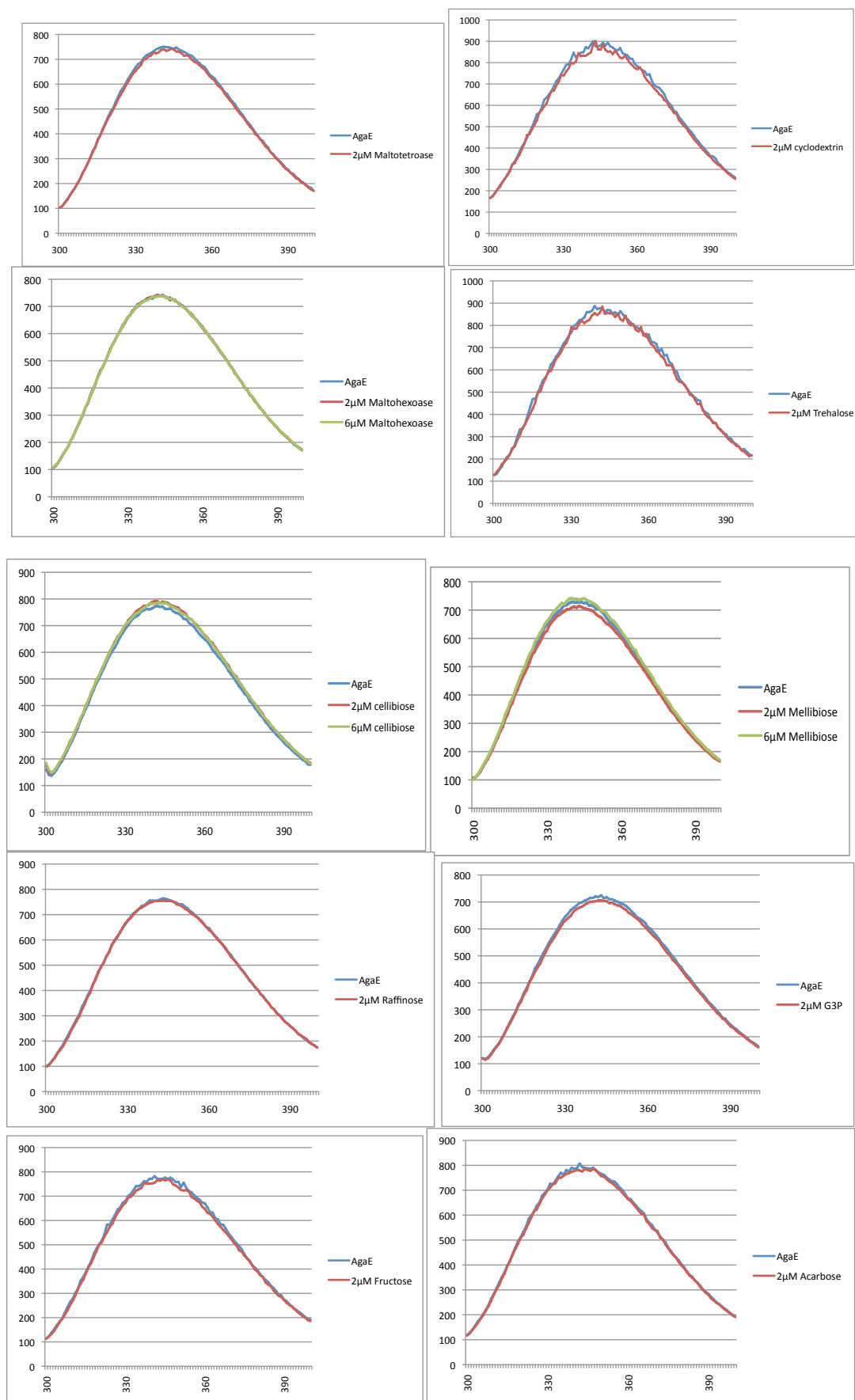


Figure 6.1 Traces of tryptophan fluorescent assay. A the 0.2mM AgaE in Tris –HCL pH 8.0 buffer, was excited at 280 nm, however, no emission was obtained upon sugar added.

6.3 Circular dichroism (CD)

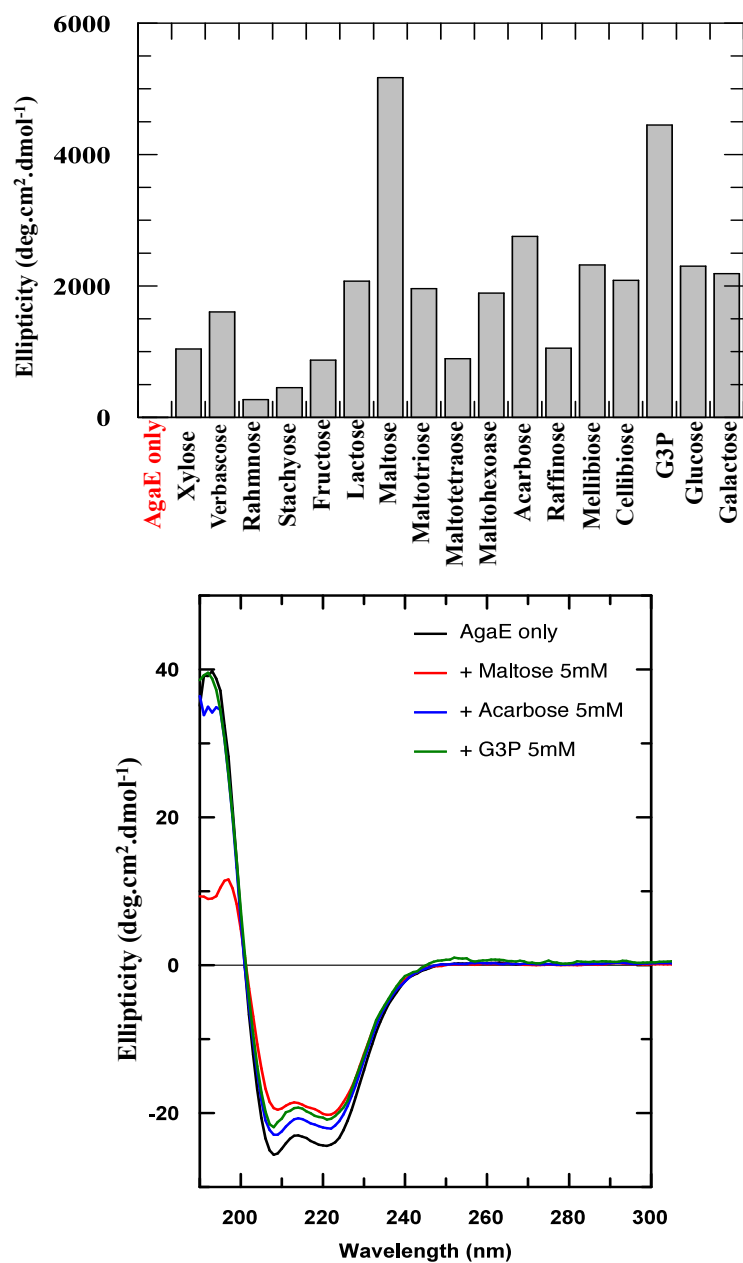
An alternative method was thus employed, to try to measure ligand binding. The alteration in the secondary structure of a protein can be measured using Circular dichroism (CD), which is produced by the interaction between protein molecules and UV circularly polarized light. As proteins are chiral molecules they absorb left and right circularly polarized light differently [154]. Furthermore, the presence of secondary structure (α -helices, β -sheets, turns) in the protein will produce a distinct CD signal [154]. Therefore, the alteration of the circular polarized light between left and right hands upon ligand binding are measured and this can result in a different CD signal as the protein structure changes on ligand binding. The CD spectrum of a protein is usually determined at a wavelength and ellipticity length scales appropriate to measure the new signal form the complex.

6.3.1 Methodology

The CD spectrophotometer (Jasco 715 model) was used to test the binding affinity of AgaE with several mono, di and oligosaccharides. The CD machine is provided with temperature cooler, nitrogen gas cylinder and sample changer. The output data was processed using an attached computer. The protein sample was prepared in a solution, which had no intrinsic CD signal at the wavelength used (250-190 nm). The protein and buffer concentrations were adjusted in order to produce a good signal. Therefore, 2UM of AgaE protein was prepared in 4ml of 2mM Tris-HCL buffer pH 8.0 only. For each cycle of protein-sugar binding affinity experiment, data were collected in the presence and absence of a single sugar at different concentrations. Thus, 190U1 of AgaE solution was added to a 1.0 mm cuvette and incubated for 30s before data collection. Then, a sugar of different concentrations (0.03, 1.0, 5.0 and 50 mM) was added into the protein solution and data was collected again. The CD signal was measured between 210 nm and 222 nm and at a temperature of 20+/- 0.5°C, CD speed time 50 nm/min with one spectrum accumulation.

6.3.2 CD Data analysis

Circular dichroism spectroscopy has been used to obtain information on the protein-ligand interaction by measuring the change in the environment of protein structure upon adding sugar in the solution. Firstly, the results revealed no significant indication upon adding 5 mM sugar concentration of the monosaccharide, such as glucose, galactose, xylose, and rhamnose, which suggest that AgaE is not a monosaccharide binding protein (Figure 6.3). Therefore they were not included in the lower concentration CD measurements. Secondly, binding of AgaE with tri or oligosaccharides, such as stachyose, verbascose, lactose, raffinose, and maltooligosaccharides, such as maltotriose, maltotetraose and maltohexoase, also show no significant change in the CD spectrum. However, raffinose does show some interaction with AgaE in only high concentration of sugar (50mM) (Figure 6.3). Furthermore, disaccharide sugars, such as maltose, trehalose, cellobiose and mellibiose, have been also attempted to measure their binding affinity to AgaE structure, and results revealed that AgaE structure conformation has been highly changed due to the addition of maltose more than other disaccharides. As the highest peak was indicated for maltose in 5mM sugar concentration followed by the CD signal peaks for glycerol 3-phosphate and acarbose, respectively, within the same concentration. To sum up, the results of CD experiment were performed first in 5mM sugar and the suspected binding sugar were attempted again in lower concentration (0.03mM) for higher affinity sugars, however, the CD signals from lower concentration sugar added were not significant as the 5 mM concentration, which caused a significant change in the structure occurred by adding 5mM maltose and 5 mM glycerol 3-phosphate (Figure 6.2).



Figures 6.2 The CD spectrum of AgaE only and in complex with 5mM maltose, acarbose and glycerol 3-phosphate. None of these sugars showed any indication of binding to the protein.

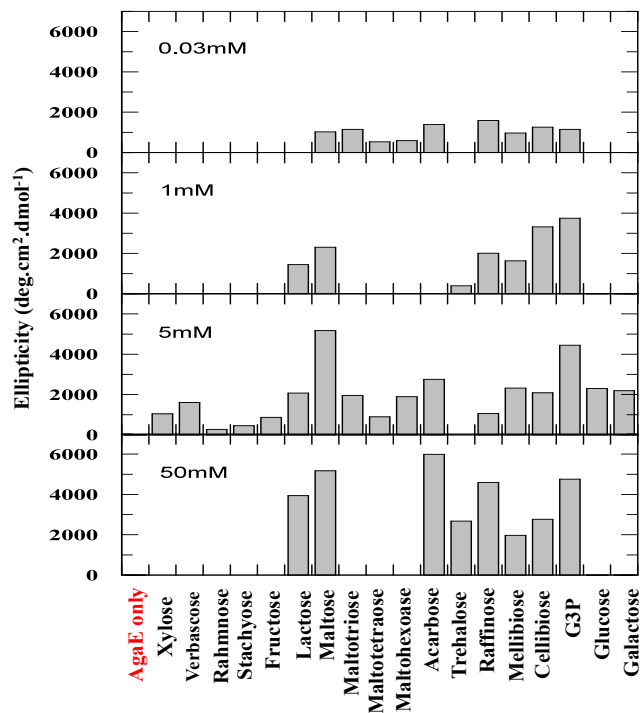


Figure 6.3 The CD spectrum representation of the AgaE in complex with different sugar concentrations 0.3 mM, 1mM, 5mM and 50 mM at 222 nm intensity. Alteration in the spectrums was observed with two sugars glycerol 3 phosphate and maltose of 5 mM concentration.

6.4 AgaE – sugar complex

6.4.1 Co-crystallization of AgaE with sugar

The structural comparison of AgaE with other solute binding proteins revealed that AgaE is likely to be a carbohydrate binding protein of the ABC transporter family. Therefore, attempts to co-crystallize AgaE with the most likely sugars were performed in order to determine the structure of AgaE with its bound sugar. AgaE protein was expressed and purified as described in section 4.3. A total protein concentration of 10 mg/ml was prepared in 10 mM of Tris-HCL buffer, pH 8.0 for co-crystallization experiment as described in section 4.4. The Hydra II robot with commercial crystallization screens (PACT, JCSG, PEG, Ammonium sulfate and Classic) was used to crystallize AgaE in presence of 50 mM and 300mM concentrations of various sugars. All plates were incubated at 17°C.

6.4.2 Crystallization results

The identification of successful hits was carried out by viewing each drop under a microscope. The crystallization plates were viewed after 24 hours, 3 days and 7 days of crystal growth. After 3 days of incubation, the AgaE protein started to form crystals under a number of different conditions in all screens. The trays were then left to equilibrate further, and the crystals grew bigger. (Tables 6.1 and 6.2)

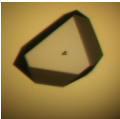

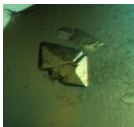
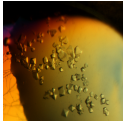
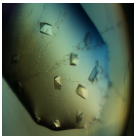


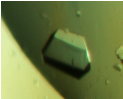
6.4.3 Data collection and processing of co-crystallized AgaE

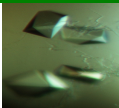

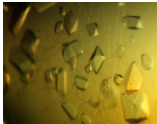
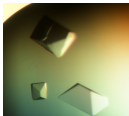
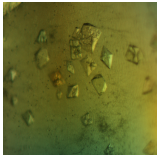
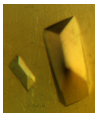

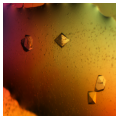
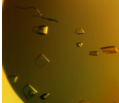

For each different potential AgaE complex, a single crystal was briefly washed in a cryoprotectant consisting of 25% ethylene glycol in the crystallization buffer and 5 mM added sugar, flash cooled to 100 K and stored in liquid nitrogen prior to data collection on the Diamond synchrotron light source. Crystals grown from the AgaE – G3P complex, were washed in 50% of glycerol instead of ethylene glycol.

Data were collected at the Diamond light source in Oxford (I03 beam line) as described in section 4.4.5 (Table 6.3). All datasets were processed using the Xia2 software at diamond and structures were determined using the molecular replacement method. The wild type AgaE model was used as a search model in the Phaser program [83].

6.4.4 Structure determination and analysis

All structures were rebuilt with several rounds of refinement cycles, after adding water molecules, ethylene glycol and PEG molecules to the related electron density in the map of each structure using Coot and Refmac5 [134]. However, there was no indication for any sugar bound in the cleft of the protein as no electron density for a potential ligand could be observed in this region. In addition, all structures had the same cell dimensions as the apo AgaE structure, and were all in a similar conformation.

AgaE sugar complex	Crystal	Crystallization condition	Resolution (Å)
AgaE-glucose		JCSG-D12 (0.04 M Potassium phosphate, 16% PEG 8000, 20% Glycerol)/ JCSG-H8 (0.2 M Na chloride, 0.1 M bis-Tris pH 5.5, 25 % PEG 3350	1.24/2.2
AgaE-galactose		JCSG-B9 (0.1 M citric acid pH 5, 20% PEG 6000	1.54
AgaE-fructose		JCSG-B9 (0.1 M citric acid pH 5, 20% PEG 6000	1.41
AgaE-fucose		JCSG-B9 (0.1 M citric acid pH 5, 20% PEG 6000	1.37
AgaE-Ribose		JCSG-H8 (0.2 M Na chloride, 0.1 M bis-Tris pH 5.5, 25 % PEG 3350	1.44
AgaE- maltose		JCSG-D12 (0.04 M Potassium phosphate, 16% PEG 8000, 20% Glycerol	1.38
AgaE-maltotriose		JCSG-D12 (0.04 M Potassium phosphate, 16% PEG 8000, 20% Glycerol	1.37
AgaE-maltotetraose		JCSG-D12 (0.04 M Potassium phosphate, 16% PEG 8000, 20% Glycerol	1.63

AgaE-maltohexoase		JCSG-D12 (0.04 M Potassium phosphate, 16% PEG 8000, 20% Glycerol)	1.34
AgaE-maltodextrin		JCSG-D12 (0.04 M Potassium phosphate, 16% PEG 8000, 20% Glycerol)	1.21
AgaE-Cyclodextrin		JCSG-B9 (0.1 M citric acid pH 5, 20% PEG 6000)	1.45
AgaE-maltopentose		PACT-A1 (0.1M SPG buffer pH 4, 25% PEG 1500)	1.57
AgaE-Acarbose		PACT-A1 (0.1M SPG buffer pH 4, 25% PEG 1500)	1.52
AgaE-trehalose		JCSG-D12 (0.04 M Potassium phosphate, 16% PEG 8000, 20% Glycerol)	1.33
AgaE-Raffinose		JCSG-D12 (0.2 M Na chloride, 0.1 M Phosphate-citrate pH 4.2, 20% PEG 8000)	1.64
AgaE-melibiose		JCSG-D12 (0.04 M Potassium phosphate, 16% PEG 8000, 20% Glycerol)/ PACT-A1 (0.1M SPG buffer pH 4, 25% PEG 1500)	1.34/1.28
AgaE-cellobiose		JCSG-B9 (0.1 M citric acid pH 5, 20% PEG 6000)/ PACT-A1 (0.1M SPG buffer pH 4, 25% PEG 1500)	1.7/1.89
AgaE-G3P		JCSG-B9 (0.1 M citric acid pH 5, 20% PEG 6000)	1.85

Tables 6.1 Photographs of the AgaE-sugar complex crystals. The tables represent crystals that were sent to Diamond for data collection with their crystallization conditions and data resolution.

DATA SET	AgaE-Glucose	AgaE-Galactose	AgaE-Fructose	AgaE-Fucose	AgaE-Ribose
Wavelength (Å)	0.9763 12.7	0.9763 12.7	0.97620 12.7	0.9763 12.7	0.9763 12.7
Energy (KeV)					
Space group	P 2 ₁ 2 ₁ 2 ₁	P 2 ₁ 2 ₁ 2 ₁	P 2 ₁ 2 ₁ 2 ₁	P 2 ₁ 2 ₁ 2 ₁	P 2 ₁ 2 ₁ 2 ₁
Unit cell parameters					
a (Å)	64.12	64.10	64.05	64.14	63.0
b (Å)	69.43	69.31	69.24	69.30	69.11
c (Å)	101.27	101.9	101.1	101.1	101.8
$\alpha = \beta = \gamma$ (°) =	90.0	90.0	90.0	90.0	90.0
Resolution range (Å)	30.6-1.24 (1.27-1.24)	69.31-1.54 (1.58-1.54)	34.6-1.41 (1.45-1.41)	50.6-1.37 (1.41-1.37)	53.5-1.44 (1.48-1.44)
Unique observation	126739 (8369)	66423 (4229)	87206 (6361)	583196 (6733)	80845 (5893)
Rmerge	0.030 (0.425)	0.041 (0.495)	0.037 (0.66)	0.040 (0.509)	0.043 (0.60)
Rpim	0.015 (0.340)	0.020 (0.298)	0.17 (0.30)	0.019 (0.336)	0.022 (0.290)
Completeness (%)	99.0 (90.1)	98.6 (96.5)	99.9 (100)	99.7 (97.0)	99.9 (99.9)
Anomalous completeness (%)	95.8 (70.8)	97.2 (78.6)	99.1 (99.8)	98.1 (87.9)	99.3 (98.6)
Multiplicity	5.8 (2.7)	6.2 (4.1)	6.5 (6.7)	6.1 (3.6)	6.4 (5.6)
Anomalous multiplicity	2.3 (2.3)	3.2 (2.0)	3.3 (3.3)	3.0 (1.7)	3.3 (2.7)
Mean (I)/σ(I)	23.8 (2.3)	20.9 (2.4)	21.1 (3.0)	18.8 (2.3)	18.6 (2.8)

DATA SET	AgaE- Maltose	AgaE- Maltotriose	AgaE- Maltotetraose	AgaE- Maltopentose	AgaE- Maltodextrin	AgaE- Cellibiose
Wavelength (Å)	0.9763 12.7	0.9763 12.7	0.9763 12.7	0.9763 12.7	0.9763 12.7	0.96861 12.7
Energy (KeV)						
Space group	P 2 ₁ 2 ₁ 2 ₁	P 2 ₁ 2 ₁ 2 ₁	P 2 ₁ 2 ₁ 2 ₁	P 2 ₁ 2 ₁ 2 ₁	P 2 ₁ 2 ₁ 2 ₁	P 2 ₁ 2 ₁ 2 ₁
Unit cell parameters						
a (Å)	63.83	64.00	64.27	64.35	63.05	63.77
b (Å)	69.18	69.13	69.35	69.46	69.37	69.26
c (Å)	101.65	101.5	101.4	101.51	101.23	101.49
$\alpha = \beta = \gamma$ (°) =	90.0	90.0	90.0	90.0	90.0	90.0
Resolution range (Å)	31.92-1.38 (1.42-1.38)	24.17-1.37 (1.41-1.37)	57.24-1.63 (1.68-1.63)	30.56-1.57 (2.50 -1.57)	18.57-1.21 (1.24-1.21)	23.58-1.7 (1.74-1.7)
Unique observation	91282 (5890)	94948 (6928)	56719 (4134)	60995 (4229)	137346 (9795)	49655 (74362638)
Rmerge	0.038 (0.545)	0.039 (0.64)	0.082 (0.91)	0.041 (0.495)	0.03 (0.53)	0.1 (0.668)
Rpim	0.018 (0.301)	0.019 (0.298)	0.040 (0.41)	0.020 (0.298)	0.015 (0.341)	0.053 (0.327)
Completeness (%)	98.9 (87.8)	99.8 (99.9)	99.8 (100)	98.6 (96.5)	99.7 (97.6)	99.0 (98.7)
Anomalous completeness (%)	97.6 (80.6)	98.7 (99.7)	96.6 (99.1)	97.2 (78.6)	98.4 (89.4)	93.5 (97.2)
Multiplicity	6.2 (4.3)	6.4 (6.3)	6.2 (6.6)	6.2 (4.1)	6.2 (3.8)	5.8 (6.1)
Anomalous multiplicity	3.2 (2.1)	3.3 (3.2)	3.1 (3.3)	3.2 (2.0)	3.1 (1.8)	2.9 (3.0)
Mean (I)/σ(I)	20.2 (2.5)	19.6 (2.9)	15.4 (2.6)	20.9 (2.4)	23.2 (2.3)	9.7 (2.0)

DATA SET	AgaE- Maltose	AgaE- Cyclodextrin	AgaE- Trehalose	AgaE- Acarbose	AgaE- Raffinose	AgaE- Mellibiose
Wavelength (Å)	0.9763 12.7	0.9763 12.7	0.9763 12.7	0.9763 12.7	0.96861 12.7	0.96861 12.7
Energy (KeV)						
Space group	P 2 ₁ 2 ₁ 2 ₁	P 2 ₁ 2 ₁ 2 ₁	P 2 ₁ 2 ₁ 2 ₁	P 2 ₁ 2 ₁ 2 ₁	P 2 ₁ 2 ₁ 2 ₁	P 2 ₁ 2 ₁ 2 ₁
Unit cell parameters						
a (Å)	63.93	64.22	64.07	64.2	64.31	64.42
b (Å)	69.28	69.34	69.2	69.2	69.37	69.39
c (Å)	101.26	101.14	101.26	101.0	101.56	100.86
$\alpha = \beta = \gamma$ (°)	90.0	90.0	90.0	90.0	90.0	90.0
Resolution range (Å)	28.61.34 (1.37-1.34)	64.22-1.45 (1.49-1.45)	27.94-1.33 (1.37-1.33)	34.59-1.52 (1.56-1.52)	27.50-1.64 (1.68-1.64)	30.26-1.34 (1.37-1.34)
Unique observation	101503 (7438)	79539 (5656)	101723 (6537)	69102 (4998)	56357 (4087)	101930 (7479)
Rmerge	0.035 (0.615)	0.047 (0.611)	0.027 (0.401)	0.04 (0.6)	0.077 (0.675)	0.037 (0.675)
Rpim	0.016 (0.29)	0.022 (0.31)	0.013 (0.31)	0.022 (0.34)	0.037 (0.313)	0.017 (0.314)
Completeness (%)	100.0 (100.0)	98.7 (96.1)	98.6 (87.5)	99.1 (98.6)	99.9 (99.9)	99.9 (99.9)
Anomalous completeness (%)	99.3 (3.2)	97.7 (95.4)	95.5 (71.8)	94.9 (94.1)	98.5 (99.3)	98.7 (98.5)
Multiplicity	6.5 (6.5)	6.5 (6.1)	5.9 (3.1)	4.4 (4.5)	6.3 (6.5)	6.4 (6.3)
Anomalous multiplicity	3.2 (3.2)	3.3 (3.1)	3.0 (1.4)	2.2 (2.2)	3.2 (3.2)	3.2 (3.1)
Mean (I)/ σ (I)	22.0 (3.0)	17.9 (2.9)	26.8 (2.4)	19.9 (2.6)	13.0 (2.5)	22.7 (2.7)

Tables 6.2 Data collection statistics for all AgaE-sugar complex crystals, values in parentheses refer to the high resolution shell.

6.5 Determination of more crystal structures of AgaE potential complex

The CD spectra assay suggested that AgaE has some affinity for maltose, glycerol-3-phosphate and acarbose, and for these sugars, crystals were grown in a number of different conditions. Therefore, more crystals of AgaE grown in the presence of these three sugars, and in different crystallization conditions, were sent to Diamond for data collection, in case the different conditions favoured complex formation. Surprisingly, data were collected in different $P 2_12_12_1$ derivative space group and different cell dimensions (Table 6.3). For the AgaE in complex with maltose, the crystal grew in PACT-b1 condition and data were collected to 2.85 Å and in space group $P 2_12_12_1$ and cell dimensions $a=64.5$, $b=107.0$ and $c=113.4$. For the AgaE in complex with acarbose, two data were collected from crystals obtained with different crystallization condition. The first crystals grew in the same crystallization conditions as the AgaE - maltose complex crystal, and data were collected to 2.34 Å and in different space groups $P 2_12_12$ and different cell dimensions $a= 64.7$, $b =112.5$ and $c= 53.3$ Å. The second crystal grew in similar crystallization condition of the apo AgaE and data were collected to 1.22 Å and similar space group $P 2_12_12_1$ and cell dimensions $a= 62.3$, $b = 68.3$ and $c= 100.4$ Å. Data statistics for these three datasets are shown in table 6.3. Thus, structures were determined using the molecular replacement Phaser using the apo AgaE structure with water molecules deleted as a search model and refined using *refmac5* (table 6.4), as mentioned in section 6.1.3.

DATA SET	AgaE-Maltose	AgaE-Acarbose	AgaE-Acarbose
Crystallization condition	PACT-b1 (0.1M MIB buffer pH 4, 25% PEG 1500)	PACT-b1 (0.1M MIB buffer pH 4, 25% PEG 1500)	JCSG -A9 (0.2 M Ammonium chloride pH 6.3, 20% PEG 3350)
Wavelength (Å)	0.9763	0.9763	0.9763
Energy (KeV)	12.7	12.7	12.7
Space group	P 2 ₁ 2 ₁ 2 ₁	P 2 ₁ 2 ₁ 2 ₁	P 2 ₁ 2 ₁ 2 ₁
Unit cell parameters			
a (Å)	64.5	64.7	62.3
b (Å)	107.0	112.5	68.3
c (Å)	113.4	53.3	100.4
$\alpha = \beta = \gamma (^{\circ}) =$	90.000	90.000	90.000
Resolution range (Å)	56.7-2.85 (2.92-2.85)	24.9-2.34 (2.40-2.34)	53.0-1.22 (1.25-1.22)
Unique observation	18948 (1362)	16958 (1243)	120904 (6801)
Rmerge	0.23 (0.8)	0.081 (0.54)	0.041 (0.6)
Rpim	0.012 (0.351)	0.04 (0.25)	0.02 (0.3)
Completeness (%)	99.0 (99.0)	100 (100)	95.0 (74.0)
Anomalous completeness (%)	98.5 (94.4)	98.0 (98.6)	93.2 (66.5)
Multiplicity	6.2 (6.4)	6.3 (6.2)	6.4 (5.5)
Anomalous multiplicity	3.2 (3.2)	3.3 (3.2)	3.3 (2.6)
Mean (I)/σ (I)	6.6 (2.1)	15.1 (3.1)	19.4 (2.6)

Table 6.3 Data collection statistics for all data sets of AgaE crystals in complex with acarbose and maltose, values in parentheses refer to the high resolution shell.

Model	Crystal I	Crystal II	Crystal III
Resolution (Å)	2.85 Å	1.34 Å	1.22 Å
Number of reflections	17932	25635	27836
Protein molecules per asymmetric unit	2	1	1
Number of atoms	6946	3220	3111
Number of waters	393	140	113
Number of PEG	1	1	1
Ramachandran favored (%)	98.4	98.1	97.3
Ramachandran outliers (%)	0	0.3	0
Poor rotamers (%)	1.9	0.6	1.0
RMSD bond (Å)	0.006	0.007	0.011
RMSD angle (°)	1.06	1.08	1.34
Average B-factors (Å²)			
Main chain (Å²)	20.7	38	14.4
Side chain (Å²)	22	55	16.6
PEG	31.35	33	-
R-factor	0.15	0.22	0.19
R-Free	0.21	0.29	0.25
Molprobit score	0.93	0.71	0.99
	100 th percentile	100 th percentile	100 th percentile

Table 6.4 The refinement statistics for three-crystal structures with potential sugar bound.

6.6 Structures analysis

6.6.1 AgaE structure with potential binding of acarbose

Two data sets were collected for AgaE in complex with acarbose, from different crystallization conditions JCSG-a9 and PACT-b1. The second data (b1) has been collected in different space group (Table 6.3) and both are in different cell dimensions than the native crystal. Crystal structures were determined using phaser and refined through several cycles using Refmac5. Both structures superimpose well with the apo structure and with each other with r.m.s.d value of about 0.4Å (Figure 6.4). This indicated that no significant movement of the molecule was made upon alteration of the cell dimensions and different packing in the crystal. However, the binding site of the potential AgaE and acarbose complex structure has revealed some poor density for only one ring of the acarbose compound. This density was refined as the acarvision ring of acarbose as attempts failed to fit a glucose ring (Figures 6.5, 6.6). However, structure comparison of AgaE in complex with acarbose with other acarbose binding proteins, such as MalE from *E.coli* and GacH from *S.glaucescens*, revealed that acarbose binds to MalE and GacH via the maltose moiety of the protein, and in both structures the reducing rings of maltose are bound to the inner moieties of maltose, however, in AgaE binding site, the acarvision ring of acarbose showed to be bound tightly to the internal maltose moiety instead of the glucose ring of its maltose. This might be due to the presence of PEG molecule that is bound tightly, which prevents the acarbose from binding in the same way it does in solution (Figures 6.7 and 6.8).

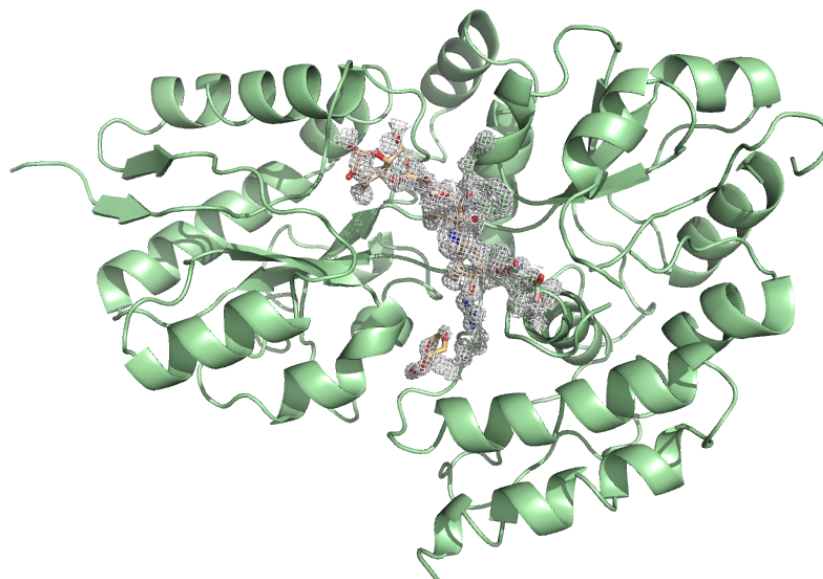


Figure 6.4 The 3D structure of AgaE (green) in complex with acarbose (wheat) and PEG molecule with electron densities.

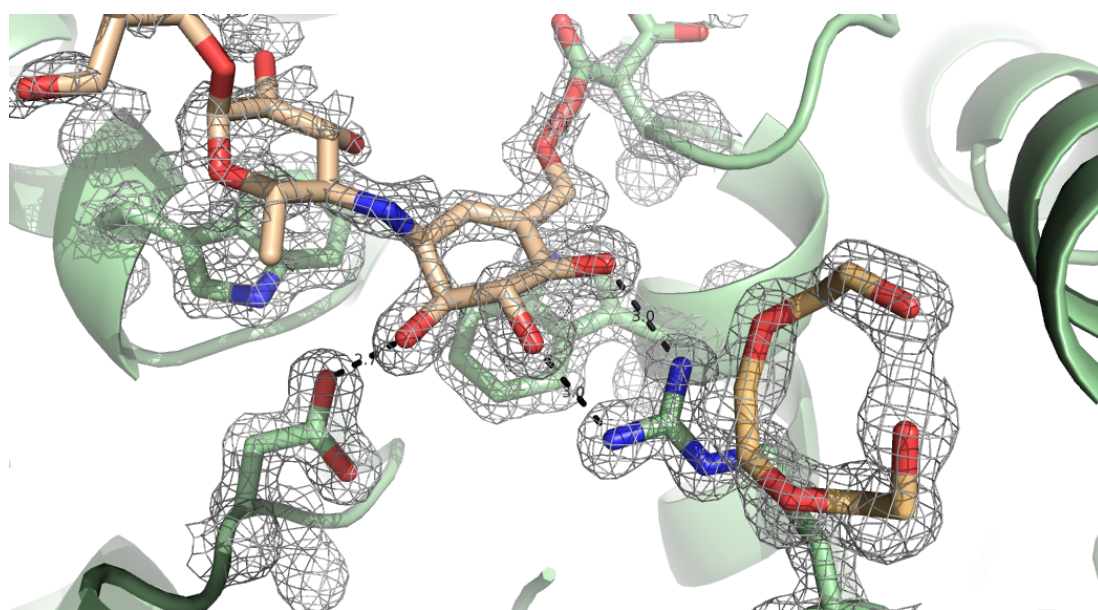


Figure 6.5 The binding site of AgaE with the acarbose bound. Residues that coordinate the acarviosin ring of acarbose are shown as stick. PEG molecule is shown as yellow stick.

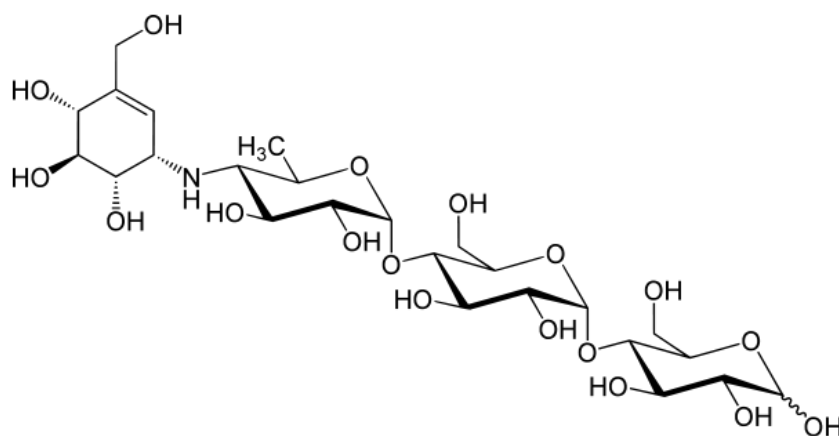
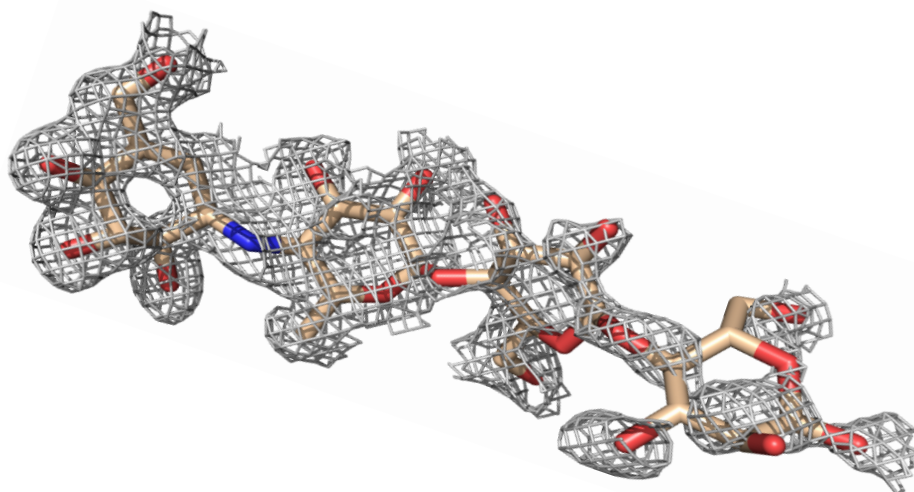
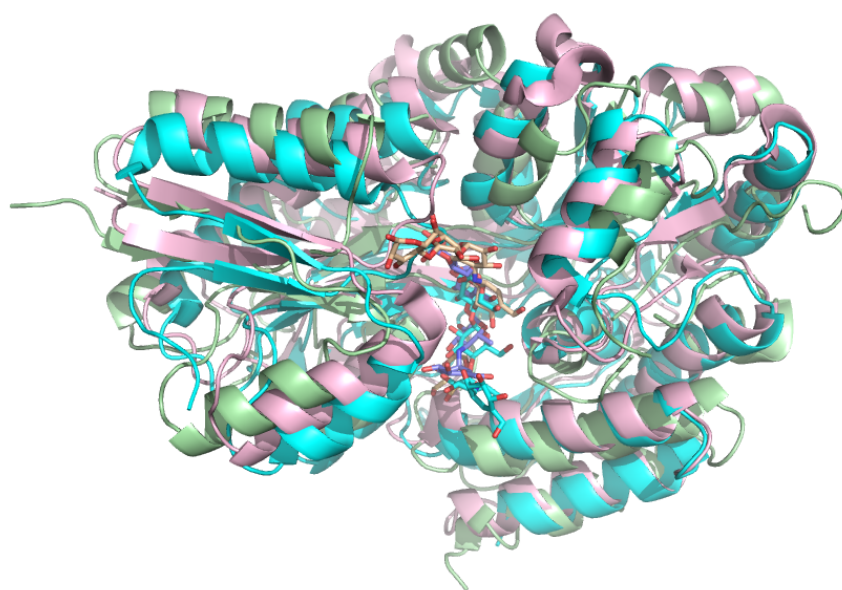
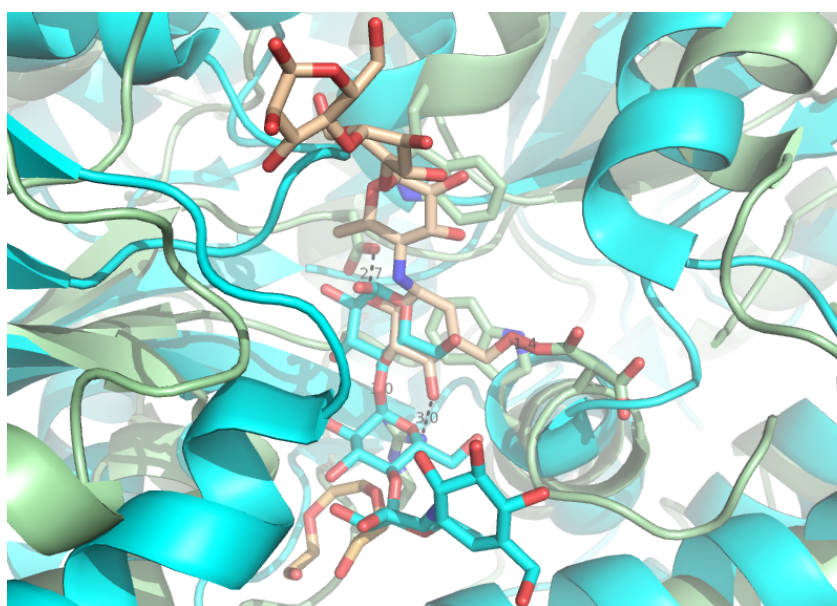


Figure 6.6 Acarbose structure with electron density contoured to 0.6 σ .



(a)



(b)

Figure 6.7 Superposition of 3D structures of AgaE (green), GacH (cyan) and MalE (pink) in complex with acarbose. B) The binding site of the superimposed structures AgaE (green) and GacH (cyan) in complexes with acarbose. Acarbose and PEG molecule in AgaE are shown as wheat stick and coordinated residues as green stick.

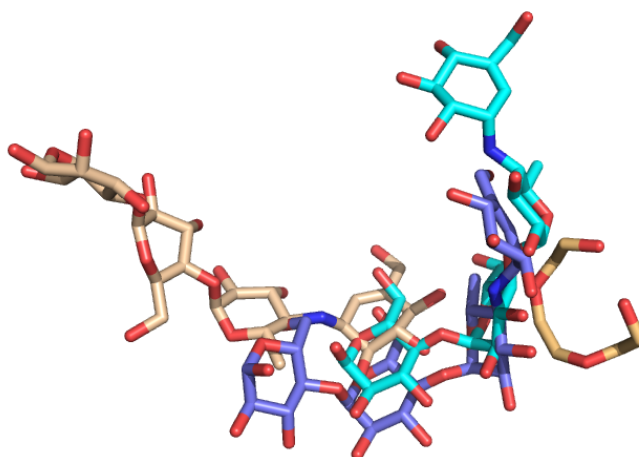


Figure 6.8 Superposition of bound acarbose from the binding sites of AgaE (wheat), GacH (cyan) and MalE (slate). The PEG molecule that occupied the acarbose binding site is shown as yellow stick.

6.6.2 AgaE Structures with potential binding of G3P and maltose

The binding site of AgaE structure in complex with G3P revealed no density for G3P, which indicates that G3P did not bind to AgaE. This might be due to the binding of ethylene glycol molecule in the one of the possible glycerol binding site. To test this, AgaE was co crystallized in ammonium sulfate precipitant screen, however, single crystals grew from the trials and washed with 50% of glycerol and stored in liquid nitrogen for data collection experiment. Unfortunately, data have not been collected due to the poor diffraction from the crystal.

Moreover, the binding site of AgaE in complex with maltose has not revealed any clear density for maltose, despite the movement in the side chains that are thought to be involved in the maltose binding.

6.7 Final discussion

The nutritional resource that *M.tuberculosis* needs for its growth was first to be studied intensively since its discovery [28]. Examples of these molecules are fatty acids, amino acids, carbohydrates and alcohol. Since the emergence of molecular biology the use of one of these small molecules were examined by testing the measurements of oxygen used by the cell [155]. Moreover, this necessary nutrition is believed to be transported into the cells of mycobacterium by proteins that attached to the cell membrane. These proteins are called transporters. The vast majority of these transporter lipoproteins are still unknown; even with the progress of genetic analysis methods [156, 157], especially in *M.tuberculosis* and *M.smegmatis*. Also, it has been shown that mycobacterium utilize the glyoxylate cycle to survive [158]. These findings indicated that the pathogenic mycobacterium uses lipids as the carbon supply throughout the infection. In addition, proteins that are responsible for the transport of disaccharides have been shown to be encoded by genes that are essential for the growth of mycobacterium[75]. This observation suggests that mycobacterium might use lipids instead of carbohydrates during host attack.

The non- pathogenic *M.smegmatis* is used as a fast growing model organism for the pathogenic *M.tuberculosis* to understand their physiological and pathogenesis aspects. Several studies have emerged in order to shed light on the transport system of both *M.tuberculosis* and *M.smegmatis*. There has also been some focus on the

nutrients transported and used for the carbon source. Most studies made to date have centered around the identification and characterization of some solute and amino acid transporters, based on homology with other bacterial transporters. Examples of these transporters are the phosphate and sulphate uptake proteins from *M.smegmatis* [159]. The composition of the mycobacterium cell wall has an effect on the solute transport in and out of the cell, which in turn has an effect on the life of mycobacterium and its pathogenicity. Thus the project described in this thesis has focused on different targets of lipid-anchored protein with different putative functions in *M.smegmatis*. The 3D structure of one target (Msmeg_0515) has been successfully determined out of 8 target proteins studied.

Structural and biochemical analysis of Msmeg_0515 (AgaE) suggested that it is most likely to be a maltose binding protein of the ABC transport system. Previous bioinformatics study on the sugar transporters of *M.smegmatis* suggested that AgaE is a α -galactoside sugar binding protein [3], however, attempts to test the binding affinity of such α -galactosides sugars, such as raffinose, melibiose, stachyose and verbascose, using two different assays have failed to indicate any significant interaction with AgaE. Also, a co-crystallization of AgaE in the presence of these four sugars revealed an open form structure with no indication for electron density of any bound sugar. These findings suggest that AgaE is not a α -galactoside sugar binding protein. Furthermore, ABC solute binding proteins usually have a high affinity of > 1 μ M to its natural substrate, and the 1mM sugar concentration used in the co-crystallization experiment, should be sufficient for binding. In the case of AgaE a > 50 mM concentration of sugar was used, however, all crystal structures that have been determined were in an open form with no sugar bound in the binding site. Although, the initial CD spectrum showed the highest binding affinity for 5mM maltose, G3P and acarbose to AgaE, more data is required to support these results, such as an ITC experiment.

The AgaE structure is most similar to other malto - oligosaccharide structures, and a sequence based structural alignment shows that the binding residues Y182 and W254 of GacH, that protrude from the C-domain (site II) and form hydrophobic contacts with the ligand, are conserved in all other malto-oligosaccharide structures. These two residues corresponded to the W202 and W273 residues in AgaE; W172 and S250 in *E.coli* UgpB (G3P binding protein) and C182 and W257 in *Thermococcus litoralis* MBP (trehalose binding protein). These two residues are suggested to be

critical for substrate specificity in these structures and thus mutagenesis experiments could be attempted on these two residues in AgaE to test its role for binding malto-oligosaccharide.

Chapter 7

Structural studies on phosphoglucose isomerase (PGI) protein mutants from *Pyrococcus furiosus*.

7.1 Introduction

In parallel with the studies on proteins from *M.smegmatis* reported in this thesis, the effect of mutating residues in the model enzyme of phosphoglucose isomerase (PGI) from *Pyrococcus furiosus* was investigated.

Although mutagenesis has long been used as a technique for studying the active site of enzymes, attempts to rationally design enzymes with improved activity, have not been very successful, despite obvious biotechnological applications [160, 161]. There are some examples where random mutagenesis has increased the activity of enzymes, but again these are limited [162-164]. More success has been seen in altering the substrate specificity of enzymes, [162] but in many cases these mutant enzymes show lower activity than the wild type.

An alternative method to identify possible candidate residues for mutation has been proposed [165]. The correlation between sequences in proteins belonging to the same enzyme family is mapped [162]. Residues that show a high correlation of a similarity can be identified, and then subjected to mutagenesis to see if the activity of the enzyme can change.

To test this correlated mutagenesis theory, the sequences of the cupin superfamily enzymes (of which *Pyrococcus furiosus* phosphoglucose isomerase (*PfPGI*) is a member) were aligned and a number of sites identified in the sequences, where the type of residue present at one position was correlated with the type of residue at another position (John Raedts and Jasper Akerboom, Wageningen University). Using this procedure, a number of correlated key residues were identified in *PfPGI* as a target for mutagenesis experiments to investigate their role in enzyme function, through activity measurements. The correlated positions identified were P132 and Y133 in the *PfPGI* sequence. All possible mutants were made at these two positions and, surprisingly, two of these mutant PGIs showed increased levels of activity compared to the wild type (Figure 7.2). The aim of this study was therefore, to determine the 3D structures of these PGI mutants and two control mutants, to investigate the alteration in enzyme activity in terms of the protein structure.

7.2 *P.furiosus* Phosphoglucose isomerase (PGI) structure

Phosphoglucose isomerase (PGI) is a catalytic enzyme found in several organisms including *P.furiosus*, which is a hyperthermophilic organism of archaea [166, 167]. PGI functions as reversible sugar catalytic enzyme, that converts fructose 6 phosphate (F6P) to Glucose 6 phosphate (G6P) [168, 169]. The structures of PfuPGI in complex with its substrate (F6P) and inhibitor (5PAA) were determined to 2.0 Å [170, 171] (Figure 7.1). *PfuPGI* structure composed of two monomers of 21.5 kDa with 189 amino acids each and function as homodimer by gel filtration [172]. Although, *PfuPGI* shares similar structure of other organisms PGIs, its primary sequence shows no homology [173, 174]. Crystal structure analysis of *PfuPGI* revealed that, its fold belongs to the cupin superfamily that is formed of two separate β -sheets; each contains three β -strands and six β -strands respectively [175]. Cupin superfamily is a widespread family involved in different functions, such as oxidoreductases and isomerases.

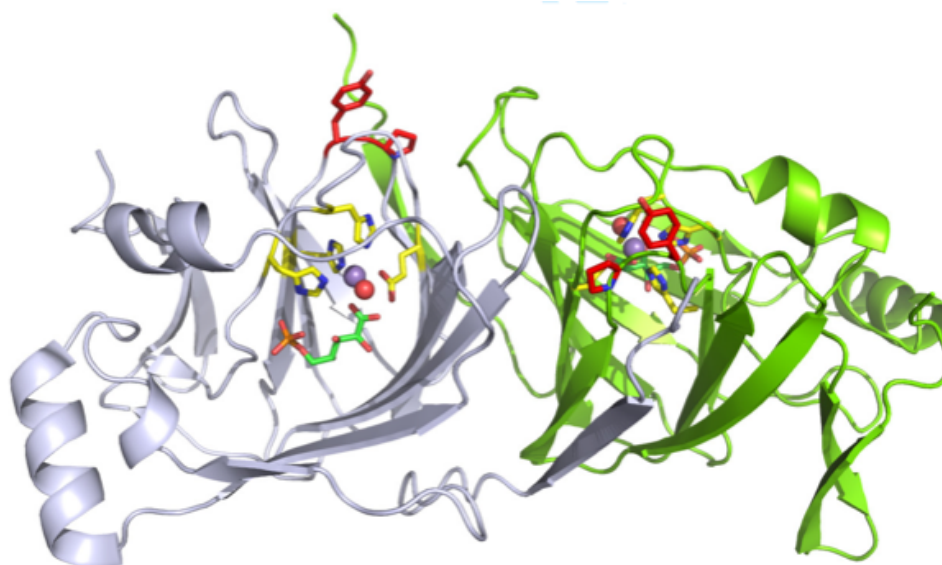


Figure 7.1 Cartoon representation of the dimer (white and green subunits) of the wild type PfuPGI $Mn^{+2}/5PAA$ 3D-structure (PDB code: 1X7N). The correlated amino acid pair “PY” is indicated in red. Shown in yellow are those residues involved in metal ion (purple sphere) binding, including a water molecule (red sphere). The inhibitor 5-phospho-D-arabinonate (5PAA) is shown as a stick model (green).

7.3 Active site and function of *PfuPGI*

The main function of *PfuPGI* is to catalyze the reverse conversion of fructose 6 phosphate and glucose 6 phosphate [168, 169], this mainly occurs through two processes, either by hydride shift directly or via a process called cis-enediol intermediate [171]. The first process, was based on a hydrogen atom that transfers between the C1 and C2 of fructose, however, this process was not observed in NMR experiment [171]. The second process where the hydrogen transfer occurs by process called cis-enediol intermediate. In this process a proton is transferred between the two carbons via a negative charged amino acid (E197) through chemical interaction with the solvent.

The active site of *PfuPGI* is located in the core of β -barrel fold with three rings of histidine (H88, H90 and H136) atoms and glutamate (E97) in contact with bound metal ion (12,13).

7.4 The identification of possible PGI mutants

A *PfPGI* library has been generated based on predictions made using the Comulotor CMA algorithm, as previously described [162]. The refined structure-based multiple sequence alignment of the cupin super-family, containing a total of 1711 sequences was used. The amino acids with the highest pair-wise correlated mutation score were Pro132 and Tyr133 in the *PfPGI*. These two residues are placed in a conserved structural loop across the family [162] (Figure 7.2). These two residues have been mutated into three different double mutants, (P132A, Y133G), (P132R, Y133G), (P132A, Y133D), and one with single mutation (P132V). Functional activity for these four mutants *PfPGI* has been performed by Raedts, J *et al*, (Wageningen University) and revealed increasing in the activity of *PfPGI* in the case of AG and RG, and less activity for the *PfPGI* with single mutation VY [162] (Figure 7.2). As these two mutated residues are not in contact with the active site residues, the changing in the activity of *PfPGI* mutants is not clear, thus, attempts to crystallize *PfPGI* mutants to determine their 3D structures were carried out in order to investigate these mutants and their effect on the activity at the atomic level.

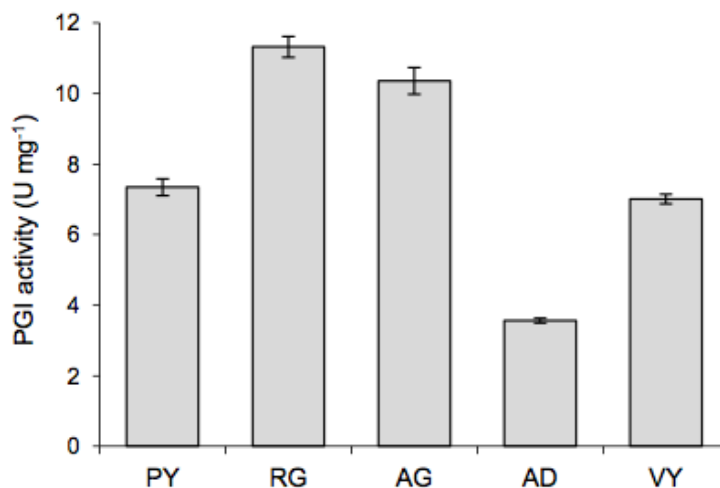


Figure 7.2 Specific activity of wild type PfPGI (PY) compared to selected high occurrence/activity mutants RG and AG, and low occurrence/activity mutants AD and VY. Manganese was used as co-factor and added via titration (not published).

7.5 Overexpression of *PfuPGI* mutants

The gene that encoded *PfPGI* mutants was cloned into pET24d by Raedts, John et al, (Wageningen University), as described previously [169]. Glycerol stocks for each *PfPGI* mutant was provided and used to inoculate primary cultures in LB medium supplemented with 50 $\mu\text{g ml}^{-1}$ kanamycin in a 37°C shaker. The overnight culture was used to inoculate (0.2% v/v) sterile glass tubes containing 10 milliliter LB/Km medium. When the optical density of the culture reached $A_{600} = 0.5$, gene expression was induced by addition of 0.1 mM isopropyl-1-thio- β -D-galactopyranoside (IPTG). Growth was continued overnight at 37°C, after which the cells were harvested by centrifugation (4,600 x g for 15 min). Pelleted *E.coli* cells were resuspended in 20 mM Tris- HCl buffer (pH 8.0) and disrupted by sonication. Cell debris was removed by centrifugation (16,000 x g for 15 min). *E.coli* proteins were denatured by heating the cell free extract at 65°C for 30 min, and removed by centrifugation (16,000 x g for 15 min). The result was a heat-treated cell free extract containing mainly PfPGI. Its purity was checked by SDS-PAGE. Protein concentrations were determined by Coomassie Brilliant Blue G250, using bovine serum albumin as reference and analysis by SDS-PAGE (Quantity One®, Bio-Rad).

7.6 Purification of *PfuPGI* mutants

The cell extract of 3g cell pellets was loaded into a 10 ml of DEAE sepharose fast flow column (Amersham Pharmacia Biotech). The column was equilibrated with 50 mM Tris-HCl (pH 8.0). PGI activity eluted at 180 mM of NaCl during a linear gradient of 0 to 1 M NaCl. The fraction with the highest activity was loaded on a pre-equilibrated Superdex 200 GL gel filtration column (GE Healthcare) and eluted in 20 mM Tris-HCl (pH 7.0) containing 100 mM NaCl. Protein purification progress was checked using SDS PAGE.

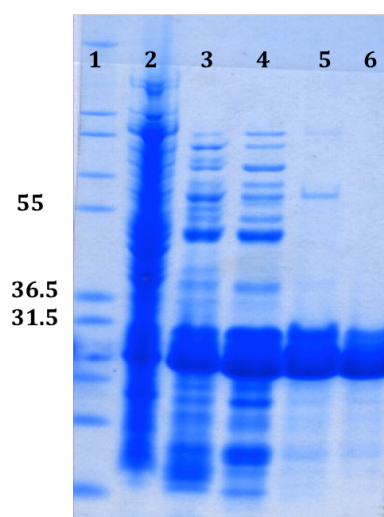


Figure 7.3 SDS-PAGE gel analysis of the purification of PfuPGI. SDS represents samples taken through the purification process of PfuPGI (AG) mutant. Lane 1: Mark12; lane 2: cell extracted (supernatant); lane 3: heat-treated fraction; lane 4: fraction of DEAE column; lane 5: combined fractions before gel filtration; lane 6: fraction of gel filtration column. Similar Purification processes were carried out for all PfuPGI mutants and all resulted in equivalent yield protein with similar behavior.

7.7 Crystallization of PfuPGI mutants

For crystallization, PfuPGI was overexpressed and purified as described previously. For each mutant, protein was concentrated to 11.5 mg ml⁻¹ in a solution of 10 mM Tris-HCl, pH 8.0, and 50 mM F6P and 5 mM MnCl₂ (Figure 7.4). Mutants RG and AG crystallized from hanging drops by mixing equal volumes of protein solution with a reservoir solution containing 0.35 M MgCl₂, 0.1 M sodium acetate pH 5.5 and 10-35% PEG4000. For mutants AD and VY, crystals were grown using a Hydra plus One robot, and commercial screens. AD crystallized from a solution of 0.2 M calcium acetate, 0.1 M sodium acetate pH 6.5, 40% PEG300, whereas VY crystallized from solutions of 0.2 M sodium nitrate, 0.1 M Bis Tris Propane pH 6.5, 20% PEG4000. 50 mM F6P was added to the AD and VY crystals, before mounting. For each different mutant, a single crystal was briefly washed in a cryoprotectant consisting of 25% ethylene glycol in the crystallization buffer, flash cooled to 100 K and stored in liquid nitrogen prior to data collection on the Diamond synchrotron light source. Data were processed using the Xia2 software and structures determined by molecular replacement using the wild type PfuPGI coordinates as a search model (PDB_code 1X82) [176] and the program Phaser [81]. Rounds of building using Coot [86] and refinement in Refmac5 [134] gave acceptable models, verified using Molprobit [135]. For each structure, electron density was present for all the polypeptide chain and the models had no missing residues. However, density was weak for the side chains of Lys21, Lys188, Lys189 (AG); Arg25, Glu114, Asp116, Lys118, Lys188 and Lys189 (RG chain A); Glu114, Lys188 and Lys189 (RG chain B) and Lys188 and Lys189 (AD and VY, chains A and B). Data collection and refinement statistics are given in Tables 7.1 and 7.2. The four mutant structures were compared to the wild type Mn/5PAA Structure (IX7N) by superposition of all the protein atoms of the residues that coordinate the Mn²⁺ (His88, His90, Glu97 and His136) (Table 7.3).

Hanging drop Methods	Robot hit screen
RG	VY
0.35M Mgcl2, 0.1M Na acetate PH5.5 19% PEG 4000 5mM Mn + 50 mM F6P	0.2 M Na nitrate, 0.1M Bis Tris Propane PH6. 20% PEG 4000 5mM Mn + 50 mM F6P
AG	AD
0.35 M Mgcl2, 0.1M Na acetate PH5.5 17% PEG 4000 5mM Mn + 50 mM F6P	0.2 M Ca Acetate, 0.1M Na acetate PH6.5 40% PEG 300 5mM Mn + 50 mM F6P

Figure 7.4 Photographs of PfPGI mutants crystal.

MODEL	AG	RG	AD	VY
Space group	C2	P2 ₁ 2 ₁ 2 ₁	P2 ₁ 2 ₁ 2 ₁	P1
Unit cell parameters (Å)				
a	87.85	72.95	41.3	44.9
b	43.28	74.47	60.03	45.0
c	58.45	75.90	146.7	48.7
α	90.0°	90.0°	90.0°	87.8°
β	122.0°	90.0°	90.0°	89.8°
γ	90.0°	90.0°	90.0°	75.5°
Molecules per ASU	1	2	2	2
Resolution (Å)	28.92-1.41 (1.44-1.41)	26.58-2.04 (2.09 -2.04)	25.58-1.89 (1.94 -1.89)	18.49-1.79 (1.84-1.79)
Wavelength Å	0.97630	0.97630	0.97630	0.97630
Unique observations	32797 (1460)	26994 (1969)	29381 (2144)	33074 (2386)
R_{pim}	0.036 (0.349)	0.074 (0.219)	0.028 (0.352)	0.086 (0.349)
R_{merge}	0.038 (0.364)	0.154 (0.475)	0.037 (0.529)	0.109 (0.459)
Completeness	91.2 (55.4)	99.6 (99.9)	98.5 (99.2)	95.7 (94.0)
Multiplicity	3.1 (2.4)	6.3 (6.4)	3.5 (3.5)	1.8 (1.8)
Mean(I)/sd(I)	13.0 (2.1)	7.0 (3.7)	15.9 (2.4)	4.3 (2.2)
Anomalous completeness	79.6 (42.6)	98.1 (99.2)	86.2 (91.5)	63.8 (60.7)
Anomalous multiplicity	1.6 (1.3)	3.3 (3.3)	1.9 (1.9)	1.0 (1.0)

Table 7.1 Data collection statistics of PfPGI mutants crystals.

Model	AG	RG	AD	VY
Resolution (Å)	1.4	2.0	1.9	1.8
Number of reflections	31134	25635	27836	31398
Protein molecules per asymmetric unit	1	2	2	2
Number of atoms	1739	3221	3142	3308
Number of waters	197	141	112	218
Number of Mn ions	1	2	2	2
Number of F6P	0	0	0	2
Number of 5PAA	1	2	0	0
Ramachandran favoured (%)	98.4	87.1	97.3	97.9
Ramachandran outliers (%)	0	0.3	0.3	0
Poor rotamers	1.3	1.6	0.6	0.3
RMSD bond (Å)	0.006	0.007	0.011	0.009
RMSD angle (°)	1.06	1.08	1.34	1.26
Average B-factors (Å ²)				
Main chain	21	35	31	24
Side chain	31	24	34	27
Waters	37	36	33	26
Mn	13	27	31	21
5PAA/F6P	15	33	-	26
R-factor	0.14	0.22	0.19	0.17
R-FREE	0.20	0.29	0.25	0.22
Molprobit score	0.80	0.94	0.99	1.03
	100 th percentile	100 th percentile	100 th percentile	100 th percentile

Table 7.2 Final refinement Statistics.

Mutant	HIS88-N2	HIS90-N2	HIS136-N2	GLU97-O1
PY WT (1X81)	2.45	2.25	2.29	2.26- O2
RG	2.20	2.30	2.29	2.17-O1
AG	2.32	2.25	2.23	2.33-O1
AD	2.24	2.21	2.35	1.98-O1
VY	2.31	2.31	2.28	2.09-O1

Table 7.3 Mn²⁺ ligands and coordination distances (Å).

7.8 Structure analysis of the mutation-carrying loop

To further examine possible conformational changes, for instance in the active site structure and metal coordination, crystallization trials were initiated for the four PfPFI mutants. First crystallization attempts were set up with manganese as incorporated cofactor and F6P as substrate. These co-crystallization trials successfully yielded well-diffracting crystals for both the mutants RG and AG (Table 7.1). The AG mutant structure is the highest resolution (1.4 Å) for any PfPFI variant, and interestingly, the electron density map for this structure (and also for the RG structure) clearly showed that 5-phosphoarabinonic acid (5PAA) (Figure 7.5), rather than F6P (as added to the crystallization mixture) was bound in the active site (Figure. 7.8). Thus an unexpected conversion had occurred: during the experiment F6P had been, at least partially, oxidized to 5PAA, resulting in preferential binding for 5PAA in the active site of both mutant structures. To confirm that 5PAA had indeed been produced, a solution of the same composition as the crystallization solution was analyzed by mass spectrometry after being left at room temperature for one week. The spectra contained a small peak of m/e ratio 259, (F6P), but also many other peaks with m/e ratios less than F6P, including a large peak with m/e ratio of 245 corresponding to 5PAA, clearly indicating that a breakdown of the sugar had taken place (Figure 7.6). The oxidation of F6P to 5PAA in the presence of permanganate has been observed before [177] and thus we presume a similar reaction occurs in the crystallization solution. For the other two mutants (VY and AD), no ternary complex structures were obtained by co-crystallization, and thus protein crystals were grown in the presence of MgCl₂, and subsequently soaked in a solution of F6P for two hours prior to X-ray data collection. This approach was successful for mutant VY, resulting in a structure with F6P in the active site. For

mutant AD, the only structure that could be obtained contained solely Mn^{2+} in the active site (Figure 7.7).

The crystal structures of the four mutants reveal that in each structure clear electron density remains remarkably consistent, with only minor changes in the position of the main chain atoms. However, the changes in the side chains of Pro132 and Tyr133 have more marked effects on the positions of residues 92-94 of the 90-96 loop, and also on the position of the N-terminal 5 residues from the adjacent subunit of the dimer. There are also some small consequential changes in the positions of second shell residues packing against these two loops (Figure 7.9).

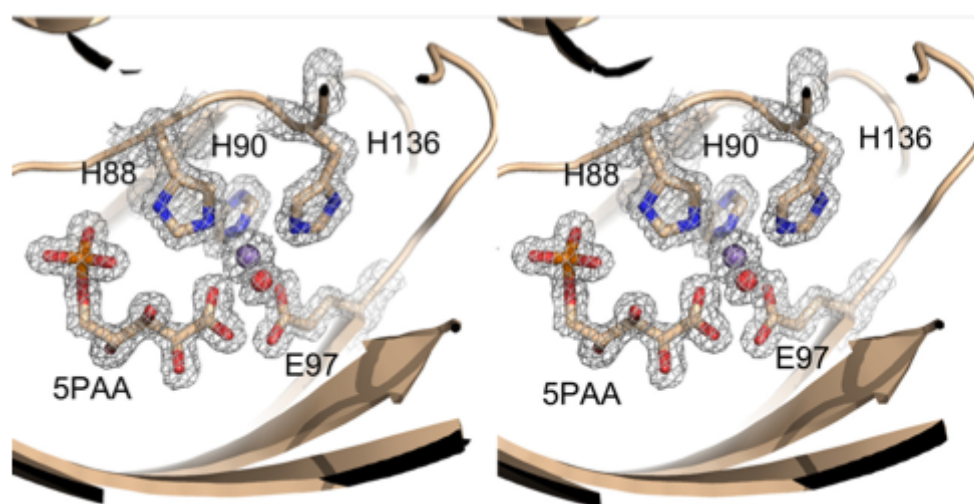


Figure 7.5 A stereo representation of the $mF_0 - DF_c$ electron density (grey mesh), contoured at 1.5σ , for the AG mutant PfPGI structure, showing the manganese (purple sphere), coordinating water (red sphere), bound 5PAA and metal coordinating residues (sticks).

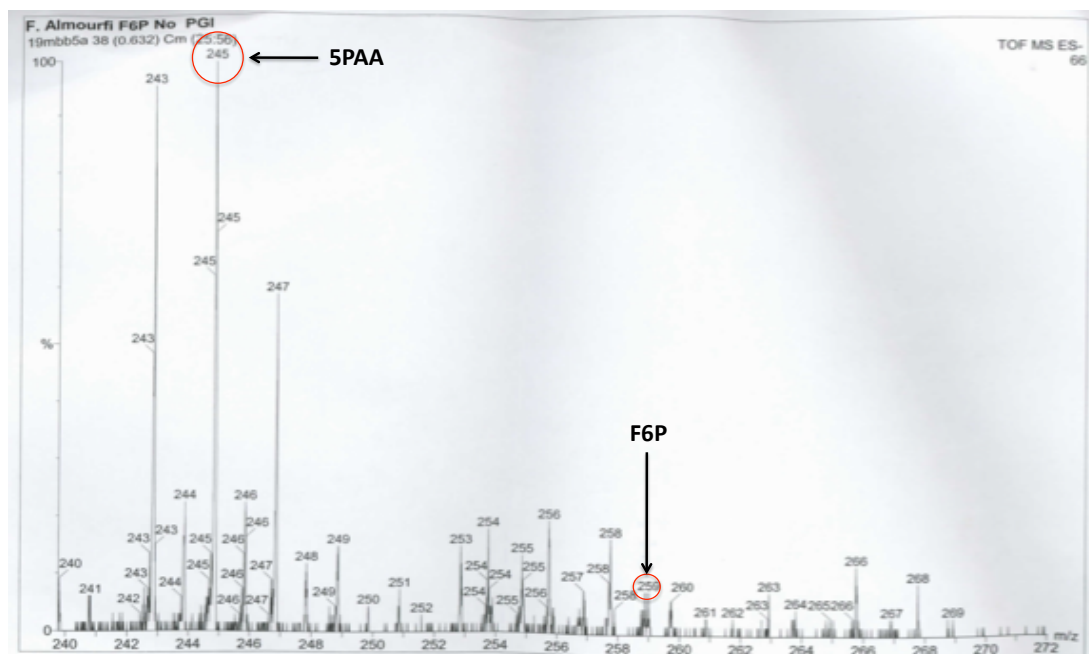


Figure 7.6 Mass spectrometry experiment result of the crystallization buffers of AG and RG mutants crystals. The results confirmed the conversion of the F6P to 5PAA. The spectra contained a small peak of m/e ratio 259, (F6P), but also many other peaks with m/e ratios less than F6P, including a large peak with m/e ratio of 245 corresponding to 5PAA, clearly indicating that a breakdown of the sugar had taken place.

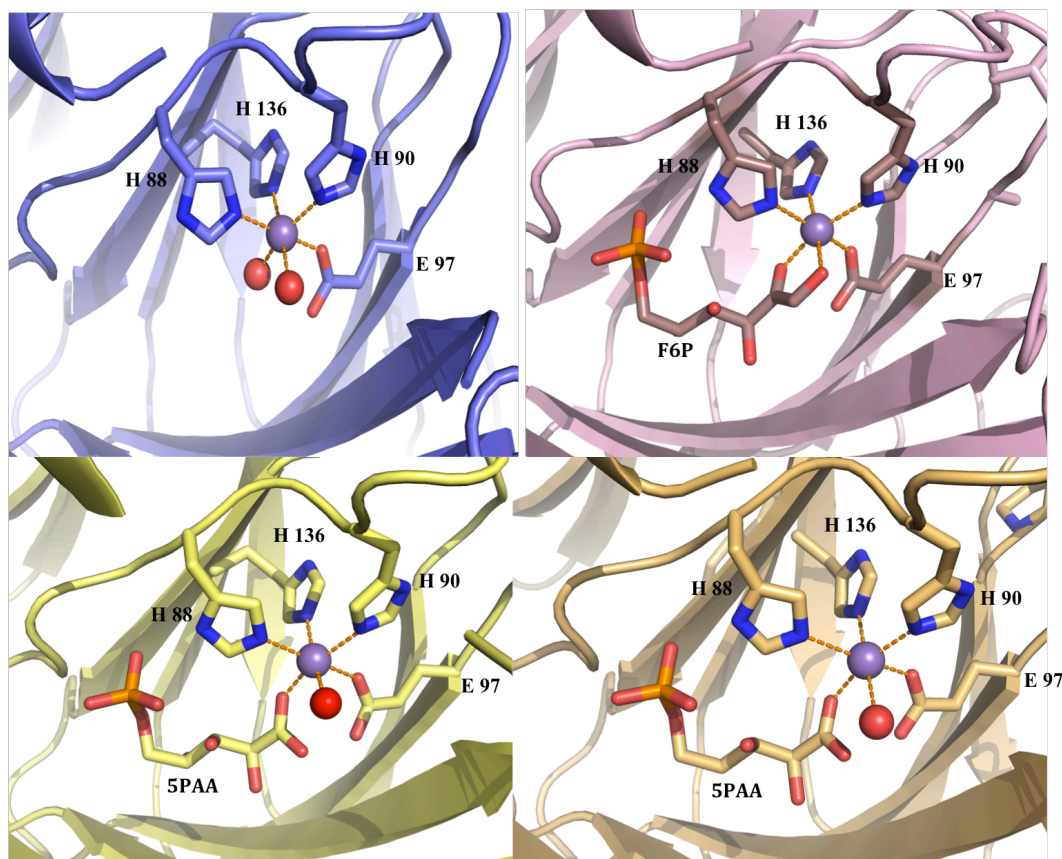


Figure 7.7 Mn^{+2} coordination in the mutant PfPGI structures. For all mutant structures, the Mn^{+2} ion (purple sphere) is coordinated in an octahedral arrangement: mutant AD (blue), mutant VY (pink), mutant AG (yellow) and mutant RG (wheat). Ligands to the metal are highlighted in stick representation, water molecules are shown as red spheres.

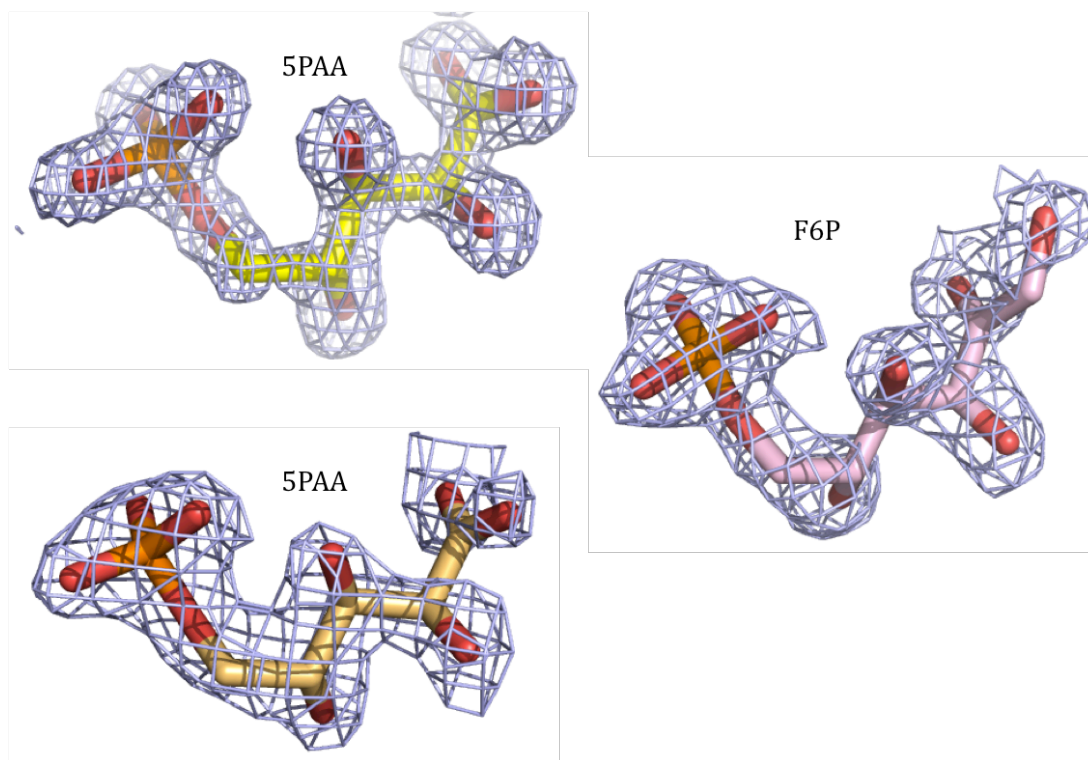


Figure 7.8 Electron densities of substrate and inhibitor of PfPGI mutants. The F6P (pink) binds to VY, 5PAA (yellow) and 5PAA (wheat) bound to AG and RG mutants, respectively.

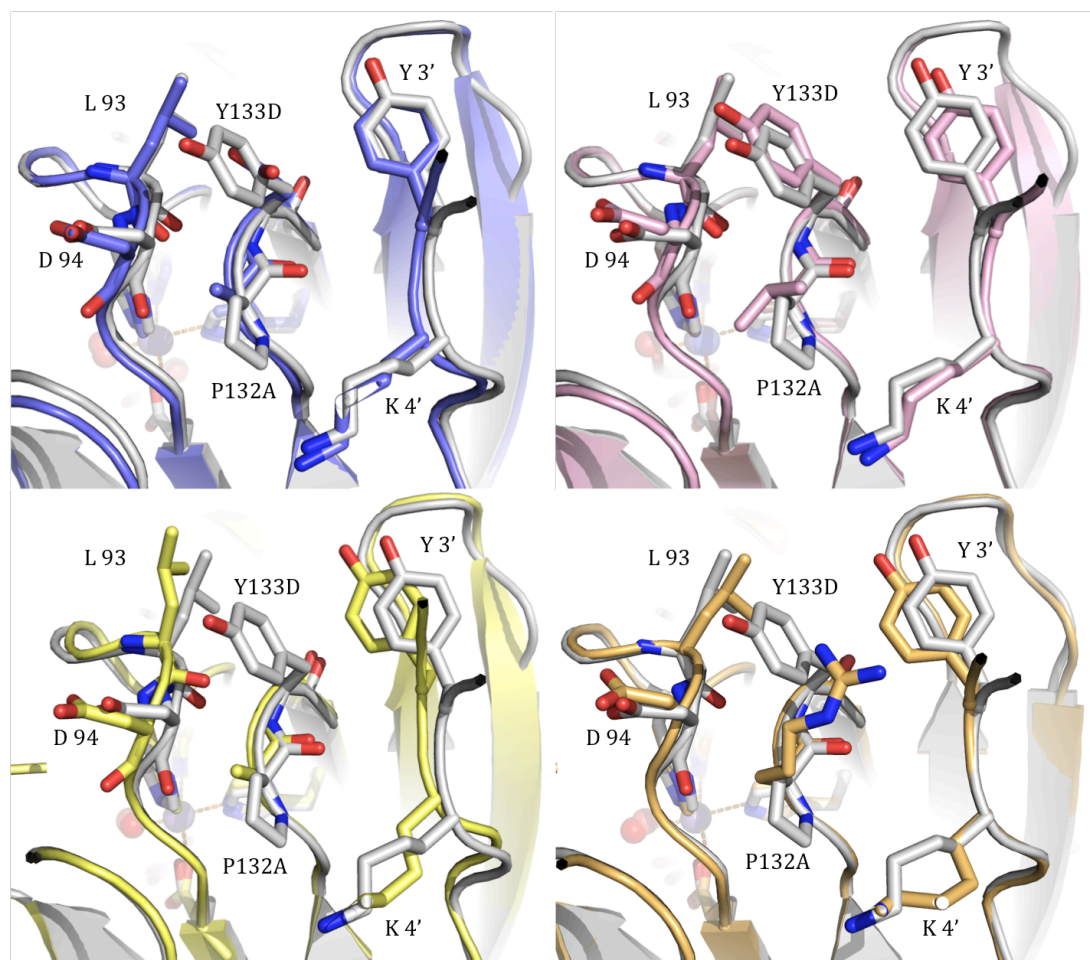


Figure 7.9 Comparison of the structure of the loops adjacent to the mutation position in the four mutant structures. In each panel the wild type structure (1X7N) is shown in white, with relevant residues highlighted in stick representation and labeled. Y3' and K4' refer to the N- terminal strand from the adjacent subunit in the dimer. (a) Mutant AD (blue), (b) mutant VY (pink), (c) mutant AG (yellow) and (d) mutant RG (wheat).

For the P132R-Y133G mutant, the loss of the tyrosine side chain at position 133, is somewhat alleviated by the side chain of R132 occupying approximately the same position in the structure. There are movements of up to 0.5Å in the positions of both the Leu93-Asp94 and Tyr3-Lys4 loops, compared to the wild-type structure. In the P132A-Y133G mutant, the changes to the position of Leu93 and Asp94 are more marked, with movements of 1.8Å and 1.6Å for the alpha carbons of Leu93 and Asp94 compared to the wild type structure. Given that both mutations in this AG structure are to smaller residues than those in the wild type, these fairly large movements are, somewhat counter-intuitively, away from the 132-133, presumably making the packing worse. In the P132V single mutant (VY), the change of the proline side chain to valine pushes the Leu93-Asp94 loop away, in order to accommodate the larger bifurcated side chain of valine. Movements of 0.4Å and 0.7Å are seen between the C α s of Leu93 and Asp94, respectively. In mutant AD, the movement of the alpha carbons of Leu93 and Asp94 away from 132-133 is 0.7Å and 0.9Å, respectively. In this mutant AD, the change from tyrosine to the negatively charged aspartic acid has had little effect on the position of the side chain of Tyr3 from the adjacent subunit.

As might be expected, the small (2-fold increase) in activity seen for the AG and RG mutants did not result in major differences in the structure of the enzyme. Indeed, understanding the precise role of individual residues in the activity of enzymes is fraught with difficulty, as small changes in position of residues far from the active site can marginally influence the binding of the substrate and/or the stabilization of the transition state, thus altering the activity. In addition, a two-fold increase in enzyme activity is not usually regarded as significant, as this degree of change can be within the error of the experiment. Nevertheless, in a biotechnological environment, a two-fold increase in activity and therefore productivity could be a very worthwhile result. Thus the method of correlated mutagenesis could therefore be a very valuable technique in producing optimized enzymes for biotechnological processes, and requires moiré testing to evaluate.

Appendix

Symbols and abbreviations

Abbreviation

Symbols	Definition
Crystallographic	
ASU	Asymmetric unit.
A, b, c, α , β & γ	Dimensions and angles of the real space unit cells in the crystal.
Å	Angstrom (10^{-10} m).
d_{hkl}	Inter-planar spacing in the reciprocal lattice
MAD	Multiple-wavelength anomalous dispersion
SAD	Single-wavelength anomalous dispersion
R	R-factor
R_{free}	Free R-factor
R_{merge}	Merging R-factor.
R_{pim}	Precision-indicating merging R-factor
F	Structure factor
F'	Structure factor from the atomic dispersive scattering.
F''	Structure factor from the atomic anomalous scattering.
F ₀	Structure factor from the normal atomic scattering.
F _{hkl}	Structure factor for a single reflection with indices hkl.
F _A	Structure factor of the anomalous contribution of all atoms.
F _H	Structure factor of the atoms in a heavy metal substructure.
F _P	Structure factor of the atoms in a protein structure.
F _{PH}	Structure factor of a protein structure containing heavy atoms.
F _T	Structure factor of the total contribution of all atoms.
F _{hkl}	Structure factor amplitude for the reflection with indices hkl.
F _{calcs}	Calculated structure factor amplitude.
F _{obs}	Observed structure factor amplitude.
I _{hkl}	Intensity of a reflection with indices hkl.
I / σ I	Signal to noise ratio.
f'	Dispersive atomic scattering factor.
f''	Anomalous atomic scattering factor.
f _{anomalous}	Atomic scattering from an anomalously scattering atom
hkl	Miller indices.
X,y,z	Real space coordination.
V _m	Mathew's coefficient number.
Z	Number of equivalent positions in the unit cell
λ	Wavelength of X-ray
P(u, v, w)	Patterson space coordinates.
α	Phase angle.
α_A	Calculated phase.
ρ	Electron density.
Chemicals & Biological	
DNA	Deoxyribonucleic acid.
DNTP	Deoxyribonucleotide deoxyribonucleic acid.
EDTA	Ethylene diaminetetraacetic acid.
SDS	Sodium dodecyl sulfate.
TRIS	Tris (hydroxymethyl) aminomethane.
ATP	Adenosine triphosphate
Nad(P)H	Nicotinamide adenine dinucleotidew (phosphate).
IPTG	Isopropyl-D-1-thiogalactopyranoside

EDTA	Ethylenediaminetetraacetic acid
PEG	Polyethylene glycol

Miscellaneous

bp	Nucleic acid base pair.
PCR	Polymerase chain reaction.
LB media	Lysogeny broth media: 1 % (w/v) tryptone, 0.5 % (w/v) yeast extract, 1 % (w/v) NaCl.
AU	Absorbance unit.
pI	Isoelectric point.
OD	Optical density.
PAGE	Polyacrylamide gel electrophoresis.
RMSD	Root mean square deviation.
PDB	Protein data bank.

References

1. Madan Babu, M. and K. Sankaran, *DOLOP--database of bacterial lipoproteins*. Bioinformatics, 2002. **18**(4): p. 641-3.
2. Henderson, H.E., et al., *Frameshift mutation in exon 3 of the lipoprotein lipase gene causes a premature stop codon and lipoprotein lipase deficiency*. Mol Biol Med, 1990. **7**(6): p. 511-7.
3. Titgemeyer, F., et al., *A genomic view of sugar transport in Mycobacterium smegmatis and Mycobacterium tuberculosis*. J Bacteriol, 2007. **189**(16): p. 5903-15.
4. *WHO publishes Global tuberculosis report 2013*. Euro Surveill, 2013. **18**(43).
5. Cook, G.M., et al., *Physiology of mycobacteria*. Adv Microb Physiol, 2009. **55**: p. 81-182, 318-9.
6. *[Tuberculosis Annual Report 2011--(4) Tuberculosis treatment and treatment outcomes]*. Kekkaku, 2013. **88**(9): p. 677-86.
7. Ducati, R.G., et al., *The resumption of consumption -- a review on tuberculosis*. Mem Inst Oswaldo Cruz, 2006. **101**(7): p. 697-714.
8. Donoghue, H.D., *Human tuberculosis--an ancient disease, as elucidated by ancient microbial biomolecules*. Microbes Infect, 2009. **11**(14-15): p. 1156-62.
9. Comstock, G.W., *Simple, practical ways to assess the protective efficacy of a new tuberculosis vaccine*. Clin Infect Dis, 2000. **30 Suppl 3**: p. S250-3.
10. Russell, D.G., *Mycobacterium tuberculosis: here today, and here tomorrow*. Nat Rev Mol Cell Biol, 2001. **2**(8): p. 569-77.
11. Tascon, R.E., et al., *Mycobacterium tuberculosis-activated dendritic cells induce protective immunity in mice*. Immunology, 2000. **99**(3): p. 473-80.
12. Ehlers, S., *Lazy, dynamic or minimally recrudescant? On the elusive nature and location of the mycobacterium responsible for latent tuberculosis*. Infection, 2009. **37**(2): p. 87-95.
13. van Crevel, R., T.H. Ottenhoff, and J.W. van der Meer, *Innate immunity to Mycobacterium tuberculosis*. Clin Microbiol Rev, 2002. **15**(2): p. 294-309.
14. Korbel, D.S., B.E. Schneider, and U.E. Schaible, *Innate immunity in tuberculosis: myths and truth*. Microbes Infect, 2008. **10**(9): p. 995-1004.
15. North, R.J. and Y.J. Jung, *Immunity to tuberculosis*. Annu Rev Immunol, 2004. **22**: p. 599-623.
16. Russell, D.G., *Who puts the tubercle in tuberculosis?* Nat Rev Microbiol, 2007. **5**(1): p. 39-47.
17. de Chastellier, C., *The many niches and strategies used by pathogenic mycobacteria for survival within host macrophages*. Immunobiology, 2009. **214**(7): p. 526-42.
18. Fox, W., G.A. Ellard, and D.A. Mitchison, *Studies on the treatment of tuberculosis undertaken by the British Medical Research Council tuberculosis units, 1946-1986, with relevant subsequent publications*. Int J Tuberc Lung Dis, 1999. **3**(10 Suppl 2): p. S231-79.
19. Caminero, J.A., et al., *Best drug treatment for multidrug-resistant and extensively drug-resistant tuberculosis*. Lancet Infect Dis, 2010. **10**(9): p. 621-9.

20. Dye, C. and B.G. Williams, *The population dynamics and control of tuberculosis*. Science, 2010. **328**(5980): p. 856-61.
21. Bishai, J.D., W.R. Bishai, and D.M. Bishai, *Heightened vulnerability to MDR-TB epidemics after controlling drug-susceptible TB*. PLoS One, 2010. **5**(9): p. e12843.
22. Jain, A. and P. Dixit, *Multidrug-resistant to extensively drug resistant tuberculosis: what is next?* J Biosci, 2008. **33**(4): p. 605-16.
23. Navin, T.R., S.J. McNabb, and J.T. Crawford, *The continued threat of tuberculosis*. Emerg Infect Dis, 2002. **8**(11): p. 1187.
24. Orme, I.M., D.N. McMurray, and J.T. Belisle, *Tuberculosis vaccine development: recent progress*. Trends Microbiol, 2001. **9**(3): p. 115-8.
25. Brennan, P.J. and H. Nikaido, *The envelope of mycobacteria*. Annu Rev Biochem, 1995. **64**: p. 29-63.
26. Ouellet, H., et al., *Mycobacterium tuberculosis CYP125A1, a steroid C27 monooxygenase that detoxifies intracellularly generated cholest-4-en-3-one*. Mol Microbiol, 2010. **77**(3): p. 730-42.
27. Uhia, I., et al., *Characterization of the KstR-dependent promoter of the gene for the first step of the cholesterol degradative pathway in Mycobacterium smegmatis*. Microbiology, 2011. **157**(Pt 9): p. 2670-80.
28. Koch, R., *Classics in infectious diseases. The etiology of tuberculosis: Robert Koch. Berlin, Germany 1882*. Rev Infect Dis, 1982. **4**(6): p. 1270-4.
29. Yegian, D. and R.J. Vanderlinde, *The Nature of Acid-Fastness*. J Bacteriol, 1947. **54**(6): p. 777-83.
30. Jarlier, V. and H. Nikaido, *Mycobacterial cell wall: structure and role in natural resistance to antibiotics*. FEMS Microbiol Lett, 1994. **123**(1-2): p. 11-8.
31. Hirschfield, G.R., M. McNeil, and P.J. Brennan, *Peptidoglycan-associated polypeptides of Mycobacterium tuberculosis*. J Bacteriol, 1990. **172**(2): p. 1005-13.
32. Krogh, A., et al., *Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes*. J Mol Biol, 2001. **305**(3): p. 567-80.
33. Sander, P., et al., *Lipoprotein processing is required for virulence of Mycobacterium tuberculosis*. Mol Microbiol, 2004. **52**(6): p. 1543-52.
34. Drews, J., *Drug discovery: a historical perspective*. Science, 2000. **287**(5460): p. 1960-4.
35. Mahfoud, M., et al., *Topology of the porin MspA in the outer membrane of Mycobacterium smegmatis*. J Biol Chem, 2006. **281**(9): p. 5908-15.
36. Sangiorgio, V., et al., *GPI-anchored proteins and lipid rafts*. Ital J Biochem, 2004. **53**(2): p. 98-111.
37. Newton, A.C., *Protein kinase C: structure, function, and regulation*. J Biol Chem, 1995. **270**(48): p. 28495-8.
38. Babu, M.M., et al., *A database of bacterial lipoproteins (DOLOP) with functional assignments to predicted lipoproteins*. J Bacteriol, 2006. **188**(8): p. 2761-73.
39. Hayashi, S. and H.C. Wu, *Lipoproteins in bacteria*. J Bioenerg Biomembr, 1990. **22**(3): p. 451-71.
40. Rezwan, M., et al., *Lipoprotein synthesis in mycobacteria*. Microbiology, 2007. **153**(Pt 3): p. 652-8.

41. McDonough, J.A., et al., *The twin-arginine translocation pathway of Mycobacterium smegmatis is functional and required for the export of mycobacterial beta-lactamases*. J Bacteriol, 2005. **187**(22): p. 7667-79.
42. Sutcliffe, I.C. and D.J. Harrington, *Pattern searches for the identification of putative lipoprotein genes in Gram-positive bacterial genomes*. Microbiology, 2002. **148**(Pt 7): p. 2065-77.
43. Klein, P., R.L. Somorjai, and P.C. Lau, *Distinctive properties of signal sequences from bacterial lipoproteins*. Protein Eng, 1988. **2**(1): p. 15-20.
44. Reglier-Poupet, H., et al., *Maturation of lipoproteins by type II signal peptidase is required for phagosomal escape of Listeria monocytogenes*. J Biol Chem, 2003. **278**(49): p. 49469-77.
45. Mei, J.M., et al., *Identification of Staphylococcus aureus virulence genes in a murine model of bacteraemia using signature-tagged mutagenesis*. Mol Microbiol, 1997. **26**(2): p. 399-407.
46. Petit, C.M., et al., *Lipid modification of prelipoproteins is dispensable for growth in vitro but essential for virulence in Streptococcus pneumoniae*. FEMS Microbiol Lett, 2001. **200**(2): p. 229-33.
47. Leskela, S., et al., *Lipid modification of prelipoproteins is dispensable for growth but essential for efficient protein secretion in Bacillus subtilis: characterization of the Lgt gene*. Mol Microbiol, 1999. **31**(4): p. 1075-85.
48. Sutcliffe, I.C. and R.R. Russell, *Lipoproteins of gram-positive bacteria*. J Bacteriol, 1995. **177**(5): p. 1123-8.
49. Narita, S., S. Matsuyama, and H. Tokuda, *Lipoprotein trafficking in Escherichia coli*. Arch Microbiol, 2004. **182**(1): p. 1-6.
50. Seydel, A., P. Gounon, and A.P. Pugsley, *Testing the '+2 rule' for lipoprotein sorting in the Escherichia coli cell envelope with a new genetic selection*. Mol Microbiol, 1999. **34**(4): p. 810-21.
51. Yamaguchi, K., F. Yu, and M. Inouye, *A single amino acid determinant of the membrane localization of lipoproteins in E. coli*. Cell, 1988. **53**(3): p. 423-32.
52. Sutcliffe, I.C. and D.J. Harrington, *Lipoproteins of Mycobacterium tuberculosis: an abundant and functionally diverse class of cell envelope components*. FEMS Microbiol Rev, 2004. **28**(5): p. 645-59.
53. Cole, S.T., et al., *Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence*. Nature, 1998. **393**(6685): p. 537-44.
54. Wu, H.C., *Biosynthesis of lipoproteins*. In *Escherichia coli and Salmonella typhimurium*. American Society for Microbiology., 1996.
55. Sankaran, K. and H.C. Wu, *Lipid modification of bacterial prolipoprotein. Transfer of diacylglycerol moiety from phosphatidylglycerol*. J Biol Chem, 1994. **269**(31): p. 19701-6.
56. Diaz-Silvestre, H., et al., *The 19-kDa antigen of Mycobacterium tuberculosis is a major adhesin that binds the mannose receptor of THP-1 monocytic cells and promotes phagocytosis of mycobacteria*. Microb Pathog, 2005. **39**(3): p. 97-107.
57. Chang, Z., et al., *The immunodominant 38-kDa lipoprotein antigen of Mycobacterium tuberculosis is a phosphate-binding protein*. J Biol Chem, 1994. **269**(3): p. 1956-8.

58. Boshoff, H.I., et al., *The transcriptional responses of Mycobacterium tuberculosis to inhibitors of metabolism: novel insights into drug mechanisms of action*. J Biol Chem, 2004. **279**(38): p. 40174-84.
59. Sulzenbacher, G., et al., *LppX is a lipoprotein required for the translocation of phthiocerol dimycocerosates to the surface of Mycobacterium tuberculosis*. EMBO J, 2006. **25**(7): p. 1436-44.
60. Keenan, J., et al., *Immune response to an 18-kilodalton outer membrane antigen identifies lipoprotein 20 as a Helicobacter pylori vaccine candidate*. Infect Immun, 2000. **68**(6): p. 3337-43.
61. Romano, M., et al., *Evaluation of the immunogenicity of pBudCE4.1 plasmids encoding mycolyl-transferase Ag85A and phosphate transport receptor PstS-3 from Mycobacterium tuberculosis*. Vaccine, 2006. **24**(21): p. 4640-3.
62. Niederweis, M., *Nutrient acquisition by mycobacteria*. Microbiology, 2008. **154**(Pt 3): p. 679-92.
63. Hancock, R.E., et al., *Interaction of aminoglycosides with the outer membranes and purified lipopolysaccharide and OmpF porin of Escherichia coli*. Antimicrob Agents Chemother, 1991. **35**(7): p. 1309-14.
64. Nikaido, H., *Porins and specific diffusion channels in bacterial outer membranes*. J Biol Chem, 1994. **269**(6): p. 3905-8.
65. Braun, V. and H. Killmann, *Bacterial solutions to the iron-supply problem*. Trends Biochem Sci, 1999. **24**(3): p. 104-9.
66. Miles, L.R., et al., *Effect of polypurine tract (PPT) mutations on human immunodeficiency virus type 1 replication: a virus with a completely randomized PPT retains low infectivity*. J Virol, 2005. **79**(11): p. 6859-67.
67. McDevitt, D. and M. Rosenberg, *Exploiting genomics to discover new antibiotics*. Trends Microbiol, 2001. **9**(12): p. 611-7.
68. Camus, J.C., et al., *Re-annotation of the genome sequence of Mycobacterium tuberculosis H37Rv*. Microbiology, 2002. **148**(Pt 10): p. 2967-73.
69. Arcus, V.L., et al., *The potential impact of structural genomics on tuberculosis drug discovery*. Drug Discov Today, 2006. **11**(1-2): p. 28-34.
70. Lin, T.W., et al., *Structure-based inhibitor design of AccD5, an essential acyl-CoA carboxylase carboxyltransferase domain of Mycobacterium tuberculosis*. Proc Natl Acad Sci U S A, 2006. **103**(9): p. 3072-7.
71. Ehebauer, M.T. and M. Wilmanns, *The progress made in determining the Mycobacterium tuberculosis structural proteome*. Proteomics, 2011. **11**(15): p. 3128-33.
72. Rogall, T., et al., *Towards a phylogeny and definition of species at the molecular level within the genus Mycobacterium*. Int J Syst Bacteriol, 1990. **40**(4): p. 323-30.
73. Cole, S.T., *Comparative and functional genomics of the Mycobacterium tuberculosis complex*. Microbiology, 2002. **148**(Pt 10): p. 2919-28.
74. Kyte, J. and R.F. Doolittle, *A simple method for displaying the hydropathic character of a protein*. J Mol Biol, 1982. **157**(1): p. 105-32.
75. Sasseti, C.M. and E.J. Rubin, *Genetic requirements for mycobacterial survival during infection*. Proc Natl Acad Sci U S A, 2003. **100**(22): p. 12989-94.

76. Teh, J.S., T. Yano, and H. Rubin, *Type II NADH: menaquinone oxidoreductase of Mycobacterium tuberculosis*. *Infect Disord Drug Targets*, 2007. **7**(2): p. 169-81.
77. Rhodes, G., *Crystallography Made Crystal Clear* Academic Press February 2006, London, USA.
78. Kantardjieff, K.A. and B. Rupp, *Matthews coefficient probabilities: Improved estimates for unit cell contents of proteins, DNA, and protein-nucleic acid complex crystals*. *Protein Sci*, 2003. **12**(9): p. 1865-71.
79. Leslie, A.G.W., *Recent changes to the MOSFLM package for processing film and image plate data. Recent Changes to the MOSFLM Package for Processing Film and*

Image Plate Data. 1992.

80. Kabsch, W., *Integration, scaling, space-group assignment and post-refinement*. *Acta Crystallogr D Biol Crystallogr*, 2010. **66**(Pt 2): p. 133-44.
81. G., W., *xia2: an expert system for macromolecular crystallography data reduction*. *Journal of Applied Crystallography*, 2010. **43**(1): p. 186-190.
82. Evans, P., *Scaling and assessment of data quality*. *Acta Crystallogr D Biol Crystallogr*, 2006. **62**(Pt 1): p. 72-82.
83. McCoy, A.J., et al., *Phaser crystallographic software*. *J Appl Crystallogr*, 2007. **40**(Pt 4): p. 658-674.
84. Sheldrick, G.M., *Experimental phasing with SHELXC/D/E: combining chain tracing with density modification*. *Acta Crystallogr D Biol Crystallogr*, 2010. **66**(Pt 4): p. 479-85.
85. Pape, T. and T. Schneider, *{HKL2MAP}: a graphical user interface for phasing with {SHELX} programs*. *J. Appl. Cryst.*, 2004. **27**: p. 843-844.
86. Emsley, P. and K. Cowtan, *Coot: model-building tools for molecular graphics*. *Acta Crystallogr D Biol Crystallogr*, 2004. **60**(Pt 12 Pt 1): p. 2126-32.
87. Emsley, P., et al., *Features and development of Coot*. *Acta Crystallogr D Biol Crystallogr*, 2010. **66**(Pt 4): p. 486-501.
88. Winn, M.D., et al., *Overview of the CCP4 suite and current developments*. *Acta Crystallogr D Biol Crystallogr*, 2011. **67**(Pt 4): p. 235-42.
89. Kelley, L.A. and M.J. Sternberg, *Protein structure prediction on the Web: a case study using the Phyre server*. *Nat Protoc*, 2009. **4**(3): p. 363-71.
90. Kanehisa, M., et al., *Data, information, knowledge and principle: back to metabolism in KEGG*. *Nucleic Acids Res*, 2014. **42**(1): p. D199-205.
91. Fisher, N., et al., *The malaria parasite type II NADH:quinone oxidoreductase: an alternative enzyme for an alternative lifestyle*. *Trends Parasitol*, 2007. **23**(7): p. 305-10.
92. Ana M. P. Melo, T.M.B., and Miguel Teixeira, *New Insights into Type II NAD(P)H:Quinone Oxidoreductases*. *Microbiol Mol Biol Rev*, Dec 2004. **68**: p. 603-616.
93. Weinstein, E.A., et al., *Inhibitors of type II NADH:menaquinone oxidoreductase represent a class of antitubercular drugs*. *Proc Natl Acad Sci U S A*, 2005. **102**(12): p. 4548-53.
94. Kerscher, S.J., *Diversity and origin of alternative NADH:ubiquinone oxidoreductases*. *Biochim Biophys Acta*, 2000. **1459**(2-3): p. 274-83.

95. Mattevi, A., et al., *Three-dimensional structure of lipoamide dehydrogenase from Pseudomonas fluorescens at 2.8 Å resolution. Analysis of redox and thermostability properties.* J Mol Biol, 1993. **230**(4): p. 1200-15.
96. Melo, A.M., T.M. Bandejas, and M. Teixeira, *New insights into type II NAD(P)H:quinone oxidoreductases.* Microbiol Mol Biol Rev, 2004. **68**(4): p. 603-16.
97. Iwata, M., et al., *The structure of the yeast NADH dehydrogenase (Ndi1) reveals overlapping binding sites for water- and lipid-soluble substrates.* Proc Natl Acad Sci U S A, 2012. **109**(38): p. 15247-52.
98. Schrödinger, L., *The PyMOL Molecular Graphics System, Version 1.3.* 2010.
99. Heikal, A., et al., *Structure of the bacterial type II NADH dehydrogenase: a monotopic membrane protein with an essential role in energy generation.* Mol Microbiol, 2014.
100. Feng, Y., et al., *Structural insight into the type-II mitochondrial NADH dehydrogenases.* Nature, 2012. **491**(7424): p. 478-82.
101. Sonnhammer, E.L., G. von Heijne, and A. Krogh, *A hidden Markov model for predicting transmembrane helices in protein sequences.* Proc Int Conf Intell Syst Mol Biol, 1998. **6**: p. 175-82.
102. Petersen, T.N., et al., *SignalP 4.0: discriminating signal peptides from transmembrane regions.* Nat Methods, 2011. **8**(10): p. 785-6.
103. Lee, Y.H., et al., *Treponema pallidum TroA is a periplasmic zinc-binding protein with a helical backbone.* Nat Struct Biol, 1999. **6**(7): p. 628-33.
104. Armoa, G.R., et al., *A highly immunogenic putative Mycobacterium kansasii lipoprotein.* Microbiology, 1995. **141 (Pt 10)**: p. 2705-12.
105. Altschul, S.F., et al., *Basic local alignment search tool.* J Mol Biol, 1990. **215**(3): p. 403-10.
106. Higgins, C.F., *ABC transporters: from microorganisms to man.* Annu Rev Cell Biol, 1992. **8**: p. 67-113.
107. Saurin, W., M. Hofnung, and E. Dassa, *Getting in or out: early segregation between importers and exporters in the evolution of ATP-binding cassette (ABC) transporters.* J Mol Evol, 1999. **48**(1): p. 22-41.
108. Linton, K.J. and C.F. Higgins, *The Escherichia coli ATP-binding cassette (ABC) proteins.* Mol Microbiol, 1998. **28**(1): p. 5-13.
109. Quentin, Y., G. Fichant, and F. Denizot, *Inventory, assembly and analysis of Bacillus subtilis ABC transport systems.* J Mol Biol, 1999. **287**(3): p. 467-84.
110. Braibant, M., P. Gilot, and J. Content, *The ATP binding cassette (ABC) transport systems of Mycobacterium tuberculosis.* FEMS Microbiol Rev, 2000. **24**(4): p. 449-67.
111. Jiang, D., et al., *Structural analysis of Mycobacterium tuberculosis ATP-binding cassette transporter subunit UgpB reveals specificity for glycerophosphocholine.* FEBS J, 2014. **281**(1): p. 331-41.
112. Ehrmann, M., et al., *The ABC maltose transporter.* Mol Microbiol, 1998. **29**(3): p. 685-94.
113. Locher, K.P., *Review. Structure and mechanism of ATP-binding cassette transporters.* Philos Trans R Soc Lond B Biol Sci, 2009. **364**(1514): p. 239-45.

114. Ames, G.F., C.S. Mimura, and V. Shyamala, *Bacterial periplasmic permeases belong to a family of transport proteins operating from Escherichia coli to human: Traffic ATPases*. FEMS Microbiol Rev, 1990. **6**(4): p. 429-46.
115. Gilson, E., et al., *Evidence for high affinity binding-protein dependent transport systems in gram-positive bacteria and in Mycoplasma*. EMBO J, 1988. **7**(12): p. 3971-4.
116. Ross, J.I., et al., *Inducible erythromycin resistance in staphylococci is encoded by a member of the ATP-binding transport super-gene family*. Mol Microbiol, 1990. **4**(7): p. 1207-14.
117. Felmler, T., S. Pellett, and R.A. Welch, *Nucleotide sequence of an Escherichia coli chromosomal hemolysin*. J Bacteriol, 1985. **163**(1): p. 94-105.
118. Bulut, H., et al., *Crystal structures of receptors involved in small molecule transport across membranes*. Eur J Cell Biol, 2012. **91**(4): p. 318-25.
119. Chen, J., et al., *Trapping the transition state of an ATP-binding cassette transporter: evidence for a concerted mechanism of maltose transport*. Proc Natl Acad Sci U S A, 2001. **98**(4): p. 1525-30.
120. Higgins, C.F., et al., *Binding protein-dependent transport systems*. J Bioenerg Biomembr, 1990. **22**(4): p. 571-92.
121. Tam, R. and M.H. Saier, Jr., *Structural, functional, and evolutionary relationships among extracellular solute-binding receptors of bacteria*. Microbiol Rev, 1993. **57**(2): p. 320-46.
122. Boos, W., *Bacterial transport*. Annu Rev Biochem, 1974. **43**(0): p. 123-46.
123. Ames, G.F., C. Prody, and S. Kustu, *Simple, rapid, and quantitative release of periplasmic proteins by chloroform*. J Bacteriol, 1984. **160**(3): p. 1181-3.
124. Schneider, E., *ABC transporters catalyzing carbohydrate uptake*. Res Microbiol, 2001. **152**(3-4): p. 303-10.
125. Albers, S.V., et al., *Glucose transport in the extremely thermoacidophilic Sulfolobus solfataricus involves a high-affinity membrane-integrated binding protein*. J Bacteriol, 1999. **181**(14): p. 4285-91.
126. Bertram, R., et al., *In silico and transcriptional analysis of carbohydrate uptake systems of Streptomyces coelicolor A3(2)*. J Bacteriol, 2004. **186**(5): p. 1362-73.
127. Pao, S.S., I.T. Paulsen, and M.H. Saier, Jr., *Major facilitator superfamily*. Microbiol Mol Biol Rev, 1998. **62**(1): p. 1-34.
128. Park, J.H. and M.H. Saier, Jr., *Phylogenetic characterization of the MIP family of transmembrane channel proteins*. J Membr Biol, 1996. **153**(3): p. 171-80.
129. Reizer, J., A. Reizer, and M.H. Saier, Jr., *A functional superfamily of sodium/solute symporters*. Biochim Biophys Acta, 1994. **1197**(2): p. 133-66.
130. Brinkkotter, A., et al., *Pathways for the utilization of N-acetyl-galactosamine and galactosamine in Escherichia coli*. Mol Microbiol, 2000. **37**(1): p. 125-35.
131. Stephan, J., et al., *The growth rate of Mycobacterium smegmatis depends on sufficient porin-mediated influx of nutrients*. Mol Microbiol, 2005. **58**(3): p. 714-30.

132. Kabsch, W., *Evaluation of single-crystal X-ray diffraction data from a position-sensitive detector*. J. Appl. Crystallogr, 1988. **21**: p. **916-924**.
133. Stein, N., *CHAINSAW: a program for mutating pdb files used as templates in molecular replacement*. Journal of Applied Crystallography, 2008. **41**: p. 641-643.
134. Murshudov, G.N., et al., *REFMAC5 for the refinement of macromolecular crystal structures*. Acta Crystallogr D Biol Crystallogr, 2011. **67**(Pt 4): p. 355-67.
135. Chen, V.B., et al., *MolProbity: all-atom structure validation for macromolecular crystallography*. Acta Crystallogr D Biol Crystallogr, 2010. **66**(Pt 1): p. 12-21.
136. Laskowski, R.A., et al., *AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR*. J Biomol NMR, 1996. **8**(4): p. 477-86.
137. Krissinel, E. and K. Henrick, *Inference of macromolecular assemblies from crystalline state*. J Mol Biol, 2007. **372**(3): p. 774-97.
138. Jones, S. and J.M. Thornton, *Protein-protein interactions: a review of protein dimer structures*. Prog Biophys Mol Biol, 1995. **63**(1): p. 31-65.
139. Holm, L. and P. Rosenstrom, *Dali server: conservation mapping in 3D*. Nucleic Acids Res, 2010. **38**(Web Server issue): p. W545-9.
140. Diez, J., et al., *The crystal structure of a liganded trehalose/maltose-binding protein from the hyperthermophilic Archaeon Thermococcus litoralis at 1.85 Å*. J Mol Biol, 2001. **305**(4): p. 905-15.
141. Vahedi-Faridi, A., et al., *Crystal structures of the solute receptor GacH of Streptomyces glaucescens in complex with acarbose and an acarbose homolog: comparison with the acarbose-loaded maltose-binding protein of Salmonella typhimurium*. J Mol Biol, 2010. **397**(3): p. 709-23.
142. Matsumoto, N., et al., *Crystal structures of open and closed forms of cyclo/maltodextrin-binding protein*. FEBS J, 2009. **276**(11): p. 3008-19.
143. Quijcho, F.A., J.C. Spurlino, and L.E. Rodseth, *Extensive features of tight oligosaccharide binding revealed in high-resolution structures of the maltodextrin transport/chemosensory receptor*. Structure, 1997. **5**(8): p. 997-1015.
144. Campobasso, N., et al., *Crystal structure of thiaminase-I from Bacillus thiaminolyticus at 2.0 Å resolution*. Biochemistry, 1998. **37**(45): p. 15981-9.
145. Pei, J., B.H. Kim, and N.V. Grishin, *PROMALS3D: a tool for multiple protein sequence and structure alignments*. Nucleic Acids Res, 2008. **36**(7): p. 2295-300.
146. Fukami-Kobayashi, K., Y. Tateno, and K. Nishikawa, *Domain dislocation: a change of core structure in periplasmic binding proteins in their evolutionary history*. J Mol Biol, 1999. **286**(1): p. 279-90.
147. Boos, W., and J. M. Lucht, *Periplasmic binding protein-dependent ABC transporters*. American Society for Microbiology., 1996: p. 1175-1209.
148. Berntsson, R.P., et al., *A structural classification of substrate-binding proteins*. FEBS Lett, 2010. **584**(12): p. 2606-17.

149. Duan, X., et al., *Crystal structures of the maltodextrin/maltose-binding protein complexed with reduced oligosaccharides: flexibility of tertiary structure and ligand binding*. J Mol Biol, 2001. **306**(5): p. 1115-26.
150. Beasley, F.C., et al., *Characterization of staphyloferrin A biosynthetic and transport mutants in Staphylococcus aureus*. Mol Microbiol, 2009. **72**(4): p. 947-63.
151. Suzuki, R., et al., *Structural and thermodynamic analyses of solute-binding Protein from Bifidobacterium longum specific for core 1 disaccharide and lacto-N-biose I*. J Biol Chem, 2008. **283**(19): p. 13165-73.
152. Abbott, D.W., et al., *The molecular basis of glycogen breakdown and transport in Streptococcus pneumoniae*. Mol Microbiol, 2010. **77**(1): p. 183-99.
153. Wallace, A.C., R.A. Laskowski, and J.M. Thornton, *LIGPLOT: a program to generate schematic diagrams of protein-ligand interactions*. Protein Eng, 1995. **8**(2): p. 127-34.
154. Greenfield, N.J., *Determination of the folding of proteins as a function of denaturants, osmolytes or ligands using circular dichroism*. Nat Protoc, 2006. **1**(6): p. 2733-41.
155. Edson, N.L., *The intermediary metabolism of the mycobacteria*. Bacteriol Rev, 1951. **15**(3): p. 147-82.
156. Kana, B.D. and V. Mizrahi, *Molecular genetics of Mycobacterium tuberculosis in relation to the discovery of novel drugs and vaccines*. Tuberculosis (Edinb), 2004. **84**(1-2): p. 63-75.
157. Machowski, E.E., S. Dawes, and V. Mizrahi, *TB tools to tell the tale-molecular genetic methods for mycobacterial research*. Int J Biochem Cell Biol, 2005. **37**(1): p. 54-68.
158. McKinney, J.D., et al., *Persistence of Mycobacterium tuberculosis in macrophages and mice requires the glyoxylate shunt enzyme isocitrate lyase*. Nature, 2000. **406**(6797): p. 735-8.
159. Gebhard, S., S.L. Tran, and G.M. Cook, *The Phn system of Mycobacterium smegmatis: a second high-affinity ABC-transporter for phosphate*. Microbiology, 2006. **152**(Pt 11): p. 3453-65.
160. Lutz, S., *Beyond directed evolution--semi-rational protein engineering and design*. Curr Opin Biotechnol, 2010. **21**(6): p. 734-43.
161. Sullivan, B.J., et al., *Stabilizing proteins from sequence statistics: the interplay of conservation and correlation in triosephosphate isomerase stability*. J Mol Biol, 2012. **420**(4-5): p. 384-99.
162. Kuipers, R.K., et al., *Correlated mutation analyses on super-family alignments reveal functionally important residues*. Proteins, 2009. **76**(3): p. 608-16.
163. Kowarsch, A., et al., *Correlated mutations: a hallmark of phenotypic amino acid substitutions*. PLoS Comput Biol, 2010. **6**(9).
164. Gloor, G.B., et al., *Mutual information in protein multiple sequence alignments reveals two classes of coevolving positions*. Biochemistry, 2005. **44**(19): p. 7156-65.
165. Kundrotas, P.J. and E.G. Alexov, *Predicting residue contacts using pragmatic correlated mutations method: reducing the false positives*. BMC Bioinformatics, 2006. **7**: p. 503.

166. Dunwell, J.M., S. Khuri, and P.J. Gane, *Microbial relatives of the seed storage proteins of higher plants: conservation of structure and diversification of function during evolution of the cupin superfamily*. Microbiol Mol Biol Rev, 2000. **64**(1): p. 153-79.
167. Chica, R.A., N. Doucet, and J.N. Pelletier, *Semi-rational approaches to engineering enzyme activity: combining the benefits of directed evolution and rational design*. Curr Opin Biotechnol, 2005. **16**(4): p. 378-84.
168. Hansen, T., M. Oehlmann, and P. Schonheit, *Novel type of glucose-6-phosphate isomerase in the hyperthermophilic archaeon Pyrococcus furiosus*. J Bacteriol, 2001. **183**(11): p. 3428-35.
169. Verhees, C.H., et al., *The phosphoglucose isomerase from the hyperthermophilic archaeon Pyrococcus furiosus is a unique glycolytic enzyme that belongs to the cupin superfamily*. J Biol Chem, 2001. **276**(44): p. 40926-32.
170. Berrisford, J.M., et al., *Crystal structure of Pyrococcus furiosus phosphoglucose isomerase. Implications for substrate binding and catalysis*. J Biol Chem, 2003. **278**(35): p. 33290-7.
171. Berrisford, J.M., et al., *Evidence supporting a cis-enediol-based mechanism for Pyrococcus furiosus phosphoglucose isomerase*. J Mol Biol, 2006. **358**(5): p. 1353-66.
172. Tokuriki, N. and D.S. Tawfik, *Stability effects of mutations and protein evolvability*. Curr Opin Struct Biol, 2009. **19**(5): p. 596-604.
173. Jochens, H. and U.T. Bornscheuer, *Natural diversity to guide focused directed evolution*. Chembiochem, 2010. **11**(13): p. 1861-6.
174. Kuipers, R.K., et al., *3DM: systematic analysis of heterogeneous superfamily data to discover protein functionalities*. Proteins, 2010. **78**(9): p. 2101-13.
175. Mintseris, J. and Z. Weng, *Structure, function, and evolution of transient and obligate protein-protein interactions*. Proc Natl Acad Sci U S A, 2005. **102**(31): p. 10930-5.
176. Berrisford, J.M., et al., *The structures of inhibitor complexes of Pyrococcus furiosus phosphoglucose isomerase provide insights into substrate binding and catalysis*. J Mol Biol, 2004. **343**(3): p. 649-57.
177. Azmat R, N.R., Qamar N, Malik I., *Kinetics and mechanisms of oxidation of d-fructose and d-lactose by permanganate ion in acidic medium*. Natural Science, 2012. **4**(1):466-478.
178. Arthur, C.W., *Function of the Laboratory in the Epidemiological Control of Syphilis*. Am J Public Health Nations Health, 1935. **25**(7): p. 845-7.



Correlated mutation analysis as a tool for smart library design to improve protein performance

Journal:	<i>PROTEINS: Structure, Function, and Bioinformatics</i>
Manuscript ID:	Draft
Wiley - Manuscript type:	Research Article
Date Submitted by the Author:	n/a
Complete List of Authors:	Raedts, John; Wageningen University, Laboratory of Microbiology Almourfi, Feras; University of Sheffield, Department of Molecular Biology and Biotechnology Joosten, Henk-Jan; Bio-Product, - Hendriks, Sjon; Wageningen University, Laboratory of Microbiology Sedelnikova, Svetlana; University of Sheffield, Department of Molecular Biology and Biotechnology Kengen, Servé; Wageningen University, Microbiology Hagen, Wilfred; Delft University of Technology, Department of Biotechnology Schaap, Peter; Wageningen Universiteit, Microbiology Baker, Patrick J.; University of Sheffield, Department of Molecular Biology and Biotechnology van der Oost, John; Wageningen Universiteit, Microbiology
Key Words:	protein engineering, Comulator, cupin superfamily, phosphoglucose isomerase, protein structure

SCHOLARONE™
Manuscripts

1
2
3
4 1 **Correlated mutation analysis as a tool for smart library design to**
5
6
7 2 **improve protein performance**
8
9
10 3

11 4 John Raedts¹, Feras Almourfi², Henk-Jan Joosten³, Sjon Hendriks¹, Svetlana E. Sedelnikova²,
12 5 Servé W.M. Kengen¹, Wilfred R. Hagen⁴, Peter J. Schaap⁵, Patrick J. Baker^{2*}, John van der
13 6 Oost^{1*}
14
15
16
17
18
19
20
21
22
23
24
25
26
27

28 8 ¹*Laboratory of Microbiology, Wageningen University, Dreijenplein 10, 6703 HB,*
29 9 *Wageningen, The Netherlands*
30
31
32
33
34

35 11 ²*The Krebs Institute for Biomolecular Research, Department of Molecular Biology and*
36 12 *Biotechnology, University of Sheffield, Firth Court, Western Bank, Sheffield S10 2TN, UK*
37
38
39
40
41
42
43
44
45
46

47 14 ³*Bio-Product, Castellastraat 116, 6512 EZ, Nijmegen, The Netherlands*
48
49
50
51
52
53

54 16 ⁴*Department of Biotechnology, Delft University of Technology, Julianalaan 67, 2628 BC,*
55 17 *Delft, The Netherlands*
56
57
58
59
60

61 19 ⁵*Laboratory of Systems and Synthetic Biology, Wageningen University, Dreijenplein 10, 6703*
62 20 *HB, Wageningen, The Netherlands*
63
64
65
66

67 22 Work performed at ¹⁻⁵, experimental work mostly at ¹⁻²

68 23 *To whom correspondence should be addressed:

69 24 John van der Oost, Tel. +31-317-483108, Fax +31317-483829, john.vanderoost@wur.nl

70 25 Patrick J. Baker, Tel. +44-114-2222725, Fax +44-114-2222800, p.baker@sheffield.ac.uk

1
2
3
4 26 **Abstract**
5

6 27 To enable rational approaches in protein engineering, various bioinformatics tools are being
7
8 28 developed. In this work we used structure-based multiple sequence alignments (MSAs) and a
9
10 29 correlated mutation analysis (CMA) tool to identify target amino acid residues for
11
12 30 mutagenesis enabling ‘smart library design’. CMA analysis of the cupin super-family
13
14 31 revealed a set of correlated amino acids. Using the phosphoglucose isomerase from
15
16 32 *Pyrococcus furiosus* (PfPGI) as model enzyme, we varied the strongly correlated residues
17
18 33 Pro132 and Tyr133 by saturation mutagenesis. Although this amino acid pair is located in a
19
20 34 loop relatively distant from the active site, their predicted relevance could be confirmed by
21
22 35 activity measurements of the PfPGI substitution mutants. Screening of the generated
23
24 36 substitution library revealed a positive correlation between the prevalence of correlating
25
26 37 amino acid pairs in the superfamily and the specific activity of the corresponding PfPGI
27
28 38 mutants. All tested mutants retained protein stability like wild type PfPGI. Crystal structures
29
30 39 of a selection of the mutants were determined to increase our understanding of the molecular
31
32 40 basis of the observed differences in activity. Interestingly, the obtained crystal structures of
33
34 41 the four selected PfPGI variants did not reveal major changes in substrate and metal binding.
35
36 42 This could be confirmed by electron paramagnetic resonance (EPR). This study suggests that
37
38 43 CMA can play an important role in predicting non-obvious mutations that could lead to subtle
39
40 44 optimization of protein performance however without necessarily introduction of structural
41
42 45 changes.
43
44
45
46
47
48
49
50

51 46
52
53 47 **Keywords:**
54

55 48 protein engineering, Comulator, cupin superfamily, phosphoglucose isomerase, protein
56
57 49 structure
58
59
60

1
2
3 51 **Abbreviations used:**
4

5 52 5PAA, 5-phosphoarabinonic acid; CMA, correlated mutation analysis tool; EPR, electron
6

7 paramagnetic resonance; F6P, D-fructose 6-phosphate; MSA, multiple sequence alignments;
8

9
10 54 PfPGI *Pyrococcus furiosus* phosphoglucose isomerase; PGI, phosphoglucose isomerase
11

12
13 55

14
15 56 **Introduction**
16

17
18 57 The variety of enzymes to be found in nature is enormous, thereby providing a rich source of
19

20 58 potential biocatalysts for industrial purposes. However, in the course of natural evolution
21

22 59 these enzymes have been optimized to function optimally in *in vivo* environments, which may
23

24
25 60 differ substantially from *in vitro* industrial conditions. Therefore, often there is a need for
26

27 61 optimization of proteins for their applicability in an industrial setting. Generally this is
28

29 62 achieved by the generation of large libraries of protein variants, from which mutants with
30

31 63 improved features are selected. As screening of large libraries typically is costly and time
32

33 64 inefficient, reductions in library-size by “smart library design” is an appreciable step forward
34

35 65 ^{1,2}. The use of smart library design of a small set of promising candidates might be as equally
36

37 66 effective to improve enzyme activity, as a complete (random) library.
38
39

40
41 67 Smart library design requires identification of key residues. Such amino acids either
42

43 68 can be identified through experimental analyses, or via functional predictions using
44

45 69 bioinformatics. Comulator, a stand-alone extension of the 3DM software suite, is a recently
46

47 70 developed bioinformatics tool that uses a correlated mutation analysis (CMA) algorithm to
48

49 71 identify co-evolved residues in large structure based multiple sequence alignments (MSAs)³⁻
50

51 72 ⁷. Two very distinct roles have been linked to CMA-based residues prediction. On the one
52

53 73 hand, correlated residues are proven to represent contact positions in the protein structure and
54

55 74 as such these residues can be used for the prediction of amino acid side-chain interactions of
56

57 75 the correlated residues ⁸⁻¹⁰. However, many of these correlated residues do not contact each
58
59
60

1
2
3 76 other, and are therefore not directly involved in packing within the protein structure¹¹. On the
4
5
6 77 other hand, instead of physical contacts in the tertiary structure, correlated residues can also
7
8 78 play a role in the protein function^{6,11,12}. The functional relevance of correlated residues makes
9
10
11 79 CMA also suited for the identification of functionally relevant residues.

12
13 80 The subject of this study is the cupin super-family, a large group of structurally
14
15 81 related proteins present in all three domains of life¹³. Members of this super-family cover a
16
17 82 wide range of functions, including isomerases, dioxygenases, oxidoreductases and storage
18
19 83 proteins. The name cupin has been derived from the latin word “cupa” (small barrel),
20
21 84 reflecting the conserved beta-barrel structure. High resolution structures have been obtained
22
23 85 for many members of the cupin superfamily¹⁴. One of the best characterized members is the
24
25 86 phosphoglucose isomerase of the archaeon *Pyrococcus furiosus* (PfPGI; EC 5.3.1.9). PfPGI
26
27 87 is a glycolytic enzyme that catalyses the reversible isomerization of glucose-6-phosphate to
28
29 88 fructose-6-phosphate (F6P)^{15,16}. Several crystal structures of this homodimer (monomeric
30
31 89 subunit is 21.5 kDa) have been solved, the coordination of the catalytic metal ion has been
32
33 90 elucidated using electron paramagnetic resonance (EPR) analysis, and ample insight in the
34
35 91 catalytic mechanism of the enzyme has been gained¹⁷⁻¹⁹. Moreover, several practical features
36
37 92 make PfPGI an ideal candidate for our engineering analysis: very efficient expression in *E.*
38
39 93 *coli*, straightforward purification and activity assay, and last but not least, the native enzyme
40
41 94 is very stable.

42
43 95 In this study we describe the generation of a small library in which we randomly
44
45 96 substituted two correlated amino acids at a previously identified “hot spot” in our model
46
47 97 enzyme PfPGI (Fig.1). The predicted relevance of the correlated residues was validated by a
48
49 98 detailed analysis of selected mutants, in which specific activity, metal coordination, and
50
51 99 overall 3D structure were compared. This study shows that CMA may be useful as a guide to
52
53
54
55
56
57
58
59
60

1
2
3 100 functionally relevant sites in the periphery of a protein structure, and as such that it can be
4
5
6 101 used as a tool to identify non-obvious “hot spots” for mutagenesis.
7
8
9 102

10 103 **Materials and Methods**

11
12
13 104 Yeast glucose-6-phosphate dehydrogenase was purchased from MP biomedical. Chemicals
14
15
16 105 were purchased from Sigma-Aldrich and Roche. The PfPGI mutant library was created by
17
18 106 BaseClear (The Netherlands), the genes were cloned in expression vector pET24d (Novagen).
19
20 107

21 22 108 ***PfPGI mutant library***

23
24
25 109 The cloning of the gene *pgiA* has been described previously¹⁶. A site saturation library was
26
27 110 designed and created based on CMA using the Comulator software. The constructed library
28
29
30 111 consisted of *pgiA* variants that had alterations in two strongly correlated amino acids; proline
31
32 112 132 and tyrosine 133. The corresponding numbering in the 3DM alignment was Pro27 and
33
34 113 Tyr28 (Fig.1). The created *pgiA* variants were cloned in expression vector pET24d and used
35
36 114 to transform *E. coli*.
37
38
39 115

40 41 116 ***PfPGI expression and purification***

42
43
44 117 Starter cultures of the PfPGI mutants were inoculated from a glycerol stock and grown in
45
46 118 Luria Bertani medium supplemented with 50 $\mu\text{g ml}^{-1}$ kanamycin (LB/Km) in a 37°C shaker.
47
48
49 119 The overnight culture was used to inoculate (0.2% v/v) sterile glass tubes containing 10
50
51 120 milliliter LB/Km medium. When the optical density of the culture reached $A_{600} = 0.5$, gene
52
53 121 expression was induced by addition of 0.1 mM isopropyl-1-thio- β -D-galactopyranoside
54
55
56 122 (IPTG). Growth was continued overnight at 37°C, after which the cells were harvested by
57
58 123 centrifugation (4,600 x g for 15 min). Pelleted *E. coli* cells were resuspended in 20 mM Tris-
59
60 124 HCl buffer (pH 8.0) and disrupted by sonication. DNase was added to degrade the DNA in

1
2
3 125 the cell lysate to reduce viscosity. Cell debris was removed by centrifugation (16,000 x g for
4
5
6 126 15 min). *E. coli* proteins were denatured by heating the cell free extract at 70°C for 30 min,
7
8 127 and removed by centrifugation (16,000 x g for 15 min). The result was a heat treated cell free
9
10 128 extract containing mainly PfPGI. Its purity was checked by SDS-PAGE. Protein
11
12 129 concentrations were determined by Coomassie Brilliant Blue G250²⁰, using bovine serum
13
14 130 albumin as reference and analysis by SDS-PAGE (Quantity One®, Bio-Rad).

15
16
17 131 PfPGI was purified to homogeneity using an FPLC method similar as described
18
19 132 before⁶. Heat treated cell free extract was diluted to lower the salt concentration, filtered
20
21 133 through a 0.45 µm filter and loaded on a Q-sepharose fast flow column (Amersham
22
23 134 Pharmacia Biotech). The column was equilibrated with 20 mM Tris-HCl (pH 8.0). PGI
24
25 135 activity eluted at 180 mM of NaCl during a linear gradient of 0 to 1 M NaCl. The fraction
26
27 136 with the highest activity was loaded on a pre-equilibrated Superdex 200 GL column and
28
29 137 eluted in 20 mM Tris-HCl (pH 7.0) containing 100 mM NaCl. Protein concentrations and
30
31 138 purity were determined after which the purified enzyme fraction was used for activity assays.
32
33
34
35

36 139

37 38 39 140 ***PfPGI activity assay***

40
41 141 Any divalent metal was stripped from the purified PfPGI using 50 mM EDTA by incubated
42
43 142 at 50°C for 20 min just prior to the activity measurement. PfPGI activities were determined
44
45 143 by measuring NADPH formation in a coupled enzyme assay with yeast glucose-6-phosphate
46
47 144 dehydrogenase. This yeast enzyme was present in excess to ensure that the detection of
48
49 145 NADPH absorbance at 340 nm ($\epsilon = 6.3 \text{ mM}^{-1}\text{cm}^{-1}$) corresponded to PfPGI activity. The assay
50
51 146 mixture contained 0.5 mM NADP, 5 mM F6P and 0.35 units of D-glucose-6-phosphate
52
53 147 dehydrogenase, all in 20 mM Tris-HCl buffer (pH 7.0). All assays were performed using a
54
55 148 Hitachi U2001 spectrophotometer with a temperature controlled cuvette holder set at 50°C.
56
57
58
59
60

1
2
3 149 The optimal activity was measured after careful titration with MnCl₂, while an excess of this
4
5
6 150 salt resulted in enzyme inhibition.

7
8 151

9
10 152 ***PfPGI crystallization***

11
12
13 153 For crystallization, PfPGI was overexpressed and purified as described previously ²¹. For
14
15 154 each mutant, protein was concentrated to 11.5 mg ml⁻¹ in a solution of 10 mM Tris-HCl, pH
16
17 155 8.0, 50 mM F6P and 5 mM MnCl₂. Mutants RG and AG crystallized from hanging drops by
18
19 156 mixing equal volumes of protein solution with a reservoir solution containing 0.35 M MgCl₂,
20
21 157 0.1 M sodium acetate pH 5.5 and 10-35% PEG4000. For mutants AD and VY, crystals were
22
23 158 grown using a Hydra plus One robot, and commercial screens. AD crystallized from a
24
25 159 solution of 0.2 M calcium acetate, 0.1 M sodium acetate pH 6.5, 40% PEG300, whereas VY
26
27 160 crystallized from solutions of 0.2 M sodium nitrate, 0.1 M Bis Tris Propane pH 6.5, 20%
28
29 161 PEG4000. 50 mM F6P was added to the AD and VY crystals, before mounting. For each
30
31 162 different mutant, a single crystal was briefly washed in a cryoprotectant consisting of 25%
32
33 163 ethylene glycol in the crystallization buffer, flash cooled to 100 K and stored in liquid
34
35 164 nitrogen prior to data collection on the Diamond synchrotron light source. Data were
36
37 165 processed using the Xia2 software ²² and structures determined by molecular replacement
38
39 166 using the wild type PfPGI coordinates as a search model (pdb code 1X82)¹⁷ and the program
40
41 167 Phaser ²³. Rounds of building using Coot ²⁴ and refinement in Refmac ²⁵ gave acceptable
42
43 168 models, verified using Molprobit ²⁶. For each structure, electron density was present for all
44
45 169 the polypeptide chain and the models had no missing residues. However, density was weak
46
47 170 for the side chains of Lys21, Lys188, Lys189 (AG); Arg25, Glu114, Asp116, Lys118,
48
49 171 Lys188 and Lys189 (RG chain A); Glu114, Lys188 and Lys189 (RG chain B) and Lys188
50
51 172 and Lys189 (AD and VY, chains A and B). Data collection and refinement statistics are given
52
53 173 in Tables 1 and 2. The four mutant structures were compared to the wildtype Mn/5PAA

1
2
3 174 structure (IX7N) by superposition of all the protein atoms of the residues that coordinate the
4
5
6 175 Mn (His88, His90, Glu97 and His136).
7

8 176

9
10 177 ***EPR spectroscopy***

11
12 178 Electron paramagnetic resonance spectra were obtained from circa 5 mg ml⁻¹ samples of
13
14 179 PfPGI mutants in 10 mM Tris-HCl, pH 8.0, to which 0.2 mM of MnCl₂ was added
15
16 179 PfPGI mutants in 10 mM Tris-HCl, pH 8.0, to which 0.2 mM of MnCl₂ was added
17
18 180 anaerobically. Spectra were also taken for the PfPGI ternary complexes produced by 10 min
19
20 181 incubation with 10 mM F6P. X-band spectra were collected on a Bruker ECS-106
21
22 182 spectrometer using a microwave frequency of 9.45 GHz, a microwave power of 0.126 mW or
23
24 183 126 mW, a modulation frequency of 100 kHz, a modulation amplitude of 6.3 gauss, and a
25
26 184 sample temperature of 13 K.
27
28

29 185

30
31 186 ***Protein Data Bank accession codes***

32
33 187 The structure factors and coordinates for the four mutant PfPGI structures have been
34
35 188 deposited in the protein data bank with accession 4LTA (RG), 4LUK (AG), 4LUL (AD),
36
37 189 4LUM (VY).
38
39

40 190

41
42
43 191 **Results and Discussion**

44
45 192 A PfPGI library has been generated based on predictions made by using the Comulotor CMA
46
47 193 algorithm, as previously described ⁶. We used a refined structure-based MSA of the cupin
48
49 194 super-family, containing a total of 1711 sequences. The amino acids with the highest pair-
50
51 195 wise correlated mutation score were Pro132 and Tyr133 in PfPGI (3DM-numbers 27 and 28)
52
53 196 (Fig.1). This amino acid pair is located in a structurally conserved surface loop ⁶ that can be
54
55 197 found in most members of the cupin super-family including PfPGI (Fig.2). Previous
56
57 198 experiments have shown that a PfPGI double mutant exhibited elevated PGI activity-levels,
58
59
60

1
2
3 199 while the two single mutants were less active than wild type PfPGI ⁶. This result is not
4
5
6 200 obvious since this peripheral surface loop is not in close proximity to the catalytic residues.

7
8 201

9
10 202 ***PfPGI mutant activity levels***

11
12 203 To examine the effect on the PGI activity of the correlated residues Pro132 and Tyr133 in
13
14
15 204 more detail, we selected fifteen mutants (out of the 400 possible) that correspond to amino
16
17 205 acid pairs that are either (highly) abundant or (almost) absent within the refined cupin
18
19 206 superfamily alignment (Fig.1). Cultures of these mutants could be grown as described
20
21
22 207 previously⁶ and PfPGI expression could be induced successfully for any of the mutants. As a
23
24 208 control we included *E. coli* harbouring the empty vector (plasmid pET24d), to have a
25
26 209 correction for background protein concentrations and to exclude possible background
27
28 210 activity. Most *E. coli* proteins could be removed from the cell lysate by a heat treatment step
29
30 211 and subsequent centrifugation. PfPGI was stripped with EDTA to remove any bound divalent
31
32 212 cations and subsequently titrated with Mn²⁺ as cofactor, as this cation results in highest *in*
33
34 213 *vitro* activity ¹⁸. The resulting heat stable cell free extract was used for PfPGI activity
35
36 214 measurements, to compare activity of the selected mutants with wild type PfPGI (Fig.3).

37
38
39 215 We could detect PGI activity in the lysates of all the fifteen mutants; none of the
40
41 216 PfPGI mutants completely lost activity. The negative control (strain with empty vector) was
42
43 217 free of background activity, hence the measured PGI activity originated from PfPGI only.
44
45 218 Significant differences were detected between the specific activities of the examined PGI
46
47 219 mutants. Interestingly, we observed elevated activities for those amino acid combinations that
48
49 220 based on the CMA are abundant in the protein family alignment (for pair frequencies see
50
51 221 Figures 1 and 3). For those combinations of amino acids that are absent or less abundant than
52
53 222 wild type PfPGI in the MSA, typical activity levels were observed that were comparable or
54
55 223 lower than wild type PfPGI (amino acid pair PY). These findings suggest a positive

1
2
3 224 correlation between the natural prevalence of an amino acid pair, and the activity of the
4
5
6 225 corresponding mutant.
7

8 226 To validate these values, we selected the two mutants that had highest activity and
9
10 227 two mutants that performed similar to, or less than, wild type PfPGI. These four mutants and
11
12 228 the wild type enzyme were purified to homogeneity, to enable a precise analysis of their
13
14
15 229 specific activity. A total of five large batch cultures were grown; wild type PfPGI
16
17 230 (P132/Y133) and mutants P132A/Y133G (AG), P132R/Y133G (RG), P132A/Y133D (AD)
18
19 231 and P132V (VY). Based on the CMA, the first two mutants contain an amino acid
20
21
22 232 combination that is highly abundant, while the other two mutants represent an amino acid
23
24
25 233 combination that is not found in the MSA (VY) or at a low frequency (AD) compared to wild
26
27 234 type. The five PfPGI variants were purified to homogeneity by heat treatment and two
28
29 235 subsequent chromatography steps. The resulting pure samples were used to determine the
30
31
32 236 specific activity of each of the PfPGI variants (Fig.4).
33

34 237 Comparing the specific activities measured for the purified PfPGI variants, there is
35
36 238 good agreement with the values as described above (Fig.3), although the relative activity
37
38 239 presented there was slightly overrated, likely due to difficulties in obtaining accurate protein
39
40
41 240 concentrations in cell lysate. In comparison with the wild type PfPGI (PY) we again observed
42
43 241 an increased activity for both mutant RG and mutant AG, while both mutant VY and mutant
44
45 242 AD have a similar or decreased specific activity, respectively, compared to PY.
46
47

48 243 The intriguing question to address is the underlying molecular basis of the observed
49
50 244 differences in PGI activity. Obviously, the catalytic site should be examined. In addition, the
51
52 245 differences in activity might relate to the PGI metal binding site. Removal of the divalent
53
54
55 246 metal co-factor results in complete loss of PGI activity, which can be restored by the addition
56
57
58 247 of divalent metals. Despite the fact that the surface loop carrying the correlated mutations is
59
60 248 located rather distant from the metal binding site and the catalytic site, it is tempting to

1
2
3 249 speculate that the mutations in this loop have an effect on the metal binding site and the
4
5 250 catalytic site and hence lead to the observed differences in activity.
6
7

8 251

9
10 252 ***Structure analysis of the mutation carrying loop***

11
12 253 To further examine possible conformational changes, for instance in the active site structure
13
14 254 and metal coordination, crystallization trials were initiated for the four PfPGI mutants. First
15
16 255 crystallization attempts were set up with manganese as incorporated cofactor and F6P as
17
18 256 substrate. These co-crystallization trials successfully yielded well-diffracting crystals for both
19
20 257 the mutants RG and AG (Table 1). The AG mutant structure is the highest resolution (1.4 Å)
21
22 258 for any PfPGI variant, and interestingly, the electron density map for this structure (and also
23
24 259 for the RG structure) clearly showed that 5-phosphoarabinonic acid (5PAA), rather than F6P
25
26 260 (as added to the crystallization mixture) was bound in the active site (Fig. 5). Thus an
27
28 261 unexpected conversion had occurred: during the experiment F6P had been, at least partially,
29
30 262 oxidized to 5PAA, resulting in preferential binding for 5PAA in the active site of both mutant
31
32 263 structures (Table 2). To confirm that 5PAA had indeed been produced, a solution of the same
33
34 264 composition as the crystallization solution was analysed by mass spectrometry after being left
35
36 265 at room temperature for one week. The spectra contained a small peak of m/e ratio 259,
37
38 266 (F6P), but also many other peaks with m/e ratios less than F6P, including a large peak with
39
40 267 m/e ratio of 245 corresponding to 5PAA, clearly indicating that a breakdown of the sugar had
41
42 268 taken place. The oxidation of F6P to 5PAA in the presence of permanganate has been
43
44 269 observed before ²⁷ and thus we presume a similar reaction occurs in the crystallization
45
46 270 solution.

47
48 271 For the other two mutants (VY and AD), no ternary complex structures were obtained by co-
49
50 272 crystallization, and thus protein crystals were grown in the presence of MgCl₂, and
51
52 273 subsequently soaked in a solution of F6P for two hours prior to X-ray data collection. This
53
54
55
56
57
58
59
60

1
2
3 274 approach was successful for mutant VY, resulting in a structure with F6P in the active site.
4
5
6 275 For mutant AD, the only structure that could be obtained contained solely Mn^{2+} in the active
7
8 276 site. (Table 2).
9

10 277 The crystal structures of the four mutants reveal that in each structure clear electron
11
12 278 density was present for the 132-133 loop that carries the substitutions. This demonstrates that
13
14
15 279 this loop is not disordered in any of the structures, and that the native fold of PfPGI can
16
17 280 accommodate the substitutions within its structure, as predicted by the Comulotor software.
18
19 281 However, as might be expected, a number of differences are noted when comparing the loop
20
21 282 structure in detail (Fig.6). In the wild type structure (P132/Y133) one face of the side chain of
22
23 283 Pro132 packs against the Arg95, Ala96 peptide and the carbonyl, $C\alpha$ and $C\beta$ of Asp94, with
24
25 284 the carboxyl of Asp94 pointing away from Pro132. The side chain of Tyr133 packs against
26
27 285 the main chain of Leu93, with additional interactions forming between the edge of the phenyl
28
29 286 moiety and the side chain of Leu93. The other face of the Pro132-Tyr133 loop packs against
30
31 287 the side chains of residues Tyr3 and the aliphatic part of Lys4, from the second subunit in the
32
33 288 PGI dimer. In the four mutant structures, the main chain conformation of the 131-134 loop
34
35 289 remains remarkably consistent, with only minor changes in the position of the main chain
36
37 290 atoms. However, the changes in the side chains of Pro132 and Tyr133 have more marked
38
39 291 effects on the positions of residues 92-94 of the 90-96 loop, and also on the position of the N-
40
41 292 terminal 5 residues from the adjacent subunit of the dimer. There are also some small
42
43 293 consequential changes in the positions of second shell residues packing against these two
44
45 294 loops.
46
47
48
49
50
51
52

53 295 For the P132R-Y133G mutant, the loss of the tyrosine side chain at position 133, is
54
55 296 somewhat alleviated by the side chain of R132 occupying approximately the same position in
56
57 297 the structure. There are movements of up to 0.5\AA in the positions of both the Leu93-Asp94
58
59 298 and Tyr3-Lys4 loops, compared to the wild-type structure. In the P132A-Y133G mutant, the
60

1
2
3 299 changes to the position of Leu93 and Asp94 are more marked, with movements of 1.8Å and
4
5
6 300 1.6Å for the alpha carbons of Leu93 and Asp94 compared to the wild type structure. Given
7
8 301 that both mutations in this AG structure are to smaller residues than those in the wild type,
9
10 302 these fairly large movements are, somewhat counter-intuitively, away from the 132-133,
11
12 303 presumably making the packing worse. In the P132V single mutant (VY), the change of the
13
14 304 proline side chain to valine, pushes the Leu93-Asp94 loop away, in order to accommodate the
15
16 305 larger bifurcated side chain of valine. Movements of 0.4Å and 0.7Å are seen between the Cαs
17
18 306 of Leu93 and Asp94, respectively. In mutant AD, the movement of the alpha carbons of
19
20 307 Leu93 and Asp94 away from 132-133 is 0.7Å and 0.9Å, respectively. In this mutant AD, the
21
22 308 change from tyrosine to the negatively charged aspartic acid has had little effect on the
23
24 309 position of the side chain of Tyr3 from the adjacent subunit.
25
26
27
28
29
30
31

311 *Structure analysis of the manganese coordination*

32
33
34 312 In all four PGI mutant structures, the manganese is 6-coordinated in an octahedral geometry
35
36 313 (Fig.7). Three of the ligands are the imidazole nitrogens of residues His88, His90 and His136.
37
38 314 The fourth ligand is one of the carboxyl oxygens of Glu97. The 5th and 6th ligands are
39
40 315 different depending on the substrate. In mutant AD, water molecules provide these two
41
42 316 ligands. In the two structures with 5PAA, mutants RG and AG, the 5th ligand is one of the
43
44 317 carboxylate oxygens of 5PAA and the 6th ligand a water molecule. In the F6P soaked crystal
45
46 318 structure, mutant VY, both the 5th and 6th ligands are provided by the F6P substrate, one is
47
48 319 the C2 carbonyl and the other the C1 hydroxyl. In this short F6P soak crystal structure, there
49
50 320 is no indication that any of the F6P has been turned over to G6P, as the electron density
51
52 321 clearly shows the C2 carbon to have trigonal (sp^2) geometry, and the C1 carbonyl to be
53
54 322 tetrahedral (sp^3), indicating the presence of the ketone isomer of the substrate.
55
56
57
58
59
60

1
2
3 323 Despite these quite large changes in the relative positions of the 132-133 and 93-94
4
5
6 324 loops between the different mutants, the position of the manganese coordinating residue
7
8 325 His90, which lies only three residues away from the moving residue Leu93, is very similar in
9
10 326 all the structures (Table 3). The same is true for His88, Glu97 and His136. Thus, these
11
12 327 mutations, whilst altering the structure local to the mutated residues, seem to have little effect
13
14
15 328 on the coordination of the manganese or the general architecture of the active site. The
16
17 329 changes seen in the activities between these different mutants are thus presumably due to the
18
19
20 330 accumulative effect of many small changes between the structures.
21

22 331 Based on the obtained crystal structures of the wild type PGI and the 4 variants, we
23
24 332 can conclude that differences in manganese coordination are, at most, very subtle. For a more
25
26 333 sensitive investigation of the coordination state of the bound manganese during catalysis,
27
28 334 EPR spectral analyses have been performed of the four PfPGI mutants, comparing PfPGI
29
30 335 without substrate as well as in complex with F6P (Fig.8). In agreement with previous results
31
32 336¹⁹, we find that the addition of F6P to any of the manganese-containing PfPGI mutants, leads
33
34 337 to a collapse of the hexacoordinate manganese signal. Additionally, there is a considerable
35
36 338 change of the signal from pentacoordinate manganese, leading to a substantial increase in the
37
38 339 pentacoordinate over hexacoordinate ratio. All this indicates that the manganese coordination
39
40 340 of the apo-enzyme shifts towards pentacoordinate upon F6P binding, very similar to what has
41
42 341 been observed and reported previously for the wild type PfPGI¹⁹. Therefore, it can be
43
44 342 concluded that the metal binding site is not significantly changed by any of the substitutions.
45
46
47
48
49

50 343

53 344 **Conclusion**

54
55
56 345 Rational design-based protein engineering, aiming for improved enzyme activity, often builds
57
58 346 on changes in (non-catalytic) residues located in close proximity to the catalytic site and/or
59
60 347 substrate/cofactor binding residues, as these residues generally are more easy to predict²⁸.

1
2
3 348 Although multiple examples exist where this has been proven successful ²⁹⁻³², focussing on
4
5
6 349 only these residues misses additional opportunities for protein improvement. In our specific
7
8 350 case, the loop containing residues 132/133 is not an obvious pick based on its distance to the
9
10 351 catalytic site and cofactor binding site. Nonetheless, alterations in this loop do have the
11
12 352 potential to improve PGI activity levels.

13
14
15 353 Comparing the mutant PGI activity levels, a general trend could be observed.
16
17 354 Although the differences in activity are rather subtle, it could be shown that those amino acid
18
19 355 correlations that have a high occurrence typically result in a variant with a higher activity
20
21 356 than wild type, whereas those amino acid correlations that have a low occurrence generally
22
23 357 result in a mutant with a decreased activity compared to wild type.
24
25
26

27 358 Admitted, the mutations we introduced did only result in a rather small increase in
28
29 359 specific activity. However, selection of a mutant with a multi-fold increase in activity was not
30
31 360 the primary goal. Remarkably, none of the 15 tested PfPGI mutants completely lost activity.
32
33 361 Moreover, despite changes in the activity, the protein structure and stability seemed
34
35 362 consistently without detectable changes. For the 4 crystallized mutants there is no evidence
36
37 363 for disorder in the mutated residues or for the adjacent stretches of the polypeptide. This
38
39 364 contrasts favourably with other studies where generally many mutations of this magnitude
40
41 365 result in (local) disorder in the mutated residues, or for the adjacent stretches of the
42
43 366 polypeptide. Site-directed mutagenesis of residues that lead to improved catalytic properties
44
45 367 often affect the protein-structure stability ³³⁻³⁵. Therefore, additional compensatory mutations
46
47 368 are often required to neutralize this destabilization ^{36,37}. Evaluating the data presented here,
48
49 369 we can conclude that this is a very promising result for Comulator-based CMA predictions,
50
51 370 validating the Comulator software to predict stable mutations within the PGI structural
52
53 371 framework.
54
55
56
57
58
59
60

1
2
3 372 No significant differences were found in the structures, neither in the substrate
4
5
6 373 binding pocket, nor in the coordination of the catalytic manganese. However, minor
7
8 374 differences in the conformation and packing of the loops surrounding the mutated residues
9
10 375 were observed between the different PfPGI variants. These subtle differences in packing
11
12 376 presumably propagate through the structure to result in small differences in substrate binding
13
14
15 377 and/or catalytic efficiency, resulting in the differences in activity seen between the mutants.
16
17 378 Another explanation for the changed activity may be that the substituted loop residues
18
19 379 somehow affect (either positively or negatively) the enzymes' flexibility; something that
20
21 380 cannot be observed in a static condition like a crystal structure.
22
23

24
25 381 To conclude, we showed that CMA based predictions have the potential to identify
26
27 382 hotspots that are possibly interesting target residues for substitutions, as they may lead to
28
29 383 improved protein performance. This is a very different application of CMAs, compared to
30
31 384 using CMAs for predictions of residue contact points in a protein structure⁸⁻¹⁰, and as such as
32
33 385 a tool for the *ab initio* prediction of protein structures³⁸⁻⁴¹. We think that caution is required
34
35 386 in the selection of correlated residues for such structural prediction tools, as we show that this
36
37 387 specific, strongly correlated, amino acid pair in the cupin superfamily has no important effect
38
39 388 in the protein structure, thereby providing an obvious example of a CMA-based correlated
40
41 389 residue prediction containing functional information rather than structural information. All in
42
43 390 all, the presented results show that CMAs also have the potential to be used for generation of
44
45 391 small size, "smart" libraries containing beneficial variants. The selected mutations are based
46
47 392 on changes in highly correlated amino acids. As these residues may be distantly located from
48
49 393 the enzyme's active site, and as such have no obvious relation to the enzymes' performance,
50
51 394 they may easily be overlooked in other rational design approaches.
52
53
54
55
56
57
58
59
60

396 **References**

- 397 1. Chica RA, Doucet N, Pelletier JN. Semi-rational approaches to engineering enzyme
398 activity: combining the benefits of directed evolution and rational design. *Curr Opin*
399 *Biotechnol* 2005;16(4):378-384.
- 400 2. Lutz S. Beyond directed evolution--semi-rational protein engineering and design.
401 *Curr Opin Biotechnol* 2010;21(6):734-743.
- 402 3. Alcolombri U, Elias M, Tawfik DS. Directed Evolution of Sulfotransferases and
403 Paraoxonases by Ancestral Libraries. *Journal of Molecular Biology* 2011;411(4):837-
404 853.
- 405 4. Jochens H, Bornscheuer UT. Natural diversity to guide focused directed evolution.
406 *ChemBiochem* 2010;11(13):1861-1866.
- 407 5. Kuipers RK, Joosten HJ, van Berkel WJ, Leferink NG, Rooijen E, Ittmann E, van
408 Zimmeren F, Jochens H, Bornscheuer U, Vriend G, dos Santos VA, Schaap PJ. 3DM:
409 systematic analysis of heterogeneous superfamily data to discover protein
410 functionalities. *Proteins* 2010;78(9):2101-2113.
- 411 6. Kuipers RK, Joosten HJ, Verwiël E, Paans S, Akerboom J, van der Oost J, Leferink
412 NG, van Berkel WJ, Vriend G, Schaap PJ. Correlated mutation analyses on super-
413 family alignments reveal functionally important residues. *Proteins* 2009;76(3):608-
414 616.
- 415 7. Sullivan BJ, Nguyen T, Durani V, Mathur D, Rojas S, Thomas M, Syu T, Magliery
416 TJ. Stabilizing Proteins from Sequence Statistics: The Interplay of Conservation and
417 Correlation in Triosephosphate Isomerase Stability. *Journal of Molecular Biology*
418 2012;420(4-5):384-399.
- 419 8. Kundrotas PJ, Alexov EG. Predicting residue contacts using pragmatic correlated
420 mutations method: reducing the false positives. *BMC Bioinformatics* 2006;7:503.

- 1
2
3 421 9. Mintseris J, Weng Z. Structure, function, and evolution of transient and obligate
4
5
6 422 protein-protein interactions. *Proc Natl Acad Sci U S A* 2005;102(31):10930-10935.
7
8 423 10. Pazos F, Helmer-Citterich M, Ausiello G, Valencia A. Correlated mutations contain
9
10 424 information about protein-protein interaction. *J Mol Biol* 1997;271(4):511-523.
11
12
13 425 11. Kowarsch A, Fuchs A, Frishman D, Pagel P. Correlated mutations: a hallmark of
14
15 426 phenotypic amino acid substitutions. *PLoS Comput Biol* 2010;6(9).
16
17
18 427 12. Gloor GB, Martin LC, Wahl LM, Dunn SD. Mutual information in protein multiple
19
20 428 sequence alignments reveals two classes of coevolving positions. *Biochemistry*
21
22 429 2005;44(19):7156-7165.
23
24
25 430 13. Dunwell JM, Khuri S, Gane PJ. Microbial relatives of the seed storage proteins of
26
27 431 higher plants: conservation of structure and diversification of function during
28
29 432 evolution of the cupin superfamily. *Microbiol Mol Biol Rev* 2000;64(1):153-179.
30
31
32 433 14. Dunwell JM, Purvis A, Khuri S. Cupins: the most functionally diverse protein
33
34 434 superfamily? *Phytochemistry* 2004;65(1):7-17.
35
36
37 435 15. Hansen T, Oehlmann M, Schonheit P. Novel type of glucose-6-phosphate isomerase
38
39 436 in the hyperthermophilic archaeon *Pyrococcus furiosus*. *J Bacteriol*
40
41 437 2001;183(11):3428-3435.
42
43
44 438 16. Verhees CH, Huynen MA, Ward DE, Schiltz E, de Vos WM, van der Oost J. The
45
46 439 phosphoglucose isomerase from the hyperthermophilic archaeon *Pyrococcus furiosus*
47
48 440 is a unique glycolytic enzyme that belongs to the cupin superfamily. *J Biol Chem*
49
50 441 2001;276(44):40926-40932.
51
52
53 442 17. Berrisford JM, Akerboom J, Brouns S, Sedelnikova SE, Turnbull AP, van der Oost J,
54
55 443 Salmon L, Hardre R, Murray IA, Blackburn GM, Rice DW, Baker PJ. The structures
56
57 444 of inhibitor complexes of *Pyrococcus furiosus* phosphoglucose isomerase provide
58
59 445 insights into substrate binding and catalysis. *J Mol Biol* 2004;343(3):649-657.
60

- 1
2
3 446 18. Berrisford JM, Akerboom J, Turnbull AP, de Geus D, Sedelnikova SE, Staton I,
4
5
6 447 McLeod CW, Verhees CH, van der Oost J, Rice DW, Baker PJ. Crystal structure of
7
8 448 Pyrococcus furiosus phosphoglucose isomerase. Implications for substrate binding
9
10 449 and catalysis. *J Biol Chem* 2003;278(35):33290-33297.
11
12 450 19. Berrisford JM, Hounslow AM, Akerboom J, Hagen WR, Brouns SJ, van der Oost J,
13
14 451 Murray IA, Michael Blackburn G, Waltho JP, Rice DW, Baker PJ. Evidence
15
16 452 supporting a cis-enediol-based mechanism for Pyrococcus furiosus phosphoglucose
17
18 453 isomerase. *J Mol Biol* 2006;358(5):1353-1366.
19
20 454 20. Bradford MM. A rapid and sensitive method for the quantitation of microgram
21
22 455 quantities of protein utilizing the principle of protein-dye binding. *Anal Biochem*
23
24 456 1976;72:248-254.
25
26 457 21. Akerboom J, Turnbull AP, Hargreaves D, Fisher M, de Geus D, Sedelnikova SE,
27
28 458 Berrisford JM, Baker PJ, Verhees CH, van der Oost J, Rice DW. Purification,
29
30 459 crystallization and preliminary crystallographic analysis of phosphoglucose isomerase
31
32 460 from the hyperthermophilic archaeon Pyrococcus furiosus. *Acta Crystallogr D Biol*
33
34 461 *Crystallogr* 2003;59(Pt 10):1822-1823.
35
36 462 22. Winter G. xia2: an expert system for macromolecular crystallography data reduction.
37
38 463 *Journal of Applied Crystallography* 2010;43(1):186-190.
39
40 464 23. McCoy AJ, Grosse-Kunstleve RW, Adams PD, Winn MD, Storoni LC, Read RJ.
41
42 465 Phaser crystallographic software. *Journal of Applied Crystallography* 2007;40(4):658-
43
44 466 674.
45
46 467 24. Emsley P, Cowtan K. Coot: model-building tools for molecular graphics. *Acta*
47
48 468 *Crystallogr D Biol Crystallogr* 2004;60(Pt 12 Pt 1):2126-2132.
49
50
51
52
53
54
55
56
57
58
59
60

- 1
2
3 469 25. Murshudov GN, Vagin AA, Dodson EJ. Refinement of Macromolecular Structures by
4
5
6 470 the Maximum-Likelihood Method. *Acta Crystallographica Section D* 1997;53(3):240-
7
8 471 255.
- 9
10 472 26. Chen VB, Arendall WB, 3rd, Headd JJ, Keedy DA, Immormino RM, Kapral GJ,
11
12
13 473 Murray LW, Richardson JS, Richardson DC. MolProbity: all-atom structure
14
15 474 validation for macromolecular crystallography. *Acta Crystallogr D Biol Crystallogr*
16
17 475 2010;66(Pt 1):12-21.
- 18
19
20 476 27. Azmat R, Naz R, Qamar N, Malik I. Kinetics and mechanisms of oxidation of d-
21
22 477 fructose and d-lactose by permanganate ion in acidic medium. *Natural Science*
23
24 478 2012;4(1):466-478.
- 25
26
27 479 28. Morley KL, Kazlauskas RJ. Improving enzyme properties: when are closer mutations
28
29 480 better? *Trends in Biotechnology* 2005;23(5):231-237.
- 30
31 481 29. Hill CM, Li WS, Thoden JB, Holden HM, Raushel FM. Enhanced degradation of
32
33 482 chemical warfare agents through molecular engineering of the phosphotriesterase
34
35 483 active site. *J Am Chem Soc* 2003;125(30):8990-8991.
- 36
37
38 484 30. Koga Y, Kato K, Nakano H, Yamane T. Inverting enantioselectivity of Burkholderia
39
40 485 cepacia KWI-56 lipase by combinatorial mutation and high-throughput screening
41
42 486 using single-molecule PCR and in vitro expression. *J Mol Biol* 2003;331(3):585-592.
- 43
44
45 487 31. Machielsen R, Looger LL, Raedts J, Dijkhuizen S, Hummel W, Hennemann H-G,
46
47 488 Dausmann T, van der Oost J. Cofactor engineering of *Lactobacillus brevis* alcohol
48
49 489 dehydrogenase by computational design. *Eng Life Sci* 2009;9(1):38-44.
- 50
51
52 490 32. Reetz MT, Bocola M, Carballeira JD, Zha D, Vogel A. Expanding the Range of
53
54 491 Substrate Acceptance of Enzymes: Combinatorial Active-Site Saturation Test.
55
56 492 *Angewandte Chemie International Edition* 2005;44(27):4192-4196.
57
58
59
60

- 1
2
3 493 33. Meiering EM, Serrano L, Fersht AR. Effect of active site residues in barnase on
4
5
6 494 activity and stability. *J Mol Biol* 1992;225(3):585-589.
7
8 495 34. Shoichet BK, Baase WA, Kuroki R, Matthews BW. A relationship between protein
9
10 496 stability and protein function. *Proc Natl Acad Sci U S A* 1995;92(2):452-456.
11
12 497 35. Tokuriki N, Stricher F, Serrano L, Tawfik DS. How protein stability and new
13
14 498 functions trade off. *PLoS Comput Biol* 2008;4(2):e1000002.
15
16
17 499 36. Studer RA, Dessailly BH, Orengo CA. Residue mutations and their impact on protein
18
19 500 structure and function: detecting beneficial and pathogenic changes. *Biochem J*
20
21 501 2013;449(3):581-594.
22
23
24 502 37. Tokuriki N, Tawfik DS. Stability effects of mutations and protein evolvability. *Curr*
25
26 503 *Opin Struct Biol* 2009;19(5):596-604.
27
28
29 504 38. Burger L, van Nimwegen E. Disentangling direct from indirect co-evolution of
30
31 505 residues in protein alignments. *PLoS Comput Biol* 2010;6(1):e1000633.
32
33
34 506 39. Jones DT, Buchan DW, Cozzetto D, Pontil M. PSICOV: precise structural contact
35
36 507 prediction using sparse inverse covariance estimation on large multiple sequence
37
38 508 alignments. *Bioinformatics* 2012;28(2):184-190.
39
40
41 509 40. Marks DS, Colwell LJ, Sheridan R, Hopf TA, Pagnani A, Zecchina R, Sander C.
42
43 510 Protein 3D structure computed from evolutionary sequence variation. *PLoS One*
44
45 511 2011;6(12):e28766.
46
47
48 512 41. Marks DS, Hopf TA, Sander C. Protein structure prediction from sequence variation.
49
50 513 *Nat Biotechnol* 2012;30(11):1072-1080.
51
52
53
54 514
55
56
57 515
58
59 516
60

1
2
3
4 517 **Figures**

5
6 518 Figure 1. Residue pair frequency table of amino acid couple 27 and 28 (3DM numbering). In
7
8 519 wild type PfPGI these residues are Pro132 and Tyr133. The occurrences are relative to the
9
10 520 number of unique sequences in the super family alignment.
11
12

13 521
14
15 522 Figure 2. Cartoon representation of the dimer (white and green subunits) of the wild type
16
17 523 PfPGI Mn²⁺/5PAA 3D-structure (PDB code: 1X7N). The correlated amino acid pair “PY” is
18
19 524 indicated in red. Shown in yellow are those residues involved in metal ion (purple sphere)
20
21 525 binding, including a water molecule (red sphere). The inhibitor 5-phospho-D-arabinonate
22
23 526 (5PAA) is shown as a stick model (green).
24
25
26

27 527
28
29 528 Figure 3. Graphical representation of the relative activity of each PfPGI mutant (Y-axis)
30
31 529 compared to the amino acid pair occurrence according the Comulotor based CMA predictions
32
33 530 (X-axis). All PGI activities are relative to wild type PfPGI (PY; in bold).
34
35
36

37 531
38
39 532 Figure 4. Specific activity of wild type PfPGI (PY) compared to selected high
40
41 533 occurrence/activity mutants RG and AG, and low occurrence/activity mutants AD and VY.
42
43 534 Manganese was used as co-factor and added via titration.
44
45
46

47 535
48
49 536 Figure 5. A stereo representation of the mF_O – DF_C electron density (grey mesh), contoured
50
51 537 at 1.5σ, for the AG mutant PfPGI structure, showing the manganese (purple sphere),
52
53 538 coordinating water (red sphere), bound 5PAA and metal coordinating residues (sticks).
54
55

56 539
57
58 540 Figure 6. Comparison of the structure of the loops adjacent to the mutation position in the
59
60 541 four mutant structures. In each panel the wild type structure (1X7N) is shown in white, with

1
2
3 542 relevant residues highlighted in stick representation and labelled. Y3' and K4' refer to the N-
4
5
6 543 terminal strand from the adjacent subunit in the dimer. (a) Mutant AD (slate), (b) mutant RG
7
8 544 (wheat), (c) mutant AG (yellow) and (d) mutant VY (pink).

9
10 545
11
12
13 546 Figure 7. Mn^{2+} coordination in the mutant PfPGI structures. For all mutant structures, the
14
15 547 Mn^{2+} ion (purple sphere) is coordinated in an octahedral arrangement: (a) mutant AD (blue),
16
17 548 (b) mutant RG (wheat), (c) mutant AG (yellow) and (d) mutant VY (pink). Ligands to the
18
19 549 metal are highlighted in stick representation, water molecules are shown as red spheres.

20 550
21
22
23
24
25 551 Figure 8. Invariance of EPR spectra from the PfPGI mutants. The red traces were taken at
26
27 552 low microwave power (0.126 mW) for a full record of the six-line pattern around 3300 gauss
28
29 553 typical for hexacoordinate Mn^{2+} . The black traces were taken at high microwave power (126
30
31 554 mW) to emphasize the broad features over the whole magnetic-field range from
32
33 555 pentacoordinate Mn^{2+} . The left-hand traces are in the absence of substrate, while the right-
34
35 556 hand traces are in the presence of 10 mM F6P.

36
37
38
39 557
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Table 1 - Data collection statistics

Data set	AG	RG	AD	VY
Spacegroup	C2	P2 ₁ 2 ₁ 2 ₁	P2 ₁ 2 ₁ 2 ₁	P1
Unit cell parameters (Å)				
a	87.9	73.0	46.2	80.7
b	43.3	74.5	46.2	42.6
c	58.5	75.9	186.2	88.9
α	90.0°	90.0°	90.0°	90.0°
β	122.0°	90.0°	90.0°	104.9°
γ	90.0°	90.0°	120.0°	90.0°
Molecules per ASU	1	2	2	2
Resolution (Å) ¹	28.92-1.41 (1.44 - 1.41)	26.58-2.04 (2.09 -2.04)	25.58-1.89 (1.94 -1.89)	18.49-1.79 (1.84-1.79)
Wavelength (Å)	0.97630	0.97630	0.97630	0.97630
Unique observations ¹	32797 (1460)	26994 (1969)	*****	33074 (2386)
R _{pim} ¹	0.036 (0.349)	0.074 (0.219)	0.028 (0.352)	0.086 (0.349)
Completeness (%) ¹	91.2 (55.4)	99.6 (99.9)	98.5 (99.2)	95.7 (94.0)
Multiplicity ¹	3.1 (2.4)	6.3 (6.4)	3.5 (3.5)	1.8 (1.8)
Mean(I)/sd(I)	13.0 (2.1)	7.0 (3.7)	15.9 (2.4)	4.3 (2.2)

¹ Numbers in parentheses indicate values for the highest resolution shell

Table 2 - Refinement Statistics

Model	AG	RG	AD	VY
Resolution (Å)	1.4	2.0	1.9	1.8
Number of reflections	31134	25635	27836	31398
Protein molecules per asymmetric unit	1	2	2	2
Number of atoms	1739	3221	3142	3308
Number of waters	197	141	112	218
Number of Mn ²⁺ ions	1	2	2	2
Number of F6P	0	0	0	2
Number of 5PAA	1	2	0	0
Ramachandran favoured (%)	98.4	87.1	97.3	97.9
Ramachandran outliers (%)	0	0.3	0.3	0
Poor rotamers (%)	1.3	1.6	0.6	0.3
RMSD bond (Å)	0.006	0.007	0.011	0.009
RMSD angle (°)	1.06	1.08	1.34	1.26
Average B-factors (Å ²)				
Main chain	21	35	31	24
Side chain	31	24	34	27
Waters	37	36	33	26
Mn ²⁺	13	27	31	21
5PAA/F6P	15	33	-	26
R-factor	0.14	0.22	0.19	0.17
R-Free	0.20	0.29	0.25	0.22
MolProbity score	0.80	0.94	0.99	1.03
	100 th percentile	100 th percentile	100 th percentile *	100 th percentile

Table 3 - Mn²⁺ ligands and coordination distances (Å).

Mutant	His 88-N2	His 90-N2	His 136-N2	Glu 97-
PY WT (1X81)	2.45	2.25	2.29	2.26-O2
RG	2.20	2.30	2.29	2.17-O1
AG	2.32	2.25	2.23	2.33-O1
AD	2.24	2.21	2.35	1.98-O1
VY	2.31	2.31	2.28	2.09-O1

Figure 1.

Position 28

%	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	0.12	0	0.18	0.12	0	11.40	0	0	0	0	0	0.06	0	0	0	0.18	0.06	0	0	0
C	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
D	0	0	0	0	0	0.12	0	0	0	0	0	0.12	0	0	0	0	0	0	0	0
E	0	0	0	0	0	7.80	0	0	0	0	0.12	0	0	0	0	0.06	0	0.06	0	0
F	0.43	0	0	0	0	0.06	0	0	0	0	0	0	0	0	0	0	0	0	0	0
G	1.65	0	26.26	0.18	0	1.10	0.06	0	0	0.24	0	0.30	0	0	0	0.61	1.16	0	0	0.06
H	0	0	0	0	0	0.06	0	0	0	0	0	0	0	0.06	0	0	0	0	0	0
I	0	0	0.18	0	0	1.16	0	0	0	0	0	0.85	0	0	0	0	0.06	0	0	0
K	0.61	0	0	0	0	4.51	0	0	0	0	0	0.12	0	0	0	0.30	0	0	0	0
L	0	0	0	0	0	0.30	0.06	0	0	0	0	0	0	0	0	0	0	0	0	0
M	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.06	0	0	0	0
N	0.18	0.06	0.06	0	0.30	0.06	0.06	0	0	0.85	0.06	0.06	0	0.06	10.42	0.18	0.06	0.12	0	0.18
P	0	0	0	0	0.12	3.41	0	0	0	0	0	0	0	0	0	0.06	0.06	0	0	0.79
Q	0.18	0	0	0	0	3.53	0.06	0	0	0.12	0	2.25	0.06	0	0	0.12	0	0	0	0.06
R	0.12	0	0.06	0	0	10.24	0.06	0	0	0	0	0.24	0	0	0.06	0	0	0	0	0
S	0	0	0	0	0	0.73	0	0	0	0	0	0.06	0	0	0	0	0	0	0	0
T	0.24	0	0	0	0	0.06	0	0	0	0	0	0	0	0	0	0	0	0	0	0
V	0	0	0.12	0	0	1.77	0	0	0	0	0	1.40	0	0	0	0.12	0	0	0	0
W	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Y	0	0	0.06	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Figure 2.

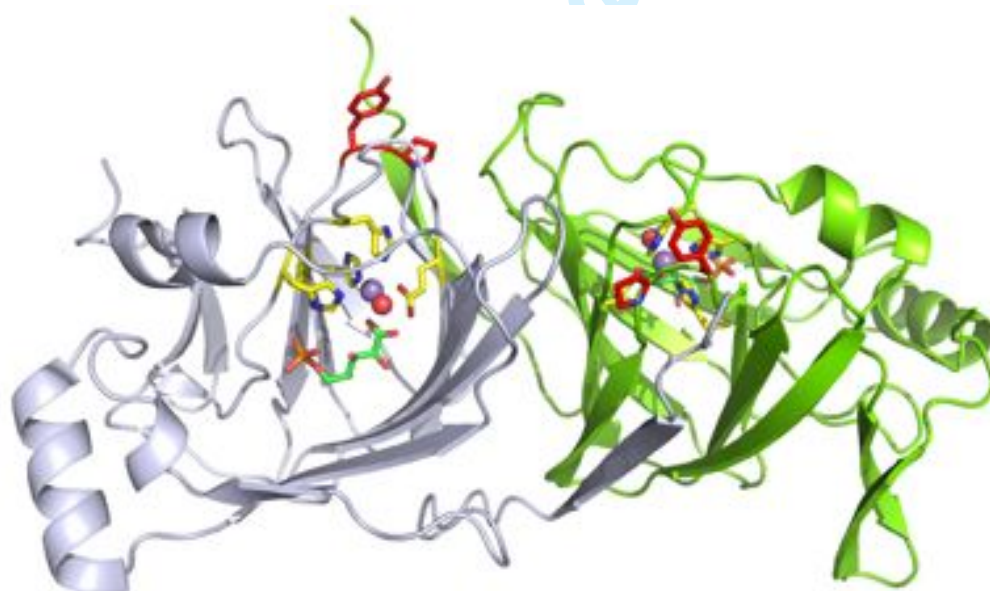


Figure 3.

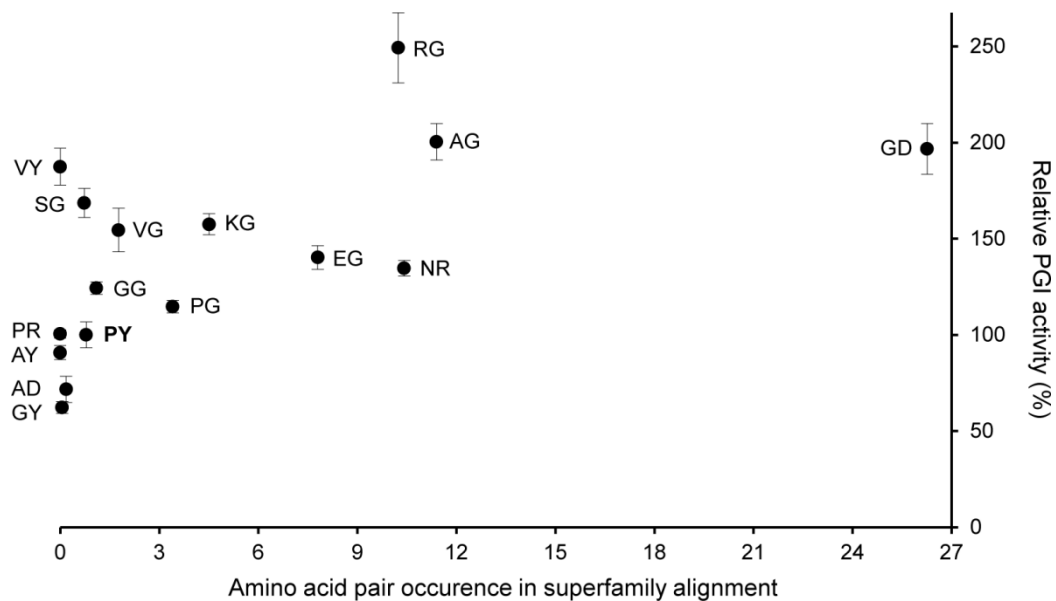


Figure 4.

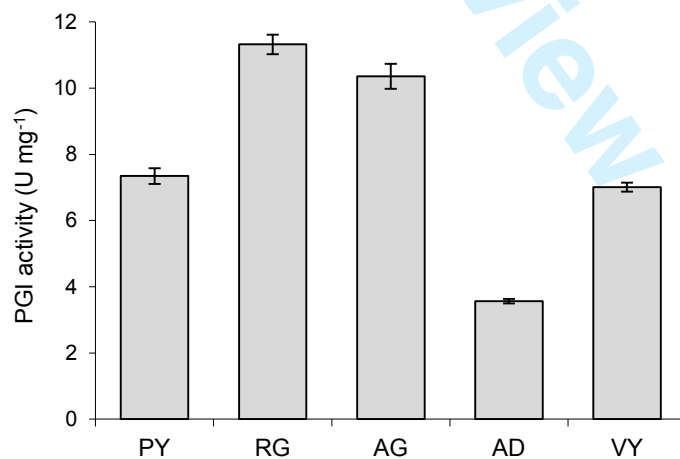
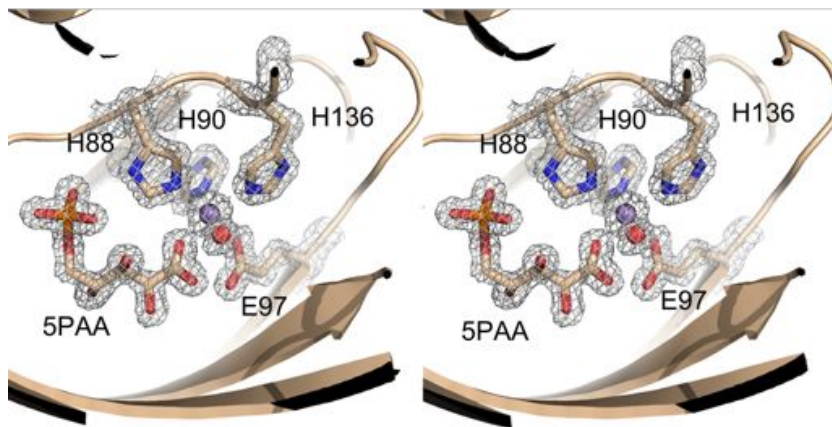


Figure 5.



For Peer Review

Figure 6.

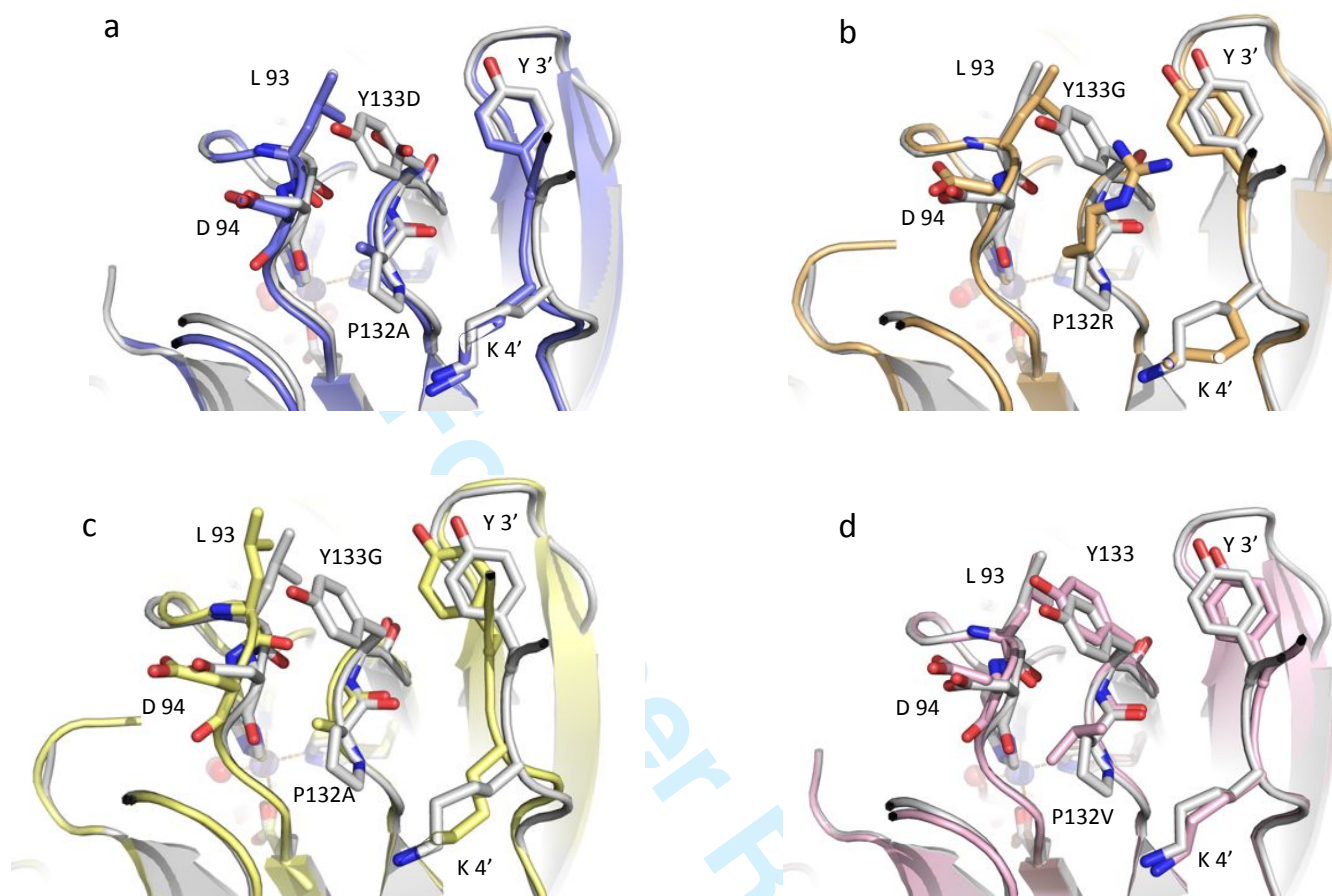


Figure 7.

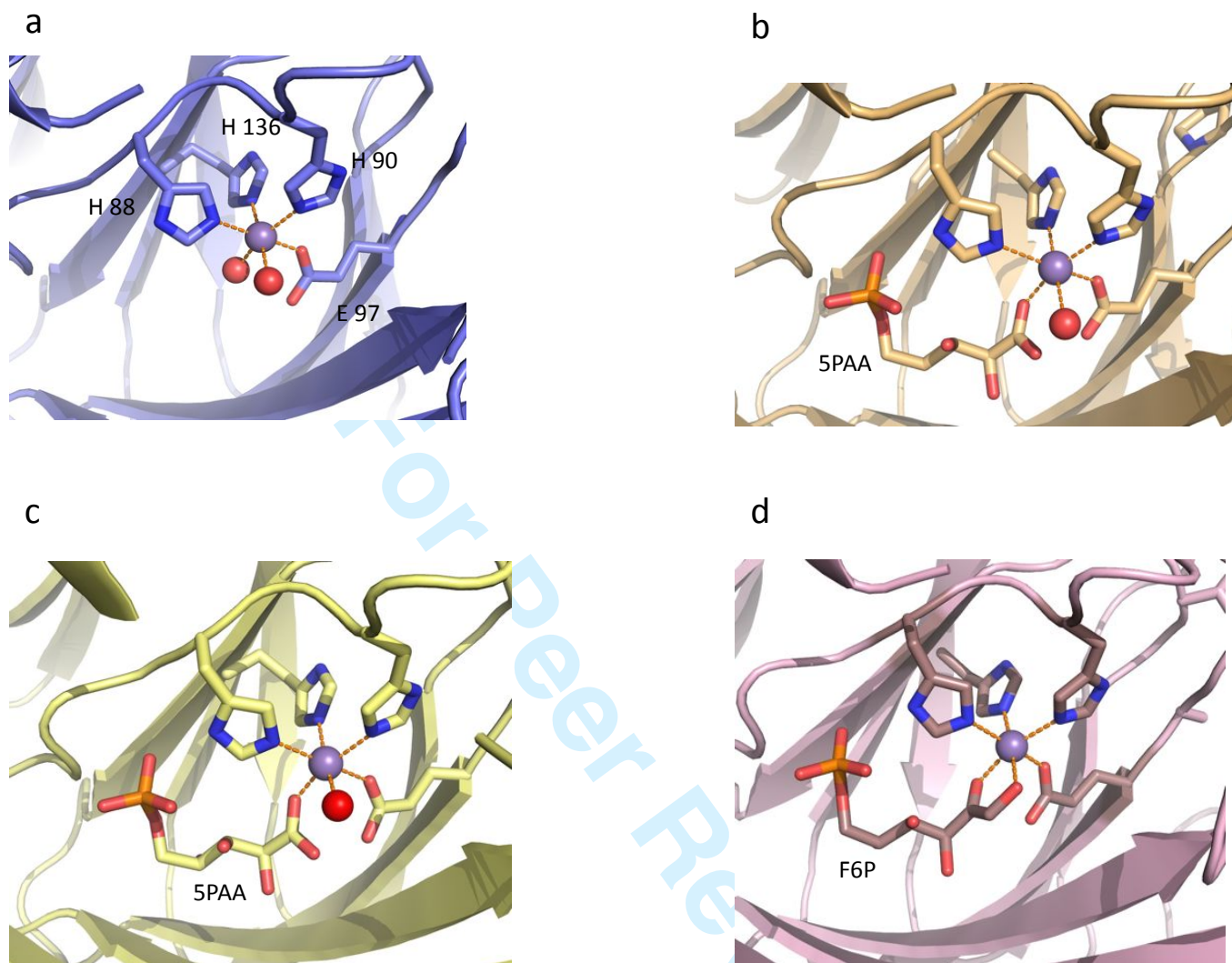
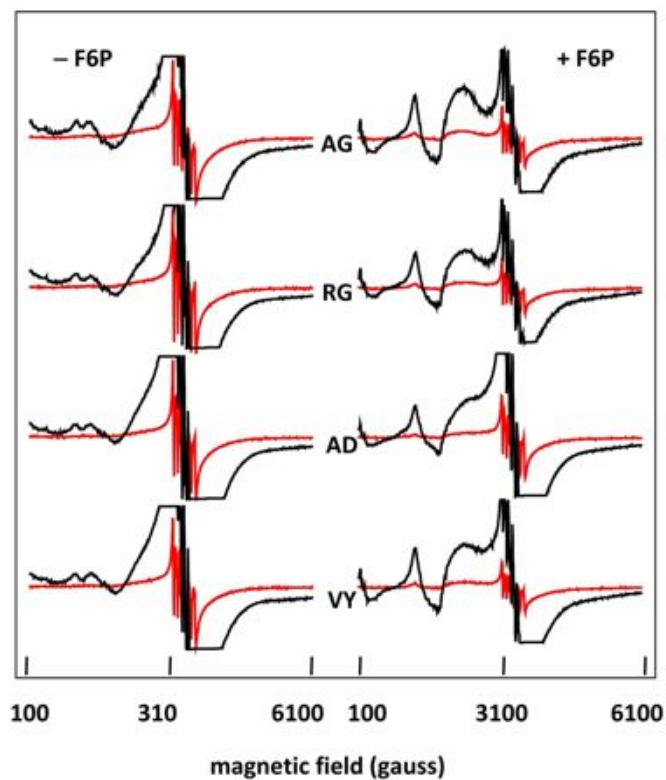


Figure 8.



1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60